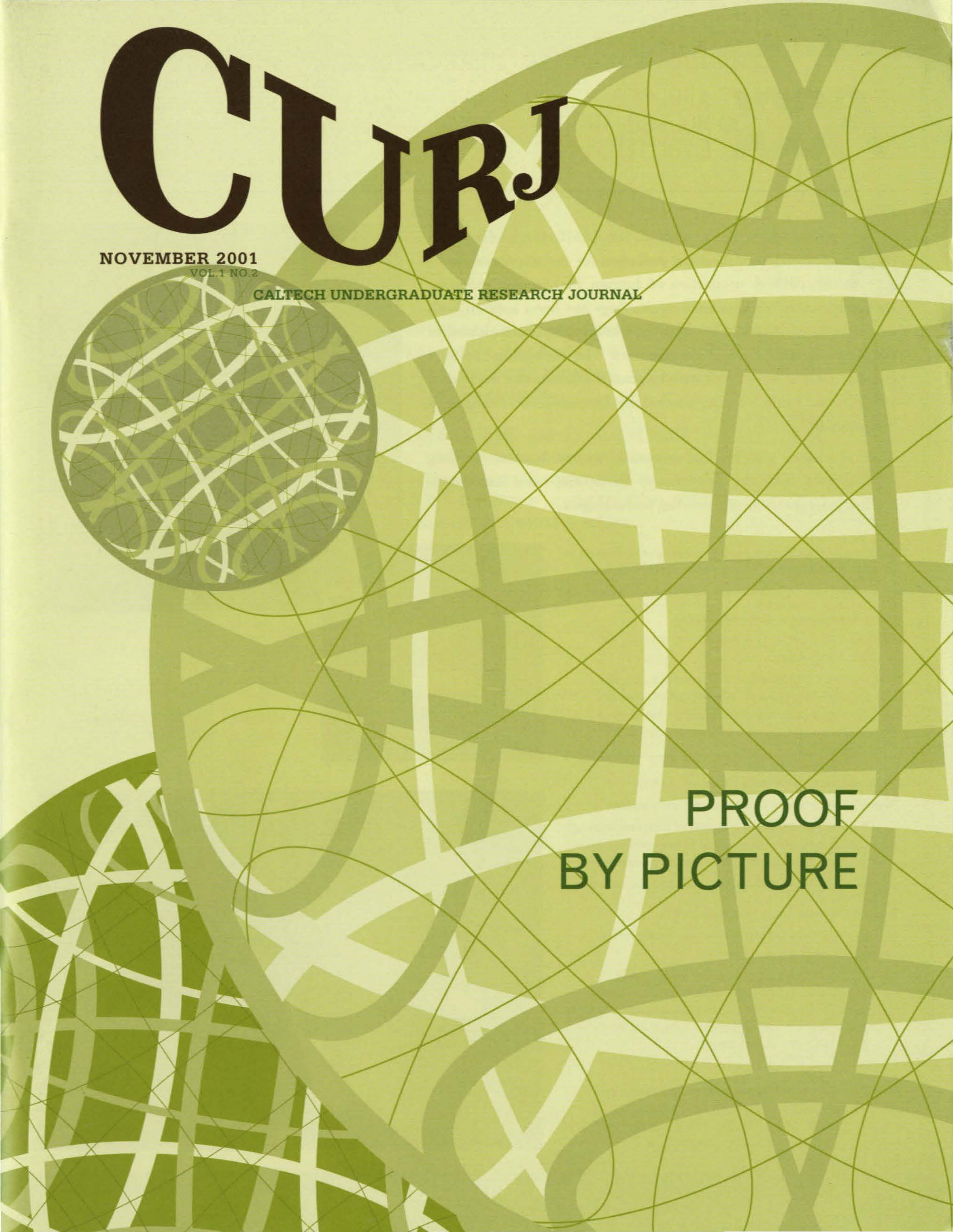


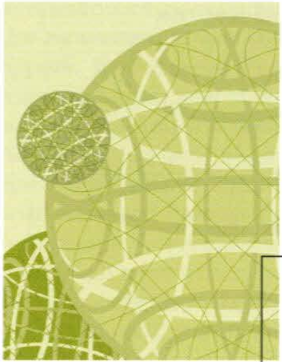
CURJ

NOVEMBER 2001
VOL.1 NO.2

CALTECH UNDERGRADUATE RESEARCH JOURNAL

PROOF BY PICTURE

CURJ



LETTERS

REVIEWS

RESEARCH

FINIS

Ap *Applied Physics*
B *Biology*
Ec *Economics*
Gc *Geochemistry*
Me *Mechanical Engineering*
M *Mathematics*

FROM THE EDITOR

In the wake of recent terrorist attacks, attention to national defense overshadowed economic, social, and technological interests. At the newsstands, war coverage blocked out most pages of the major papers. In Congress, Republican and Democratic leadership allowed unprecedented loss of civil liberties when they passed the Patriot Act surveillance and anti-terrorism bill. Amidst the drama of military strikes, America's domestic wounds festered. The economy had been in decline since the last presidential election and continued its descent even with renewed defense spending. Mismanagement of energy resources, reduction of government funding for alternative energy research, environmental destruction, wage disparity, and a decrepit public education system were forgotten maladies.

Health still suffers neglect. Anthrax has claimed four lives and poses a threat of unknown proportions, consequently displacing attention away from the millions of Americans that suffer from pressing concerns like unaffordable prescription drugs and dysfunctional healthcare. If anthrax has replaced cancer, drug abuse, and AIDS as the nation's foremost health concern, then neither our healthcare system nor the Patriot Act will fix it. With bioterrorism now a permanent threat, the anthrax problem will need systematic integration into a repaired healthcare system rather than ad hoc emergency-response measures.

At last, the tide is turning. With increasing public demand to attend to matters outside of anthrax and Afghanistan, supporters of the status quo can no longer hope that the federal government's attack on terrorism will remain a blinding diversion. Waking from the trance of defense, businessmen recall that the danger of economic depression and layoffs is as serious for the country's future as the potential fallout of a biological war. Scientists plow forward with research, remembering that their discoveries today will rifle through society for years to come. America's specialized agencies are dealing with terrorism; the nation no longer needs to focus exclusively on that issue.

Moving beyond its posture of self-defense, America averts obsolescence. With national security still unsettled, we cannot forget terrorism, but a transition is evolving to a time in which counter-terrorism is only one of many important issues. In reconsidering national priorities, we have not sacrificed our age-old commitment to economic and intellectual prosperity for a defense issue that has no short-term solution. Our journal also reaffirms its commitment to scientific communication and social policy with each letter, review, and research article. By remembering its values, America steers clear of Afghanistan's folly, where obsession with war has halted social and economic progress for centuries.

Ram Srinivasan

Ram Srinivasan
Editor-in-Chief

COVER IMAGE

Surfaces Marta S. Kaltreider

CURJ DONORS

Caltech President's Office
Caltech Engineering and Applied Sciences Division
Typecraft, Inc.
Associated Students of Caltech

Special thanks to:
Steve Madden, for always saying the right thing at the right moment.

CALTECH UNDERGRADUATE RESEARCH JOURNAL VOL.I NO.II

www.curj.caltech.edu

HELP WANTED: FUTURE AMERICAN PIONEERS

BY DANIEL S. GOLDIN

If America is to maintain its leadership role in the global and competitive environment, we must sustain our high-tech industry. This industry drives the economic machinery in America that gives us the unprecedented quality of life and national security that we sometimes take for granted.

In the twenty-first century we must have bright young minds to lead us, as we did in the 1960s, when many students were driven by their desire to go to the Moon and embraced the Space Frontier.

That generation will be retiring soon and I don't see our best and brightest young people entering the field to meet the incredible demand for their services. Who is worrying about America 20 to 30 years from now? Some of our most talented young people are not going into science and technology, threatening the very essence of our productivity. The demand is up, but the interest is down. In the end, this will not be good for America.

For the last nine years I have been privileged to be the leader of one of the country's premiere research and development agencies, the National Aeronautics and Space Administration. Before that I spent over 25 years in the aerospace industry. During this time I have seen the rise of the space age, the emergence of microelectronics and most recently the impact of information technology. However, in the last several years I have seen signs that our country's technological superiority is at risk.

To confirm our leadership position we must renew interest in science and engineering among the younger generations in K-12 and those just entering colleges, and in our current workforce. We have many exciting challenges—energy, health care, communications, transportation, space exploration, national defense—that are attracting some of our country's brightest minds, but overall our production of scientists and engineers is decreasing.

This is not just a problem for NASA, not just a problem for the aerospace industry, but for the country as a whole. Over the past two decades, the U. S. college-age population declined by more than 21%, from 21.6 million in 1980 to 17 million last year. The trend of decade-long declining enrollment in degrees in several fields of natural science and engineering reflects this demographic situation.

I got a glimpse of the future a little more than a year ago, when the Mars Polar Lander crashed onto the planet's surface and the Mars Climate Orbiter got lost in space. We had very dedicated and intelligent people, but we did not have enough of them and they did not have all the experience they needed.

This was a wake-up call, and I took steps to secure NASA's future by assuring adequate training and mentoring for our younger minds. However, the "graying" and subsequent retirement of our most experienced engineers and scientists is accelerating, not slowing. Unless we have a large increase in degree production and work the supply side of scientists and engineers—not just the demand side—the average age of the engineering workforce in America will continue to rise. About 25% of the current science and engineering workforce are more than 50 years old, and at NASA, we have twice as many people over 60 than under 30.

What is the demographic composition of the workforce? Women make up 50% of the total workforce but comprise only 23% of the scientists and engineers. The situation is worse in some specific areas. For example, only 9% of engineers are women and only 6% of aerospace engineers are women. If we can convince young women over the next decade to become engineers and scientists in representative numbers, we will have almost half the problem solved.

African Americans, Hispanics, and Native Americans comprise only 7% of the total science and engineering workforce, although they make up 24% of the population and, demographically, that number is rising. Additionally, these groups tend to be concentrated in social science areas, not the hard sciences. They comprise less than 3% of the workforce in the physical and life sciences and in engineering.

These trends are getting worse. Between 1996 and 1997, the number of first-year graduate enrollments of African Americans in the science and engineering fields dropped by 20%. The number of Hispanics entering graduate school in science and engineering declined by 18% during the same period. Moreover, minorities receive fewer than 3% of the doctorates annually in all of the natural sciences, mathematics, and engineering. Incredibly, since 1974, only 25 African-American women have received doctorates in physics. We can no longer ignore these facts.

Highly educated scientists and engineers are part of the answer, but so is an enlightened investment community that sees our future clearly. Our horizon is focused on quarters in the fiscal year, but when I ask who's responsible for America 20 years from now, not many hands go up. In America, we continue to swim upstream.

Daniel S. Goldin is the longest-serving Administrator of the National Aeronautics and Space Administration. Before joining NASA in 1992, he spent 25 years at the TRW Space and Technology Group.

OPEN DAY AT IVORY TOWERS?

BY PAUL WYMER

The Wellcome Trust, the United Kingdom's largest charity funding medical research, recently commissioned the Office for Public Management to explore ways of improving the relationship between science and the public. Given the Trust's sphere of influence, the nature of this exercise raises questions about efforts in general to "sell science" to the public.

The groundwork for the Office's analysis has already been done. A group of experts from a wide range of key organizations attended a workshop at the Trust that involved a complex process leading to alternative scenarios of what the future relationship between science and the public might be like. Members of the Public Understanding of Science cartel are currently invited to comment on these scenarios. Once amended, the scenarios will be the basis for "an open behavioral simulation of the new relationships between science and the public."

Dr. Mike Dexter, the present director of the Wellcome Trust, remarked at the beginning of his tenure that he thought current efforts at reaching the public were misdirected. "We're communicating with the chattering classes but not the public," he said. Coincidentally, this was precisely the comment I made to the Trust's senior management a few years earlier during my time as their Head of Communication and Education. I wasn't particularly surprised to be ignored, but judging by the Trust's latest approach to the problem, Dexter is suffering the same fate.

Paradoxically, the Public Understanding of Science community has in many cases produced a wedge between the two elements it is trying to marry together. Its approach is characterized by the membership of the flagship Committee on the Public Understanding of Science (COPUS). This boasts various dignitaries and professors but, perversely, the "public" barely gets a look in. While many innovative projects have been developed, the movement itself is largely self-serving, incoherent and pathologically introspective. However fashionable in academic and "edutainment" circles, "public understanding of science" is a largely ineffective idiom, which is perhaps why a recent Parliamentary Select Committee Report called for a new identity for this disparate sphere of activity.

Name change or not, there seems a significant chance that the plot will continue to be lost. Rather than attempting to learn from previous activities, the Wellcome Trust seems intent on an increasingly detached approach to the problem. This is particularly ironic when the Trust is probably the only organization that could, if it chose to, make a real difference to the public's perception of science.

What lessons can be learned from the track record? Confusion about aims and objectives must be near the top of the list. The interpretations of "public," "understanding," and "science" have been debated interminably, but to no apparent conclusion. As far as science is concerned, it seems unlikely that we are dealing with "the greatest, most beautiful and enlightening achievement of the human spirit" that Karl Popper talked about. As much is acknowledged by *Excellence and Opportunity*, a recent government report on science and innovation. At least this proposal makes no bones about it; members of the public are consumers and science is simply a means of creating new products. If nothing else, this provides focus, but it would seem to take the skids from underneath the larger part of the Public Understanding of Science movement at a stroke.

The vagueness of the term "understanding" is emphasised by the number of alternatives that have been suggested to replace it. While these have shown an encouraging evolution from "appreciation" and "engagement" to "consultation" and even "dialogue," there is still an underlying tone of condescension. This is inescapable when the public is apparently defined as anyone who is not a scientist. As perhaps one would expect of scientists, several models have been put forward as to how this amorphous mass may be classified and thereby managed. Similarly, attempts at consultation and dialogue, characterized by focus groups and citizens panels, sometimes seem analogous to experiments with laboratory animals.

Since scientists themselves are inescapably members of the public, the dichotomy seems unhelpful. Conservation groups such as Greenpeace or Friends of the Earth wouldn't dream of dissociating themselves from the public. Regardless of whether one agrees with them or not, they are very much "by the people for the people." This stance, coupled with highly competent PR and media skills, has doubtless led to the ascendancy of these groups at the expense of science in the public consciousness.

To their credit, the Wellcome Trust and other science funding agencies have assiduously pursued the media in recent years as a means of conveying science in a form that is accessible to the public. Scientists are encouraged to issue press releases at every opportunity and given media training in order to do so effectively. This has certainly resulted in an increase in column inches for science, but given the nature of the media, stories have often been sensationalized or have concentrated on contentious issues.

In summary, the openness and transparency in science policy promised by the British Government in the recent government report will require more than cosmetic changes to the way in which science is communicated. It will require the opportunity for the public to be directly and demonstrably involved in policy making and the way in which money is allocated. Clearly this is easier said than done, but the Wellcome Trust is probably in the best position to initiate the process. Public admittance to the funding panels would be as healthy a move for science as the admittance of women to the Long Room was for cricket. Unfortunately, adopting this as a scenario for open behavioral simulations of the new relationships between science and the public might be sailing a little too close to the wind.

Paul Wymer was head of Communication and Education at the Wellcome Trust from 1990 to 1997. He is currently an Education Advisor in the British National Health Service and a teacher in London.

OUTBREAK:
THE EBOLA VIRUS

BY R. AARON ROBISON

IN JULY OF 1976, A SUDANESE STOREKEEPER BECAME THE FIRST recorded victim of a terrifying new hemorrhagic fever that emerged from its lair in the rain forests of central Africa. The virus that killed him eventually became known as Ebola, named after a river in what was then Zaire, where one of the first large-scale epidemics occurred in that same year. Those epidemics, which occurred simultaneously in Sudan and Zaire, eventually resulted in at least 550 reported cases and over 330 deaths, with a fatality rate of nearly 90% in Zaire. Sporadic outbreaks have been recurring since that time, with another severe epidemic in Kikwit, Zaire in 1995. In September of 2000, an epidemic broke out in Uganda that killed 173 people (see Figure 1). Earlier this year in Toronto, a woman traveling from Africa started exhibiting Ebola-like symptoms, sparking a brief media bonanza before she was finally diagnosed with a form of malaria.

The event that brought Ebola into the minds of most Americans took place in a monkey facility in Reston, Virginia, just ten miles west of Washington, D. C. (see Figure 2). It was there, in the latter part of 1989, that some of the facility's simian inhabitants began showing signs of hemorrhagic fever. Those in charge became alarmed when more and more of the monkeys began to die of the inexplicable illness. Panic set in when the virus responsible was discovered to be a strain of the same Ebola virus that had killed hundreds of people in Sudan and Zaire. It was quickly discovered that the Reston strain could not infect humans, but the scare left the nation with a haunting awareness of the vulnerability of America's urban centers to a deadly and highly contagious pathogen.

