Data Mining

DAPT 631

Vishal Patel

Spring 2023

DERIVE
www.derive.io

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

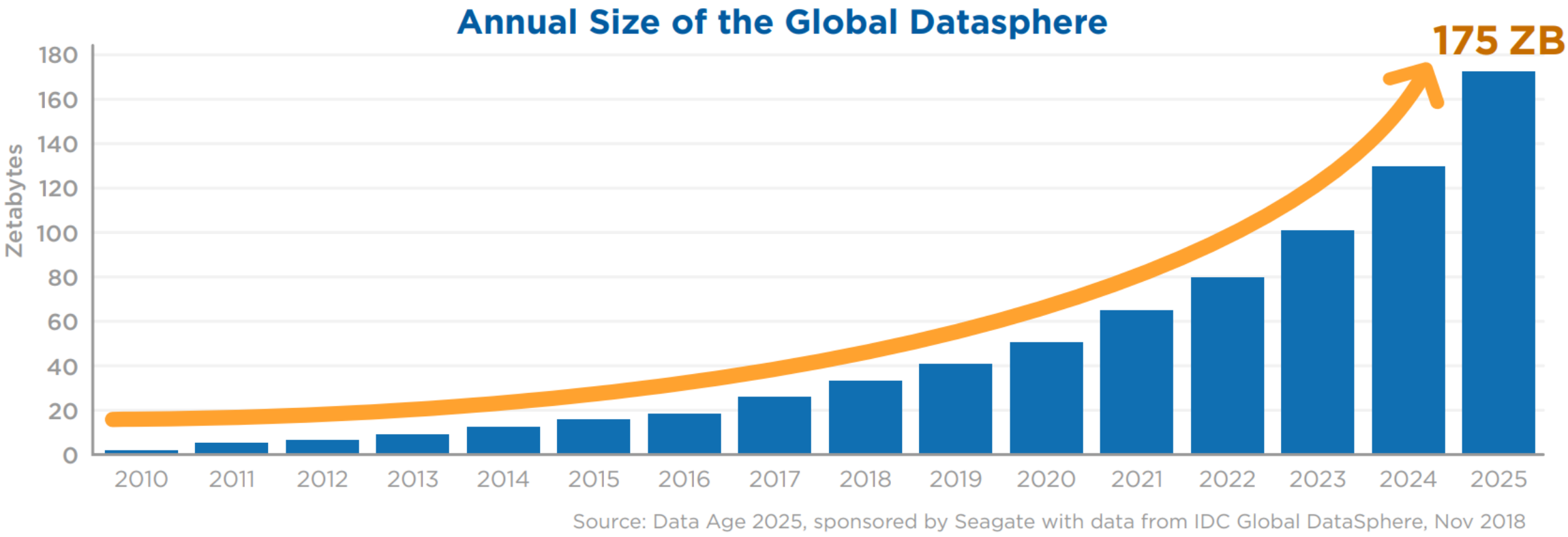
Synchronoss
BERG FREUNDE.DE
ActiveCampaign
response LABS
Michaels
MOEN
AMERICA'S TEST KITCHEN
BOSS HUGO BOSS
Verdata
The Container Store
Neutrogena
biogen idec
PERFETTI van Melle
SoFi
BABYLON harvested here.
Virgin mobile
McDonald's
PETRO HOME SERVICES
KEURIG GREEN MOUNTAIN
Pizza Hut
IBM
BLOCKBUSTER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

