Data Mining

DAPT 631

Vishal Patel

Spring 2024

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
BERG FREUNDE.DE
ActiveCampaign
response LABS
Michaels
The Container Store
AMERICA'S TEST KITCHEN
BOSS HUGO BOSS
Verdata
MOEN
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

AI / Generative AI: Hype of Tripe?

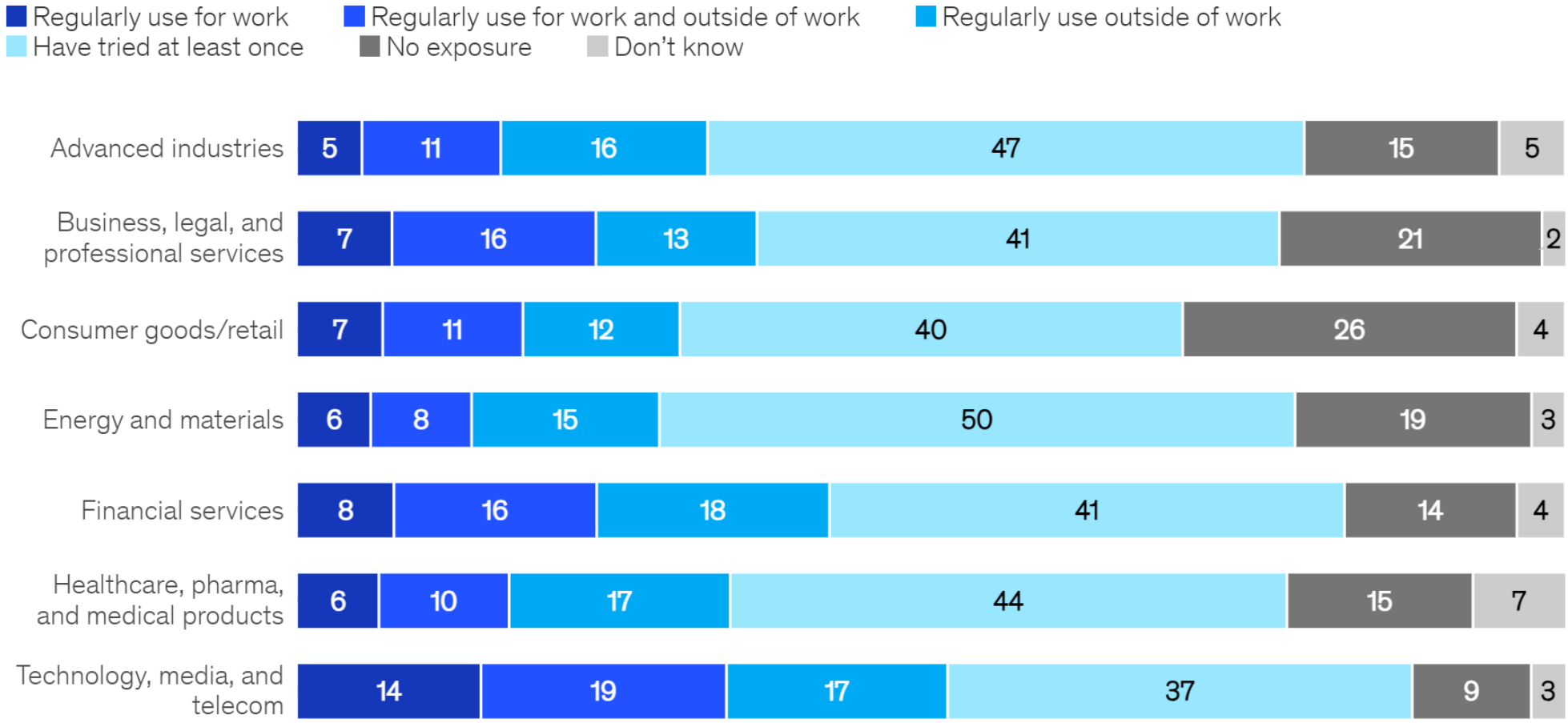
How do you feel about the current AI trends?

A. Excited / Optimistic
B. Nervous / Skeptic
C. It’s a fad
D. Indifferent
E. Other

Respondents across regions, industries, and seniority levels say they are already using generative AI tools.

Reported exposure to generative AI tools, % of respondents

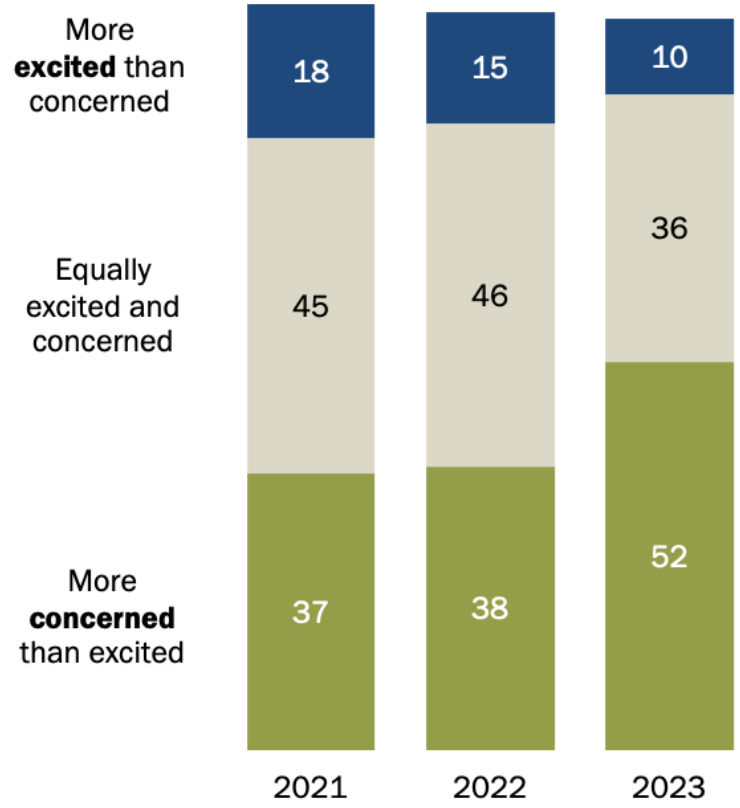
Select demographic: By industry



August 2023

Concern about artificial intelligence in daily life far outweighs excitement

% of U.S. adults who say the increased use of artificial intelligence in daily life makes them feel ...