"Ebola may accomplish in a matter of days what HIV does to victims over the course of a decade."

FIGURE 1. Suspected Ebola patients wait anxiously in an isolation ward of Lacor Hospital in the Gulu region of Uganda during an Ebola epidemic in late 2000. In the space of three months, there were more than 400 cases of Ebola hemorrhagic fever in Uganda, and over half were fatal.
Source: Associated Press, 2000.

FIGURE 2. Image from the movie Outbreak, loosely based on a 1989 incident at a monkey facility in Reston, Virginia where monkeys mysteriously began dying from hemorrhagic fever. The cause was later discovered to be a new form of Ebola, luckily one that could not infect humans.
Source: Motion picture Outbreak, 1995.

FEELING THE SYMPTOMS

The virus responsible for wreaking this havoc and inspiring such fear is a member of the *filoviridae* family of viruses, a distinction shared only by the Marburg virus, a similar hemorrhagic fever virus originating in Africa. These viruses are so named for their characteristic long, filamentous appearance when viewed at high magnification (see Figure 3). Each virus holds one molecule of RNA, a single-stranded relative of the double-stranded DNA that composes our genes. A shell of protein surrounds this strand of RNA that encodes all of the virus' genes. These proteins are enclosed in a membrane from which protrude glycoproteins, a type of viral protein that accounts for the virus' ability to enter host cells.

Once a person contracts the virus, three to eighteen days pass before symptoms start. When they do begin, the symptoms come on abruptly, quickly progressing beyond the initial headache and fever to more serious symptoms by the third day. In fatal cases, the sick individual begins to bleed internally within one week. Once this hemorrhaging begins, the condition of the patient deteriorates rapidly. Bleeding escalates and the victim's organs begin literally to disintegrate. Finally, as a reaction to the severe loss of blood, the individual goes into hypotensive shock and dies. This aspect of the illness is the most horrifying because the contents of the victim's body appear to drain from every orifice, even the eyes. If the victim survives, the recovery period is extensive, and serious complications can arise in the infection's wake, including arthritis, vision and hearing loss, and conjunctivitis. In Africa, there is also an intense social stigma associated with Ebola infection, and the survivors are often traumatized by the social repercussions.

"Even though the original cases have been identified in almost every Ebola epidemic, we are still no closer to finding the reservoir."

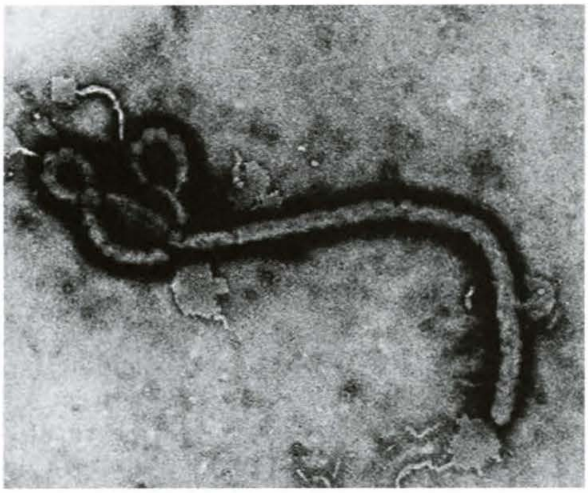


FIGURE 3. This photo, taken in 1976 using an electron micrograph at 160,000 times magnification, was the first ever of the Ebola virus. The virus is long and filamentous, giving its family the name *filoviridae*. The outer shell of the virus consists of a membrane studded with proteins. A protein shell surrounds the inner core of RNA.
Source: F. A. Murphy.

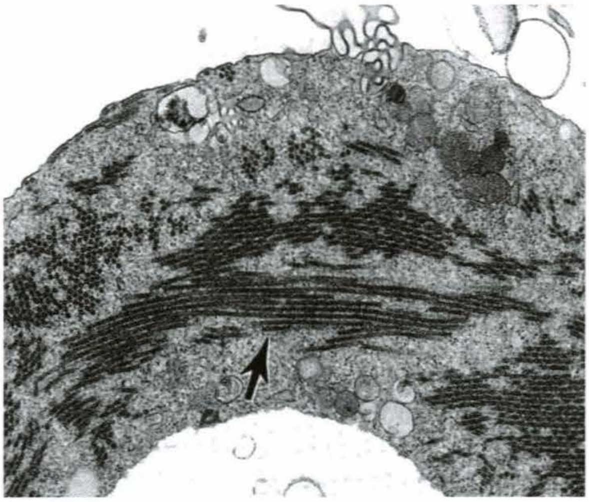


FIGURE 4. The Ebola virus forms inside an infected cell. An arrow marks the creation of viral filaments. An Ebola-infected cell becomes a virus factory producing thousands of new viruses that bud off and infect other cells. A single drop of infected blood can contain millions of viruses.
Source: S. R. Zaki, C. S. Goldsmith. Pathological Features of Filovirus Infections in Humans. Current Topics in Microbiology and Immunology 235, 97-116 (1999).

REPLICATE AND ATTACK

The course of Ebola infection has provided a bewildering puzzle for scientists, who are only beginning to understand the precise nature of the effect the virus has on our bodies. The entry and growth of the virus in human cells appears to be rather typical. Once the virus comes in contact with the surface of the cell, the viral proteins present on the virus surface interact with receptors on the surface of the cell. This causes the cell's outer membrane to fold inward and form a cavity with the virus inside, enabling the contents of the virus to be released into the cell. The virus disrupts the host cell's normal functions and commandeers the cellular machinery to begin making viral proteins and copies of the viral genome. As the infection proceeds, new virus structures are formed and appear as long filaments in the infected cell (see Figure 4). These viruses eventually bud from the outer membrane of the cell, which the virus incorporates as its own outer membrane. The newly unleashed virus then proceeds to diffuse and rapidly infects neighboring cells.

Unlike most viruses, which tend to be restricted to particular types of cells or a particular anatomical region, the Ebola virus appears to infect a broad range of host cells, such that the infection often pervades the body. Nevertheless, examinations of the bodies of deceased Ebola victims and laboratory animals infected with the virus suggest that the kidneys, liver, and spleen, all essential elements of the immune system, may be particularly targeted. The massive tissue death observed in these organs indicates that they are centers of viral replication. Ebola's presence in these organs suggests an attempt to defeat the immune system.

There are other indications that Ebola may do just that, prompting some scientists to remark that Ebola can accomplish in a matter of days what HIV does to its victims over the course of a decade. Soon after the virus was first uncovered, scientists hypothesized that such immunosuppressive activity did occur, but they did not understand how. One study compared the immune response over time in fatal cases to those that recovered. They discovered a marked difference in the immune activity between the two groups. Surviving patients exhibited high levels of antibody early on in infection and those levels increased over time. Antibodies are molecules produced by the immune system that help to clear viruses from the body. The antibodies found in patients who survived were able to bind to Ebola virus proteins, indicative of an effective advanced immune response. These patients also had significantly increased activity of killer T-cells, which destroy infected cells. In contrast, patients with fatal cases showed virtually no expression of antibodies specific to Ebola at any point during infection, and while killer T-cells showed early activity, these cells rapidly disappeared as infection progressed. This suggests a race early in infection between the immune system and the viral replication. If the immune system does not respond quickly enough, it is overtaken and destroyed.

How the virus interacts with the immune system is still not fully understood, but some of the factors contributing to the immunosuppressive effect are known. One important feature is the virus' ability to produce a soluble form of the protein that is normally found on its surface and secrete it in enormous quantities from infected cells. Recent discoveries indicate that this protein selectively binds to and inactivates neutrophils, a type of immune cell that is essential for the primary inflammatory immune response. This is the activity normally responsible for the redness and swelling we associate with infections, and helps to activate other branches of the immune system.

In addition, the surface protein on the virus appears to have an affinity for the cells lining the blood vessels, which have an important role in signaling responses by the immune system. Viral infection of these cells has been shown to interfere with this signaling process and disrupt the cells' ability to counteract the infection. A specific portion of this protein is also known to inhibit the proliferation of immune cells. However, while this could permit secondary infections in Ebola patients, we still have no evidence that this happens.

SEARCHING FOR THE VIRAL SOURCE

Ebola infections are short and deadly. Thus the virus cannot maintain itself in a human population without quickly killing all its victims and becoming extinct. Because viruses are incapable of being outside a host for more than a brief duration, there must be a natural reservoir, another living host in which the virus normally resides without destructive side effects. When humans encounter this natural reservoir, the virus is transmitted back into the human population. However, even though the original cases supposedly responsible for this transmission have been identified in almost every Ebola epidemic, we are still no closer to finding the reservoir.

Significant efforts have been made to trace these primary cases in hopes of determining where the contact originated, a search which has taken investigators from the decaying rot of bat caves to the remains of dead monkeys. In one of the most vigorous attempts to uncover the source, over 3000 animals were taken from the region of the 1995 outbreak in Zaire and checked for signs of the virus, to no avail. The search is further complicated by the infrequency of outbreaks, suggesting that the source is one that humans rarely encounter. However, once the reservoir is discovered, efforts can be made to control contact between humans and the host to curtail epidemics altogether.

"The course of Ebola infection has provided a bewildering puzzle for scientists, who are only beginnig to understand the effect the virus has on our bodies."

RACE FOR A CURE

There have been many approaches to finding a treatment for this formidable virus. All of these investigations have been hampered by the significant danger and expense involved in working with Ebola, which is a Class 4 biological hazard. This rating, given only to the most dangerous biological agents, mandates the use of special facilities and procedures for handling the virus. The earliest and most rudimentary means of dealing with it was to treat patients in the early stages of infection with the blood serum of recovering patients, hoping that the high concentration of anti-Ebola antibodies in the hyperimmune serum would give the immune systems of the patients a fighting chance. However, when used in the chaotic early outbreaks, this method met with little success in reducing the fatality rate. Later studies in laboratory animals found that the hyperimmune serum taken from non-susceptible animals infected with the virus was highly effective at preventing infection in guinea pigs and baboons. The same treatment also seemed to help animals that were already infected. The tactic may still be effective, but it has yet to be used successfully in patients.

Antiviral drugs have had great success in slowing the growth of HIV into AIDS, and it is possible that the same type of drugs could slow Ebola as well. Ebola researchers at the Army Medical Research Institute of Infectious Diseases in Fort Detrick, Maryland discovered a drug that is effective at inhibiting a protein the virus needs to grow. Inhibiting the mechanism by which this protein attaches sugar groups to viral proteins slows viral replication. Unfortunately, as with most antiviral drugs, the drug itself is toxic and cannot safely be administered except in minute quantities. The compromised livers of Ebola victims, increasingly susceptible to toxins, further aggravate this problem. However, below toxic levels, the drug was effective in treating mice infected with lethal doses of the virus: 100% of the mice recovered when treatment commenced within one day of the initial infection. When treatment was not started until three days after infection, the recovery rate was reduced to 40%, possibly because victims' compromised systems were more susceptible to the toxicity of the compound by that point. The drug has yet to be successfully tested in monkeys, let alone in humans. While it has shown promise in the laboratory setting, its viability as therapy for humans is uncertain.

VACCINES TO THE RESCUE

Vaccines have existed in various forms for centuries, long before the viruses that cause disease were at all understood. Many modern vaccines are made from killed or attenuated virus, which are either dead virus particles or a form of the virus that has been passed for many generations through a non-human organism, such as mice or chickens, until it is no longer capable of causing a serious infection in humans. Unfortunately, neither of these vaccine types is suitable for combating Ebola in humans because of the risk of exposing individuals to a potentially infectious virus.

The alternative is to create a vaccine that incorporates only a small piece of the virus, sufficient to raise immunity, but insufficient to enable a complete virus to form. Tests in guinea pigs found that when circular pieces of DNA (plasmid cDNA) that contained the genes for three Ebola viral proteins were injected into the animals, they were protected against a lethal dose of the virus. Last year, using a similar scheme, a group from the Vaccine Research Center at the National Institutes of Health in Bethesda, Maryland and the Centers for Disease Control in Atlanta, Georgia injected monkeys with Ebola cDNA and an attenuated adenovirus. The adenovirus was a harmless strain of the virus responsible for the common cold, modified to incorporate Ebola proteins in its genome. The DNA vaccine is traditionally more effective at producing cellular immunity, whereas the adenovirus also induces production of protective antibodies. All four monkeys immunized with this vaccine and subsequently injected with the virus survived. In three monkeys, the response was so effective that there was no evidence of any viral replication. All non-immunized monkeys died.

While a vaccine offers great promise in the battle against Ebola, there are still a number of concerns surrounding its use, and a vaccine is not a simple end-all to the problem of Ebola. In the vaccine study mentioned above, the quantity of virus injected into each monkey was miniscule, and the vaccine may not be effective against larger initial infections. In addition, the vaccine only protects against one of four subtypes of the virus that are dangerous to humans. In order to be effective, the vaccine would need to protect against all four strains. Once a successful vaccine exists, there is also doubt as to how it can be administered. It is economically unfeasible to create a vaccine program large enough to immunize anyone who has a chance of coming in contact with the virus, and attempts to limit the group to those in contact with the natural reservoir are not possible until the reservoir has been identified.

THE VIRUS IN PERSPECTIVE

The Ebola virus has attracted global attention and inspired both awe and dread. However, while millions of people have been infected with HIV or chronic malaria, approximately one thousand people have died from Ebola since its discovery in 1976. Some civic leaders have questioned the prudence of expending resources on a disease that affects so few people. Even though the current scope of the virus may be relatively small, there is a possibility that a new form of Ebola may arise that is not as limited in extent as its predecessors. Current research will then have significant implications. The current duel with Ebola reflects an increasingly frequent theme in medical science, where with each new disease of increasing complexity and diversity, our ingenuity is tested as scientists and members of a global health community. C

R. Aaron Robison is a fourth year undergraduate in Biology at the California Institute of Technology. *The author wishes to thank David Baltimore, President and Professor of Biology at Caltech.*

FURTHER READING

1 H. D. Klenk. Marburg and Ebola Viruses. *Current Topics in Microbiology and Immunology* **235**, 1-225 (1999).

2 C. J. Peters, J. W. LeDuc. Ebola: The Virus and the Disease. *Journal of Infectious Diseases* **179 (S1)**, IX-XVI (1999).

3 C. J. Peters *et al.* Filoviridae: Marburg and Ebola Viruses. In *Fields Virology*. B. N. Fields *et al.* (Lippincott-Raven, Philadelphia, 1996), pp. 1161-1176.

4 N. J. Sullivan, A. Sanchez, P. E. Rollin, Z. Y. Yang, G. J. Nabel. Development of a Preventive Vaccine for Ebola Virus Infection in Primates. *Nature* **408**, 605-607 (2000).

RACING THROUGH WATER: SUPERCAVITATION

BY VICTORIA STURGEON

LAST DECEMBER, AN AMERICAN BUSINESSman, sick with cancer, was sentenced to twenty years of hard labor in a Russian prison. The businessman, Edmund Pope, was charged with attempting to purchase the propulsion plans for the Shkval, a Russian torpedo which some western military experts believe was responsible for the sinking of the Kursk submarine last year. Pope's trial was conducted under utmost secrecy; no Americans were ever permitted to see the charges against him. Even his Russian defense team was forbidden from seeing the information upon which the prosecution based its case.

The reason for all this secrecy was the Shkval torpedo, an underwater missile that shatters speed records by using a newly discovered phenomenon known as supercavitation.

WATER AND DRAG

Any child diving off the board at her neighborhood swimming pool knows the effect that water has on a fast-moving object. A fall from twenty feet onto a hard surface—like the bottom of a pool—would break a child's leg at the very least, but a ten-year-old might jump off the high dive a dozen times on any warm summer afternoon. As soon as she breaks the surface of the water, her velocity decreases, bringing her to a near stop before she hits the bottom of the deep end.

Similarly, when a projectile is fired into the ocean, because of the strong drag force, the projectile decelerates and loses virtually all of its forward momentum in only a few thousand feet. Older torpedoes, in fact, were reusable, floating to the surface and stopping if they missed their targets.

Drag affects any object moving through water. Competitive swimmers know the difference that even a small reduction in drag can have on performance in a race. The locker room at a major high school swim meet is full of young men shaving their legs, arms, and heads in the hopes of cutting another second off their time.

Although the basic concepts powering them are strikingly similar, drag slows submarine travel to well below airplane speeds. Water limits even nature's strategies, and the fastest bird moves twice as quickly as the fastest fish. Drag imposes significant constraints on marine engineering that ultimately limit the speed and range of submarines, ships, and projectiles alike.

