Data Mining

DAPT 631

Vishal Patel

Spring 2026

DERIVE
www.derive.io

- Vishal
- I run my own Data Science practice at DERIVE, LLC since 2016
- MS in Computer Science 2003 (IIT, Chicago), and MS in Decision Sciences 2012 (VCU, Richmond)
- Mining data since 2003


Synchronoss
good sam
BERG FREUNDE.DE
ActiveCampaign
citi
Michaels
The Container Store
AMERICA'S TEST KITCHEN
BOSS HUGO BOSS
response LABS
synchrony FINANCIAL
MOEN
Neutrogena
biogen idec
Verdata
SoFi
BABYLON harvested here.
Virgin mobile
PERFETTI van Melle
McDonald's
PETRO HOME SERVICES
IBM
KEURIG GREEN MOUNTAIN
Pizza Hut
BLOCKBUSTER
SEARS
AARP
Sprint
JAGUAR
Michelob ULTRA
Humana
sanofi aventis
NEXTEL

- Introduction
- History
- Course Structure

AI: Hype or Tripe?

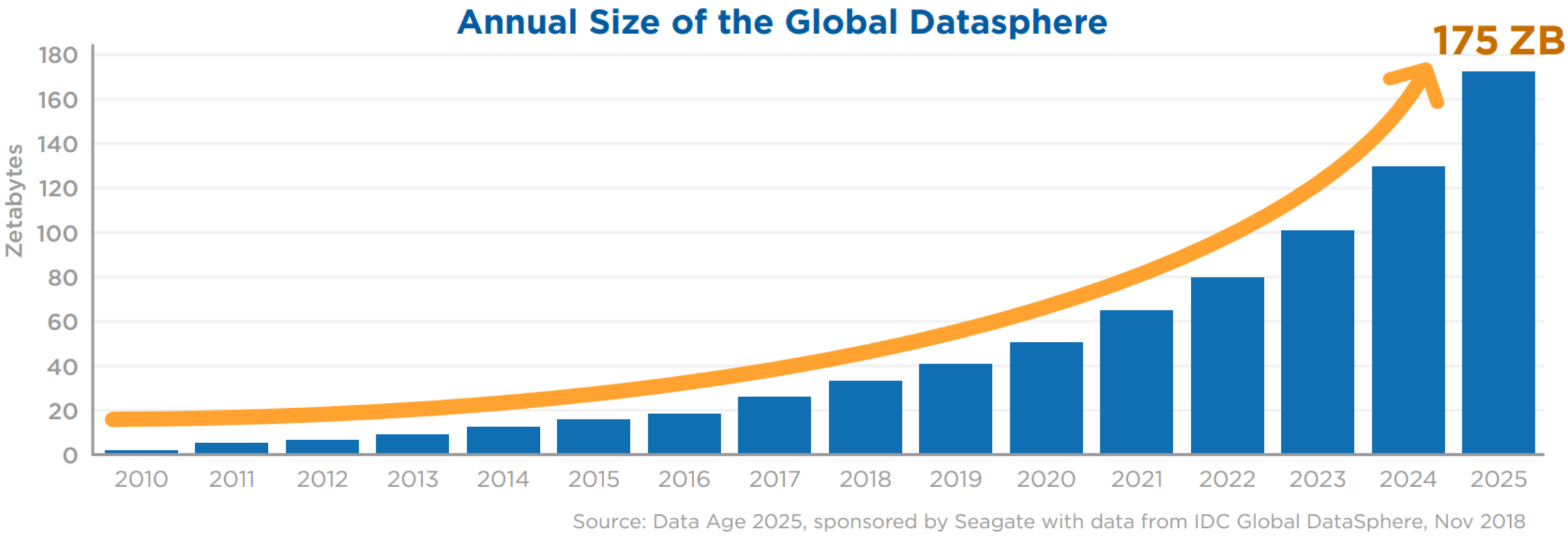
How do you feel about AI?

A. Excited
B. Concerned
C. Both
D. Other

Introduction

- History
- Course Structure

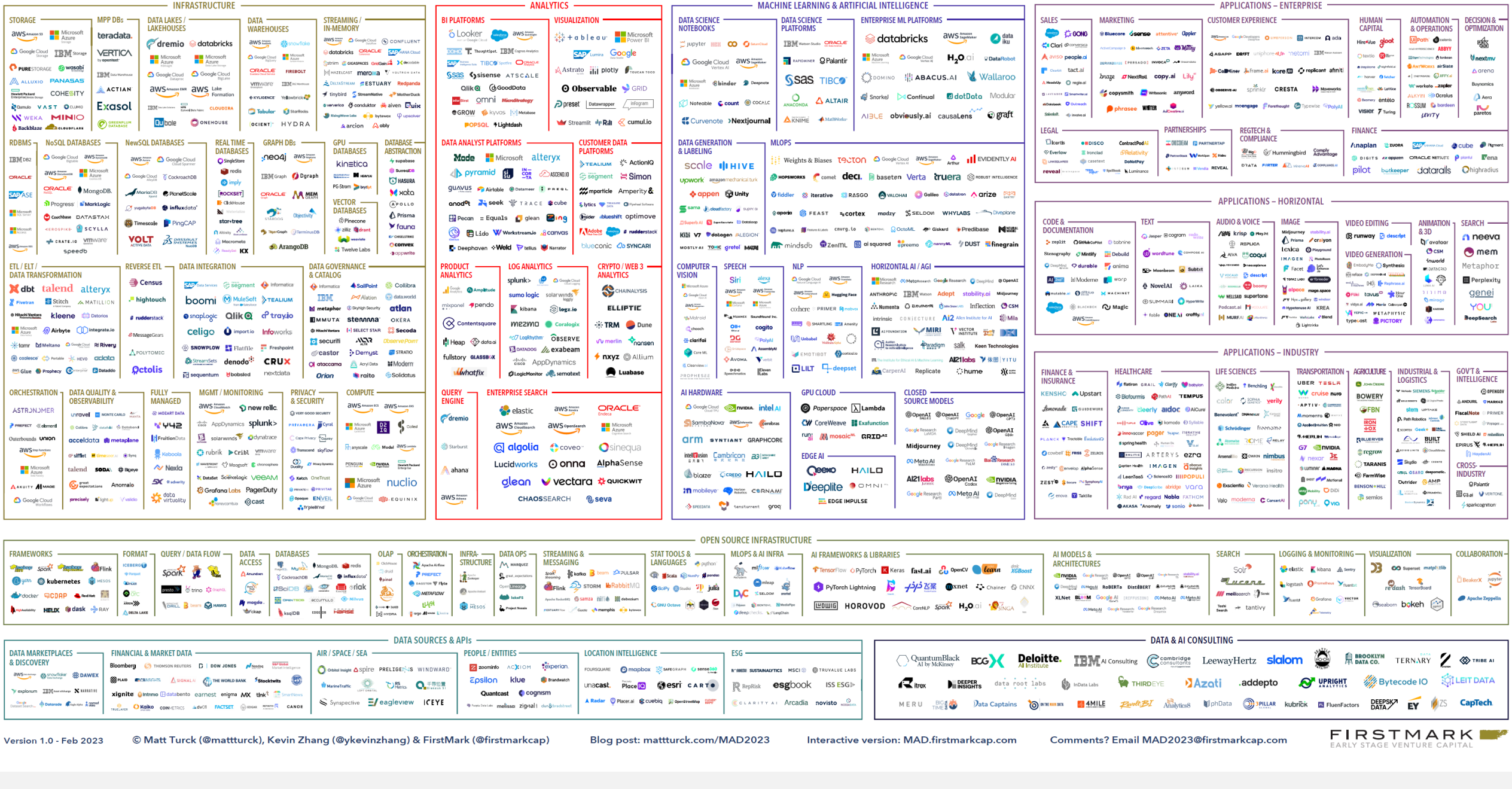

Cambrian Era



One zettabyte = One trillion gigabytes = One sextillion (10^{21}) bytes

For comparison, the universe is 4×10^{17} seconds old.

