Notes

Notes for Chapter 1

The Main Themes. It is obviously an over statement to say that which technology vendors survive over a five year period can best be viewed as a random walk. The perspective of this book is to focus on the underlying structure of the industry as a whole and from this perspective whether Company A, B, or C brings relational databases to market is less important than the fact that once relational database technology is developed it is relatively easy to predict that some company will bring it to market and much harder to predict which company.

Companies themselves can be modeled and these models can be used to predict how likely a company is to remain in the market during the next year. When models like these are built for technology companies, although predictions can be made based upon internal and external factors, it is not a bad approximation over a decade to use a random walk model in which there are probabilities each year that the company will either grow larger, grow smaller, stay the same, or be acquired.

A Billion IP Addresses for Each of Your Children. Routes on the Internet change all the time. Figure 1.1 contains a portion of a traceroute from 2009. As described later in the book, data sent between two computers on the network is divided into what are called packets and these packets are sent from one device to another on the Internet in a series of what are called hops. The Linux traceroute command sends three probes from one device to another. The traceroute commands provides the hop number (column 1 in the figure), the time in milliseconds for each of the three probes (columns 2, 3 and 4), the IP address of the device (column 5), and the Internet address of the device (column 6).

The current Internet uses 32 bits to specify an IP address. The IP address is divided into three components. The first specifies the type (Class A begins with a zero, Class B with a 10, and Class C with a 110), the second the id of the network, and the third the ID of the host or computer. The maximum number of hosts that can be addressed in this way is 2^{32} or about 4 billion, which seemed a lot in 1984 when this scheme was first introduced.

Class	Network	Hosts
0	7 bits	24 bits
10	14 bits	16 bits
110	21 bits	8 bits

The IPv6 scheme uses 128 bits, which can be thought of as divided into eight 16 bit segments. Each 16 bit segment can be written as 4 hexadecimal digits. For example,

```
FEDC:BA98:7654:3210:FEDC:BA98:7654:3210
```

is an IPv6 address. The maximum number of hosts that can be addressed in this way is 2^{128} or about 10^{38}, which seems to be a lot today.

Network	Hosts
64 bits	64 bits

The estimates for the world population for 1981 and 2003 are from the US Census web site (www.census.gov). The description of the IPv4 internet protocol is defined in in the DARPA Internet Program Protocol Specification [125]. The description of the IPv6 internet protocol is from [63].

Three Ages of Computing. This description of the alphabetic Greek number system is taken from [112, Hist-Topics/Greek_numbers.html].

The SAD History of Computing. The number system familiar to us is called the Hindu-Arabic number system and requires the symbol zero. The earliest historical evidence for the use of the symbol zero are from 9th Century manuscripts from India [23].

The following two formulas, which were developed in about the 17th century, turn multiplication and division into addition and subtraction:

$$ab = \exp(\log a + \log b)$$

$$a/b = \exp(\log a - \log b)$$

Why Symbols Matter. This section is based in part on [158].

Algorithms as Recipes for Manipulation Symbols. Here is a simple five line Python function to illustrate how square roots can be computed iteratively:

```
def square\_root(a):
    x[0]=a/2.0
    for i in range(1,20):
x[i] = (x[i-1] + (a/x[i-1]))/2.0
print x[i]


x={}
print 'sqrt of 5934939'
square\_root(5934939)
```

Here is another example of an algorithm that generalizes the algorithm for finding square roots discussed in this section. In the 1660's Newton introduced the following simple algorithm for computing the solutions of equations such as $f(x) = 0$. As an example, to find the square root of a, let $f(x) = x^2 - a$. To find the solutions of a quadratic equation, let $f(x) = ax^2 + bx + c$. To find the cube root of a, let $f(x) = x^3 - a$. And to find the solutions of a cubic equation, let $f(x) = ax^3 + bx^2 + cx + d$.

The algorithm depends upon the function $f'(x)$ which Newton introduced and specifies how quickly $f(x)$ changes. For example, if $f(x)$ is a formula for the path of a trajectory, then $f'(x)$ is a formula for its velocity.

1. Take a guess, no matter how bad. Call the guess x_0.
2. Let $x_{n+1} = x_n - f(x_n)/f'(x_n), \qquad \text{for } n \geq 0.$
3. If $x_{n+1} - x_n$ is small, stop (the answer is x_{n+1}); otherwise, goto Step 2.

Before Newton's algorithm, there was no simple general method to solve algebraic equations, no matter how good the symbols you had nor how fast the computing device. Newton's algorithm changed all that.

Computing Devices: From Paper to Chips. For more information about the Rhind papyrus, see [12]. For more information about Euclid's Elements, see [44].

Case Study: The Slide Rule. A still useful and influential reference book of mathematical tables is the *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* [1]. The handbook is over 1000 pages long and contains a wide variety of different mathematical tables.

For the next few paragraphs we discuss in more detail how the multiplication of two numbers can be computed by adding two related numbers. For example, in a first course in trigonometry, one usually sees the following formula:

$$\sin(A) * \cos(B) = (1/2)\sin(A + B) + (1/2)\sin(A - B).$$

Formulas like these, together with a table of sines and cosines, can be used to reduce multiplications to additions as follows:

1. To multiply a times b, first use a tables of sines to find numbers A and B such that $a = \sin(A)$ and $b = \sin(B)$.

2. Compute $A + B$ and $A - B$.

3. Use the table of sines to find $\sin(A + B)$ and $\sin(A - B)$.

4. Add the two numbers together from Step 3 and divide by 2. By the formula above, this is the desired product of a and b.

In these days of calculators and spreadsheets, it is hard to appreciate why in many cases it is easier to multiply two numbers together using these four steps involving additions and mathematical tables instead of multiplying the numbers directly. Perhaps the best way to understand this is to try one night when you are having trouble sleeping, to multiple the numbers 1.84825884 and 489.83238535 both ways.

Today's logarithm function and its inverse, the exponential function, are quite close to the functions that John Napier introduced in the 16th century. Their are two essential properties of these functions: First, multiplication is reduced to addition:

$$ab = \exp(\log a + \log b).$$

Second, division is reduced to subtraction:

$$a/b = \exp(\log a - \log b)$$

Here exp can be thought of as a function that undoes a logarithm, in the sense that, if in a table of look up values, the logarithm function is the look-up from left to right, then the exp function is the complementary look-up from right to left. These two formulas and a table of logarithms were used to add and multiple numbers for sometime before the invention of the slide rule.

The history of the logarithm is adapted in part from [40]. Information about Edmund Gunter is adapted in part from [112, Biographies/Gunter.html].

From Mainframes to Devices. The quote from Winston Churchill is from a speech he gave to the Royal College of Physicians in London, in 1944. The quote from Ludwig Wittgenstein is from [153, page 1]. For information about the IBM System 360, see [128]. For data about the diffusion of personal computers, see [46].

Case Study: Punch Cards. This case study is adapted from [83].

The Second Era: Desktop Software Applications and the PC. Information about VisiCalc is from [14].

The Fourth Era: Clouds of Devices and Services. It would be convenient if the fourth era had a well recognized name, such as the term "web," which so nicely characterized the third era. Instead, various terms are used today for the fourth era, such as "Device Net" and "EmNet" [100]. Also, sometimes the term "Cloud" is used, although this is generally used for something else [84]. It is still very early in the Fourth Era and I expect a name will catch on before long.

