Introduction aux sciences des données

Kavé Salamatian

UNIVERSITÉ SAVOIE MONT BLANC

LISTIC

Data Science – A Definition

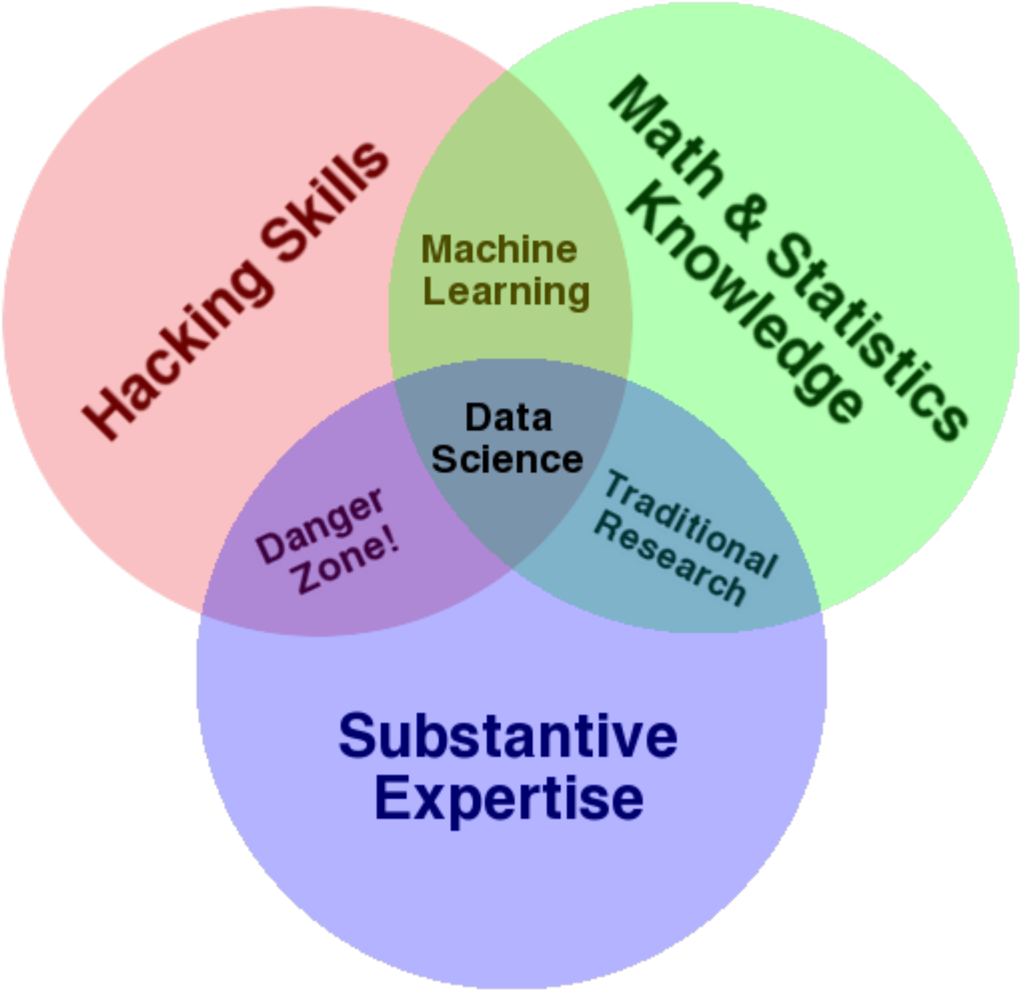
Data Science is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, interact with data to create data products.

Turn data into data products.

