A Systematic Approach to Data Science

M. Tamer Özsu
University of Waterloo
tamer.ozsu@uwaterloo.ca

World’s Most Valuable Resource

“**Data** is the new oil.”

Clive Robert Humby
mathematician, entrepreneur, and Chief Data Scientist, Starcount

“**Data** is the new currency.”

Antonio Neri, *President Hewlett Packard Enterprise*

“Data is a commodity like gold.”

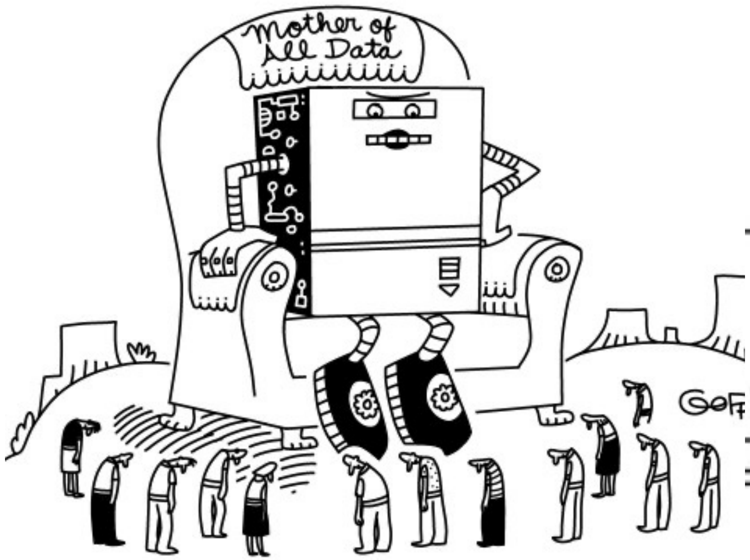
Matt Shepherd
Head of Data Strategy, BBH London

“At the heart of the digital economy and society is the explosion of insight, intelligence and information – data. Data is the lifeblood of the digital economy.

World Economic Forum
A New Paradigm for Business of Data BRIEFING PAPER - JULY 2020

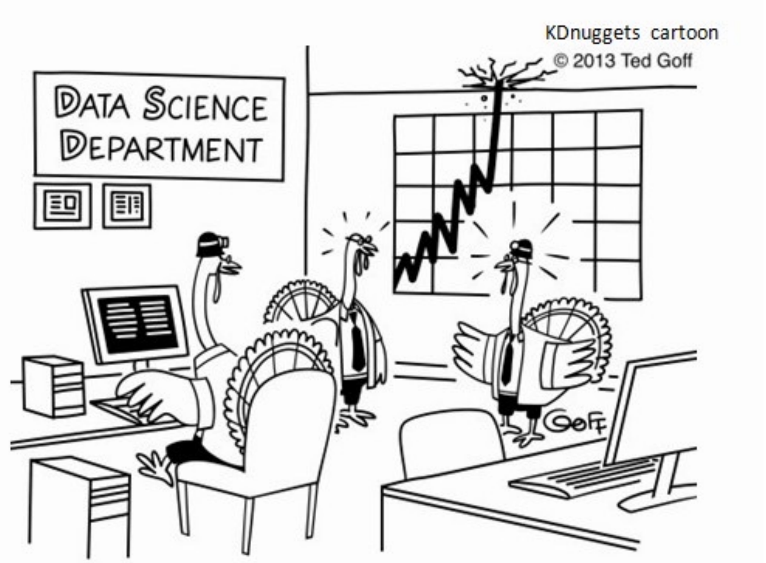
Data Science/Big Data in the News…

Data Science Everywhere!...



"No candy? No flowers? No cards? Big Data predicted that 67.53% of you would remember!"

"You can't keep adjusting the data to prove that you would be the best Valentine's date for Scarlett Johansson."



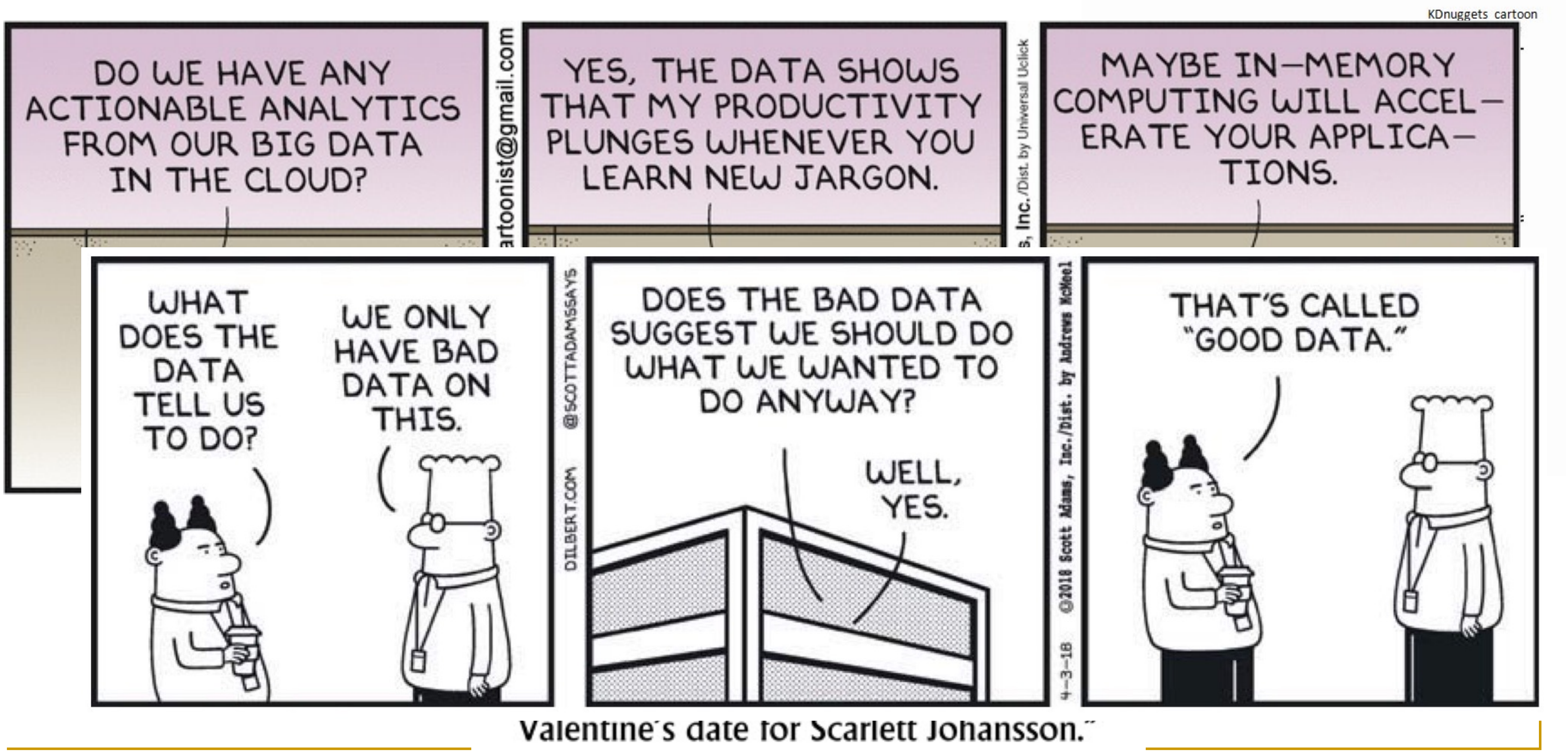
"I don't like the look of this. Searches for gravy and turkey stuffing are going through the roof !"

