3 Foundational Myths

of Data Science

Ryan Shaw
School of Information and Library Science
University of North Carolina at Chapel Hill

Lecture for the School of Management Science and Engineering, CUFE, May 27, 2022

What is data science?

- Most human practices (social, economic, cultural, etc.) that extend over time and space are **mediated by documents** (messages with some persistence)
- With the widespread adoption of **networked computing** and **electronic storage and transmission** of documents:
    1. more kinds of **messages can be made persistent** (turned into documents)
    2. it becomes feasible to **accumulate these documents** on a large scale
- Data science seeks to **understand and improve these practices** by analyzing these large-scale accumulations of documents

Usually when we think of science, we think of…

Data science is not like that. Instead it is a response to technological disturbances

Technological disturbances

Late 19th–early 20th century

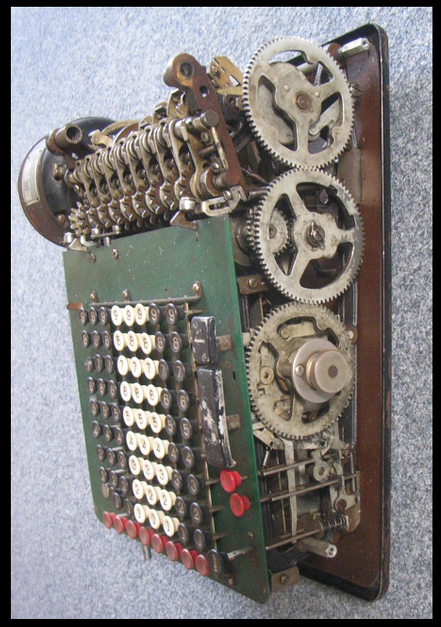
- Cash registers
- Tabulating machines
- Calculating machines

Abbott (1988) The system of professions

late 19th century: merchants hiring stranger to work in their stores
the Ritty brothers of Dayton, Ohio - led to founding of NCR - Edward Deeds / Charles Kettering (invented the credit card) - 1914, Engineers Club

1880 census: six years 1890 census: six weeks, saved $5 million (in 1890 dollars), $140 million today. Later, with people from NCR, formed Computing-Tabulating-Recording Company (CTR), in 1924 became IBM

Burroughs Adding Machine, 1885



Monroe Division Machine, 1912

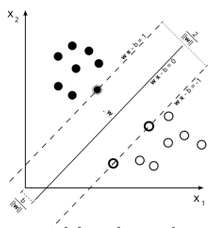
Burroughs now is known as UNISYS, Monroe Systems for Business still around

heir to the Burroughs fortune: William S. Burroughs

create large quantities of accurate data

aggregate data in a way that it can be easily queried

enable rapid and sophisticated analysis



smartphones

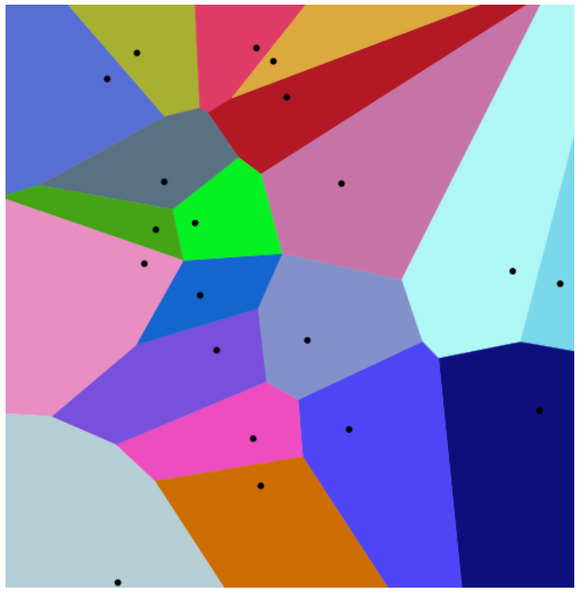
large-scale databases

machine learning

Battles for jurisdiction
among emerging professions

"Each profession has its activities under various kinds of jurisdiction … **Jurisdictional boundaries are perpetually in dispute** … Jurisdictional claims furnish the impetus and the pattern to organizational developments."

Andrew Abbott, *The system of professions*



Abbott (1988) The system of professions

The quantitative information professions

up until the 1960s

Accounting

Statistics

Management engineering

Economics

Marketing

Operations research

Systems analysis

Abbott (1988) The system of professions

"No coherent set of people has in fact emerged to take jurisdiction in this area … It continues to be extremely permeable, with … careers following wildly diverging patterns. There are certain small and relatively elite groups in the area … [but] they have yet to institutionalize coherent training programs and to create secure links of jurisdiction.

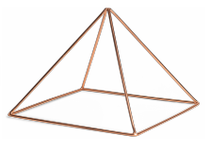
Andrew Abbott, *The system of professions*

Abbott (1988) The system of professions

3 Foundational Myths

of data science

1. The conduit metaphor 

2. The knowledge pyramid 

3. The common substrate 

1. explain the myth 2. provide an alternative account

Myth #1

The conduit metaphor

RESEARCH AND INNOVATION

Researchers use AI to unlock the secrets of ancient texts

August 5, 2021

The Abbey Library of St. Gall in Switzerland is home to approximately 160,000 volumes of literary and historical manuscripts dating back to the eighth century — all of which are written by hand, on parchment, in languages rarely spoken in modern times.

To preserve these historical accounts of humanity, such texts, numbering in the millions, have been kept safely stored away in libraries and monasteries all over the world. A significant portion of these collections are available to the general public through digital imagery, but experts say there is an extraordinary amount of material that has never been read — a treasure trove of insight into the world's history hidden within.

https://news.nd.edu/news/researchers-use-ai-to-unlock-the-secrets-of-ancient-texts/

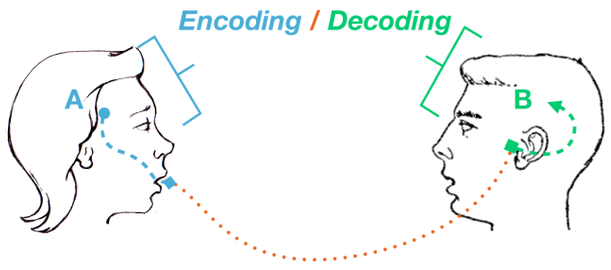
13

- The documents are containers
- Meaningful insights are locked inside these containers
- The way to extract that hidden meaning is to read the documents
- But there are too many documents
- So, the extraction process has been automated

Now consider what is implied by these statements:
Presumably, the insights or secrets referred to in the press release have to do with what the manuscripts mean.
That meaning is somehow “locked inside” the manuscripts.
The way to extract that hidden meaning from the manuscripts is to read them, but there are many manuscripts, and not enough people who can read ancient languages.
To solve that problem, the extraction process has been automated.