First explored in the 1940s, supercavitation exploits a loophole that allows underwater travel with minimal drag. For many years, scientists and naval experts studied its parent field, cavitation, because of the *problems* that it brings about. Only recently did researchers consider supercavitation as a way to build faster submarines and torpedoes.

AVOIDING CAVITATION

The same principle that keeps an airplane airborne causes cavitation around ship propellers: as velocity increases, pressure decreases. When a ship's propeller spins underwater, the blades drag liquid around with them. The pressure of this fast moving water drops, sometimes drastically. Water's state depends upon its pressure just as much as it does on its temperature, which is why water boils at lower temperatures at high altitude. Lowering a liquid's pressure below a certain point, called its vapor pressure, will cause it to boil away as if it were being heated over a stove.

If a propeller spins fast enough, the surrounding liquid will speed up, drop in pressure, and boil away. At first, the physical characteristics of boiling and cavitation are almost identical. Both involve the formation of small vapor-filled spherical bubbles that gradually increase in size. However, the bubbles produced by the two processes end in very different manners. In boiling, bubbles are stable: the hot gas inside either escapes to the surface or releases its heat to the surrounding liquid. In the latter case, the bubble does not collapse, but instead fills with fluid as the gas inside condenses.

The process is different in cavitation. Cavitation bubbles depend upon the low pressure of the surrounding fluid to survive. As the pressure of the surrounding liquid increases, the cavity suddenly collapses—a centimeter-sized cavity disappears in milliseconds. Cavities implode violently and create shock waves that dig pits in exposed metal, scarring propeller blades and pipes. Engineers around the world strive to minimize cavitation damage by constructing sluiceways, pipes, and channels that control the pressure and velocity of the passing liquid to eliminate cavitation.

When it acts upon propellers, cavitation not only causes damage but also decreases efficiency. The same decrease in water pressure that causes cavitation also reduces the force that the water can exert against the boat, causing the propeller blades to "race" and spin ineffectively. When a propeller induces significant cavitation, it is pushing against a combination of liquid water and water vapor. Since water vapor is much less

"Supercavitation exploits a loophole that allows underwater travel with minimal drag."

dense than liquid water, the propeller can exert much less force against the water vapor bubbles. With the problems it causes, it is no wonder maritime engineers try to avoid cavitation. Varying the shape, pitch, material, and placement of propellers helps reduce, but not eliminate, the damage.

THE EXCEPTION TO THE RULE

Recently, however, scientists and engineers have developed an entirely new solution to the cavitation problem. For ships traveling faster than 60 miles per hour, propeller-induced cavitation is unavoidable. Supercavitation offers a solution.

In supercavitation, the small gas bubbles produced by cavitation expand and combine to form one large, stable, and predictable bubble around the supercavitating object. The bubble is longer than the object, so only the leading edge of the object actually contacts liquid water. The rest of the object is surrounded by low-pressure water vapor, significantly lowering the drag on the supercavitating object. Modern propellers intentionally induce supercavitation to reap the benefits of lower drag.

A supercavity can also form around a specially designed projectile. The key is creating a zone of low pressure around the entire object by carefully shaping the nose and firing the projectile at a sufficiently high velocity. At high velocity, water flows off the edge of the nose with a speed and angle that prevent it from wrapping around the surface of the projectile, producing a low-pressure bubble around the object (see Figure 1). With an appropriate nose shape and a speed over 110 miles per hour, the entire projectile may reside in a vapor cavity.

Since drag is proportional to the density of the surrounding fluid, the drag on a supercavitating projectile is dramatically reduced,

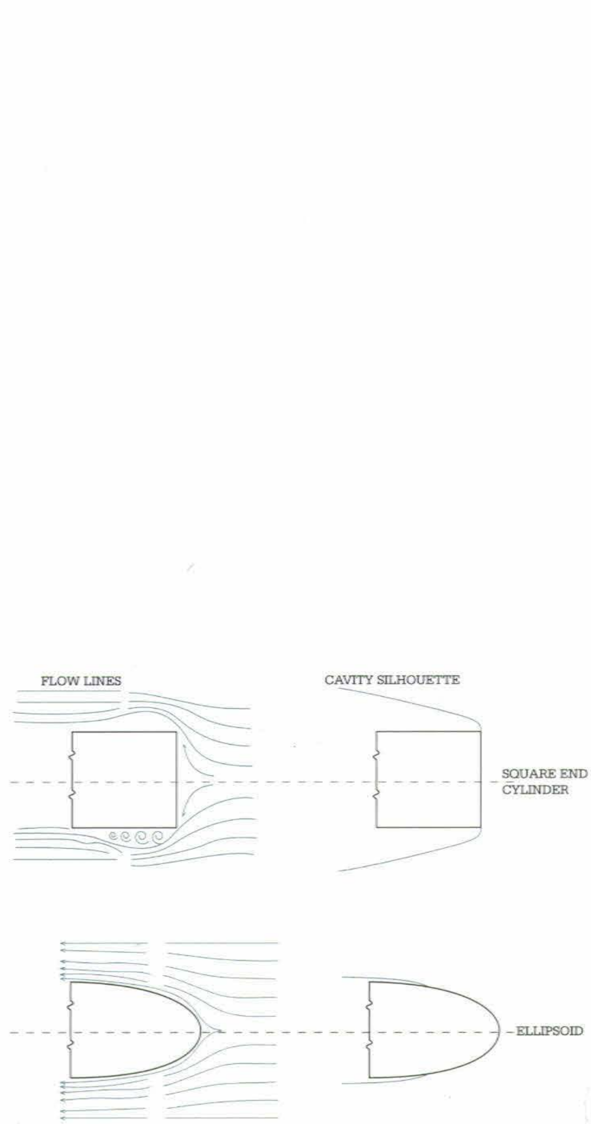


FIGURE 1. Nose shape affects fluid flow around projectiles, so engineers must design them carefully. An appropriate nose shape and enough speed can create a vapor cavity that encloses the entire projectile.
Source: J. W. Daly, Ph.D. thesis, Aerodynamics, California Institute of Technology (1945).

allowing supercavitating projectiles to attain higher speeds than conventional projectiles. In water, a rough approximation predicts that a supercavitating projectile has 200,000 times less skin friction than a normal projectile. The potential applications are impressive.

LIFE OF A CAVITY

Supercavities are classified as one of two types: vapor or ventilated. Vapor cavities are the pure type of supercavity, formed only by the combination of a number of smaller cavities. In a ventilated cavity, however, gases are released into the bubble by the supercavitating object or a nearby water surface (see Figure 2). For example, a torpedo might release its exhaust gases into its supercavity, or a projectile dropped into the ocean might take air down with it. Although gases increase the size of ventilated cavities, additional complications, like gas elasticity and leakage rates, result in a less stable cavity that fluctuates in size and shape. In other respects, ventilated cavities and vapor cavities are indistinguishable and subject to the same principles.

The cavitation number, the dimensionless quantity K, predicts the behavior of a supercavity. The number is a function of the pressure difference between the cavity and the surrounding water, the density of the surrounding fluid, and the velocity of the object. Physically, the cavitation number is a measure of the instability of a cavity. For small K, cavitation begins. As K decreases, the cavity increases in size and stability (see Figure 3). For supercavities, the value of K and the shape of the object's nose predict the shape and size of the vapor cavity, so monitoring the cavitation number of a projectile tracks the status of its cavity.

Researchers have studied the life cycle of a projectile-induced supercavity in detail, from

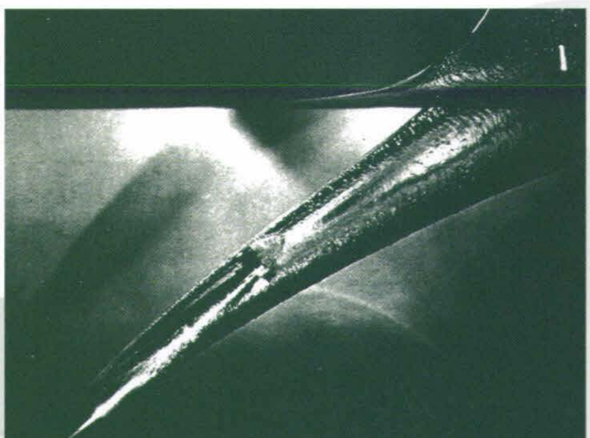


FIGURE 2. A projectile dropped into water pulls air down with it to create a ventilated supercavity. The supercavity is less stable than predicted by its size because the trapped air causes the cavity to fluctuate. Once the air leaks out, the supercavity is maintained only if the projectile is moving fast enough to create a vapor cavity.
Source: J. G. Waugh, G. W. Stubstad. Hydroballistic Modeling (Naval Undersea Center, San Diego, 1972).

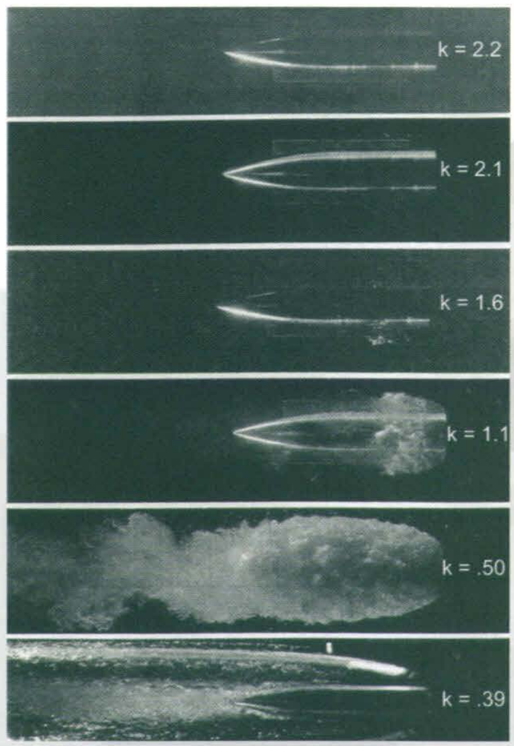


FIGURE 3. The cavitation number, calculated from physical properties of the object and fluid, describes the state of a cavity. As the cavitation number decreases for this square end cylinder, the flow over it increases and the supercavity grows larger. A larger supercavity is more stable and has less drag, allowing the projectile to travel faster. Knowing nose shape and cavitation number is enough to predict the shape of a supercavity.
Source: J. W. Daly, Ph.D. thesis, Aerodynamics, California Institute of Technology (1945).

the initial moment of supercavitation to the vapor cavity's death. Immediately after being fired, a projectile is enclosed within a vapor cavity and experiences little drag. As the projectile slows, its cavitation number increases and the size of the vapor cavity decreases until it disappears. Unlike a normal cavity, the death of a supercavity surrounding a projectile is not sudden or violent. The cavity simply shrinks around the projectile at an ever-increasing rate until the cavity no longer exists. There is little or no damage to the supercavitating object from cavity collapse, a crucial advantage over the craters left by cavitation. However, a closer look reveals some complications slowing down supercavitating torpedoes.

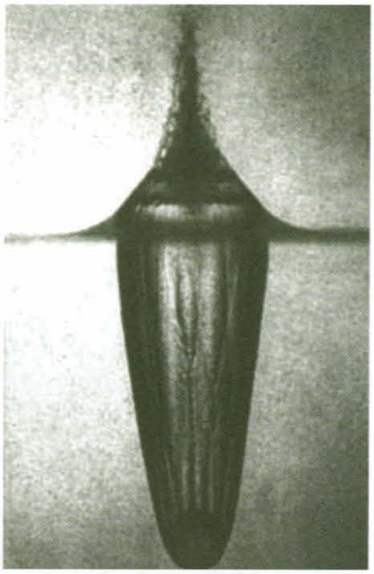


FIGURE 4. The tail end of a supercavity splashes around, banging against and damaging a projectile's tail end. The same sort of cavity splashing occurred when this sphere was dropped into water.
Source: J. G. Waugh, G. W. Stubstad. Hydroballistic Modeling (Naval Undersea Center, San Diego, 1972).

BUILDING A BETTER TORPEDO

A torpedo dropped into water draws a column of air down with it, creating a temporarily ventilated cavity that reduces drag on the torpedo. The air eventually leaks out, but if the torpedo is moving fast enough the collapsing ventilated cavity is replaced by a vapor cavity. However, the behavior of the cavity's tail end becomes a problem. The supercavity's tail end may splash violently around the projectile's rear, causing significant structural damage to control and propulsive surfaces (see Figure 4).

Projectile wobbling creates more problems. Projectiles are quite unstable within their vapor cavities. Salil Kilkarni and Rudra Pratap at the Indian Institute of Science in Bangalore, India showed that small imbalances in how a projectile enters the water cause it to wobble within its supercavity. The projectile pitches and yaws about its nose as it proceeds through the water, hitting its tail end against one side of the cavity and then another (see Figure 5). Since the surface of the vapor cavity is liquid water, these impacts increase drag on the projectile and, over time, slow the projectile. The vapor cavity shrinks (reflected by an increasing cavitation number) causing increasingly frequent impacts and thus an escalating drag force that quickly induces cavity death.

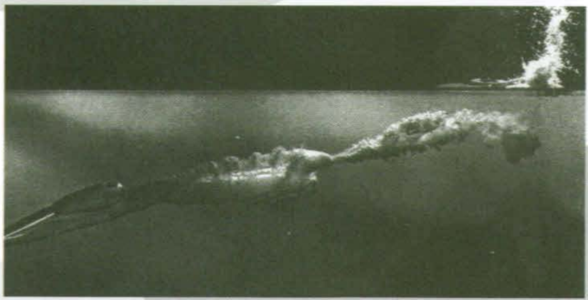


FIGURE 5. Small imbalances during water entry make a supercavitating projectile's motion extremely unstable. It pitches and yaws about its nose, slapping its tail against the sides of the supercavity. These impacts increase the drag on the object, slowing it to the point that its supercavity collapses.
Source: J. G. Waugh, G. W. Stubstad. Hydroballistic Modeling (Naval Undersea Center, San Diego, 1972).

"Supercavitating passenger submarines could **cross the Atlantic** in less than **an hour.**"

MOVING FORWARD WITH SUPERCAVITATION

Supercavitation has obvious applications. Dropped from helicopters and airplanes, new torpedoes supercavitate from the moment they enter the water. A supercavitating torpedo would contact liquid water only at its nose and could outrun any ship; the Russian Shkval torpedo has been clocked at speeds up to 230 miles per hour. Researchers at the US Naval Undersea Warfare Center were able to fire unpowered supercavitating projectiles at speeds faster than the speed of sound in water, approximately 1 mile per second!

The U. S. Navy has recently developed a supercavitating bullet to clear underwater mines. Bullets fired from a helicopter-mounted standard Gatling gun supercavitate through the water and detonate mines with little cost and even less risk to human life. As for civilian applications, supercavitating passenger submarines could cross the Atlantic in less than an hour.

However, there are significant technical problems to overcome. Foremost is the impossibility of steering a supercavitating object (although we probably wouldn't hear about it even if a solution had been found), a fact that restricts the potential utility of both torpedoes and passenger transportation. Since the only portion of a supercavitating projectile that contacts liquid water is the nose, there is no practical way to alter the trajectory without a significant loss in speed.

Additionally, as discussed above, projectile and cavity wobbling damage the tail end of the projectile. One way to reduce damage is to make the cavity larger by ventilating the tail end of the cavity with exhaust. Ventilating increases instability but protects the projectile from damage. However, even if human passenger ships could be launched safely at supercavitating speeds, the accelerations at the tail end caused by this instability would probably be too great for passengers.

An additional difficulty is size. To be economically efficient, a supercavitating submarine would need to be large enough to transport many passengers and their luggage. A high velocity is needed to maintain a large enough cavity and to provide ample lift, leading us to the problem of propulsion. Such a submarine requires a powerful source of propulsion that does not require direct contact with liquid water. A water jet might work, but that requires more power than can be presently generated under the circumstances.

FICTION BECOMES REALITY?

The Russian navy, and possibly the American navy, has apparently solved some of the technical challenges in developing supercavitating applications. However, the nature of the solutions is closely guarded, as Edmund Pope discovered last year. We can only speculate how far military technology has advanced. Supercavitation's history, now colored by espionage and intrigue, has already inspired fantastic stories of futuristic naval warfare. Squadrons of ultrafast submarines might someday clash in the oceans, abandoning traditional stealth tactics in favor of high-speed underwater acrobatics. However, although supercavitation has been widely studied since the 1940s, many questions remain unanswered. Despite what naval engineers and science fiction fans might wish, for the present, supercavitating submarines dogfighting underwater will have to wait in our imaginations. C

Victoria Sturgeon is a fourth year undergraduate in Mechanical Engineering at the California Institute of Technology. *The author wishes to thank Professor Allan Acosta, Richard L. and Dorothy M. Hayman Professor of Mechanical Engineering, Emeritus at Caltech.*

FURTHER READING

1 S. Ashley. Warp Drive Underwater. *Scientific American* **284**, 70-79 (May 2001).

2 D. Graham-Rowe. Faster than a Speeding Bullet. *New Scientist* **167**, 26-29 (July 22, 2000).

3 S. S. Kilkarni, R. Pratap. Studies on the Dynamics of a Supercavitating Projectile. *Applied Mathematical Modeling* **24**, 113-129 (2000).