THE 2023 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE



Version 1.0 - Feb 2023 © Matt Turck (@mattturck), Kevin Zhang (@ykevinzhang) & FirstMark (@firstmarkcap) Blog post: mattturck.com/MAD2023 Interactive version: MAD.firstmarkcap.com Comments? Email MAD2023@firstmarkcap.com

FIRSTMARK EARLY STAGE VENTURE CAPITAL

Robert Cohn, Peter Wang | Keynote: How Open Data Science Opens the World of Innovation

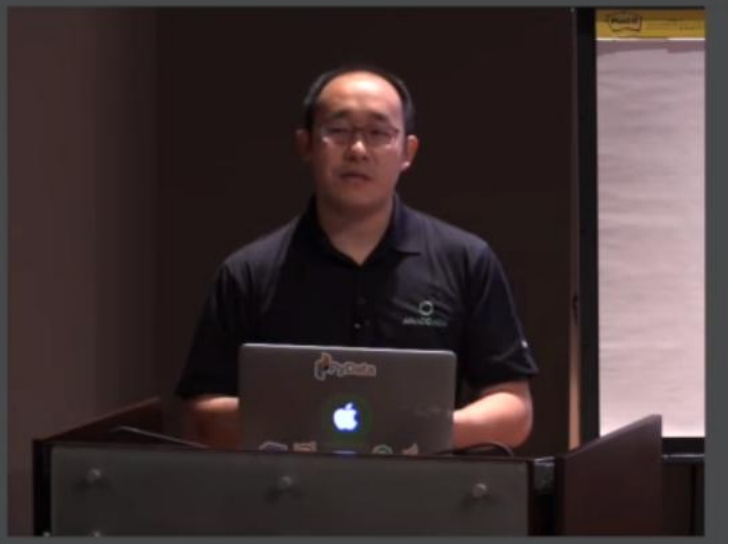
Watch later Share

CONTINUUM ANALYTICS

Era of Data Literacy

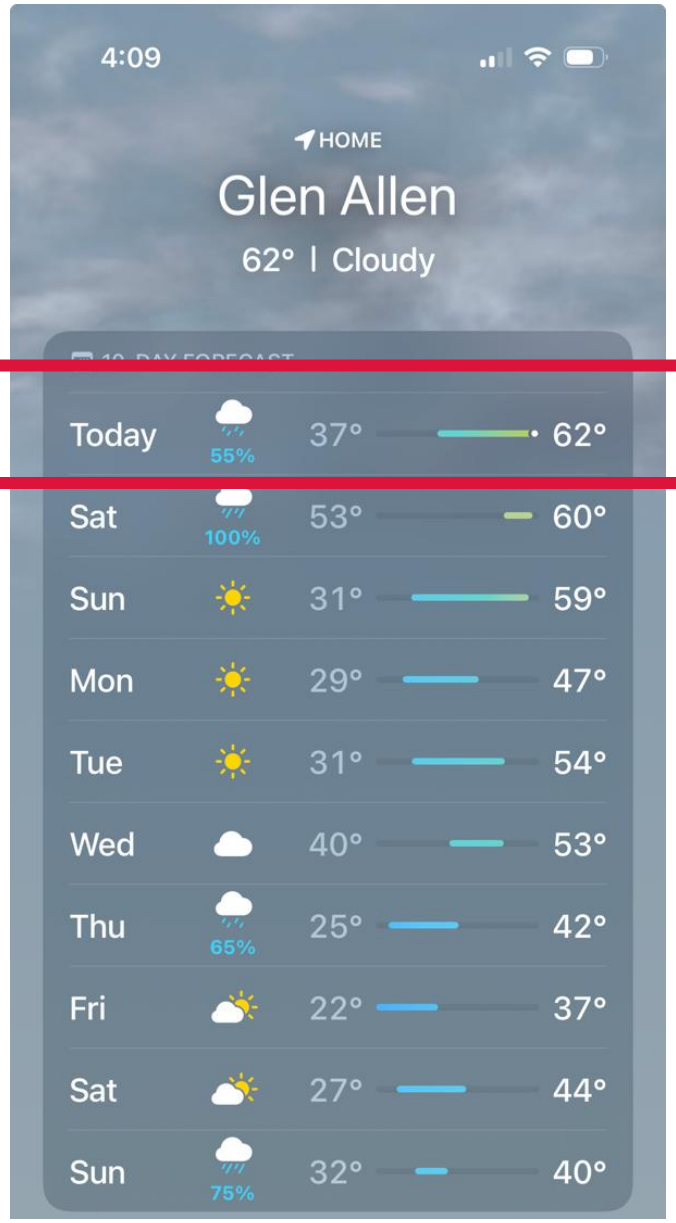
- Data exploration and analysis are going to be a new kind of **literacy** that will be required to do great work in any field.
- Language is a human instinct and is a natural path to insight. We see this in our interaction with Python/PyData users, whose passion chiefly stems from this *expressiveness* and *agility*.
- An analytical language is "**thoughtware**", not "software".

ANACONDA



PyData DC 2016

29:26 / 41:23



55% chance of rain

A. It will rain 55% of the day. ✗

- Rain is not evenly distributed in time.
- Forecasts are binary at a location: it either rains or it doesn't.

B. It will rain over 55% of the city. ✗

- Area coverage can be forecast, but that is a different metric.
- A city could get rain everywhere for 5 minutes and still count as rain.

C. 55% of forecasters think it will rain. ✗

- Forecasts are not opinion polls.
- Modern forecasting is probabilistic modeling, not voting.

D. It will rain moderately. ✗

- Probability ≠ intensity.
- A 55% chance could still mean a thunderstorm.

E. There's a 55% chance it rains where I am. ✓

There is a 55% probability that measurable rain will occur at my location today.

F. It will rain on 55 out of 100 days like today. ✓

If we had 100 days with identical atmospheric conditions, rain would occur on about 55 of those days at this location.

A probabilistic model can be correct even when the outcome surprises you.

Non-analytical thinking asks:

Did it rain or not?

Analytical thinking asks:

Was the probability well-calibrated given uncertainty?

Modern life is guided (governed?) by algorithms that:

- Don’t provide **certainty**
- Provide risk **estimates**
- Require **interpretation**, not blind trust

“Crime rate is down by 10%.”

1. **Timeframe**
   - Compared to what time period? Last month, last year, quarter last year?
   - Was the comparison period abnormal (unemployment, recession, pandemic)?
2. **Definition of “crime”**
   - Which crimes are included? Violent/property/drug-related?
   - Have legal definitions changed over time?
3. **Measurement & reporting**
   - Reported crime or convictions?
   - Did reporting methods change? Online reporting/police staffing/etc.
4. **Geography**
   - Down where? City wide/urban/suburban/nationwide?
5. **Seasonality**
   - Do crime patterns vary by season/holidays/etc.?
6. **Causality**
   - What caused the decrease? How do we know it wasn’t: economic shifts, demographics, etc.?

“Crime rate is down by 10%.”

Under some definition of crime,

in some area,

over some time period,

using some measurement method,

the reported count is 10% lower

than during some comparison period.

1. **Interrogate metrics**

   Numbers are compressed stories. Ask what got compressed away.

2. **Distinguish signal from framing**

   The statistic may be true and still misleading.

3. **Algorithmic literacy**

   Dashboards, models, and KPIs shape decisions: Policing, Lending, Hiring, Healthcare, Education

If you don’t ask *how the number was constructed*,

you are outsourcing thinking.

What is Data Mining?

Data mining is the process of
discovering patterns in large data sets
involving methods at the intersection of
machine learning, statistics, and database systems.

[Wikipedia]

Dictionary

data mining

data mining

noun COMPUTING
noun: **data mining**; noun: **datamining**
the practice of examining large pre-existing databases in order to generate new information.

Data mining is the extraction of
implicit, previously-unknown,
and potentially-useful information from data.