The conduit metaphor

Myth #1



"… suggests that communication is a pipeline transfer of units of material called 'information' from one place to another. My mind is a box. I take a unit of 'information' out of it, encode the unit (that is, fit it to the size and shape of the pipe it will go through), and put it into one end of the pipe (the medium, something in the middle between two other things). From one end of the pipe the 'information' proceeds to the other end, where someone decodes it (restores its proper size and shape) and puts it in his or her own boxlike container called a mind."

Walter Ong, *Orality and literacy*

The view of communication I've presented today can be contrasted to another view of communication that is less realistic, but that is reflected in the language in everyday talk about communication. This other view is known as the "conduit metaphor." Walter Ong, a researcher who studied how people communicate using both spoken language and written language, characterized the conduit metaphor as suggesting "that communication is a pipeline transfer of units of material called 'information' from one place to another."

According to the conduit metaphor, my mind is a box. When I want to communicate, I take a unit of "information" out of my mind, and encode it so that it will fit the size and the shape of the pipe that it needs to go through. I put it into one end of the pipe, which is the medium of communication, something that's in the middle between me and the intended receiver. And then from my end of the pipe, the "information" proceeds to the other end where somebody decodes it, meaning that they restore it to its proper size and shape, and then they place it in their own box-like container called a mind.

This is a model that really distorts the act of communication. This is **not** how communication happens. And yet we often *talk* as if this is how information happens. We say things like, "Make sure you get the message across."
"Make sure that you put your ideas into" your essay, or your e-mail, or your advertisement, so that the message can be "sent" to the receiver and they can "retrieve" the message that you intended from the document that you created. We talk like that. But that's a very unrealistic model of communication.

Documents are not "containers" for information. They are tools or equipment that we can use, successfully or unsuccessfully, to have meaningful informative experiences.

Information as an abstract substance

flowing through conduits

> "It's always there when we look for **it**, available wherever we bother to direct our attention. We can glean **it** from the pages of a book or the morning newspaper and from the glowing phosphors of a video screen. Scientists find **it** stored in our genes and in the lush complexity of the rain forest. The Vatican Library has a bunch of **it**, and so does Madonna's latest CD. And **it**'s always in the air where people come together, whether to work, play, or just gab."
>
> The information revolution, *Businessweek*, August 1998

information as a "content" divorced from any specific physical realization

Implications of the conduit metaphor

- Spoken or written **language functions like a pipe** allowing the transfer of standardized units from one person to another
- To communicate something to someone else, one must **transfer ideas from their mind to these standardized units** (“encoding”)
- These **standardized units contain ideas**
- Ideas that have been placed into these standardized units **objectively exist in an external space**, independent of any human interaction
- The person being communicated to must **transfer ideas from these standardized units to their mind** (“decoding”)

Reddy (1979) The conduit metaphor

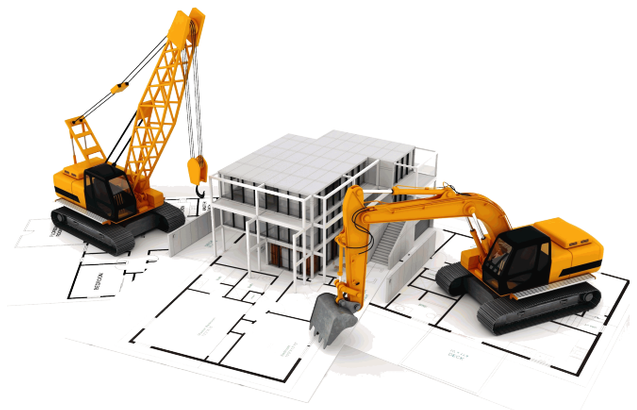
Implications of the conduit metaphor

- **Successful communication is routine.** A failure to communicate is due to a failure by the sender to properly insert his or her ideas into suitable forms. On the other hand, if this is done properly, understanding is assured.
- **Books, videos, images, Web pages, social media posts, etc. contain ideas.** The more of these things we can keep around, the more ideas we preserve. This is how we maintain our culture.

Reddy (1979) The conduit metaphor

Countering myth #1

Meaning is *constructed*



19

Linda Nylind/The Guardian

Meaning is constructed by individuals in specific situations. Documents (persistent messages) are tools that we use to construct meaning. Our understanding of those tools depends on our past experiences, and most importantly on our experience of being a participant in various social relations.

A sign

is a 3-way signifying relation

Peirce's theory of signs

Signs can be analyzed as three-way relationships. Consider this fellow writing the word “CAT” while thinking about his cat, while his actual cat is there sleeping in the corner.

A sign

is a 3-way signifying relation

Peirce's theory of signs

representamen

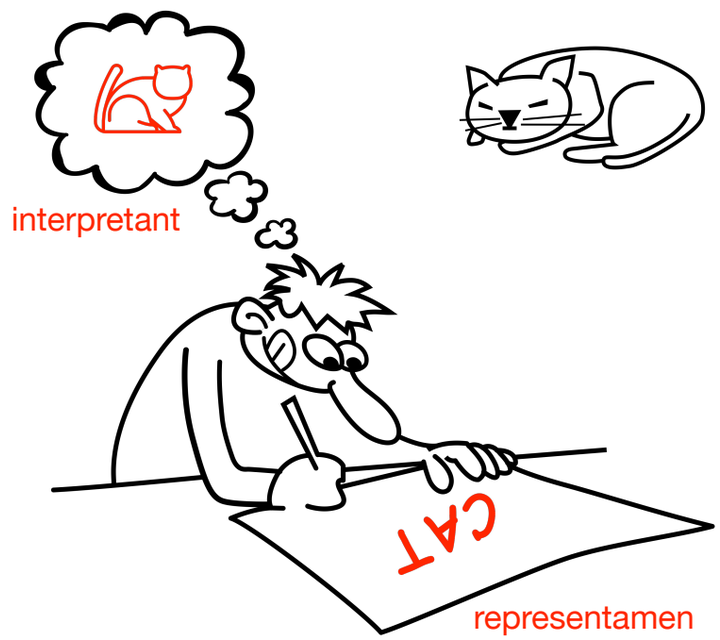
The words he's writing on the page are the *representamen* of the sign: these lines on paper that spell out C-A-T.

What is motivating him to write those words, or what happens in his mind when he reads those words, is his mental concept of the cat. Maybe he’s thinking about his own cat.