UNIVERSITÉ SAVOIE MONT BLANC

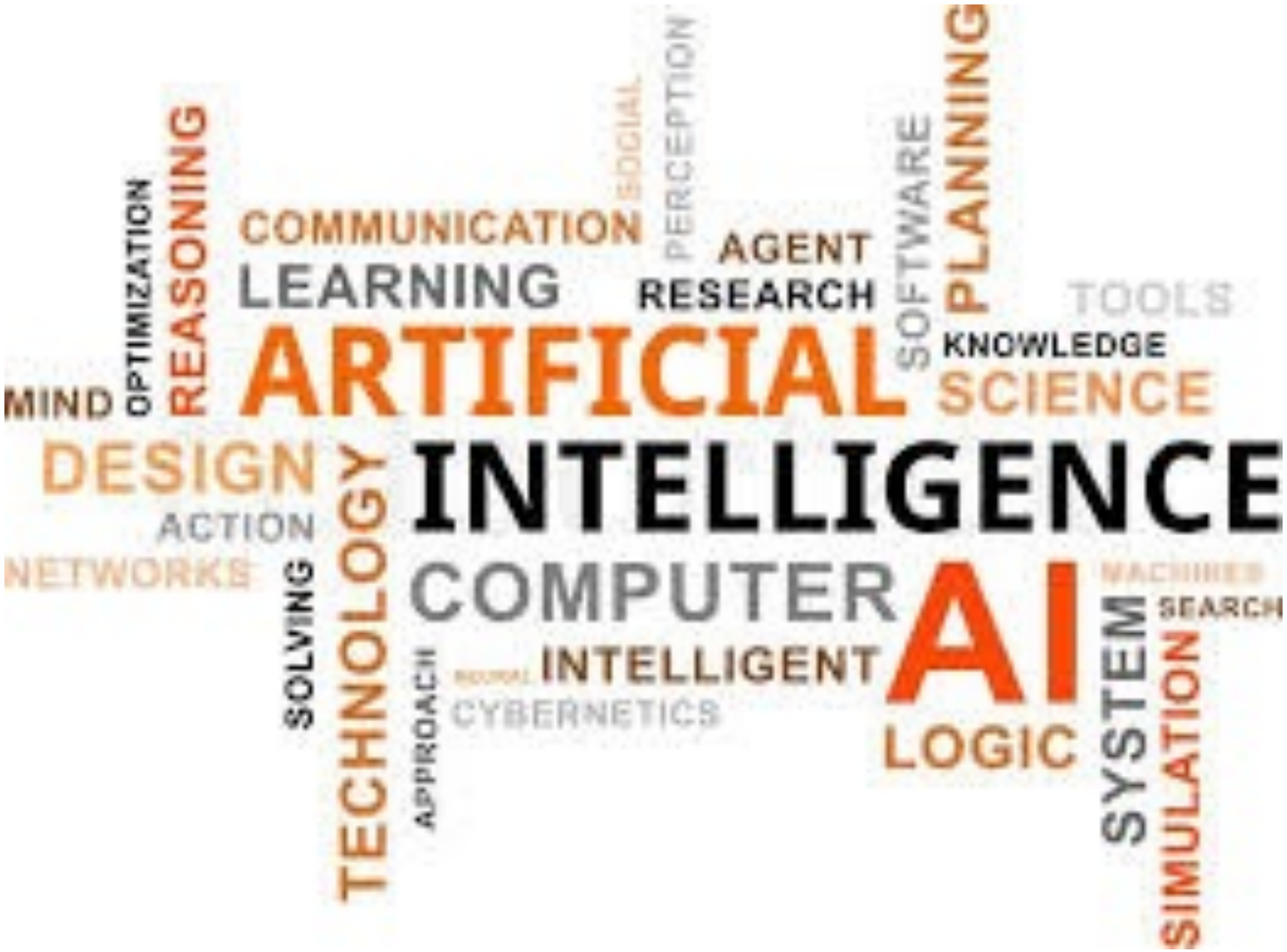
LISTIC

Data Science – A Visual Definition



Any fool consider himself as intelligent

Danish folklore



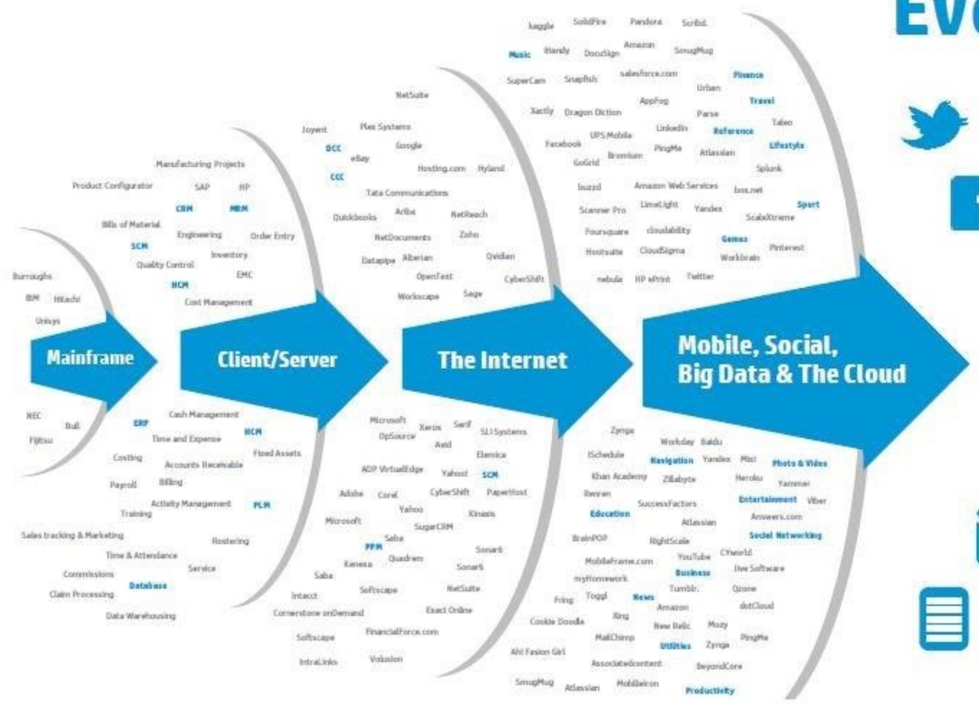
Why computers ?

- Capture and storage of massive amount of data
  - Diverse, heterogeneous, imperfect
- Data processing
  - Sanitisation
- Statistical modeling
  - Co-occurrences/correlations
  - Uncertainty structure
- Specialisation
  - Application of the model to specific individuals
- Action
  - Intelligent behaviour