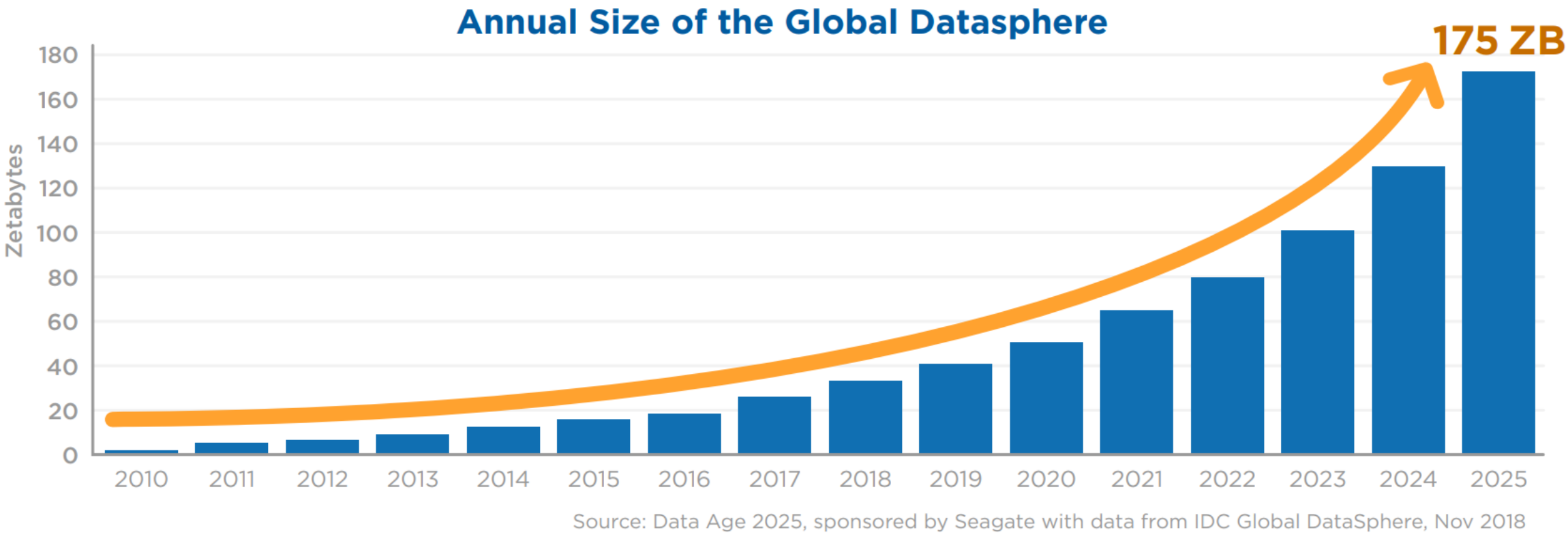
Note: Respondents who did not give an answer are not shown.
Source: Survey conducted July 31-Aug. 6, 2023.

PEW RESEARCH CENTER

Introduction

- History
- Course Structure

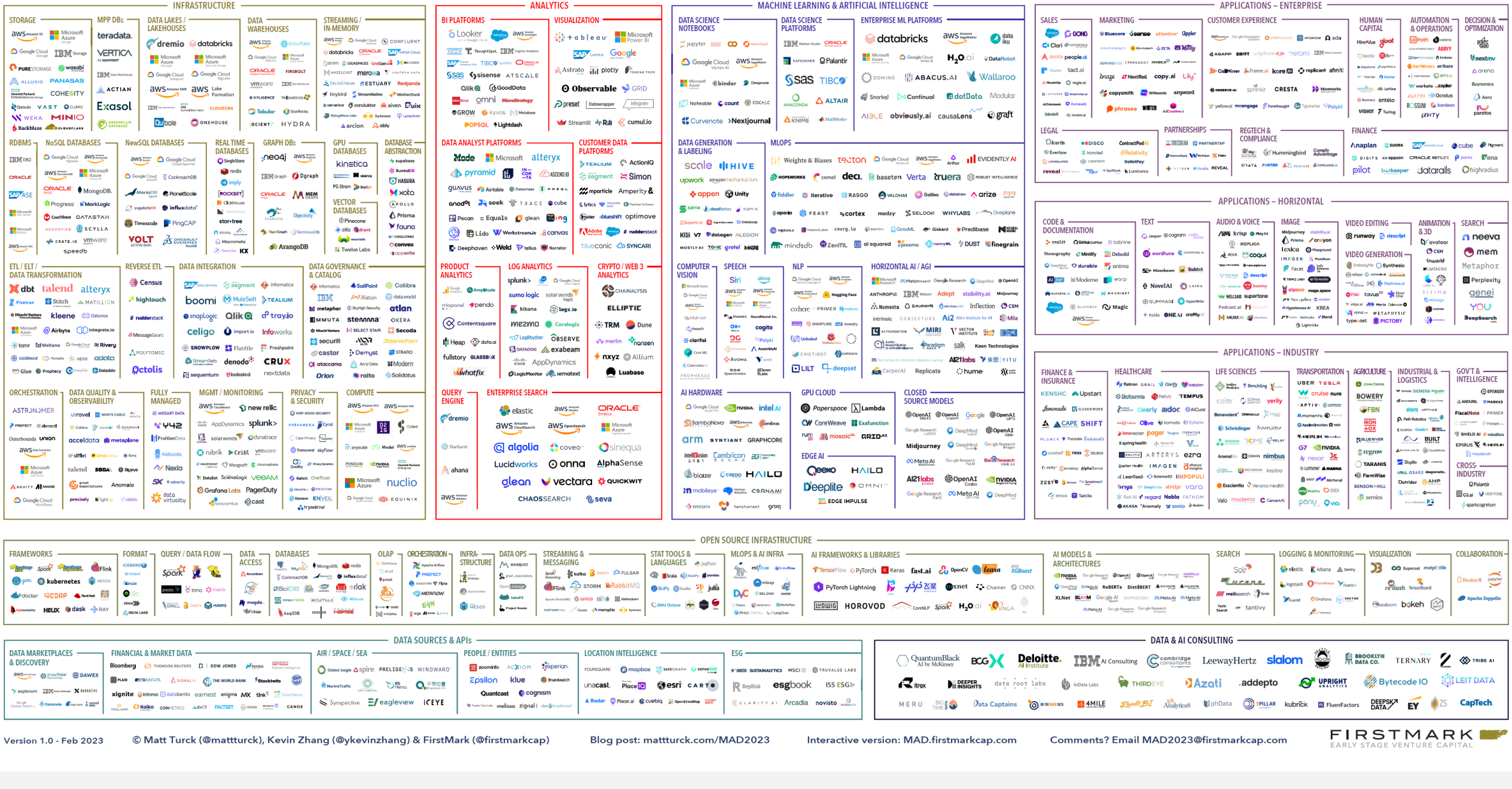

Cambrian Era



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

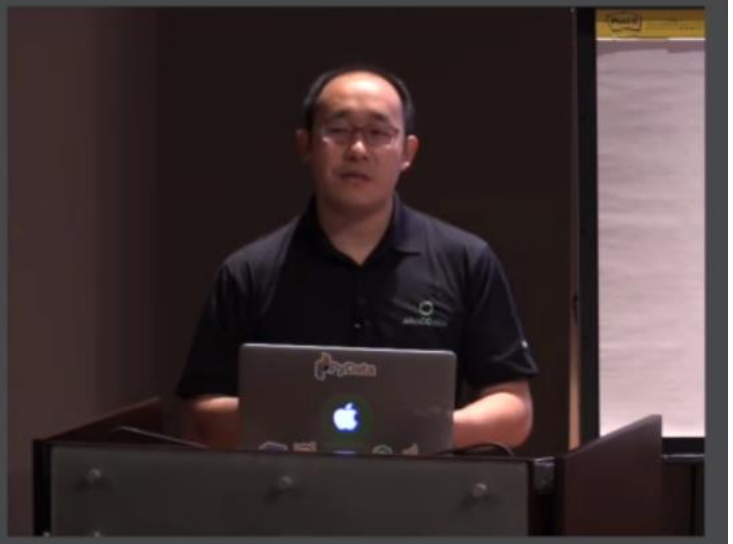
Watch later Share

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

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.

