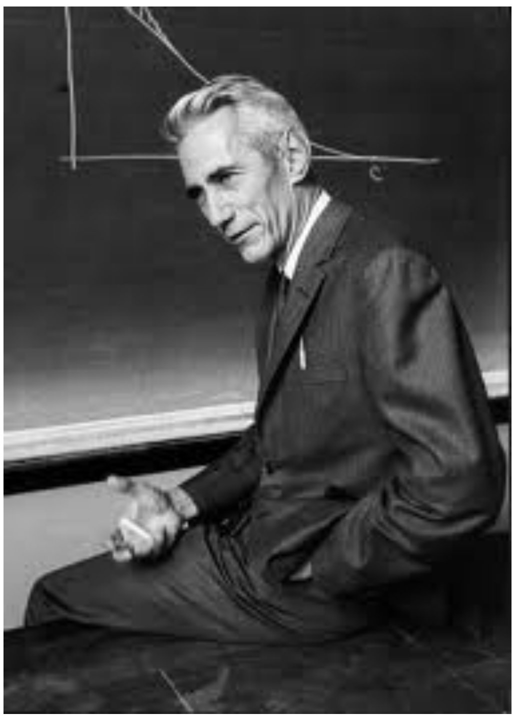
CSC105, Introduction to Computer Science
Lecture: Information, Data, and Digital Data

What is information?

Claude Shannon

"A Mathematical Theory of Communication.," (1949)



Claude Shannon (1916–2001).

- his master's thesis (1937) laid the groundwork for **digital circuit design** that was the foundation for modern electronic digital computers.
- in 1949, he published a work that transformed **cryptography** from an art to a science.
- with mathematician Ed Thorp, Shannon developed applications of **game theory** which they employed on numerous occasions to win big in gambling games in Las Vegas. Later these principles were applied even more successfully for stock market investments. The principle they devised known as "Kelly's criterion" was also employed by others including Warren Buffet.

In Shannon's view, when we express information—for example, in language—the *extent or amount of that information is a function of how ordered or organized is that signal*.

In order for the system to be *effective* (successful at communicating):

a. the most common (least informative) ideas should be expressed as concisely as possible.

   In English, for example, words like "a", "an" and "the" are the most common and shortest.

   On the other hand, less common (more informative) ideas are expressed less concisely.

   Examples: "earthquake", "turnpike", and "chemical."

b. the intelligibility of messages should be preserved even when part of the signal is lost or corrupted by noise.

In this context, information according to Shannon can be characterized as

- *what is known beyond chance prediction*.

Thus, information decreases uncertainty and aids decision-making.

A chief corollary of Shannon's analysis is that the expression of information can be understood as having three stages.

information → message → signal.

a. In order to convey **information**, I must first capture it in the form of a **message**.

b. Ordinarily, the **message** cannot be transmitted unless I convert it to some form of physical **signal**.

c. the received **signal** is then converted into its corresponding **message**, which is then interpreted to extract whatever **information** it contains.

Fundamental components of any communication system:

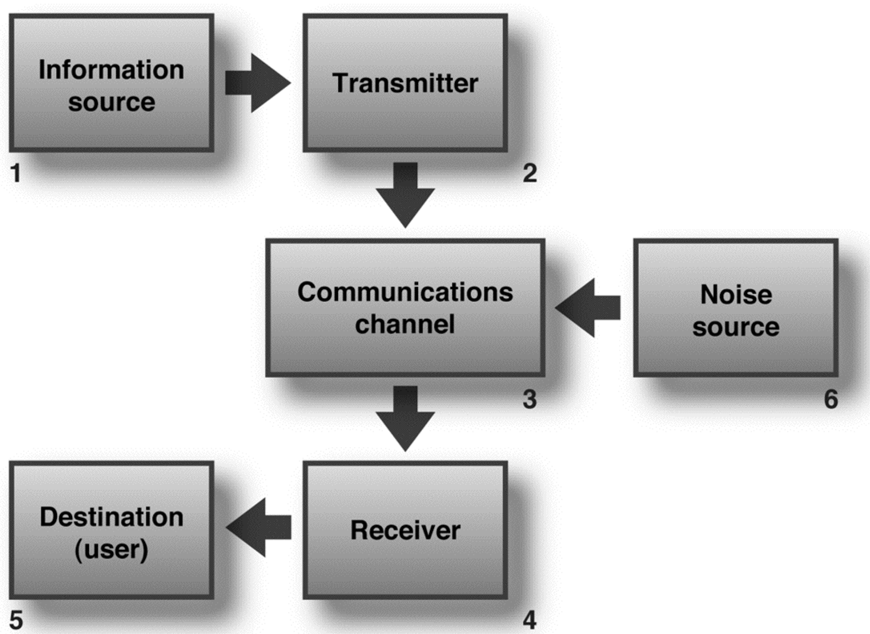


Figure 1: Shannon's model for a communication system. *NOTE*: The distance of the channel (not depicted here) is another practical factor for an effective communication system.