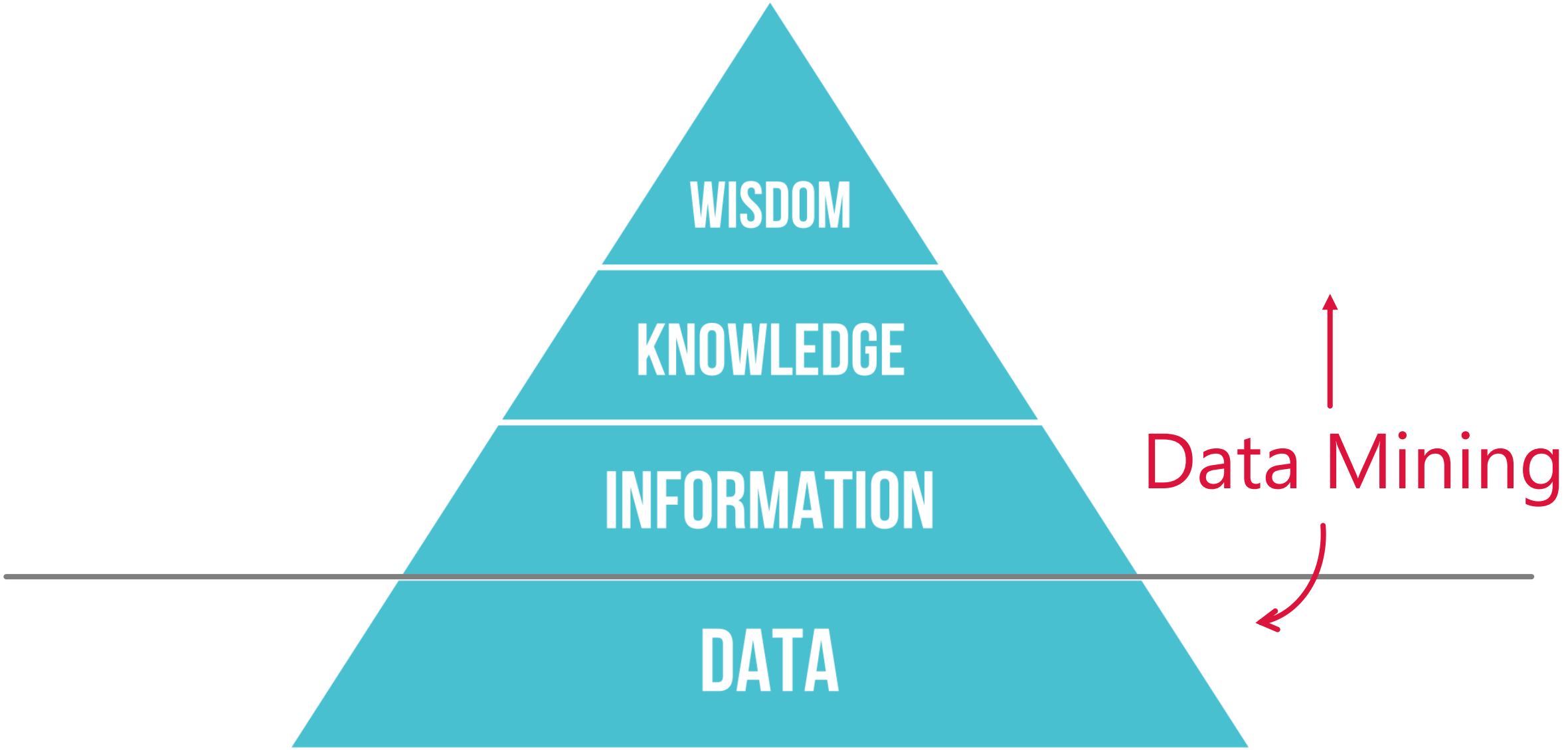
– Witten and Frank

ChatGPT

Data mining is a process of discovering patterns, trends, and valuable insights or knowledge from large volumes of data. It involves using various techniques and algorithms to analyze and extract meaningful information from datasets, often with the goal of making informed business decisions, identifying opportunities, or solving complex problems.

Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting
through large amounts of data stored in repositories,
using pattern recognition technologies as well as statistical and mathematical techniques.

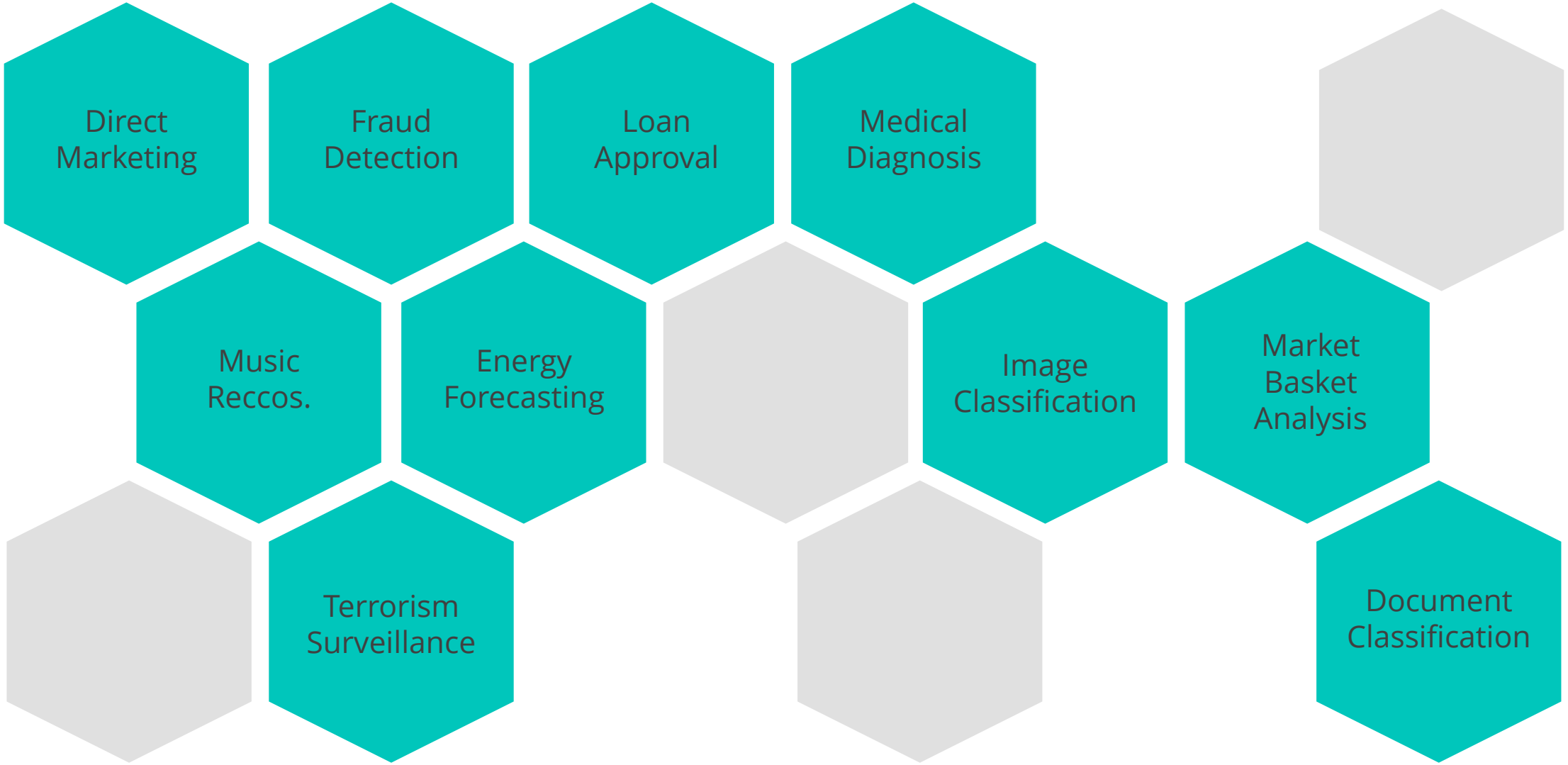
– Gartner


WISDOM
KNOWLEDGE
INFORMATION
DATA
Data Mining


Data
Data Mining
Decision
Past
Future

Data Mining Tasks

Applications of Data Mining



- Introduction
- **History**
- Course Structure

Statistics

Census

Mortality tables

Accounting

From Latin: *status* state

... teaches us what is the political arrangement

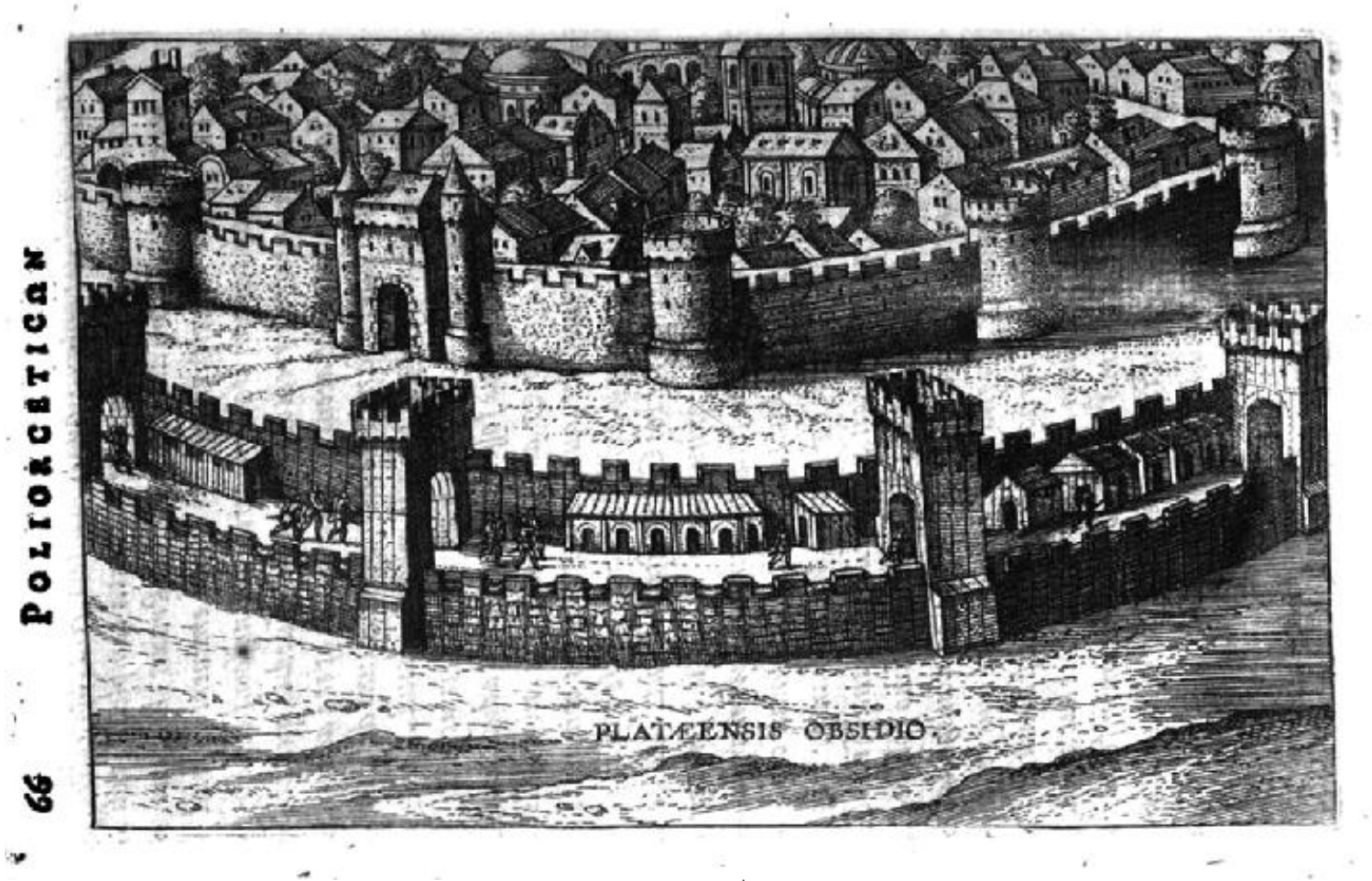
of all modern states of the world.