- Introduction
- History
- Course Structure

Introduction

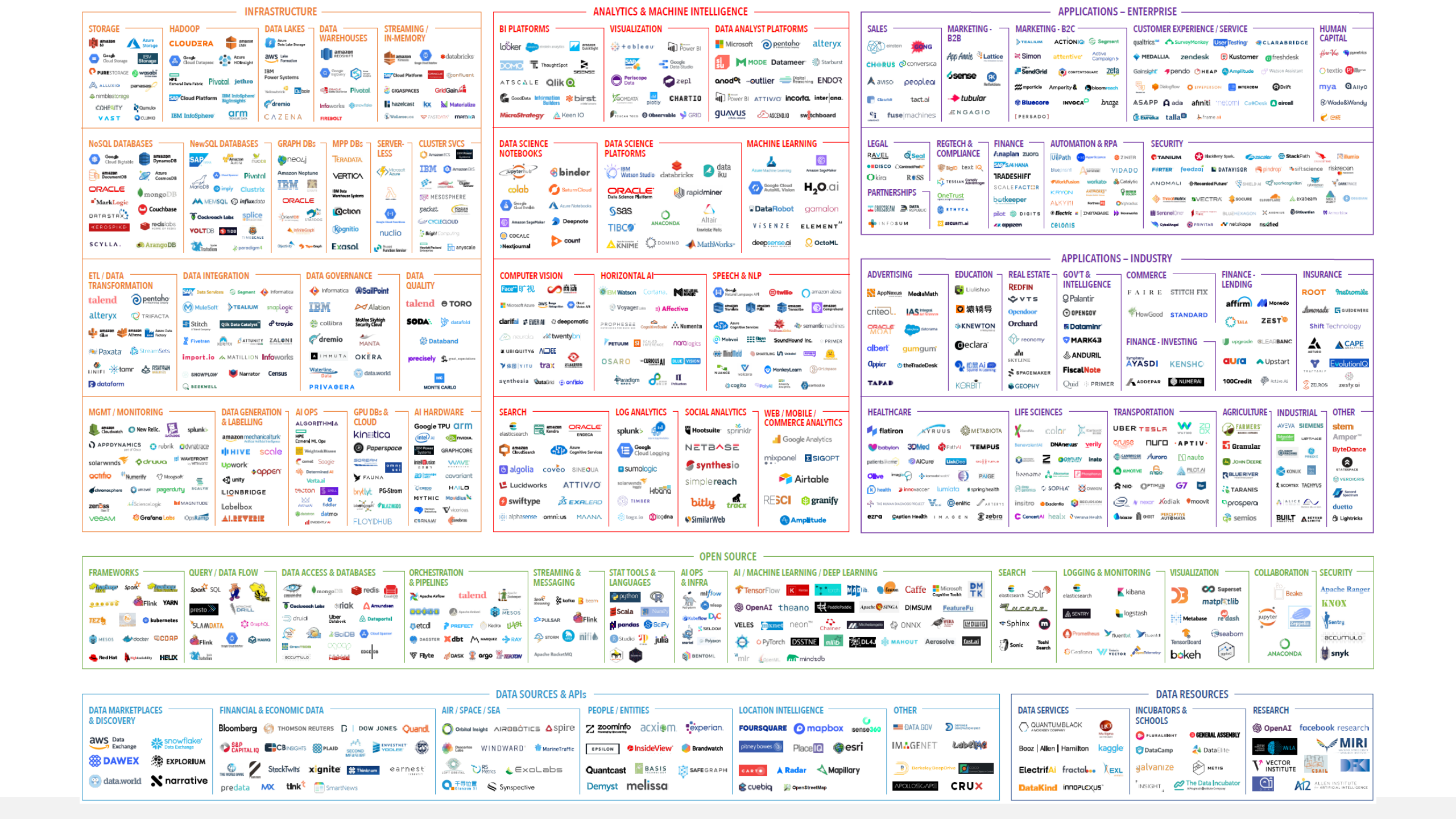
- History
- Course Structure

Cambrian Era



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

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


INFRASTRUCTURE
STORAGE
HADOOP
DATA LAKES
DATA WAREHOUSES
STREAMING / IN-MEMORY
NoSQL DATABASES
NewSQL DATABASES
GRAPH DBs
MPP DBs
SERVER-LESS
CLUSTER SVCS
ETL / DATA TRANSFORMATION
DATA INTEGRATION
DATA GOVERNANCE
DATA QUALITY
MGMT / MONITORING
DATA GENERATION & LABELLING
AI OPS
GPU DBs & CLOUD
AI HARDWARE
ANALYTICS & MACHINE INTELLIGENCE
BI PLATFORMS
VISUALIZATION
DATA ANALYST PLATFORMS
DATA SCIENCE NOTEBOOKS
DATA SCIENCE PLATFORMS
MACHINE LEARNING
COMPUTER VISION
HORIZONTAL AI
SPEECH & NLP
SEARCH
LOG ANALYTICS
SOCIAL ANALYTICS
WEB / MOBILE / COMMERCE ANALYTICS
APPLICATIONS – ENTERPRISE
SALES
MARKETING - B2B
MARKETING - B2C
CUSTOMER EXPERIENCE / SERVICE
HUMAN CAPITAL
LEGAL
REGTECH & COMPLIANCE
FINANCE
AUTOMATION & RPA
SECURITY
PARTNERSHIPS
APPLICATIONS – INDUSTRY
ADVERTISING
EDUCATION
REAL ESTATE
GOV'T & INTELLIGENCE
COMMERCE
FINANCE - LENDING
FINANCE - INVESTING
INSURANCE
HEALTHCARE
LIFE SCIENCES
TRANSPORTATION
AGRICULTURE
INDUSTRIAL
OTHER
OPEN SOURCE
FRAMEWORKS
QUERY / DATA FLOW
DATA ACCESS & DATABASES
ORCHESTRATION & PIPELINES
STREAMING & MESSAGING
STAT TOOLS & LANGUAGES
AI OPS & INFRA
AI / MACHINE LEARNING / DEEP LEARNING
SEARCH
LOGGING & MONITORING
VISUALIZATION
COLLABORATION
SECURITY
DATA SOURCES & APIs
DATA MARKETPLACES & DISCOVERY
FINANCIAL & ECONOMIC DATA
AIR / SPACE / SEA
PEOPLE / ENTITIES
LOCATION INTELLIGENCE
OTHER
DATA RESOURCES
DATA SERVICES
INCUBATORS & SCHOOLS
RESEARCH

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

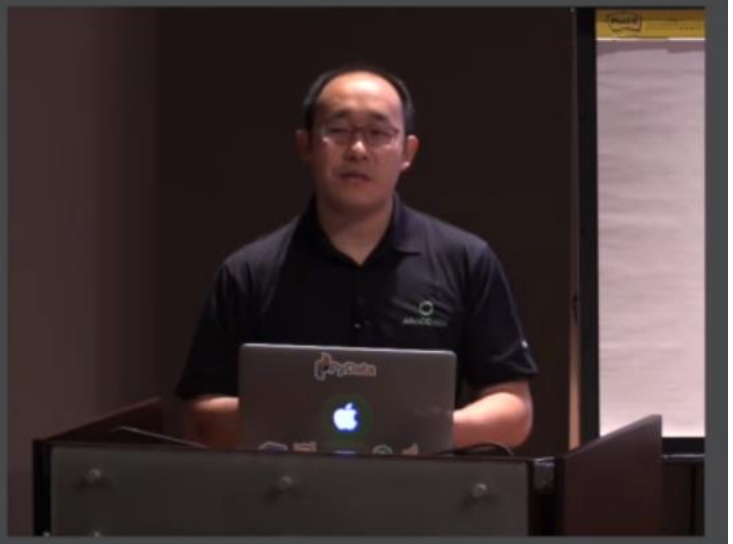
Watch later Share

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

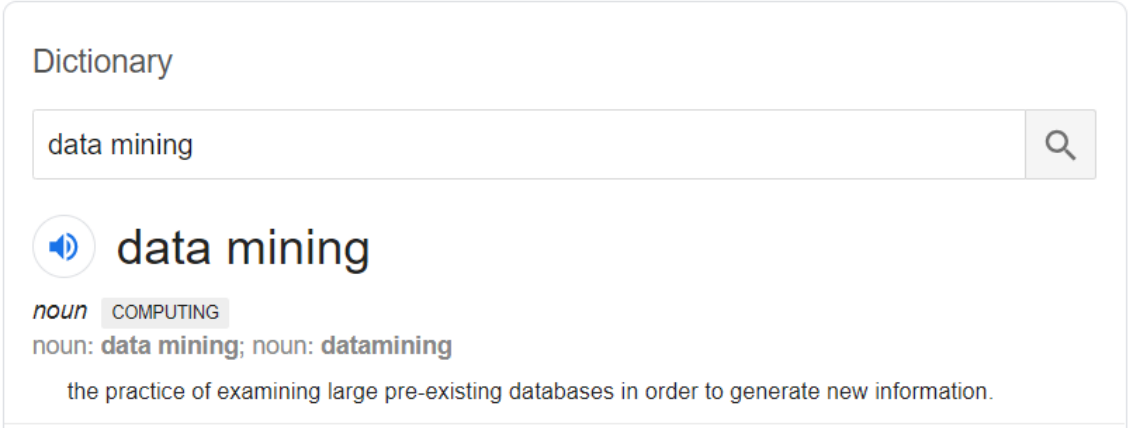
What is Data Mining?