Case Study: Routers In fact, not all computers on the internet have unique IP addresses. First, with the current IP addressing scheme (IPv4), they are simply not enough addresses. Second, for security reasons, many computers are behind firewalls and have IP addresses that are part of private networks. This is not all that different than the telephone system. Telephones that are part of a PBX system usually share a common phone number and are accessed through an extension number.

The example in this section of how routers work is adapted from [154], page 160–164.

Notes for Chapter 2

The Title. The Oxford English Dictionary defines commoditization as a noun with the meaning of commodification. Commodification is defined as "The action of turning something into, or treating something as, a (mere) commodity; the commercialization of an activity, etc., that is not by nature commercial [118]." As used in this Chapter, commoditization is perhaps better defined as "the process by which a good or service transitions from a relatively scarce and expensive item to one that is widely available and inexpensive." In short, items that are commodities are ubiquitous.

Christmas and Easter. Easter Day is often described as the first Sunday after the full moon that occurs next after the vernal equinox [146]. The subtlety is that the full moon in this description is not the full moon as observed by an astronomer but rather an ecclesiastical full moon as determined from tables agreed to by an ecclesiastical council. These tables roughly, but not exactly, follow the astronomical full moon. The reference [146] contains a good description of how these tables work and how they differ from astronomical observations.

Danti's Law - The Commoditization of Time. Meridian lines are describe on page 23 of [73]. Information about the accuracy of various types of clocks and watches is from [107]. Information about the solar year using the calendars of Julius Caesar's, Gregory XIII and Kahan is from [69].

The Commoditization of Space. This section is based upon information about Harrison and his chronometers from the following sources: [129], [67, pages 169–177], and [68, pages 26–29].

Moore's Law – The Commoditization of Processing Power. The quote by Gordon Moore is from a video transcript [94]. The projection is contained in the 1965 Electronics Magazine article [93].

Commoditization is All Around Us. The history of the meter is taken from [68, pages 278–280].

Storage and Johnson's Law. In 2011, you can get a 2 TB disk for less than $100. In 2007, you could buy a Seagate 750 GB Barracuda 3.5 inch Ulta-ATA/100 internal hard drive for $309.00 from Amazon (the retail prices is $399.99). This disk drive has four platters and spins at 7,500 rpm. is 4 inches by 5.8 inches and weighs 1.5 pounds.

The data about disks from mainframe computers is from the Disktrend web site. The first three columns of this table was extracted from www.disktrend.com on June 10, 2002. The remaining columns are computed.

Software and Stallman's Law. Richard Stallman's vision for free software is described in [136]. Stallman is a social activist. A good overview of his social activism can be found on his web site (www.stallman.org).

GNU is an abbreviation for *G*nu is *N*ot *U*nix.

Microsoft's Client Business Unit had $13.2 billion of revenue in 2006 and $12.1 billion of revenue in 2005 [91]. Microsoft's overall revenue for 2006 was $44.3 billion and $39.8 billion in 2005.

In June 2005, Sun Microsystems posted more than 5 million lines of source code from their Solaris operating system as part of an initiative that transforms a proprietary operating system that Sun estimates cost $500 million into an open source operating system called Open Solaris. [138].

A good analysis of how to estimate source lines of code (SLOC) has been done by David A. Wheeler [157]. In particular, the SLOC estimates for the Red Hat Linux distribution are from this paper. This paper also contains the COCOMO estimates. More information can also be found on his web site (www.dwheeler.com/sloc).

The size of the Debian Linux distribution are from the papers [4] and [54]. These references also contain information about the SLOC for the various Microsoft Windows distributions.

There is no agreed upon methodology for measuring

SLOC. Different tools will report somewhat different results. It is especially difficult for an outsider without access to the Microsoft source code to estimate the SLOC. For this reason, all the SLOC of code should be considered approximate, but especially those for the Microsoft Windows systems.

My comments about designs for successful open source software projects is in adapted in part from [152].

Data and the Bermuda Principles GenBank is maintained by the U.S. National Institute of Health. More information about GenBank can be found at the NIH web site (www.nih.gov), which is the source of the information here.

According to the Netcraft Web Server Survey there were 108,810,358 distinct websites in February, 2007. In December 2010, Netcraft estimated that there were approximately 266,848,493 web sites. So assuming that 0.1% of these contained some open data, then there would be over 100,000 sources of open data. This is a very, very rough estimate.

Network Effects. Since the convention when writing semi-popular books about technology is to make technical assertions, but not always to check them, we will assume in this section, for simplicity, that Metcalf's Law is true.

The quote from Bill Gates is from [48].

For additional material on Metcalfe's Law and network effects, see [116] and [117] and the references contained there.

The list of characteristics of network effects in technology markets is adapted from [131] and [149].

There is no name that I know of to refer to the network effect associated with software and data. In this book I use the names *Linus' Law* and *Pearson's Law* after Linus Torvalds, the the lead developer of Linux, and Karl Pearson, a statistician who lived from 1857 to 1936.

Notes for Chapter 3

A Case Study in Innovation: Approximating Solutions to Equations The Pythagorean Theorem, relating the shorter sides a and b of a right triangle to the longer side c: $c^2 = a^2 + b^2$, is usually one of the first theorems a student learns.

The first known statements of the relation appear on Babylonian tablets dating from the period 1900–1600 B.C. The first proof is thought to be due to Pythagoras (c.560–c.480 B.C.) or to someone from Pythagoras' school. All we know for certain is that the first written proof that is extant is due to Euclid (c. 300 B.C.) Euclid's treatment of this theorem was the standard treatment for hundreds of years.

It is important to remember that the equation as written today is relatively recent: Euclid gave a geometric (not algebraic) statement and a geometric proof.

If we substitute $a = b = 1$ into the equation $c^2 = a^2+b^2$, then $c^2 = 1^2 + 1^2 = 2$ and c is the square root of 2. More generally to find the square root of a number c, we need to need a solution x to an equation of the form:

$$f(x) = x^2 - c = 0.$$

With this notation, the Newton-Raphson iteration takes the form:

$$x_{n+1} = x_n - f(x_n)/f'(x_n), \qquad n \geq 1.$$

Continuing our example of computing the square root of c using $f(x) = x^2 - c$, we have that $f'(x) = 2x$. More generally, $f'(x)$ is what is called the derivative of $f(x)$ and measures the rate of change of the function $f(x)$.

Here is another example of the Newton-Raphson method. To find a seventh root of a number c, i.e., solutions to the following equation:

$$x^7 = c,$$

you can use the following Newton-Raphson iteration:

1. Begin with a guess, say $x_0 = 1$.
2. Compute

$$x_{n+1} = x_n - \frac{x_n^7 - c}{7x_n^6}, \qquad n \geq 1.$$

3. If x_{n+1} and x_n are close together, stop because you have found the seventh root of c. If not, return to Step 2.

Of course, computers are well suited for computing these types of iterations.

For more information, about the history of the Newton-Raphson algorithm, see [34, page 169].

The simple computer program for computing square roots would never work in practice: First, the program runs into an error if we divide by zero. So we must check for this. Second, sometimes five iterations is enough, sometimes it isn't. We must check for this. Soon we find ourselves with a longer program.

Checking the various different special cases cases quickly takes over as the major task. It turns out that it is easy to leave out various special cases and that as you add more and more cases it becomes easier and easier for them to begin to conflict. This is a simple example of why it is difficult to write even simple programs.

Most computers today follow IEEE Standard 754 for binary floating point arithmetic. Under this standard, floating point numbers are written as +/- d.ddd x 10eee. The number eee is called the exponent and the number d.ddd is called the significand or mantessa. For example, a 32 bit representation for floating point numbers allocates 1 bit for the sign, 8 bits, for the exponent, and 23 bits for the significand. This provides about 8 decimal digits of accuracy for the significand [52].