UNIVERSITÉ SAVOIE MONT BLANC

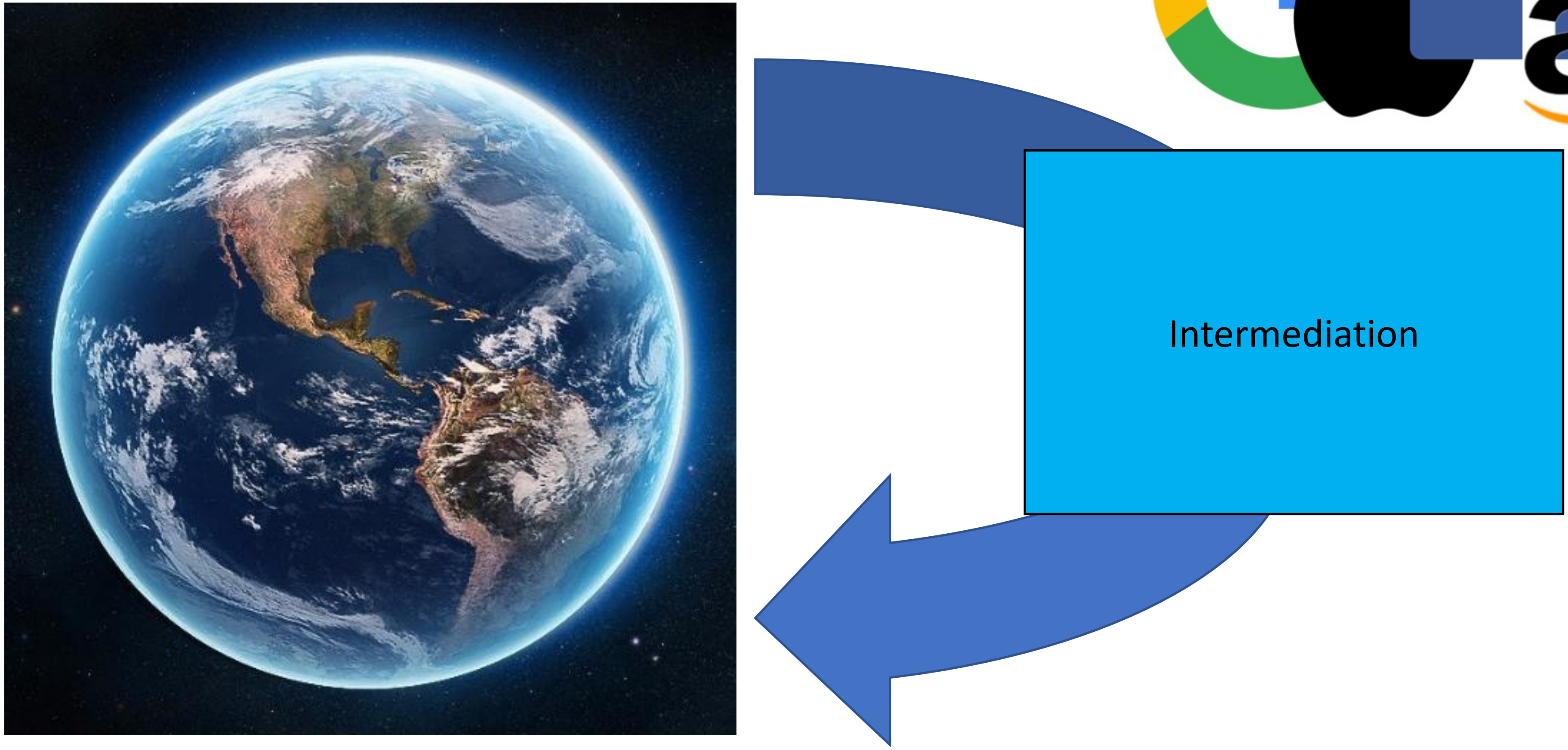
LISTIC

DATASphere AND AnthropocenE



UNIVERSITÉ
SAVOIE
MONT BLANC

Intermediation



5 Vs of Big Data

- Raw Data: Volume
- Change over time: Velocity
- Data types: Variety
- Data Quality: Veracity
- Information for Decision Making: Value

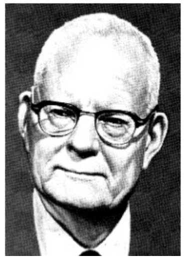
Evolution of statistics and statistics

1935: "The Design of Experiments"

R.A. Fisher

1939: "Quality Control"

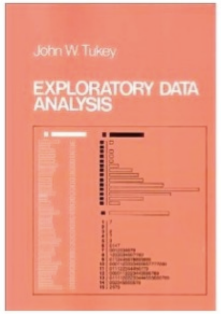
W.E. Deming



1958: "A Business Intelligence System"

Peter Luhn

1977: "Exploratory Data Analysis"

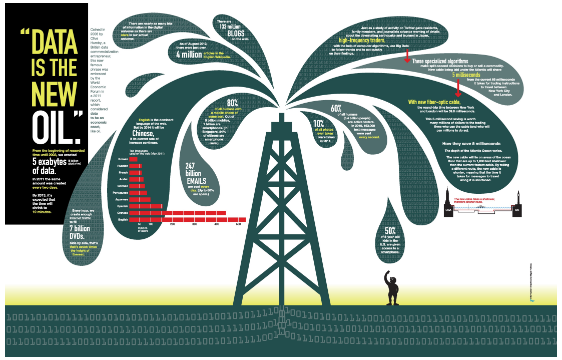


1989: "Business Intelligence"

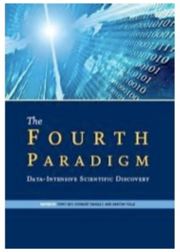
Howard Dresner

1997: "Machine Learning"

World Economic Forum 2011

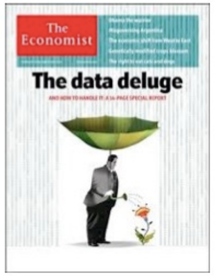


2007:"The Fourth Paradigm"



2009:"The Unreasonable Effectiveness of Data"

2010:"The Data Deluge"



UNIVERSITÉ SAVOIE MONT BLANC

LISTIC

Epistemiology: Types of scientific inferences

Deduction

All the beans in this bag are white

These beans are from this bag

These beans are white

Properties:

- The truth of the premises warrants the truth of the conclusion.
- In a deductively valid argument, it is impossible that the premises are true and the conclusion is false.
- Logically necessary
- Not ampliative (“analytic”)
- A deductive argument is called „sound“ if its premises happen to be true.

Induction

These beans are white

These beans are from this bag

All the beans in this bag are white

Properties:

- The truth of the premises does not warrant the truth of the conclusion.
- It is possible that the premises are true and the conclusion is false.
- Not necessary
- Ampliative (”synthetic”)

Abduction

(Peirce: „Hypothesis“)

All the beans in this bag are white

These beans are white

These beans are from this bag

Properties:

- The truth of the premises does not warrant the truth of the conclusion.
- It is possible that the premises are true and the conclusion is false.
- Not necessary
- Ampliative (”synthetic”)
- The conclusion „explains“ the premises.
- ***Inference to the best explanation***

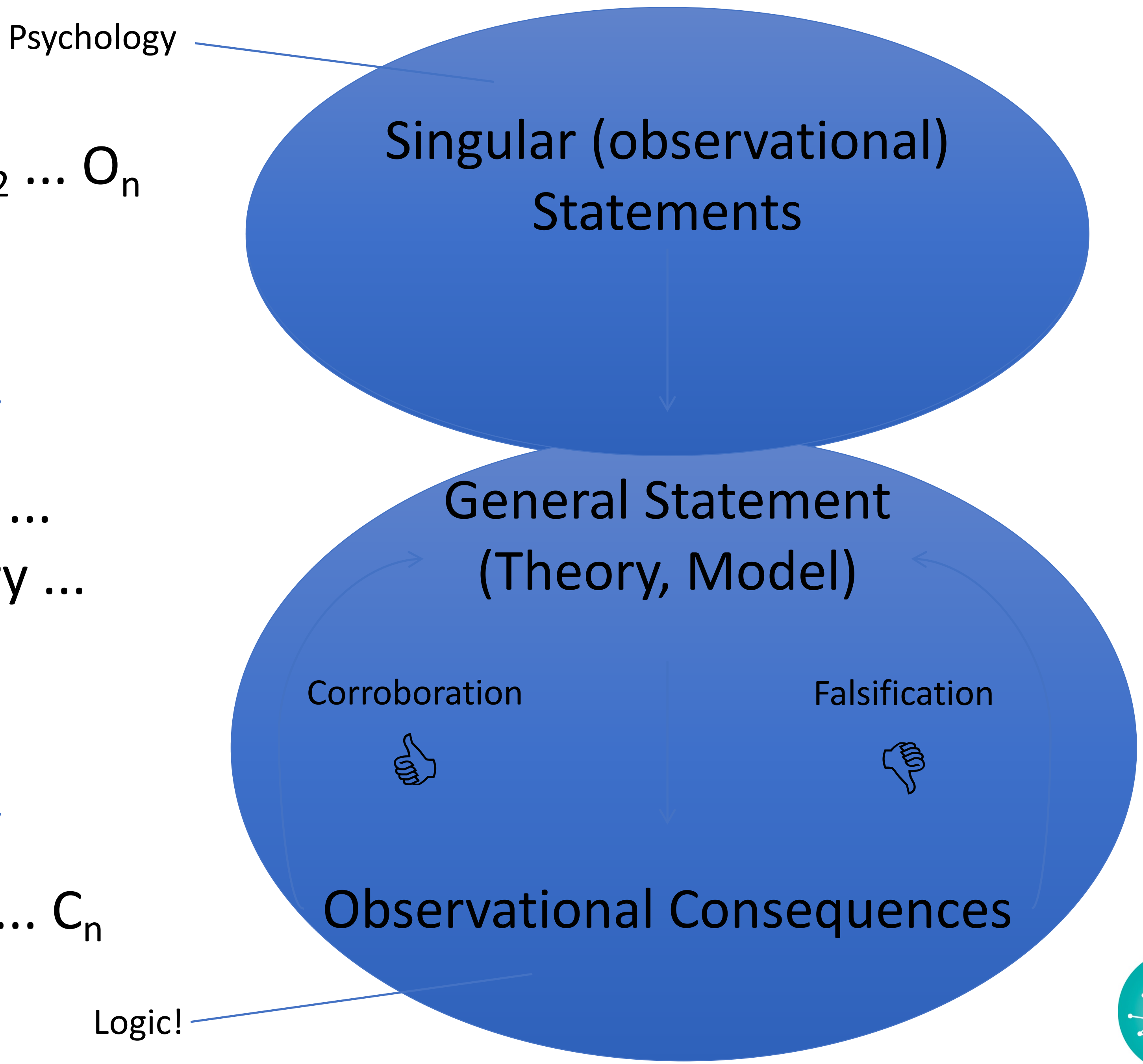
C. S. Peirce, *Deduction, Induction, Hypothesis* (1878)