A sign

is a 3-way signifying relation

Peirce's theory of signs

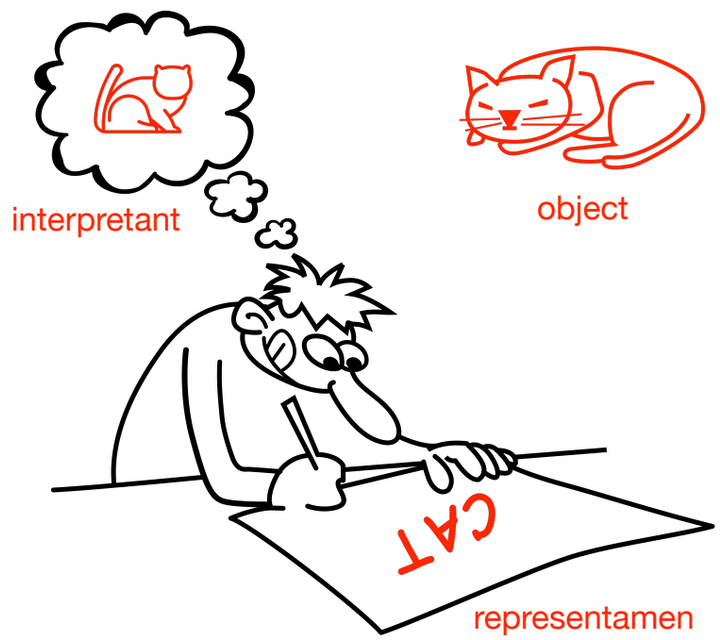


So we call this that mental image—what disposes this fellow to write those marks, or the reaction that he has upon seeing the written marks—is the *interpretant* of the sign. It's his interpretation of those marks. Individual meaning

A sign

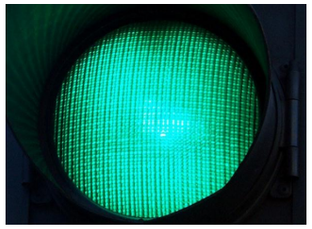
is a 3-way signifying relation

Peirce's theory of signs

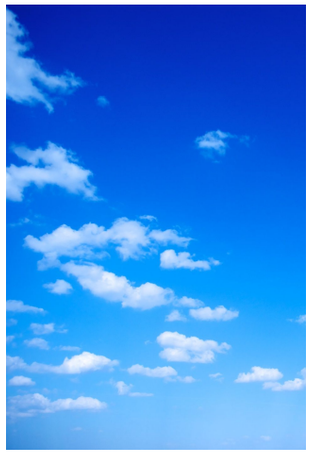


Finally, the actual cat sleeping over in the corner—or maybe the general category of *cat* to which the actual cat belongs—is the *object* in this three-way signifying relation. The object is the thing that the sign is understood to refer to. Social meaning

green



blue

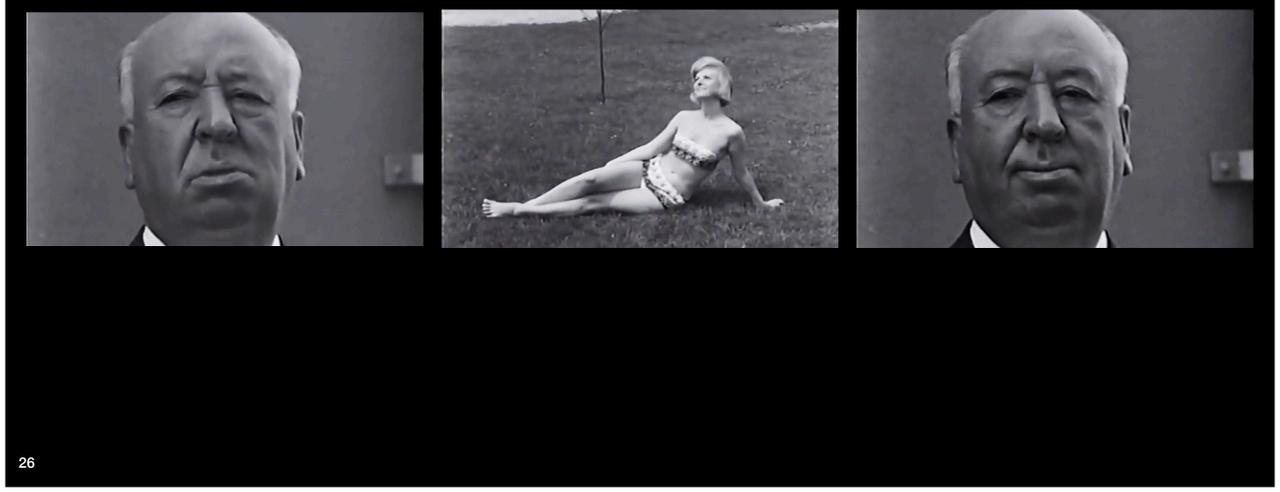


緑

青

25

Kuleshov effect. Kindly man



26

Dirty old man

Told not to say ‘gay’ in graduation speech, he made his point anyway

He said ‘curly hair’ instead

By Valerie Strauss

May 24, 2022 at 6:44 p.m. EDT

Zander Moricz speaks at the Pine View School graduation in Osprey, Fla. (Twitter)

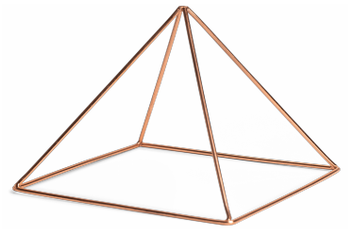
https://www.washingtonpost.com/education/2022/05/24/he-couldnt-say-gay-graduation-speech/

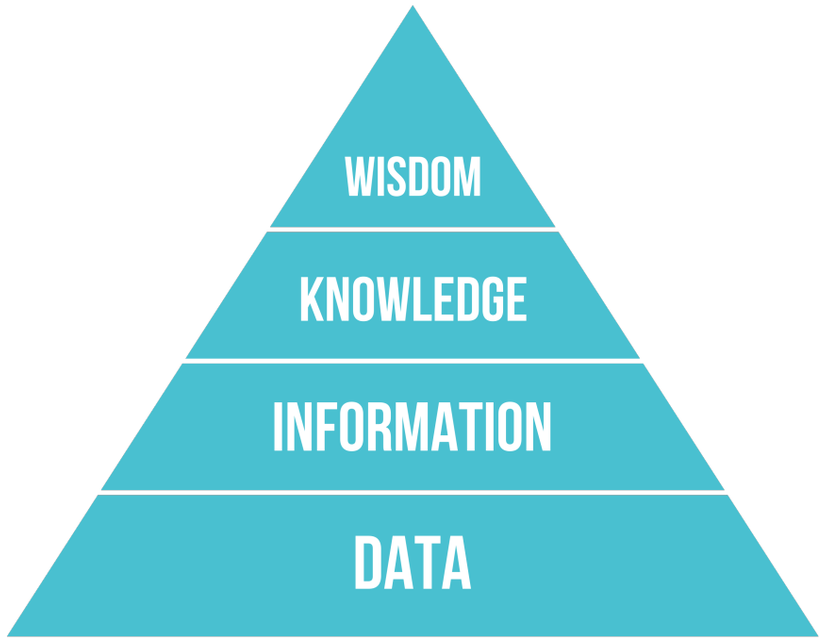
Meaning is constructed

- Meaning is not a substance that is contained inside documents
- If it were, there would be no way to explain how the same document can mean different things to different readers
- Documents are tools for constructing meaning
- One's understanding of what those tools can do is grounded in socially conditioned experience
- It's always possible to discover new uses for tools