Even though only 8 decimal digits are significant, the default behavior for simple programs is to write out more than 8 digits after the decimal point. It is important to realize that in general these are incorrect. To get more accuracy, one either needs to use a 64 bit or larger representations for the floating point numbers or to use specialized programs that use variable length representations. With the latter, operations are much slower, but much greater accuracy can be obtained.

A Case Study in Clutter: Business Intelligence. This section was written in 2003. At that time, a key word search using "business intelligence" was done using the online DM Review site www.dmreview.com on March 9, 2003; 1392 documents were returned. A key word search using "business intelligence" was done on Google on March 9, 2003; Google reported that 1,1600,000 were identified containing this phrase. In 2011, a search on Google for "business intelligence" returned approximately 124,000,000 documents containing this phrase.

The Imperative to be in the Upper Right. There are a lot of industries in which a company can choose from in order to be a leader. The U.S. Department of Commerce's book describing the North American Industry Classification System (NAICS) is 1390 pages long [143]. The NAICS uses six digits to identify particular industries: the first two digits designate a business sector, the third digit designates a subsector, the fourth digit designates an industry group, and the fifth digit designates a particular industry. For example, NAICS code 511210 includes packaged computer applications software. See [143] for more details.

Who Clutters. In the Seventh Century B.C.E, standard technology for predicting the future included the entrails of beasts and the motion of the stars. Virgil was a Roman poet who was born in 70 B.C.E, died in 19 B.C.E., and who wrote an epic, which consolidated some of the legends about the ancestors of the Romans. Here is a description of a seer from Book X of his Aeneid [150]:

> Third in the line was Asilas, the mighty seer
> who mediated between men and gods, and who

> knew the secrets held by the entrails of beasts, the stars in the sky, the voices of birds, and the flash of presaging thunderbolts, ...
>
> Virgil, The Aeneid, Book X, translated by G. R. Wilson Knight [150].

Sources of Clutter: Features. The five page marketing brochure "Crystal Reports XI: Feature Comparison by Version and Edition" contains over 150 features. This brochure was retrieved from the the Business Objects web site on November 20, 2005.

A Case Study in Innovation: Databases. The market research firm IDC estimate of the size of the database market in 2004 was reported in eWeek in the article "Study: 2004 Database Market Grew 12 Percent" [148]. IDC estimated the following worldwide sales of databases as follows: Oracle - $6.2 Billion (41 percent of the market); IBM - $4.59 Billion (31 percent of the market); Microsoft $2.01 Billion (13 percent of the market); Other vendors 2.21 Billion (15 percent of the market).

A brief description of the history of SQL is in [99] pages 162-164.

A Case Study in Innovation: Searching for primes. Since a natural number n always has 1 and n as factors, these are not considered when deciding if a number can be factored into a product of smaller numbers. For example, even though $7 = 1 \times 7$, 7 is still considered to be a prime number. If a number n is not prime it is called *composite* .

Historical facts about Mersenne primes are from [27] and [28], which are excellent resources. The tables listed the large known prime by year from there also. Some of the table entries were computed using a simple Python program.

Here is a Python program for computing primes using the Sieve of Erastosthenes.

```
def sieve(n):
  candidates = range(1, n+1) # [1, ..., n+1]
  candidates[0] = 0   # candidates[0]=1 is not prime
  for p in candidates:
    if p:   # skip zeros
      if p*p>n:
        break   # done
      for q in range(p*p, n+1, p): # sieving
        # candidates[q-1]=q is not prime
        candidates[q-1] = 0
        # return non-zero candidates
        return filter(None, candidates)

# print first n primes
n = 100
print sieve(n)
```

A Python implementation of the Sieve of Erastosthenes.

Lock-In or the Tyranny of Vendors and Users. The discussion of vendor lock in strategies is adopted from [131, Page 117].

A Case Study in Innovation - Routing Packets. Class A, B, and C IP addresses are being replaced by a new approach, which more efficiently allocates the address space called Classless Inter-Domain Routing or CIDR.

Notes for Chapter 4

The Basic Equation of Marketing. Innovators, early adopters, main street, and laggards are well described in [92]. The core ideas are also well described in [38].

How Long to Reach Main Street. The average price for a car in 1914 is from the Cadillac Database [26]. See also [6] and [22]. The corresponding 2003 price was calculated using the US Department of Labor's Bureau of Labor Statistics estimate of the consumer price index of 10.0 for 1914 and 181.7 for 2003 [18]. According to the US Census

Bureau, the US population is about 270 Million, so that using an estimate of 100 Million for the consumer market underestimates its size.

Case Study: The Nike Pagasus. The history about the Nike Pegasus is from the Nike web site [105].

Technology Roadmaps. The table describing the semiconductor roadmap is from the International Technology Roadmap of Semiconductors (ITRS) [66].

Case Study: Grid Computing. The following facts about SETI@home are from a 2002 article about the project [5]. SETI@home was announced in 1998 and the first software was released in May, 1999. By July 2002, over 3.8 million individuals had participated in the project by downloading the SETI@home application and donating cycles to the computation. During the 12 month period starting in July, 2001, the average throughput of the project was 27.36 Teraflops. A Teraflop is a trillion floating point operations per second, such as an addition or multiplication of two floating point numbers. According to the 2002 Top 500 List of the most powerful computers, the SETI@home virtual supercomputer ranked as one of the top ten supercomputers that year [139].

The USA Today article describing shared computing is from [70]. The business week article is from [123]. An excellent introduction to grid computing is the book The Grid: Blueprint for a New Computing Infrastructure [47].

Context. There is no term that I am aware of that refers to factors such as complexity, lock-in and standards that affect the rate at which new technology is adopted by the marketplace. For lack of a better alternative, I use the term *context* in this book.

In [58], the Boston Consulting Group analyzed various markets and pointed out that the top three vendors usually prospered, while the others struggled along.

See [116, page 37] for more discussion about the role of efficiency for inexpensive goods or services.

Forces Effecting Technology Adoption. Information about the growth in the user base of Hotmail is from a Microsoft Press Release [89].

Case Study: Adoption of Relational Databases. The description of the technology adoption of relational databases is based in in part on [99], [55] and [135]. The time line and major events are described in [99, pages 159–169], and [37].

The complexity of querying navigational databases, such as hierarchical and network databases, is described in textbooks on databases, such as [37] and [132].

The number of installations and downloads for MySQL is from the MySQL web site (www.mysql.com).

Case Study: Adoption of Open Source Linux Kernel. The email is from Linus Torvalds [140]. The estimates for the number of Linux users is from [86].

Notes for Chapter 5

Introduction. The quote "Big Data is a Big Deal" is the title of Tom Kalil's March 29, 2012 blog post (www.whitehouse.gov/blog/2012/03/29/big-data-big-deal).

Thinking About Big Data. The study by Peter Lyman and Hal R. Varian about how how much new information is created each year can be found on their web site called "How Much Information? 2003" [82]. The project carefully collected data in 1999 and 2002 and published an analysis of the data in 2000 and 2003. The quote is from their 2000 analysis. In 2003, they revised their 1999 estimate upwards from 1–2 Exabytes to 2–3 Exabytes. It would be nice to include a more recent estimate, but I am not aware of a more recent one that is as authoritative as their 2000 and 2003 studies. The 2003 study contains the estimate that new information from streaming data (such as a person to person telephone call) is about 3-3.5 times larger than the new information from data that is stored on media, such as a hard disk.