4 D. Mackenzie. Seismic Shift. *Discover Magazine* **22**, 60-67 (September 2001).

PROOF BY PICTURE:
THE POINCARÉ-HOPF INDEX THEOREM FOR SURFACES

BY STEVE T. PAIK

SUPPOSE YOU ARE AN ANT LIVING ON A SPHERE. FROM YOUR PERSPECTIVE the world is two-dimensional: you can move back and forth, left and right, but not up or down. You settle into a comfortable life on the sphere when suddenly the wind starts to blow. Air currents develop and begin to circulate around the globe much the same way they do here on earth. In an effort to escape the torrid winds, you decide to journey across the desolate plains of the globe in search of a land where the air is perfectly still. Is it possible to find such a calm place? Fortunately, the answer is always yes—there will be at least one point on the sphere where the wind speed is zero.

A friend of yours, who is also an ant, lives on another world. Unlike yours, her world is not spherical but toroidal, shaped like a donut. When the weather changes for the worse, your friend searches for a place where the wind does not blow. Unfortunately, your friend will not always fare as well as you; sometimes she will be able to find a point on the torus where the wind speed is zero, but other times she will not find a place of shelter no matter how hard she looks. This may seem like the work of a mischievous god who teases your friend from time to time but always gives you a way out. A mathematician knows better. He knows that the consequences of topology are on your side.

Only the shape of the world that the ants inhabit distinguishes the two cases. This suggests that only the *topology* of their world affects the kind of air currents that can circulate. How does topology allow or prevent an ant from finding a sheltered spot? The Poincaré-Hopf Index Theorem offers an explanation.

We begin by describing surfaces in topological terms. When talking about a sphere or torus, we refer just to its outer surface and not to its inside. The surfaces of a sphere and of a torus are two-dimensional, like paper wrapped around a ball and donut respectively. They are examples of *compact surfaces* because we can enclose each in a box of finite size, and they are called *boundaryless* since an ant walking around on the surface of a sphere or torus never encounters an edge.

"The kinds of zeros act as a fingerprint for the vector field."

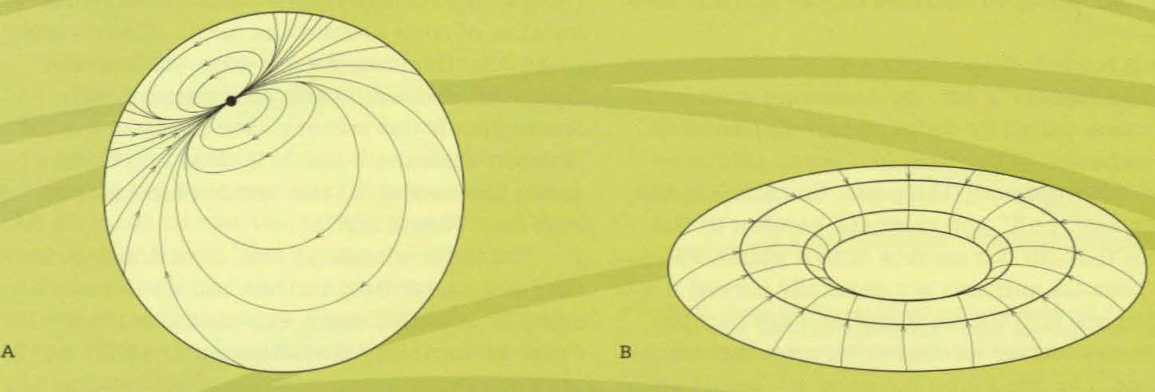


FIGURE 1. Smooth tangent vector fields are drawn on two different compact boundaryless surfaces. For the sphere (A), here is an example of a field with only one zero. (B) The torus can be easily covered with a nonzero vector field.

We assume that the wind on the surface of spherical and toroidal planets must flow tangentially, along and in contact with the surface. At each point, the wind has some magnitude and direction so it is natural to describe the weather on these surfaces by a collection of vectors tangent to the surface. This collection of tangent vectors is similar to the pattern made by iron filings near a bar magnet where the filings indicate the direction of the magnetic field lines. Our tangent vectors cover the entire surface and provide a prescription for the motion of the wind. This *tangent vector field* must also be *smooth* in the sense that the behavior of the vectors changes gradually from one point on the surface to another nearby point (see Figure 1). From here on, we will refer to smooth tangent vector fields defined on compact boundaryless surfaces simply as vector fields if we do not explicitly say otherwise. In order to talk meaningfully about these vector fields, we need a way to identify them unambiguously.

The problem is as follows. Suppose you draw a vector field on a surface. Another person can come along and ever so slightly deform a section of the field. He can continue making these tiny alterations until he has made so many of them that the field lines flow in a completely new way. The resulting vector field has no superficial resemblance to the original. This new vector field however is related to the original since the person who altered the field can always retrace his steps and get back to the field with which he started. Is there a property of a field that does not change under such smooth deformations?

Consider isolated points on the surface where the vector field vanishes—the still spots where the wind doesn't blow. These singularities are really just zero vectors in the vector field, *zeros* for short. They are immutable features of the field. This makes zeros special. We can push and pull at field lines near an isolated zero or drag a zero around over the surface, but the zero remains a zero. No amount of smooth, reversible massaging of the vector field will remove it, so the kinds of zeros in the vector field act as a fingerprint for the vector field.

Now we are ready to explore the Poincaré-Hopf Index Theorem for surfaces. This celebrated topology theorem relates a specific quantity associated with the zeros (the index) of a vector field to the number of holes in the surface. Henri Poincaré proved it in 1885 and Heinz Hopf proved the full theorem for higher dimensions in 1926 after earlier partial results by Brouwer and Hadamard. Presented here is a straightforward proof of the Poincaré-Hopf

Index Theorem designed to utilize the interplay between two simple mathematical tools and your ability to visualize curves and surfaces.

AS THE NEEDLE SPINS: INDEX AND WINDING NUMBER

We can classify a zero by assigning it an integer value called its *index*, which depends on the behavior of the field in a small neighborhood encompassing the point. To calculate the index we "pull" the relevant section of the vector field off our surface onto a plane. We do this because surfaces are generally curved but planes are not; vector directions are mathematically easier to deal with on a flat background. On this plane suppose x-y axes and an origin have been drawn. If this section of the field contains a zero, the zero is mapped to the origin of the plane.

the vector will have returned to its initial position (the smoothness of the field guarantees this), and so it must have made some whole number of trips around the origin. Each complete counterclockwise rotation adds one to the index and each clockwise rotation subtracts one. If the vector never makes a full rotation then the index is 0. This will occur when the section of the vector field has no zero (see Figure 2(A)).

The index measures *local* directional variation of a vector field but we can do the same thing on a *global* scale. For example, if we drew an arbitrary closed curve snaking its way around a sphere, it is not possible to assign it an index. Only for curves that occupy a small section of some surface does an index make sense. The reason for this is that a sur-

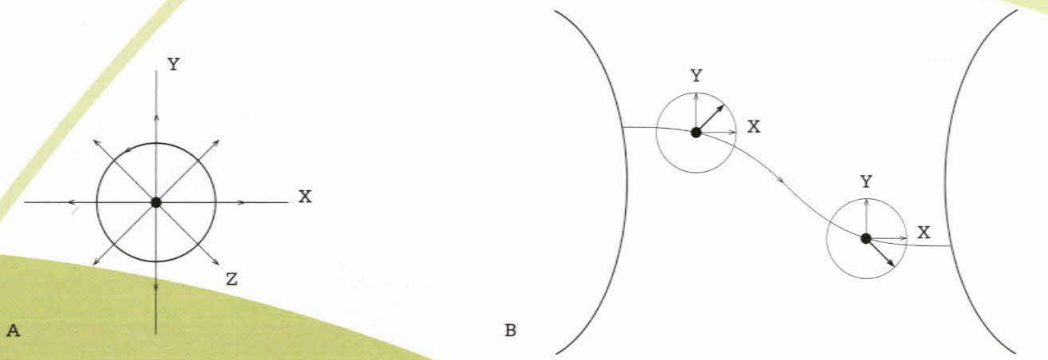


FIGURE 2. (A) A zero vector (indicated by *z*) is mapped to the origin of a plane. The index is the number of times the vectors in the vector field rotate about the origin when traversing a counterclockwise circle around the zero. The index here is 1. (B) The winding number is the number of times the vectors in the vector field rotate about the tangent vector to the path. Reversing the orientation of the path simply negates the value of the winding number.

Imagine an ant that starts from the origin and takes a tiny step outward. She walks counterclockwise on a tight circular path until she returns to her original position. Pretend you are watching this from above the plane. For each vector the ant encounters on her travel, *you* take a sheet of paper, copy the pair of x-y axes and plot the vector. Once the ant has finished one tour, make a flipbook using these pages. Flip through it fast enough and it will appear as if a vector is rotating about the origin. By the time you reach the end of the flipbook,

face appears flat only upon very close inspection. Approximate flatness is needed to ascribe an index because we need to be able to "pull" a section of the vector field off our surface and onto a plane. This can only be done if the surface is nearly flat.

There is, however, a meaningful term called the winding number that quantifies vector field variation along arbitrary loops. It is calculated in a similar manner to the index with one subtle difference. First, we draw a closed curve on a surface that does not run over any

"Singularities are just zero vectors, immutable features of the field."

zeros. Again, we call upon an ant to walk along the curve. This time *she* will be the one recording the vectors encountered on the path, not you.

The ant will stop at each point, draw a pair of x-y axes on a sheet of paper, and plot her vector. But she cannot see the entire surface as you can. How can she draw meaningful vector directions without any kind of reference? So she chooses to use the tangent vector to the path as her x-axis. The y-axis she chooses is both perpendicular to the outward normal vector (a vector that points away from the inside of the surface and is perpendicular to the tangent plane) and to the tangent vector (see Figure 2(B)).

Once her tour of the curve has been completed, the ant flips through her pages and the vectors spin about the origin. From here on the rules for calculating the winding number are the same as those for the index.

HOMOTOPY AND CURVE FAMILIES

The winding number of a closed curve has the crucial property that it does not change when we jiggle the curve. It is easy to see why this must be true: the winding number can only change by integer increments, large discontinuous jumps in value. However, slightly wiggling the curve should not have any effect on its value. Let's take advantage of this property to create a family of closed curves on a surface all with the same winding number. Starting with one curve, we wiggle it a little to get a second curve. We wiggle the second a little more to get a third and so on. In fact, we can do this as much we like to generate macroscopic movements of a closed curve over a surface while keeping the winding number constant. Mathematically speaking, we say that the winding number is invariant under a homotopy of the path and each curve in the family is *homotopic* to any other (see Figure 3).

Two things limit where or how much we may wiggle a closed curve: holes in the surface and zeros in the vector field. The first limitation is obvious, but why should a zero restrict wiggling a closed curve? Suppose we had a

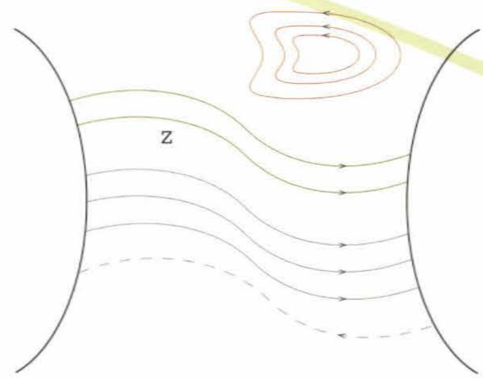


FIGURE 3. Three families of closed curves are shown. Curves of the same color are homotopic to each other and therefore have the same winding number. However, the winding number of the solid blue curves is the negative of the winding number of the dashed blue curve since it has the opposite orientation. Note that the blue curves are not homotopic to the green curves because a zero vector separates them.

curve lying on top of a zero. The technique for calculating winding number calls for the observation of the rotation of vectors as the curve is traced out. If, somewhere along this path, there were a vector with zero length, it would be ambiguous to talk about a net rotation, as a vector's direction is meaningless when its magnitude is zero.

RELATING WINDING NUMBERS AND INDICES

The index and winding number can both be formulated in terms of similar mathematical integrals. This is no coincidence. After all, each of these tools measures the rotation of the vector field about some fixed point. The choice of fixed point distinguishes the two methods: a zero vector is used for the index, and successive points along some closed curve are used for the winding number. The latter is defined for any closed path on a surface that misses a zero, so it can be studied from a local perspective as well. That is, if a curve encircles an isolated zero and is small enough, it makes sense to ask how the winding number is related to the index of that zero.

First we need to know what the *orientation* of a small curve is. Set a clock flat on a surface. According to our convention, clockwise curves curl in the direction that the clock's hands turn. It can be shown that for a small

"We can classify a zero by assigning it an integer value called its **index**, which depends on the behavior of the field in a small neighborhood encompassing the point."

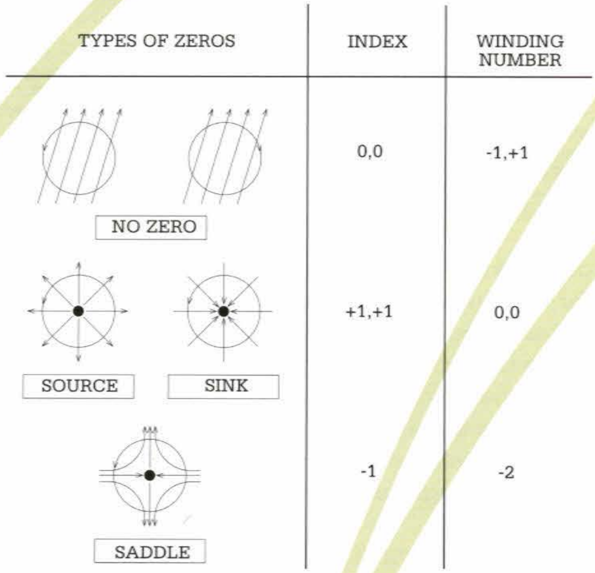


FIGURE 4. Zeros are characterized by the behavior of nearby vectors. A few types of zeros and their indices and winding numbers for the illustrated curves are listed.

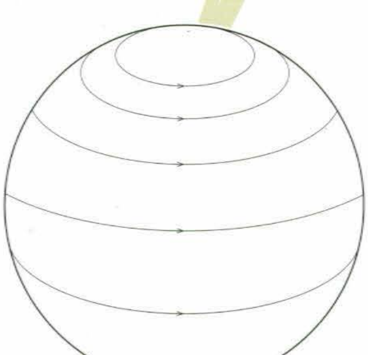


FIGURE 5. Lines of latitude for a sphere are homotopic to each other *assuming* there are no zeros in the vector field. The smallest circle at the north pole is oriented counterclockwise, but the smallest circle at the south pole is oriented clockwise so they must have different winding numbers. This is a contradiction, so there must be at least one zero vector somewhere in the vector field.

clockwise-oriented curve the winding number is equal to 1 minus the index of the zero contained within the curve. Now if we reverse the orientation of this curve, making it counterclockwise, then we need to negate the winding number so that it now equals the index minus 1. A simple but important case is when there is no zero in the area enclosed by the curve. Then the index is 0 and hence the winding number ±1 depending on the orientation (see Figure 4).

Although a few more concepts need introduction before we present the Poincaré-Hopf proof, our current arsenal of tools shows that it is impossible for the sphere to admit a completely nonzero vector field. Suppose for a moment that such a vector field did exist. Because there are no zeros to get in our way, it is possible to cover the sphere by a family of latitude circles that are homotopic to each other and thus have the same winding number (see Figure 5). But consider the very small circles near the north and south poles. These clearly have opposite orientations: one circle must have winding number –1 and the other +1. This is a contradiction, so our original assumption must have been false.