Information Technologies

The modern electronic digital computer system represents the only most recent form of what are dubbed **information technologies**. There has been a long history of various information technologies—many of which still exist today.

- a **technology** is an artificial instrument, process, or system that extends human capabilities.

Technologies are **artificial** (man-made), but they are also natural in the sense that they are **extensions of natural or customary ways in which we perform some task**.

- connect or extend what is natural for us
- **examples:** the *wheel* is an extension of the foot. It extends the physical range that we are able to transport ourselves and other good.
- Inventions such as *incandescent light* and the *microscope* are extensions of the eye. They extend the power of vision into the worlds of darkness and the very small.

Information technologies extend our natural abilities for gathering, storing, managing, and distributing information.

- *writing* is an artificial form of storing and transmitting speech.
- the *movable type printing press* extended the efficiency and economy of the written word.
- *telegraph*, *telephone*, *radio* and *television* are other obvious information technologies.

Data and Data Representation.

Information is an abstract concept; its concrete or physical complement is that of data.

- **data** is the symbolic representation of information.

Just as written language is based on physical symbols (letters, numerals, punctuation, etc.), we may represent data using other symbol systems too.

Modern computer systems are **digital** devices. This means that computers employ **digital data** representations.

Digital Data

The term "digital" refers to **numbers**. Consequently, digital encoding is based on a numeric symbol system.

Digital data possesses two important characteristics.

- **discrete**. Each symbol and token is distinct or separate, not to be confused with others.

  e.g., 2 ≠ 3.

- **precise**. Each token has a meaning or significance that is understood intersubjectively (objectively).

  e.g., John is tall vs. John's height is 6 foot, 4 inches.

We can think of the flow of information into the digital domain as having three stages:

INFORMATION	→	**DATA**	→	**DIGITAL DATA**
understood by humans		*a physical representation*		*encoded using a finite numeric representation*
thoughts, ideas, concepts, etc.		speech, writing, video signals, etc.		bit, bytes, etc.

Analog forms of data or **analog representations of information**.

- **analog data** is represented **continuously** as variations (of values) over time and/or space.
- e.g., sound, air pressure, light, electrical signals, etc. have information that is commonly represented using analog forms. Consider this example of a sound wave.

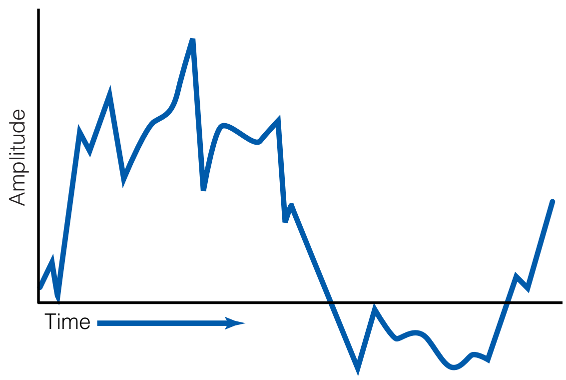


Figure 2: An analog signal represented as amplitude over time.

Digital forms of data

The process of digitizing data has two stages: **sampling** and **quantization**.

Sampling. First, we must render it as **discrete data.** This is the process of **sampling**. For example, our sound wave could be sampled at a series of intervals over time. The result would look something like this.

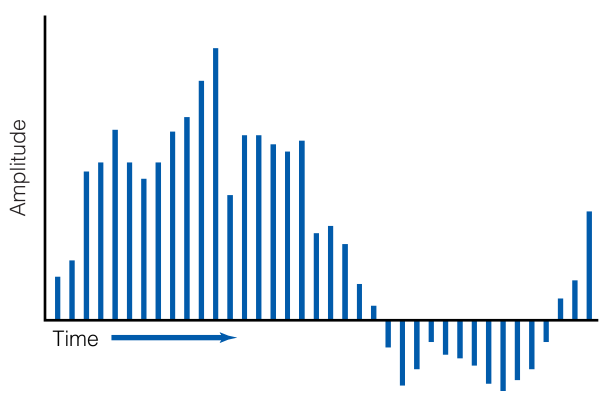


Figure 3: The previous signal is sampled producing pulse codes.

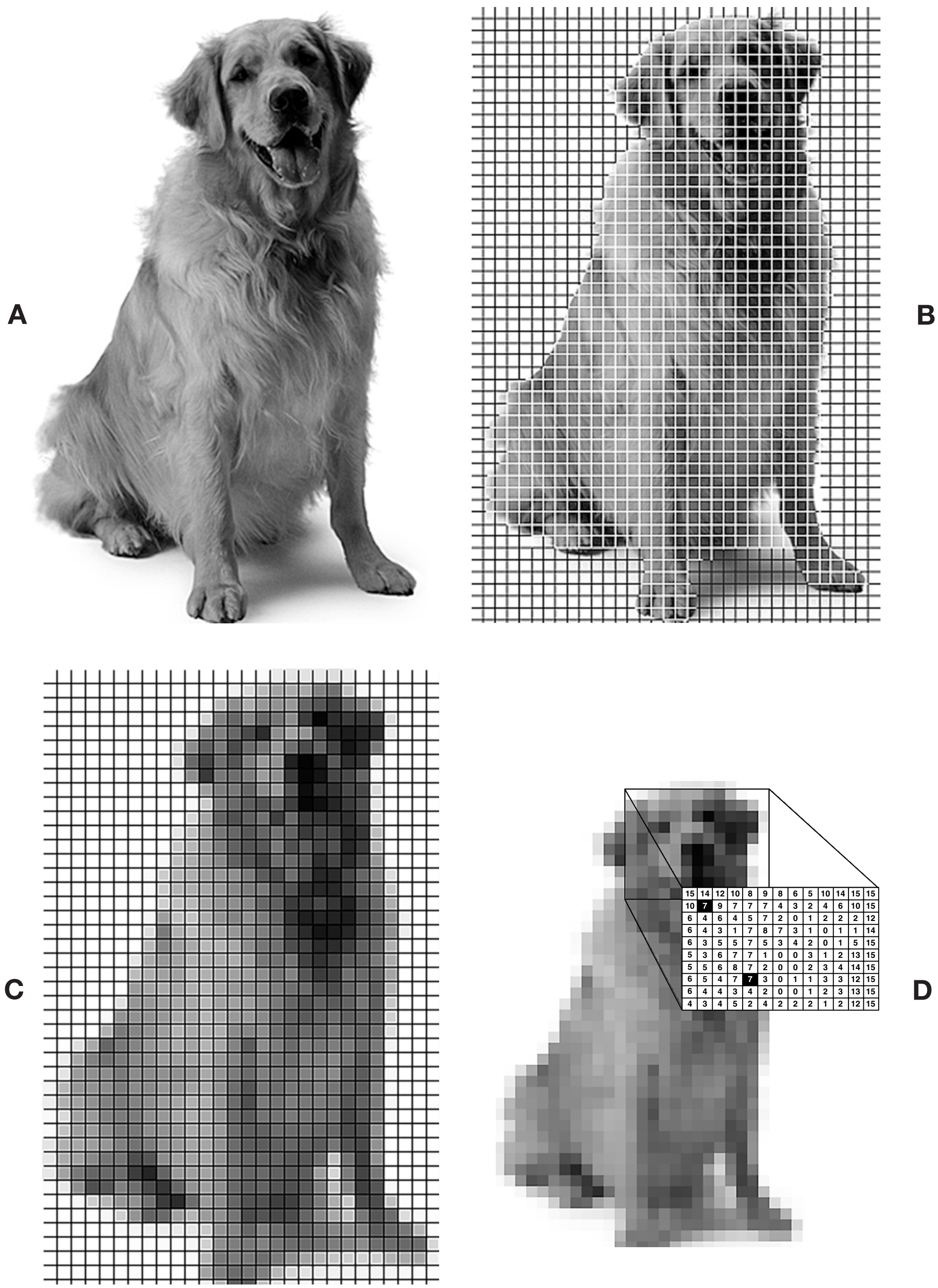
These spikes depict what is called a **pulse code**. Each instant measures the amplitude of the signal independently of other moments.

But, the question is: “Are these losses significant?”

In many cases, the answer is “No.” If we are careful in the process of sampling, we can capture data accurately enough that we can recreate the analog representation faithfully.

- analog ≈ digital

Quantizing. The final step for converting analog representations of information into digital form is called **quantizing.** Simply put, it means to substitute the quantity sampled by a discrete number.

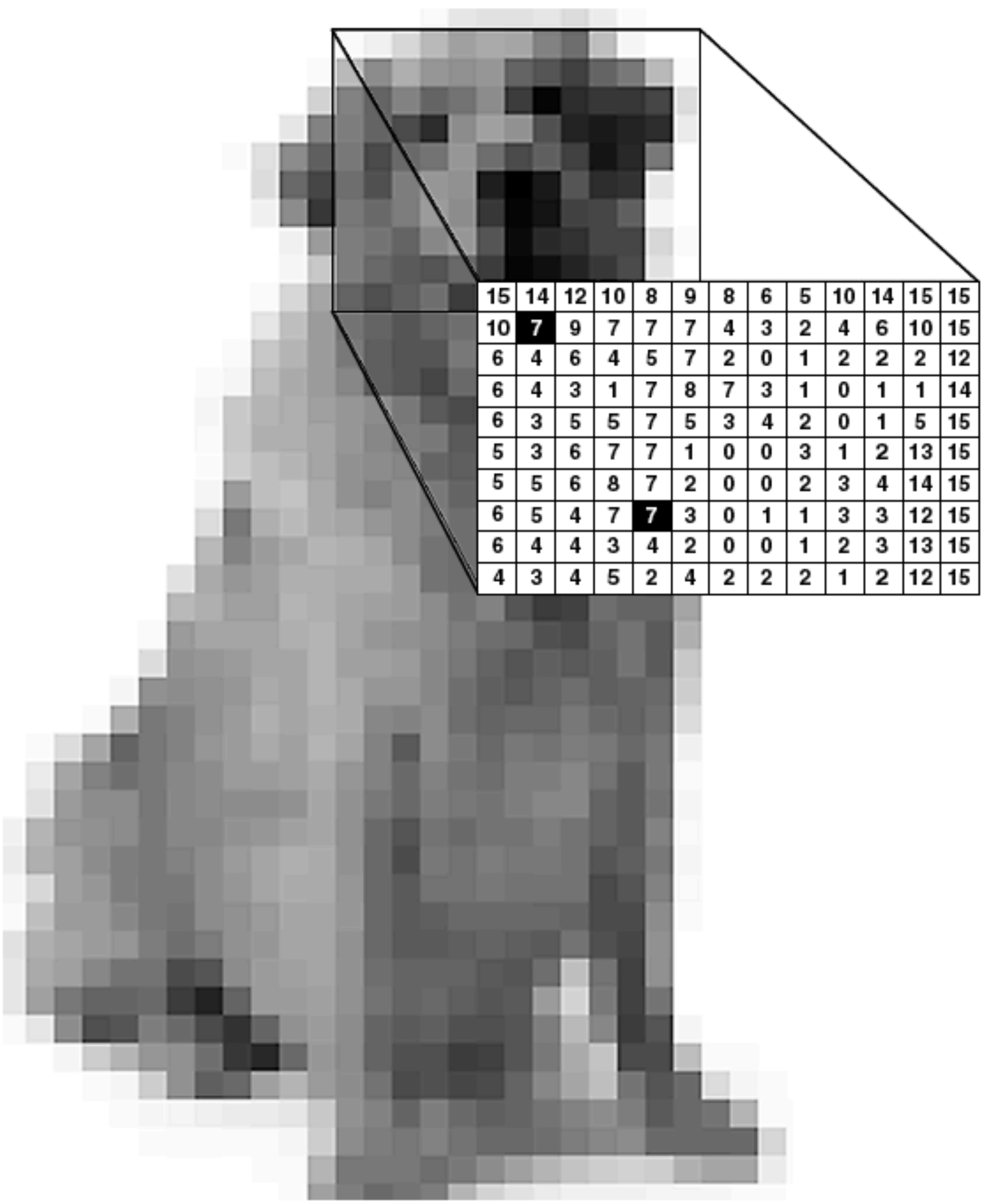


A. Here is our original black-and-white photo.

B. The image will be sampled spatially, so a grid representing the sampled areas is plotted. Each box represents a picture element or pixel, for short.

C. The resulting samples must be uniform (for precision), so "smoothing" or averaging is performed.

D. Finally, each pixel is encoded using a number that represents it relative brightness. The scale, here, is 0–15, or 16 shades of brightness from dark to light.



As the example illustrates, the process of digitizing is not always perfect. We have basically two general sources for potential error.

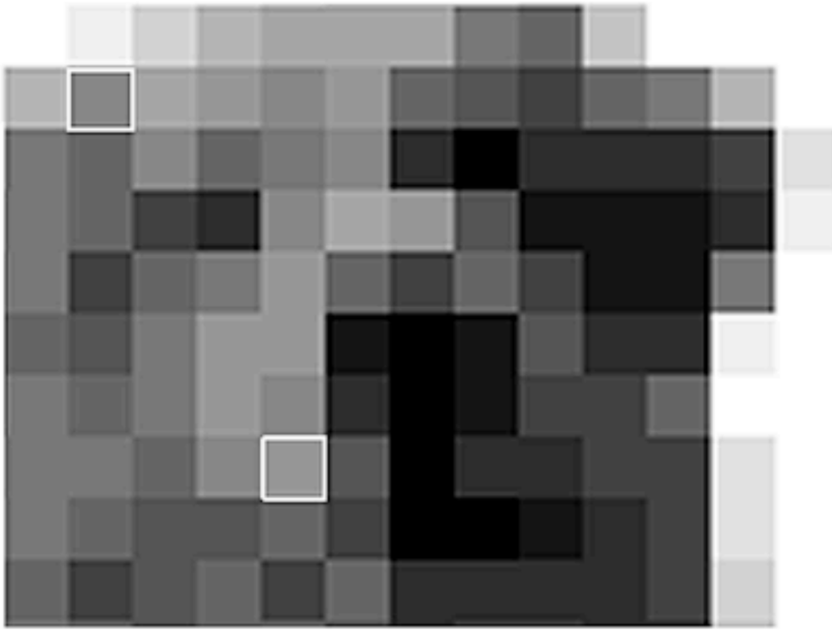
- **undersampling.** Too few samples make for inaccuracies or poor resolution. Thus, increasing the number of samples can improve the results.
- **quantizing errors**. If the range of values that can be assigned to the samples is small, then poor dynamic range can result. The results can be more sensitive when a greater range is chosen.

The Advantages of the Digital Domain.

If analog and digital forms of information are roughly interchangeable (in most instances), then the big question is "Why should we spend the time and money necessary to digitize data that is already represented by analog forms?"

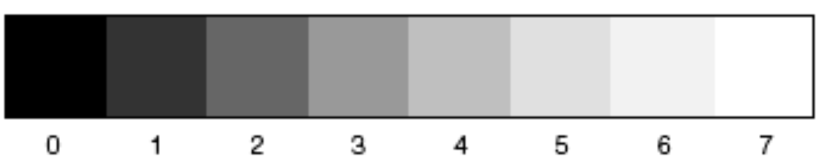
Here is the basic list for why we should.

Advantage	Source of advantage
precision	digital coding
ordinality	digital coding
more efficient storage	electronic/optical technology
faster transfer	electronic/optical technology
absolute replication	digital coding
random and selective access capabilities	digital coding, computer processing
compression	digital coding, computer processing
integrative capabilities	*common* digital coding.
content analysis and synthesis capabilities	digital coding, computer processing



15	14	12	10	8	9	8	6	5	10	14	15	15
10	**7**	9	7	7	7	4	3	2	4	6	10	15
6	4	6	4	5	7	2	0	1	2	2	2	12
6	4	3	1	7	8	7	3	1	0	1	1	14
6	3	5	5	7	5	3	4	2	0	1	5	15
5	3	6	7	7	1	0	0	3	1	2	13	15
5	5	6	8	7	2	0	0	2	3	4	14	15
6	5	4	7	**7**	3	0	1	1	3	3	12	15
6	4	4	3	4	2	0	0	1	2	3	13	15
4	3	4	5	2	4	2	2	2	1	2	12	15

Digital **precision** makes it easier to compare items that otherwise may be more difficult to discern



Digital **ordinality** makes it easier to do relative comparisons.

The Impact of Going Digital.

As with any dominant information technology, there are often changes in our lives—some significant, some subtle. Digital information technology is no different. Our reading points out several important trends. Some are desirable, some are not so much. Let's review some of them here.

- the power of digital technology is that bits are just bits. When stored on a computer, tablet, or smartphone, audio, text, and pictures are stored as bits. When transmitted across a network, the routers that push the data are pushing only bits. One packet of data is like any other packet. Only the devices that we use convert it to the things that we want.
- digitizing information makes capturing and storing easier and cheaper. This, in turn, has led to an exponential growth in the amount of data collected and stored. It has been claimed that the total amount of human knowledge acquired doubles every two years.
- as mentioned earlier, digital copies are perfect. So, if we can find storage technologies that last, it is possible to store our collected information in perpetuity.
- the ease of producing and transmitting digital information has toppled the old regime of intellectual property. But, the new regime is not exactly what most had expected.
- processing power of digital devices has grown at incredible rates over the past five decades. This is captured as Moore's Law. We will return to this idea again later when we consider how computers compute.

While technology *per se* is neither good nor bad, we can certainly say that the uses and applications of technology are not so neutral. Digital technology has advanced many values, but it has also hampered others. In that respect, it is like many of the information technologies that preceded it.

Many have called the ascendance of digital technology as a great watershed moment in our history. Certainly, digital information technology has had a significant impact, but has this impact been worth a new epoch?

Nate Silver in his book *The Signal and the Noise* (2012) points out that the evidence is not conclusive. Until the invention of the printing press, the quality of life for the average human remained largely unchanged, if we measure in terms of the global per capita GDP (gross domestic product: a measure of all of the goods and services produced). From the year 1000 to the invention of the printing press, the GDP remained relatively flat.

After its invention and with the inception of the Industrial Revolution that it helped to create, the average per capital begins a significant rise. The so-called "Information Age" has helped to fuel its continued growth, but not at any faster rate.

But, as Silver mentions, the 300 years of warfare that left millions of dead across Europe and the globe had an impact too: it reduced the growth of the world population.

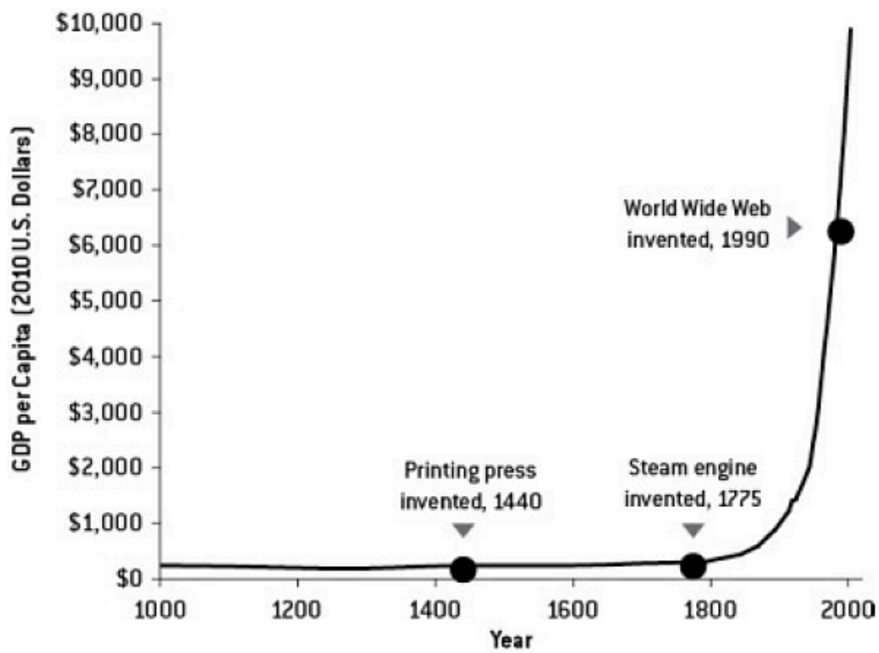


Figure 4: GDP per capita in 2010 U.S. dollars from 1000 – 2010 C.E.

Every age anoints its modernity. Ours is probably just as guilty. Examine the following excerpt.

> **"... a new communications technology was developed that allowed people to communicate almost instantly across great distances, in effect shrinking the world faster and further than ever before. A worldwide communications network [that] spanned continents and oceans, it revolutionized business practice, gave rise to new forms of crime, and inundated its users with a deluge of information. Romances blossomed over the [network]. Secret codes were devised by some users and cracked by others. The benefits of the network were relentlessly hyped by its advocates and dismissed by the skeptics. Governments and regulators tried and failed to control the new medium. Attitudes toward everything from newsgathering to diplomacy had to be completely rethought. Meanwhile, a technological a technological subculture with its own customs and vocabulary was establishing itself."**