Marissa Mayer, the Vice President for Search Products & User Experience at Google, gave a talk at Xerox PARC in Palo Alto California on August 13, 2009 (www.parc.com-/event/936/innovation-at-google.html).

The table illustrating the sizes of kilobytes, megabytes, etc. of storage is adapted from [159].

The Commoditization of Data. The announcement by Nikon that it is concentrating on manufacturing digital cameras is from January 12, 2006 edition of the New York Times.

The Data Gap. The data in the first table in this section about earned doctorates is from [103]. The information from the second table about the amount of storage required for different types of multimedia files is from [57].

The British naturalist Charles Darwin lived from 1809 to 1882. The British Penny Post was introduced in 1840. With the Penny Post, a letter could be sent anywhere within England for a penny.

The estimate of 28,800 minutes corresponding to 20 days assumes that a messenger on a horse travels approximately 50 miles per day. Of course, your mileage (carrying data on a horse) may vary.

In December 2004, Google announced that it is working with Harvard University, Stanford University, the University of Michigan, Oxford University and the New York Public Library to digitize the books in their libraries and make them available through Google [17].

Extracting Knowledge from Data. GenBank is a publicly available database containing genetic DNA sequences maintained by the U.S. National Library of Medicine and National Institute of Health. As of February, 2008 it contained approximately 85,759,586,764 bases from over 260,000 different organisms. See [97] and [9].

Here is the argument that the square root of 2 is irrational. Assume not. In other words, assume that there are integers p and q such that $(p/q)^2 = 2$. Then $p^2 = 2q^2$.

Factor both p and q into a product of primes. Then p^2 is factored into a product of the very same primes as p, but each occurs twice as often in the factorization of p^2 as it does in the factorization of p. Therefore, p^2 has an even number of prime factors. So does q^2. Now since q^2 has an even number of primes, $2 \cdot q^2$ has an odd number of primes (it has just one more). This is a contradiction, q^2 cannot have both an even and an odd number of prime factors. We conclude that the equation $(p/q)^2 = 2$ has no solution for integers p and q. This is the same thing as saying that the square root of two cannot be written as a quotient of two integers and is therefore irrational.

For more information about how simulation is used to study surface temperatures and global warming, see the National Academy of Science report on surface temperature [101].

For the story about why Ulam named the method Monte Carlo, see [61], page 238.

In October, 2004, the International Human Genome Sequencing Consortium, reduced the estimated number of human protein-coding genes from 35,000 to only 20,000–25,000 [111]. Although the number may continue to shrink, our current understanding of the coding of proteins, protein interactions, and metabolic pathways is based upon much more than just genes.

Case Study: Mar's Orbit, Brahe's Data and Kepler's Law Tycho Brahe did not publish his celestial observations, which occupied 17 volumes, but they were used by other astronomers, including Kepler. Brahe's data for the orbit of Mars can be found in [120]. See "A Brief History of Cosmology [114]" for a concise online history of cosmology describing the models of Ptolemy, Copernicus, Brahe, Kepler, and Newton.

Pearson's Law. It seems clear that the value of a column of data grows as it is included in larger and larger collections of other columns, as long as their is a common key for linking different rows in the various columns. On the other hand, whether the value grows linearly, quadratically, or at some other power is not so clear. Odlyzko and Tilly investigated the growth in power of a network containing n sources [117]. They conclude that the power of a network grows more like $n \log n$, rather than the n^2 that occurs in Metcalfe's Law. The same power may very well apply here.

Karl Pearson (1857–1936) introduced a formula for quantifying the correlation between two columns of data and the chi-squared test, both of which are used extensively in statistics. For more information about him, see [124]. Note that if two columns of data are normalized using their mean and standard deviation (into what are called z-values), then the Pearson correlation coefficient of two columns of data is simply the average of the row-by-row product of the columns.

The analysis about voting in Palm Beach County is adapted from [133]. The data is from: election.dos.state.fl.us.

Lessig's Law: Data Just Wants to Be Free The gene (UID 31657232) was obtained from the online GenBank database

http://www.ncbi.nlm.nih.gov/genbank/,

which is maintained by NCBI, which is part of the National Library of Medicine.

The Bayh-Doyle Act or University and Small Business Patent Procedures Act was passed by the U.S. Congress in 1980. Prior to this act, research by a university small business, or non-profit organization that was based upon federal funding could not be easily commercialized since the government retained the ownership of all patents and other intellectual property resulting from the research. This changed with the Bayh-Doyle Act, as the first paragraph of the Act makes clear:

> It is the policy and objective of the Congress to use the patent system to promote the utilization of inventions arising from federally supported

> research or development; to encourage maximum participation of small business firms in federally supported research and development efforts; to promote collaboration between commercial concerns and nonprofit organizations, including universities; to ensure that inventions made by nonprofit organizations and small business firms are used in a manner to promote free competition and enterprise without unduly encumbering future research and discovery; to promote the commercialization and public availability of inventions made in the United States by United States industry and labor; to ensure that the Government obtains sufficient rights in federally supported inventions to meet the needs of the Government and protect the public against nonuse or unreasonable use of inventions; and to minimize the costs of administering policies in this area.
>
> U.S. Code 200-212, Title 35, Chapter 18, Patent Rights In Inventions Made With Federal Assistance.

The central dogma of molecular biology is that DNA produces RNA in a process called transcription, and that RNA produces proteins in a process called translation. During the Human Genome Project, it began to be realized that different proteins could be produced from the same genes. For example, it turned out that genes were divided into regions that coded directly for proteins (exons) and regions that do not (introns). By selecting different exons for inclusion, a single gene can produce many different proteins. This is an example of a process called alternate splicing. So even though by the end of the project, the gene count was reduced from about 35,000 to about 26,000, the number of proteins that could be produced from these 26,000 genes was estimated to be much higher than the 35,000 or so proteins than would be produced by 35,000 genes using only the most basic form of translation. By 2004, the estimated number of human genes was reduced further to between 20,000 to 25,000.

BLAST is an abbreviation for Basic Local Alignment Search Tool. BLAST is an algorithm that compares two sequences for similar subsequences [3]. For example, the sequence of a newly identified gene in one organism containing a thousand base pairs can be compared to the entire 4 billion base pairs of the human genome in GenBank.

Case Study: The World Population. The table that estimates the world population at various historical times is from [142].

The Shape of Data. The XML description of the statistical model that classifies an Iris into three types is from www.dmg.org.

Computing statistical profiles while maintaining privacy is an active area of research today. It is important to note that although statistical profiles can be assembled without using *any* personally identifying information, this does not mean that profiles computed in this way cannot be analyzed to reveal personal information. Here is a hypothetical example of how this can occur. Assume that tables of counts are computed that count the number of students that missed 1 day of school, 2 days of school, 3 days of school, etc. Assume that another table of counts contains the average number of days of school missed by students who had various diseases. By putting these two tables together and combining it with some personal knowledge (such as one of your friends missed thirty days of school), you might be able to guess what disease he or she had, especially if there were only a few people in the school that missed that many days of school.

These types of opportunities to identify personal information from summary or aggregated data have been known to be a problem for a long time and is one of the reasons that the U.S. Census only releases information at a fairly high level of summarization.

For more information, about the Google Page Rank algorithm, see [15].