Myth #2

The knowledge pyramid





DIKW pyramid (2020) Wikimedia Commons

The knowledge pyramid

Myth #2

Data are

- symbols that represent observable properties of objects or events
- the raw products of instruments that sense those properties

Frické (2019) Knowledge pyramid, Encyclopedia of KO

The knowledge pyramid

Myth #2

Information is

- relevant answers to questions someone might ask
- the results of processing raw data to make it significant to someone

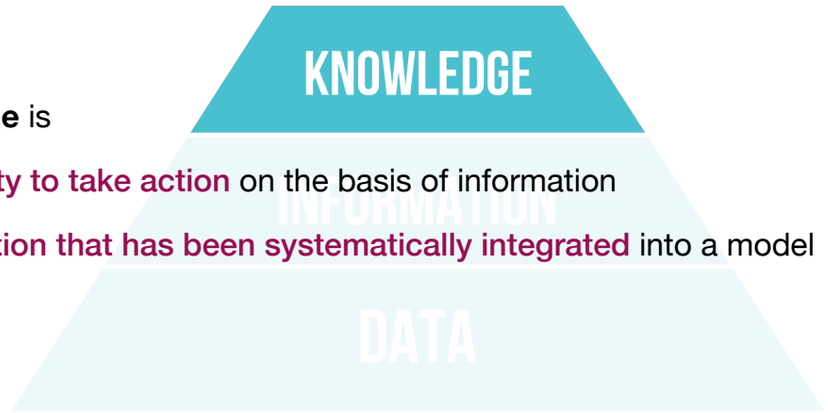
Frické (2019) Knowledge pyramid, Encyclopedia of KO

The knowledge pyramid

Myth #2

Knowledge is

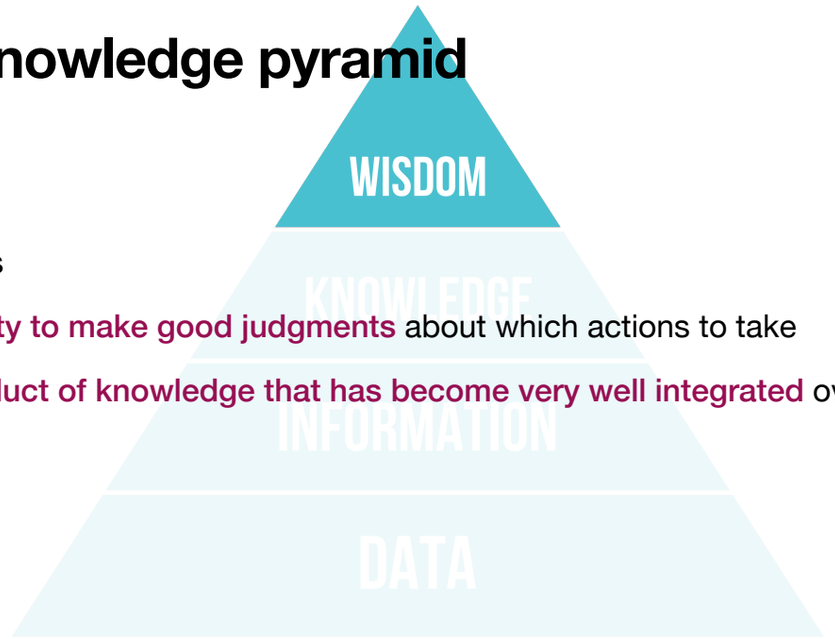
- the ability to take action on the basis of information
- information that has been systematically integrated into a model



Frické (2019) Knowledge pyramid, Encyclopedia of KO

The knowledge pyramid

Myth #2



Wisdom is

- the ability to make good judgments about which actions to take
- the product of knowledge that has become very well integrated over time

Frické (2019) Knowledge pyramid, Encyclopedia of KO

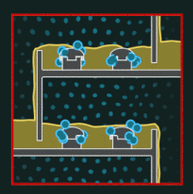

The Economist
Crunch time in France
Ten years on: banking after the crisis
South Korea's unfinished revolution
Biology, but without the cells
MAY 6TH–12TH 2017
The world's most valuable resource
Data and the new rules of competition
Google
amazon
UBER
Google
TESLA
TESLA

DISTILLATION

Crude oil contains a variety of **hydrocarbons** that have different boiling points. To separate these compounds, the oil is first sent to a boiler where it is heated into a super-hot mixture of liquid and vapour called the feed.

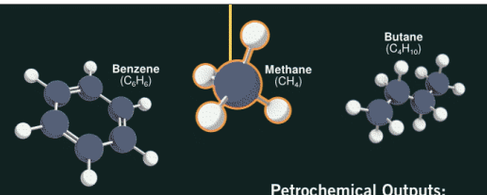
The mixture is then fed into a **distillation tower.** In here, the compounds with a lower boiling point rise up as vapours, while the compounds with a higher boiling point fall downwards as liquids.

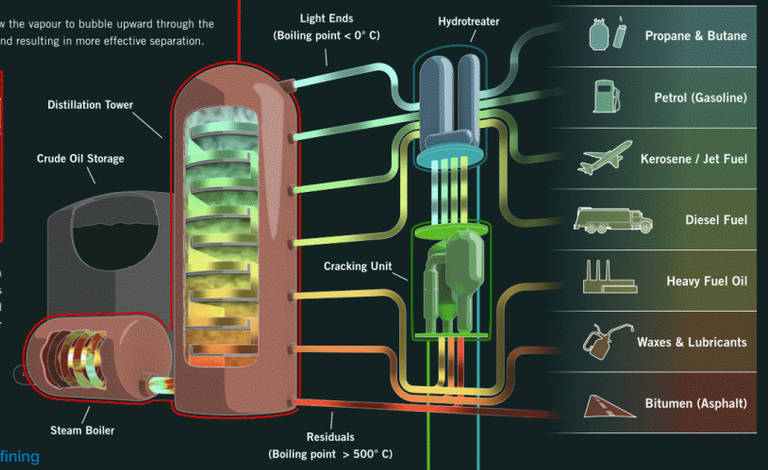
The tower contains trays that allow the vapour to bubble upward through the liquid, helping to exchange heat and resulting in more effective separation.