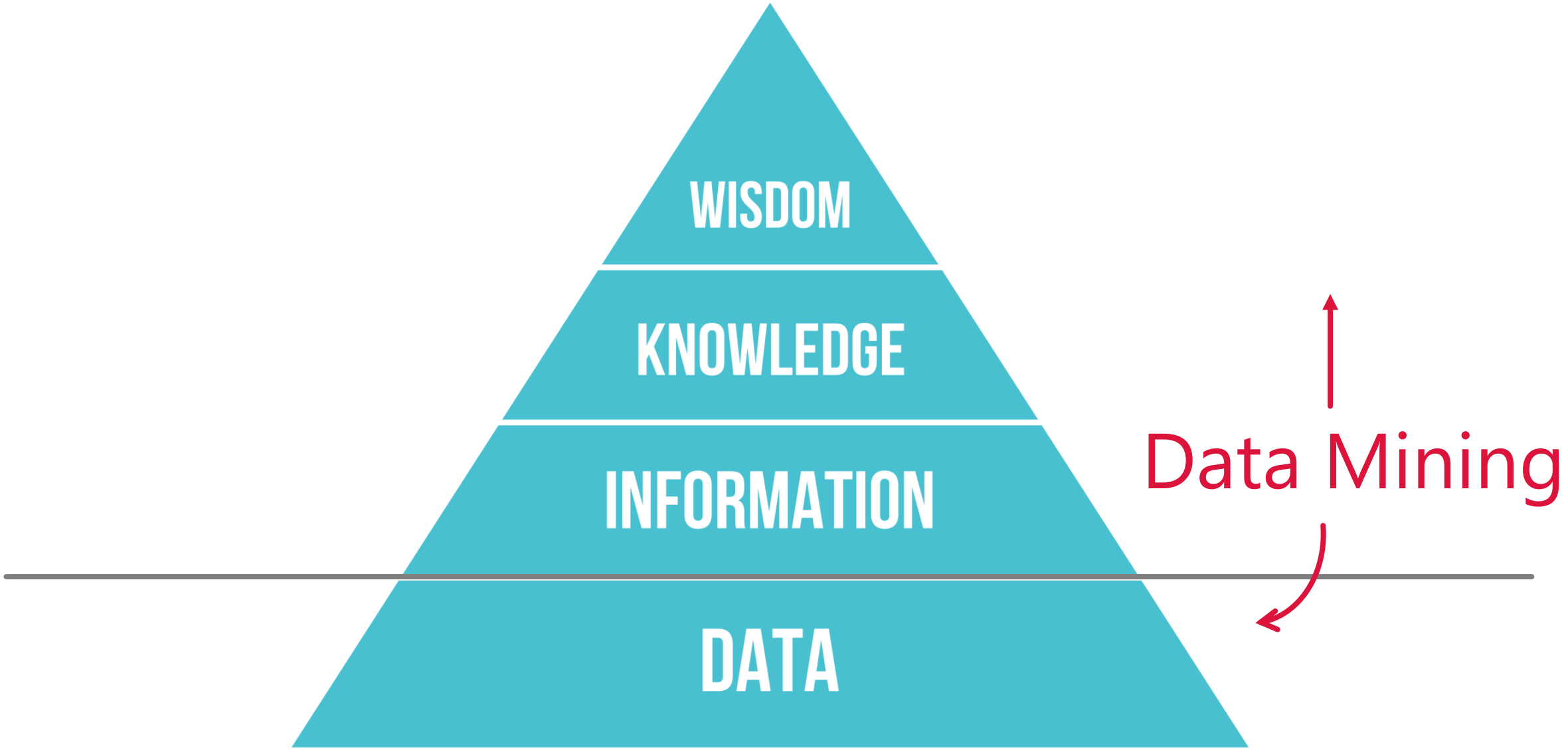
– Witten and Frank

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.