ZERO TRANSITIONS AND HOLE TRANSITIONS

Consider the surface shown in Figure 6(A) on which are defined two closed curves F_1 and F_2 that would be homotopic were it not for the zero, z, separating them. Suppose we know the winding number of F_1. How can we determine the winding number of F_2? Start by smoothly deforming the curves until they intersect, then reverse the orientation of F_2 as shown in Figure 6(B). Since they cross at points *a* and *b*, we may define each curve as being made up of two smaller curves: one component going from *a* to *b* across the front of the surface, and the complementary curve from *b* to *a* traveling around the back. Now cut the curves at *a* and *b* so that we get four pieces and reattach them so that we get a closed curve going around the zero and one big curve surrounding nothing as shown in Figure 6(C). In the same way that you can contract the loop of a lasso, we can wiggle G_2 into smaller closed curves. This means that G_2 is homotopic to a small clockwise-oriented circle. No zeros lie within it so its winding number must be +1. Contracting G_1 results in a small counterclockwise-oriented loop surrounding a zero. We know exactly how to treat this case. The winding number of G_1 equals the index of that zero minus 1! It follows that the sum of the winding numbers of G_1 and G_2 equal those of F_1 and the reversed F_2 since they are made up of the same pieces. Reversing orientation simply negates the winding number of F_2, so we get:

$$W(F_1) - W(F_2) = ind(z) \quad (1)$$

where *W( )* indicates winding number and *ind( )* stands for index.

The same idea is used to overcome the difficulty that arises when two curves are separated by a hole rather than a zero (see Figure 7(A)).

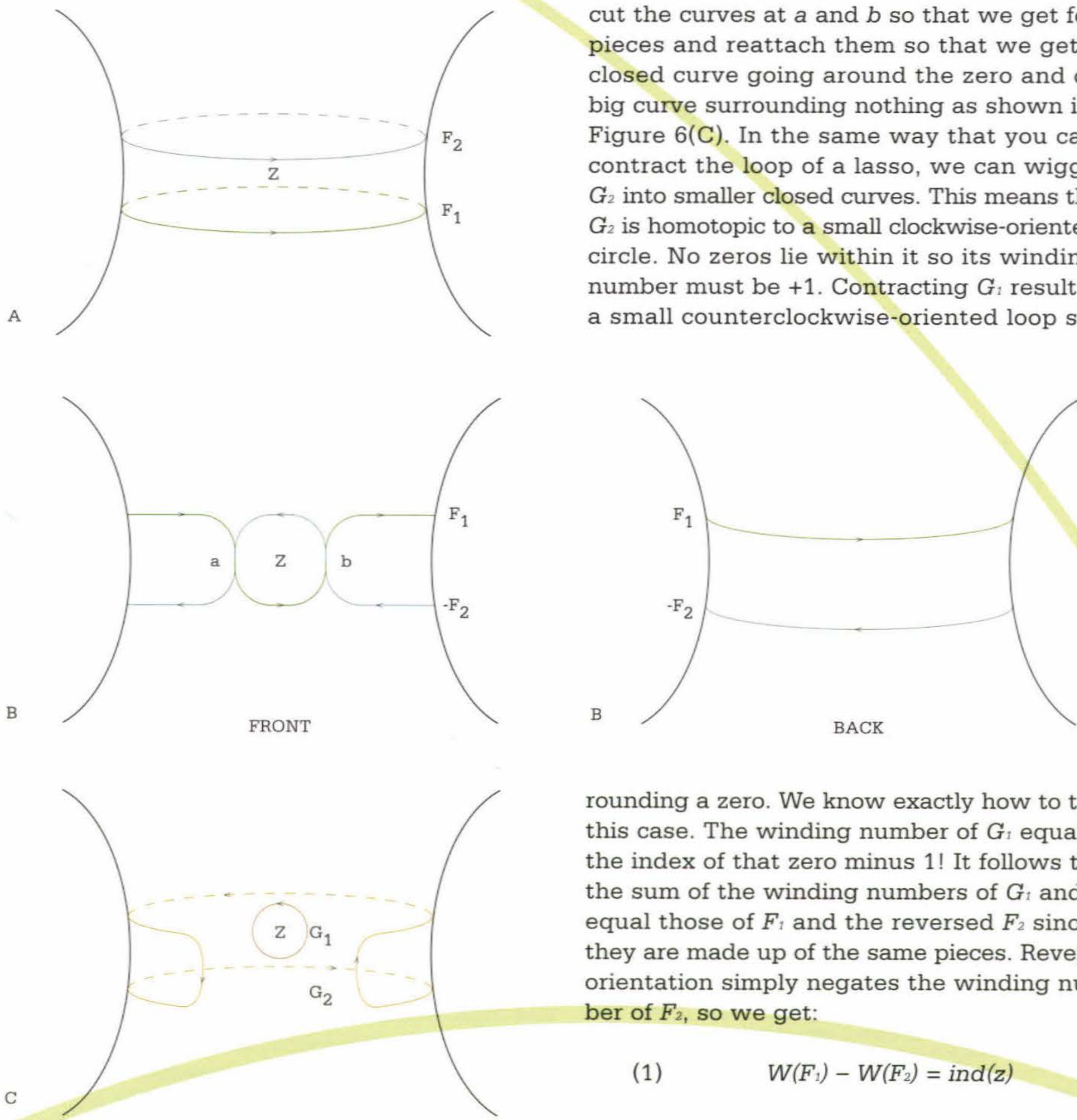
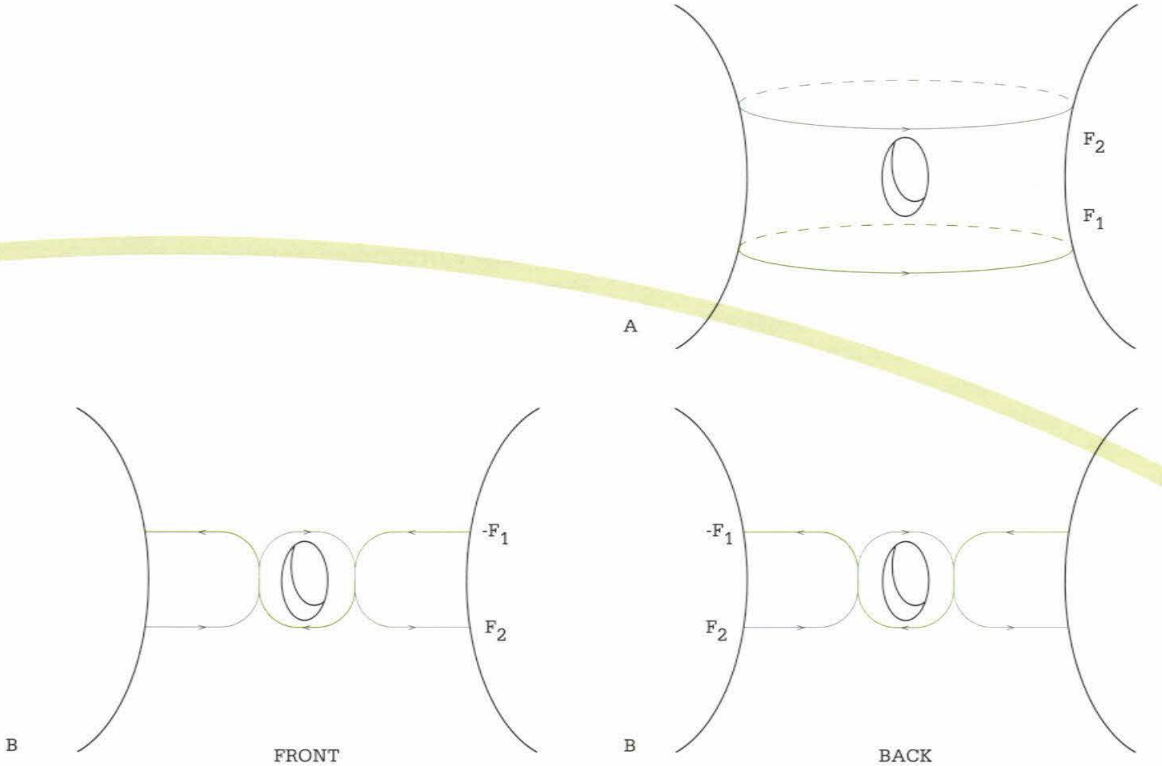


FIGURE 6. (A) A zero vector separates two closed curves. (B) The curves, F_1 and F_2, can be deformed until they intersect on the front, then (C) cut and reattached into two new curves, G_1 and G_2, whose winding numbers are easy to calculate. We can then relate the winding numbers of the two curves to the index of the zero.



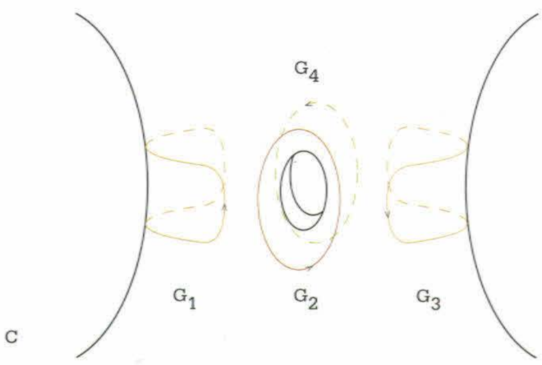


FIGURE 7. (A) A hole separates two closed curves. (B) The curves, F_1 and F_2, can be deformed until they intersect on the front and back, then (C) cut and reattached into four new curves, G_1 through G_4. Note that G_2 and G_4 are homotopic to each other but have opposite orientations so their winding numbers cancel. We can then relate the winding numbers of the two remaining curves to each other.

F_1 and F_2 are smoothly deformed until they intersect (see Figure 7(B)). The orientation of F_1 is then reversed and the curves are cut at the points of intersection and reattached to give the four loops shown in Figure 7(C). Loops G_1 and G_3 are oriented counterclockwise so their winding numbers are each –1. G_2 and G_4 encircle the hole like a lasso. Visualize sliding G_4 out through the hole until it meets G_2. They are homotopic to each other but have opposite orientations so their winding numbers (whatever they are) must have opposite sign. When we sum the four winding numbers from the loops, the contributions from G_2 and G_4 cancel! We conclude that:

$$W(F_2) - W(F_1) = -2 \qquad (2)$$

PROVING THE POINCARÉ-HOPF INDEX THEOREM

Now that we have all this machinery in place, the proof of the Poincaré-Hopf Index

"[The Theorem] says that the kinds of vector fields that may exist on a surface are determined only by the number of holes in that surface."

Theorem is within reach. Consider a general case: a surface with g holes and a vector field with some number of isolated zeros. To simplify things we drag all the zeros smoothly around the surface and line them up below the first hole as in Figure 8. This does not change their indices. Add two small circles at the ends of the surface: the bottom one has winding number +1 and the top one −1. Start with the bottom circle and wiggle it upward along the surface; in this process, the curve will make multiple transitions across zeros and holes, thus altering its winding number. When it reaches the peak, the loop is homotopic to the tiny circle we left waiting there; their winding numbers must be equal to maintain consistency. This conservation of winding number is the key to the Poincaré-Hopf Index Theorem.

Let's do the math. Equation (1) tells us that each time we move a loop over a zero we need to decrement the winding number of the curve by the zero's index. By the end of the first phase, the winding number will be 1 minus the sum of the indices of all zeros in the vector field. Equation (2) then tells us to decrement the winding number by 2 each time we move across a hole; since there are g holes we subtract $2g$ by the end of the second phase. For winding number to be conserved, the following must therefore hold:

$$1 - \Sigma\, ind(z) - 2g = -1$$

where the sum is over all zeros in the field. Rearranging the formula gives the Poincaré-Hopf Index Theorem:

$$\Sigma\, ind(z) = 2 - 2g$$

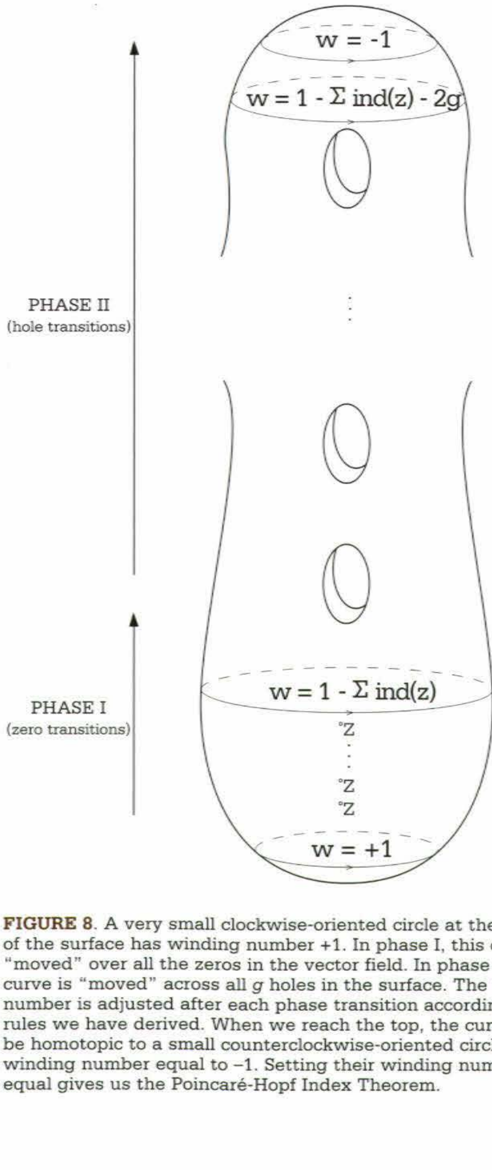


FIGURE 8. A very small clockwise-oriented circle at the bottom of the surface has winding number +1. In phase I, this curve is "moved" over all the zeros in the vector field. In phase II, the curve is "moved" across all g holes in the surface. The winding number is adjusted after each phase transition according to the rules we have derived. When we reach the top, the curve must be homotopic to a small counterclockwise-oriented circle with winding number equal to −1. Setting their winding numbers equal gives us the Poincaré-Hopf Index Theorem.

"An orange and an egg will admit similar vector fields because they both have no holes."

This is quite an impressive equation. It says that only the number of holes in a surface determine the kinds of vector fields that may exist on that surface. For example, an orange and an egg will admit similar vector fields because they both have no holes. The story told at the beginning of this article should now make sense. Because the sphere has no holes ($g = 0$) the index sum of a vector field on a sphere must equal 2. A nonzero number implies that there is at least one zero somewhere in the vector field. An ant on a windy, spherical world is lucky. But the torus has one hole ($g = 1$) so the sum of the indices equals 0. This does not mean that all vector fields on the torus are free of zeros. Rather, it means that it is *possible* to draw a nonzero vector field, such as in Figure 1(B). If such a vector field describes the wind on a donut-shaped world, then an ant on this world will never find a sheltered spot. For this ant, there is no safety in numbers. C

Steve Paik is a third year undergraduate in Physics at the California Institute of Technology. *This work was completed with Rahul Pandharipande, Professor of Mathematics at Caltech, and funded by the 2000 Caltech Physics 11 Research Program. The author wishes to thank Rahul Pandharipande and Marzia Polito.*

FURTHER READING

1 V. Guillemin, A. Pollack. *Differential Topology* (Prentice Hall, Upper Saddle River, NJ, 1974).

2 H. Hopf. *Differential Geometry in the Large* (Springer-Verlag, Heidelberg, Germany, 1989).

3 J. Milnor. *Topology from the Differentiable Viewpoint* (Princeton University Press, Princeton, NJ, 1997).

LACKING CHEMICAL SIGNS,

ALH84001

MARS METEORITE WAXES LIFELESS

BY MELISSA J. STRAUSBERG

THE POSSIBILITY THAT LIFE MIGHT EXIST on other planets is compelling. Science fiction writers have dramatized the existence of "little green men" since the early days of space flight. One scientific theory postulates that extraterrestrial materials, delivered here by comets or meteors, were the basis for life on Earth.

Life on Earth began roughly 3.9 billion years ago, the oldest undisputed age of a fossil. In the 1950s at the University of Chicago, Stanley Miller and Harold Urey tried to understand the genesis of life by chemically simulating in glassware the conditions on Earth roughly 3.8 to 2.5 billions years ago. These chemistry experiments demonstrated that during this Archaean Eon, amino acids, the building blocks of life, could form with no more than five ingredients: ammonia, methane, hydrogen gas, water, and an electric spark (see Figure 1).

FIGURE 1. An artist's rendition of Archean Earth, thought to be the environment necessary for life to originate.
Source: S. Liebes, E. Sahtouris, B. Swimme. A Walk Through Time (John Wiley ans Sons, New York, 1998).

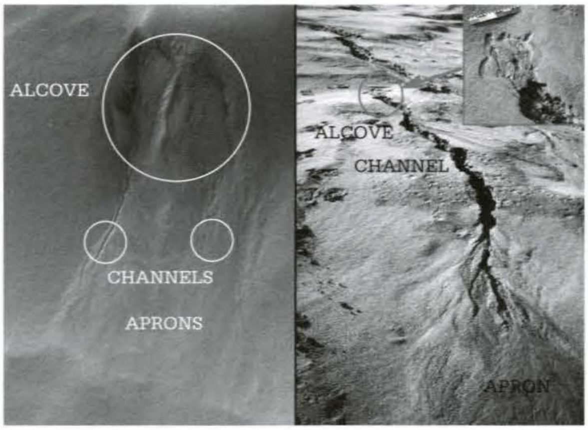


FIGURE 2. A comparison of gullies on Mars (left) and Earth (right). It is yet unknown whether the Martian gullies were caused by flowing water or by some other erosional force at work on the planet's surface.
Source: NASA/JPL/Malin Space Science Systems, 2000.