Data Science Everywhere!...

KDnuggets cartoon

"You can't keep adjusting the data to prove that you would be the best Valentine's date for Scarlett Johansson."

Data Science Everywhere!...



Data Science Needs Positioning

What is Data Science

Data Science Applications

Data Science Ecosystem

Data Science Lifecycle

Data Science System Architecture

Who Owns Data Science

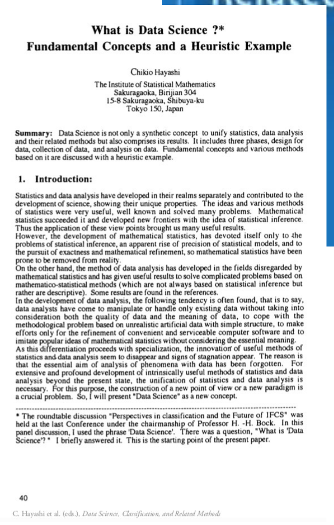
What is Data Science?

“**Data science**, also known as **data-driven science**, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.”

WIKIPEDIA
The Free Encyclopedia

What is Data Science?

“Data science intends to **analyze and understand actual phenomena with ‘data’**. In other words, the aim of data science is to reveal the features or the hidden structure of complicated natural, human, and social phenomena with data from a different point of view from the established or traditional theory and method.”



What is Data Science ?*
Fundamental Concepts and a Heuristic Example

Chikio Hayashi
The Institute of Statistical Mathematics
Sakuragaoka, Birijian 304
15-8 Sakuragaoka, Shibuya-ku
Tokyo 150, Japan

Summary: Data Science is not only a synthetic concept to unify statistics, data analysis and their related methods but also comprises its results. It includes three phases, design for data, collection of data, and analysis on data. Fundamental concepts and various methods based on it are discussed with a heuristic example.

1. Introduction:

Statistics and data analysis have developed in their realms separately and contributed to the development of science, showing their unique properties. The ideas and various methods of statistics were very useful, well known and solved many problems. Mathematical statistics succeeded it and developed new frontiers with the idea of statistical inference. Thus the application of these view points brought us many useful results.
However, the development of mathematical statistics, has devoted itself only to the problems of statistical inference, an apparent rise of precision of statistical models, and to the pursuit of exactness and mathematical refinement, so mathematical statistics have been prone to be removed from reality.
On the other hand, the method of data analysis has developed in the fields disregarded by mathematical statistics and has given useful results to solve complicated problems based on mathematico-statistical methods (which are not always based on statistical inference but rather are descriptive). Some results are found in the references.
In the development of data analysis, the following tendency is often found, that is to say, data analysts have come to manipulate or handle only existing data without taking into consideration both the quality of data and the meaning of data, to cope with the methodological problem based on unrealistic artificial data with simple structure, to make efforts only for the refinement of convenient and serviceable computer software and to imitate popular ideas of mathematical statistics without considering the essential meaning.
As this differentiation proceeds with specialization, the innovation of useful methods of statistics and data analysis seem to disappear and signs of stagnation appear. The reason is that the essential aim of analysis of phenomena with data has been forgotten. For extensive and profound development of intrinsically useful methods of statistics and data analysis beyond the present state, the unification of statistics and data analysis is necessary. For this purpose, the construction of a new point of view or a new paradigm is a crucial problem. So, I will present "Data Science" as a new concept.

* The roundtable discussion "Perspectives in classification and the Future of IFCS" was held at the last Conference under the chairmanship of Professor H. -H. Bock. In this panel discussion, I used the phrase 'Data Science'. There was a question, "What is 'Data Science'?" I briefly answered it. This is the starting point of the present paper.

40

C. Hayashi et al. (eds.), *Data Science, Classification, and Related Methods*

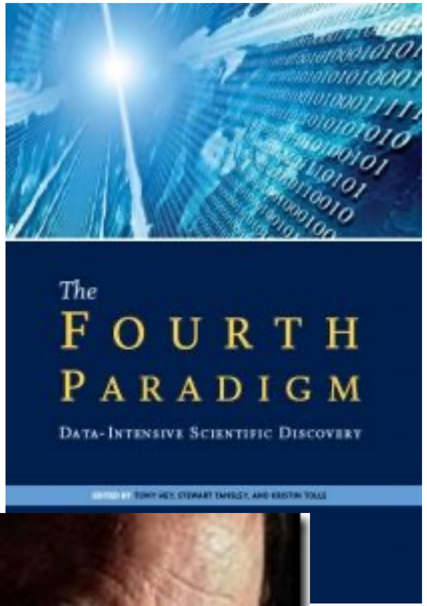
What is Data Science?