Poperian Approach

$O_1, O_2 \dots O_n$

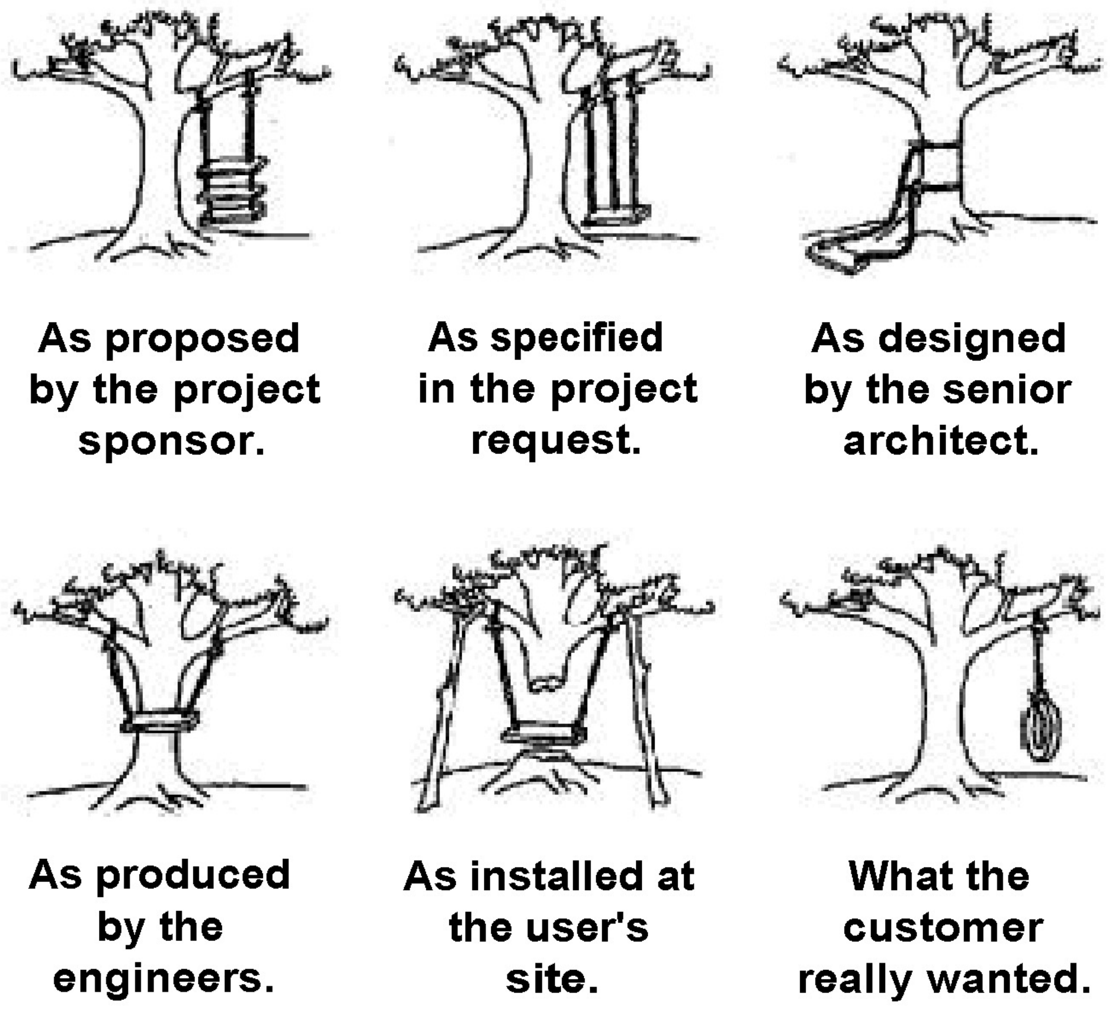
All ...
Every ...