W Hooper, 1770

DATA COLLECTIONS + ANALYSIS + DECISION MAKING

Statistics

EXAMPLE #1: UNCERTAINTY



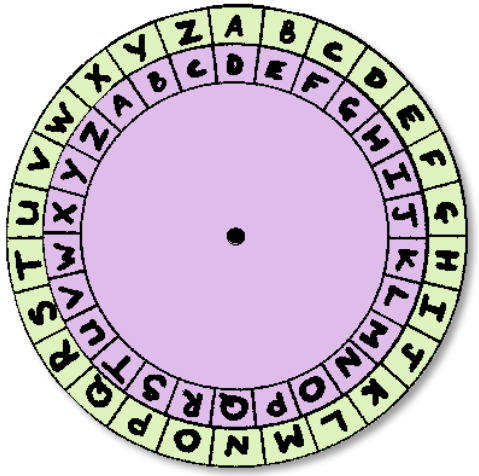
Siege of Plataea (5th Century BCE)

What do you think they used as the best estimate for the height of the wall?

A. Mean
B. Median
C. Mode
D. Max

Statistics

EXAMPLE #2 FREQUENCY ANALYSIS, CRYPTOANALYSIS

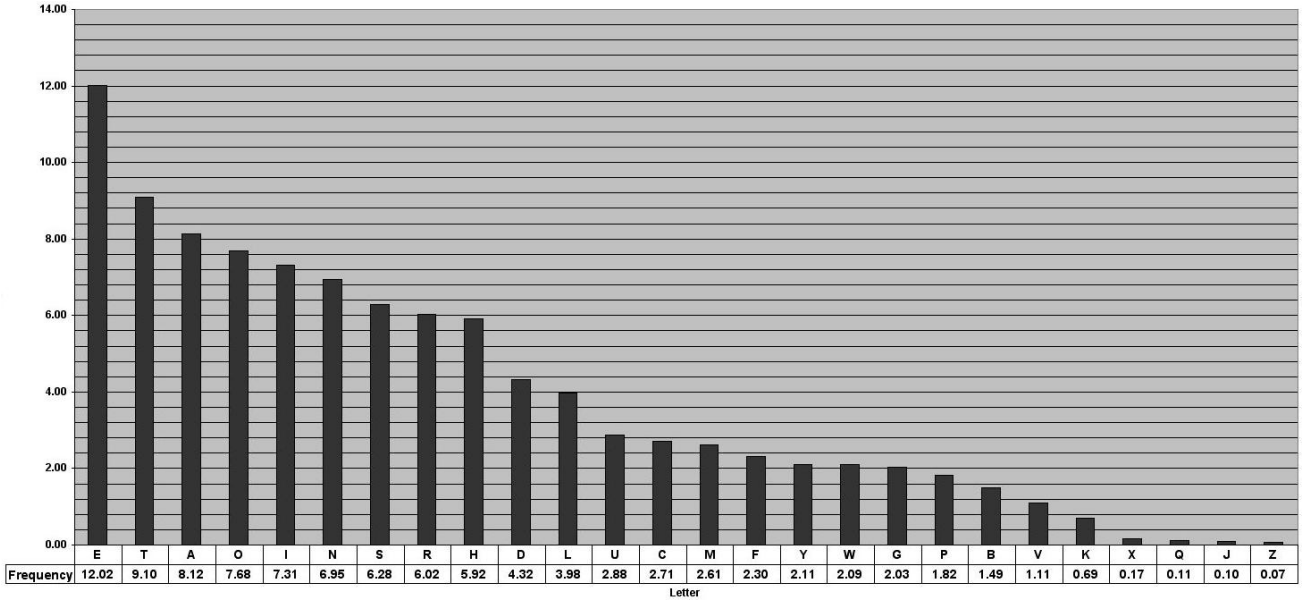


Caesar Cipher

Original message: Et tu, Brute?

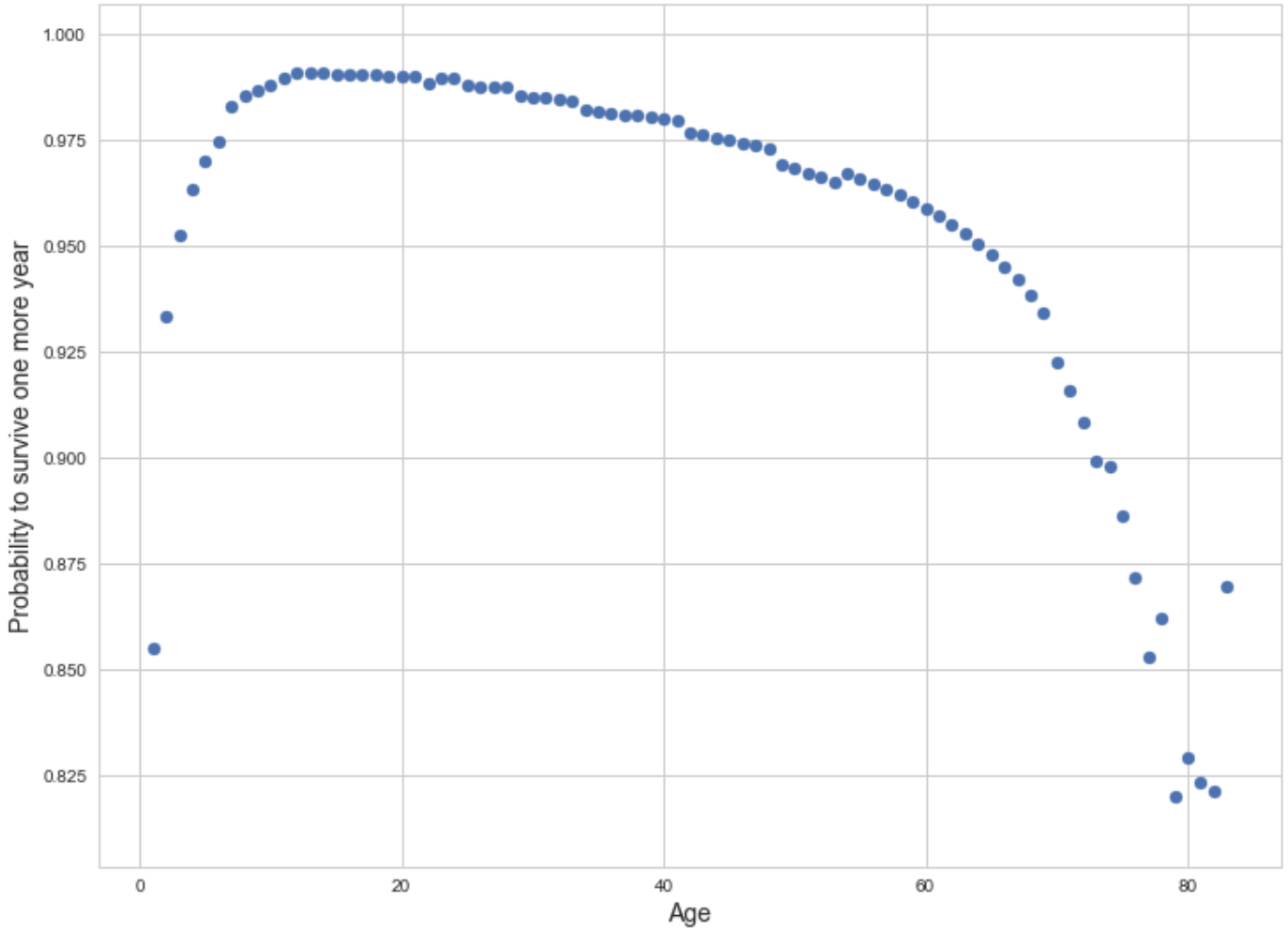
Encrypted message: Hw wx, Euxwh?