The distilled products are then piped off from the different levels of the tower. These separated products are called **fractions** or **distillates.**

This process may take place along multiple distillation towers.





FES Tanks (2015) Oil refining

Countering myth #2

Data and information are derived from knowledge

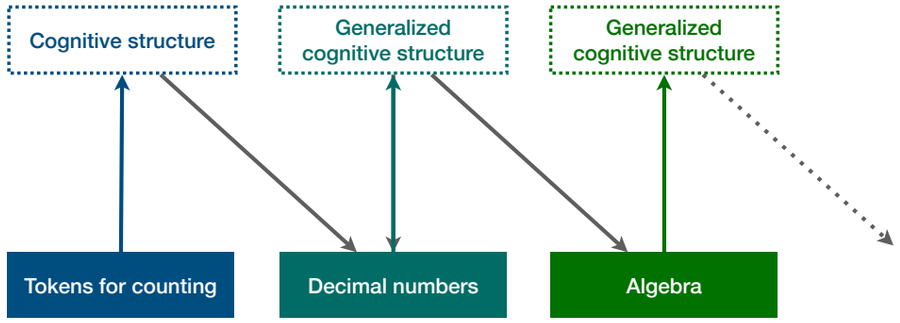
Knowledge

An alternative account

- Only **individuals** can truly have knowledge
- Knowledge **arises from experience**
- Knowledge consists of the **anticipated possible consequences of actions**
- Knowledge is **represented by cognitive structures** that develop over time
- These cognitive structures **connect past experience with a current situation**

Renn (2020) The evolution of knowledge

Internal knowledge representations (cognitive structures)



External knowledge representations (documents)

Damerow (1996) On the relationship between ontogenesis and historiogenesis of the number concept

Documents as tools for thinking

- Writing is not primarily a technology for recording speech, but a tool for **thinking about, making sense of, and reorganizing** what is being spoken about
- It is certainly possible to use spoken language without writing
- But, for literate people, **the use of writing has influenced our relationship to language so much** that we can no longer easily differentiate between them
- Similarly, notations for scoring music or explicating chemical structure **profoundly influence the practice** of music or chemistry—they do not simply record or represent the practices as it was before
- **We are not finished inventing tools for thinking**

Noë (2015) Strange tools

writings origin in making marks to help with counting

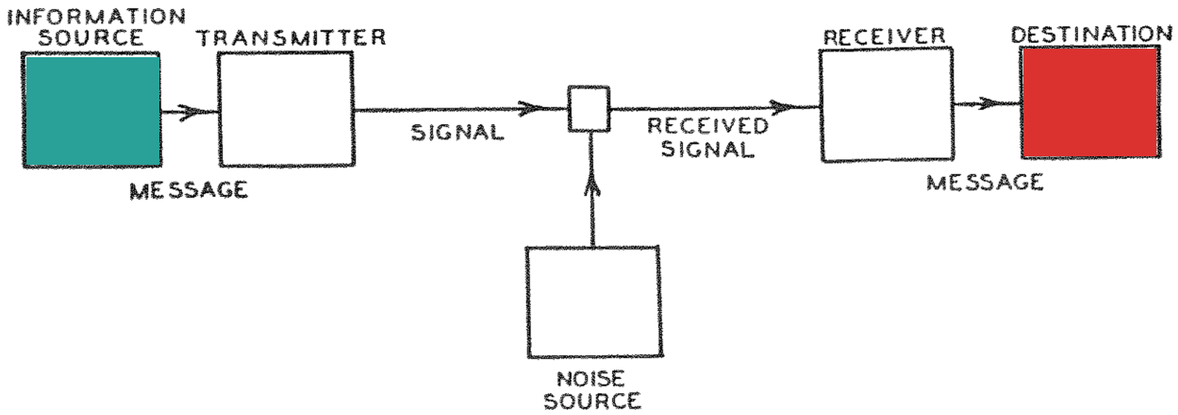
Documents of knowledge

Real material objects and actions taken with them

- Tools, artifacts, models, rituals, sound, language, music, and images, as well as symbol systems and writing
- **Secure and stabilize** the transmission of knowledge over time and space
- Tools for thinking and so **potential sources of change** and innovation
- There are **material constraints on a representation's possible uses**
- These possibilities are also **constrained by cognitive development**
- Nevertheless it is **always possible to discover new uses** of a tool

Renn (2020) The evolution of knowledge

Generative ambiguity



"The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point."

Shannon published his mathematical theory of communication as a stochastic process in 1949, after WWII, and he included in that document this diagram, which you'll notice looks very much like the classified diagram from his earlier work on the encrypted telephone system. Shannon wrote: "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point." Note that Shannon's conceptualization of the problem of communication was heavily influenced by the kinds of problems that he was working on during the war. You have some source of information, you are encoding that information into a message that is transmitted as a signal, received someplace else, and then decoded into a message on the other side. And in the middle, you have some source of noise that might affect the signal, which could be due faulty equipment, could be due to enemy sabotage, whatever—but it makes it possible that the received signal is not the same as the signal that was sent.

Notice that there's something kind of strange about Shannon's language here. He doesn't say the problem of communication is reproducing a message that was "created" or "written" at another point. He says a message "selected" at another point.

Information

Knowledge encoded for transmission

- Meaning is the use value of knowledge
- Information is the exchange value of knowledge
- Documents unite meaning and information
- They serve as both
  - representations of the meaning of knowledge for its individual appropriation (tools for thinking)
  - means for the social transmission of knowledge (information moving from sender to receiver)

Renn (2020) The evolution of knowledge

Information theory as a generalized way of thinking about the transmission role

"Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. **These semantic aspects of communication are irrelevant to the engineering problem.**"

Shannon, The mathematical theory of communication

Exchange value, not use value

Data

The monetary form of information

- a standardized document
- mobile, countable, combinable
- serves as a universal standard and measure of information
- “big data” is capital accumulation

The ability to digitally encode information and store and transmit it electronically make data possible

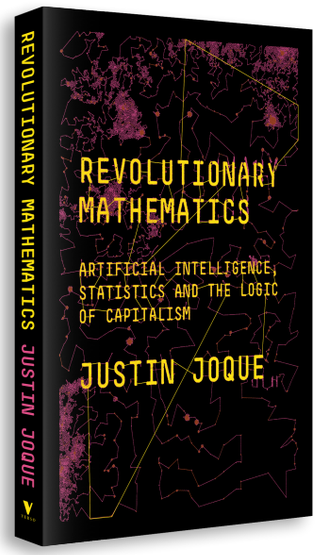
Renn (2020) The evolution of knowledge

“… it is mobile (once in small pieces), it is immutable (once in metal), it is countable (once it is coined), combinable, and can circulate from the things valued to the center that evaluates and back … Money is used to code all states of affairs …” V&C, 29