$C_1, C_2 \dots C_n$



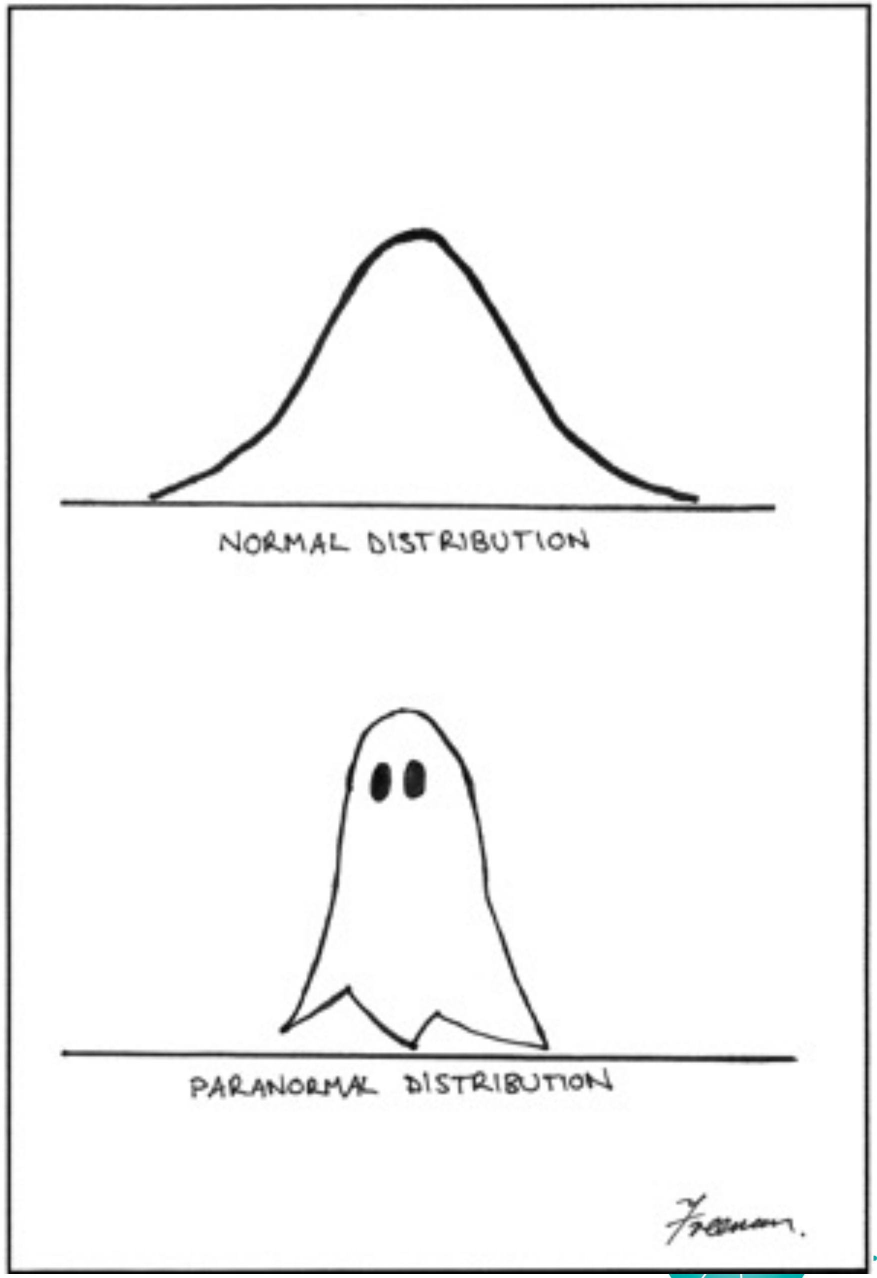
Epistemic steps

- Going from reality to model and back
    - how to not get lost in translation
- Building model steps
    - Separate what is « *in* » from « *out* »
    - Characterize the model
        - Simplify it
    - Evaluate the model
- Issues
    - Separation
    - Characterization
    - Evaluation
- We will focus on evaluation



On statistics

- There is no probability but probabilistic models
  - Some are useful, all are wrong
- At beginning there is a concrete reality
  - We move from the reality to realm of the probability model
  - We need to come back to the real world
  - Statistics defines how to not get lost in translation
- No free lunch !
  - How much I pay ?



Should we regress when we do regression ?

- Rationality conflict
  - Is the model logic dominating or the problem (real world) logic
- Oprah event
  - Discrimination
- Evaluation and transparency
  - COMPAS case

The quest for objectivity

- Finally what is objective in statistics
  - What can be transferred back in real world from the realm of statistics
- We throw a coin
  - The probability of the model is p ✓
  - Pr{10 th throw is face}=p ✗
- We construct a model of customer interests
  - The average interests of customers is characterized by given parameters ✓
  - The customer X will be interested to Y with probability p ✗
- Application of statistical models to individuals can only be subjective

Correlations vs. Causation example

- ‘‘The correlation between workers’ education levels and wages is strongly positive”
- Does this mean education “causes” higher wages?
  - We don’t know for sure !
- Recall: Correlation tells us two variables are related BUT does not tell us why

UNIVERSITÉ
SAVOIE
MONT BLANC

LISTIC

Correlation vs. Causation

- Possibility 1
  - Education improves skills and skilled workers get better paying jobs
    - Education causes wages to ↑
- Possibility 2
  - Individuals are born with social environment *A* which is relevant for success in education and on the job
    - Social environment (NOT education) causes wages to ↑

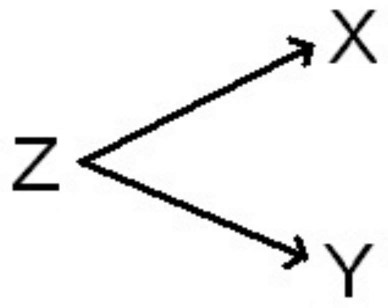
Consider the following research undertaken by the University of Texas Health Science Center at San Antonio appearing to show a link between consumption of *diet* soda and weight gain.

- *The study of more than 600 normal-weight people found, eight years later, that they were 65 percent more likely to be overweight if they drank one diet soda a day than if they drank none.*
- *And if they drank two or more diet sodas a day, they were even more likely to become overweight or obese.*

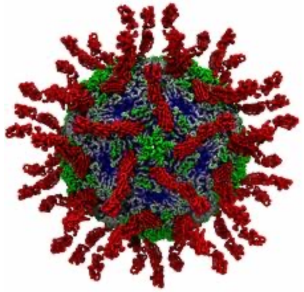
- A strong relationship between two variables does not always mean that changes in one variable causes changes in the other.
- The relationship between two variables is often influenced by other variables which are lurking in the background.
- There are two relationships which can be mistaken for causation:

1. Common response
2. Confounding

- ***Common response*** refers to the possibility that a change in a lurking variable is causing changes in both our explanatory variable and our response variable



- ***Confounding*** refers to the possibility that either the change in our explanatory variable is causing changes in the response variable OR that a change in a lurking variable is causing changes in the response variable.



Example
Correlation v.s. Causation

One study during the polio epidemic in the 1920s showed a strong correlation between ice cream consumption and cases of polio. As a result, the public was warned to avoid eating ice cream as it increased the risk of contracting the disease.

Thoughts?

- Again, there was a strongly confirmed correlation. However, it turned out that there was NO causation. With a properly controlled experiment, it could have been easily shown that increased ice cream consumption did NOT increase the risk of polio.
- Again, there was a **lurking variable** hiding in the background. It turns out that the virus that causes polio (a virus of the *picornoviridae* family for anyone who cares) thrives in warmer weather. So the lurking variable here was temperature!