Al-Kindi
(801–873 AD)



Statistics

EXAMPLE #3 MORTALITY TABLES, DEMOGRAPHY



Data from Edmond Halley's *An Estimate of the Degrees of Mortality of Mankind (1693)*, table p.600.

The graph shows the probability of surviving one of more year(s) at a certain age.

Modern Statistics

Normal distribution

t distribution

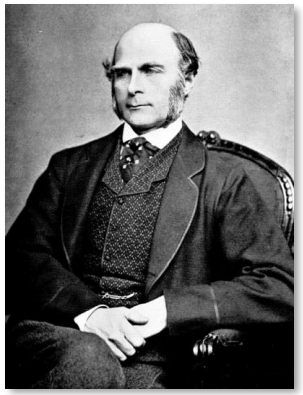
Random sampling

Design of-experiments

Bayesian Statistics

A rigorous mathematical discipline

for analysis, decision making, and inference



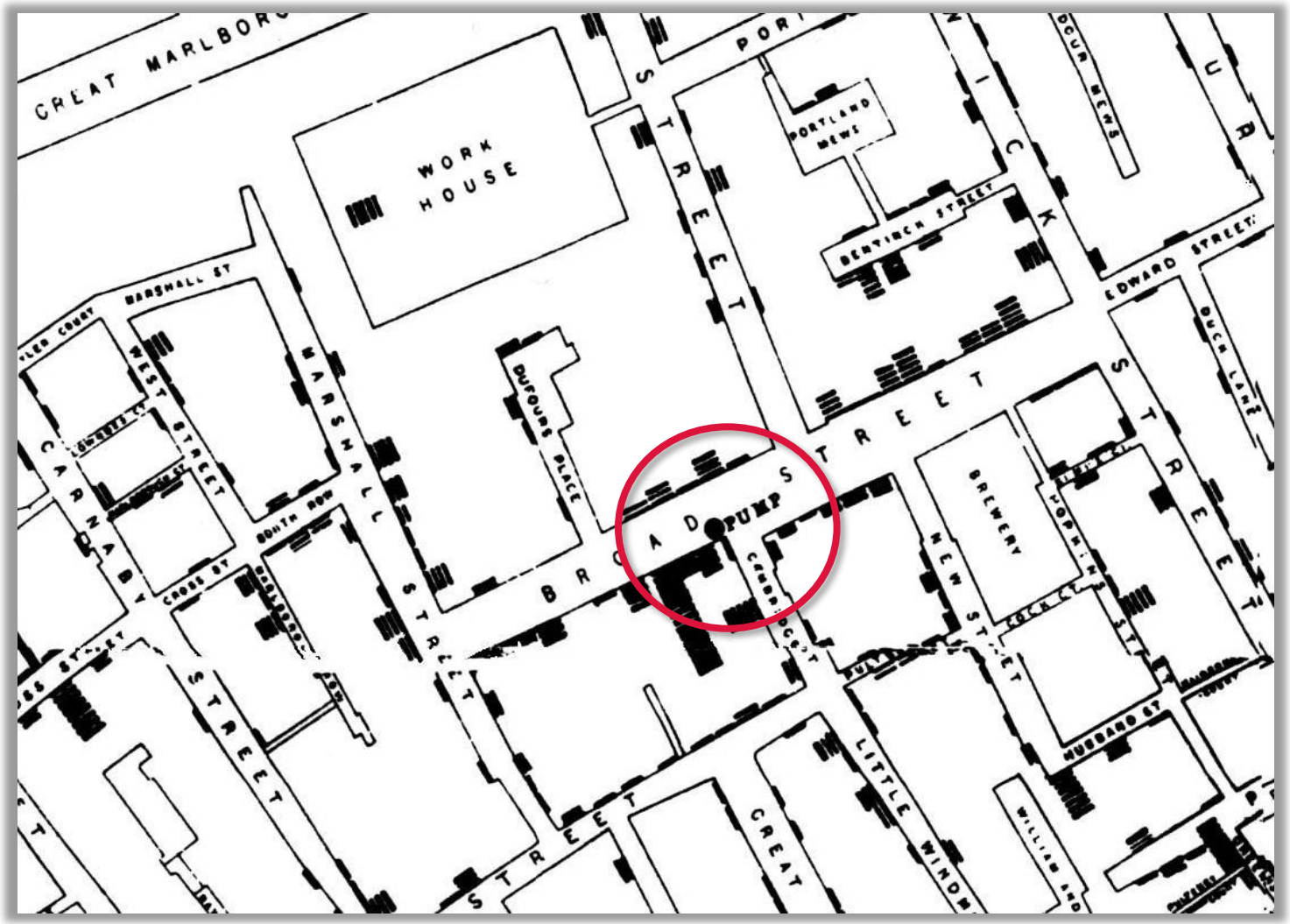
Sir Francis Galton
(1822–1911)
Correlation, regression

Carl Pearson
(1857–1936)
Founder of mathematical statistics

R A Fisher
(1890–1962)
ANOVA, Maximum Likelihood, DOE

Modern Statistics

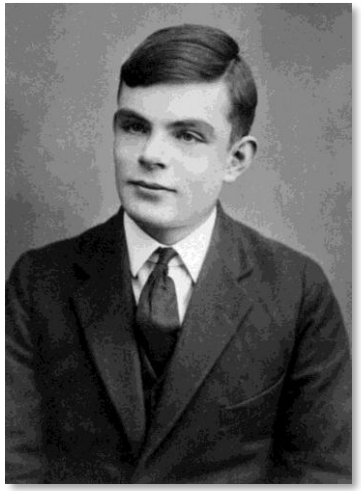
EXAMPLE: DATA VISUALIZATION



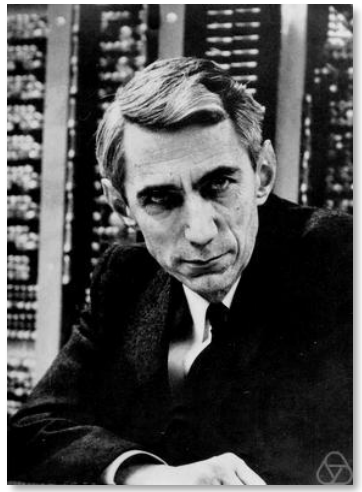
Original map by **John Snow** showing the clusters of cholera cases in the **London epidemic of 1854** [Source]

Data Mining

Algorithms &
Computation
Computer Science
Neural Networks
Decision Trees
Genetic Algorithms
Relational Databases



Alan Turing
(1912 –1954)
Theoretical Computer Science



Claude Shannon
(1916 –2001)
Information Theory

- Warren McCulloch and Walter Pitts created a computational model for neural networks. (1943)
- John Holland introduced Genetic Algorithm based on the concept of Darwin's theory of evolution. (1960)
- E. F. Codd published an important paper to propose the use of a relational database model. (1970)

Data Science

Gradient Boosting

Random Forests

Support Vector-Machines

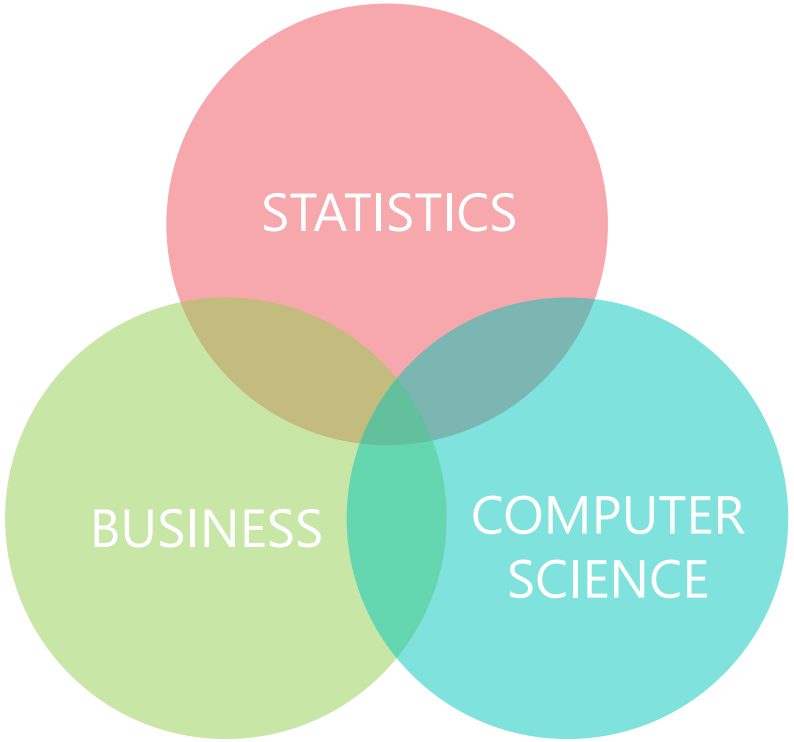
Recommender-systems

Unstructured data

Open source

Big Data

Data science is an interdisciplinary field
that uses scientific methods, processes, algorithms and systems
to extract knowledge and insights
from data in various forms,
both structured and unstructured,
similar to data mining.†