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]

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

– Witten and Frank

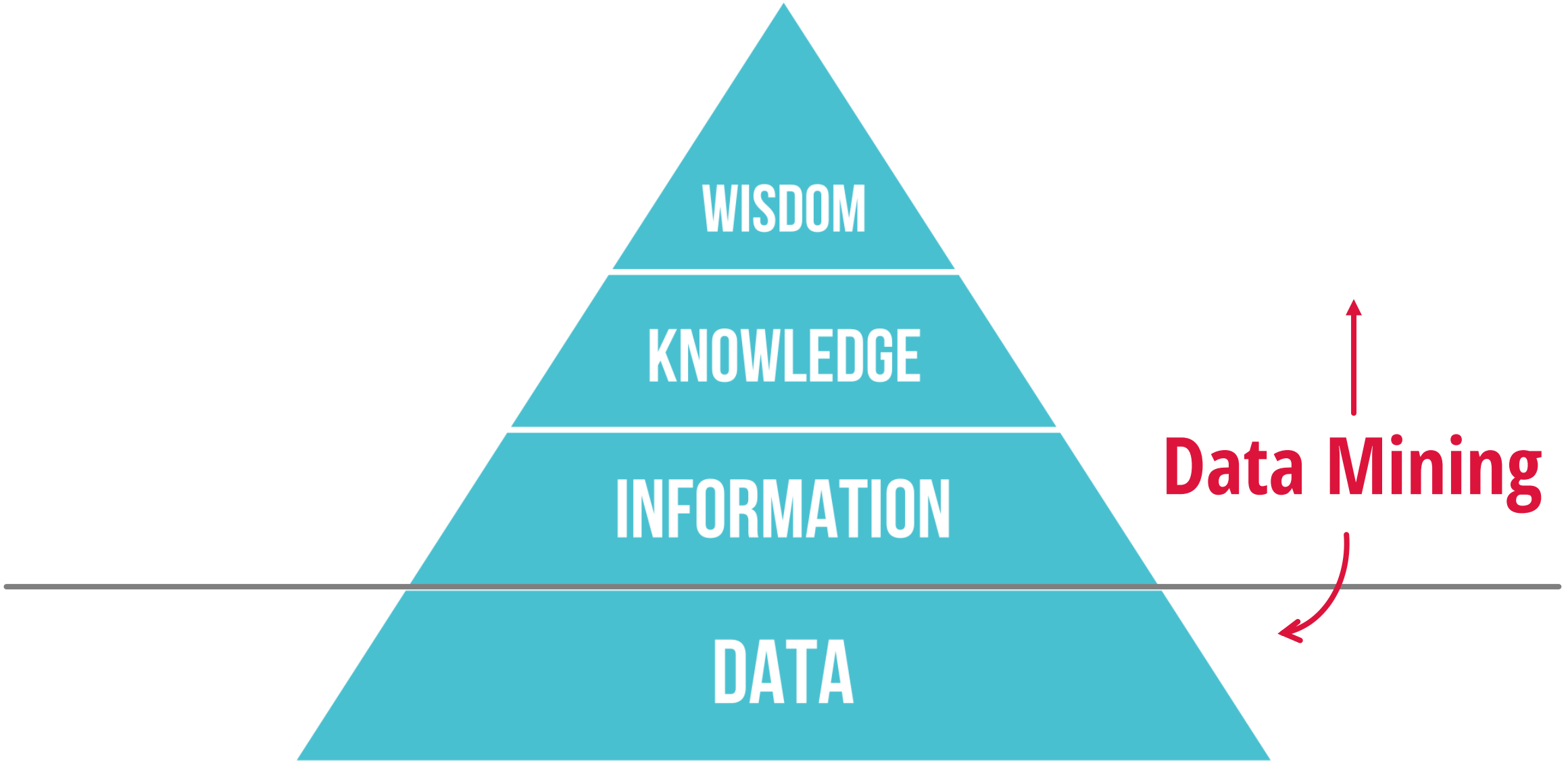


Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting
through large amounts of data stored in repositories,
using pattern recognition technologies as well as statistical and mathematical techniques.

– Gartner

What is Data Science?

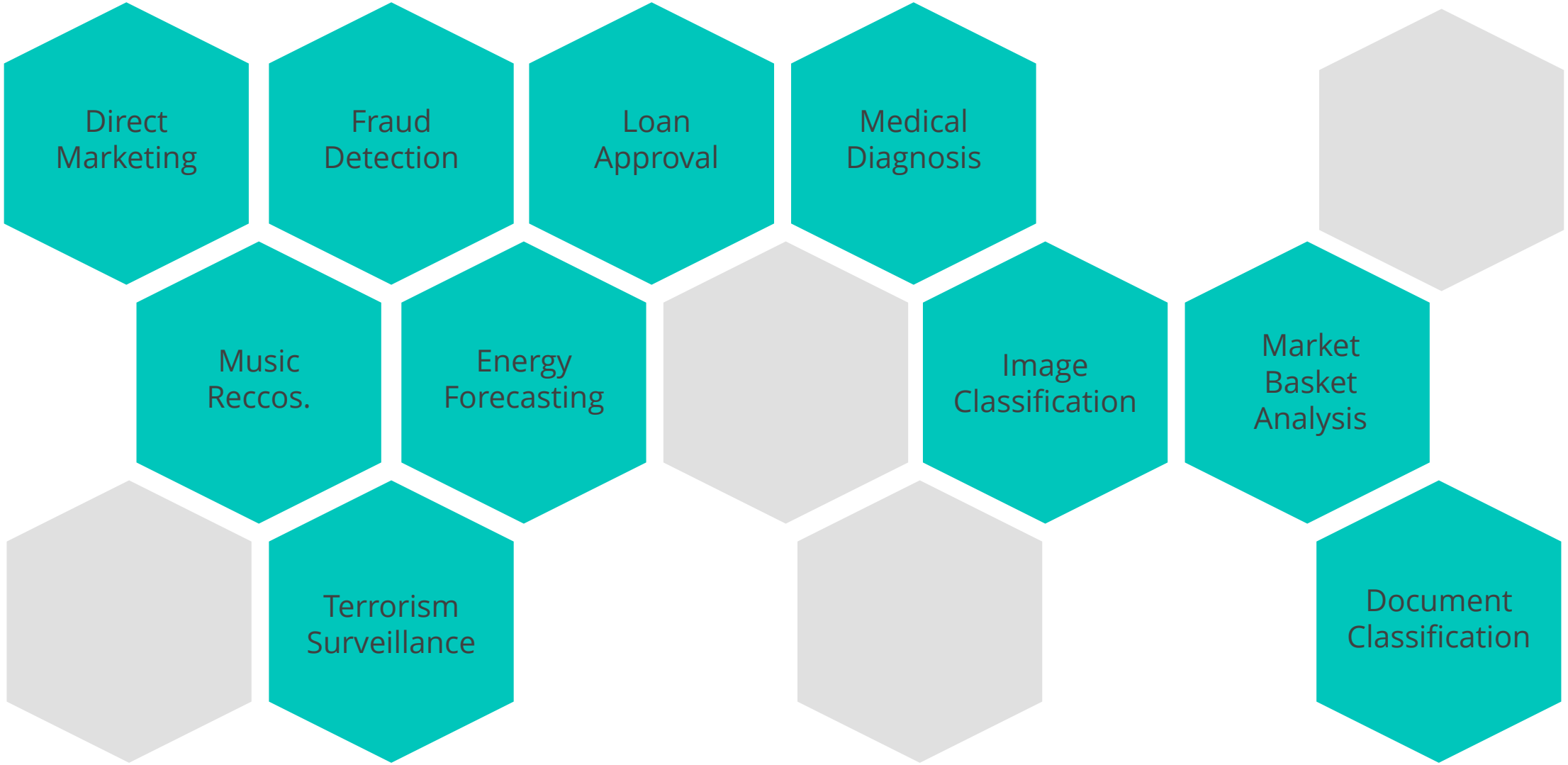
Data science is a field that involves using scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. It is a multidisciplinary field that combines domain expertise, programming skills, and knowledge of math and statistics to extract meaningful insights from data. Data scientists work with large and complex datasets to identify patterns, trends, and relationships that can help organizations make better decisions and improve their operations. They use a variety of tools and techniques, such as machine learning and statistical analysis, to analyze and interpret data, and communicate their findings through reports, visualizations, and other forms of presentation.