UNIVERSITÉ SAVOIE MONT BLANC

LISTIC

Example

STUDY: In a study, babies of women who bottle feed and women who breast feed are compared, and it is found that the incidence of gastroenteritis, as recorded in medical records, is lower in the babies who are breast-fed.

RANDOM ERROR

By chance, there are more episodes of gastroenteritis in the bottle-fed group in the study sample. (When in truth breast feeding is not protective against gastroenteritis).

Or, also by chance, no difference in risk was found. (When in truth breast feeding is protective against gastroenteritis).

MISCLASSIFICATION

Lack of good information on feeding history results in some breast-feeding mothers being randomly classified as bottle-feeding, and vice-versa.

If this happens, the study *underestimates either of the two groups.*

BIAS

The medical records of bottle-fed babies *only* are *less complete* (perhaps bottle fed babies go to the doctor less) than those of breast fed babies, and thus record fewer episodes of gastro-enteritis in them only.

This is called bias because the observation itself is in error.

In this case the error was not conscious.

CONFOUNDING

The mothers of breast-fed babies are of higher social class, and the babies thus have *better hygiene, less crowding* and perhaps other factors that protect against gastroenteritis.

Less crowding and better hygiene are truly protective against gastroenteritis, but we mistakenly attribute their effects to breast feeding.

This is called confounding, because the observation is correct (breast-fed babies have less gastroenteritis), but its explanation is wrong.

UNIVERSITÉ SAVOIE MONT BLANC

LISTIC

Overfitting or something else?

- Cars climbing up trees (at CMU)…
- Road sides look like parallel lines.
- But, unfortunately, so do trees!
- Related “incident”:
- Task: Given pictures of wooded areas,
- find pictures where tanks are hidden
- (late 1960s, DARPA challenge)

- Remarkably good performance on test set (like 99% correct).
- Then, general noticed, pictures containing tanks were taken in the late afternoon. Without tanks, in the morning. ☹
- What is the issue here? Good or bad learning? Overfitting? Or something else?
- ML is “good” but: The training data itself “flawed”!!!!
- Also, problem with “algorithmic bias.” ML methods learn / reinforce “unwanted” biases eg in hiring or loan decisions. But you may not realize it!

UNIVERSITÉ SAVOIE MONT BLANC

LISTIC

Data Science

Interdisciplinary

...[DS includes] mathematics, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products

Mout

The field of data science is emerging at the intersection of the fields of social science and statistics, information and computer science, and design

Berkeley School of Information

New kinds of Data

Extraction of knowledge from large volumes of data that are structured or unstructured, which is a continuation of the field data mining and predictive analytics, also known as knowledge discovery and data mining (KDD). "Unstructured data" can include emails, videos, photos, social media, and other user-generated content.

Wikipedia

First, the raw material, the “data” part of Data Science, is increasingly heterogeneous and unstructured. Second, computers interpret data automatically, making them active agents in the process of sense making.

Dhar

Data as product

...merely using data isn’t really what we mean by “**data science**.” A data application acquires its value from the data itself, and creates more data as a result. It’s not just an application with data; it’s a data product. Data science enables the creation of data products

Loukadis (O’Reilly Media)

New methods for making-sense to data

Data science is the study of where information comes from, what it represents and how it can be turned into a valuable resource in the creation of business and IT strategies

Rouse

At its core, data science involves using automated methods to analyze massive amounts of data and to extract knowledge from them.

New York University

Data Science landscape

Data Science

(Wikipedia)

Fields

- Nanotechnologies
- Physics
- Robotics
- Mathematics
- Statistics
- Information theory
- Information technology
- AI

Techiniques

- Signal processing
- Probability models
- Machine learning
- Statistical learning
- Data mining
- Database
- Data engineering
- Pattern recognition
- Visualization
- Predictive analytics
- Uncertainty modeling
- Data warehousing
- Data compression
- Computer programming
- High Performance Computing

Objects

Methods that scale to Big Data are of particular interest in data science, although the discipline is not generally considered to be restricted to such data.

Approaches

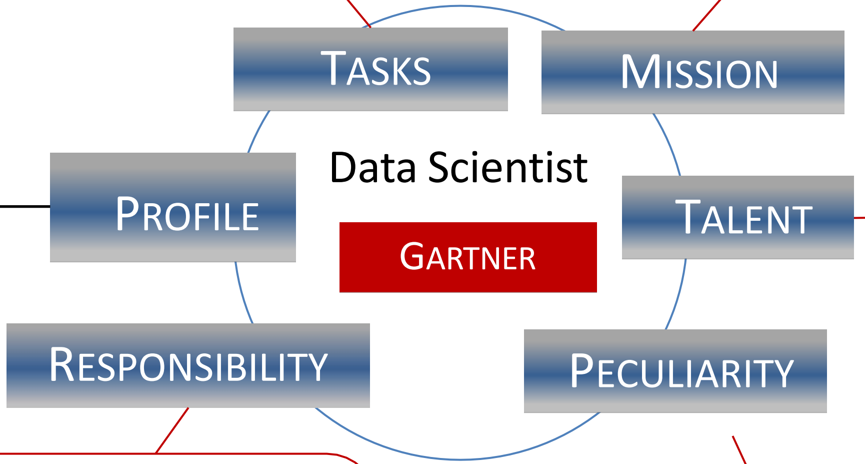
The development of machine learning, a branch of artificial intelligence used to uncover patterns in data from which predictive models can be developed, has enhanced the growth and importance of data science.

Who is a Data Scientist?

In addition to advanced analytic skills, this individual is also proficient at integrating and preparing large, varied datasets, architecting specialized database and computing environments, and communicating results.

A data scientist may or may not have specialized industry knowledge to aid in modeling business problems and with understanding and preparing data.