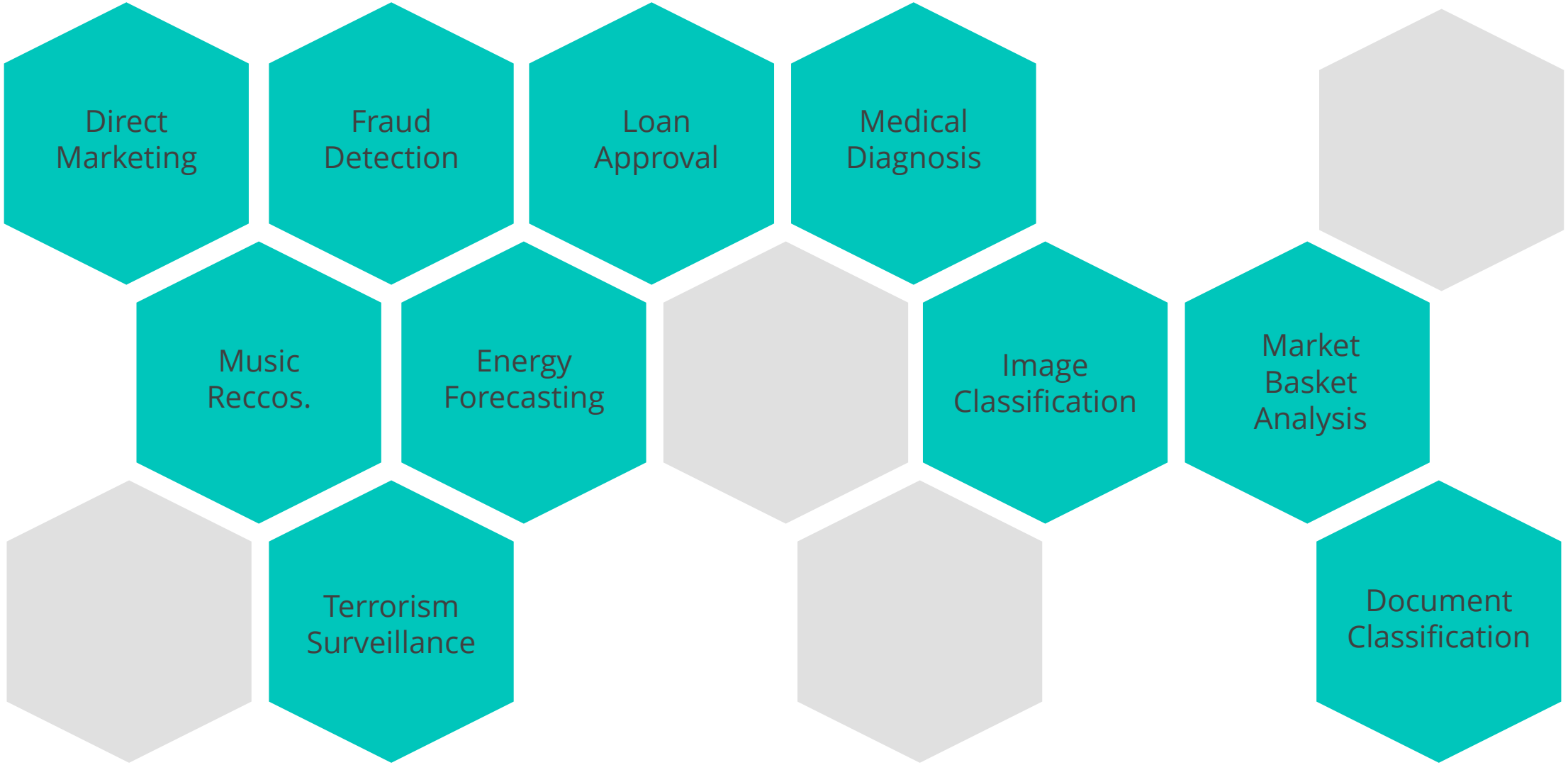
– Gartner


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

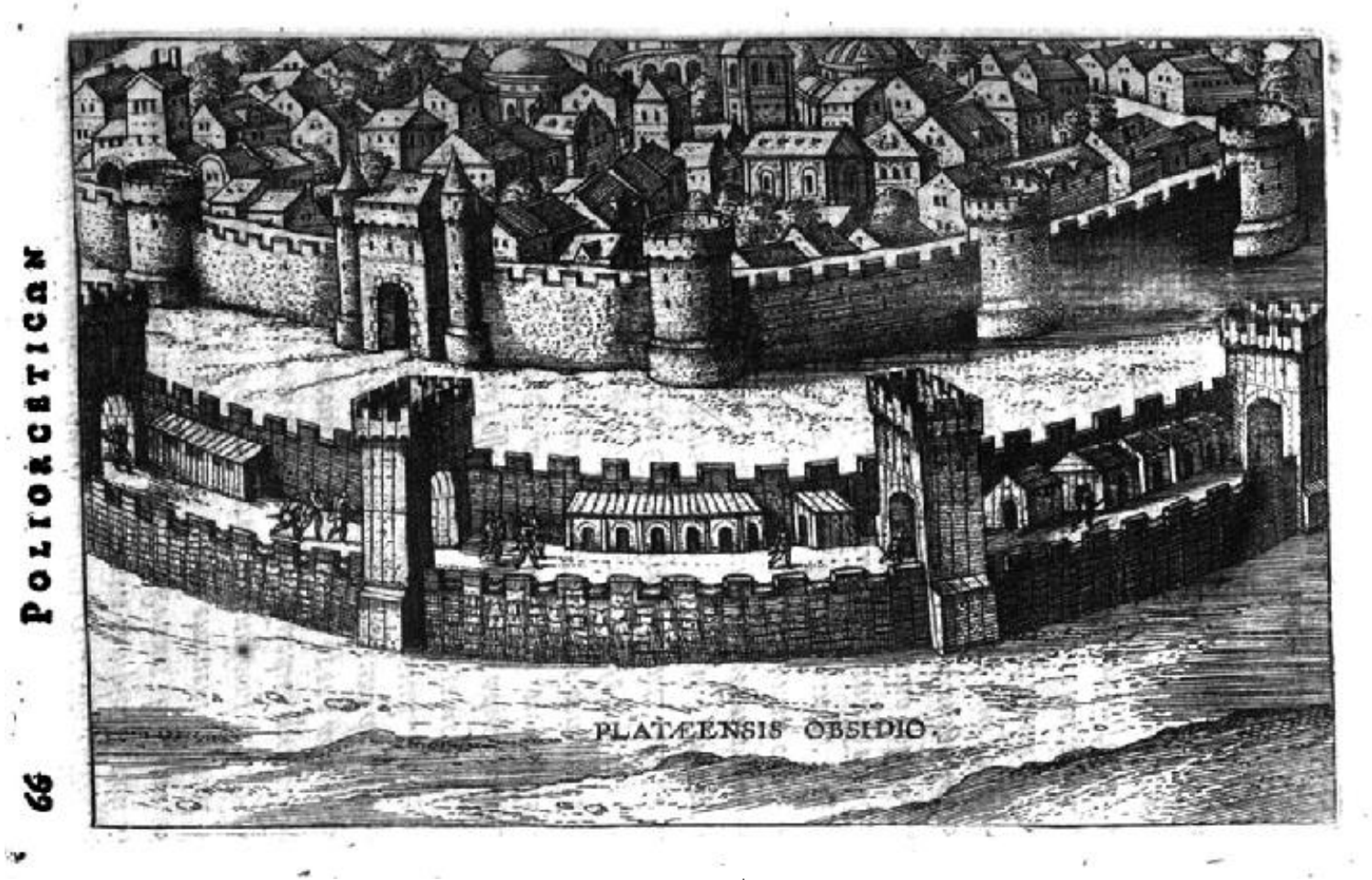
of all modern states of the world.

W Hooper, 1770

DATA COLLECTIONS + ANALYSIS + DECISION MAKING

Statistics

EXAMPLE #1: UNCERTAINTY



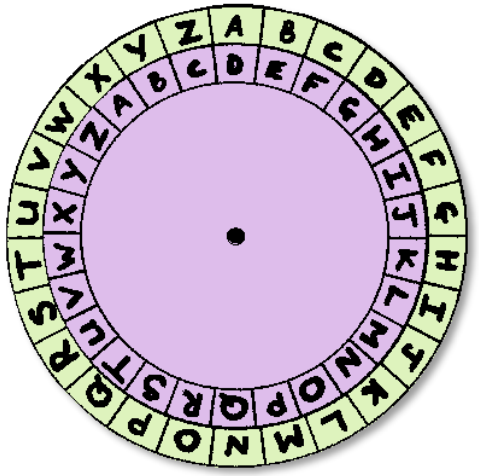
Siege of Plataea (5th Century BCE)

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

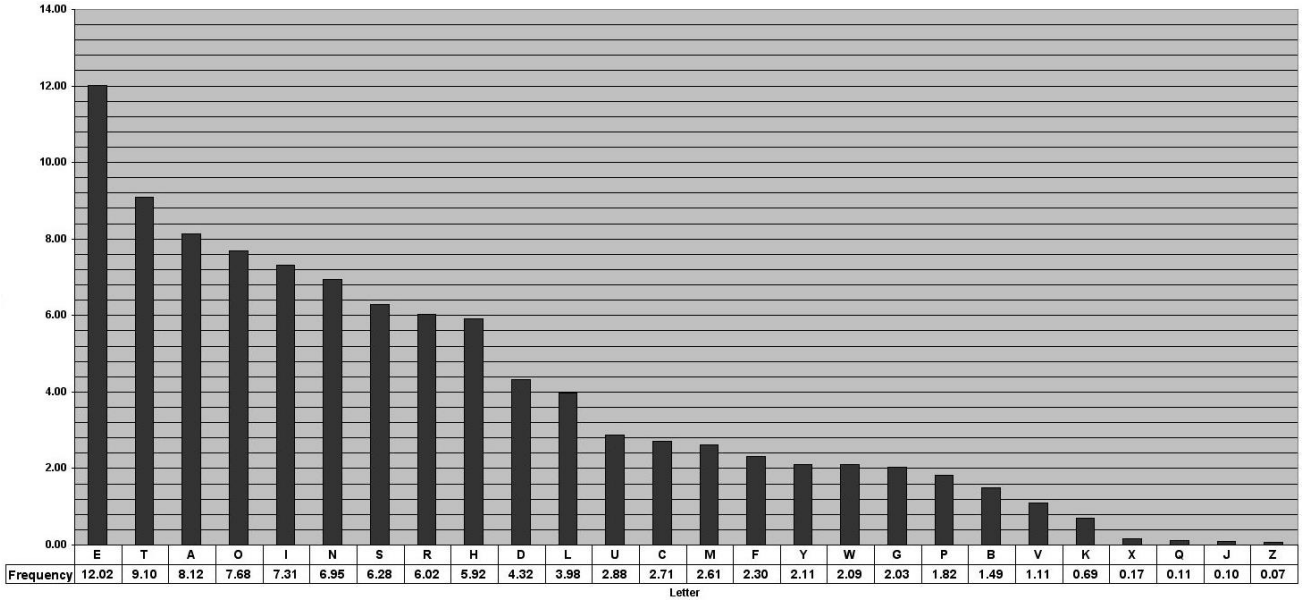


Caesar Cipher

Original message: Et tu, Brute?

Encrypted message: Hw wx, Euxwh?