† From Wikipedia

Artificial Intelligence

Deep learning

Reinforcement-Learning

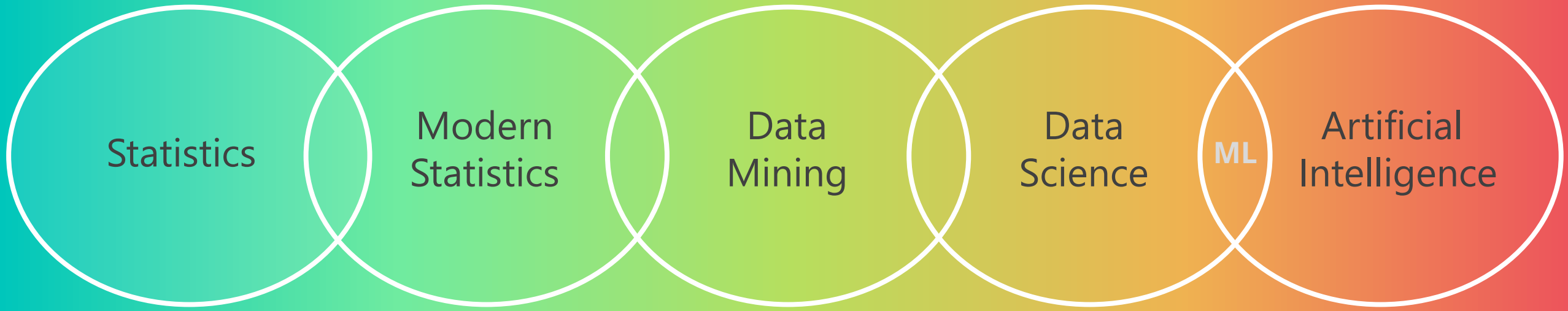
Speech recognition

Natural Language-Processing

Computer vision

Jan 2023

Jun 2023



Regression
Correlation
Frequency analysis
Descriptive statistics
ANOVA

Normal distribution
t distribution
Random sampling
Design of Experiments
Bayesian statistics

Algorithms & Computation
Computer Science
Neural Networks
Decision trees
Genetic algorithms
Relational Databases

Gradient Boosting
Random Forests
Support Vector Machines
Recommender systems
Unstructured data
Open source
Big Data

Deep learning
Reinforcement-Learning
Speech recognition
Natural Language-Processing
Computer vision

Prehistory – 18th Century | Late 19th / Early 20th Century | Mid-Late 20th Century | 21st Century

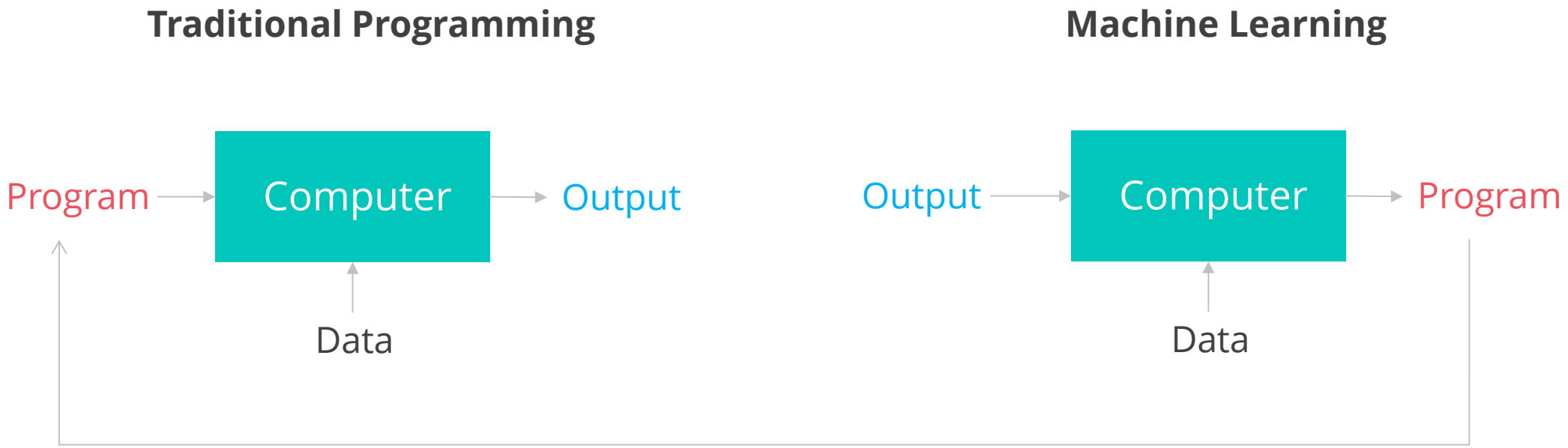
Calculations by hand → ***Distributed computing***

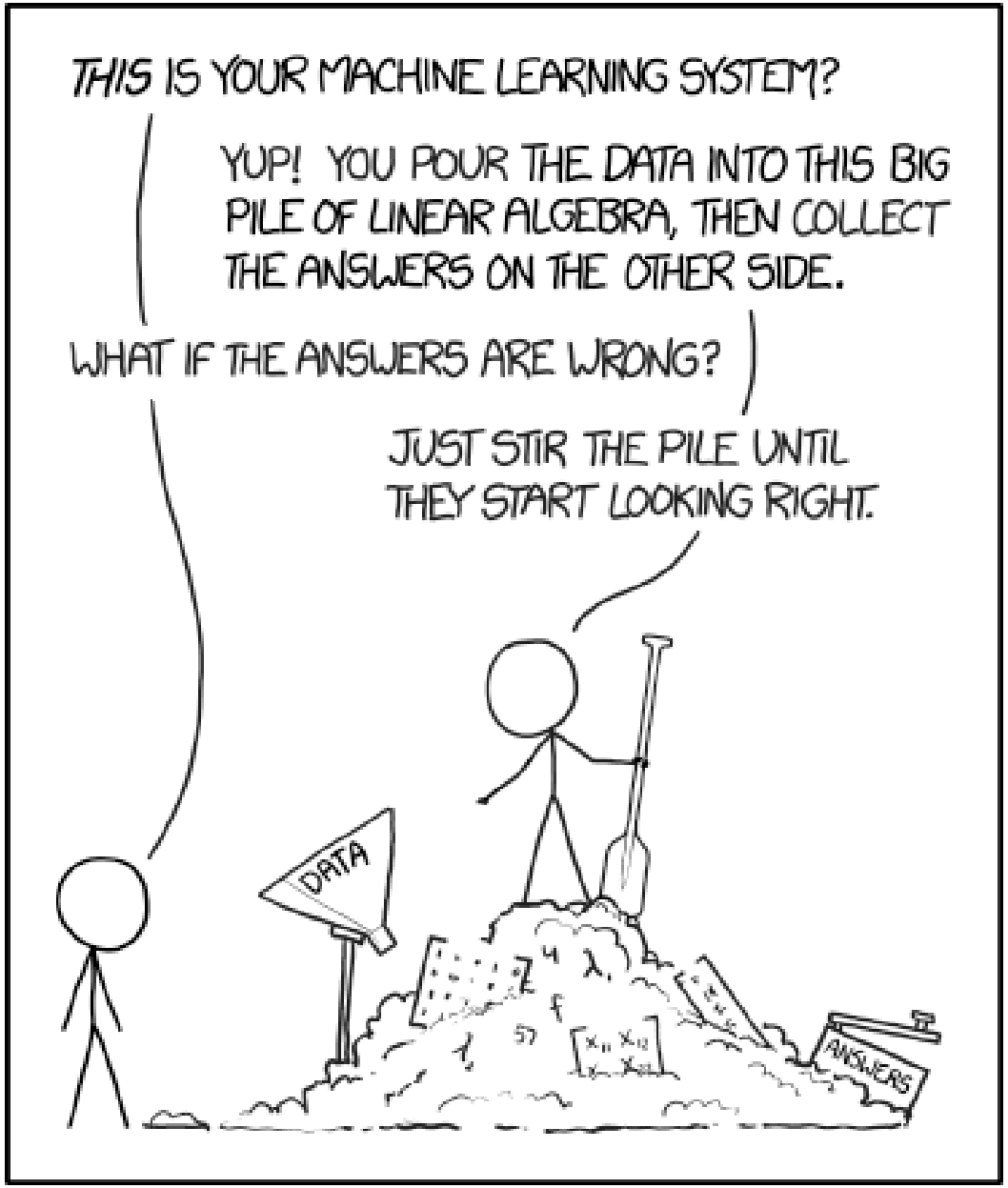
Evolution of techniques and technology

Machine Learning

Field of study that gives computers the ability to learn without being explicitly programmed.

Artur Samuel,1959




THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

- Introduction
- History

- **Course Structure**

Data Science ≈ Data Mining

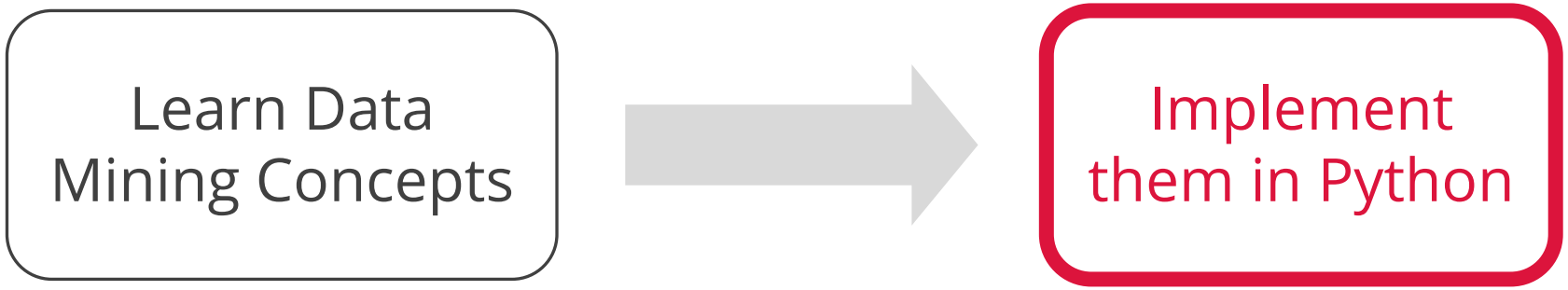
- The specific definitions and boundaries between these disciplines remain fuzzy.
- For the purpose of this class, I will use the terms ‘Data Science’ and ‘Data Mining’ interchangeably (with a preference to the former).
- We will cover several Data Science techniques in this class, e.g., Gradient Boosting.

Course Outline

1. Introduction
2. The Data Science Process
3. Supervised Learning
4. Unsupervised Learning
5. Wrap Up

1. Ask questions at any time!
2. Collaboration is encouraged.
3. All course content will be available on Canvas and GitHub.
4. Data Mining + Python
5. Homework assignments in Python

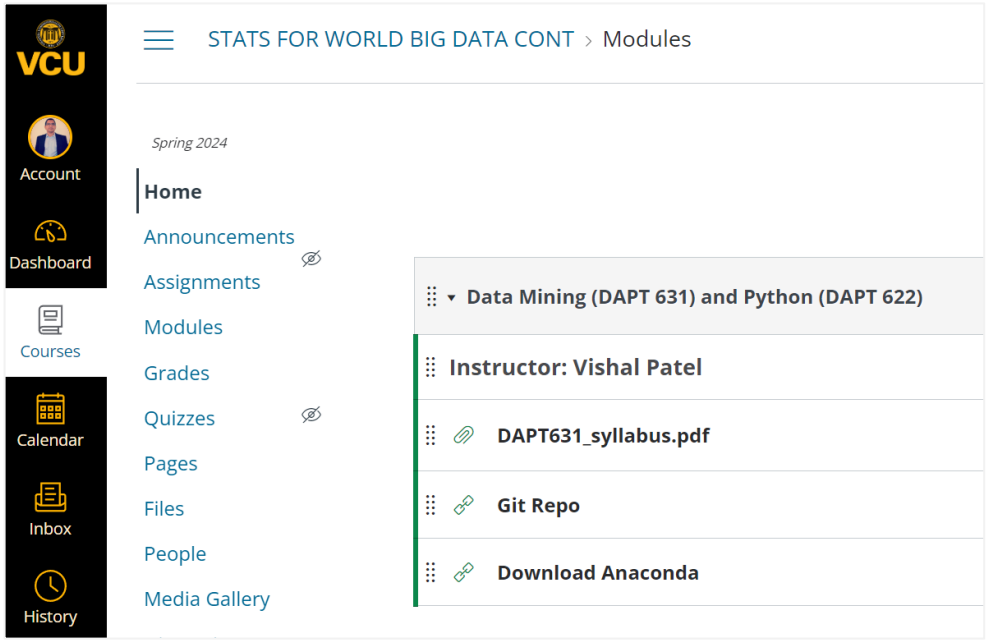
Data Mining + Python



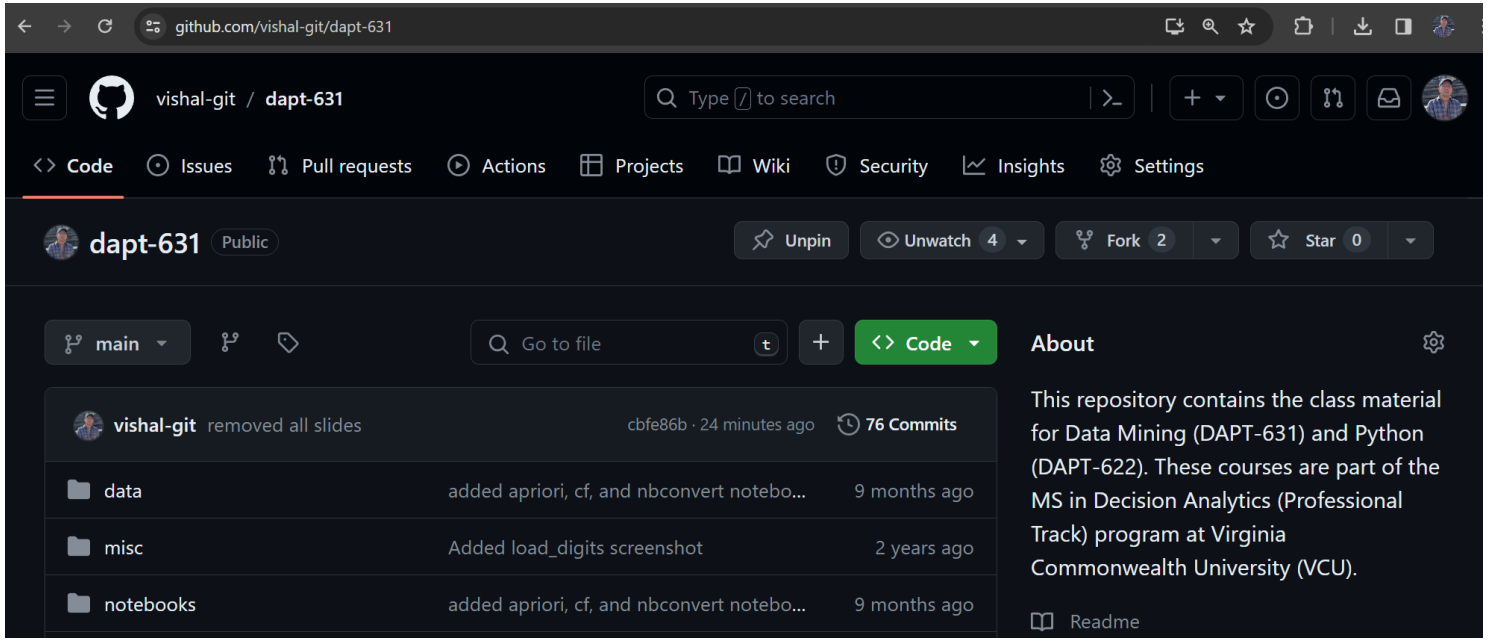
Course Objectives

1. Provide a practical knowledge of data mining algorithms.
2. Give a broader perspective to help understand what role data mining plays in the decision-making process.
3. Help you develop an appreciation for the beauty of the theoretical foundations underlying data mining.
4. Help you think more like a Data Scientist.
5. (For myself) Continue learning.

Course Material



VCU Canvas



GitHub

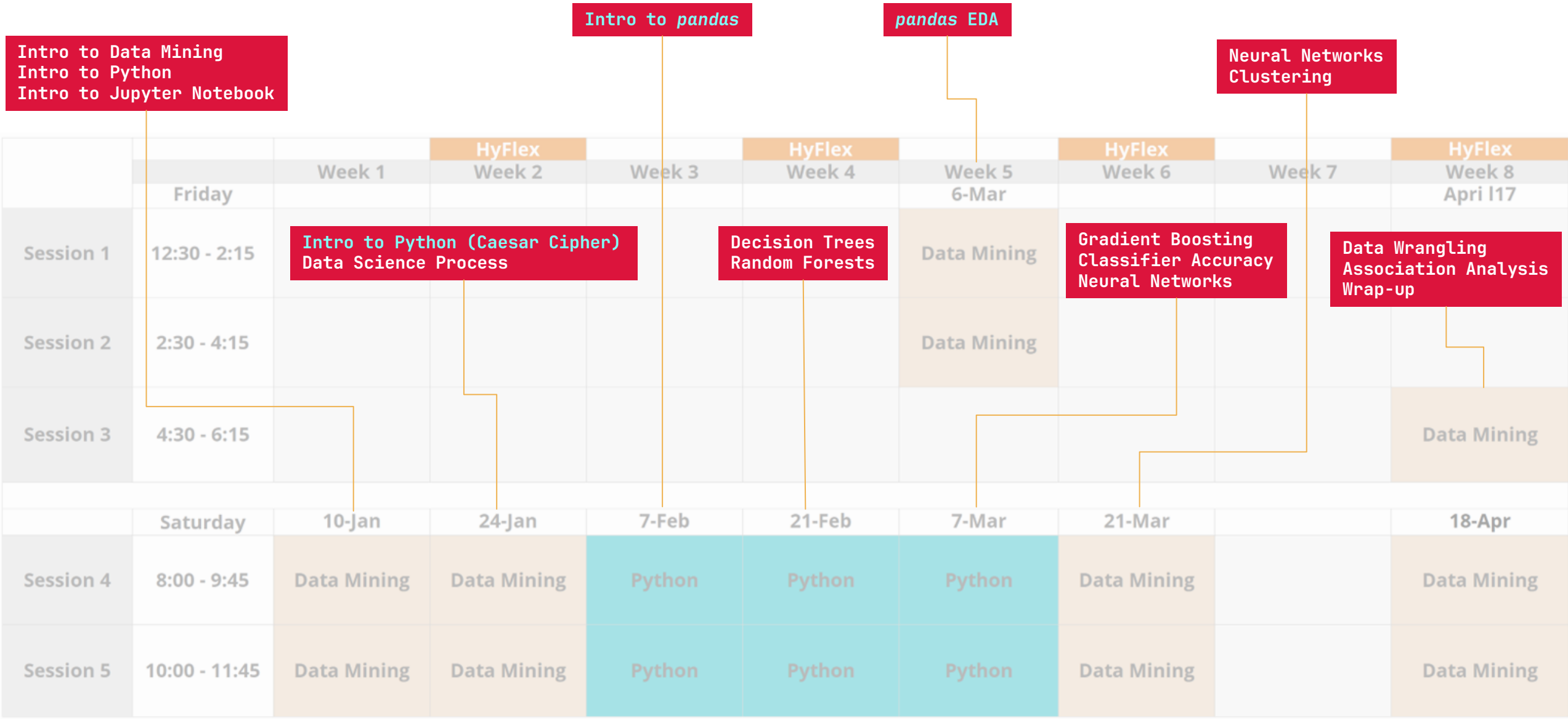
			HyFlex		HyFlex		HyFlex		HyFlex
		Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8
	Friday					6-Mar			Apri l17
Session 1	12:30 - 2:15					Data Mining			
Session 2	2:30 - 4:15					Data Mining			
Session 3	4:30 - 6:15								Data Mining
	Saturday	10-Jan	24-Jan	7-Feb	21-Feb	7-Mar	21-Mar		18-Apr
Session 4	8:00 - 9:45	Data Mining	Data Mining	Python	Python	Python	Data Mining		Data Mining
Session 5	10:00 - 11:45	Data Mining	Data Mining	Python	Python	Python	Data Mining		Data Mining

~30 hours!

What We Will Cover

1. Introduction to Data Mining
2. The Data Science Process
3. Introduction to Regression
4. Linear Regression model
5. Build a model using partitioning
6. Decision trees
7. Classification Trees
8. Random Forests
9. Gradient Boosting Trees
10. Hyper-parameter optimization
11. Introduction to Neural Networks
12. Introduction to Clustering
13. Agglomerative Clustering
14. k-means Clustering
15. DBSCAN Clustering
16. PCA
17. Association Analysis (Apriori algorithm)
18. Collaborative Filtering
19. Data Wrangling

+ 20 Jupyter Notebooks



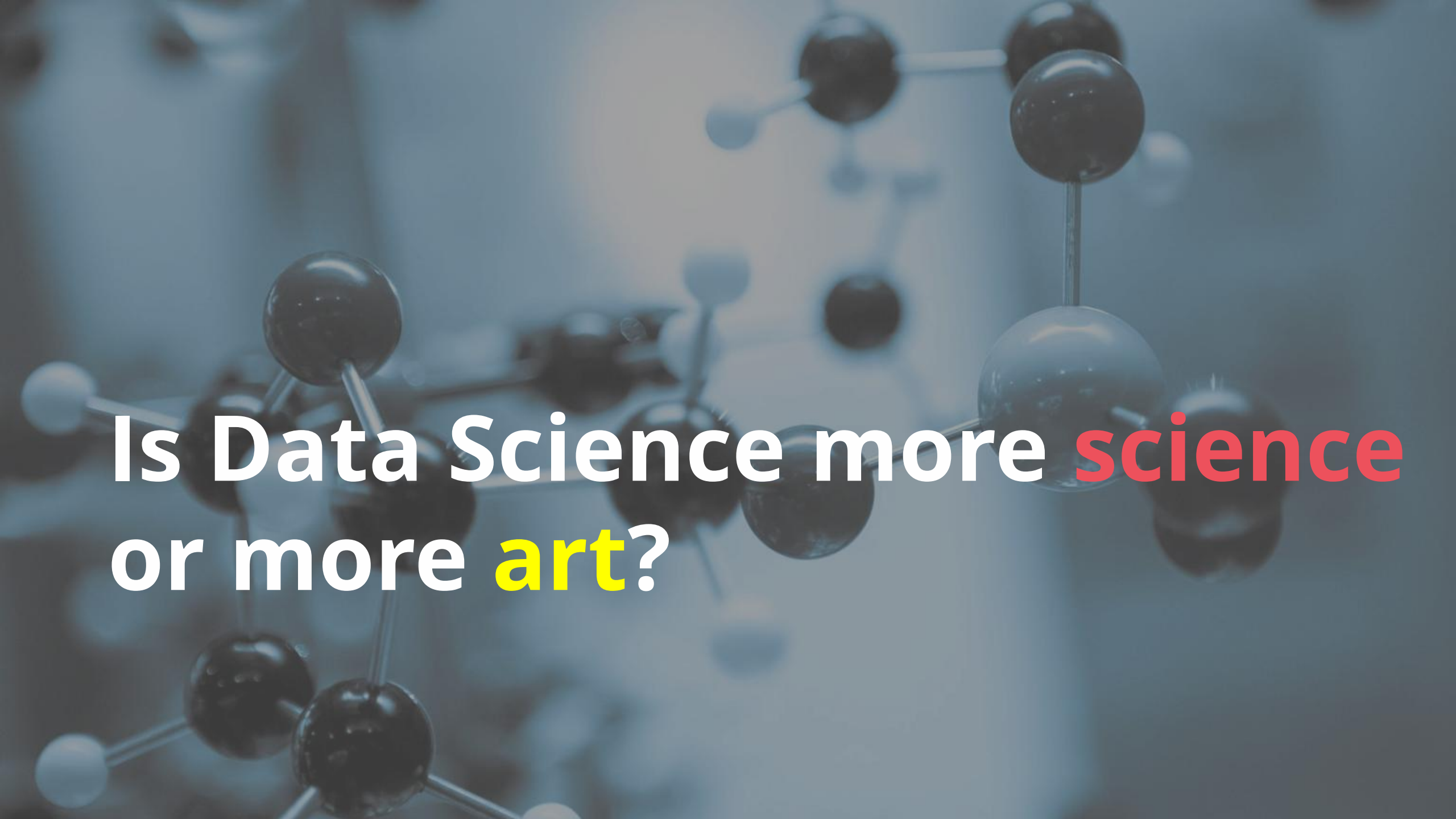
Homework Assignments:

4 Data Mining (DAPT-631)
4 Python (DAPT-618)

vishal@derive.io

www.linkedIn.com/in/VishalJP

@derive_io


Is Data Science more science
or more art?

The Science Side

- **Rigorous Methods:** Data science relies on mathematical models, statistical techniques, and algorithms to extract insights. This part is objective, measurable, and replicable.
- **Experimentation:** Like any scientific field, it involves hypothesis testing, controlled experiments, and iterative refinement.
- **Tools and Technology:** Programming languages (Python, R), frameworks, and statistical software form the backbone of its scientific nature.
- **Data-Driven Decisions:** The focus is on evidence-based conclusions.

The Art Side

- **Problem Framing:** Choosing the right questions to ask and translating real-world problems into data problems requires creativity and intuition.
- **Feature Engineering:** Selecting and transforming data features often involves a deep understanding of the domain, combined with a bit of trial and error.
- **Visualization and Storytelling:** Crafting narratives and visual representations of data is an art that requires empathy for the audience and an ability to simplify complexity.
- **Trade-offs:** Knowing when to stop fine-tuning models or balancing accuracy with interpretability involves subjective judgment.

Verdict: Both!

- It's **science** when you're working with models, data pipelines, and algorithms.
- It's **art** when applying context, making sense of results, or communicating insights effectively.

Source: ChatGPT