The data scientist has emerged as a new role, distinct from — but with similarities to — those of business intelligence (BI) analysts and statisticians

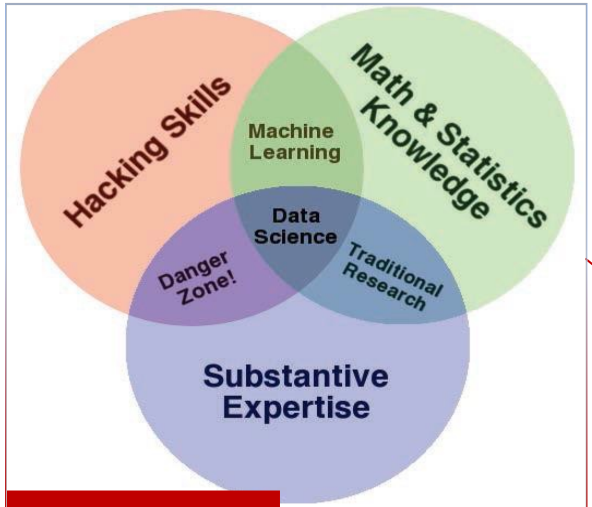


Creating value from data requires a range of talents: from data integration and preparation, to architecting specialized computing/database environments, to data mining and intelligent algorithms

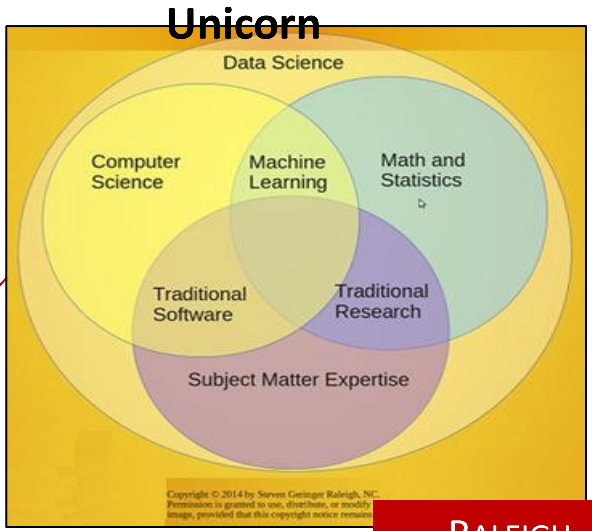
An individual responsible for modeling complex business problems, discovering business insights and identifying opportunities through the use of statistical, algorithmic, mining and visualization techniques.

Data scientists can be invaluable in generating insights, especially from "big data;" but their unique combination of technical and business skills, together with their heightened demand, makes them difficult to find or cultivate.