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

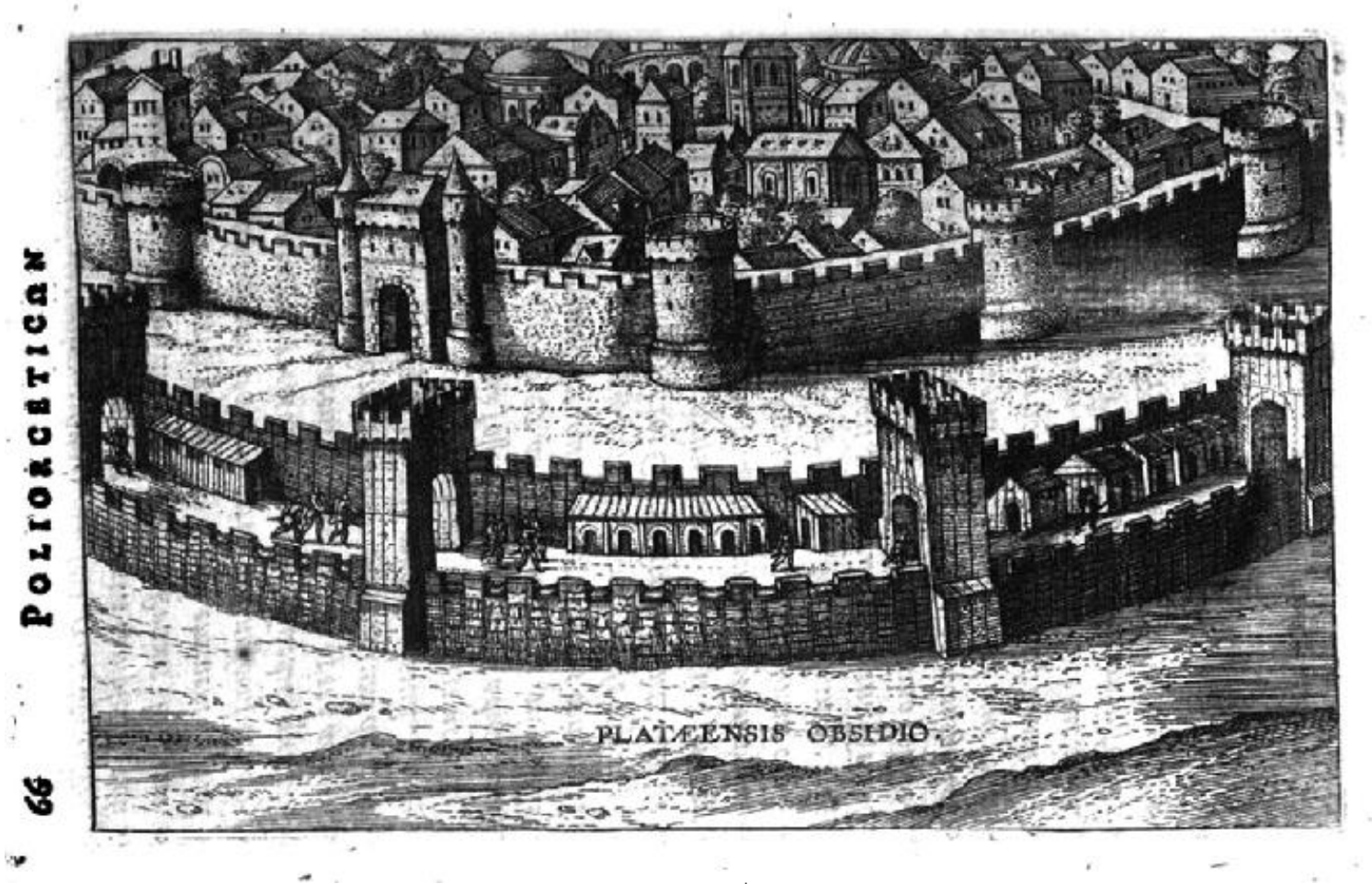
of all modern states of the world.

W Hooper, 1770

DATA COLLECTIONS + ANALYSIS + DECISION MAKING

Statistics

EXAMPLE #1: UNCERTAINTY



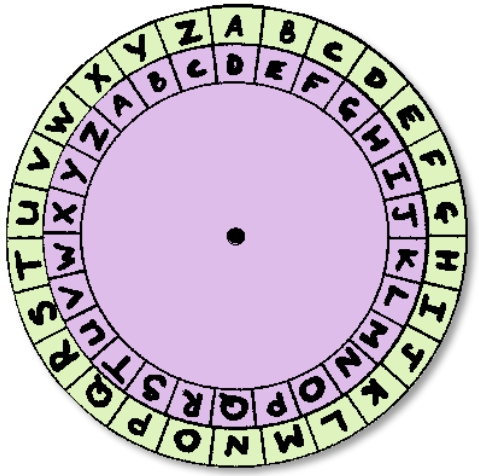
Siege of Plataea (5th Century BCE)

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

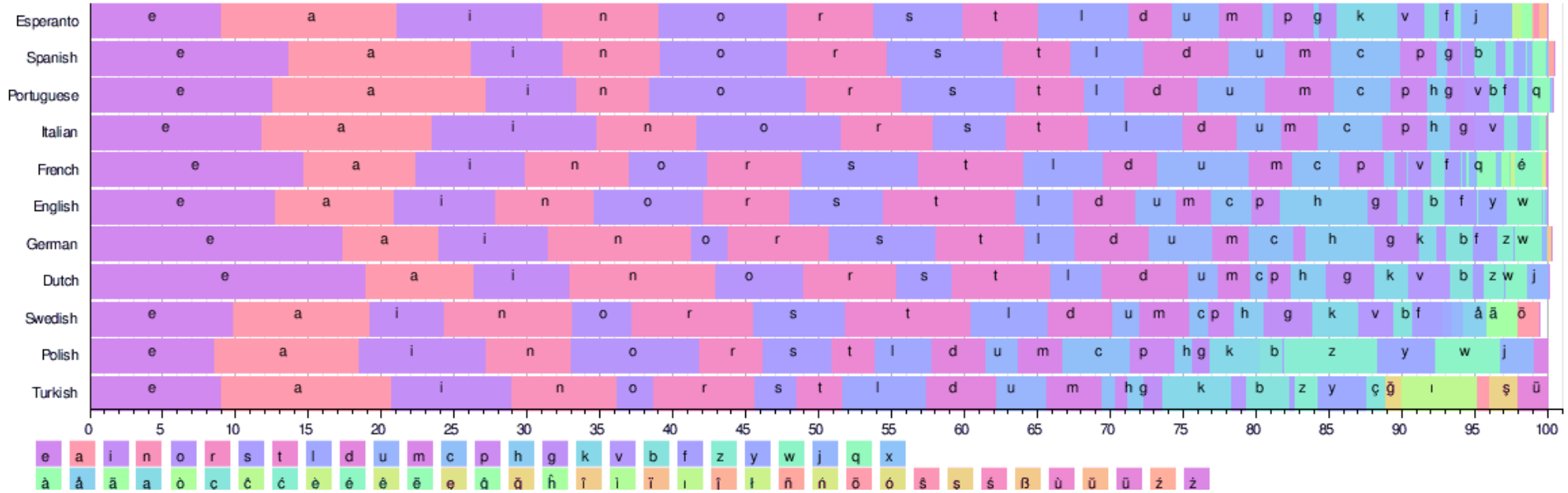


Caesar Cipher

Original message: Et tu, Brute?

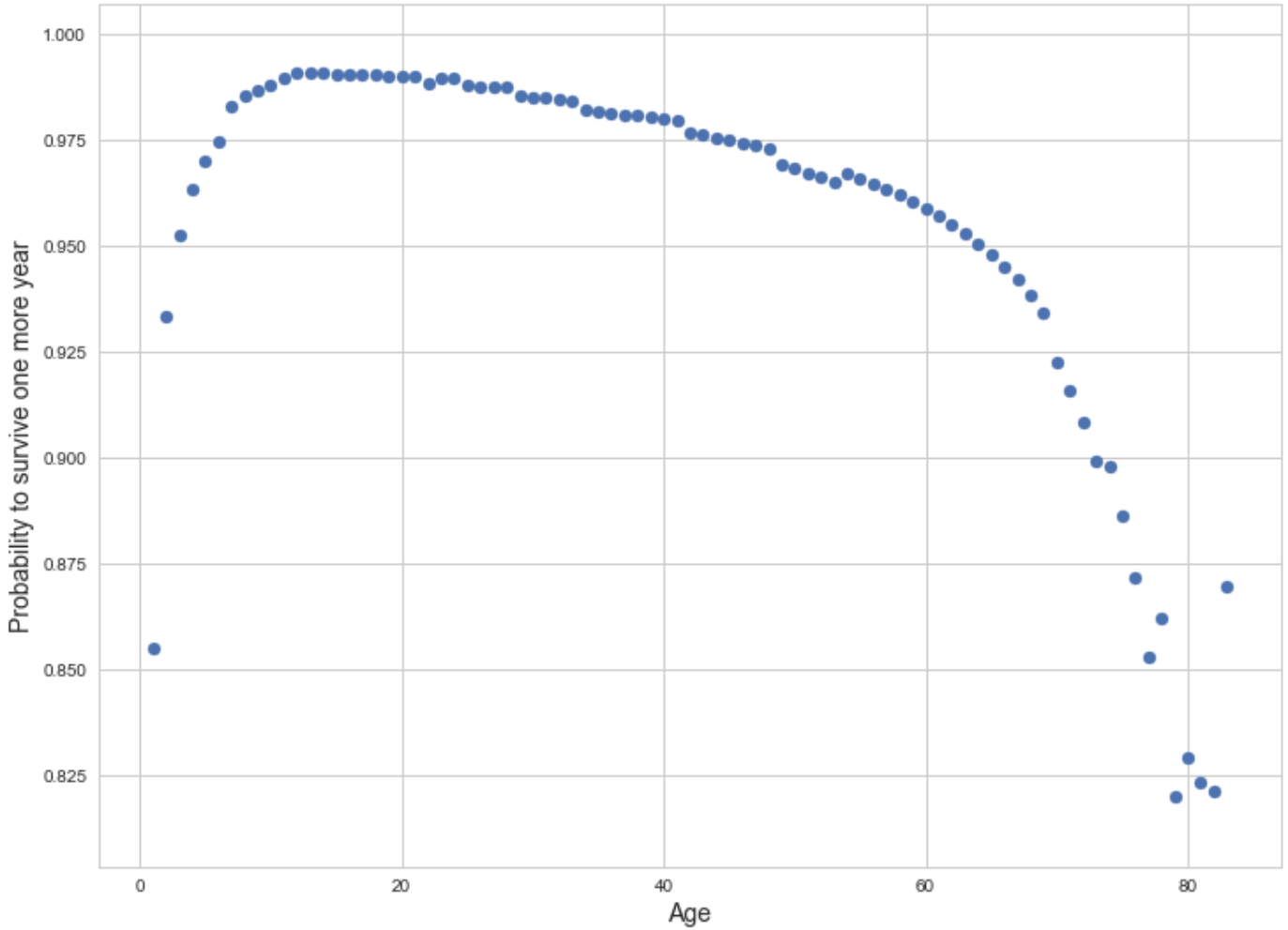
Encrypted message: Hw wx, Euxwh?

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

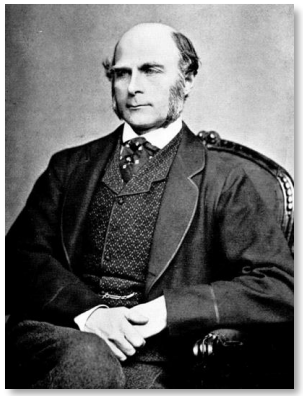
Random sampling

Design of-experiments

Bayesian Statistics

A rigorous mathematical discipline

for analysis, decision making, and inference



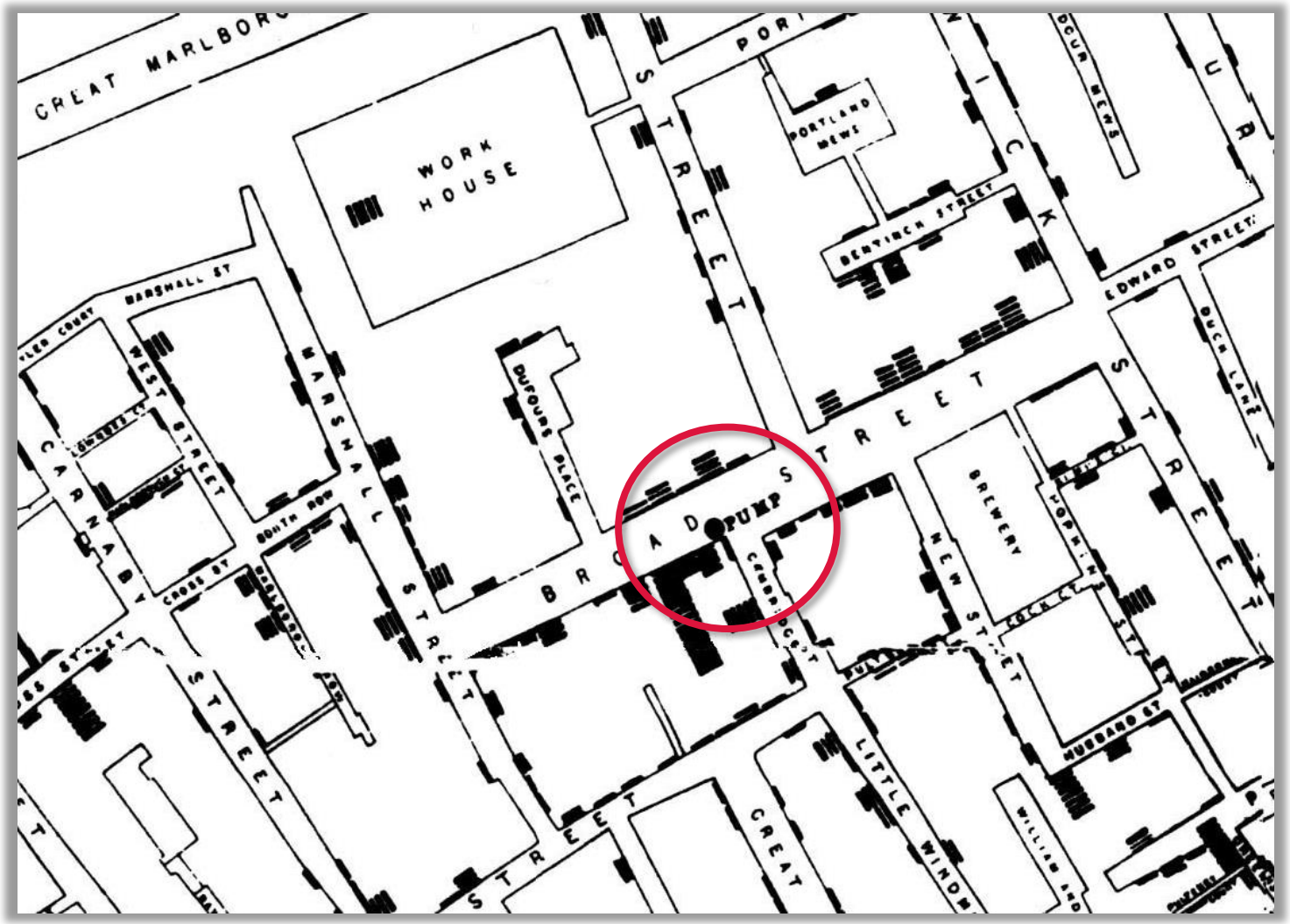
Sir Francis Galton
(1822–1911)
Correlation, regression

Carl Pearson
(1857–1936)
Founder of mathematical statistics

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

Modern Statistics

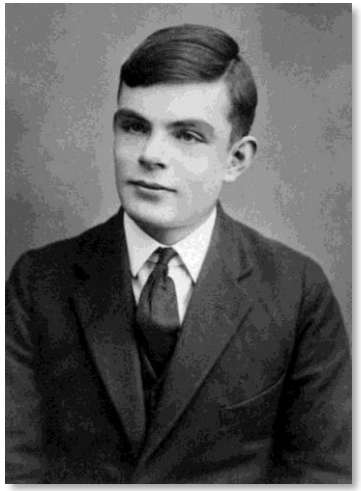
EXAMPLE: DATA VISUALIZATION



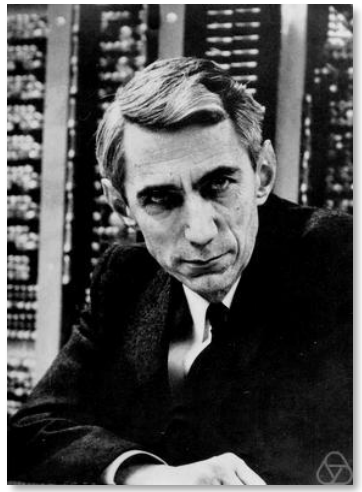
Original map by **John Snow** showing the clusters of cholera cases in the **London epidemic of 1854** [Source]

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

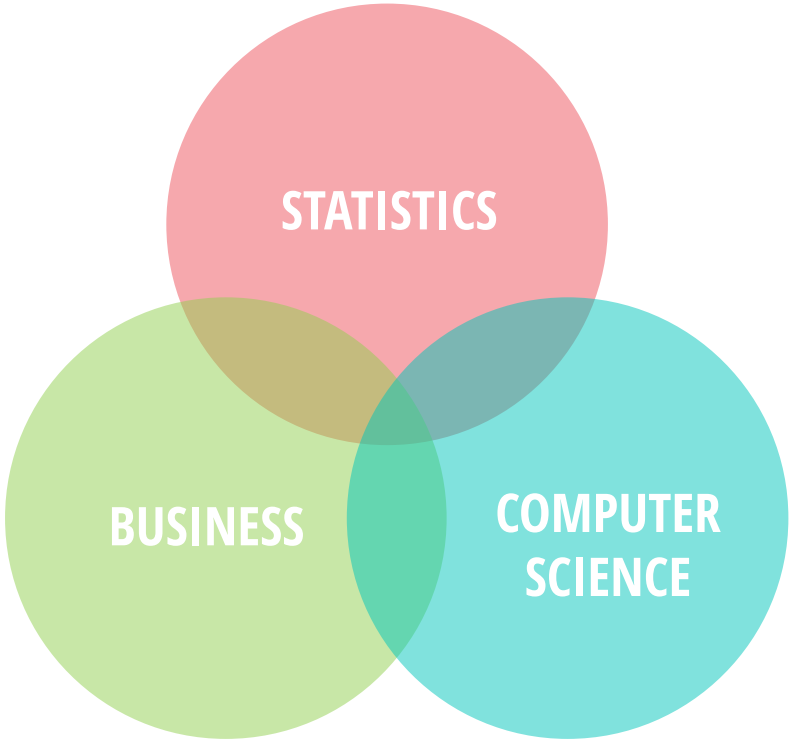
Recommender-systems

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

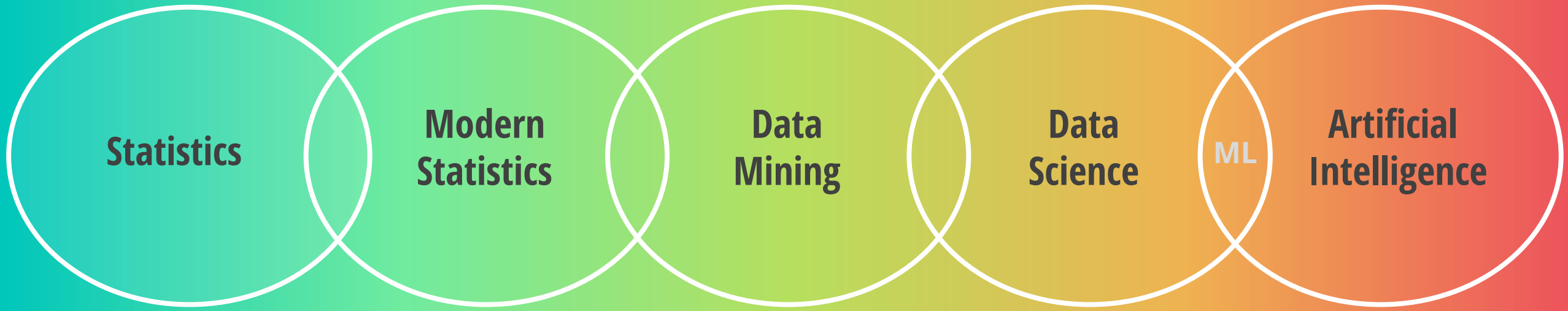
Deep learning

Reinforcement-Learning

Speech recognition

Natural Language-Processing

Computer vision



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

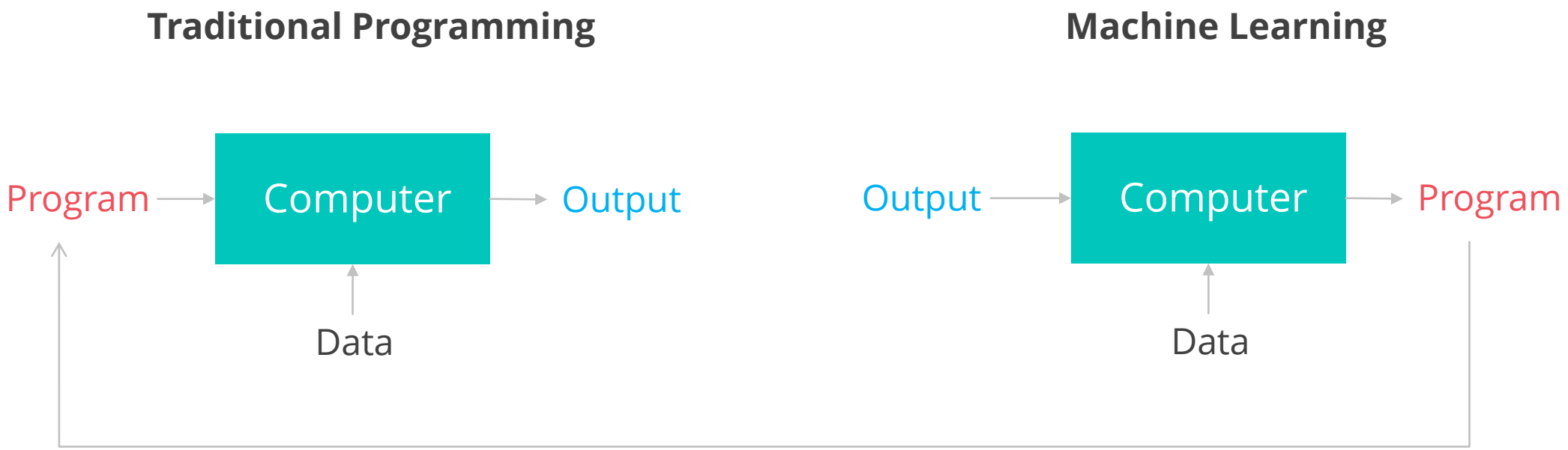
Calculations by hand → ***Distributed computing***

Evolution of techniques and technology

Machine Learning

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

Artur Samuel,1959



- 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.

Two Cultures

Statistics		Data Science
THEORETICAL		
		PRACTICAL
INFERENCE		
		PREDICTION
ASSUMPTIONS		
		EMPIRICAL
MANUAL		
		AUTOMATION

Course Outline

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

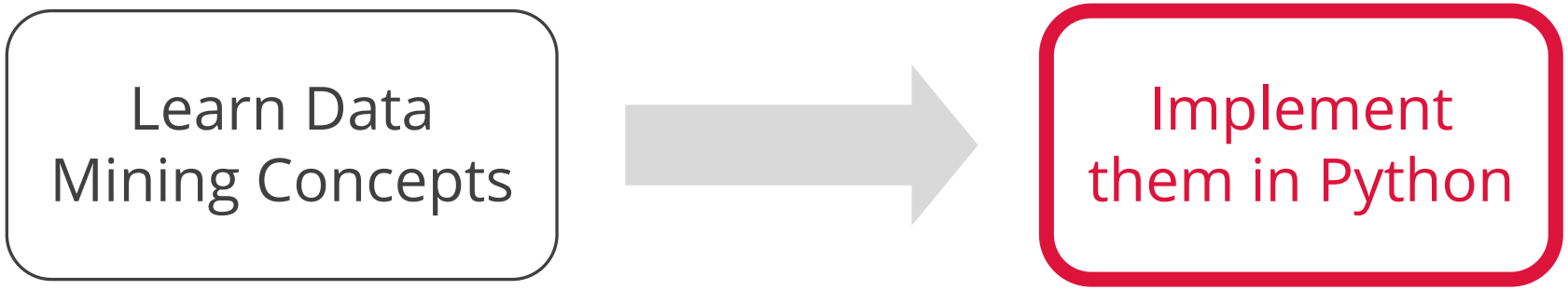
Class Structure

1. Ask questions at any time!
2. Collaboration is encouraged.
3. All content (course material) will be available on Blackboard (and on a git repository).
4. Data Mining + Python
5. Homework assignments in Python

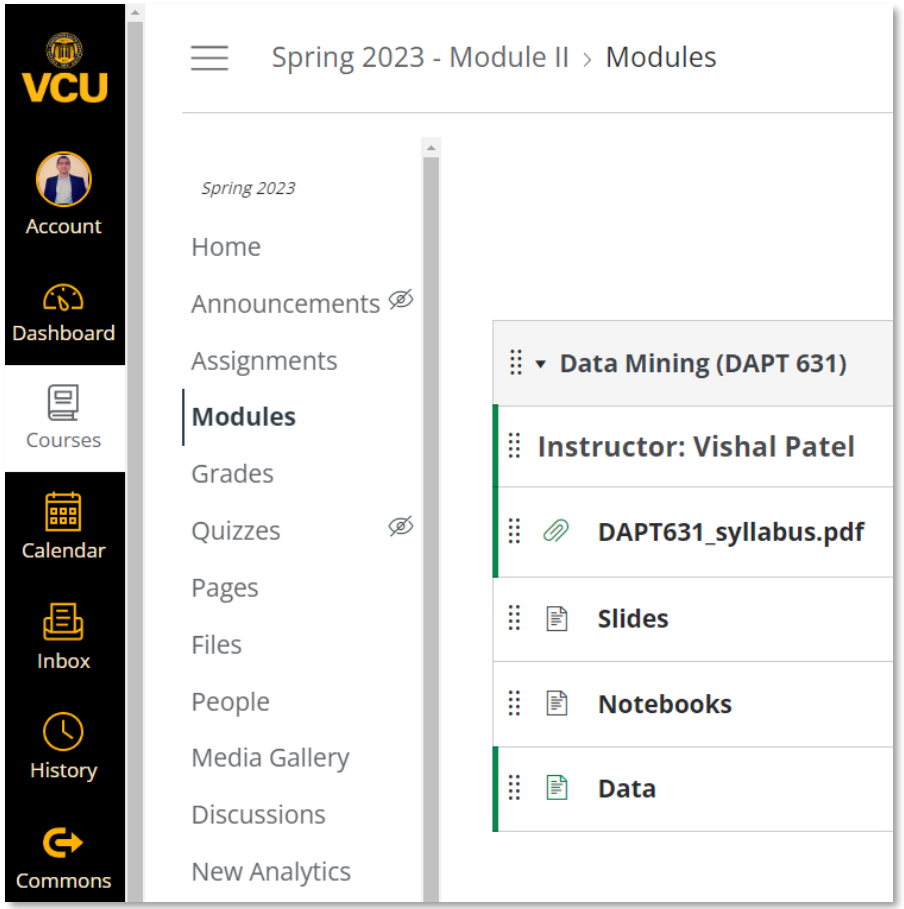
My Objectives

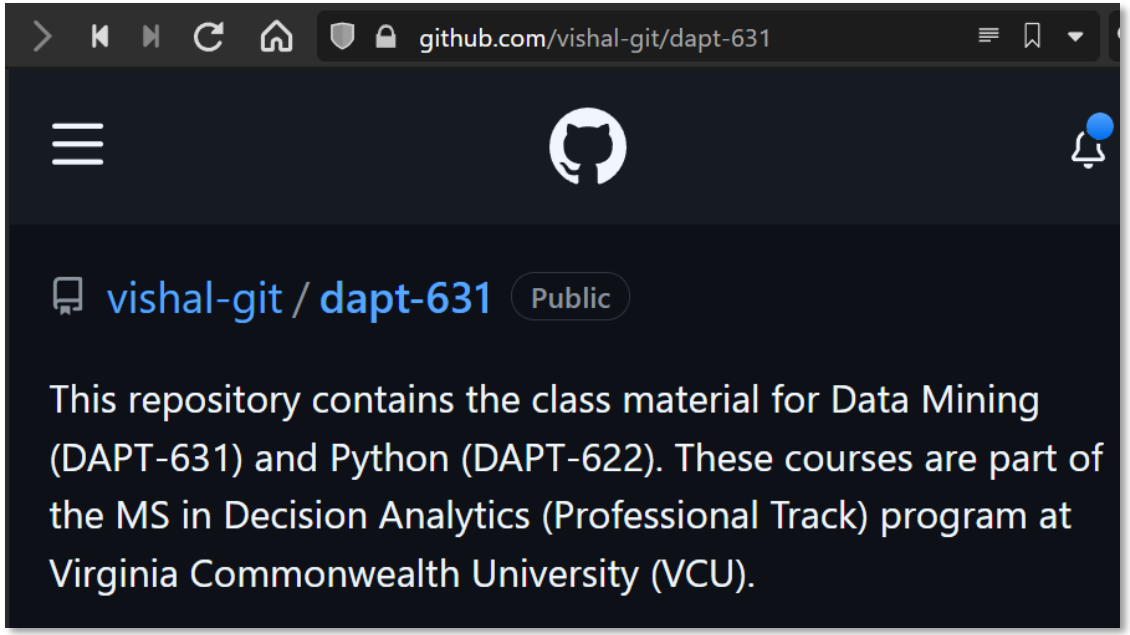
1. Provide a practical knowledge of data mining algorithms.
2. Give a broader perspective to help understand what role data mining plays in the decision-making process.
3. Help you develop an appreciation for the beauty of the theoretical foundations underlying data mining.
4. Help you think more like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

Data Mining + Python



Course Material





			HyFlex		HyFlex		HyFlex		HyFlex	
		Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9
	Friday	44567	20-Jan	44595	17-Feb	44623	44637	44651	14-Apr	44679
	11:30 - 12:30	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch
Session 1	12:30 - 2:15	Statistics	Data Mining	Practicum Client Meetings	Data Mining	Statistics	Risk Analysis	Statistics	Statistics	Statistics
Session 2	2:30 - 4:15	Statistics	Statistics	Forecasting	Forecasting	Forecasting	Forecasting	Forecasting	Python	Risk Analysis
Session 3	4:30 - 6:15	Forecasting	Statistics	Forecasting	Forecasting	Forecasting	Tableau	Forecasting	Python	Risk Analysis
Special Events	6:45 - 8:00	Social		Social		Social		Spring Gala		
	Saturday	7-Jan	21-Jan	4-Feb	44610	4-Mar	18-Mar	1-Apr	44666	29-Apr
	7:30 - 8:00	Breakfast	Breakfast	Breakfast	Breakfast	Breakfast	Breakfast	Breakfast	Breakfast	Breakfast
Session 4	8:00 - 9:45	Data Mining	Python	Data Mining	Statistics	Python	Data Mining	Data Mining	Risk Analysis	Data Mining
Session 5	10:00 - 11:45	Data Mining	Python	Data Mining	Statistics	Python	Data Mining	Data Mining	Risk Analysis	Data Mining
	11:45 - 12:30	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch
Session 6	12:30 - 2:15	Practicum Introductions	Forecasting	Tableau	Tableau	Practicum Status Reports	Statistics	Tableau	Tableau	Practicum Status Reports
Session 7	2:30 - 3:45	CATME Feedback	Forecasting	Tableau	Tableau	Practicum Status Reports	Statistics	Tableau	Tableau	Practicum Status Reports

vishal@derive.io

www.linkedIn.com/in/VishalJP

@derive_io