Here is the R code used to compute the page ranks for the example in this section.

```
# initial page ranks
a <- c(0.25, 0.25, 0.25, 0.25)

# number of incoming edges
c <- c(2.0, 1.0, 1.0, 1.0)

# random surfer factor
d <- 0.85

# for simplicity, perform 10 iterations
for (k in 1:10)
{

  a[1] <- (1-d) + d*(a[3]/c[3])
  a[2] <- (1-d) + d*(a[1]/c[1])
  a[3] <- (1-d) + d*(a[1]/c[1] + a[2]/c[2] + a[4]/c[4])
  a[4] <- (1-d)

  wt <- sum(a)
  a <- a/wt

  print(a)
}
```

Case Study: Consumer Credit and Why Not all Databases About 190 Million Americans Are Bad
The data about reported fraud is from the US Federal Trade Commission [144]. The data was retrieved from the web site www.consumer.gov/sentinel on September 1, 2002 for the years 2001, 2002, and 2003; on September 10, 2006 for the years 2004 and 2005; and on March 10, 2008 for the year 2007. The data for 2004 and 2005 was revised when the 2006 data was released. Percentages are computed using

data from [144] and data from Visa about the estimated yearly dollar volume of credit card transactions [151]. Data about the growth in credit consumer credit from 1970 to 2001 is from [141].

Creating Digital Data. For more information about the OdorDB, "Databases for the Functional Analysis of Olfactory Receptors" [36].

Using Data to Make Decisions If a statistical model predicts one of two possible outcomes ($n = 2$) and each outcome can be true or false ($m = 2$), then the number of possible types is four ($4 = 2x2$). More generally, the number of true/false outcomes is $2n$.

Case Study: NASA's EOS. According to an article by Lee Gomes in the Wall Street Journal [53], in August, 2006, there 6.1 million videos, compromising 45 terabytes of digital data, that had been viewed 1.73 billion times. Approximately 50% of the viewers were under 20 years of age. The total time spent viewing the videos was 9,305 years.

References

[1] M. Abramowitz and I.A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. U.S. Department of Commerce, 1972.

[2] Alfred Adler, Reflections, Reflections Mathematics and Creativity, The New Yorker, February 19, 1972, p. 39.

[3] S. F. Altschul, W Gish, W Miller, E.W. Myers, and D.J. Lipman, Basic local alignment search tool, Journal Molecular Biology Voluume 215, Number 3, pages 403-10, 1990

[4] Juan Jose Amor, Gregorio Robles and Jesus M. Gonzalez-Barahona, Measuring Woody: The Size of Debian 3.0, Report on Systems and Communications, GSyC, December, 2004, retrieved from libresoft.dat.escet.urjc.es/debian-counting on January 10, 2007.

[5] David P. Anderson, Jeff Cobb, Eric Korpela, Matt Lebofsky, Dan Werthimer, SETI@home: An Experiment in Public-Resource Computing, Communications of the ACM, Volume 45, Number 11, November 2002, pages 56–61.

[6] Automobile Manufactures Association, Facts and Figures, New York, New York, 1950.

[7] Stephen Baker, Google and the Wisdom of Clouds, Business Week, December 13, 2007.

[8] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell and Eric W. Sayers, GenBank, Nucleic Acids Research, Volume 39, Supplement 1, D32–D37, 2011.

[9] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and David L. Wheeler, GenBank, Nucleic Acids Research, 2008, D25-D30.

[10] Thomas J. Bergin, The History of Computing, 1985, retrieved from www.computinghistorymuseum.org on January 24, 2002.

[11] Tim Berners-Lee, Information Management: A Proposal, May 1990, retrieved from www.w3.org on March 20, 2006.

[12] Carl B. Boyer, A History of Mathematics, Princeton University Press, 1985.

[13] Boston Consulting Group, Perspectives on Experience, 1968.

[14] Daniel Bricklin, Software Arts and VisiCalc, Copyright 2003, retrieved from www.bricklin.com on March 20, 2006.

[15] Sergey Brin and Larry Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, Stanford InfoLab Technical Report, 1998, retrieved from dbpubs.stanford.edu on September 10, 2005.

[16] Frederick P. Brooks, Jr., The Mythical Man-Month: Essays on Software Engineering Second Edition, Addison Wesley, 1995.

[17] Andreas von Bubnoff, Science in the web age: The Real Death of Print, Nature, Volume 438, pages 550-552, 2005, doi:10.1038/438550a.

[18] US Department of Labor, Bureau of Labor Statistics, retrieved from www.bls.gov on November 14, 2003.

[19] Encyclopedia Britannica, Automotive Industry, www.britannica.com/eb/article?eu=114513. retrieved on July 7, 2002.

[20] Encyclopedia Britannica, Computers, www.britannica.com/eb/article?eu=130076, retrieved on August 18, 2002.

[21] Encyclopedia Britannica, Euclidian Geometry, www.britannica.com/eb/article-235561, retrieved on March 12, 2006.

[22] Encyclopedia Britannica, Motor Cars, Volume 15, page 881, Encyclopedia Britannica, Inc., Chicago, 1957.

[23] Encyclopedia Britannica, Numerals and Numeral systems, www.britannica.com/eb/article-233818 retrieved on February 2, 2006.

[24] Encyclopedia Britannica, Transportation, History of, www.britannica.com/eb/article?eu=120012. retrieved on July 7, 2002.

[25] Vannevar Bush, As We May Think, The Atlantic, July, 1945.

[26] Yann Saunders, The Cadillac Database, retrieved from http://www.car-nection.com/yann/ on November 10, 2003.

[27] Chris Caldwell, The Largest Known Prime by Year: A Brief History, retrieved from http://primes.utm.edu on December 23, 2001.

[28] Chris Caldwell, Mersenne Primes: History, Theorems and Lists, http://primes.utm.edu on September 12, 2002.

[29] M. Campbell-Kelly, M. Croarken, R. Flood and E. Robson, editors, The History of Mathematical Tables: From Sumer to Spreadsheets, Oxford University Press, 2003.

[30] Chao-Kuei, Categories of Free Software, Free Software Foundation, 2001. Retrieved from /www.gnu.org/philosophy/categories.html on March 8, 2002.

[31] The Mammogram Screening Controversy: When Should You Start? www.cnn.com, September 27, 1999, retreived from cnn.com on Dec 10, 2006.

[32] M. Collett, T. S. Collett, S. Bisch, and R. Wehner, Local and global vectors in desert ant navigation, Nature 394, pages, 269 - 272, 16 July 1998.

[33] E. F. Codd, A Relational Model of Data for Large Shared Data Banks, Communications of the ACM, Volume 13, No. 6, June 1970, pages 377-387.

[34] Jean-Luc Chabert, editor, A History of Algorithms, Springer-Verlag, Beidelberg, 1999.

[35] Francis S. Collins, Michael Morgan, Aristides Patrinos, The Human Genome Project: Lessons from Large-Scale Biology, Science, Volume 300, page 286-290, 2003.

[36] C. J. Crasto, N. Liu and G. M. Shepherd, Databases for the Functional Analyses of Olfactory Receptors, Neuroscience Database: A Practical Guide, Rolf Kotter, editor, Kluwer Academic Publishers, Dusseldorf, 2003, pages 37–50.

[37] C. J. Date, An Introduction to Database Systems, Addison Wesley Longman, 7th edition, 1999.

[38] William H. Davidow, Marketing High Technology, Free Press, New York, 1986.

[39] Chris DiBona, Sam Ockman, and Mark Stone, Open-Sources, Voices from the Open Source Revolution, O'Reilly, Cambridge, 1999.