Liquid water is essential to life on this planet. Water vapor is abundant in the universe, but liquid water only exists in the moderate temperature range of 0 to 100 degrees Celsius. Our galaxy alone contains a mass of water millions of times the mass of the Sun, but only a few locations in our solar system are thought to have the delicate temperature balance necessary for liquid water, the most likely of which are Europa and Mars.

Europa is the fourth largest of Jupiter's moons. Images from NASA's Galileo Orbiter show sheets of ice on Europa's surface, branded with cracks and tipped ridges. These features may be the result of a sub-surface water ocean, beginning 1 kilometer down and extending 60 kilometers deep. NASA hopes that their 2008 Europa Orbiter mission will confirm these inferences, but for now, the existence of liquid water on Europa remains an open question.

Mars too may contain liquid water. While much of the Martian water is frozen in polar ice caps, images from NASA's Mars Global Surveyor released last year suggest recent runs of water on the planet's surface from subterranean stores (see Figure 2). The European Space Agency hopes to discover Martian water in 2003 with their Mars Express orbiting spacecraft. Equipped with ground-penetrating radar, Mars Express will prospect for water as deep as 5 kilometers into the Martian crust.

Work on these space missions rocketed forward in 1996 when scientist David McKay of the NASA Johnson Space Center in Houston announced that his team of researchers from five institutions had found evidence of life in the Martian meteorite ALH84001 (see Figure 3). The news drew the world's attention and NASA Administrator Daniel Goldin briefed President Clinton at the White House on Martian life. One of only twelve recovered Martian meteorites, the fist-sized ALH84001 had chemical compounds and fossil-like structures that might have been biological. In the summer of 2000, we received the now-famous meteorite for study. After analyzing the meteorite with a variety of geochemical techniques, our research suggested that the contents of the ALH84001 meteorite were not biologically produced.

FIGURE 3. A piece of the Martian meteorite ALH84001, recovered from the Allan Hills ice fields in 1984. The cube in the lower right corner is 1 cm in height. ALH84001 became well-known in 1996 when scientists purported to have found life on Mars in the form of ancient biological fossils in this meteorite.
Source: Lunar and Planetary Institute, 1996.

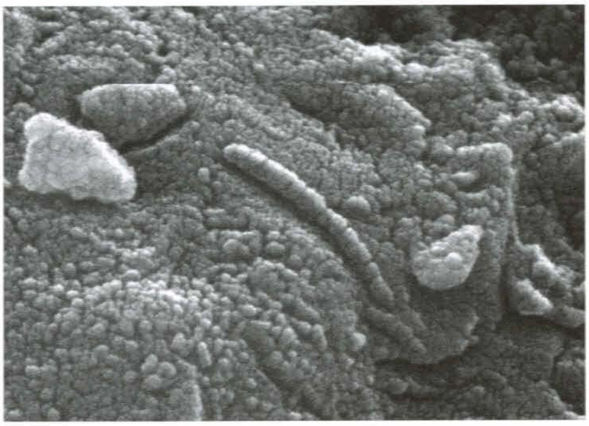


FIGURE 4. Electron micrograph of elliptical and tubular structures in Martian meteorite ALH84001, a meteorite from Mars. These odd formations could be fossilized bacteria.
Source: Lunar and Planetary Institute, 1996.

MARS COULD HAVE ONCE SUPPORTED LIFE

The Martian environment may have once been able to sustain life similar to that found on Earth. After the formation of the Martian volcano Tharsis more than four billion years ago, Mars' atmosphere became filled with the gasses produced by volcanism – ammonia, methane, and hydrogen. Lightning also hit the Martian surface, and with water, the Martian environment would have resembled that of the Archaean Earth when life formed. NASA began its search for Martian life by probing for organic matter in the planet's surface soil and underlying crust. From 1976 to 1982, two NASA Viking spacecraft sought evidence of Martian life through programmed experiments. Such experiments included monitoring carbon dioxide output from soil fed with simple sugars and measuring the soil's organic compound output after heating. The Viking experiments did not reveal any organic matter, but this did not rule out the possibility of life on Mars. In 1996, McKay proposed that a harsh oxidizing agent or high quantities of UV radiation could have left the surface of the planet devoid of organic matter. Still, traces of organic matter might remain below the surface.

With the possibility for Martian life still open, scientists focused attention on the twelve discovered Martian meteorites. In 1996, McKay and colleagues analyzed ALH84001, observing elliptical and tubular structures that resembled bacteria fossils (see Figure 4) and detecting long-lived organic molecules called polycyclic aromatic hydrocarbons. His paper concluded that these features

"The fist-sized ALH84001 had **chemical compounds** and **fossil-like structures** that might have been **biological.**"

were evidence for primitive bacterial life on Mars, but the data was inconclusive. Both elliptical or tubular structures and polycyclic aromatic hydrocarbons can be produced through inorganic as well as living processes. In 1998, Luann Becker of the Hawaii Institute of Geophysics and Planetology and colleagues found amino acids in ALH84001. While the study showed that those amino acids seeped into the meteorites over 13,000 years in Antarctic ice fields, the possibility could not be excluded that trace amounts of some amino acids might be Martian.

The geological history of ALH84001 suggests that the meteorite itself met basic conditions for sustaining life while on Mars. Radioactive dating revealed that the meteorite crystallized from magma 4.5 billion years ago. Fractures on the meteorite showed that it was broken up after it crystallized, perhaps by a nearby asteroid impact. Globules of carbon compounds formed in the rock at this time, a strong indication that the meteorite interacted with air and water while still on Mars. Cosmic ray analysis reveals that the meteorite was ejected from Mars 16 million years ago, landing on Earth 3 million years later.

HYDROGEN FINGERPRINTS

When samples of ALH84001 are heated to temperatures between 300°C and 1000°C, hydrogen containing compounds within the meteorite burn, releasing hydrogen. Like Becker's characterization of the meteorite's amino acids, determining the origin of the hydrogen in ALH84001, terrestrial or Martian, biological or inanimate, would further suggest the presence or absence of traces of life.

We can determine if some of this hydrogen is originally from Mars by looking at the ratio of deuterium to hydrogen in the gas. Deuterium is an isotope of regular hydrogen that contains an extra neutron, and the deuterium to hydrogen ratio varies from location to location. The Martian atmosphere has a deuterium to hydrogen ratio roughly five times that of Earth's atmosphere. This isotopic fingerprint makes Martian hydrogen easy to identify. Any compound from Mars containing hydrogen, including all organic compounds, has a much higher deuterium to hydrogen ratio than a structurally identical compound from Earth.

Experiments by Laurie A. Leshin and colleagues at Caltech showed in 1996 that ALH84001 releases deuterium-rich, Martian hydrogen when heated. However, those experiments did not identify the compounds that initially held the extraterrestrial hydrogen. These so-called hydrous phases might be organic molecules like hydrocarbons, but

HYDROGEN ASSAY ON UNTREATED METEORITE SAMPLE

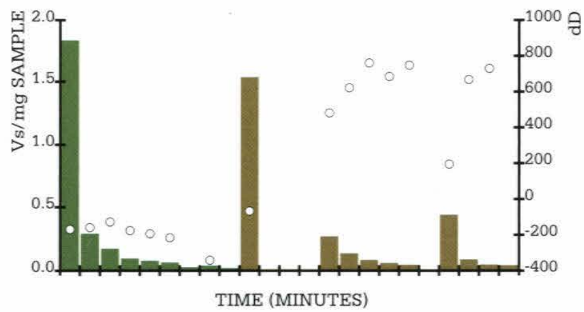


FIGURE 5. Hydrogen released from the stepped heating of Martian meteorite ALH84001 without chemical treatment of the sample. The graph displays both the total amount of hydrogen released and its deuterium to hydrogen ratio. The colored bars represent the amount of water released at each temperature step (blue for 250°C, red for 450°C, and green for 650°C). The dots represent the deuterium to hydrogen ratio of the hydrogen collected. The first temperature step releases terrestrial, deuterium-poor atmospheric water that is loosely held in the rock's structure. The second and third temperature steps release Martian hydrogen, which has a much higher deuterium to hydrogen ratio than that found on Earth.

could also be clay minerals, salts, or silicate minerals. Identifying the hydrous phases would elucidate whether the source of this Martian hydrogen was biological or inanimate. The ALH84001 meteorite does contain organic compounds, and linking the Martian hydrogen to those organic compounds would suggest that the organic compounds were Martian and possibly biologically made. On the other hand, Martian hydrogen could have been released from any of the compounds in the meteorite, not necessarily the organic compounds. Our research focused on establishing the source of the Martian hydrogen in ALH84001, whether organic or just an errant bit of dust or clay.

CHEMICAL ASSAYS

Different hydrogen containing compounds have vastly different chemical properties, providing the key to uncovering which of the meteorite's hydrogen containing compounds evolve the observed Martian hydrogen when dried and heated. For example, hydroxides and salts react strongly with acids, while organic molecules and clays are relatively inert to acid treatment. Various other chemical reactions remove particular hydrogen containing compounds from the sample. By heating a sample after removing a particular type of compound from the sample, we could establish whether that compound had been responsible for evolving Martian hydrogen. In characterizing ALH84001, we used a few basic chemical reactions to selectively remove particular hydrogen containing compounds. Hydrochloric acid selectively removes salts and hydroxides, while hydrogen peroxide was used to selectively remove organic matter.

We also characterized the Martian hydrogen containing compounds' particular chemical properties, including exchangeability, the ease with which a hydrogen-containing compound replaces its hydrogen with external hydrogen. Exchangeability was quantified by placing a sample in water of a known deuterium to hydrogen ratio and then drying and heating the sample to measure the new deu-

"The geological history of ALH84001 suggests that the meteorite itself met basic conditions for sustaining life while on Mars."

HYDROGEN ASSAY ON
HYDROCHLORIC ACID TREATED METEORITE SAMPLE

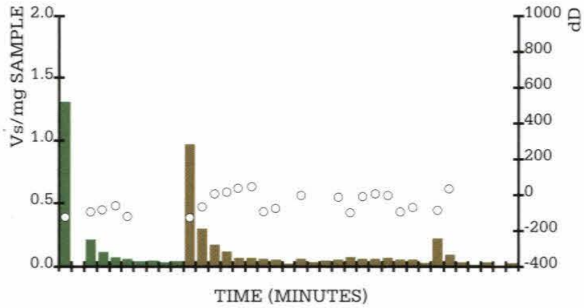


FIGURE 6. Hydrogen released from the stepped heating of acid treated Martian meteorite ALH84001. Less hydrogen was released at the second temperature step and it was of a lower deuterium to hydrogen ratio than in untreated ALH84001 (Figure 5). This difference indicates that the Martian hydrogen-containing compound is soluble in a weak acid. Temperature steps are in blue for 250°C, red for 450°C, and green for 650°C; white dots indicate deuterium to hydrogen ratio.

HYDROGEN ASSAY ON
FOUR-DAY WATER-IMMERSED METEORITE SAMPLE

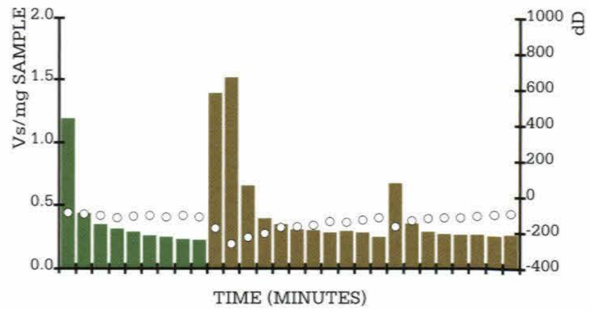


FIGURE 7. Hydrogen released from the stepped heating of samples of Martian meteorite ALH84001 immersed in water for four days. Again, the deuterium to hydrogen ratio of the hydrogen released in the second and third steps is nearer the terrestrial value than the Martian one. This indicates that ALH84001's Martian hydrogen-containing compound is exchangeable in water. Temperature steps are in blue for 250°C, red for 450°C, and green for 650°C; white dots indicate deuterium to hydrogen ratio.

terium to hydrogen ratio. Another chemical property tested was thermal stability; we preheated some samples to test how stable the hydrogen containing phases were to thermal variations.

After a particular chemical reaction, a sample and its untreated control samples underwent heating to release hydrogen to determine the deuterium to hydrogen ratio. The heating was stepped, meaning that temperatures were held at 250°C, 450°C, and 650°C. The amount of hydrogen gas liberated and its deuterium to hydrogen ratio at each temperature step is graphed for one trial in Figure 5.

A HANDFUL OF SALTY DUST

Our experiments suggest that a fresh sample of ALH84001 releases a significant amount of deuterium-rich Martian hydrogen when heated to roughly 450° C (see Figure 5). When a sample is first treated with hydrochloric acid, the hydrogen it releases upon heating has a nearly terrestrial deuterium to hydrogen ratio, as shown in Figure 6. This result conclusively demonstrates that the Martian hydrogen-containing compound is soluble in even a weak acid.

The exchangeability experiment shows that the deuterium to hydrogen ratio of hydrogen released from treated samples went below the generally accepted terrestrial ratio (see Figure 7), indicating that the Martian hydrogen carrier is exchangeable. Similar trials indicate that the Martian hydrogen carrier is insoluble in a weak base and decomposes when pre-heated to around 450°C.

The above assays determined critical chemical properties of the Martian hydrogen-containing compound: solubility in a weak acid, insolubility in a moderate base, thermal instability at 450°C and high exchangeability. These properties suggest that the Martian hydrogen in ALH84001 does not come from organic matter or clay, but is more likely a salt or an acid-soluble hydroxide. Both salt and acid-soluble hydroxides are suspected to be principal constituents of surface dust in the modern Martian environment. This surface dust may have become trapped in ALH84001 sometime before its ejection.

Although our results suggest that none of the organic matter on ALH84001 is Martian, this "negative result" conclusion is very difficult to prove entirely. It cannot be concluded

"The Martian atmosphere has a deuterium to hydrogen ratio roughly five times that of Earth's atmosphere, so this isotopic fingerprint makes Martian hydrogen easy to identify."

that there is no needle in a haystack without turning over every straw in the haystack. Similarly, a counter argument to our conclusions might suggest that perhaps our methods were not sensitive enough to detect trace amounts of Martian organic compound. There will almost always still be a possibility that life did exist on Mars.

If the Martian hydrogen-containing compound in ALH84001 is in fact a salt rather than an acid-soluble hydroxide, we could say with some confidence that there was once liquid water on Mars. But recalling the 1950s chemical simulations by Stanley Miller and Harold Urey, the origin of life requires a coincidence of many factors, including the presence of liquid water. Still, optimistic scientists continue to search for proof that terrestrial life is not unique. In 2000, McKay and colleagues from five other institutions reported finding what could be the remains of magnetic bacteria in ALH84001. NASA is planning a mission that will go to Mars and bring a soil sample back to Earth, and more meteorites from Mars are found on Earth every year. Perhaps one of these approaches will yield the proof that has yet been so elusive; for now, Martian life must remain an intriguing possibility. C

Melissa Strausberg is a second year undergraduate in Planetary Science at the California Institute of Technology. *This work was completed with John Eiler, Assistant Professor of Geochemistry at Caltech, and funded by the 2000 Axline Caltech Summer Undergraduate Research Fellowship. The author wishes to thank John Eiler and Nami Kitchen.*

FURTHER READING

1 L. Becker, B. Popp, T. Rust, J. L. Bada. The origin of organic matter in the Martian meteorite ALH84001. *Earth and Planetary Science Letters* **167**, 71-79 (1999).

2 L. A. Leshin, S. Epstein, E. M. Stolper. Hydrogen Isotope Geochemistry of SNC Meteorites. *Geochemica et Cosmochimica Acta* **60-14**, 2635-2650 (1996).

3 D. S. McKay *et al.* Search for Past Life on Mars: Possible Relic Biogenic Activity in Martian Meteorite ALH84001. *Science* **273**, 924-930 (1996).

4 H. Y. McSween. What we have learned about Mars from SNC meteorites. *Meteoritics* **29**, 757-779 (1994).

5 L. L. Watson, I. D. Hutcheon, S. Epstein, E. M. Stolper. Water on Mars: Clues from Deuterium/Hydrogen and Water Contents of Hydrous Phases in SNC Meteorites. *Science* **265**, 86-90 (1994).

SIMULATING STOCK OPTIONS

BY MONA A. SHEIKH

JOE BUCKS WORKS FOR WISEINVESTORS, INC. BECAUSE OF HIS outstanding performance last year, his boss gave him a bonus—an option of 2000 shares of company stock. The option exercise price is \$12 a share. Joe will pay this amount for the stock if he exercises his option, called a call option. A year later, the stock's market price has soared to \$20 a share, and Joe decides to exercise the option. He pays \$24,000 for stock that is currently worth \$40,000.

The incentive stock option is common, often awarded as a bonus for good performance. A stock option is a contract that gives the option-holder the right to trade (buy or sell) a quantity of stock in the future for a predetermined price (the exercise price). At expiry, if the stock price S is greater than the option price K, Joe Bucks exercises his option. He pays an amount K for stock worth S, making a profit of S minus K—\$16,000 for Joe Bucks. If K exceeds S at expiry, he lets the option expire. A high interest rate makes putting money in the bank more profitable than buying options, so option values are discounted to reflect what interest alone could make. With a high interest rate and a long time to expiry, options must be steeply discounted to be attractive.

Option contracts can be classified by when an option can be exercised. European options can only be exercised at the expiry date, whereas American options may be exercised at any time before expiry. Analyzing the European option requires only determining the stock's final price. The Black-Scholes formula, developed in 1973, provided an easy way to do that, revolutionizing the world of option trading. Once options could be accurately priced, more of the world's financial markets began listing stock options. Option trading has since grown to an annual volume of 670 million options last year. Scholes picked up a Nobel Prize for his work in 1997 (Black unfortunately passed away in 1995).

The American option's additional complication—finding the optimal time for exercise—makes pricing it one of the most challenging problems in finance. A faster and more accurate valuation of the American option would change the world of option trading like the Black-Scholes formula did nearly three decades ago. Least squares regression might be the answer.

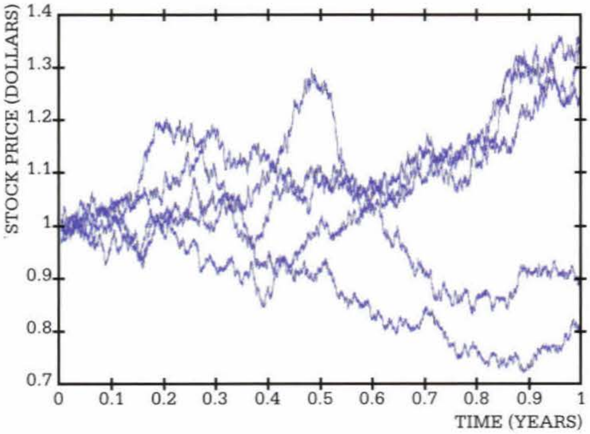


FIGURE 1. Lognormal random walks can simulate stock behavior. A "coin" flipped at discrete intervals increases or decreases stock price. This coin can be weighted to reflect expected long-term behavior (rising or falling). Many walks together produce a good model for the behavior of a stock's price.

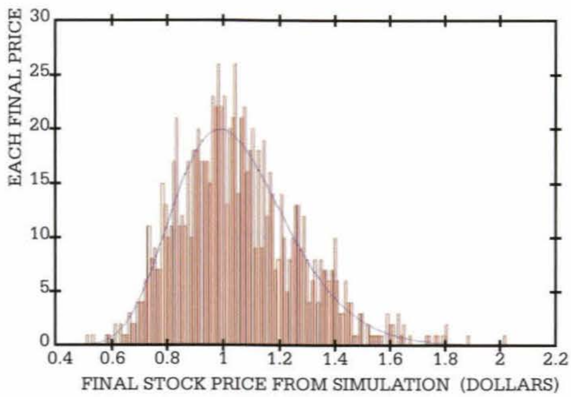


FIGURE 2. Stock option price is computed based on each lognormal random walk—a single simulation of a stock's behavior. Many trials together produce a lognormal distribution of stock prices. An option price is determined for each trial, and the value of the option is the average over all trials. More trials improve the estimate, but the simulation takes longer to run.

SIMULATION PREDICTS EUROPEAN OPTION PRICES

The Monte Carlo method predicts stock price behavior with lognormal random walks (see Figure 1). Future price is a function of current price, interest rate, how volatile a stock is, and a random factor. We used this method to simulate the European option, checking our results against the theoretical predictions of the Black-Scholes equation. In the real world, risky investments need greater incentives than safe ones, even with the same expected values. We ignore this bias, as does the Black-Scholes equation, and assume risk-neutrality.

Many lognormal random walks together produce a lognormal distribution of final stock prices (see Figure 2). We evaluate the option for each run and average our results to predict option price. Our predictions agree well with the Black-Scholes equation for the same interest rate, volatility, original price, and time to expiry (see Figure 3).

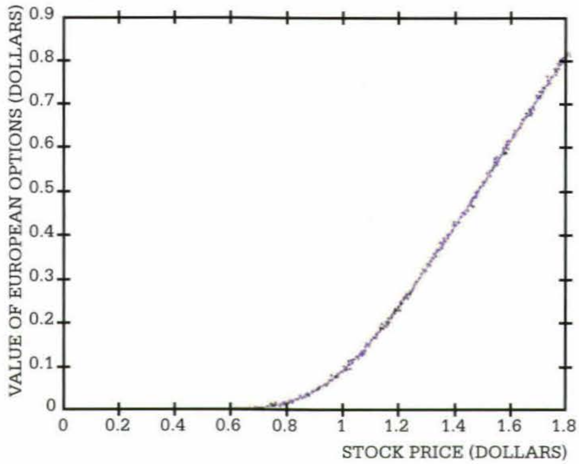


FIGURE 3. For the same interest rate, volatility, and time to expiry, Monte Carlo simulations of the European call option (points) agree with the theoretical predictions of the Black-Scholes equation (solid line), the accepted method for European option pricing. The option exercise price is 0.6 for this example, so when stock price is lower than 0.6, the option is worthless.

For the American option, direct application of the Monte Carlo technique is impossible. Since the owner of an option can exercise the option at any time before expiry, modeling requires an infinite number of points along the path. Furthermore, this method cannot predict exercise time—an American option will be exercised early if its holder thinks his profits are greatest before expiry. The decision to exercise early requires an expected future stock price based on current price. The Least-Squares Monte Carlo algorithm, recently developed by Longstaff and Schwartz, provides a means for this prediction.

SIMPLIFYING AN IMPOSSIBLE CALCULATION

Joe Bucks buys a put option. Unlike the call option his company gave him, this contract gives him the right to *sell* stock at a predetermined price to the option's issuer. Even if the stock price plummets, Joe can still sell at a high price. In a general economic downturn, his profits could cheaply buy other stocks. We want to know how much Joe should be charged for that contract.

We again simulate stock prices with lognormal random walks. We consider discrete times along each path—Joe can only sell on one day a year. This reduced flexibility will undervalue the option price since the best time to sell will undoubtedly be some other day of the year. More discrete times improve the estimate but increase computing time.

Each January 1, Joe decides whether to exercise his option or hang onto it for another year. We predict his behavior. To begin, we start at the end. For this example, our initial stock price is \$100, the exercise price is \$110, the interest rate is 5% per year compounded continuously ($e^{\text{rate x time}}$), and the contract lasts three years (see Figure 4). At expiry, after three years, the decision to sell is obvious: if exercise price exceeds stock price, Joe sells. This certainty allows us to step backward and predict what Joe was thinking the year before.

We calculate the payoff (*K* minus *S*) for each of five simulations after three years (paths 1, 4, and 5 are profitable). After two years, four paths are profitable (2, 3, 4, 5) and we determine the expected future payoff for each path (path 1 isn't profitable after two years so Joe doesn't have to make a decision—he waits). We pair these four Year 2 prices with the payoffs at Year 3, discounted by the interest rate ($e^{-.05}$ = .95 per year). To make waiting worthwhile, the payoff must be greater next year than what selling now and putting our profits in the bank would make. We fit a quadratic function to our points to relate discounted future payoff to present stock price.

We plot our function beside the Year 2 payoffs (see Figure 5), and assume that Joe sells only when the immediate payoff exceeds the expected future payoff. For example, the Year 2 payoff for path 5 is \$16, but Joe waits an extra year because he can earn \$2 more. Now we know what Joe will do for all paths that reach Year 2—sell on 2 and 4, wait on 3 and 5.

We now discount these payoffs and pair them with the prices for Year 1 (again, only those that are profitable need decisions). We use Year 2 payoffs for paths 2 and 4, discounted one year's interest, and Year 3 payoffs for paths 1 and 5, discounted two years' interest. As before, we find a quadratic function that estimates expected future payoff and tells us if Joe sells or waits (see Figure 6).

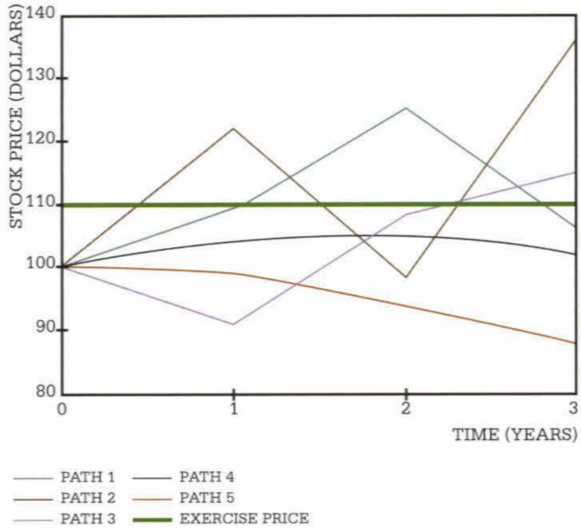


FIGURE 4. Five simulated price paths for the American put option exercisable at three discrete times (expiry after three years). A put option is the right to sell stock at a predetermined price and is profitable when that price exceeds the stock price. The Least-Squares method works backward from the date of expiry, predicting future payoffs from present prices to determine whether an investor sells or hangs onto his stock.

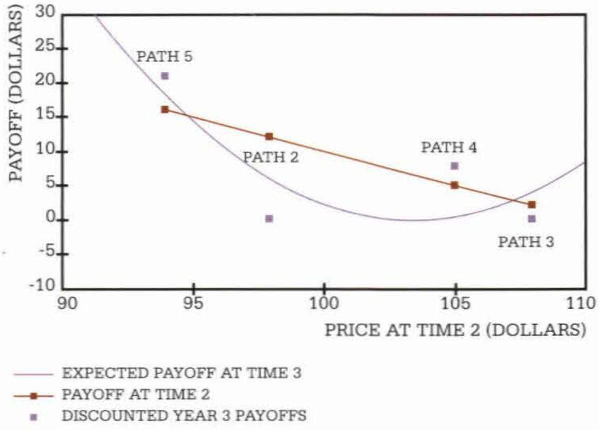


FIGURE 5. A curve is fitted to points marking the discounted Year 3 payoffs and corresponding Year 2 prices for all paths with positive Year 2 payoffs. This curve is compared to the payoffs for selling at Year 2. On paths where the payoff exceeds the expected payoff of waiting (2, 4), the option-holder sells.

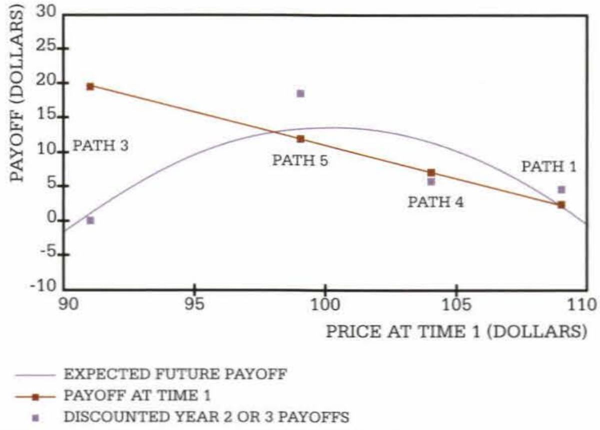


FIGURE 6. The Least-Squares Monte Carlo method is recursive. The results shown in Figure 5 allow us to determine expected payoffs for waiting at Year 1. Year 1 prices are paired with Year 2 and Year 3 payoffs, depending on how long the investor expects to wait and discounted for what interest would have made. This curve and the payoffs from selling at Year 1 allow us to predict the investor's behavior. He sells at Year 1 on path 3, when the immediate payoff exceeds the expected future payoff. At Year 2, he sells on paths 2 and 4. He sells on paths 1 and 5 at expiry. The payoffs from the options at each time ($3.43, $10.83, $18.05, $4.51, $18.86; see Figure 4) are averaged to give an option price of $11.14. This is how much the investor should be charged for the option contract.

After one year, Joe sells only on path 3. After two years, he sells on 2 and 4. On paths 1 and 5, Joe exercises the option at expiry. We calculate the option payoffs at the optimal exercise times and discount them depending on how long we had to wait ($3.43, $10.83, $18.05, $4.51, $18.86; see Figure 4). The average prices the option at $11.14. This is how much Joe should have been charged for the option.

FINANCIAL MARKETS RELY ON SPEED

Predicting the future is always difficult, and pricing options is no easy problem. It's been around for thousands of years—since the ancient Greeks sold options on goods shipped from their ports. Financial markets always need faster and more accurate predictions. When stock prices jump, new option prices must be recalculated immediately; the fastest firm has the advantage.

Least-Squares provides a straightforward way to evaluate the American option. However, although computers have no difficulty fitting curves to data points in present simulations, the large number of paths and time intervals required for accurate pricing demand too much computing time to be practical. Complex options with more complicated dependencies on time and stock price, like the Bermuda, Asian, and Barrier options, will be impossible. Neural networks promise a fast replacement for polynomial fitting. If they live up to their promise, the Least-Squares Monte Carlo algorithm and neural networks together could solve one of finance's most difficult problems, transforming the pricing of American options the way the Black-Scholes formula revolutionized trading of European options.

Mona Sheikh is a second year undergraduate in Electrical Engineering at the California Institute of Technology. *This work was completed with Yaser Abu-Mostafa, professor of Electrical Engineering and Computer Science at Caltech, and funded by the 2000 Caltech Summer Undergraduate Research Fellowship. The author wishes to thank Yaser Abu-Mostafa, Amir Atiya, Malik Magdon-Ismail, and Gentian Buzi.*

FURTHER READING

1 F. Longstaff, E. Schwartz. Valuing American Options by Simulation: A Simple Least-Squares Approach. *Review of Financial Studies* **14**, 113-147 (2001).

2 P. Wilmott, S. Howison, J. Dewynne. *The Mathematics of Financial Derivatives* (Cambridge University Press, Cambridge, 1995).

3 C. Bishop. *Neural Networks for Pattern Recognition* (Oxford University Press, Oxford, 1996).

LIQUID CRYSTALS ALLOW LARGE-SCALE ALIGNMENT OF

CARBON NANOTUBES

BY ABRAHAM HARTE

CARBON NANOTUBES INCLUDE A REMARKABLY versatile set of molecules, which possess extraordinary strength, flexibility, and thermal conductivity, as well as many novel electronic characteristics. Depending on the specific arrangement of its constituent atoms, a nanotube can be a conductor or semiconductor, and a miniature wire, a diode, or a transistor. Some types of nanotubes generate orders of magnitude less heat than comparable copper wires when a current is passed through them. With all of these useful electrical and mechanical properties, nanotubes also have only one-sixth the density of steel.

With a proper technique for assembling nanotubes, such a unique material could have diverse technological applications, including more efficient flat-panel displays and longer lasting batteries. Smaller electronic components made from nanotubes could increase computer speed and memory capacity far beyond current levels. Miniaturization with nanotubes could also lead to unprecedented medical technologies, such as probes and repair devices that could be injected into the bloodstream to diagnose and treat patients.

Formed from a cylinder-shaped arrangement of carbon atoms, a nanotube resembles a rolled sheet of graphite. The typical single-walled nanotube (see Figure 1) is only a few nanometers in diameter and can extend to over 100 microns long. Nanotubes have very different physical properties depending on the orientation of their carbon atoms. Nearly all applications require that groups of nanotube rods be distributed in some specific pattern, constrained by orientation dependence of physical properties or packaging constraints. A useful wire, for example, must be longer than it is wide, and large enough that many nanotubes must go into its construction. Developing nanotube-based technologies therefore requires the ability to easily manipulate which direction a nanotube faces.

Efforts to align carbon nanotubes have made modest advances. Up to a few individual nanotubes can be oriented with an atomic force microscope, but this is inadequate to construct larger devices, which require the quick and cheap alignment of large numbers of nanotubes. Eventually, several alignment methods will be needed to form the nanotube patterns necessary for different applications. Our research develops a technique to quickly and precisely align millions of nanotubes using liquid crystals, the same technology that forms your digital watch display.

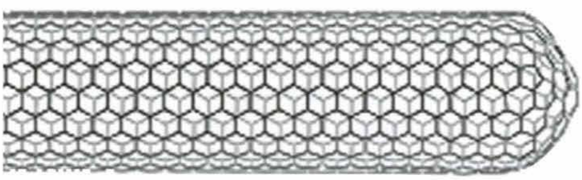


FIGURE 1. Illustration of a typical single-walled carbon nanotube. Although it is only a few nanometers wide, it can be up to 100,000 times longer than it is wide.

LIQUID CRYSTALS INSPIRE ALIGNMENT IDEAS

Laptop screens, alarm clocks, and many other everyday devices use liquid crystals. The liquid crystals we used are composed of small, rod-like organic molecules that maintain their orientations with respect to one another under specific temperature and pressure conditions. As with carbon nanotubes, liquid crystal properties change with their orientation. For example, liquid crystal display (LCD) technology locally rotates liquid crystal molecules to transmit or absorb light and form patterns like the numbers on a digital clock display. Precisely aligning cylindrical liquid crystals for commercial applications has become commonplace. The similarities between cylindrical liquid crystals and carbon nanotubes suggest that analogous methods might work to align carbon nanotubes.

Applying an electric or magnetic field to align liquid crystals is the basic guiding principle behind digital watch displays. In the simplest type of liquid crystal configuration, light from the sun or a light bulb shines through a thin layer of liquid crystal sandwiched between two glass plates with a polarizing filter on the opposite side. Just as a spring absorbs motion in the direction it stretches, cylindrical liquid crystal molecules block out light polarized along their main axis. When no voltage is applied, liquid crystal molecules orient themselves parallel to the glass, so that light transmitted through the liquid crystal layer becomes polarized. The polarizer is oriented perpendicular to this blocking angle, so it absorbs all of the light transmitted by the liquid crystal layer and the liquid crystal layer appears dark. Conversely, to light up the cell, electrodes in the glass apply the electric or magnetic field necessary to rotate the molecules perpendicular to the glass. Since the incoming light now sees the liquid crystal molecules head-on, light transmitted through the liquid crystal layer is not polarized. The polarizing filter transmits all of the unpolarized light from the liquid crystal layer, and the cell appears lit.

The most obvious method to align nanotubes is to simply apply an electric or magnetic field. This approach is not feasible with solid samples because intermolecular forces in a solid are difficult to break. Instead, we suspended nanotubes in water to allow them to rotate freely and applied magnetic fields as high as 2 Tesla. We saw no measurable effect, implying that aligning large quantities of nanotubes requires magnetic fields stronger that we are capable of producing.

Attempting a variation on the first experiment, we dissolved the nanotubes in the common liquid crystal 4-pentyl-4-cyanobiphenyl, abbreviated 5CB. 5CB is a thin liquid at room temperature, making it easy to dissolve nanotubes and create cells. By aligning the liquid crystals using a magnetic field, we expected them to exert a torque on the nanotubes, and consequently align the nanotubes in the same

"[Miniaturization] from nanotubes could increase computer speed and memory capacity far beyond current levels, leading to unprecedented medical technologies."

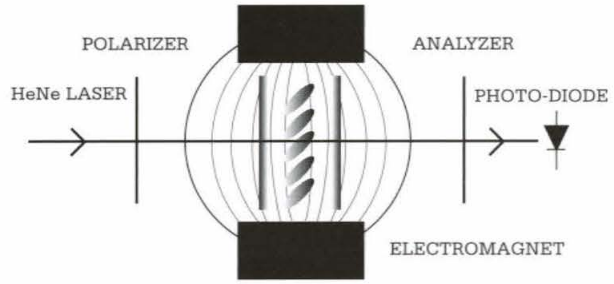


FIGURE 2. Schematic of the setup used to measure the Freedericksz transition. The output intensity of the system is used to interpret the degree of rotation of the liquid crystal molecules away from the optic axis.

direction. Our results suggest that this indeed happened, despite the fact that one liquid crystal is thousands of times smaller and less massive than a nanotube. Since liquid crystals can be as easily aligned with electric fields as with magnetic fields, our results should be identical if electric fields, which are often cheaper to create, are used.

In this experiment, the nanotube solubility in 5CB liquid crystals was measured to be only 10^{-6} grams of nanotube per gram of liquid crystal, translating to 1 nanotube molecule per 10^{9} liquid crystal molecules. Since nanotubes do not dissolve well in any known substance, this poor solubility is not surprising but practical applications would be prohibitively inefficient at these low concentrations. While we did not investigate ways of increasing solubility, there are several promising avenues for further research, suggesting that this is unlikely to be a major hurdle.

MEASURING NANOTUBE ALIGNMENT

To find the magnetic field's effect on the nanotubes, we needed a method to measure nanotube alignment. The resulting setup can measure both the alignment of the liquid crystals and detect its effect on the nanotubes in solution (see Figure 2).

To prepare the experimental setup, we produced several batches of thin glass liquid crystal cells coated with dimethyldichlorosilane. This chemical aligns all the liquid crystals so that liquid crystal molecules directly in contact with a glass surfaces remain perpendicular to the surface. Intermolecular interactions keep the other liquid crystals in the cell oriented with each other, so all of the liquid crystals in the cell will be perpendicular to the glass surface in the absence of a magnetic field.

When a sufficiently strong magnetic field is applied perpendicular to the molecules, they start to turn away from the surface. The response of the molecules is not directly proportional to the field strength, so below a certain critical field there is no observable effect. A field slightly stronger than this will dramatically rotate the molecules. This sharp change, called the Freedericksz transition, can easily be observed through differences in the indices of refraction along different directions in the cell—a property called birefringence. Liquid crystal is highly birefringent: as the molecules

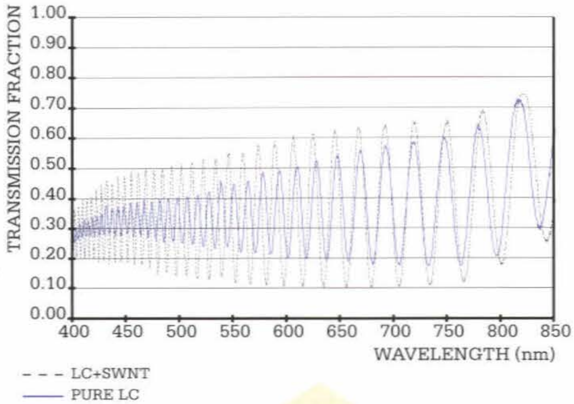


FIGURE 3. Light transmission through liquid crystal cells, with and without nanotubes in solution. At troughs most of the polarized light is perpendicular to the aligned liquid crystal and nanotubes. The polarized light at peaks is parallel to most of the molecules. The larger amplitude of the system with nanotubes implies that the nanotubes are aligned with the liquid crystal and increase the amount of light blocked or transmitted.

line up with the magnetic field, the index of refraction changes significantly in that direction, "compressing" passing light most in one direction.

The Freedericksz transition is a simple indicator of microscopic interactions: any change in the properties of the liquid crystal cell alters the transition field. If adding carbon nanotubes changed the transition field, this would indicate a strong interaction between the materials, suggesting the liquid crystal meets resistance from the nanotubes as it aligns with the field.

Saturating the solution with nanotubes produced a consistent 10 to 30 percent increase in the transition field. It is also interesting to note that liquid crystal supersaturated with nanotubes required three times the field strength to induce the same amount of rotation as liquid crystal without nanotubes, suggesting a reaction of some kind even if the nanotubes are not properly dissolved. The increase in the Freedericksz transition implies that the liquid crystal is indeed acting on the nanotubes, since at a concentration of 10^9 grams of nanotube per gram of liquid crystal the mere presence of nanotubes and dilution of the liquid crystal certainly could not account for the observed variations!

SCATTERED LIGHT RESOLVES MEASUREMENT

The upward shift we saw in the Freedericksz transition field could have been caused either by the nanotubes aligning along with the rotating liquid crystals, or by a randomly-oriented collection of nanotubes impeding the motion of the liquid crystal. Further measurements were needed to distinguish between these possibilities. As with liquid crystals, nanotubes absorb light polarized along their axis. If the nanotubes were aligning with the liquid crystals, then we might observe a larger difference between light transmittances of cells for different polarizations after nanotubes were added to the solution.

To test this, two thick quartz cells were produced in which the liquid crystals were aligned parallel to the cell walls. One of these was filled with pure 5CB liquid crystals and the other contained a saturated 5CB-nanotube solution. These samples were placed between two crossed-polarizing filters, and the transmission of the system was measured as a function of wavelength. The percent of light transmitted in each setup is shown in Figure 3. Birefringence effects determine where the minima and maxima will be, while the total amplitude of oscillation derives from losses in the cell, which in turn will depend on the alignment of the liquid crystal and nanotubes.

"As with carbon nanotubes,

liquid crystal properties change with their orientation."

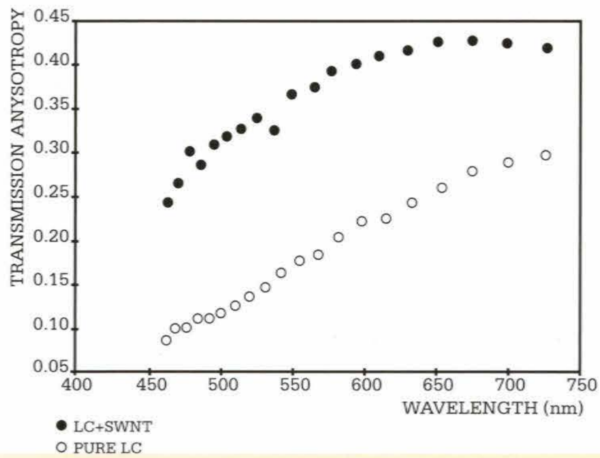


FIGURE 4. Peak to trough ratio of light transmissions from Figure 3. The higher direction dependence of the nanotube solution also indicates that the nanotubes are aligned and blocking light preferentially in one direction.

Figure 3 shows that in both cases the locations of minima and maxima are similar, indicating that the nanotubes did little to change the cell's birefringence, but the much greater amplitude of the nanotube solution oscillation shows that this cell has very different polarization properties, suggesting that the nanotubes are in fact aligned.

The differential loss parameters of the two solutions as a function of wavelength are shown in Figure 4. This quantity represents anisotropy, the difference in the brightness of light shined through the cell from different directions. If the liquid crystals or nanotubes in the cell were aligned randomly, the anisotropy would be low since the molecules would not be absorbing more light in one direction than another. If, on the other hand, the molecules were mostly aligned in one direction, polarized light perpendicular to the main axis of the molecules would be mostly absorbed. When the light comes into the cell from a different direction, there would be much less absorption, so a higher anisotropy indicates a greater alignment of molecules in the cell. Figure 4 shows that adding nanotubes to the liquid crystal did indeed increase direction-dependent losses in the cell, providing even more evidence that the liquid crystal did in fact align the nanotubes.

ENVISIONING PRACTICAL NANOTUBE ASSEMBLY

Our experiments strongly support the idea that liquid crystal molecules can be used to align carbon nanotubes. We expect that the techniques described here for carbon nanotube alignment will be extended to the smallest scales by depositing liquid crystal-carbon nanotube solution on a substrate with conventionally-etched microelectrodes to apply the fields and allow input/output capability. The liquid crystal could then be evaporated away while holding the nanotubes in place. Without self-constructing circuits, this fabrication technique may not be precise enough for nanoelectronics, but dynamic alignment of nanotubes in liquid crystal is still a major step forward.

The problem of preserving a nanotube assembly might be partly solved if nanotubes could retain their alignment after the liquid crystal was frozen. For example, aligning and freezing large concentrations of nanotube in liquid crystal could produce a large crystal with some nanotube properties. Since creating a large chunk of nanotubes is currently impossible, freezing may prove an effective alternative for generating strong materials or ultra-efficient heat sinks. Exactly how nanotubes will be assembled for practical applications remains to be seen, but each new advance in nanotube assembly brings us closer to the day that computers fit on fingertips and full-sized televisions run on clock batteries.

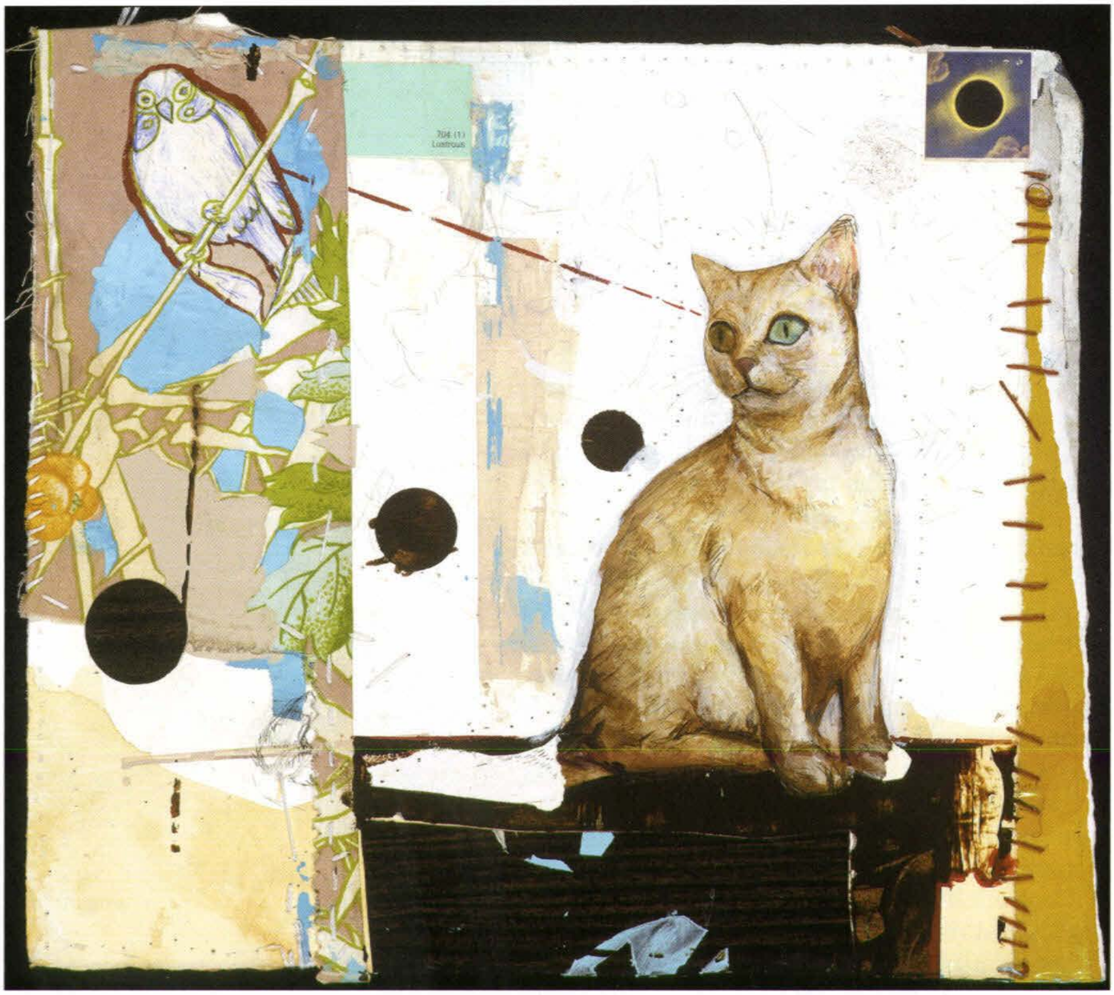
Abraham Harte is a fourth year undergraduate in Physics at the California Institute of Technology. *This work was completed with Yuen-Ron Shen, Professor of Physics at the University of California, Berkeley, and funded by the 2000 Caltech Summer Undergraduate Research Fellowship and the National Science Foundation. The author wishes to thank Y. R. Shen, Seok-Cheol Hong, Masahiro Ishigami, and Alex Zettl.*

FURTHER READING

1 Z. F. Ren *et al.* Synthesis of Large Arrays of Well-Aligned Carbon Nanotubes on Glass. *Science* **282**, 1105-1107 (1998).

2 B. I. Yakobsen, R. E. Smalley. Fullerene Nanotubes: C1,000,000 and Beyond. *American Scientist* **85**, 324-337 (1997).

CAT EYES BY RYAN SMITH



Cat Eyes
Mixed Media, 13"x 13"
Artist's Private Collection

A wild animal's glowing eyes looked back at me from the darkness of the garage, viewing the world from an ethereal angle. The light switch revealed my cat, but those glowing eyes made me wonder how cats see the world.

Ryan Smith
Art Center College of Design