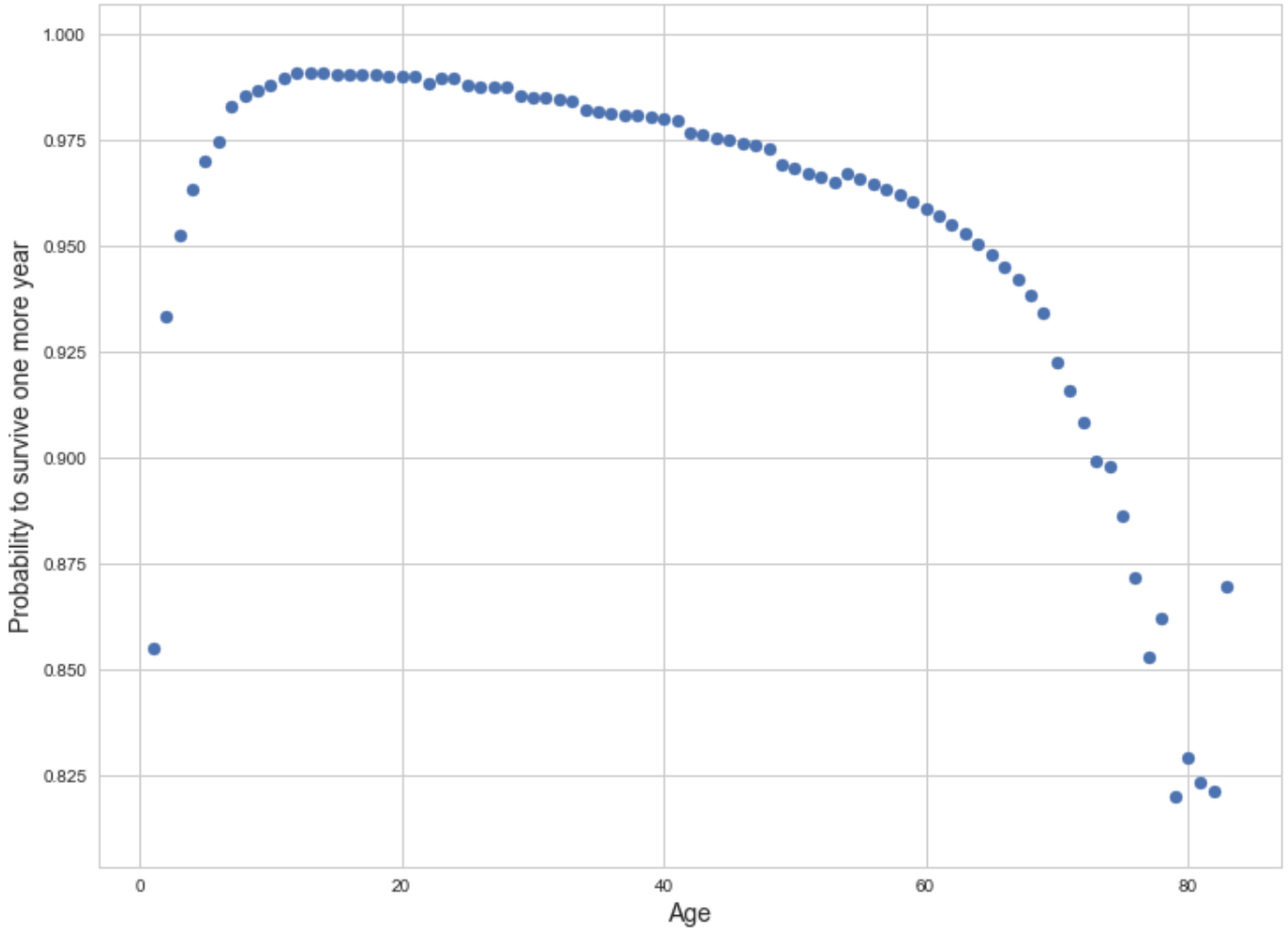
Al-Kindi
(801–873 AD)



	E	T	A	O	I	N	S	R	H	D	L	U	C	M	F	Y	W	G	P	B	V	K	X	Q	J	Z
Frequency	12.02	9.10	8.12	7.68	7.31	6.95	6.28	6.02	5.92	4.32	3.98	2.88	2.71	2.61	2.30	2.11	2.09	2.03	1.82	1.49	1.11	0.69	0.17	0.11	0.10	0.07

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

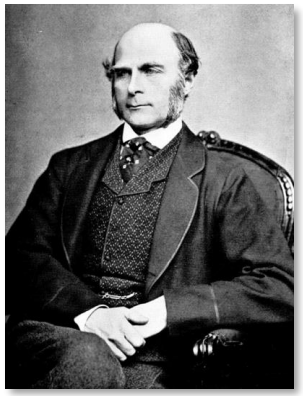
Random sampling

Design of-experiments

Bayesian Statistics

A rigorous mathematical discipline

for analysis, decision making, and inference



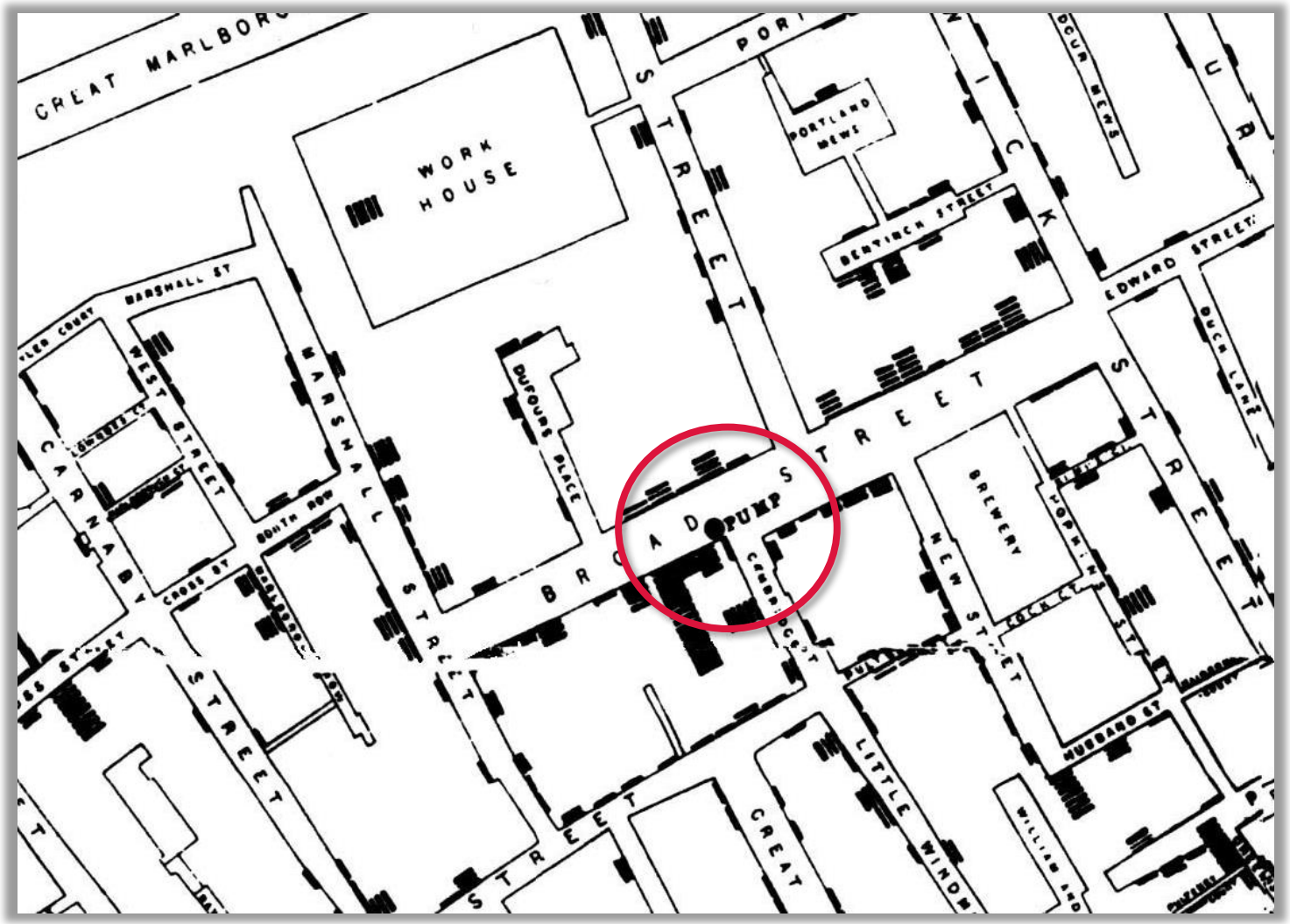
Sir Francis Galton
(1822–1911)
Correlation, regression

Carl Pearson
(1857–1936)
Founder of mathematical statistics

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

Modern Statistics

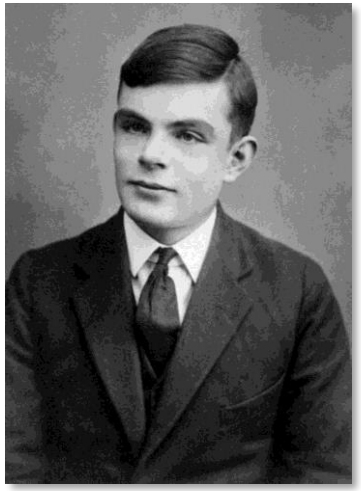
EXAMPLE: DATA VISUALIZATION



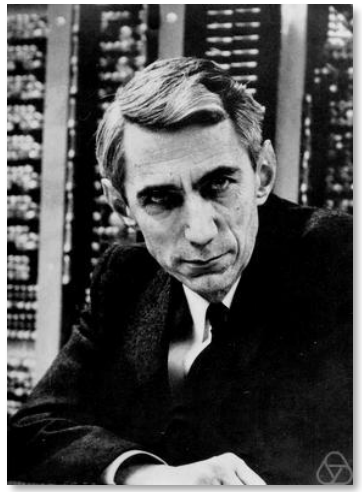
Original map by **John Snow** showing the clusters of cholera cases in the **London epidemic of 1854** [Source]

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

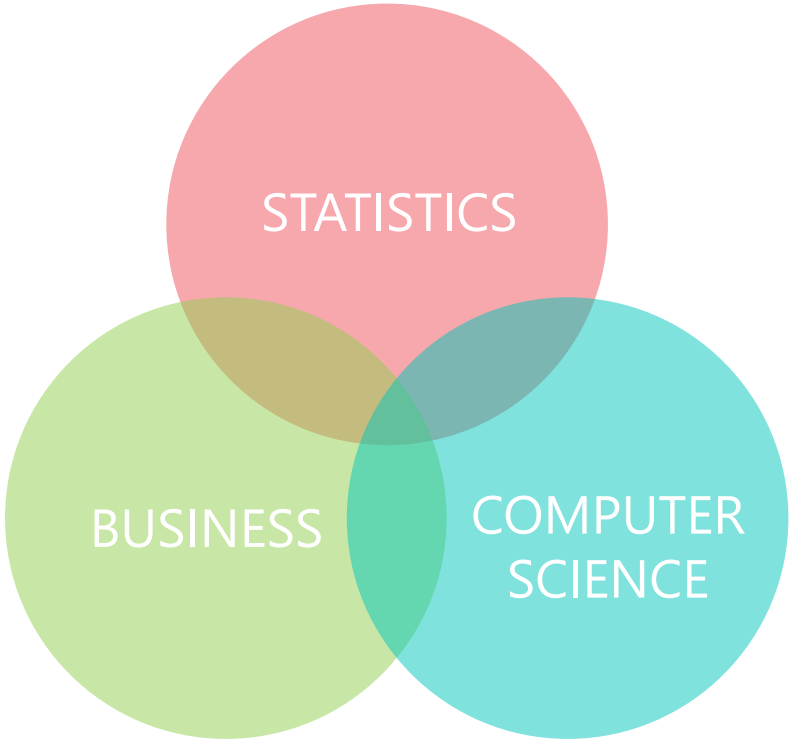
Recommender-systems

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

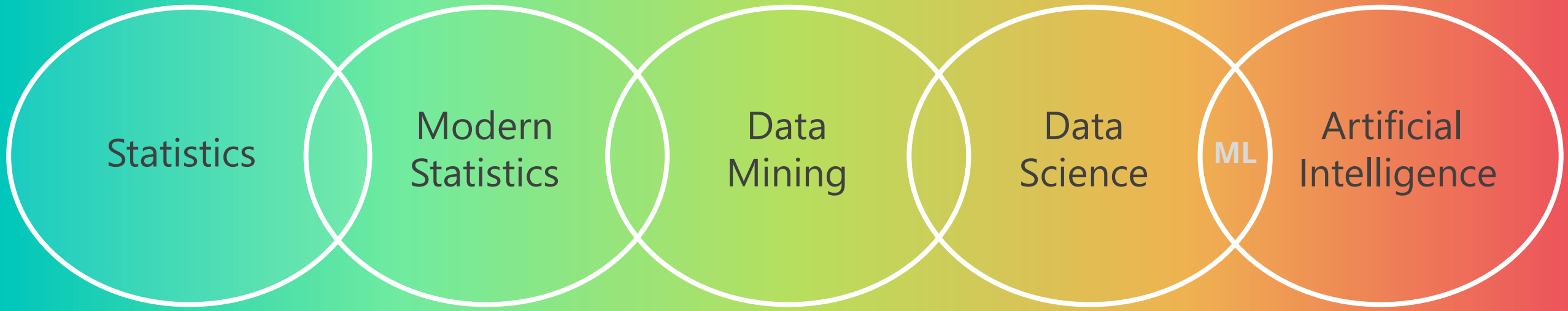
Deep learning

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

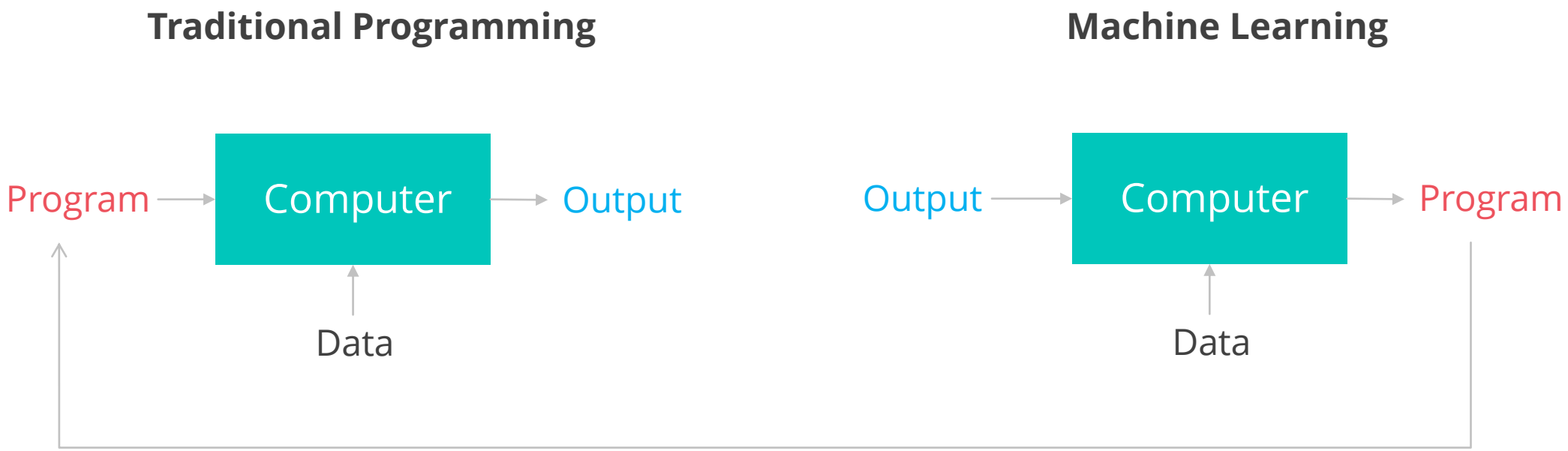
Calculations by hand → ***Distributed computing***

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

THEORETICAL

INFERENCE

ASSUMPTIONS

MANUAL

PRACTICAL

PREDICTION

EMPIRICAL

AUTOMATION

Data Science

Course Outline

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

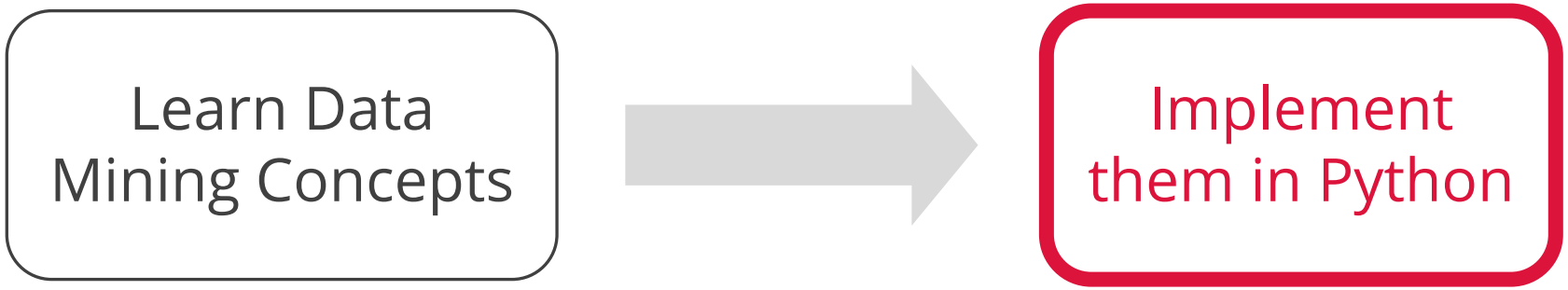
Class Structure

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

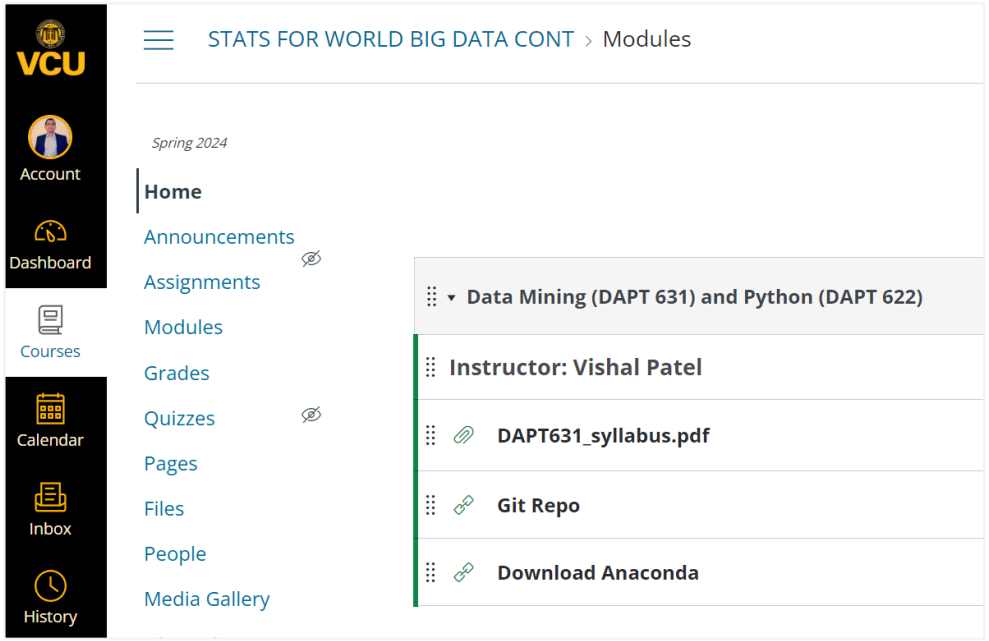
My Objectives

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.

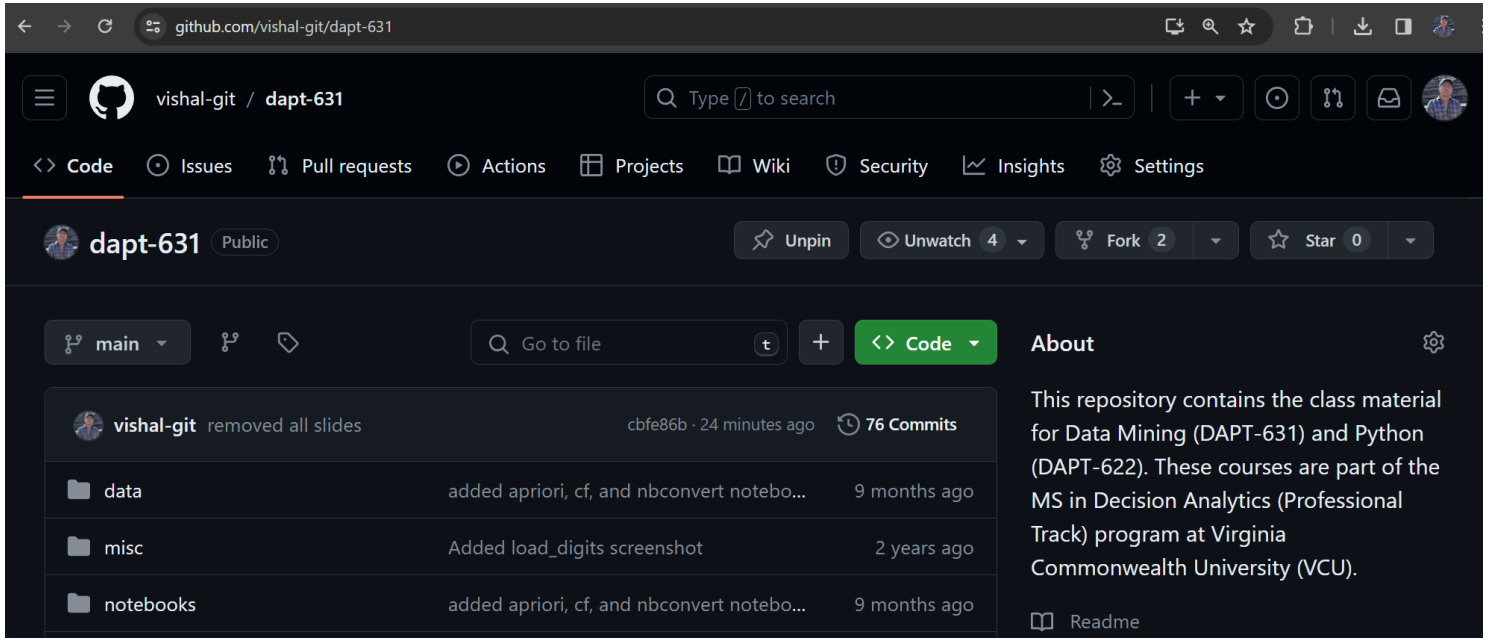
Data Mining + Python



Course Material



VCU Canvas

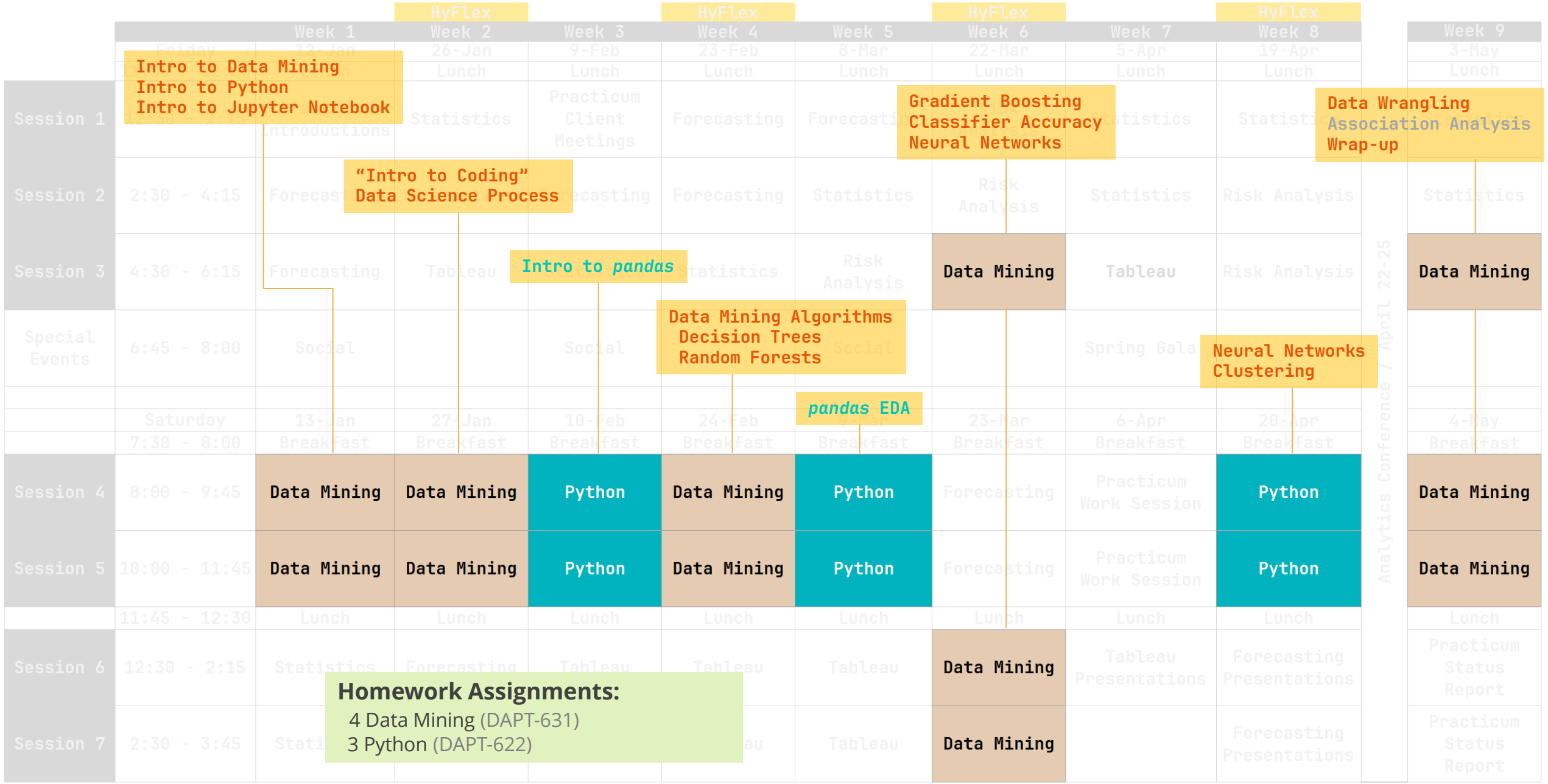


GitHub

			HyFlex		HyFlex		HyFlex		HyFlex	
		Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9
	Friday	12-Jan	26-Jan	9-Feb	23-Feb	8-Mar	22-Mar	5-Apr	19-Apr	3-May
	11:30 - 12:30	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch
Session 1	12:30 - 2:15	Practicum Introductions	Statistics	Practicum Client Meetings	Forecasting	Forecasting	Risk Analysis	Statistics	Statistics	Statistics
Session 2	2:30 - 4:15	Forecasting	Statistics	Forecasting	Forecasting	Statistics	Risk Analysis	Statistics	Risk Analysis	Statistics
Session 3	4:30 - 6:15	Forecasting	Tableau	Statistics	Statistics	Risk Analysis	Data Mining	Tableau	Risk Analysis	Data Mining
Special Events	6:45 - 8:00	Social		Social		Social		Spring Gala		
	Saturday	13-Jan	27-Jan	10-Feb	24-Feb	9-Mar	23-Mar	6-Apr	20-Apr	4-May
	7:30 - 8:00	Breakfast	Breakfast	Breakfast	Breakfast	Breakfast	Breakfast	Breakfast	Breakfast	Breakfast
Session 4	8:00 - 9:45	Data Mining	Data Mining	Python	Data Mining	Python	Forecasting	Practicum Work Session	Python	Data Mining
Session 5	10:00 - 11:45	Data Mining	Data Mining	Python	Data Mining	Python	Forecasting	Practicum Work Session	Python	Data Mining
	11:45 - 12:30	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch
Session 6	12:30 - 2:15	Statistics	Forecasting	Tableau	Tableau	Tableau	Data Mining	Tableau Presentations	Forecasting Presentations	Practicum Status Report
Session 7	2:30 - 3:45	Statistics	Forecasting	Tableau	Tableau	Tableau	Data Mining		Forecasting Presentations	Practicum Status Report

Analytics Conference / April 22-25

31 ½ Hours!


HyFlex
HyFlex
HyFlex
HyFlex
Week 1
Week 2
Week 3
Week 4
Week 5
Week 6
Week 7
Week 8
Week 9
Intro to Data Mining
Intro to Python
Intro to Jupyter Notebook
"Intro to Coding"
Data Science Process
Intro to pandas
Data Mining Algorithms
Decision Trees
Random Forests
pandas EDA
Gradient Boosting
Classifier Accuracy
Neural Networks
Neural Networks
Clustering
Data Wrangling
Association Analysis
Wrap-up
Session 1
Session 2
Session 3
Special Events
Session 4
Session 5
Session 6
Session 7
Data Mining
Python
Tableau
Homework Assignments:
4 Data Mining (DAPT-631)
3 Python (DAPT-622)
Analytics Conference / April 22-25

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

vishal@derive.io

www.linkedIn.com/in/VishalJP

@derive_io