The capitalist calculation problem

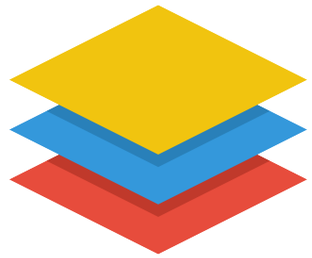
Justin Joque, *Revolutionary mathematics*

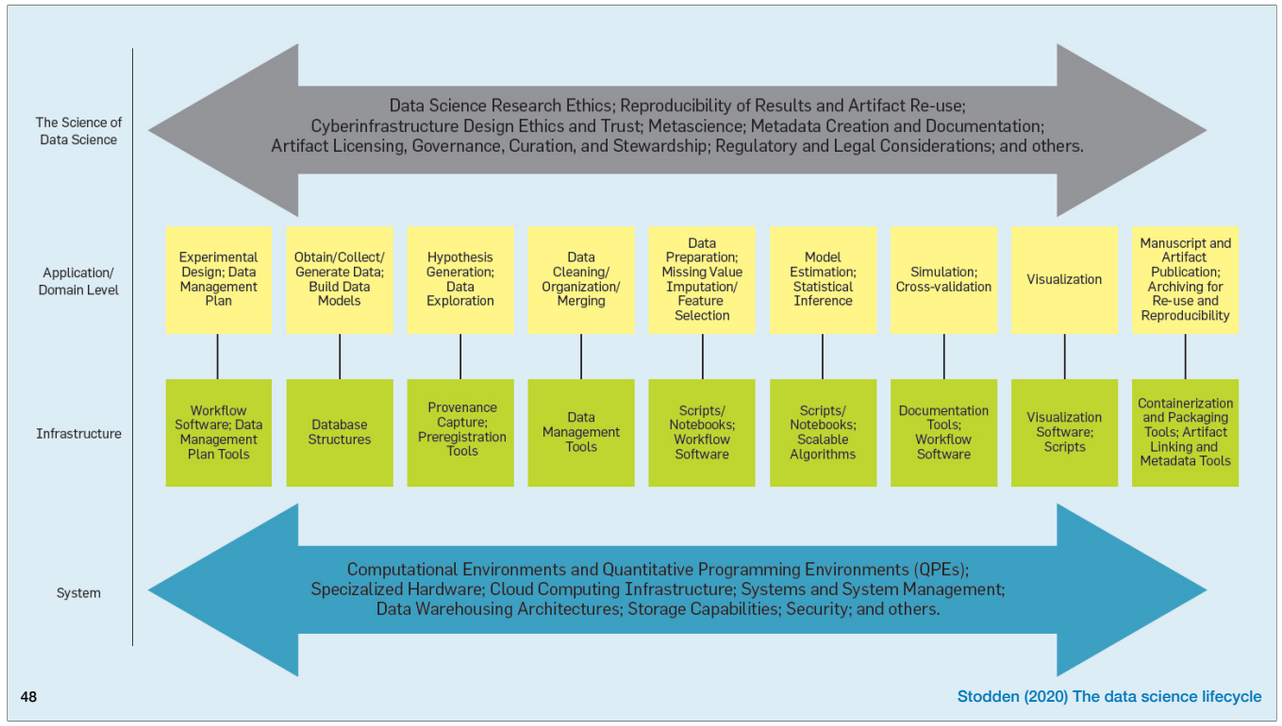
- All individual knowledge is developed through shared knowledge (documents)
- The promise of data science lies in the accumulation of shared pools of knowledge (documents)
- Data science is about the development of new tools for thinking, an inherently social process
- But data-as-capital incentivizes the private enclosure of knowledge
- Data science becomes a tool for cheating and lying to best one's competitors



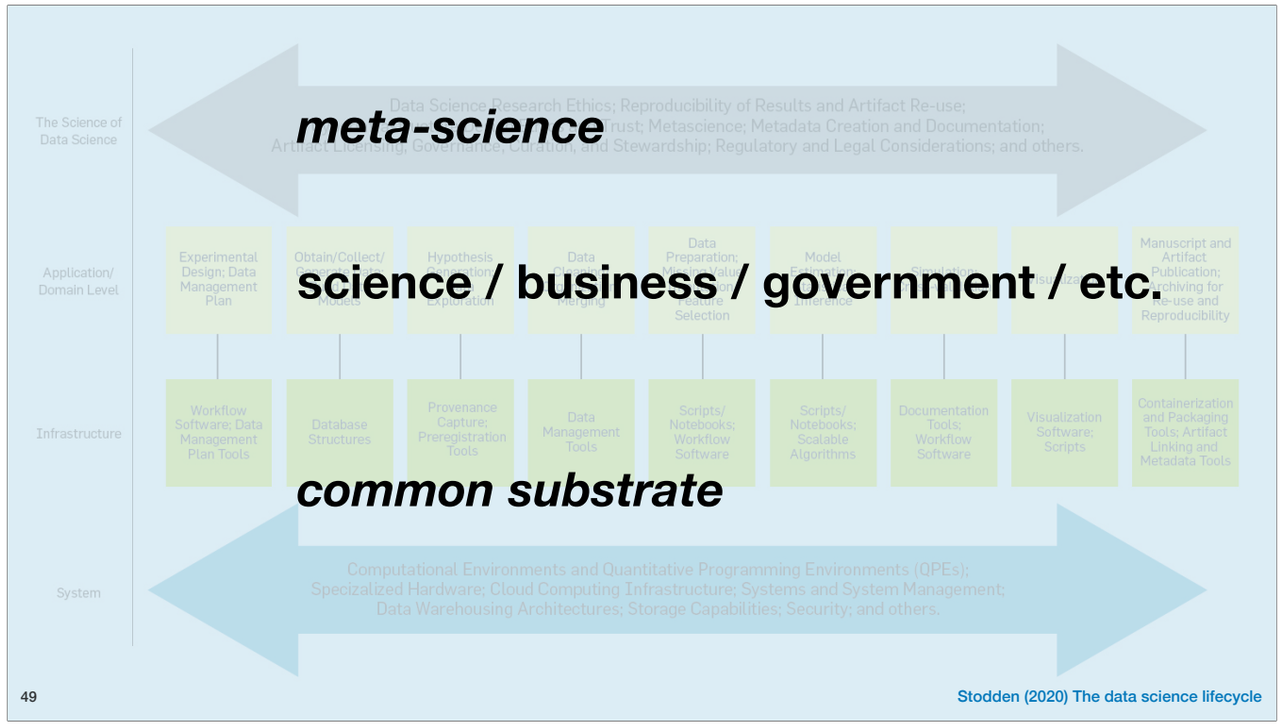
Myth #3

The common substrate





Stodden (2020) The data science lifecycle



Stodden (2020) The data science lifecycle

The ideology of data science

- Positions **data science as a neutral profession**
- The science and engineering of data science are **assumed to operate at different levels** from their application
- Data scientists **do not get involved in disputes at the "application" level**
- Differences among kinds of documents are ignored; **everything is just more data**
- The strategy is to **provide generically defined services to diverse institutional customers**, making only limited accommodations for the specific structures and ideologies of each