Venn diagram

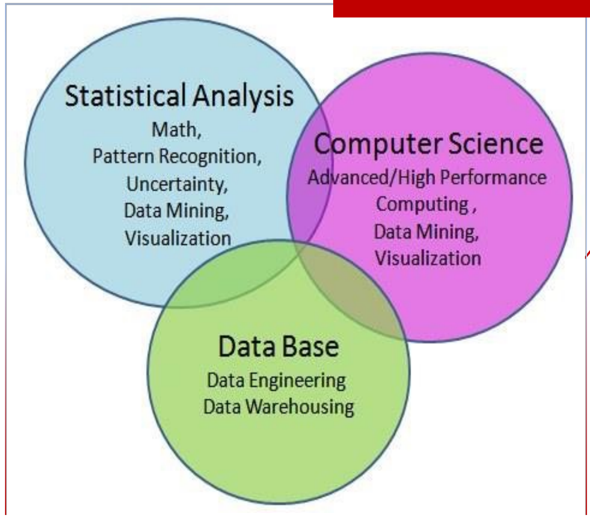


CONWAY

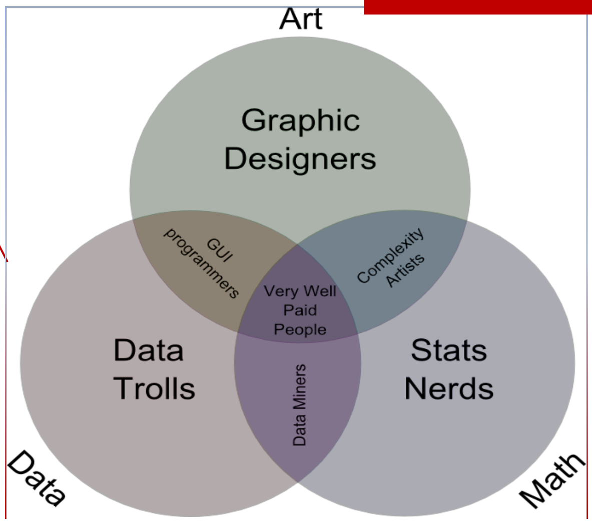


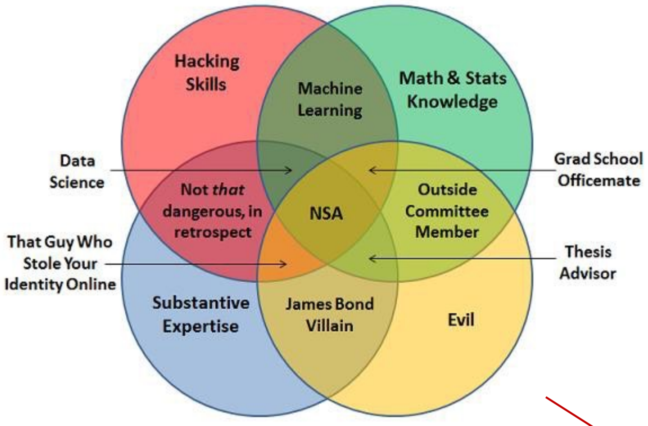
RALEIGH

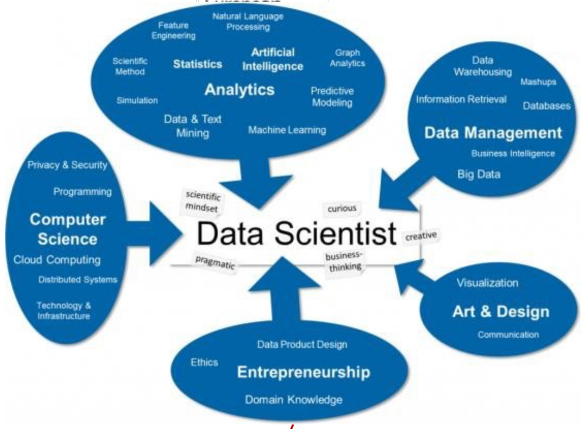
MOUT



ERICKSON







5

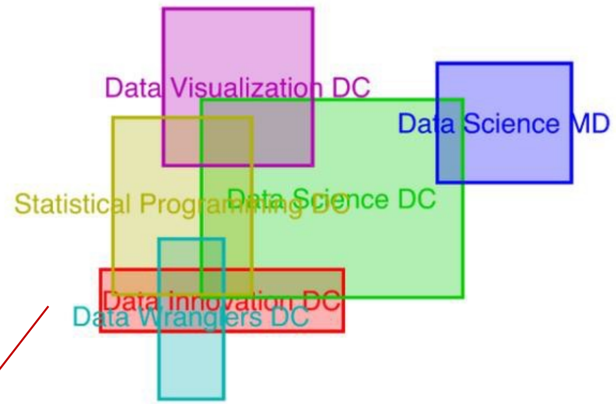
4

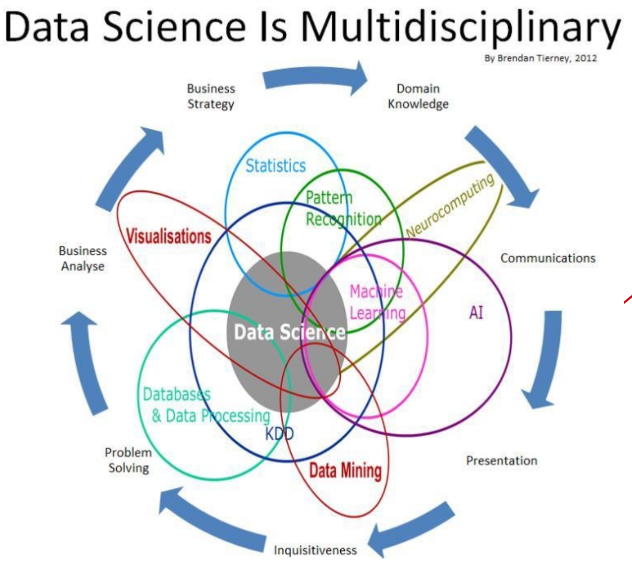
6

Venn diagram

>7

7





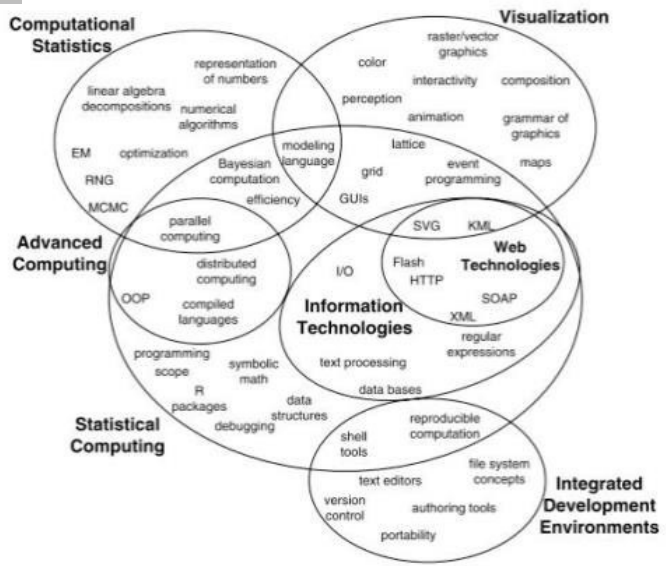
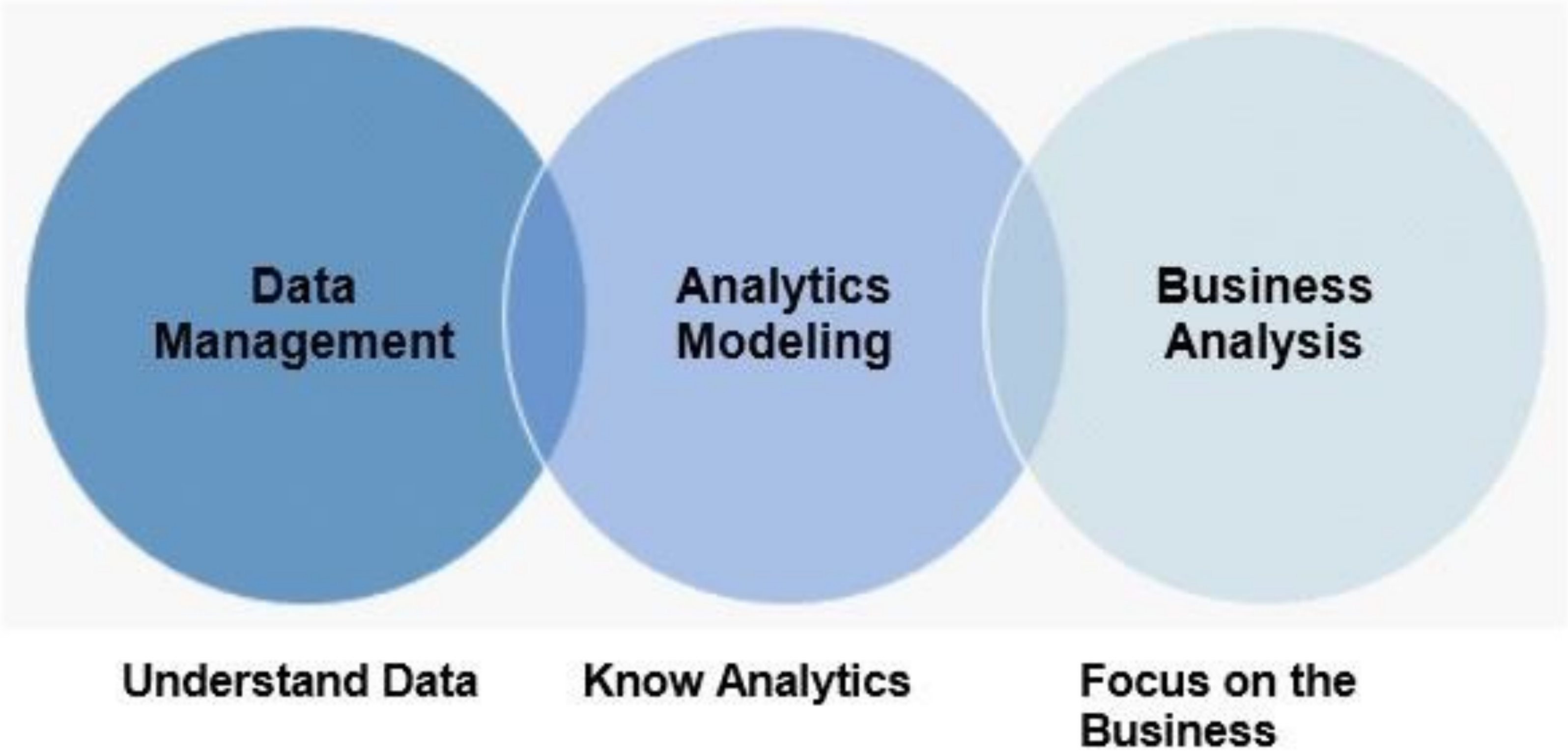


Figure 3. Core Data Scientist Skills



Source: Gartner (March 2012)

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau