Contents

List of Tables

List of Figures

Chapter 1

Introduction

"Somewhere, something incredible is waiting to be known"

—Sharon Begley (though often misattributed to Carl Sagan)[1]

Six years in the making, my PhD in Applied Physics has focused on understanding how electrons interact in thin films of complex oxide materials. Although the day-to-day labor felt mundane at times, my research took place at the frontier of human knowledge and ability. I synthesized new materials never before seen in history, in layers so thin that I counted their atoms one by one. Then I took those new materials and measured them in extreme conditions: at temperatures less than a degree above absolute zero, in magnetic fields stronger than on the surface of the sun, and in pressures lower than those faced by spacewalking astronauts.

However, before I describe my research projects in detail (Chapters 3 & 4), I'll discuss why this sort of research is worth doing (Chapter 1) and the modern experimental methods that have made it possible (Chapter 2). At the end, I'll tie these threads together and discuss their implications for the future (Chapter 5).

1.1 Technological innovation

The ostensible purpose of my PhD research is to better understand materials in order to one day develop new technologies. This first section is devoted to a broad discussion of technological innovation: why it matters, how it is constrained, how it can be analyzed, how it can be predicted, and where my research fits in.

1.1.1 Technological innovation drives human prosperity

By nearly every metric, we humans are better off today than at any prior point in history. Life expectancy is longer[2, 3], farms are more fertile[4, 5, 6], violence has declined[7], and poverty has

become the exception rather than the rule[8].

Notably, this tremendous human progress, both economic[9] and moral[10], was *not* a story of slow and steady improvement across the ages. Rather, the majority of progress was concentrated into an extraordinary era so brief and so recent that it spans fewer than ten generations. Called the Industrial Age by some scholars, this era is the one we live in now. Although it is difficult to explain why the Industrial Revolution began around 1800 in the United Kingdom,[1] its impact on the world is indelible. Supported by new scientific understanding, new technologies were invented to amplify human labor, allowing people to flourish in unprecedented ways. The development of new technologies is the source of our modern prosperity, and developing yet newer technologies will be the source of our future prosperity.

Technological innovation drives economic growth

In the neoclassical Solow-Swan model of economic growth, the long-run growth of an economy is driven by the rate of technological innovation.[4][22, 23] Although the Solow-Swan model treats technological innovation as an exogenous variable, it's clear that the rate of technological innovation

[1]Explanations for the spark that lit the Industrial Revolution include newly high levels of literacy, printing, population, seafaring, scientific method, inherited intelligence, state competition, cultural superiority, surface coal, high wages, and even fashion[11, 12]. However, none of these explanations are by themselves unambiguously convincing, and any satisfactory explanation should also explain why the Industrial Revolution did not begin in China, which shared many of these features centuries earlier, during the Ming and Qing dynasties.[13, 14] History is often too multifactorial to accurately compress into simple narratives. Even the concept of the Industrial Revolution is not as simple or discrete as its name suggests.

[2]Technically, this should be world GP (gross product), rather than world GDP, to reflect the fact that that some wealth is generated outside of national borders, such as in international waters, Antarctica, and space. Even world GNP slightly mismeasures total output, since some workers hold no citizenship anywhere. But because all of these measures are about equal, I chose to title the plot with GDP for its familiarity.

[3]Surprisingly, even information capital appears to depreciate, prompting the invention of concepts such as the half-life of knowledge,[17] the half-life of facts,[18] and software rot[19] (not to be mistaken for actual physical degradation of storage media).

[4]Actually, in the Solow-Swan model, there is one other exogenous driver of the steady-state economy: population growth. Because of this second variable, it is more correct to say that technological innovation drives *per capita* long-run economic growth. And while technology-driven economic growth dwarfs population-driven economic growth today, for most of recorded human history it was actually the other way around. Population growth, not technological innovation, contributed the most to pre-industrial economies. (Then again, much of the explosion in human population after

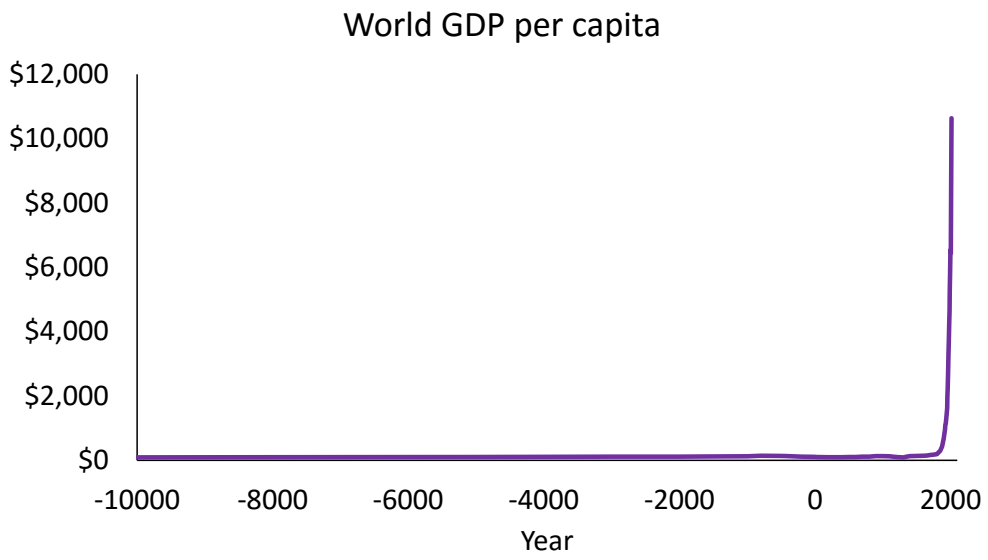


Figure 1.1: World GDP[2] over the past 12,000 years (measured in 2000 Int'l.$).[15, 16] Although reducing the complexity of human experience down to a single number is impossibly crude and full of uncertainty, the shape of this plot still illustrates the superexponential progress of modern humans. Even when plotted on a log scale, this curve retains the same hockey stick-shape, with the same inflection point around 1850. (Fun fact: although the Gregorian calendar has no year 0, astronomical year numbering defines 1 BCE to be year 0.)

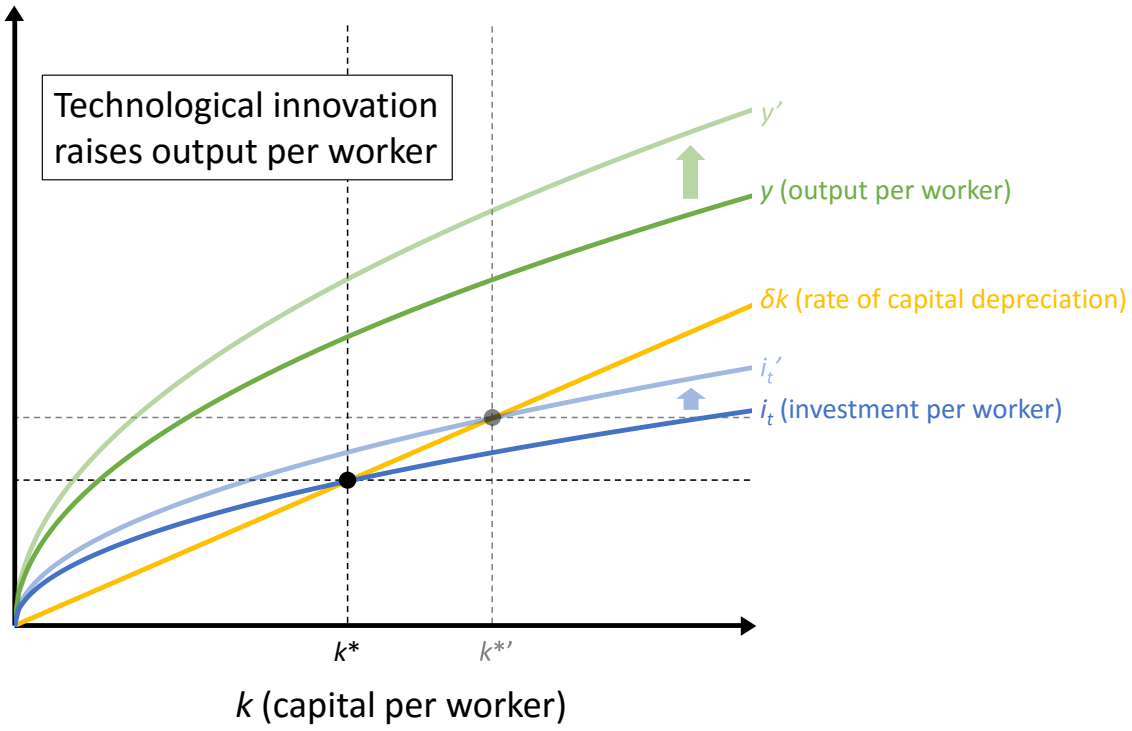


Figure 1.2: Because capital has diminishing returns and also depreciates,[3] economic growth cannot be permanently sustained by capital accumulation. According to the neoclassical Solow-Swan model of economic growth, the steady-state output per capita of an economy can only be increased through improved technology.

depends on how much a society chooses to invest in technological R&D. Societies that invest more in technological R&D will tend to be richer in the long run.

Nevertheless, technological innovation is not purely a function of investment. Throwing money at a problem is no guarantee that a solution will be found: indeed, some problems may have no solution. A perpetual motion machine will never be built regardless of how much money is invested, because the laws of physics limit what can be achieved.[5,6] Therefore, it is important for a society to understand these physical limits and strategically target R&D investment toward technologies with the greatest potential.

1.1.2 Technological innovation is constrained by physics

> *"When a distinguished but elderly[a] scientist states that something is possible, he [or she] is almost certainly right. When he [or she] states that something is impossible, he [or she] is very probably wrong." "[a] In physics, this means over thirty."*
>
> — Arthur C. Clarke, *Hazards of Prophecy*[25]

What are the ultimate physical limits of technology? Answering this question with confidence

5000 BCE[20, 21] was driven by agriculture, itself a technological innovation.)

[5] Then again, let's not be *too* hasty. Perhaps even free energy will be possible in the far future. The continuous expansion of the universe is constantly increasing the energy of bound matter, at least according to today's best models.

[6] Also, and this may be a bit of a tangent, it depends what you mean by perpetual motion machines. One clever way of classifying perpetual motion machines is by which of the first two laws of thermodynamics they violate: the first law, the second law, or neither law (or, for completeness, both laws, but such excessive disregard for the laws of physics is rare).[24] A perpetual motion machine of the first kind violates the first law of thermodynamics, doing work without input of energy. A perpetual motion machine of the second kind violates the second law of thermodynamics, doing work with the input of thermal energy (this violation is more subtle because while it obeys energy conservation it does not obey entropy/information conservation). And a perpetual motion machine of the third kind is one that simply moves forever without violating any laws of thermodynamics. In my opinion, a hydrogen atom qualifies as a perpetual motion machine of the third kind. Friction cannot slow down a quantum mechanical system in its ground state. Some might argue that a hydrogen atom still doesn't count as a perpetual motion machine because eventually it, like everything else, will decay or be sucked into a black hole or meet some other fatal end. However, I dislike this particular criticism because it implies that nothing can be a perpetual anything, robbing the word perpetual of its value.

warrants caution.[7] Many answers given by brilliant minds of history have proved to be comically wrong, sometimes within mere years of their assertion.[26] Even when apparent limits are justified by correct laws of physics, the limits have a funny way of being circumvented in unanticipated ways.

In 1920, a New York Times editorial[27] proclaimed that space rockets were impossible because nothing can push against a vacuum. The editorial correctly understood Newton's third law and the conservation of momentum, but failed to imagine expelled propellant.[8] A few years later, in 1926, Professor A. W. Bickerton also asserted the impossibility of space rockets but in a different way, arguing: "The energy of our most violent explosive—nitro-glycerine—is less than 1,500 calories per gramme. Consequently, even had the explosive nothing to carry, it has only one-tenth of the energy necessary to escape the earth.... Hence the proposition appears to be basically impossible."[25][9] To his credit, Professor Bickerton correctly interpreted the laws of gravity and energy conservation, but what he failed to imagine is that the fuel need not be carried to space along with the payload. Despite these and similar proclamations of spaceflight's impossibility, only a few decades later, *and with no fundamental paradigm shifts in mechanics or chemistry*, humankind set foot on the moon.

[7] Arthur C. Clarke wrote a marvelous essay, Hazards of Prophecy, which looks at why and how so many experts were wrong about the future, with many examples coming from predictions made about flight and spaceflight in the early 1900s. He classifies poor pessimistic predictions into two categories: failures of imagination, in which the relevant laws of physics were not yet known (but still could have been imagined), and failures of nerve, in which the relevant laws of physics were known (but other barriers were perceived).

[8] The New York Times eventually printed a retraction in 1969, as Armstrong, Aldrin, and Collins were on their way to the moon.[28]

[9] Although Bickerton's errant prediction has been pointed out by Arthur C. Clarke and other authors, according to Wikiquote, it's possible that this quotation of his was arguing specifically against projectile launches into space rather than rocket launches into space. And if so, while Bickerton's prediction was still technically wrong—in 1966, a projectile was indeed launched into space (though not orbit)—in practical terms, Bickerton was correct. Despite a history of attempts, projectile launches into space are not presently feasible due to the tremendous drag (which costs energy and makes aiming hard), the tremendous forces experienced by the projectile (meaning the payload needs to be extremely rugged and human-free), and the remaining requirement for a propulsion system so that the payload can change course into an orbit (otherwise it will crash, which is guaranteed to happen in a gravitationally bound two-body system without propulsion).[29]

These stories[10] illustrate that even when the relevant laws of physics are known, claims of impossibility by eminent experts are by no means guaranteed to be true.[11] These stories also raise an important epistemological question: If expertise in physics is not enough to avoid mistaken proclamations of physical impossibility, then upon what foundation can today's physics experts ground their confidence in their own judgments, given that their judgments are similarly undergirded by expertise in physics? Are there any technological impossibilities we can be 100% sure of? Or is it just too fundamentally difficult to reason about technologies not yet invented?

Perhaps the most classic example of physics-limited technology is the heat engine, which is limited by the laws of thermodynamics to operate below the Carnot efficiency ($1 - T_{cold}/T_{hot}$). Yet even the esteemed Carnot efficiency limit may not be absolutely ironclad. In principle, a good enough measuring system may enable work to be extracted from waste heat, thereby circumventing the Carnot limit. This idea was first conceived by James Clerk Maxwell in 1867 and later came to be known as Maxwell's demon.[31][12]

A second well-known example of a physical limit is the speed of light,[13] which sets a minimum

[10]Despite including these stories about poor spaceflight predictions, I wish to emphasize how challenging it actually is to learn by grading predictions of the past. So many people have made so many claims throughout history, that it's equally easy to find to foolish quotations as prescient ones. Without a systematic and unbiased aggregation of all past predictions, everything we see is coming through a selective filter. And without knowing how that filter of history distorts the data, it's difficult to know what people believed and how often they were wrong.

[11]Of course, many claims of impossibility *have* stood the test of time, including those regarding time travel, faster-than-light communication, invisibility, antigravity, etc. Antigravity research has a particularly interesting history, having been the subject of the Gravity Research Foundation, which was established by Massachusetts businessperson Roger Babson, who blamed gravity for his sister's childhood drowning.[30] After Babson's death, the pseudoscientific foundation transitioned into legitimacy and now exists to award annual prizes for essays on gravity. These prizes have been won by a number of scientific luminaries, including Stephen Hawking, Roger Penrose, Julian Schwinger, Frank Wilczek, and George Smoot III (the only winner of both a Nobel prize and the TV show *Are You Smarter than a 5th Grader*).

[12]In fairness, I am aware that a number of different experiments have recently claimed to measure the Landauer limit, which has been proposed as a way of reconciling Maxwell's demon with the second law of thermodynamics. However, from an admittedly less-than-thorough reading, I remain personally unconvinced.

[13]Confusingly, the speed of light is a speed limit on everything, not merely light. And extra confusingly, light can travel below the speed of light, and in fact, because no true plane waves exist, it always does. (Though it sort of depends on how you define speed. If I run a quarter-mile loop at a track in one minute, is my speed nonetheless zero, given that I haven't gone anywhere? There

on the time it takes to get from one place to another.[14] As anyone who has driven down a highway during rush hour knows, we are far from hitting the speed-of-light limit when it comes to the transportation of matter. However, when it comes to the transportation of information, our technology is already brushing against this limit, as you might notice when communicating with someone on another continent. Because it takes about 100 milliseconds for light to go around the Earth, any communication to someone on the opposite side of the Earth will always have a minimum lag of 100 milliseconds, no matter what future clever device our descendants can come up with. Even a billion years hence we can be confident that communication technology will be a little laggy on Earth-scales ($\sim$ 0.1 seconds), annoyingly laggy on Earth-to-moon scales ($\sim$ 1 second), and awfully laggy on Earth-to-Mars scales ($\sim$ 1000 seconds).[15]

It's worth noting that the speed of light is not only a limit on technologies that traverse long distances. It's also a limit on technologies that go very fast. In the time it takes your computer's 3 GHz computer chip to go through one clock cycle (1/3 of a nanosecond), an electric signal only has time to travel along a few centimeters of wire.

In any case, it seems hard to imagine a world where the speed of light can be exceeded. And yet, in the 1920s, it was hard for many to imagine a world where rockets could ferry people to the moon. But a few decades later that impossibility became reality. Though it sounds utterly ridiculous today, who are we to say that the speed of light will not be circumvented in the future? And more generally, what truths of today can we epistemologically guarantee to be immune to unforeseen scientific developments of the future?[16]

In general, it is difficult to predict the possibilities of technology. Sometimes the laws of physics change,[17] and even when they don't, technological innovation has a curious way of circumventing

was no displacement over a duration, after all.)[32]

[14]By the way, pay no mind to any claims that entanglement or quantum teleportation or wave function collapse prove that the speed of light of light can be exceeded. Many interpretations of quantum mechanics (such as many worlds/decoherence) can explain these phenomena in a local way that never violates the speed of light.

[15]Then again, even this minimum lag may one day be circumvented in ways that still obey the known laws of physics. Perhaps a computer could predict what your conversation partner would say ahead of time. Or perhaps our conscious perception of time could be slowed, making a 10-minute delay unnoticeable. Or perhaps we might even attain the power to warp space itself.

[16]Here, I want to emphasize that I don't regard the truth of the 1920s to be equivalent to the truth of today. Today, at least in my present-biased judgment, we are unambiguously more correct, further along, and closer to the absolute truth than in the 1920s. Asimov wrote a terrific essay on this topic titled The Relativity of Wrong.[33]

[17]Rather, our understanding of them changes. Though for all we know they might change too.[34, 35]

apparent obstacles. However, not all laws of physics are equally vulnerable to being overturned by unforeseen scientific revolutions. And in particular, there is a certain class of physical law that is especially impervious to unanticipated changes: conservation laws.

1.1.3 The most fundamental constraints on technology are conservation laws

"It is only slightly overstating the case to say that physics is the study of symmetry" —Phil Anderson, More is Different[36]

In physics, there are some quantities that are conserved and some quantities that are not. Conserved quantities, such as energy or momentum, are quantities that can never be created nor destroyed,[18] only transferred from one form or object to another. These conserved quantities stand in contrast with quantities like redness or photon number or temperature, whose totals we *can* change over time. So while cleverness in engineering may finagle a technology to produce red or photons or temperature, conservation laws mean that no cleverness will ever invent a technology that can truly produce energy (i.e., output more energy than is put in).[19]

The question of why some quantities are conserved and some are not has turned out to be one of the deepest and most fundamental questions in physics. Although the mathematical machinery needed to answer this question existed since the time of Newton, it was not until 1915 that Emmy Noether, while working on Albert Einstein's theory of relativity, recognized that conserved quantities have a deep origin: they are a mathematical consequence of symmetry. Specifically, she showed that if a system's dynamics can be described by a Langrangian, then every continuous symmetry will have an associated conserved quantity, which remains constant over time.[20][38]

Unlike other laws of physics, which derive from observation and experimentation, conservation laws are a mathematical fact. This means that no matter what future experiments might discover, as long as the universe's laws of physics stay symmetric, the conservation laws will never be overturned.[21] Conservation laws are some of the surest limits we have on technology.

[18]Except by big bangs and dark energy and other mysterious things that warp the space-time fabric.

[19]I mean, to be fair, nothing is created for free. Even increasing the redness of an object usually requires work of some sort. But this work is incidental and not essential to the changing redness of the object. (It's actually quite hard to talk about doing things without also referencing the energy needed to do them.)

[20]For a proof of Noether's theorem, see Appendix B.

[21]Here, I'd like to point out that this argument has a subtle loophole, or at least a subtle qualification. Although it's true that conserved quantities are guaranteed to exist if physical laws are

Below is a table of eight symmetries and their associated conserved quantities.[22] These eight conserved quantities lay a foundation for thinking about the budgets that all technologies must work within.

Conserved Quantity	**Associated Symmetry**
Mass-energy	Time invariance
Momentum	Position invariance
Angular momentum	Rotation invariance
Charge-parity-time symmetry	Lorentz invariance
Electric charge	Gauge invariance
Color charge	SU(3) gauge invariance
Weak isospin	$SU(2)_L$ gauge invariance
Probability	Probability invariance

Table 1.1: The eight exactly conserved quantities and their associated symmetries.

In addition to these eight conserved quantities, there is one more that deserves mention, even though it does not arise from a continuous symmetry: the conservation of information. Jokingly termed the minus first law of thermodynamics by Jos Uffink and popularized by Leonard Susskind,[23][39] the conservation of information is the idea that distinctions between states cannot be lost or erased over time. What this means practically is that when you erase a bit on your hard drive, that information is not truly lost from the universe, but rather ejected into the rest of the world as scrambled thermal vibrations. The conservation of information directly follows from the

symmetric, there is still room for our understanding of these conserved quantities to evolve. For example, the translational symmetry of physical laws implies that momentum is conserved. But it took physicists a long time to realize that true momentum is not just the mechanical momentum, $m\vec{v}$, but includes a contribution from the electromagnetic field as well. So although the law of momentum conservation never changed, our understanding of what momentum is did in fact change.

[22]There are approximately conserved quantities and symmetries beyond these eight, but these eight are currently believed to be exactly conserved. (Though I confess, the split between exact and inexact laws may be somewhat arbitrary. As far as I can tell, the existence of dark energy means that conservation of mass-energy is slightly violated, yet for some reason mass-energy is grouped in as one of the eight exact laws. And I suppose the Big Bang violates all of these laws.)

[23]At first, thermodynamics had only three laws (the first, the second, and the third). Later, when physicists recognized a deeper principle of thermodynamics, they inserted it as the zeroeth law of thermodynamics, ahead of the first three. By that logic, if the conservation of information is a yet deeper realization, then it ought to be instated as the minus first law.

laws of physics being deterministic[24] (or, stated in a narrower way, from the unitarity of the time-evolution operator in a Hamiltonian mechanics, leading to conservation of phase space distribution in a bounded phase space by Liouville's theorem).

Together, these nine conserved quantities (the eight symmetry-derived quantities plus information) act as fixed budgets, limiting what we and our technology can do. For example, energy conservation means that the energy budget of our solar system is fixed. We can try to harvest more of the energy budget that's out there, and we can try to use our energy more efficiently, but the ultimately available energy budget will never grow, no matter what fantastic technologies we might invent in the future.

Of these nine budgets, some are more constraining than others, at least in today's world. From now on, I will narrow the focus to the three budgets that seem to be the most relevant today: matter, energy, and information.[25]

Because we cannot create matter, energy, or information out of thin air, many of our technologies are devoted to transporting these things around to where they're needed, in the form they're needed, at the time they're needed. This insight brings us to the next section, which discusses a technology framework matrix based on the outer product of these conserved quantities and ways to get them to where they're needed.[40, 41, 42, 43]

1.1.4 Technology can be classified by constraint and function

Classifying technology is difficult. Even putting aside the thorny practical problems presented by definitions and their edge cases, classifying technology remains philosophically challenging because there are so many apparent ways to do it.

One perfectly valid classification of technology might group technologies by the materials they're made of (e.g., stone tools, wooden tools, metal tools, etc.).[26] This classification is sensible for archaeologists. Another option might classify technologies by the time they were invented (e.g.,

[24]Though some people incorrectly believe that quantum mechanics is necessarily non-deterministic, there *are* fully deterministic interpretations of quantum mechanics (such as the many worlds interpretation, which argues that all 'possible' outcomes occur simultaneously in a superposition).

[25]The momentum budget is less important because the Earth acts as a giant momentum sink, allowing us to 'create' apparent momentum by merely pushing off the ground. Also, because momentum is a signed, directional quantity, that means we can 'create' large amounts of momentum in one direction as long as we 'create' large amounts of opposite momentum too. These two factors make momentum less of a constraint on technology than energy. Nevertheless, in space, unlike on Earth, it is important to manage momentum budgets carefully.

[26]This is how archaeologists have done it. The periods of history we know as the stone age, the bronze age, and the iron age are not demarcated social structure or wealth or knowledge, but simply

prehistory, classical antiquity, industrial revolution, etc.). This might work well for historians. And a third possible classification might be to group technologies by the first letter of their names (e.g., starts with A, starts with B, starts with C). This third scheme is obviously less useful - unless you a reference librarian - but it is nonetheless logically valid. And when you consider schemes that may be semantically silly but logically valid, the classification possibilities become endless.[27]

So what should make us prefer one scheme over any another? Well, the answer to that question depends on what we are trying to accomplish. Ultimately, the purpose of any classification system is to make it easier to describe patterns by compressing the space of things (technologies, in this case) into the smaller space of classes. And like any good lossy compression algorithm, we'd prefer to lose as much noise as possible while retaining as much useful signal.[28] I suggest that a maximally useful technology categorization scheme would possess the following qualities:

- Its categories would be mutually exclusive (so no technology is twice classified) and collectively exhaustive (so no technology is left unclassified).
- Its categories would cluster technologies that behave similarly (otherwise, the categories would have little use).[29]
- Its categories would be comparable in size and importance (too many tiny and unimportant categories would be distracting and unhelpful).
- Its categories would not number too many (as to be overcomplicated) or too few (as to be trivial).
- Its categories would be abstract enough to apply to technologies of the past and the future (so that progress can be tracked by comparing them over time).

---

by the materials of excavated artifacts.

[27]Well, maybe not technically endless. In a mathematical sense, a technology classification scheme is a function that maps elements from the space of individual technologies to the smaller space of technology classes. If one assumes that the space of technology classes, c, is no greater than the space of individual technologies, t, then the number of isomorphically unique technology classification schemes is bounded by c^t. However, such a simple expression belies its truly gargantuan exponential size. You only need to have 80 technologies or so before the combinatorial explosion of classification schemes outnumbers baryons in the observable universe, commonly guessed to be 10^{80} or so.[44]

[28]In fact, with the insight that classification is a form of information compression, one can use Shannon's information entropy to mathematically formalize a way to rank arbitrary classification schemes.

[29]For example, a classification scheme that categorized technologies by their first letter would mostly satisfy the other objectives and yet be completely useless.

Finding a classification scheme that satisfies these six simple properties sounds easy enough, right? After all, there are a gazillion possible schemes and only six constraints. Unfortunately, it turns out to be impossible to construct a classification scheme that simultaneously satisfies all of these priorities because hidden within them lies a fundamental incompatibility.

In general, there will always be tradeoffs whenever you try to construct a taxonomy of technology. On the one hand, to make your categories most useful, you want them equally populated and detailed. This entails basing your classification scheme on the technologies of the present day. However, the weakness of this approach is that when things change (e.g., the digital revolution and the rising important of electronics/computers), your classification scheme falls behind and loses descriptive power.

To counter this downside, you might want a classification scheme based on abstract theoretical principles (i.e., technologies based on matter vs energy vs information) that works for all time periods. But then you end up with the opposite problem, where some of your categories are far more important and detailed than others.

Overall, it's a problem with no perfect solution. Even a scheme that tries to be flexible by adapting over time will suffer downsides, such as a lack of continuity, which stymies the ability to make meaningful temporal comparisons.

Despite there being no perfect classification scheme, I do want to share one scheme that is particularly elegant, devised by German engineer-philosophers in the late 70s.[40]

Remember how in the prior section we explored the idea that conserved quantities act as fixed budgets, constraining what we can do? Technology, in a sense, is like an accountant, manipulating these budgets by shifting entries in the columns of God's universal spreadsheet. For example, an airplane's turbojet engine is a technology that shifts chemical potential energy of ancient pressurized algae into the gravitational potential energy of me and my duffel bag.

Given this metaphor of technology as an accountant who manipulate budgets, we can classify the accountant's work in a straightforward way just by examining the phrase 'manipulates budgets.' First, what are the important budgets and second, what are the important ways they can be manipulated? As discussed before, there are three important budgets for technology: matter, energy, and information. And you can consider three important methods of manipulation: transformation, transportation, and storage. With these two dimensions—budgets and manipulations—and each of their three categories, we can construct a 3x3 matrix of technology categories, as shown below.

Figure 1.3: Luca Pacioli, also known as "The Father of Accounting and Bookkeeping."[45] He was the first publisher of the double-entry system of bookkeeping, as well as a teacher and friend to Leonardo da Vinci.

	Transformation	Storage	Transport
Matter	Mills	Warehouses	Ships
	Blast furnaces	Silos	Trains
	Refineries	Tanks	Planes
	All manufacturing	Containers	Trucks/cars
Energy	Generators	Dams	Wires
	Engines/motors	Thermal storage	Hydraulics
	Lighting	Batteries	Belts
	HVAC	Capacitors	Axles
Information	CPUs	Hard drives	Wires
	Sensors	Optical discs	Fiberoptics
	Brains	DRAM	Radio waves
	Slide rule	Books	Sound waves

Table 1.2: A 3x3 classification of technology, based on constraint and function. The nine categories are filled with examples of common technologies.

Weaknesses of the constraint-function classification of technology

Although the above classification of technology is attractive for its simple, fundamental approach, I don't want to pretend that it is free of weaknesses.

One problem I see with this classification scheme is the distinction it draws between energy and matter. While this makes sense for electrical energy and a few other types,[30] which are carried without mass,[31] most forms of energy (fossil fuels, food, nuclear fuel) *are* carried by mass. Chemical energy requires chemicals. Nuclear energy requires nuclei. Therefore, transporting these forms of energy necessarily requires transporting chemicals and nuclei. So when you attempt to classify a technology like coal-carrying train cars, it's not so clear whether it belongs under transportation of matter or transportation of energy.

Blurry division can even be a problem when it comes to information and matter. Although plenty of information is carried masslessly over copper wires and fiberoptic cables and even through the air, information is also sometimes carried in forms with mass, such as books or hard drives. (Semi-related fun fact: when it comes to transporting large amounts of data, a truck full of hard drives is often faster and cheaper than a lightspeed fiberoptic cable.[46])

Along the other axis of classification (the classification by function), the distinctions are also

[30] Along with electricity, energy can also be carried masslessly by light, belts, water waves, or sound waves, to name a few.

[31] Well, mostly without mass. Einstein's $E = mc^2$ says that energy by definition always carries a little mass.

potentially blurry. For instance, storage just seems to to be the base case of the other two: transformation with zero change, or transportation across zero distance. In fact, in real life, transportation is often just a storage container with wheels and a diesel engine. With all the diversity of manufacturing stuffed into the transformation column, is it consistent or useful to give two separate columns to boxes and boxes with wheels?[32] Maybe yes, but it's just not so clear to me.[33]

And lastly, where do service sector jobs fit into this classification? For instance, into which of these nine boxes does a butler fall? Does laying a fork on the dinner table fall under transportation of matter? Again, the answers to these questions are not so clear.[34]

Despite these weaknesses, I nonetheless like this classification scheme. Because it's so abstract, it's an especially good framework for thinking about future technology. If a whole new type of technology is invented, odds are it won't fit neatly into the NAICS (North American Industry Classification System),[51] but it is likely to fit into this 3x3 matrix framework.

How is technology classification relevant to my graduate research?

At this point, you may be wondering why my dissertation has digressed so far from its stated purpose of describing my graduate research. How did we get to technology classification and how is it in any way relevant to my research?

Physicists know the power of reasoning from first principles. So when it comes to thinking about how to improve the lives of humans, my reasoning starts from first principles. As I alluded to in the first section about economic growth, technological innovation is the reason we humans have it so good today. Human well-being was revolutionized with technologies like agriculture, cars, and

[32]Speaking of boxes and boxes with wheels, if you'd like to read more about the global shipping industry I recommend the three following books, in order of readability: *Ninety Percent of Everything: Inside Shipping, the Invisible Industry That Puts Clothes on Your Back, Gas in Your Car, and Food on Your Plate*,[47] *The Box: How the Shipping Container Made the World Smaller and the World Economy Bigger*,[48] and *Prime Movers of Globalization: The History and Impact of Diesel Engines and Gas Turbines*.[49] In their own way, each changed the way I see the world.

[33]Perhaps the best approach I see for defining these categories from first principles is as follows: Storage is the category where positions in space don't change over time. Transportation is the category where positions in space change over time macroscopically. And transformation is the category where positions in space change over time microscopically (or at least on a component-level scale such that it changes the identity of the material or object).

[34]In general, classifying services is harder than classifying manufactured goods. This is likely why industry classification schemes such as ISIC have manufacturing subcategories as deeply detailed as "manufacture of wood and cork, except for furniture" and yet group all of "wholesale and retail trade" into a single service subcategory.[50]

clean water.[35] As we gaze to the future, what technologies can humanity hope for next?

Physics first principles are a great guide to predicting the future. Not only can we use physics first principles to tell us what technologies are likely impossible, but we can also use them to construct a technology classification scheme that makes it easier to reason about future technology. And reasoning about future technology is how our nation chooses to fund scientific research, such as mine.

With this classification of technologies by constraint and function, we can now begin to look at the history of each category, and to more concretely conceive of what improvements might be possible.

1.1.5 Classifying information technology highlights opportunities for innovation

> *"Man is the lowest-cost, 150-pound, nonlinear, all-purpose computer system which can be mass-produced by unskilled labor."*
>
> *—Often misattributed to a 1965 NASA report, it apparently originated from test pilot Albert Scott Crossfield in 1954[53]*

From now on, I will focus simply on information technology, the most likely application of my research. I will examine past innovation in the three main functions of information technology: information storage, information transportation, and information transformation. From there, I will look at possible improvements, which will lead into Chapter 2 and my graduate research.

Storing information

Before computers, before the printing press, before written language itself, humans had few methods for storing information. For basic counting, notches could be carved into tally sticks,[54] but for remembering anything more complex, early humans could rely on nothing but their brains.[36] Still, by any standard, brains are impressive. There is evidence that some Aboriginal myths predate rising sea levels around 10,000 years ago, demonstrating the impressive longevity of collective memory based on oral tradition.[55, 56] And in the Pacific Northwest, a Klamath myth tells of the eruptions

[35] Actually, there is serious archaeological evidence that agricultural transitions *lowered* human quality of life (as measured by health markers such as skeleton size).[52] Whether this was a good or a bad thing depends on your utilitarian calculus, considering that agriculture also increased population density.

[36] Though in a way, our genes also store information. My natural fear of heights wasn't instilled into me by a teacher or some formative experience—it was hard-coded into me from my beginning, thanks to evolution.

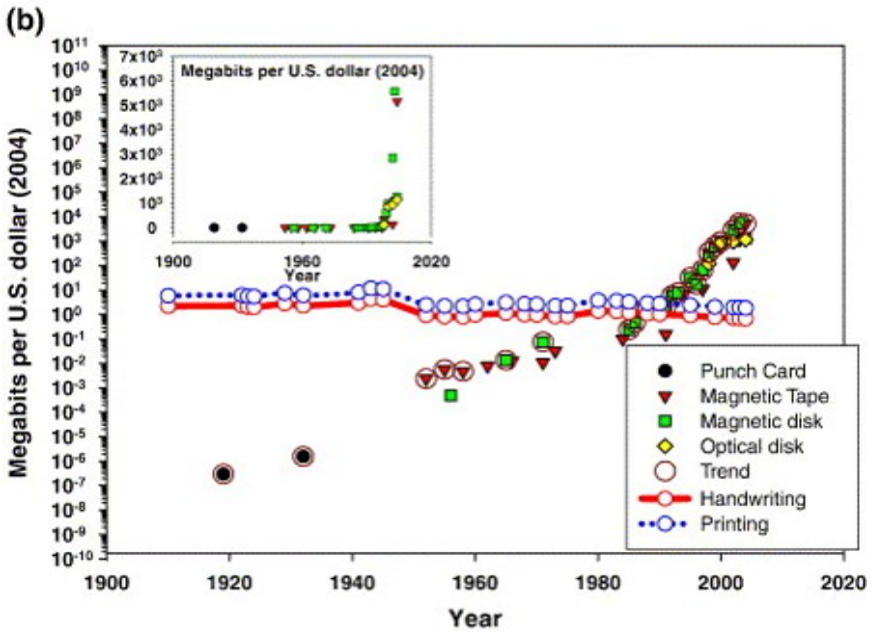


Figure 1.4: Progress in information storage over the last century. Surprisingly, perhaps, it took until 1990 or so until computer storage was cheaper than paper storage. Reproduced from Koh (2006)[42] with permission.[37]

of Mount Shasta and Mount Mazama, which last erupted around 8,000 years ago, according to volcanologists.[57]

Since then, humanity has come a long way. Figure 1.4 shows the last 100 years of progress in information storage technology, as measured by megabits stored per dollar. As the plot shows, computer memory was actually more expensive than printed paper until around 1990! However, today computer memory is orders of magnitude better, whether stored on optical discs, magnetic drives, or in solid-state flash memory.

What are the prospects for further improvement?[38] Progress in optical discs, magnetic disks,

[37]Reprinted from Technological Forecasting and Social Change, Volume 73, Heebyung Koh & Christopher L. Magee, A functional approach for studying technological progress: Application to information technology, 1061-1083 Copyright (2006), with permission from Elsevier.[42]

[38]In the ultimate limit, the covariant entropy bound[58] (related to the Bekenstein bound[59]) roughly limits the density of information to one bit per Planck square on its bounding surface. For a 1 cm sphere, this works out to be about 10^{66} bits, incomprehensibly more dense than today's technology, which can store about 10^{12} bits in the same volume. DNA has an information density of 10^{18} bits per cubic centimeter,[60] and if we were to store one per atom in a solid,[61] then we could reach around 10^{23} bits per cubic centimeter.

and solid-state drives appears to have slowed in the past few years, but it hasn't stopped.

Magnetic hard drives have been growing exponentially denser for sixty years. Relative to the very first commercial hard disk drive in 1956, a modern drive can store more than 100 million times more bits per area. Progress has been relatively steady, but the fastest improvement came in the 1990s and 2000s, when read heads sequentially exploited anisotropic magnetoresistance, giant magnetoresistance,[62] and then tunnel magnetoresistance,[63, 64] as well as perpendicular recording. However, since 2010, the improvement trend has stalled, and industry experts forecast a permanently reduced rate.[65] Today, disk drive companies are still trying to squeeze out more marginal gains in density with approaches like shingled recording, heat-assisted magnetic recording,[39] helium filling, and bit-patterned media. The general forecasting consensus is that the future rate of improvement will be perhaps 10% per year, far lower than the historical rates of 50%–100% per year enjoyed in the 1990s and early 2000s. One interesting fact regarding hard drive progress is that the reading speed of hard drives has not kept pace with the stunning advancements in bit density. Sixty years after the IBM's RAMAC spun its disks at 1200 RPM[66], commercial disks today commonly spin at only 5400 RPM or 7200 RPM (though 15,000 RPM has been demonstrated).[40] As a result, memory access algorithms originally developed for specialized use on magnetic tape are now finding greater use with magnetic disk drives.

Solid-state drives also seem to be reaching maturation. In *The Bleak Future of NAND Flash*

[39] A bit more on heat-assisted magnetic recording (HAMR): The main challenge with increasing bit density on hard drives is that as you make the magnetic bits smaller, they become less thermally stable. You can increase their stability by increasing their magnetic coercivity, but the tradeoff is that the bits become harder to write. Heat-assisted magnetic recording is a technology that makes bits easier to write, thereby making it possible to use smaller (higher coercivity) magnetic bits. This idea is not new, and in fact, enjoyed brief commercial realization in magneto-optic discs back in the 1980s. Today, Seagate is a leader in HAMR technology, but their release dates keep getting pushed back. Right now the latest word is that they'll be commercially available in ~~2017~~ 2018, but who knows whether that date will stick. I've heard the main problem with HAMR now is the thermal stability of the plasmonic antenna. Because the magnetic bits are much smaller than the wavelength of laser light, conventional optics cannot be used. Instead, the head has a vertical waveguide with a plasmonic antenna at the end, which concentrates the light at the bit being written. Unfortunately, a side-effect of heating the magnetic bits is that the antenna also gets hot, and it can fail after just 100 writes. Engineers are working to solve this and it seems likely they'll succeed, but right now this is the primary problem delaying HAMR's commercial deployment, at least from what I've heard.

[40] Although rotation speeds today are only 4.5–6 times faster than 1956, reading bandwidth is still much, much higher. As linear bit density increases, so too will reading bandwidth for a constant rotation speed. However, areal bit density scales quadratically with linear density, leading to a widening gap between disk capacity and reading bandwidth.

Memory, Laura Grupp et al. write: "The technology trends we have described put SSDs in an unusual position for a cutting-edge technology: SDDs will continue to improve by some metrics (notably density and cost per bit), but everything else about them is poised to get worse."[67] Ultimately flash memory suffers from a tradeoff between density and reliability, and as silicon transistors scale down, we will get better at trading off along that price-reliability frontier, but not at pushing that frontier outward.

Despite signs that magnetic hard drives and solid-state drives are slowing their rate of improvement, many speculative technologies are peeking over the horizon, including resistive RAM, magnetic RAM, phase change memory, memristive memory, ferroelectric memory, among others.

In one sense, it's not surprising that people have conceived of so many diverse memory technologies. For a technology to encode information, the only property it needs (at minimum) is two distinguishable states. Even a line of LEGO pieces turned either to the left or to the right could be used to encode information if you really wanted them to (and in fact, in 2012 such a machine was built in honor of Alan Turing's 100th birthday[68]).

However, in another sense, it is somewhat surprising (at least to me), that there are so many competitive memory technologies. Although almost anything can be used to encode information, in order to do it *well* a technology must simultaneously satisfy two goals that appear opposed: it must be easy to change when you want to (writing) and it must be hard to change when you don't want to (retention).[41]

Stone tablets, for example, do well at retention because stone is so durable, but do poorly at ease of writing because chiseling takes so much labor. In contrast, wet sand is easy to write on, but once written remains easy to disturb. Clay tablets were a great technological innovation because they took the best of both: while wet, clay is easy to imprint, like sand, but once dried or fired, it becomes far more durable, like stone. For the most part clay tablets were merely a read-only memory, but the tablets could be soaked in water to be recycled into new blank tablets.[42]

For modern microscopic memory technologies, reliability is limited by thermal stability (though other factors, such as cosmic radiation, matter too). If the energy barrier between memory states is too low, thermal fluctuations may randomly flip the stored bit. Although it can depend on specifics, a general rule of thumb used in magnetic recording is that a barrier of 25 k_BT will last one minute, a barrier of 40 k_BT will last eight years, and a barrier of 60 k_BT will last one billion years.[69, 70]

[41] This apparent tension also exists for knots. A good knot must not come undone on its own, but still must be easy to tie and untie. Giant tangles of rope are not considered good knots, because although they might never jiggle loose, they are not easy to tie or untie.

[42] In a strange way, this method of recycling clay tablets is analogous to modern flash memory, which can be read in a random-access fashion but must be erased in large blocks for rewriting.

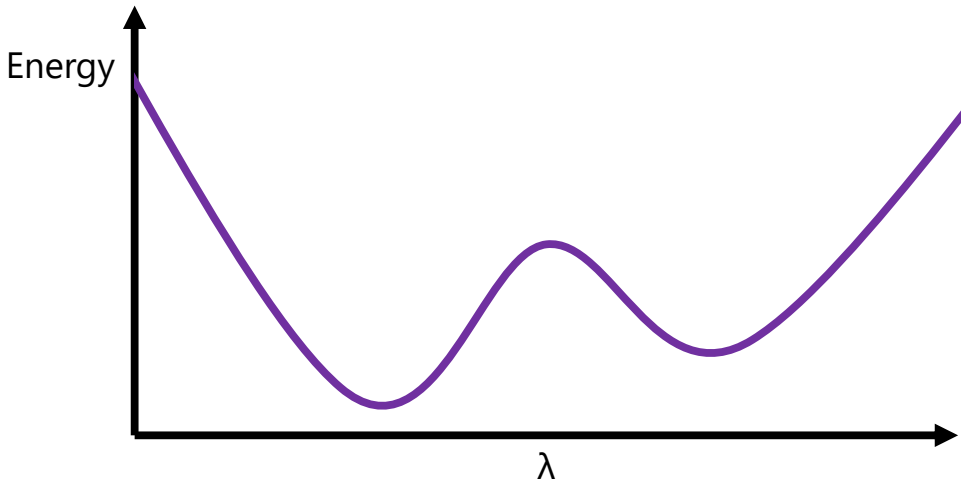


Figure 1.5: An arbitrary two-state memory technology can be modeled as a bistable potential well. Each energy well corresponds to a metastable memory state and the height of the barrier between the two states determines the memory's thermal stability in quasi-static equilibrium. λ is any parameter that characterizes the continuous transformation from one memory state to the other (in a magnetic bit, for example, it could be magnetization).

Transporting information

The trends and outlook for information transportation technologies are quite different than for information storage. Unlike information storage, which has a diverse cornucopia of competing technology options, information transportation has very few technology options worth considering. Because we'd like information to travel as fast as possible (i.e., the speed of light), right off the bat our options are limited to electromagnetic waves.[43] Of course, electromagnetic waves still come in a few flavors, such as free space radio transmission, free space optical transmission (requiring direct line of sight), fiber optic transmission, or wire cables.

Because electromagnetic waves are already capable of carrying signals at the universe's maximum speed, signal speed is not a very useful indicator of technological progress. Rather, bandwidth per dollar is a much better metric for evaluating progress in information transportation technology. Figure 1.6 shows progress in bandwidth per dollar, normalized by cable length, over the past 150 years.

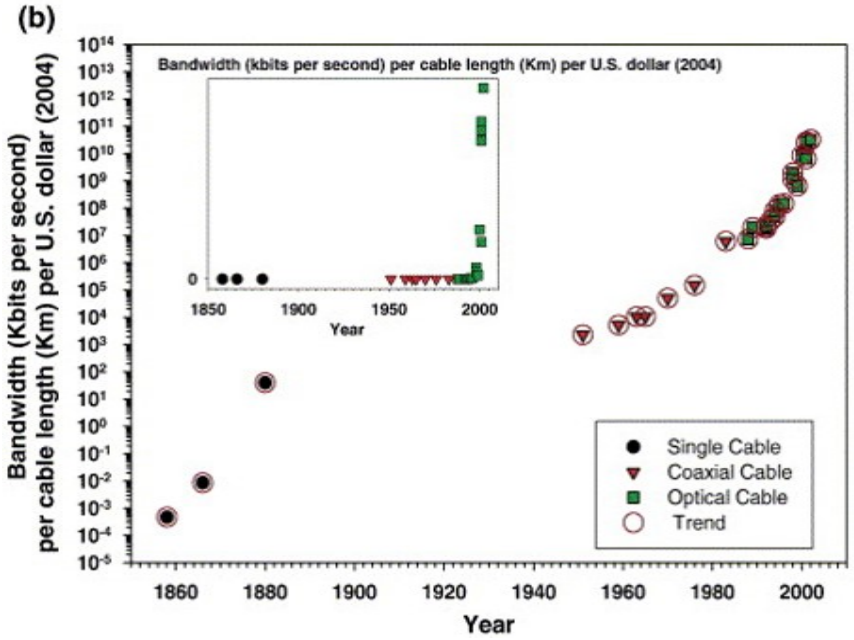


Figure 1.6: Progress in information transportation over the last century. Reproduced from Koh (2006)[42] with permission.[44]

[43]Of the three other forces, the weak and strong forces are limited by their short range, and gravitational waves require vast amounts of energy and mass for even our most sensitive detectors.[71]

[44]Reprinted from Technological Forecasting and Social Change, Volume 73, Heebyung Koh & Christopher L. Magee, A functional approach for studying technological progress: Application to

The outlook for information transportation technologies appears relatively simple. For signals that travel at the speed of light, fiberoptic cables will continue to dominate as the cheapest option between two points. Engineers will continue to research how to inject and distinguish frequencies on a finer and finer scale, allowing more parallel channels of communication down a single fiber. Further progress seems likely to be evolutionary rather than revolutionary.[45]

Transforming information

The third and final functional category of information technology is information transformation. Figure 1.7 shows the last century of technological process in this category.

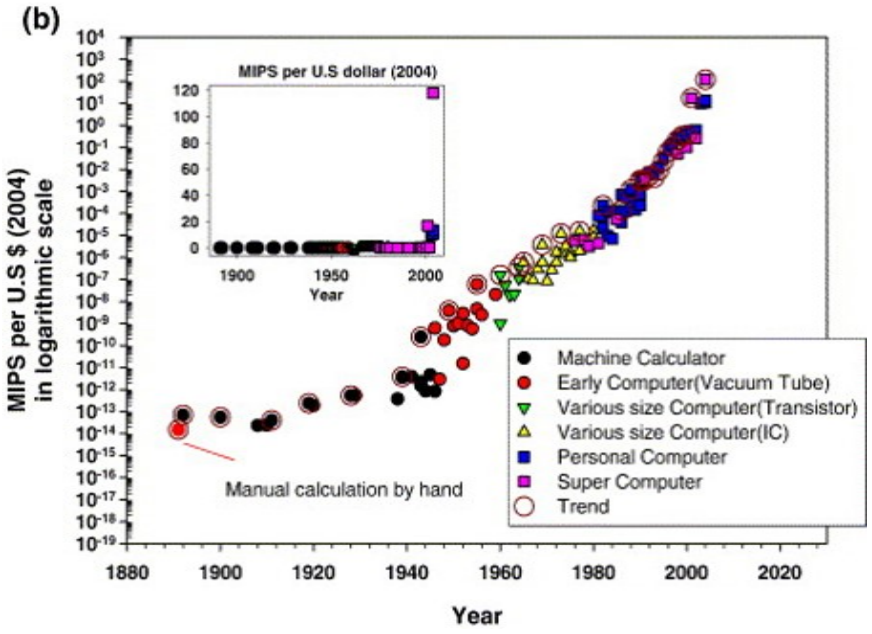


Figure 1.7: Progress in information transformation over the last century. Reproduced from Koh (2006)[42] with permission.[46] Note: while digging through the sources for this plot,[72] I discovered that the authors accidentally switched the red 'Manual calculation by hand' point with its nearest black neighbor. Despite their error, the main picture is unchanged.

---

information technology, 1061-1083 Copyright (2006), with permission from Elsevier.[42]

[45]Well, further progress *along this single performance dimension* is likely to be evolutionary. New wireless technologies will continue to be developed for their advantages along other performance dimensions like mobility or ease of installation.

[46]Reprinted from Technological Forecasting and Social Change, Volume 73, Heebyung Koh & Christopher L. Magee, A functional approach for studying technological progress: Application to information technology, 1061-1083 Copyright (2006), with permission from Elsevier.[42]

Many people are familiar with Moore's law,[73, 74] which famously prescribed the rapid exponential progress of integrated circuits. However, fewer are aware that rapid exponential progress in information technology predated the integrated circuit and Moore's 1965 prediction by many decades.

The technological enablers of transistor scaling

A common misconception about transistor scaling is that the driving force behind it was lithography changing to shorter and shorter wavelengths.[47] While it's true that shorter wavelengths have indeed enabled smaller feature sizes, shorter wavelengths have actually provided the least improvement of the three factors in the Raleigh equation.

$$R = k_1 \frac{\lambda}{NA} \tag{1.1}$$

	Resolution (R)	$\mathbf{k_1}$	**Wavelength** (λ)	**Numerical aperture** (NA)
1975[75]	2700 nm	1.00	436 nm	0.16
2016[76]	40 nm	0.28	193 nm	1.35
Improvement	68x	3.5x	2.3x	8.4x

Table 1.3: Four decades of improvement in lithography resolution. Contrary to common belief, improvement in resolution has not come primarily from shrinking wavelengths. In fact, the largest contributor in reaching smaller resolutions has been numerical aperture, whose gains were partly enabled by water immersion. (Immersing a lens in water will not by itself raise the numerical aperture of that lens; but rather, immersion allows lenses with higher numerical aperture to be usable.)

1.1.6 Moore's law is ending

> *"All good things must come to an end." —Chaucer, heavily paraphrased, 1374.[77]*

Moore's law has been so consistent for so long that some people take it as an article of faith that any engineering obstacles can and will be quickly overcome. Each time in the past that skeptics cried wolf, claiming that some limit would halt transistor scaling, the ingenuity of the semiconductor

[47] A second common misconception about lithography is that the resolution limits the feature size of devices, thereby limiting their density. In fact, this isn't quite true. What the resolution limits is the *spacing* of the devices (known as pitch in the semiconductor industry), which indeed limits their density. But the actual sizes of features can be made arbitrarily small with sensitive photoresists or reliable etch.

industry proved those cries wrong. When shrinking aluminum wires became too susceptible to electromigration, the industry switched to copper. When shrinking SiO_2 gate oxides became too electrically leaky, the industry switched to HfO_2. So it seems entirely reasonable, when confronted with warning signs today, to shrug off the worries and keep faith in the billions of dollars being spent by the semiconductor industry on R&D.

However, just as the boy who cried wolf eventually *did* encounter a wolf (and was disbelieved), eventually Moore's law will encounter a wolf of its own. And ironically, the better we are at learning that prior wolf cries were false, the more vulnerable we become to the wolf's true arrival. So it's important not to generalize too much from the cries of past skeptics. We must always remain vigilant to the possibility that this time is different. And in my informed opinion, this time *is* different.

Of all the steps needed to make a computer chip, lithography is the most critical.[48] Lithography is what, for now, limits the density of transistors, the major driver of their cost. Historically, as transistors scaled down, the cost of lithography at each node rose. But because the cost of lithography rose more slowly than the gains from greater density, overall the move was profitable. However, as it gets harder and harder to push lithography smaller, we may be reaching the point where continued scaling fails to make economic sense.

What are the prospects for shrinking the resolution of lithography? As shown in Equation 1.1 and Table 1.3, the resolution of lithography is driven by three factors: k_1, wavelength, and numerical aperture. Let's look at each in turn. Right now, k_1 is already at 0.28, very close to the theoretical maximum of 0.25 for two-beam imaging. So at best, there's only room for 12% more improvement in k_1. Numerical aperture also seems to be as good as it's going to get, with limited room for improvement. Although it is possible reach values slightly higher than 1.35 by using oil instead of water, it is very difficult to find oils that meet the stringent requirements of transparency, lack of bubbles, lack of impurities, high flow speed, and photoresist compatibility.[75] Given that k_1 and NA appear maxed out, the most promising route to better resolution seems to be shrinking the wavelength further. 157-nm $F_2{}^*$ lasers were attempted many years ago but were eventually abandoned after they proved too problematic for a variety of reasons (such as their incompatibility with water immersion). So instead, the industry has placed big hopes (and big bets) on an even more extreme option, 13.5 nm, which practically falls into X-ray territory, but, perhaps for marketing appeal, is known as extreme ultraviolet (EUV). Using a wavelength as small as 13.5 nm is a challenge because it requires many components of lithography to be simultaneously reinvented.

[48] One way to quantify this is to compare the market capitalization of ASML, the dominant seller of lithography tools, with Applied Materials, the dominant seller of most non-lithography tools. As of May 2016, ASML's market capitalization is $42 billion and Applied Materials's market capitalization is $22 billion. Even combining Applied Materials with Tokyo Electron (about $10 billion) and Lam Research (also about $10 billion) just manages to equal ASML. It's telling that the market value of lithography tools is comparable to all other tools combined.

Transmissive optics are not compatible with such short wavelengths, so now everything, including the mask itself, must be made from reflective multilayer mirrors. Old photoresists won't work with this new wavelength, so new photosensitive polymers must be invented with all the properties needed to be spun homogeneously and then chemically developed. Lastly, entirely new sources are needed to generate this short wavelength radiation. Although some have suggested building miniature synchrotrons, the leading technology for producing radiation involves shooting a CO_2 laser pulse at individual drops of molten tin as they fall in a vacuum. The drops explode into a plasma, releasing high energy radiation which is then filtered and collected into a beam for lithography.

Despite the billions of dollars poured into extreme ultraviolet lithography, the technology remains many years behind schedule. In 2003, Intel expected EUV would be ready by 2007.[78] In 2007, Intel thought EUV would be ready by 2011.[79] And by 2011, the target for EUV had slipped to the 2013–2015 range.[80] Although progress is still being made, many components of the system are behind where they need to be. The biggest problem in particular seems to be getting the power of the light source high enough for high throughput. Coupled with the general risk and cost of radical process changes, it's not clear whether EUV lithography will *ever* make it to production. And if it doesn't, then we may be at the end of the road for lithography. Further scaling may be possible by double, triple, or quadruple patterning, but these methods seem less likely to generate the cost savings that finer lithography has achieved in the past.

So if this truly is the end of the road for lithographic scaling and Moore's law, how could we tell? What might it look like? Well, if Moore's law were ending, I would expect to see more delays, more R&D layoffs, and more mergers. What do we actually see? Let's review: Intel's 10-nm node is severely delayed (until late 2017 at best). Their famous tick-tock development model has been abandoned. And recently in April 2016, Intel laid off 12,000 employees during a period of economic expansion. On the GPU side, NVIDIA and AMD have been stuck on the 28 nm node for five straight years.[49] EUV lithography remains years behind schedule and is far from production readiness. The ITRS roadmap is vaguer than it's ever been. Giant mergers are up (Intel+Altera, Applied Materials+Tokyo Electron,[50] KLA Tencor+Lam Research, etc.), concentrating the industry more than ever. And ultimately, a 5-year-old PC still works just fine. I submit that this is what the end of Moore's Law looks like.[51]

[49] Though, to be fair, they are at last on the cusp of getting 14/16 nm products to market.

[50] The proposed merger between Applied Materials and Tokyo Electron was blocked by the US Department of Justice over antitrust concerns, but I still count the attempt as a sign of a maturing, concentrating industry.

[51] Moore's law means many different things to different people. Although Moore defined the law to refer to the number of components in an integrated circuit,[73] many have broadened its meaning to refer to the consistent exponential progression of information technology.[81] And even Moore himself later adjusted his law's exponential rate.[74] When I say that Moore's law is dying or Moore's law is

If Moore's law really is ending, then what comes next? Will our computing technology plateau, still slowly improving, but at a much reduced rate? Or, as in past eras, will we find a new technology to replace our mature incumbents and thereby continue the long-term improvement trend in information technology?

No one knows.

1.1.7 What comes next?

What will future information technology be built of? At a fundamental level, there aren't many options. Apart from dark matter, everything we know of comes from the 17 particles of the Core Theory, shown in Figure 1.8.[52]

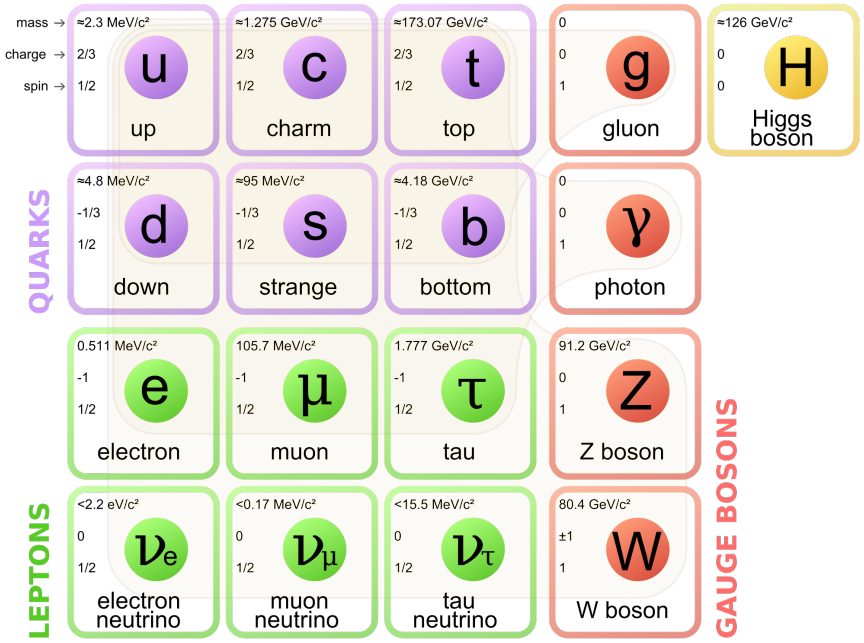


Figure 1.8: The known particles of the Standard Model (or, as Frank Wilczek and Sean Carroll prefer to call it, the Core Theory). Picture by MissMJ, distributed under a CC-BY-SA 3.0 license, via Wikimedia Commons.

From these 17 particles arise stable bound neutral clumps that we know as atoms. Triplets of

ending, I don't mean that all progress will stop for all time. What I mean is that the rate of progress in the near future will fall significantly and distinctly below the rates of the past few decades. No doubt, in the short term, progress will continue along many dimensions. In the long term, I expect nothing specific, only that I'll be surprised.

[52] Or, if you're a string theorist, strings.

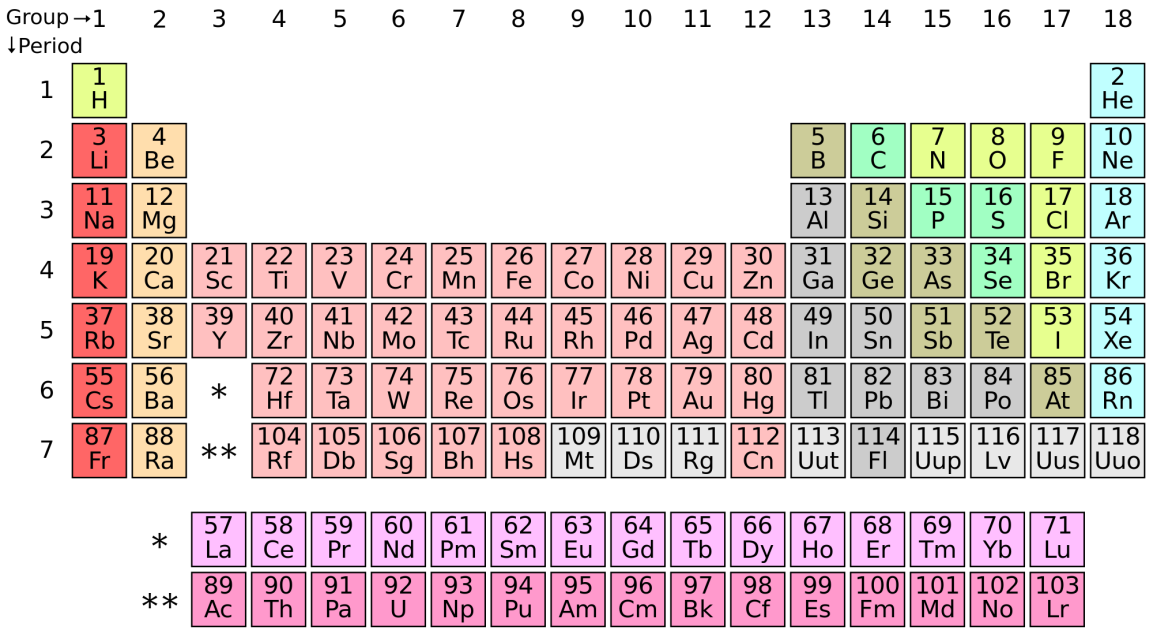


Figure 1.9: The periodic table. Almost all technology comes from combinations of these hundred or so building blocks. Picture by Sandbh, distributed under a CC-BY-SA 4.0 license, via Wikimedia Commons.

quarks are bound together to make protons and neutrons, which themselves are bound together into nuclei, around which negatively charged electrons orbit. Of the 118 elements ever observed, around 80 are radioactively stable. Almost all technology comes from different combinations and configurations of these hundred or so elements.

Further limiting the atomic building blocks available for technology are the rarity of some heavy elements, which make them expensive to extract. Figure 1.10 shows the relative abundances of elements in the Earth's crust. Abundance is an imperfect measure of cost, since cost also depends on the local concentration in ore deposits and the selectivity of various physical and chemical purification methods, but it nevertheless reflects the general trend. Rare earth elements in particular are more expensive than their abundance suggests, due in part to the fact that their highest energy f electrons barely participate in chemical bonding, causing them all to react very similarly (their chemical similarity makes them more difficult to purify).[53]

Ideally, any new information technology will be composed of abundant atomic matter. But apart from that physical and economic constraint, there is an exponentially large space of materials to search for new technological possibilities. Furthering this search (in a small and specific way) is the

[53]In fact, the chemical similarity of the rare earth elements is one reason why it took so long to identify them as separate elements.

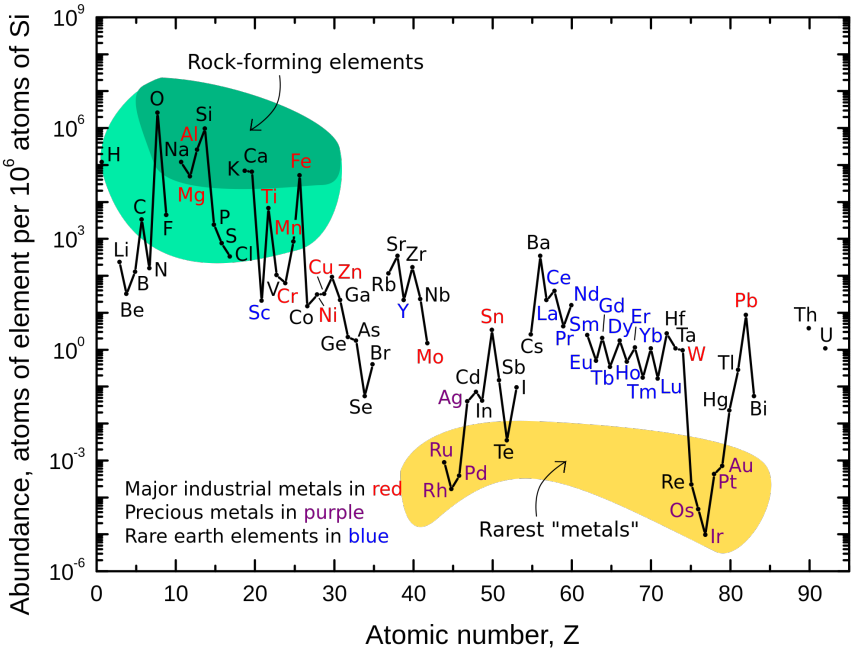


Figure 1.10: The abundances of elements in the Earth's crust vary tremendously. Oxygen, the crust's most abundant element, is about a trillion times more common than Iridium, the crust's rarest element. Rarer elements are less common in technology because they typically take more work to find and purify. Interestingly, the abundances of elements in the Earth's crust do not mirror those of the rest of the solar system, partly because heavier elements sank to the core of the Earth. With asteroid mining, these heavy elements may become cheaper someday.[82] Picture adapted by michbich from USGS Fact Sheet 087-02, distributed into the public domain, via Wikimedia Commons.

focus of this dissertation and the past six years of my life.

1.2 Searching for new materials and new physics

As we search for new materials to build new technology, it's worth thinking about the features that made past materials so successful. For instance, what was it about the semiconductor that made it so useful for information technology? The answer to this question, in my opinion, is the power of "sometimes."

1.2.1 The power of "sometimes"

If you owned a computer that only "sometimes" turned on, placed Amazon orders that only "sometimes" arrived, and had friends who only "sometimes" kept their promises, you'd live in a state of frequent frustration. Fortunately for most us, we don't live in such a world. All around us, products and processes have been designed to be reliable.[54]

However, when it comes to electronic and magnetic materials, reliability - at its literal extreme - is not necessarily a good thing.[55] Transistors that are always conducting won't make a good computer chip. Magnetic bits that always point up won't make a good hard drive. And photochromic lenses that always stay dark won't make good eyeglasses. Common to each of these technologies are

[54]In my opinion, the significance of reliability is too often underappreciated in our modern world. Let me explain why. Complex technology, nearly by definition, is composed of smaller, more fundamental parts that work together. So for a complex piece of technology to function, it typically needs all of its components functional at once. For instance, for a car to operate, it needs wheels AND tires AND an engine AND gasoline AND a driver AND so on, for thousands of critical components. If any one of those components fails, the entire car will fail. Because complex technology depends on components whose functionally is logically ANDed together, reliability becomes vastly more important as technology becomes more interdependent. A single component failing will not only halt its own output, but it will halt the output of every other component that indirectly relies on it. For this reason, the reliability of components acts as a limit on how complex and interdependent our technology can get. Without reliable components, the human race would be stuck with relatively primitive technologies. This is why reliability is so important. Without it, the combinatorial[83, 84] technological explosion of the industrial revolution(s) might never have been possible.

[55]In fact, the power of "sometimes" even extends to biology. Some researchers believe that thermal noise is not necessarily a hindrance to cells, but actually helps in some fundamental ways.[85] There is evidence for this in genetics[86] and neurons[87], and some researchers even hypothesize that cell size has evolved to be in the Goldilocks sweet spot of thermal noise, big enough to overcome the major problems of noise but still small enough to derive some benefit.

materials that "sometimes" work one way and "sometimes" work another way. This is the power of the "sometimes."

1.2.2 A brief history of semiconductors

No class of materials is more emblematic of the power of "sometimes" than the semiconductor,[56] which sometimes conducts electricity and sometimes does not. The controllable conductivity of semiconductors forms the foundation of modern electronics, from computer chips to photovoltaics to wireless communications to the laser.

However, the semiconductor was not always such a star material. It wasn't until 1910 that the term was coined,[91] and even decades afterward, many physicists questioned whether semiconductors even existed. In 1931, future Nobel laureate Wolfgang Pauli wrote to his colleague Rudolf Peierls: "On semiconductors one should not do any work, that's a mess, who knows whether there are semiconductors at all!"[92] His attitude was hardly uncommon at the time. Another physicist of that era complained "What are semiconductors good for? They are good for nothing. They are erratic and not reproducible."[93] Among scientists, semiconductors had gained a nasty reputation for being difficult to study because their electrical properties seemed to vary uncontrollably from

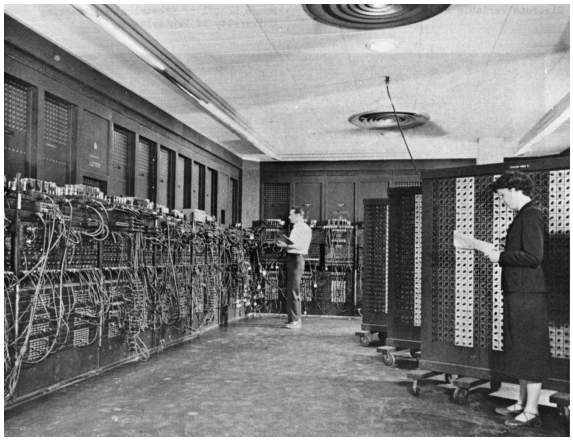
[56]Surprisingly, the definition of semiconductor is not wholly agreed upon. A fair number of professors and textbooks define a semiconductor to be a material with conductivity in between a conductor and an insulator.[88] Others define a semiconductor to be a material with a small band gap,[89] where small is perhaps 3 eV or less. However, these definitions are fuzzy and fail to drive home what makes a semiconductor fundamentally distinct from an insulator. The definition that I prefer and advocate is as follows: a semiconductor is a material that can be doped with charge carriers.[57] This definition emphasizes the true essence of semiconductors, which is that their conductivity can be controlled by chemical doping and electrostatic gating. And this is the fundamentally unique property of semiconductors that leads to the more superficial properties described by the other definitions. Because charge carriers can only persist when their energy is less than the energy of compensating electrostatic defects, the band gap typically needs to be small (and how small is not an arbitrary cutoff like 3 eV, but will vary depending on the material's particular defect chemistry). Furthermore, having a small band gap will typically cause a material's conductivity to lie in between that of conductors (which have no band gap) and that of insulators (which have a large band gap).

[57]For materials that can only be doped one way, with electrons but not holes, or holes but not electrons, I propose the following term: *semi*semiconductor. $SrTiO_3$ is an example of a semisemiconductor, because although it can be doped n-type by La^{3+} or Nb^{3+} or oxygen vacancies, attempts to dope non-transient holes into the system will be stymied by the formation of compensating oxygen vacancies. Because of this mechanism of compensating oxygen vacancy formation, a fair number of complex oxides are semisemiconductors.[90]

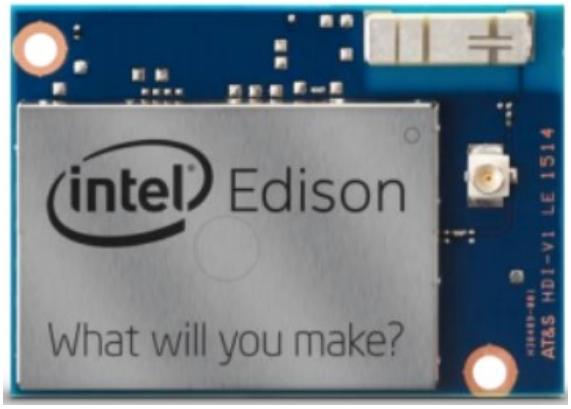
sample to sample, sometimes differing by as much as 7 orders of magnitude. Considering that reproducibility is a bedrock principle of the scientific method, it was extremely challenging to study materials that could not be reproduced reliably.

Progress in the field of semiconductors remained sluggish until World War II,[94, 95] when the military's well-funded desire for radar systems stoked demand for semiconductor diode rectifiers. Around the same time, scientists finally began to reliably control the doping level of semiconductors, focusing on materials like silicon and germanium, which were doped by chemical impurities, rather than self-doping materials like copper oxide, which were doped by chemical vacancies.

And the rest is history.



(a) ENIAC required a staff of six full-time engineers and broke down about once a day.



(b) Edison requires a staff of zero full-time engineers and breaks down about zero times a day.

Figure 1.11: ENIAC and Intel's Edison board, side by side. (Not even close to scale.)

	Year	Cost (2015$)	Power consumption	Tubes/transistors
ENIAC[96]	1946	$6,000,000	174,000 W	17,468
Intel's Edison[97]	2014	$50	0.5 W[98]	5,000,000,000[99]

Table 1.4: 70 years of progress stimulated by the semiconductor transistor.

1.2.3 What will be the semiconductor of the 21st century?

As Moore's law ends and semiconductor-based computing matures, we face two possible futures: either we settle for a reduced rate of technological progress or, alternatively, we invent new materials and paradigms that are capable of circumventing the physical limits of traditional semiconductor transistors. Although the second option of new technology and high performance sounds more

exciting, it's quite possible that it simply won't pan out.[58] Perhaps there really isn't anything better than the semiconductor transistor. And even if we do invent a radically superior transistor, it will still take a long, long time to catch up to the semiconductor industry in terms of supply chain management, factory tooling, worker education, technological compatibility, economies of scale, etc. So in the short run, over the next 20 years at least, I'd put my money on the continued dominance of the semiconductor transistor. But projecting far out, say, in 100 or 1,000 years, are we really going to be using the same general transistor technology as today?[59] Nobody knows. But framed this way, it certainly seems quite plausible that we could discover a superior alternative.

Therefore, if I had to frame the goals of my research field into a single question, it would be this:

What will be the semiconductor of the 21st century?

That is, what will be the next paradigm shift,[100] the next disruptive[101] materials platform that leaps ahead of the semiconductor transistor? And in answering that question, in seeking new materials for future computers, what generalizable lessons can we learn from the success of semiconductors to guide our search? And, perhaps just as important, but less commonly asked, what lessons *shouldn't* we learn from semiconductors? I.e., which idiosyncratic features of the semiconductor might we profitably discard in constructing a new and improved technological paradigm?

Here, I wish to highlight two important lessons from the semiconductor, and then illustrate how they apply to the family of materials known as complex oxides. Lesson one: Controllable variation is valuable. The properties that made the semiconductor so successful were its variability and our eventual control of that variability by chemical doping. Therefore, in a research program looking for future computing materials, it makes sense to seek materials with high variability (perhaps along different properties than conductance) and then learn to control them. Complex oxides, as a family, possess a tremendously wide range of properties, properties that can be controlled not only by chemical doping, but also by structural modification. Lesson two: Poor understanding of a material is actually a positive signal of its usefulness, because the materials with the most variability are naturally the last to be understood. As I discussed earlier, semiconductors were known for many

[58] In all of history there has been nothing quite like the blistering technological progress of integrated circuits over the past 60 years. The astounding success of the semiconductor transistor has instilled confidence in many people, who extrapolate out and expect progress to continue. However, viewed from another perspective, the astounding success of the semiconductor transistor makes it that much harder to be surpassed by a new challenger. As we target technological R&D, how much sense does it make to aim at supplanting one of the most successful technologies of all time? Given the transistor's dominance, perhaps we'd better off investing less in alternatives and more in its continued development.

[59] Some technologies, like the wheel, have been dominant for thousands of years. There is a nonzero probability that the same will be true of the semiconductor transistor.

decades before finally being put to use in diodes and transistors. The reason this took so long is twofold. First, the variability of semiconductors made them naturally harder to study. And second, because they were harder to study, very few scientists even attempted to study them, which further slowed the pace of scientific understanding. As a result, this materials family with so much technological potential was sitting right under our noses for decades. Similarly, complex oxides have been known for decades, yet remain relatively difficult to understand. It seems quite possible that they will follow a similar arc as the semiconductor, where at first their complexity stymies our understanding until eventually that complexity is understood and harnessed for new purposes unachievable by the materials that came before.

1.3 Harnessing the complexity of complex oxides

A complex oxide is a chemical compound that contains oxygen and at least two other elements (or oxygen and just one other element that's in at least two oxidation states).[102, 103] Complex oxide materials are notable for their wide range of magnetic and electronic properties, such as ferromagnetism, ferroelectricity, colossal magnetoresistance, and high-temperature superconductivity. Although the space of complex oxides is far too combinatorially large to survey in a single paragraph, or even in an entire book, I will highlight a few classic examples that illustrate the diversity of correlated-electron phenomena. Most of the world's commercial hard ferromagnets, such as those in electric generators and motors, are ferrites, such as $SrFe_{12}O_{19}$.[104] Commercial ferroelectrics, used commonly in actuators and positioning systems, are commonly titanates,[105] such as $PbZi_xTi_{1-x}O_3$[106] and $BaTiO_3$. Less commercially valuable but still interesting is the phenomenon of colossal magnetoresistance. Colossal magnetoresistive materials include $LaMnO_3$ and other manganites.[107] And classic examples of superconductivity include $YBa_2Cu_3O_{7-x}$[108] and other cuprates. Common to all of these materials—ferrites, titanates, manganates, and cuprates—are the facts that they possess (1) transition metal ions, (2) oxygen, and (3) complex chemical formulas. This is no coincidence. Correlated-electron phenomena often arise from semi-localized electrons in d orbitals, whose potential can be unlocked by ionic oxygen bonding and tuned by alloying and doping. The rest of this section will briefly explain why these unique circumstances make complex oxides so interesting and potentially useful in information technology.

1.3.1 The paradigm of orbital localization

In general, if you want a material to exhibit any sort of useful dynamic behavior, it must, by definition, be in different states at different times. In semiconductors, the different states that we exploit are different levels of conductivity, which are controlled by the concentrations of electrons and holes. In complex oxides, though, I wish to focus on a dichotomy that is superficially similar in ways but is fundamentally quite different: the dichotomy of localized and delocalized electron

orbitals.

All of the complexity of atomic matter and chemistry and life itself come from valence electrons, which can inhabit a very limited set of orbitals.[60] The principal quantum number, n, ranges from 1 to 7, and the azimuthal quantum number, l, ranges from 0 to 3 (often labeled by the letters s, p, d, and f for historical reasons).[61] It's amazing that such complexity arises from such simple components.

A solid's properties are determined by the electrons that bond it together. In general, a solid with electrons in delocalized orbitals (such as those in the s, p, or $5d$ orbitals) will tend to be metallic and well-described by band theory. These correspond to the blue region on Figure 1.12. On the other end of the spectrum, a solid with localized electrons (such as those in the $4f$ orbitals) will tend to be insulating with local-magnetic moments and well-described by molecular orbital theory.[62] These correspond to the red region in Figure 1.12. However, for electrons in orbitals with middling localization (such as $5f$, $3d$, and $4d$), it's much harder to predict how they will behave. In this regime, electrons cannot easily zoom by one another, so they often end up exhibiting complex, correlated behavior. This results in uncommon ordered properties such as ferromagnetism, ferroelectricity, high-temperature superconductivity, charge ordering, and more.[63] Although this correlated behavior is much harder to model, it may have potential for future technology.

1.3.2 The importance of oxygen in complex oxides

By themselves, transition metals and rare-earth metals do not necessarily exhibit correlated electron phenomena from their semi-localized d or f orbitals. This is because these metals also have extended s orbitals (or for lanthanides, $6s$ and $5d$ orbitals) that participate in bonding and conduction, masking the effects of the d or f orbitals underneath.

[60] Again, I am relying on the simple but useful fiction of single-electron orbitals. These do not exist in real life, but because they make talking about chemistry so much easier, I will continue to rely on them. In your mind, feel free to substitute the word 'orbital' with the wordier 'counterfactual low-energy multi-electron excitation.'

[61] Of course, in theory, these numbers can go higher. But such atoms are extremely unstable.

[62] Although this range of properties, from metallic to insulating, sounds superficially like semiconductors, the causes are quite different. In semiconductors, the variation in conductance is caused by the variation in the concentration of electrons or holes. However, in complex oxides, the variation in conductivity can be caused by the variation in orbital localization. As a result, there can arise many correlated electron phenomena not seen in semiconductors.

[63] Although electron correlation can lead to emergent properties of materials, it is by no means required.[109] The fractional quantum Hall effect in graphene comes from a system filled with relatively uncorrelated sp^2 electrons.

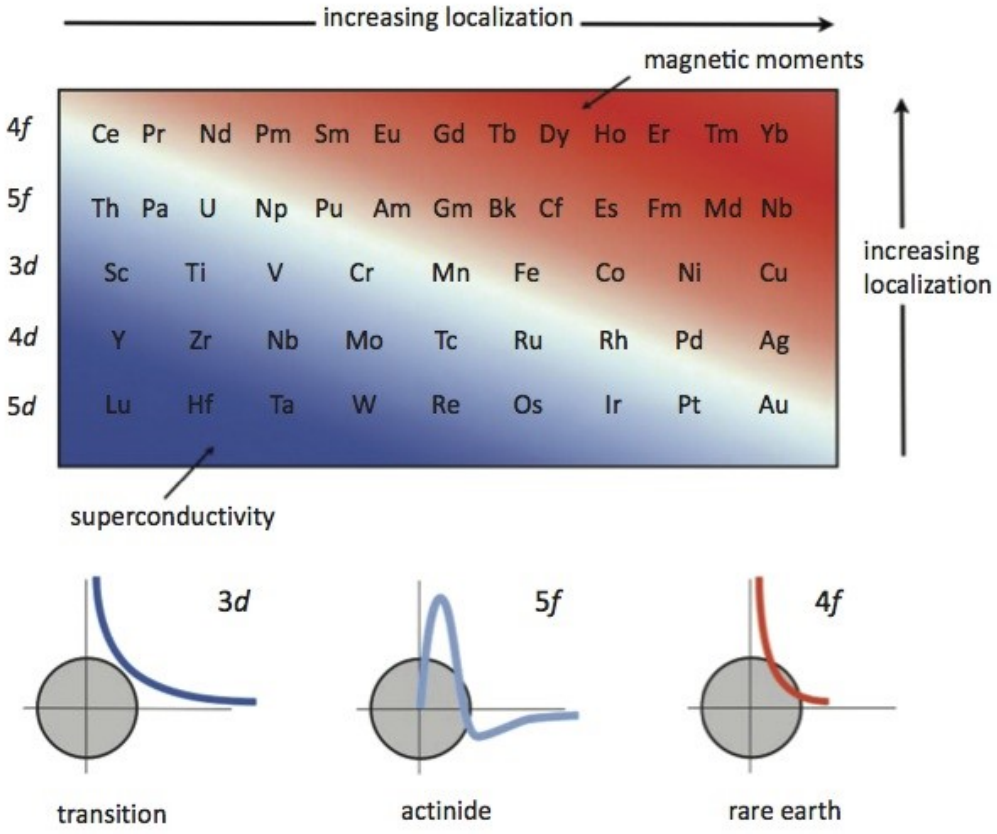


Figure 1.12: A Kmetko-Smith diagram,[110] copied from P. Coleman's *Introduction to Many-Body Physics*.[111]

The best way to pull off these extended electrons is to bond the transition or rare-earth metal with oxygen. Oxygen is the perfect element for this task: it has the highest electronegativity behind fluorine, but unlike fluorine it is safe and abundant. It also is better than fluorine at bonding with multiple atoms in a solid.

Without oxygen, the complex oxides would not exhibit complex, correlated electron behavior. Oxygen is necessary to pull off the valence s electrons, leaving the key d or f electrons behind as the new valence shell.

1.3.3 Controlling bonding in crystals of complex oxides

Having now emphasized how the chemical bonding of electrons drives materials properties, it's worth considering what options we have for designing new materials by controlling their chemical bonding and structure.

Top-down design versus bottom-up

Because atoms are so small, placing them one-by-one into an imagined crystal structure would take an inordinate amount of time to get any appreciable quantity of material. Therefore, nearly all techniques for synthesizing materials rely on bottom-up approaches, where nanoscale physical interactions help guide the atoms into their desired places.

Chemical composition

One of the most powerful tools for creating materials is thermodynamic equilibrium. A collection of atoms can be heated into a relatively homogeneous liquid and then slowly cooled, resulting in an equilibrium crystal that is controlled by the initial collection of atoms. This technique was used by early humans to produce bronze and other materials, and today it is still very common, being used to produce alloys and ingots of many different crystals. In my research, crystal substrates were synthesized using the Czochralski process,[64] a specific technique that broadly relies on equilibrium thermodynamics and chemical composition to create large, homogeneous crystals.

In general, thermodynamic techniques require lots of heat and time, which makes them relatively simple and potentially expensive.

Non-equilibrium processing

In the vast space of possible materials, those that are thermodynamically stable inhabit just a tiny fraction of it. Therefore, it is important to find non-equilibrium processing techniques that allow us to create useful meta-stable materials. Quenching is a classic technique that exploits slow kinetics

[64] As with many scientific and technological advances, the process was discovered by accident, when Czochralski dipped his pen in molten tin instead of his inkwell.[112]

as a strategy for avoiding a material's equilibrium state.[113] Quickly quenching carbon-rich steel through its eutectic point makes it much stronger by limiting the rate at which carbon precipitates out of the material.

In contrast to equilibrium techniques, non-equilibrium techniques must avoid heat or time.

One relevant technique for reaching non-equilibrium, meta-stable materials is thin-film deposition. The next few subsections discuss ways in which thin-film deposition can be exploited to reach materials states that are unreachable by conventional equilibrium techniques.

Epitaxial strain

In terms of materials properties, the most straightforward advantage of thin-film deposition is epitaxial strain. Epitaxial strain is a type of strain that occurs in thin films, caused by the bonding force of the substrate's atoms acting on the atoms of the the thin film. If the substrate has an atomic structure with wider atomic spacing than the film, the substrate's atoms will pull the thin film's atoms further apart in the two in-plane directions. Conversely, if the substrate has an atomic structure with closer atomic spacing that the substrate, the thin film's atoms will be pulled closer together in the two in-plane directions. Typically there is an elastic response in the out-of-plane direction that helps accommodate the modified spacing. These structural distortions can be quantitatively described by ϵ and σ, the second-order strain and stress tensors, respectively.

$$\epsilon = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{bmatrix} \tag{1.2}$$

$$\sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \tag{1.3}$$

The strain tensor can be additively decomposed into three parts: the volume dilation, the orthorhombic distortion, and the shear strain.

$$\begin{aligned} \epsilon = \epsilon_{volume} + \epsilon_{orthorhombic} + \epsilon_{shear} = & \begin{bmatrix} \epsilon_{11}+\epsilon_{22}+\epsilon_{33} & 0 & 0 \\ 0 & \epsilon_{11}+\epsilon_{22}+\epsilon_{33} & 0 \\ 0 & 0 & \epsilon_{11}+\epsilon_{22}+\epsilon_{33} \end{bmatrix} \\ & + \frac{1}{3}\begin{bmatrix} 2\epsilon_{11}-\epsilon_{22}-\epsilon_{33} & 0 & 0 \\ 0 & 2\epsilon_{22}-\epsilon_{11}-\epsilon_{33} & 0 \\ 0 & 0 & 2\epsilon_{33}-\epsilon_{11}-\epsilon_{22} \end{bmatrix} + \frac{1}{3}\begin{bmatrix} 0 & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & 0 & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & 0 \end{bmatrix} \end{aligned} \tag{1.4}$$

For [001] epitaxy, the type I used in my research, the applied strain is orthorhombic (with in-plane spacing differing from out-of-plane spacing) and, depending on the elastic response, volumetric as well. For epitaxy on other crystal planes, the decomposition will be different.

One of the most successful examples of strain engineering is strained silicon, incorporated into modern transistors in the mid-2000s. By depositing silicon thin films next to silicon-germanium, the spacing between the silicon atoms is increased, allowing for higher electron mobility and consequently higher electronic performance. In fact, over the past decade or so, the use of strained silicon has been a bigger driver of performance than traditional scaling.

Epitaxial structure-transfer

For crystals with a complex structure, the effects of epitaxial strain can be similarly complex. When atoms get pushed toward one another, complex crystal structures can accommodate this pushing in a number of ways. In many perovskite-based crystal structures, for example, biaxial compression is accommodated by octahedral rotations, where bond angles are what primarily change, rather than bond lengths. Bond angles are especially important in superexchange magnetism, and the paradigm of strain-induced octahedral rotations has become increasingly recognized as a tool for understanding and manipulating the properties of materials.[114]

1.3.4 The astounding success of band theory

Mass education is great. On an individual scale, mass education has provided the poor a path out of poverty, allowing the children of farmers and laborers to leverage their intelligence in more productive occupations. On a societal scale, investment in mass education has likely contributed to the rise of modern democracy, productivity, and peace.[65]

However, mass education has downsides. Efficiently standardizing myriad human accomplishments into book chapters and homework assignments often has the regrettable side effect of deadening a proper sense of surprise and awe. In my opinion, band theory is one of these victims.

Each year, thousands of undergraduate students are dryly taught band theory, which, after some simplifying assumptions, allows one to calculate how electrons act in solids. As the students advance, they may be taught how these assumptions break down in some materials, where band theory fails to apply. However, there is an aspect to band theory that few students are ever taught: it's a surprise that it works at all!

[65] Speaking of schooling and peace, some well-meaning education reformers (and others) argue that the roots of modern compulsory education lie in the Prussian 'factory' model of schools, invented at the dawn of the industrial age to make more obedient soldiers and workers out of the population. But scholars of education history dispute this narrative, calling it sloppy and crudely reductionist.[115]

To accurately calculate the properties of a material, one must solve the time-dependent Schroedinger equation:

$$i\hbar\frac{\partial}{\partial t}\psi\left(\mathbf{r},t\right)=H\psi\left(\mathbf{r},t\right) \tag{1.5}$$

where, for a solid with some assumptions,[66] the Hamiltonian H is

$$H=\sum_{i}\frac{\vec{p_i}^{\,2}}{2m_e}+\frac{1}{2}\sum_{i,i'}\frac{e^2}{4\pi\epsilon_0\left|\vec{r_i}-\vec{r_{i'}}\right|}+\sum_{j}\frac{\vec{p_j}^{\,2}}{2m_j}+\frac{1}{2}\sum_{j,j'}\frac{Z_jZ_{j'}}{4\pi\epsilon_0\left|\vec{r_j}-\vec{r_{j'}}\right|}+\frac{1}{2}\sum_{i,j}\frac{eZ_j}{4\pi\epsilon_0\left|\vec{r_i}-\vec{r_j}\right|} \tag{1.6}$$

where the terms are, in order, the kinetic energy of electrons, the potential energy from electron-electron repulsion, the kinetic energy of the nuclei, the potential energy from nucleus-nucleus repulsion, and the potential energy from electron-nucleus attraction. i is the index for electrons, $\vec{p}$ is the momentum operator, m_e is the mass of the electron, e is the elementary charge, ϵ_0 is the dielectric constant, $\vec{r}$ is the position operator, j is the index for nuclei, and Z_j is the charge of the nucleus.

The brevity of this equation belies its truly monstrous size. Even if you were to ignore all of the nuclear terms and all of the core electrons, this equation is still impossibly difficult to solve, requiring the diagonalization of Hamiltonian with roughly 10^{46} terms. And even if such a computation were magically possible, the mere task of inputting initial conditions or reading the output would be an impossibly immense task.

Therefore, many different approximations to this equation have been developed.[116, 117, 118] One of the simplest and most fundamental is band theory. Band theory's key assumption, which makes the model solvable (at the cost of being wrong in ways), is the assumption that electrons are non-interacting. With this assumption, the solid's wavefunction can be factored into single-electron states (and then rebuilt into a multi-electron state by applying the Pauli exclusion principle through Fermi-Dirac statistics).

From here, one approach is to ignore the nuclei as well, and lump the single electron's interactions into an effective periodic potential V that depends only on space and no other particles.[67] This makes the equation much simpler:

$$H=\frac{\vec{p}^{\,2}}{2m_e}+V(\vec{r}) \tag{1.7}$$

Alternatively, after making the independent electron assumption, one can perturb the Hamiltonian by introducing a tight-binding term, H_{TB}, where the atomic orbitals are the starting basis then

[66] Already this Hamiltonian ignores relativity, ignores all forces except electromagnetism, ignores the energy of the electromagnetic field, and assumes nuclei are point particles.

[67] In this approach, it is important to choose a good effective potential, and many methods have been invented to generate such potentials.

mixed by transfer integral terms proportional to t_{ij}. If only nearest-neighbor terms are considered, this approach predicts band structures with cosine dispersion.

$$H_{TB} = -\sum_{i,j} t_{ij}(c_i^\dagger c_j + c_j^\dagger c_i) \tag{1.8}$$

Overall, band theory has been wildly successful at explaining the electronic properties of many materials. But when you stop to consider how unreasonable and unphysical it is for band theory to assume that electrons are non-interacting, it's astonishing that the theory works at all. Although I regard this as a somewhat open problem, I believe that it can be answered by a deep understanding of statistical mechanics, where it turns out that the lowest-energy excitations of multi-particle wavefunctions end up being remarkably similar to the lowest-energy excitations of single-particle wavefunctions, even though the states are totally different.[68]

1.3.5 The not-so-astounding failures of band theory

Given band theory's strong assumption of electron independence, it should come as little surprise that its predictions are sometimes violated. The most classic example are Mott insulators, materials that band theory predicts to be metals but turn out to be insulators. This class of materials was highlighted by Mott in 1949 when he hypothesized that electron repulsion prevented the flow of charge in NiO.[119]

1.3.6 Progressing beyond the independent electron assumption

One of the most successful approaches for moving beyond the independent electron assumption is the Hubbard model, which extends the tight-binding model of band structure with a term that models the on-site repulsion of electrons, the strength of which is scaled by the parameter U.

$$H_H = -\sum_{i,j} t_{ij}(c_i^\dagger c_j + c_j^\dagger c_i) + \sum_i U_i n_{i\uparrow} n_{i\downarrow} \tag{1.9}$$

However, the Hubbard model, with its 'effective' repulsion scaled by U, is relatively crude. In *Condensed Matter Physics*, Michael Marder writes of the Hubbard model: "there is no better illustration of the difficulties involved in progressing systematically beyond the one-electron pictures of solids."[118, 120]

There is a vast body of literature detailing the theoretical and experimental attempts to construct models that are more accurate at the cost of more complexity. I will not attempt to describe even the sliver of that vastness within my ken, but merely wish to emphasize that cutting-edge goes

[68] I cannot personally prove this statement myself, but it comes from extended discussion a few years back with UC Berkeley Physics Professor Dung-Hai Lee.

far beyond what I've briefly shared here. However, simple models such as the Hubbard model are nonetheless important because their framing of the problem is the foundation for many of these more complicated and specialized techniques.[69]

1.3.7 The role of experiments

It is precisely because of the difficulty of theoretical prediction that scientists perform experiments. Oftentimes, the easiest way to discover a material's properties is not to calculate them, but to measure them. And once many such properties have been measured, emergent patterns can be identified, even if they are fully untethered from first principles physics equations. This is the role of experiments, performed by experimentalists like myself.

The hope is that by synthesizing and measuring new and interesting materials, we can identify technologically useful phenomena or patterns. These measurements can also constrain the space of physical models, raising the likelihood that new and better models will be developed. These were the scientific goals of my two major PhD research projects, described in detail in Chapters 3 and 4.

1.4 Magnets

Much of my PhD research involved magnetism, one of the most classic examples of correlated electron behavior. For reasons that I will touch on later, there is hope that information technology based on magnets can surpass the the semiconductor transistor in energy efficiency. This section of my dissertation provides relevant background information about magnetism and magnets, but for an in-depth explanation of these topics, I suggest reading a textbook, such as Spaldin's *Magnetic Materials: Fundamentals and Device Applications*[121] or Stoehr's *Magnetism: From Fundamentals to Nanoscale Dynamics*[122] or Coey's *Magnetism and Magnetic Materials*[123].[70]

1.4.1 Magnets are prominent in physical and information technology

The first known use of magnets in technology took place in China around 300 BCE. Spoon-shaped pieces of lodestone (a rock rich in Fe_3O_4) were used like compasses to point south. Although at first

[69] Another reason that simple models are important is that it's important to know which features of a model are necessary for modeling some complex behavior. If a simple model can model some system as well as a complex model, then it means the complexity of the complex model is not explaining very much.

[70] Spaldin's book is accessible, Stoehr's book is comprehensive, and Coey's book has decent coverage of applications and their markets. However, be aware that Coey's book is in the first edition and contains many errors, errors that he seems uninterested in correcting on his errata page, which has not been updated in years despite my repeated emails.

these spoons' only uses were geomancy, feng shui, and fortune telling, eventually improved versions found use in navigation.

Since then, magnets have spread. They are in computer hard drives, in generators, in motors, in wireless communication electronics, in voltage transformers, and more. Magnet technologies can be classified into three major categories: permanent (hard) magnets, soft magnets, and magnetic recording. Each category is roughly a third of the global market for magnets, which in 2010 totaled $30 billion.[123]

Hard magnets, defined by their high magnetic remanence and high coercivities, are useful for converting between electrical energy and mechanical energy.[71] They are used in generators, electric motors, actuators, holding devices, and sensors. Production is dominated by ferrites, though high-performance applications use rare-earth magnets.[72]

Soft magnets, defined by their low coercivities and low magnetic remanence, are useful for efficiently amplifying magnetic fields. They are used in the cores of transformers, in radiofrequency and microwave circuits, and in other electronics. Transformers typically use Fe-Si alloys, whereas high-frequency applications, which require high resistivity for low losses, use soft ferrites.

Magnets used in magnetic recording tend to have middling coercivities. Hard drive disks are perhaps the biggest market today, but this technology family includes magnetic tape (both for archival data storage as well as audio and video cassette) and magnetic stripes. If anything, this market seems less healthy now than in the past. Although data centers demand more and more information storage each year, much of this growth has been absorbed by solid-state drives, a high-performance alternative with steadily dropping costs. Audio and video cassettes are already obsolete, and it's likely that fate will similarly befall magnetic card strips, which face competition from EMV Chip and PIN technology as well as smartphones.

Although the commercial applications of my research are vague, the high cost of its materials and fabrication techniques make high-value information technology its likeliest application. This could be information storage, information processing, or sensors.

1.4.2 What is a magnet?

Like some other branches of science, magnetism lacks a full vocabulary to label its phenomena, and as a result, sometimes a single word can be overloaded with multiple meanings. For example, what does it mean for a material to be magnetic? Does it mean the material exhibits a response in an applied magnetic field? If so, then all materials are magnetic, be they diamagnets, paramagnets,

[71] Or, put another way, between magnetic fields and torque.

[72] Although you might reasonably expect rare-earth magnets to primarily consist of rare-earth elements, in fact they are mostly iron. The common rare-earth magnet neodymium iron boride ($Nd_2Fe_{14}B$) is only 12at% rare earths.

ferromagnets, or otherwise. Does it mean that the material's electron magnetic moments exhibit long-range order? Then diamagnets and paramagnets would be left out, but antiferromagnets, which also possess very little net magnetic moment, would remain included. Does it mean that the material is presently exhibiting long-range ordering? If so, then ferromagnets above their T_c would be excluded. Ultimately, the question of what is a magnetic material has no correct answer. It depends entirely on what the asker means. For this chapter, I will use the word magnet to mean ferromagnet.[73]

Magnetic materials can be categorized in many ways. They can be itinerant, where the valence electrons are mobile, or they can be local, where the valence electrons cannot easily move from atom to atom. Local magnets are easier to describe than itinerant magnets. All of their magnetic behavior can be expressed in terms of how their microscopic magnetic moments order. If they form domains and tend to align in a common direction in the absence of a magnetic field, then the material is a ferromagnet. Else, if they point in random directions in the absence of an applied magnetic field, then the material is a paramagnet.[74] If the electron spins point in no direction at all (i.e., there are no unpaired electrons), then the material is a diamagnet.[75] These three basic forms of magnetism

[73]The word 'ferromagnet' is itself overloaded. Narrowly defined, a ferromagnetic material is one whose microscopic magnetic moments all align in the same direction. But broadly defined, a ferromagnetic material is any material that exhibits a spontaneous net magnetic moment. This broader definition encompasses orderings that are not strictly ferromagnetic, such as ferrimagnetic ordering, in which unequal microscopic magnetic moments alternate their alignment, resulting in a net magnetic moment.

[74]The dividing line between ferromagnets and paramagnets is actually rather subtle (and to a newcomer, it's not even obvious that there ought to be a sharp dividing line rather than a gradual transition). In a ferromagnet, because of thermal fluctuations, the local moments will never be fully aligned, and at temperatures just below T_c the microscopic ordering can look quite patchy. In a paramagnet, on the other hand, the local moments will never be fully randomly aligned; short-range interactions may align small patches, just as in a ferromagnet. In fact, by eye, it is impossible to tell the difference between a patchy ferromagnet and a patchy paramagnet. The property that divides them is whether their ordering is long-range or just short-range. E.g., if you start at one local moment and then move a long distance away, how accurately can you predict the alignment of your new location? If you can only predict it with 50% accuracy, the material is a paramagnet. If you can predict it with more than 50% accuracy, the material is a ferromagnet. The question of where to draw the dividing line between ferromagnets and paramagnets is handled with more sophistication in the study of phase transitions and critical exponents in field of statistical mechanics.

[75]Really, any material with a magnetic susceptibility that is (mostly) well-defined, negative, and independent of temperature is a diamagnet. There are some conductors and semiconductors with unpaired electrons that are nevertheless diamagnetic because their conduction electrons have low

are plotted in Figure 1.13, though of course many more complicated patterns exist as well.

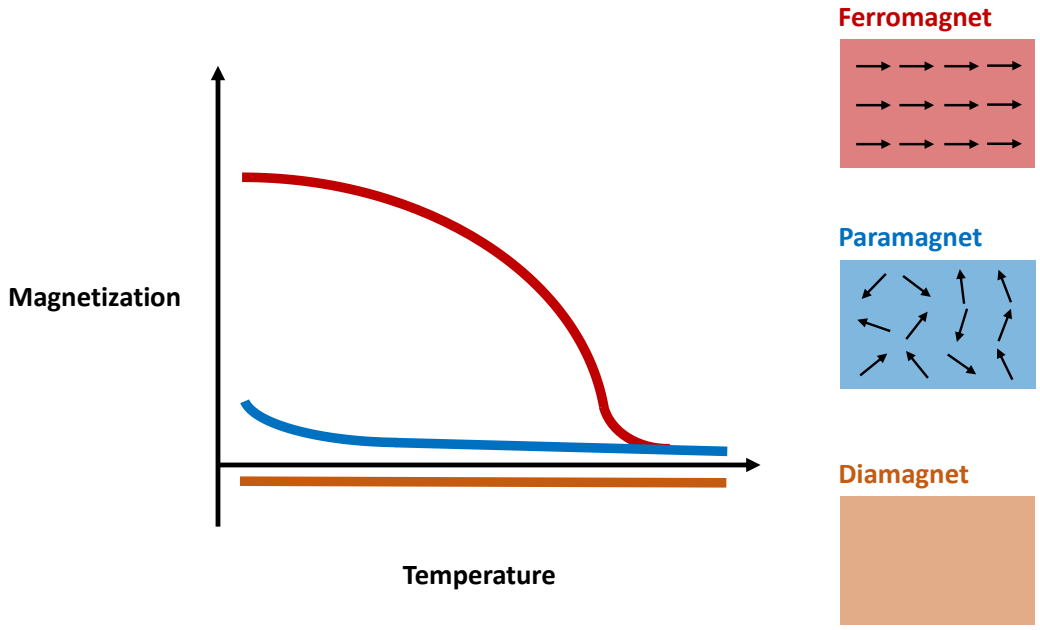


Figure 1.13: Three common forms of magnetism and their dependence on temperature.

1.4.3 Magnets require two ingredients: magnetic moments and exchange

Ferromagnetism, the alignment of many microscopic magnetic moments, requires two ingredients. First, it requires that a material has magnetic moments in the first place. And second, it requires that the exchange interaction is strong enough for those microscopic magnetic moments to cooperatively order.

1.4.4 Magnetic moments

Fundamentally, electron magnetic moments come from two sources: spin angular momentum and orbital angular momentum.[76] Spin, denoted S, is the quantum mechanical angular momentum intrinsic to particles. Orbital angular momentum, denoted L, comes from the orbit of the electron around its atom. Sometimes their total, denoted J, is a good quantum number to use.

effective masses, amplifying their diamagnetic response.

[76] A third source is from classical motion through space. This effect is weak and opposes applied magnetic fields, resulting in diamagnetism. Because it is so weak, it only clearly appears in materials with no unpaired electrons or in materials where the conduction electrons have small effective mass.

The Zeeman effect describes how electron magnetic moments feel a torque that tries to align them with an applied magnetic field. Because this torque is relatively weak, it takes very low temperatures and very high magnetic fields to overcome thermal fluctuations and fully align an individual electron magnetic moment. However, with the exchange interaction between neighboring electrons, they can stay aligned at warmer temperatures and in zero applied magnetic field.

1.4.5 Exchange

The exchange interaction is a quantum phenomenon that explains why electron spins align with one another. Because fermions have wavefunctions that are antisymmetric under identical particle exchange, either their spatial component or spin component is antisymmetric too (but not both). Antisymmetric spatial wavefunctions tend to keep the particles further apart from one another, which is energetically favorable because electrons repel. Since the spatial part of the wavefunction is antisymmetric, the spin part of the wavefunction must be symmetric. This is why wavefunctions with aligned spins have lower energy.

The strength of the exchange interaction depends strongly on the overlap between the individual electron orbitals. Electrons in orbitals with a high degree of overlap will experience a strong exchange interaction. However, if the orbital overlap is too large, the electrons will tend to delocalize and conduct, diminishing the exchange effect.

Direct exchange

When two electron orbitals overlap directly, this is known as direct exchange, or sometimes just exchange. This is the common mechanism of exchange in elemental metals like iron and gadolinium.

Superexchange

In many oxides, such as the ones I study, $3d$ transition metal ions are not adjacent in their crystal lattice, but separated by O_2^{2-} ions. Nonetheless, exchange can occur through the intermediary oxygen ion. This is known as superexchange. Whether the interaction aligns the magnetic moments parallel or antiparallel depends on the bond angles and orbital filling, and can be described by the three Goodenough-Kanamori-Anderson rules.[124, 125, 126, 127]

Antisymmetric exchange

Antisymmetric exchange, also known as the Dzyaloshinski-Moriya interaction, is another type of exchange that occurs in low-symmetry materials with spin-orbit coupling.[128, 129] It is responsible for magnetic skyrmions.[130]

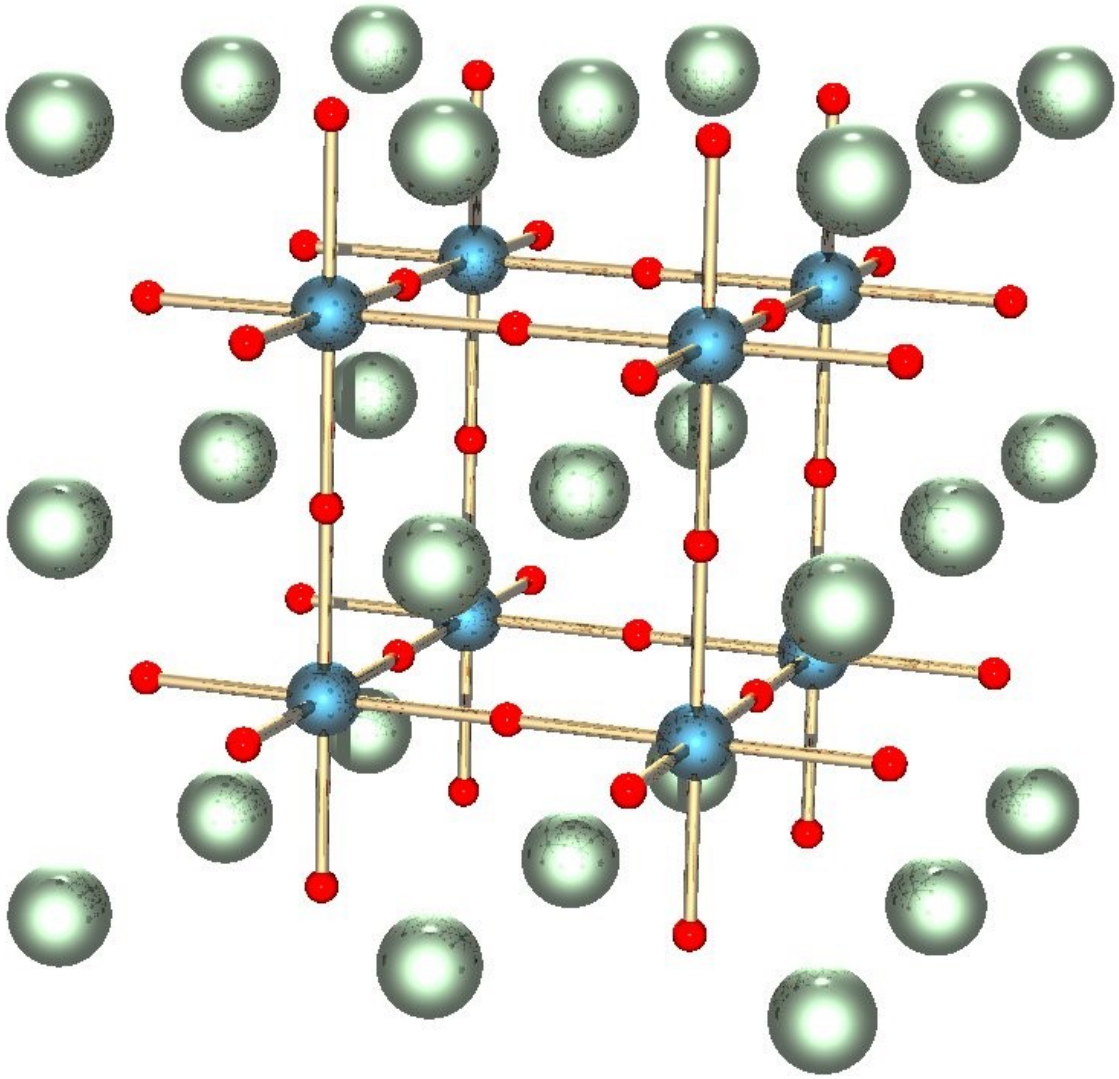


Figure 1.14: The materials I studied have the perovskite crystal structure. Perovskites are oxides with the chemical formula ABO_3. The green spheres represent the A-site atoms, the blue spheres represent the B-site atoms, and the red spheres represent oxygen atoms. For many perovskites, the interesting action takes place on the B-site atoms and depends sensitively on the chemical bonding between the B-site atom and the six oxygen atoms that surround it. In particular, the magnetism of insulating magnetic perovskites often comes from the superexchange interaction through oxygen atoms in eacheach B-O-B chain. Perovskite picture in the public domain, released by user Cadmium at English Wikipedia.

Double exchange

Double exchange is an interesting mechanism that takes place between mixed-valence $3d$ ions with both localized and itinerant electrons. A classic example are the manganese perovskites, such as $La_{0.7}Ca_{0.3}MnO_3$, which has a mixture of Mn^{3+} and Mn^{4+} ions.

1.4.6 Spintronics

Digital electronics are commonplace in modern life, and becoming more commonplace with each passing year. They have transformed the workplace, increasing productivity so much that some economists have deemed it the third industrial revolution.[77][133]

Digital electronics work by encoding information in the position of electrons.[78] Today, the best and cheapest way to shuttle electrons around is inside of conducting material (i.e., wires). Unfortunately, moving electrons through a material costs energy each time an electron collides with the material's atomic lattice.[79,80] This fundamental, nearly inescapable energy cost is why electronics consume so much energy. Alternative, lossless materials like superconductors or topological insulators or even vacuums are a possible way to circumvent this problem of electron scattering, but

[77]However, economists do not entirely agree on what makes an industrial revolution an industrial revolution. I have seen economists suggest as many as six independent industrial revolutions[131] and as few as zero,[132] though one, two, and three seem to be the commonest counts.

[78]Annoyingly, at least to me, many scientists working on spintronics introduce their research carelessly by saying that spintronics encodes information in the spin of the electron while electronics, in contrast, uses charge. But this is ridiculous—the charge of the electron is constant! What isn't constant is the wavefunction of the electron, and the wavefunction is factored into a spatial component and a spin component. Therefore, I think it's natural to say that electronics encodes information in the spatial positions of electrons (leading to voltage and current) whereas spintronics encodes information in their spins (leading to magnetism). Of course, I'm not being totally fair. A more macroscopic view might say that electronics exploits charge density whereas spintronics exploits spin density. And in this macroscopic view based on density, then the spin vs. charge dichotomy is more defensible.

[79]Incidentally, this fundamental property of electrical conduction is why voltage is needed to sustain current.

[80]Also, this explains why transmission lines are operated at high voltages and made of metal. By delivering high voltage electrons, you need to send fewer for the same amount of power, and fewer electrons sent means fewer collisions which means less power loss. Secondly, by making power lines out of conductive metal, collisions are less frequent (for a fixed current). That said, in real power line design there are other constraints too, like cost and weight, which favor aluminum over its more conductive cousins, such as copper, silver, and gold.

none have easy paths to success. So instead, another plan is to start using electron spin instead of electron position to encode information. This avoids the central power loss issue of electronics because it avoids the problem of electron scattering.

For these reasons, spintronics has the potential to be more energy efficient than electronics. However, we're a long way from a world of magnet-based computers. And even if we do get there, electronics won't be over. We'll still need ways to interface magnetic components with electrical components as way to bridge new technology with old.

Information can be encoded in many ways. In transistors, information is encoded by either many or few electrons in the transistor gate (or equivalently, by any property that is concomitantly controlled by the gate electrons). Magnets, on the other hand, store information in very different ways. Magnets are especially well suited to encoding binary information - single electrons naturally have two spin states. However, configurations of many spins can encode information in other ways too (domain wall positions, skyrmions, vortex polarity, resonance etc.).

Today's most well-known magnetic technology is the hard drive. To convert an electric signal to a magnetic signal, it uses a small electromagnet. To convert a magnetic signal back to an electric signal, it uses tunnel magnetoresistance.

Another magnetic technology, still on the horizon, is MRAM (magnetic random access memory). In MRAM, electric signals are usually converted to magnetic signals by spin transfer torque. Converting a magnetic signal back to an electric signal uses tunnel magnetoresistance, just like a hard drive.

Right now, the best way to turn a magnetic signal into an electric signal seems to be tunnel magnetoresistance, in which flipping the magnetization of a material can scale its resistance by up to a factor of 6 or so at room temperature. Going the other way, the best way to turn an electric signal to a magnetic signal is probably spin transfer torque, although the inverse spin hall effect has potential too. But can we do better?

Ferromagnetism needs two ingredients: spins and exchange (and the exchange needs to be stronger than destabilizing thermal noise). So if you want to turn ferromagnetism on and off, you need to turn either spin on/off or exchange on/off.[81]

The most common approach is to use an electric field to affect exchange (or, how exchange competes with bandwidth). Ferroelectric materials can change their crystal structure in response to an electric field, and a change in crystal structure can lead to a change in the strength of the exchange interaction.

You can classify technologies into three functional categories: transportation, storage, transformation (you can also include production as a subset of transformation). This applies to digital information technology reasonably well. Production of information (from the system's perspective)

[81] You cannot really turn exchange off, but you can make it weak enough so that the ordering it would induce is is too weak to survive thermal noise.

comes from input devices like keyboards or sensors. Transportation of information occurs along wires and fiber optic cables. Storage of information occurs in hard drives and RAM. And information is transformed in processors built of transistors.

Where can spin be used best? Let's look at each category one by one, starting with transportation.

What about transportation? For transportation, there aren't many technological options. Wires and fiber optics seem best (and in the sense that they harness electromagnetic waves, they are basically the same, just operating at different frequencies with different dielectric losses). There have been ideas of using nanomagnets in a cascading domino configuration, but in my honest opinion the reliability engineering seems totally impractical without wild jumps in our fabrication abilities. Over shorter distances, spin currents can carry spins encoding information.

How about storage? Magnets are already used in hard drives, and possibly MRAM. With both of these, the cost of production is relatively fixed, meaning that the cost per bit varies inversely with the density. So the question is, how dense can these go? With hard drives, as far as I can tell, the limits seem to be driven the size of write head, read head, and magnetic media. Assuming the first two can be made arbitrarily small, the third will limit you, and it has a limit based on how stable you want your memory to be. However, if you can keep boosting your coercivity, you can keep shrinking the bits (though high coercivity might require new writing technologies like HAMR). Also if bits can refreshed as they are in RAM, then you might also be able to go smaller (assuming the relaxation process is nonlinear).

Really though, the grand prize lies in information transformation. When people complain about the high power of electronics holding back computer chips, they are talking about electrical transistors and the subthreshold slope problem. Here, spintronics may be the next evolution. The most famous proposal is the SpinFET, proposed in 1990 by Datta and Das.[134] But 25 years later, and we're still far from making them. What else might be out there? Really it all depends on how we represent information with spins and how we get those spins to interact with each other in a low noise way. But it seems safe to assume that materials that can convert magnetism to electricity and vice versa will be essential building blocks.

1.5 Connection to my research

Although my introduction has spent time broadly examining the technological landscape for spintronic devices, my actual research has been focused at a much lower level. My goal has been to help understand the fundamental physics and materials science of complex oxides, a class of materials which exhibits many correlated-electron phenomena not seen in *sp*-band semiconductors. The hope is that by understanding these complex phenomena, we can provide the asphalt to pave the way to a superior spintronic future.

The rest of this dissertation is dedicated to discussing my specific efforts in detail. Chapter 2 will describe the major experimental tools that enabled me to deposit thin films and then measure their crystal structures, elemental compositions, surface topography, electrical properties, and magnetic properties. Chapter 3 will detail my efforts to understand the conductivity and magnetoresistance of the $LaAlO_3/SrTiO_3$ interface and modify it through the use of rare-earth dopants. Chapter 4 will then detail my efforts to understand the magnetism of cobalt perovskite thin films, and in particular how I used SQUID magnetometry and X-ray magnetic circular dichroism, in combination with X-ray diffraction, to measure how chemical pressure and epitaxial strain can affect the magnetism of these materials. Finally, Chapter 5 will summarize my major results and present some thoughts on the future.