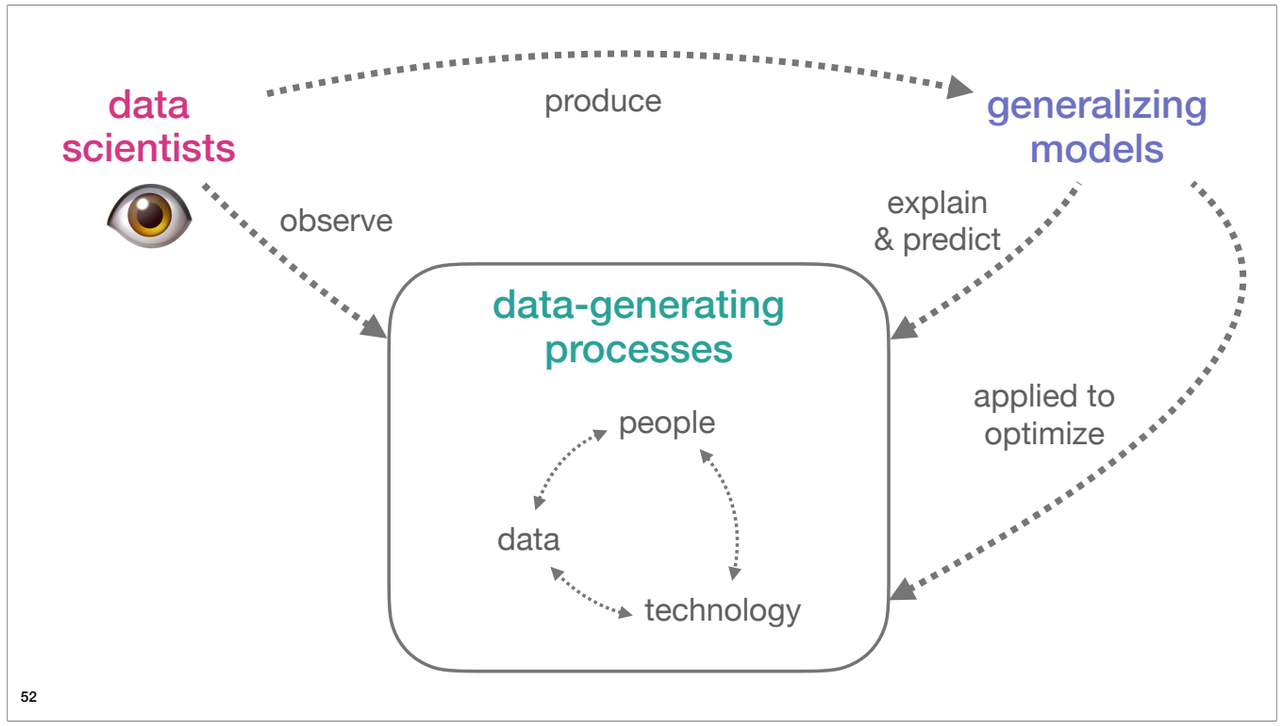
Agre (1995) Institutional circuitry

The ideology of statistics

"… statistics… has always been a small, largely academic discipline that extends vast intellectual jurisdiction by commodifying its techniques in texts, formulas, tables, and graphing tools… statisticians flood their techniques everywhere, let others use them badly, and make a living repairing bad applications and contracting their direct services to the elite clientele… One can easily imagine a large profession like public accounting, but called statistics, that comprised consultants who would swear that statistical analyses meant what the substantive authors claimed they did. In fact, however, this did not happen, for reasons which suggest a deliberate choice."

Andrew Abbott, *The system of professions*

Russian semiology of statistics; Bernhard Rieder's work on ordering techniques as "interested readings of reality"



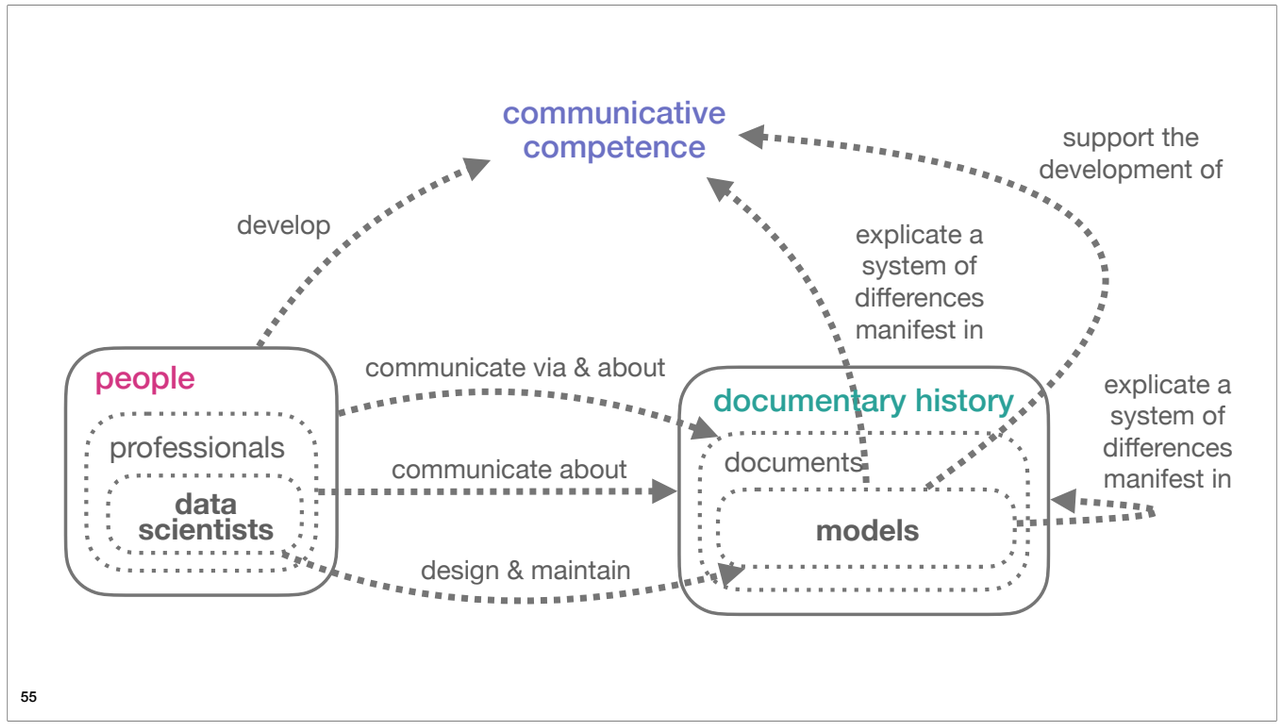
This paradigm of data science posits that there are various ongoing data-generating processes. Data scientists observe these processes, paying special attention to the form and organization of data and how these relate to the goals of the people creating and using it. Because participants in the processes are preoccupied with the *content* of data, they are blind to its form and organization. So, data scientists should stand outside the processes they observe, as it is this standpoint that enables them to develop special insight into the form and organization of data without being distracted by its content. From this meta-level, data scientists produce generalizing models. They use these models to explain and predict the data-generating process. When working in an applied mode, they use their models to optimize the process.