[40] Dr. Anthony, Logarithms: History and Use, retrieved from www.mathforum.org on November 1, 2004.

[41] Stephen Donadio, The New York Public Library: Book of Twentieth-Century American Quotations, New York, Stonesong Press, 1992.

[42] Jack Dongarra and Francis Sullivan, Guest Editors' Introduction: The Top 10 Algorithms, Computing in Science and Engineering, Volume 2, Number 1, IEEE Press, 2000.

[43] Joann G. Elmore, Mary B. Barton, Victoria M. Moceri, Sarah Polk, Philip J. Arena, and Suzanne W. Fletcher, Ten-Year Risk of False Positive Screening Mammograms and Clinical Breast Examinations, New England Journal of Medicine, Volume 338, pages, 1089-1096, 1998.

[44] Euclid, The Thirteen Books of the Elements, Translated with introduction and commentary by Sir Thomas L. Heath, Second Edition Unabridged, Dover Publications, Inc., New York, 1956.

[45] Christopher Farrell, A Better Way to Size Up Your Nest Egg: Monte Carlo models simulate all kinds of scenarios, Business Week, January 22, 2001.

[46] Mark Doms, FRBSF Economic Letter 2005-37, The Diffusion of Personal Computers across the U.S., Federal Reserve Bank of San Francisco, December 23, 2005.

[47] Ian Foster and Carl Kesselman, The Grid: Blueprint for a New Computing Infrastructure, Morgan Kaufmann Publishers, Inc., San Francisco, California, 1999.

[48] Bill Gates, The Road Ahead, Penguin Books, 1995.

[49] Categories of Free and Non-Free Software, Free Software Foundation, 2001, retrieved from http://www.gnu.org/philosophy/categories.html on March 8, 2002.

[50] George Gilder, Metcalf's Law and Legacy, Forbes ASAP, September 13, 1993. Also, George Gilder, Telecosm, Simon and Schuster, 1996. Also, George Gilder, Telecosm: How Infinite Bandwidth Will Revolutionize Our World, Free Press, 2000.

[51] Gerd Gigerenzer, Calculated Risks, Simon and Schuster, New York, 2002.

[52] David Goldberg, What Every Computer Scientist Should Know about Floating-Point Arithmetic, ACM Computing Surveys, Volume 23, Number 1, 1991, pages 5-48.

[53] Lee Gomes, Will All of Us Get Our 15 Minutes ON a YouTube Video, Wall Street Journal, August 30, 2006.

[54] Jess M. Gonzlez-Barahona, Miguel A. Ortuo Prez, Pedro de las Heras Quirs, Jos Centeno Gonzlez, Vicente Matelln Olivera, Counting potatoes: The size of Debian 2.2, http://people.debian.org/ jgb/debian-counting/counting-potatoes

[55] Jim Gray, Data Management: Past, Present, and Future, retrieved from research.microsoft.com/ gray on December 20, 2001.

[56] Eric Lee Green, Commoditizing Computers, retrieved from www.badtux.org on December 20, 2002.

[57] Brian Hayes, Terabyte Territory, American Scientist, Volume 90, Number 3, 2002, pages 212-216.

[58] Bruce D. Henderson, The Experience Curve — Reviewed, Boston Consulting Group, Inc., 1973.

[59] John L. Hennessy and David A. Patterson, Computer Architecture: A Quantitative Approach, second edition, Morgan Kaufmann Publishers, Inc., San Francisco, California, 1996.

[60] David G. Hicks, The Museum of HP Calculators, www.hpmuseum.org. Retrieved on March 10, 2003.

[61] P. Hoffman, The Man Who Loved Only Numbers: The Story of Paul Erdos and the Search for Mathematical Truth, New York, Hyperion, 1998.

[62] Internet Assigned Numbers Authority (IANA), IPv6 Address Allocation and Assignment Policy, retrieved from http://www.iana.org/ipaddress/ipv6-allocation-policy-26jun02 on December 26, 2003.

[63] Internet Engineering Task Force (IETF), IP Version 6 Working Group (ipv6), retrieved from www.ietf.org on December 10, 2003.

[64] Internet Engineering Task Force (IETF), RFC 2460, IPv6 Specification, retrieved from www.ietf.org on December 10, 2003.

[65] Internet Software Consortium, http://www.isc.org/.

[66] International Technology Roadmap of Semiconductors (ITRS), retrieved from http://public.itrs.net on November 23, 2003.

[67] Cedric Jagger, The World's Great Clocks and Watches, Hamlyn, London, 1997.

[68] James Jespersen and Jane Fitz-Randolph, From Sundials to Atomic Clocks, Dover Publications, Mineola, New York, 1999.

[69] William Kahan, Computing Days Between Dates, the Day of the Week, etc. Retrieved from www.cs.berkeley.edu/ wkahan/daydate/daydate.txt on September 10, 2002.

[70] Michelle Kessler, High tech's latest bright idea: Shared computing, USA TODAY, January 8, 2003.

[71] Rachael King, How Cloud Computing Is Changing the World, Business Week, August 4, 2008.

[72] Ray Kurzweil, The Age of Spiritual Machines, Penguin Books, New York, New York, 1999.

[73] J. L. Heilbron, The Sun in the Church, Harvard University Press, Cambridge, Massachusetts, 1999.

[74] Alan Hall, How the Web Was Wove, Business Week, October 5, 2000.

[75] Projects Prove Innovation, InfoWorld, November 20, 2001.

[76] Internet Software Consortium, Domain Survey, retrieved from www.isc.org on August 10, 2002.

[77] Barry M. Leiner, Vinton G. Cerf, David D. Clark, Robert E. Kahn, Leonard Kleinrock, Daniel C. Lynch, Jon Postel, Larry G. Roberts, and Stephen Wolff, A Brief History of the Internet, Internet Society, 2000, retrieved from www.isoc.org/internet/history/index.shtml on January 4, 2002.

[78] Michael Lesk, How Much Information is there in the World, retreived from http://www.lesk.com/mlesk/ksg97/ksg.html on December 20, 2001.

[79] Lawrence Lessig, The Future of Ideas: The Fate of the Commons in a Connected World, Random House, New York, 2001.

[80] Steve Lohr and John Markoff Windows Is So Slow, but Why?, New York Times, March 27, 2007.

[81] Peter Lyman and Hal R. Varian, How Much Information?, retrieved from www.sims.berkeley on June 30, 2001.

[82] Peter Lyman and Hal R. Varian, How Much Information? 2003, retrieved from www.sims.berkeley on November 20, 2006.

[83] David L. Margulius, When PC Still Means 'Punched Card', New York Times, February 7, 2001.

[84] Peter Mell and Tim Grance, The NIST Definition of Cloud Computing, NIST Special Publication 800-145, 2011.

[85] Scott McCartney, ENIAC: The Triumphs and Tragedies of the World's First Computer, Walker Publishing Company, 1999.

[86] Josh McHugh, For the lover of hacking, Forbes Magazine, August 10, 1998.

[87] Donella H. Meadows, Dennis I. Meadows, Jorgen Randers, and William W. Behrens III, The Limits to Growth, Club of Rome, 1972.

[88] Gregory John Michaelson, Undusting Napier's Bones, retreived from http://www.macs.hw.ac.uk/ greg/calculators/napier/ on August 10, 2003.

[89] MSN Hotmail Continues to Grow Faster Than Any Media Company in History, Microsoft Press Release, February 8, 1999. Retrieved from www.microsoft.com on July 10, 2002.

[90] Microsoft .NET, retrieved from www.microsoft.com on February 10, 2002.

[91] Microsoft Corporation Annual Report, 2006.

[92] Geoffrey A. Moore, Crossing the Chasm, Harper-Collins Publishers, New York, 1991.

[93] Gordon E. Moore, Cramming more components onto integrated circuits, Electronics, Volume 38, Number 8, April 19, 1965.

[94] Moore's Law, An Intel Perspective, video transcript, retrieved from www.intel.com on June 11, 2009.

[95] NASA's Earth Observing System, retrieved from www.nasa.gov on March 10, 2002

[96] NOAA/NASA AVHRR Oceans Pathfinder Program, retrieved from www.nasa.gov on March 10, 2002.

[97] National Center for Biotechnology Information, GenBank, www.ncbi.nlm.nih.gov/Genbank, 2008.

[98] National Research Council, An Assessment of Space Shuttle Flight Software Development Processes, National Academy Press, 1993.

[99] National Research Council, Funding a Revolution: Government Support for Computing Research, National Academy Press, 1999.

[100] National Research Council, Embedded, Everywhere, National Academy Press, Washington, D.C., 2001.

[101] Committee on Surface Temperature Reconstructions for the Last 2,000 Years, Surface Temperature Reconstructions for the Last 2,000 Years, National Research Council, The National Academies Press, Washington, DC, 2006.

[102] National Academy of Engineering, Greatest Engineering Achievements of the 20th Century, 2000.

[103] US National Science Foundation, Division of Science Resources Studies, Survey of Earned Doctorates, 2004, retrieved from NSF on June 10, 2006.

[104] Netcraft Web Server Survey, retrieved from www.netcraft.com on May 10, 2007.

[105] Nike - Our history, retrieved from www.nike.com on June 10, 2002. Nike - Our chronology, retrieved www.nike.com on June 10, 2002.

[106] Who Needs High-Accuracy Timekeeping and Why?, NIST Press Release, December 29, 1999, retrieved from www.nist.gov on September 20, 2006.

[107] National Institute of Standards and Technology, A Walk Through Time, retrieved from www.nist.gov on July 25, 2002.

[108] Nikon Plans to Stop Making Most Cameras That Use Film, New York Times, January 12, 2006.

[109] Nokia, Quarterly and Annual Information, retrieved from www.nokia.com/2008/Q1/index.html on June 10, 2008.

[110] Human Genome Project Information, retrieved from www.ornl.gov/hgmis/project/hgp.html on January 3, 2002.

[111] Human Genome Product Information, How Many Genes Are in the Human Genome?, retrieved from www.ornl.gov on June 10, 2006.

[112] John J O'Connor and Edmund F Robertson, The MacTutor History of Mathematics Archive, retrieved from www-history.mcs.st-andrews.ac.uk on March 6, 2006.

[113] John J O'Connor and Edmund F Robertson, Greek Number Systems, in [112].

[114] John J. O'Conner and E. F. Robertson, The MacTutor History of Mathematics Archive, retrieved from www-gap.dcs.st-and.ac.uk/ history/ on June 20, 2003.

[115] John J. O'Conner and E. F. Robertson, Mathematics and Architecture, in [112].

[116] Andrew M. Odlyzko, The history of communications and its implications for the Internet, June, 2000, retrieved from http://www.research.att.com/ amo/doc/networks.html on July 1, 2001.

[117] Andrew Odlyzko and Benjamin Tilly, A refutation of Metcalfe's Law and a better estimate for the value of networks and network interconnections, retreived from www.umn.edu on June 10, 2006.

[118] Oxford English Dictionary, Oxford University Press, Second Edition, 1989.

[119] Tim O'Reilly, The Open-Source Revolution, Release 1.0, November, 1998.

[120] Wayne Pafko, Visualizing Tycho Brahe's Mars Data, retrieved from www.pafko.com on June 20, 2003.

[121] Percedes Pascual, Xavier Rod, Stephen P. Ellner, Rita Colwell, Menno J. Bouma, Cholera Dynamics and El Nio-Southern Oscillation, Science, Volume 289, 2000, pages 1766–1769.

[122] Byron E. Phelps, Early Electronic Computer Developments at IBM IEEE Annals of the History of Computing, Volume 2, Number 3, July, 1980.

[123] The Party-line Approach to Supercomputing, Developments to Watch, BusinessWeek, December 5, 1994.

[124] Theodore M. Porter, Karl Pearson: The Scientific Life in a Statistical Age, Princeton University Press, 2004.

[125] Jonathan B. Postel, Darpa Internet Program Protocol Specification RFC792, September, 1981, Retrieved from http://www.rfc-editor.org on December 10, 2003.

[126] Jonathan B. Postel, Simple Mail Transfer Protocol, August 1982, RFC 821, retrieved from www.rfc-editor.org on February 2, 2002.

[127] Price Waterhouse Coopers, PWC Internet Survey, retrieved from www.pwcinternet.com on September 8, 2000.

[128] Emerson W. Pugh, Lyle R. Johnson and John H. Palmer, IBM's 360 and Early 370 Systems, MIT Press, Cambridge, 1991.

[129] John Harrison and the Longitude Problem, Royal Observatory Greenwich Online Exhibit, www.rog.nmm.ac.uk/museum/harrison, retrieved on July 5, 2002.

[130] Jean E. Sammet, Brief Summary of the Early History of COBOL, IEEE Annals of the History of Computing, Volume 7, Number 4, 1985.

[131] Carl Shapiro and Hal R. Varian, Information Rules, Harvard Business School Press, Boston, Massachusetts, 1999.

[132] Avi Silberschatz, Michael Stonebraker, and Jeff Ullman, editors, Database Systems: Achievements and Opportunities, Communications of the ACM, Volume 34, Number 10, pages 110–120, 1991.

[133] Richard L. Smith, A Statistical Assessment of Buchanan's Vote in Palm Beach County, retrieved from www.stat.unc.edu on June 10, 2006.

[134] Michael Specter, Do Fingerprints Lie?, New Yorker, May 27, 2002.

[135] Paul McJones, editor, The 1995 SQL Reunion: People, Projects, and Politics, retreieved from www.mcjones.org on January 24, 2002.

[136] Richard Stallman, The GNU Manifesto, Free Software Foundation, 1985 and 1993, retrieved from www.gnu.org on March 8, 2002.

[137] Sun Microsystem Corporate History, retrieved from www.sun.com on December 31, 2001.

[138] Sun Microsystems, No-cost Software FAQs, retrieved from www.sun.com on January 10, 2007.

[139] The Top 500 Supercomputer Sites, retrieved from www.top500.org on Oct 10, 2008.

[140] Linus Torvalds, Linux History, July 31, 1992, retrieved from www.li.org on October 10, 2002.

[141] Michael Turner, The Fair Credit Reporting Act: Access, Efficiency & Opportunity; The Economic Importance of Fair Credit Reauthorization, Information Policy Institute, 2003.

[142] U.S. Census Bureau, World POPClock Projection, retrieved from http://www.census.gov.

[143] U.S. Department of Commerce, North American Industry Classification System (NAICS), 2007.

[144] US Federal Trade Commission, Consumer Sentinel, retrieved from www.consumer.gov on September 1, 2002, September 10, 2006 and March 10, 2007.

[145] US Federal Trade Commission, Prepared Statement of the Federal Trade Commission on the Fair Credit Reporting Act Before the Senate Committee on Banking, Housing, and Urban Affairs Washington, D.C. July 10, 2003.

[146] U.S. Navel Observatory, The Date of Easter, retrieved from www.navy.mil on December 10, 2006.

[147] Hal R. Varian, How Much Does Information Technology Matter?, New York Times, May 6, 2004.

[148] Lisa Vaas, Study: 2004 Database Market Grew 12 Percent, retrieved from www.eweek.com on March 7, 2005.

[149] Cento G Veljanovski, Competition Law Issues In the Computer Industry: An Economic Perspective, QUT Law and Justice Journal, Volume 3, Number 1, 2003, pages 3–27.

[150] Virgil, The Aeneid, translated by G. R. Wilson Knight, Penguin Books, London, 1956.

[151] Visa, About Visa USA, retrieved from http://www.visa.com on August 29, 2003.

[152] Paul Vixie, Software Engineering, in DiBona, 1999, op. cit., pages 91–100.

[153] G. H. Von Wright, editor, Culture and Value, translated by Peter Winch, The University of Chicago Press, 1980.

[154] Jean Walrand and Pravin Varaiya, High Performance Communication Networks, Second Edition, Morgan Kaufmann, San Francisco, California, 2000.

[155] David A. Wheeler, The Most Important Software Innovations, May, 2001. Retrieved from http://www.dwheeler.com/ on July 7, 2001.

[156] David A. Wheeler, Estimating Linux's Size, July 26, 2001, Version 1.05, Retrieved from http://www.dwheeler.com/ on August 10, 2002.

[157] David A. Wheeler, More Than a Gigabuck, Estimating GNU/Linux's Size, July 2002. Retrieved from http://www.dwheeler.com/ on August 10, 2002

[158] Michael R. Williams, A History of Computing Technology, Prentice-Hall, 1985.

[159] Roy Williams, Data Powers of Ten, retrieved from www.davedoyle.com on May 20, 2006.

[160] Richard Saul Wurman, Information Anxiety, Bantam Books, New York, 1990. See also, Richard Saul Wurman, Information Anxiety 2, QUE, Indianapolis, 2001.

Index

air cushioning, 132
algorithms, 15
alphabet, 14, 205
Andreesen, Marc, 90
ASCII, 204
atomic clocks, 45
average selling price, 123

Bermuda Principles, 75, 189
Berners-Lee, Tim, 90
big data, 163
Bill's Law, 69, 127
Bowerman, Bill, 131
Brahe, Tycho, 179
breakthrough, 138
breast cancer, 213
business intelligence, 88

calendars, 47
chasm, 125
Cisco, 36
cloud computing, 139
clusters, 137
clutter, *see* market clutter
COCOMO cost model, 70
Codd, E. F., 107, 147
coding
    odors, 206
    text, 204
commoditization, 3, 53
    CPUs, 52
    of data, 43
    space, 50
    storage, 64
    time, 45
competition, 93
computer
    hollowed out, 35
    tapes, 104
computing
    cloud, 139
    clusters, 137
    context, 140
    data era, 161
    epochs, 10
    era, *see* era, 24
    grid, 139
    integration, 142
    lock-in, 140
    platforms, 38
    roadmap, 134
    standards, 140, 144
concordance, 204
consultants, 97
credit card fraud, 209
credit scores, 199
cycle stealing, 139

Danti's Law, 47
Darwin, Charles, 169

data
- commoditization, 166
- era of big data, 5
- event streams, 196
- gap, 168
- graph, 197
- how much, 166
- open, 189
- semi-structured, 194
- sequence, 194
- shape of, 193
- size of, 162
- streaming, 216, 219
- tabular, 182

data mining, 87

data transport
- by horse, 169

database, 145
- adoption, 145
- distributed, 108
- keys, 107
- navigational, 147
- relational, 107
- schema, 106
- System R, 148
- tables, 146
- transaction, 106

denial of service attack, 217

devices, 8
- computing, 23

doubling criterion, 58

doubling game, 58

early adopters, 123

El Nino, 183

email, 31

epicycles, 178

era
- clouds of devices, 34
- computing, 3, 9
- data, 43
- mainframe, 26
- pc, 28
- web, 30

Euclid, 109

false negative, 213

false positive, 211, 213

FCRA, 199

features, 92

fingerprints, 211

Galileo, 175

Gates, Bill, 78, 127

GenBank, 75, 189

geometric growth, 58

Gilder's Law, 66

Gilder, George, 66, 77

GIMPS, 111

GNU Manifesto, 72

GPS, 49

graph, 197

grids, 139

Harrrison, John, 51

host ID, 7

Human Genome Project, 75, 176, 189, 196

images
- digital vs analog, 168

information anxiety, 88

innovation, 3
- changing perception, 89
- rarity of, 82

IP address, 5, 116

IPv6, 6

Iterative methods, 85

Johnson's Law, 62
Joy, Bill, 127

Kepler,Johannas, 179
KKR, 30

laggards, 123
law
  Bill's, 69
  Gilder's, 66, 79
  Johnson's, 62
  Linus's, 79
  Metcalfe's, 77
  Moore's, 52
  Pearson's, 75
  Stallman's, 67
Lessig, Lawrence, 188
Linux, 156
lock-in, 140
longitude, 51

Main Street Market, 123, 127
market
  clutter, 4, 82, 96, 98
  forces, 121
  leader, 94
  submarkets, 92
  vertical, 93
marketing
  basic equation, 123
mathematical tables, 20
meridian line, 47
Metcalfe's Law, 66, 77
Metcalfe, Robert, 77
Model T, 144
Monte Carlo algorithms, 176
Moore's Law, 52
Mosaic, 136
mythical man-month, 158

network
  effects, 77, 79
  ID, 7
  latency, 57
  local area, 36
  packets, 217
  telephone, 111
Newton, Isaac, 83, 180
Newton-Raphson Method, 85
Nike, 130
number
  as limit, 87
  as solutions to equations, 84
  prime, 109
  rational, 174
  viewed as lengths, 83
  viewed geometrically, 13

Occam's razor, 179
open source software, 70, 159

Page Rank, 198
Pearson's Law, 182
PMML, 194
primes, 109
  Mersenne, 110
profiles, 216
Ptolemaic convenience, 179
punched cards, 27, 104
Pythagoras, 83, 174

real time, 216
relational database, *see* database
research
  applied, 135

basic, 134
resources
controlled vs free, 188
roadmap, 134
router, 36, 114
routing tables, 118
ruler and compass constructions, 84

SAD, 10, 22
satellites, 219
schema, 106
science
data, 173
experimental, 175
simulation, 176
sensors, 222
Sieve of Eratosthenes, 110
slide rule, 10, 19
SMTP, 32
Stallman's Law, 73
Stallman, Richard, 67
standards, 140, 144
statistical model, 209
errors, 210
Sun Microsystems, 128
symbols, 10, 11

technology
adoption cycle, 4, 40
context, 120
roadmap, 120, 134
telephones, 113
Torvalds, Linus, 156
transforming technologies, 61

VisiCalc, 29
von Neumann, John, 91

waffles, 132
winners, technology, 91

About the Author

Robert L. Grossman is a faculty member at the University of Chicago and a Partner of Open Data Group. At the University of Chicago, he is the Director of Informatics at the Institute for Genomics and Systems Biology, a Senior Fellow at the Computation Institute, and a Professor in the Division of Biological Sciences. He founded Open Data Group in 2002, and since then it has been one of the leaders in building predictive models over big data.

Cover design by Rachel Pasch. Cover image by Michal Sabala.