Countering myth #3

Data science pluralism

Data science is not a separate layer

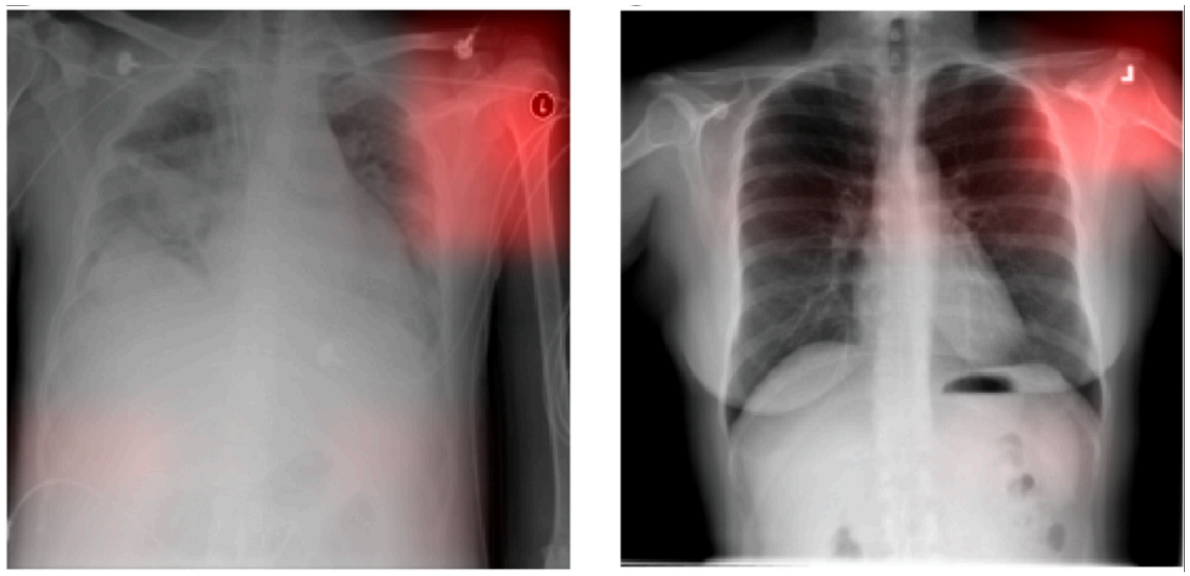
- Risk that the “application” of data science comes to be seen as solely a problem of how to cast a “domain” in mathematical terms amenable to computation
- But the world cannot easily be separated into “data science” and “application domain” levels
- The data that power data science are complex, historically-specific products of institutional, organizational, and technical infrastructures
- Doing data science well requires deep understanding of the history and ideology of specific data-generating practices
- Data science is better understood as a specialization of existing processes of generating and interpreting documents


communicative competence
develop
support the development of
explicate a system of differences manifest in
people
professionals
data scientists
communicate via & about
communicate about
design & maintain
documentary history
documents
models
explicate a system of differences manifest in

In summary

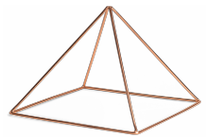
Problems with the conduit metaphor

- Meaning is conceived of as **something to be extracted** from data
- Assumes that there is **just one true meaning** (“ground truth”)
- The world is **always more than what can be represented**
- Forgetting this is **a source of flaws** (e.g. spurious correlations)



Zech et al. (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs

Problems with the knowledge pyramid



- Data is not “raw material”
- Knowledge is prior to information and data, not derived from them
- Documents are both means of transmitting knowledge and tools for generating new knowledge
- The techniques of data science are just the latest developments in a long line of technologies for creating and interpreting documents
- Treating data as capital risks destroying the benefits of data science

Problems with the common substrate

- History and structure influence “content”:
  science can’t easily be separated from meta-science
- Multiple specialized data sciences can arise
- What we need are not data scientists with some knowledge of a domain, but experts in a domain with some knowledge of data science techniques
- A major challenge is to coordinate across disciplinary boundaries (e.g., epidemiology & economics)
- But the “common substrate” model just assumes away this challenge

In summary

- **Meaning is constructed**, not extracted
  Statistics and computing can provide new constructive tools

- **Data and information derive from knowledge**, not vice versa
  Data science for the common good requires shared knowledge

- **There are many data sciences**, not just one
  Let a hundred flowers bloom

References

Abbott, A. (1988). *The system of professions: An essay on the division of expert labor.* University of Chicago Press.

Agre, P. E. (1995). Institutional circuitry: Thinking about the forms and uses of information. *Information Technology and Libraries*, 14(4), 225–230.

Damerow, P. (1996). On the relationship between ontogenesis and historiogenesis of the number concept. In *Abstraction and representation.* Springer. https://doi.org/10.1007/978-94-015-8624-5_9

Frické, M. (2019). The knowledge pyramid: The DIKW hierarchy. *Knowledge Organization*, 49(1), 33–46. Also in B. Hjørland & C. Gnoli (Eds.), *ISKO encyclopedia of knowledge organization*. http://www.isko.org/cyclo/dikw

Joque, J. (2022). *Revolutionary mathematics: Artificial intelligence, statistics and the logic of capitalism.* Verso.

Noë, A. (2015). *Strange tools: Art and human nature.* Hill and Wang.

Ong, W. (1982). *Orality and literacy: The technologizing of the word.* Routledge.

Reddy, M. J. (1979). The conduit metaphor—a case of frame conflict in our language about language. In A. Ortony (Ed.), *Metaphor and thought* (p. 284-297). Cambridge University Press.

Renn, J. (2020). *The evolution of knowledge: Rethinking science for the Anthropocene.* Princeton University Press. https://doi.org/10.1515/9780691185675

Shaw, R. (2015). Big data and reality. *Big Data & Society*, 2(2). https://doi.org/10.1177/2053951715608877