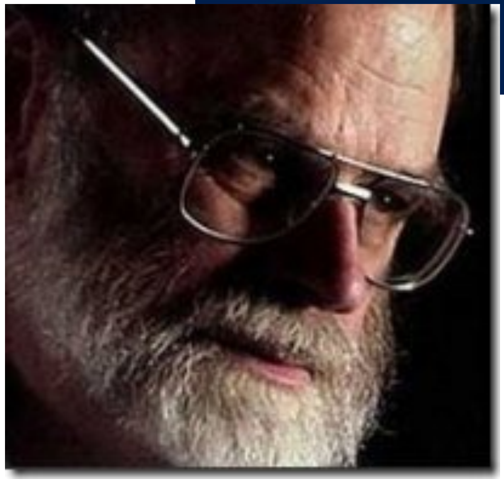
"Data science intends to **analyze and understand actual phenomena with 'data'**. In other words, the aim of data science is to reveal the features or the hidden structure of complicated natural, human, and social phenomena with data from a different point of view from the established or traditional theory and method."

* The roundtable discussion "Perspectives in classification and the Future of IFCS" was held at the last Conference under the chairmanship of Professor H. -H. Bock. In this panel discussion, I used the phrase 'Data Science'. There was a question, "What is 'Data Science'?" I briefly answered it. This is the starting point of the present paper.

What is Data Science?

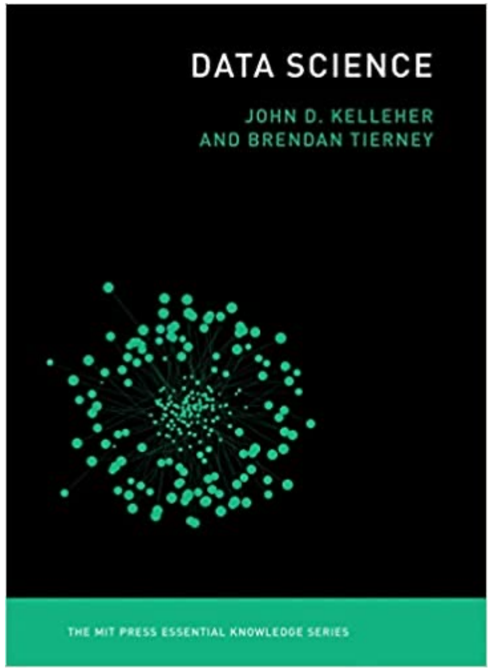
- Fourth paradigm
  - "… change of all sciences moving from observational, to theoretical, to computational and now to the 4th Paradigm – Data-Intensive Scientific Discovery"





What is Data Science?

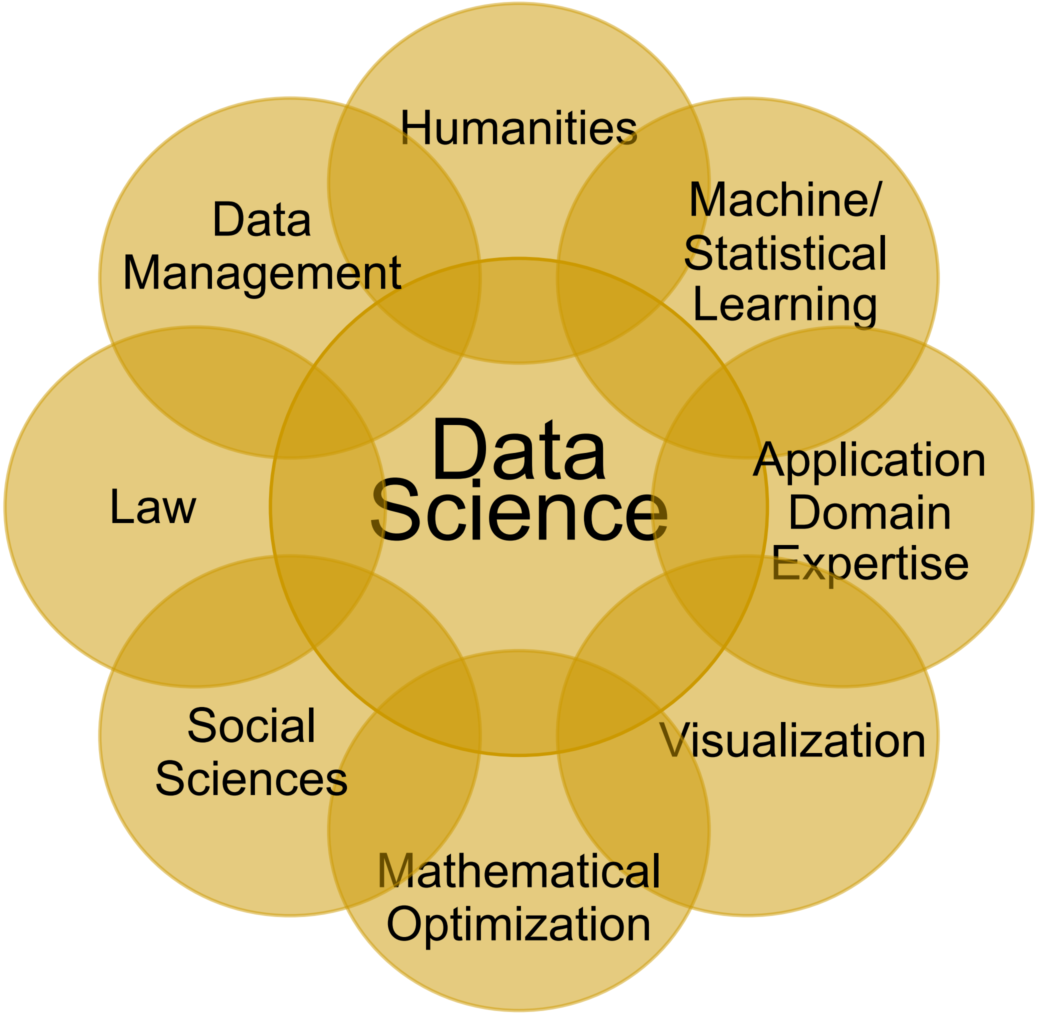
- “Data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting non-obvious and useful patterns from large data sets.”
- “…the terms *data science*, *machine learning*, and *data mining* are often used interchangeably.”
- “…although data science borrows from these other fields, it is broader in scope.”



A Working Definition

A data-driven approach to problem solving that involves the process of collecting, managing, analyzing, explaining and visualizing data and analysis results.

Data Science as a Unifier



Who is a Data Scientist?

To be revealed at the end…

Two Myths…

- Data science = Big data

Two Myths…

- Data science $\neq$ Big data
- Big data is like a raw material
- Processing it leads to data science & better understanding
- Applications are important
  - No applications $\rightarrow$ no data science

Two Myths…

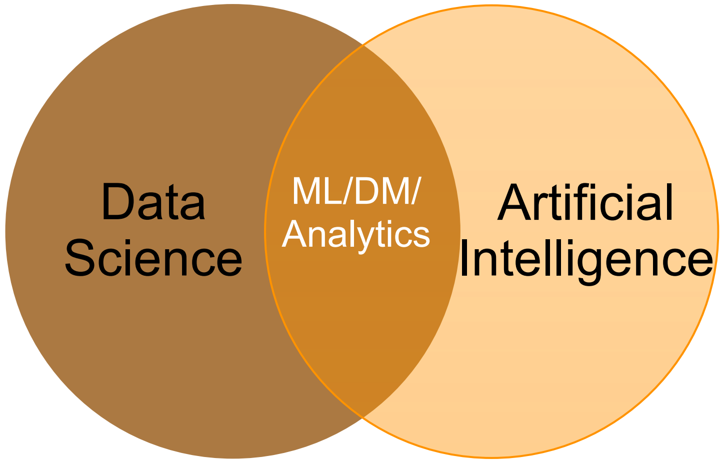
- Data science $\neq$ Big data
- Big data is like a raw material
- Processing it leads to data science & better understanding
- Applications are important
  - No applications $\rightarrow$ no data science

- Data science $\subseteq$ Machine learning $\subset$ AI

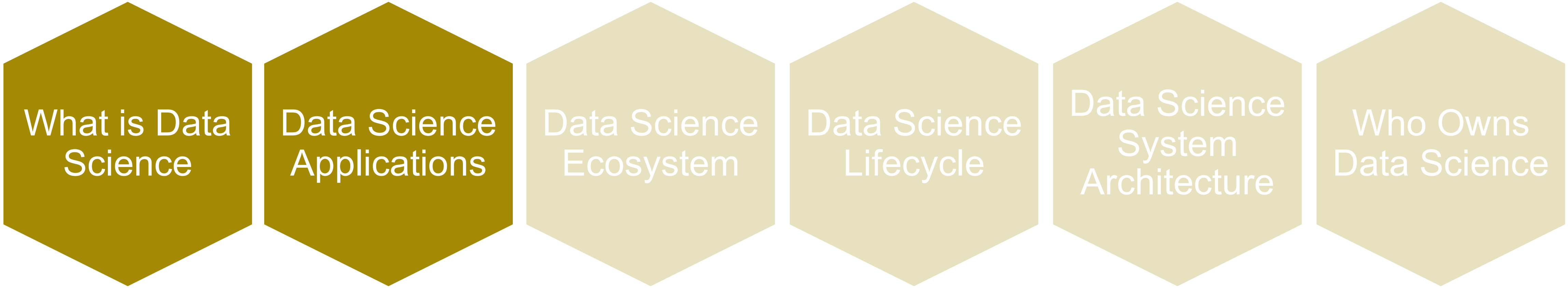
Two Myths…

- Data science $\neq$ Big data
- Big data is like a raw material
- Processing it leads to data science & better understanding
- Applications are important
  - No applications $\rightarrow$ no data science

- Data science $\not\subseteq$ Machine learning $\not\subseteq$ AI



- They are related but not the same


What is Data Science
Data Science Applications
Data Science Ecosystem
Data Science Lifecycle
Data Science System Architecture
Who Owns Data Science

Data Science Applications

- Data science is about applications
  - Applications give purpose
  - Applications inform core technologies
- Almost any field with large data sets are good candidates
- Some examples
  - Fraud detection
  - Biological & biomedical applications
  - Recommender systems
  - Health sciences & health informatics applications
  - Sustainability
  - Finance & insurance
  - Smart cities
  - Sports
  - …

Data Science Application Examples

- Fraud detection
  - Investigate fraud patterns in past data
  - Early detection is important
    - Before damage propagates
    - Harder than late detection
  - Precision is important
    - False positive and false negative are both bad
  - Real-time analytics

Data Science Application Examples

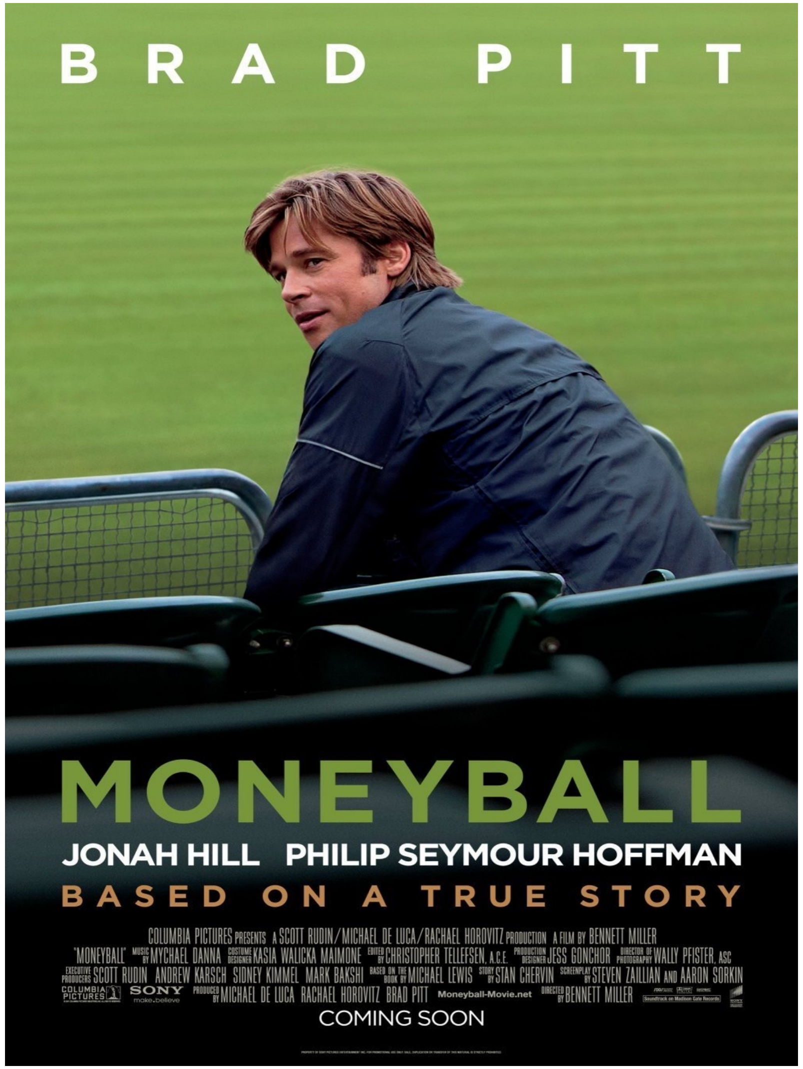
- Recommender systems
  - The ability to offer unique personalized service
  - Increase sales, click-through rates, conversions, …
  - Collaborative filtering at scale

Data Science Application Examples

- Sustainability
  - Climate variability and change
  - Ecology
  - FEW
    - Food
    - Energy
    - Water

Data Science Application Examples

- Moneyball
  - How to build a baseball team on a very low budget by relying on data
  - *Sabermetrics*: the statistical analysis of baseball data to objectively evaluate performance
  - 2002 record of 103-59 was joint best in MLB
    - Team salary budget: $40 million
  - Other team: Yankees
    - Team salary budget: $120 million



What is Data Science

Data Science Applications

Data Science Ecosystem

Data Science Lifecycle

Data Science System Architecture

Who Owns Data Science

Data Science Ecosystem

Data Science Building Blocks

Data Engineering

- Data quality
- Big Data storage and computing solutions
- Data pipelines (ETL)

Data Analytics

- Explore data (data mining)
- Build models & algorithms (machine learning)
- Visualizations & visual analytics

Data Security & Privacy

- Differential privacy
- Applications of cryptography
- Data integrity

Data Ethics

- Impact on individuals, organizations & society
- Ethical & normative concerns
- Regulatory issues

Data Engineering

Big data management
(Four Vs)

Data Engineering

Big data management (Four Vs)

- Data processing platforms
- Data integration
  - ETL process
  - Data lakes
- Data quality issues
- Data provenance

Data Engineering Essential

Data Engineering Essential

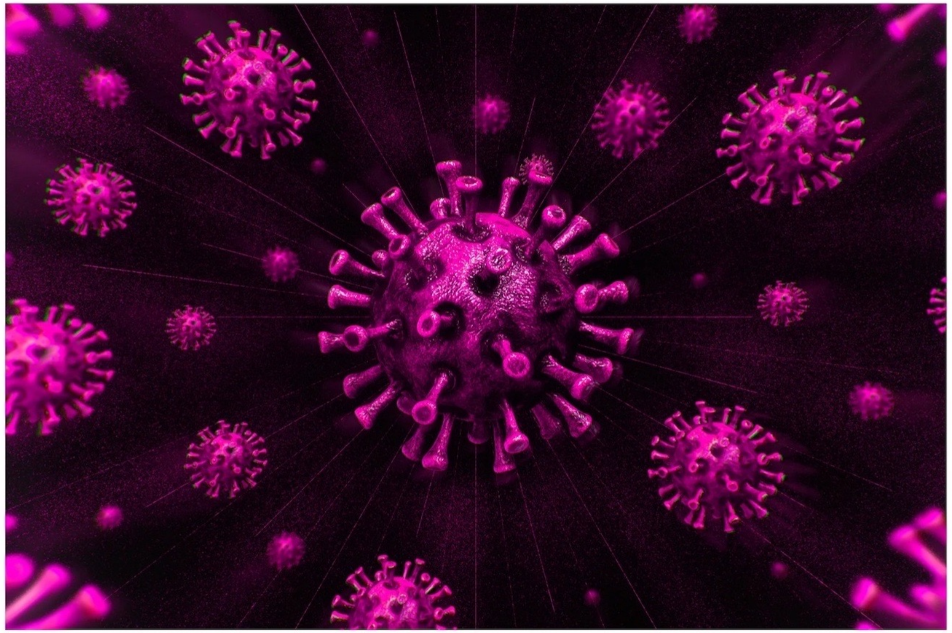
Data Engineering Essential

Excel spreadsheet error blamed for UK's 16,000 missing coronavirus cases

The case went missing after the spreadsheet hit its filesize limit

By James Vincent | Oct 5, 2020, 9:41am EDT

SHARE



Data Engineering Essential

THE VERGE TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT VIDEO MORE

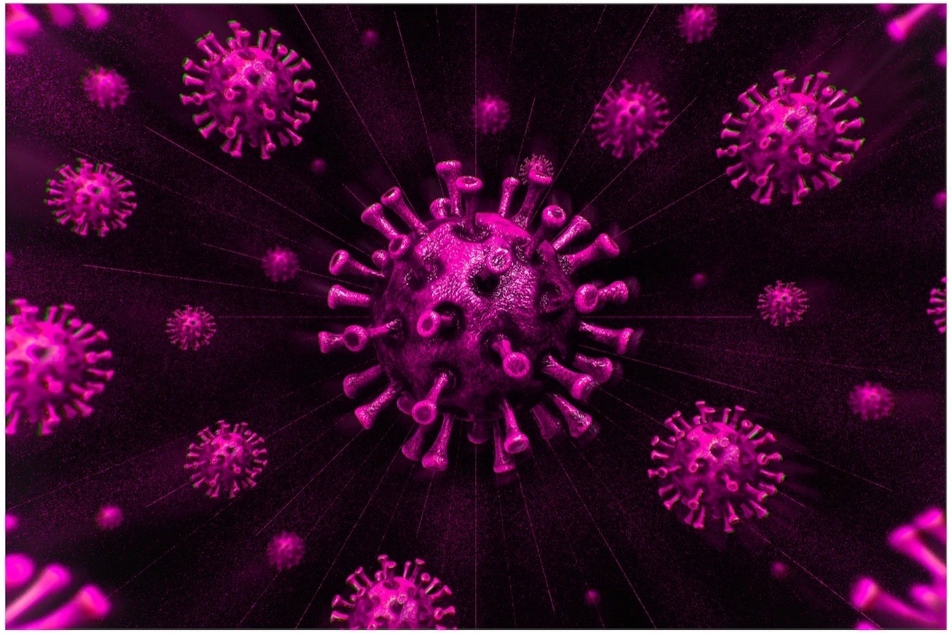
SCIENCE US & WORLD TECH

Excel spreadsheet error blamed for UK's 16,000 missing coronavirus cases

The case went missing after the spreadsheet hit its filesize limit

By James Vincent | Oct 5, 2020, 9:41am EDT

SHARE



"THE ISSUE WAS CAUSED BY THE FACT THAT SOME FILES CONTAINING POSITIVE TEST RESULTS EXCEEDED THEIR MAXIMUM FILE SIZE"

Data Engineering Essential

THE VERGE TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT VIDEO MORE

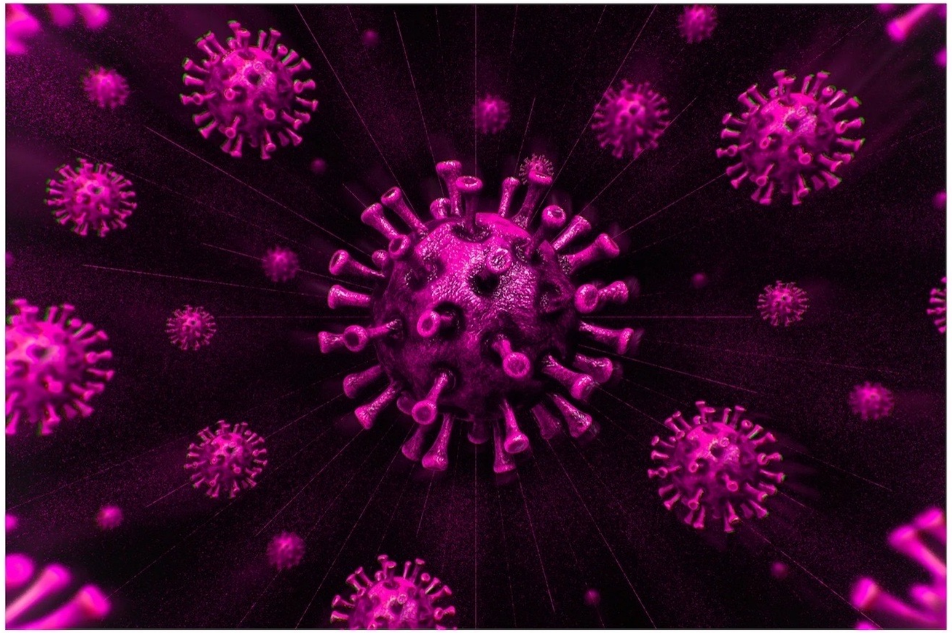
SCIENCE \ US & WORLD \ TECH

Excel spreadsheet error blamed for UK's 16,000 missing coronavirus cases

The case went missing after the spreadsheet hit its filesize limit

By James Vincent | Oct 5, 2020, 9:41am EDT

SHARE



Under-reported figures

From 25 Sept to 2 Oct

50,786

Cases initially reported by PHE

15,841

Unreported cases, missed due to IT error

8 days of incomplete data

1,980 cases per day, on average, were missed in that time

48 hours Ideal time limit for tracing contacts after positive test

Source: PHE and gov.uk

Getty Images

Data Engineering Essential

THE VERGE | TECH | REVIEWS | SCIENCE | CREATORS | ENTERTAINMENT | VIDEO | MORE

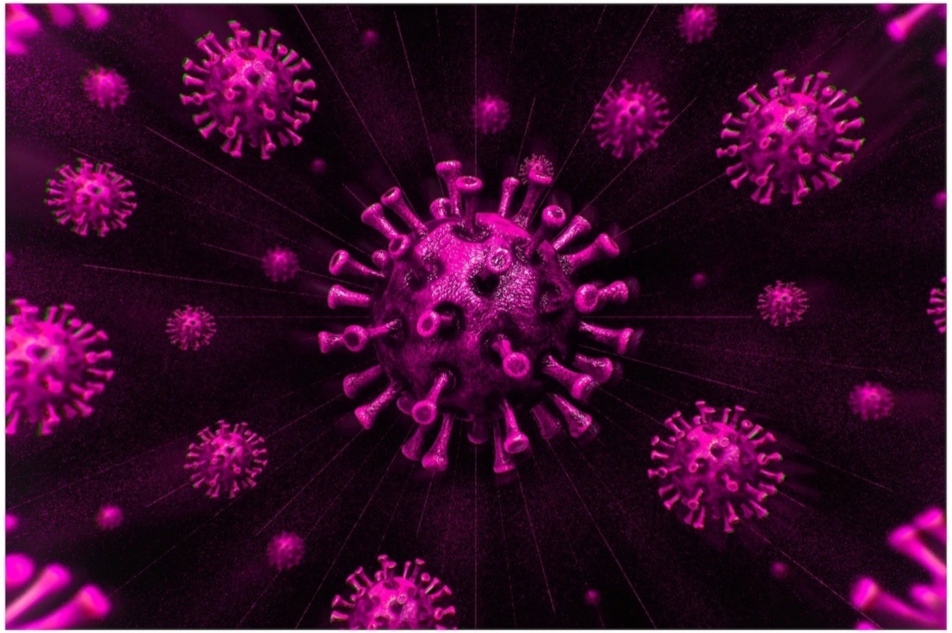
SCIENCE \ US & WORLD \ TECH

Excel spreadsheet error blamed for UK's 16,000 missing coronavirus cases

The case went missing after the spreadsheet hit its filesize limit

By James Vincent | Oct 5, 2020, 9:41am EDT

SHARE



Under-reported figures

From 25 Sept to 2 Oct

50,786

Cases initially reported by PHE

15,841

Unreported cases, missed due to IT error

8 days of incomplete data

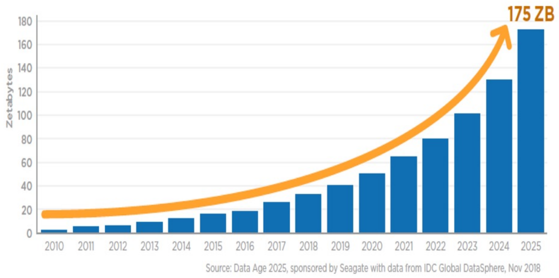
1,980 cases per day, on average, were missed in that time

48 hours Ideal time limit for tracing contacts after positive test

Source: PHE and gov.uk

Getty Images

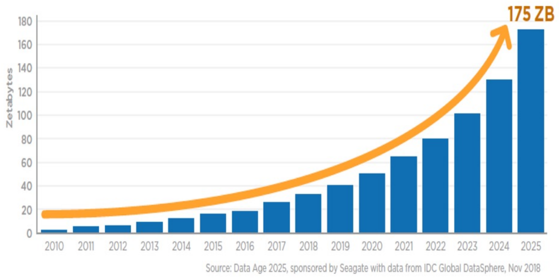
Big Data – Four Vs



Volume

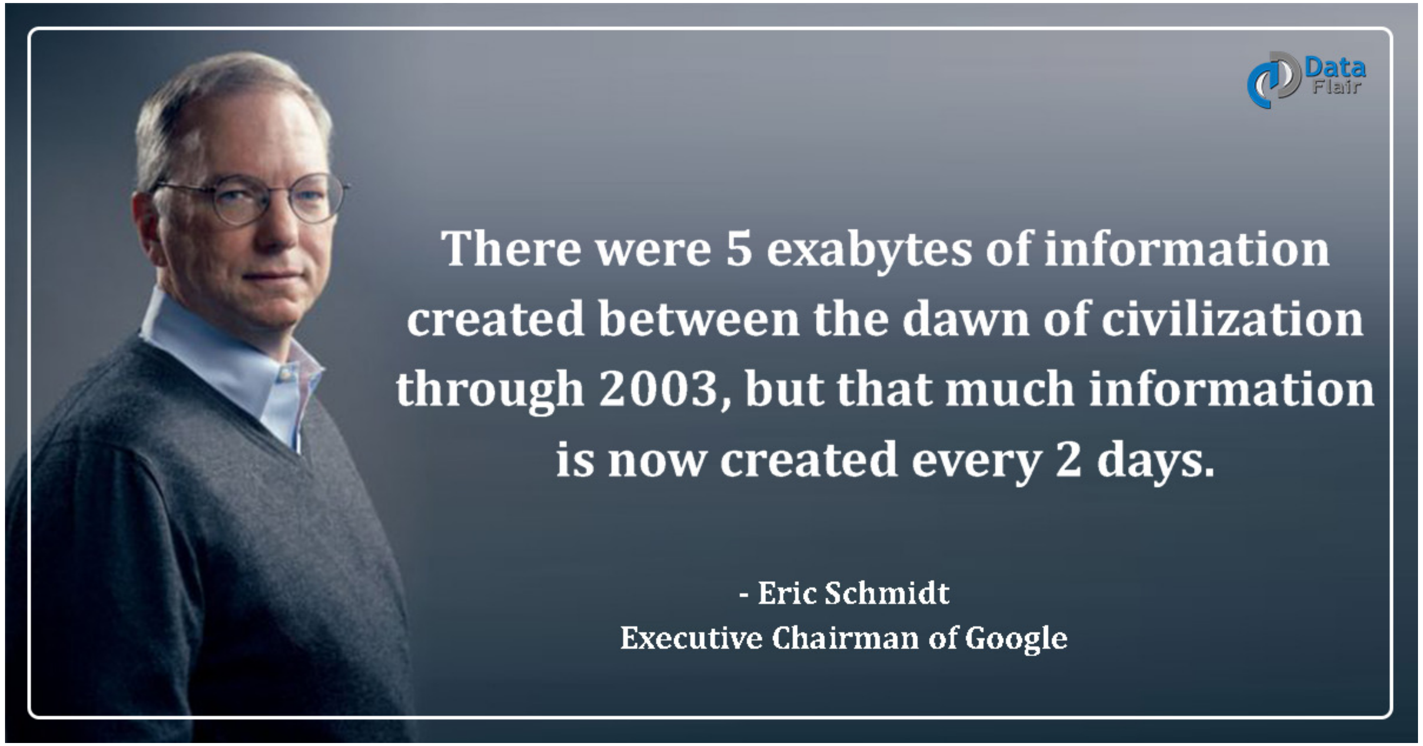
- Scale of data
- Data at rest

Big Data – Four Vs

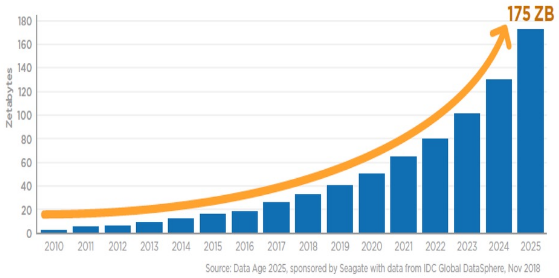


Volume

- Scale of data
- Data at rest

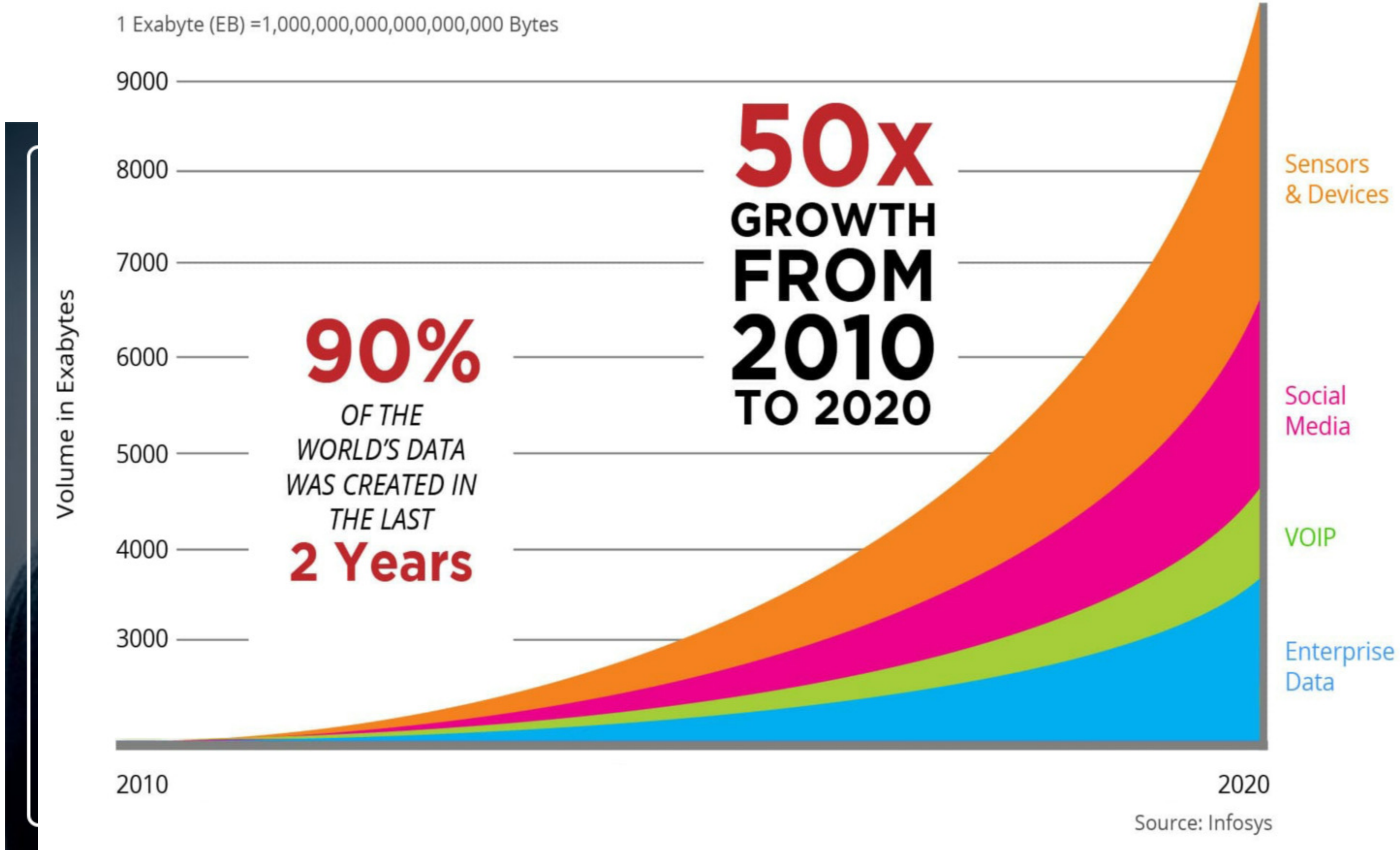


Big Data – Four Vs

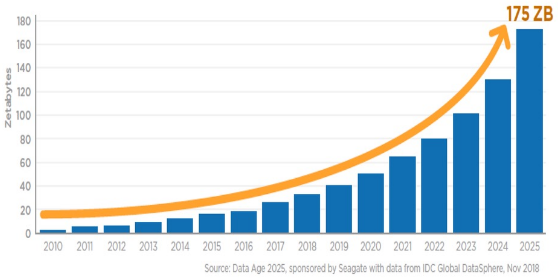


Volume

- Scale of data
- Data at rest



Big Data – Four Vs



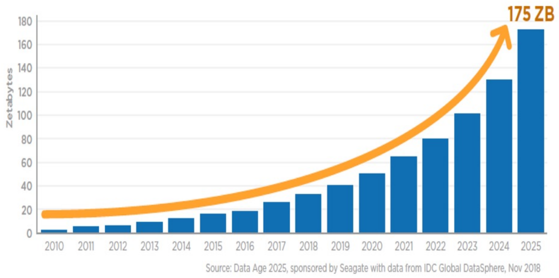
Volume

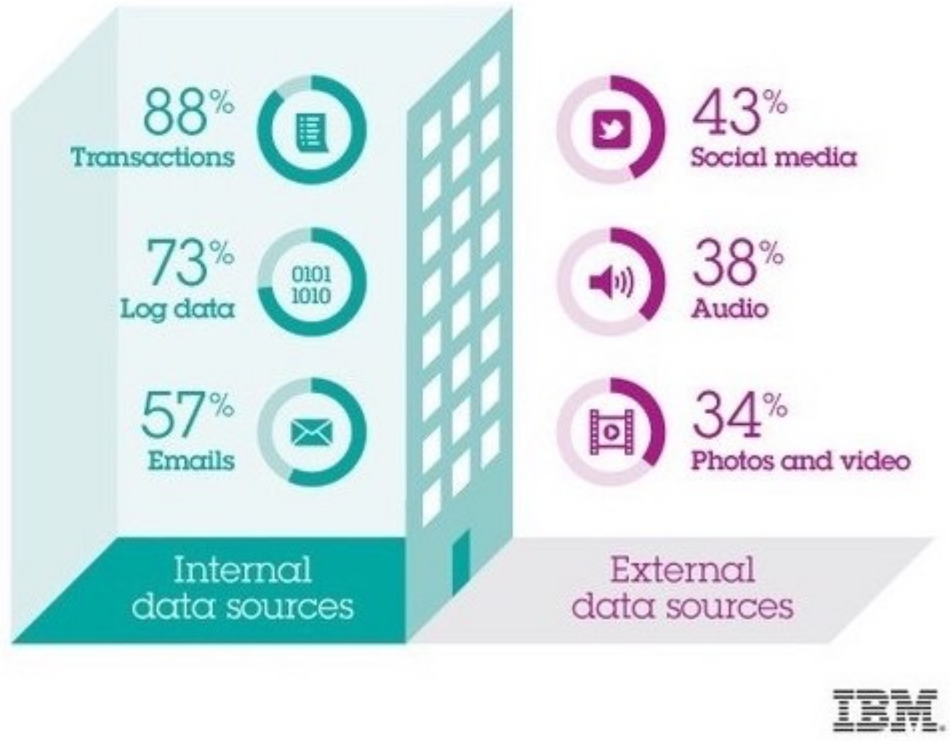
- Scale of data
- Data at rest

Variety

- Forms of data
- Unstructured challenges

Big Data – Four Vs

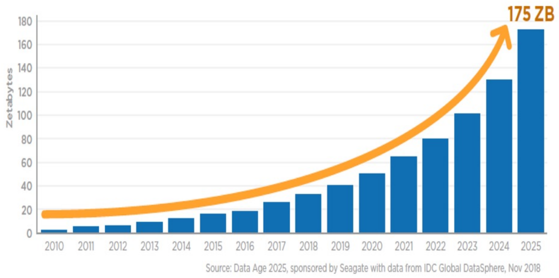




Volume
• Scale of data • Data at rest

Variety
• Forms of data • Unstructured challenges

Big Data – Four Vs



Volume

- Scale of data
- Data at rest

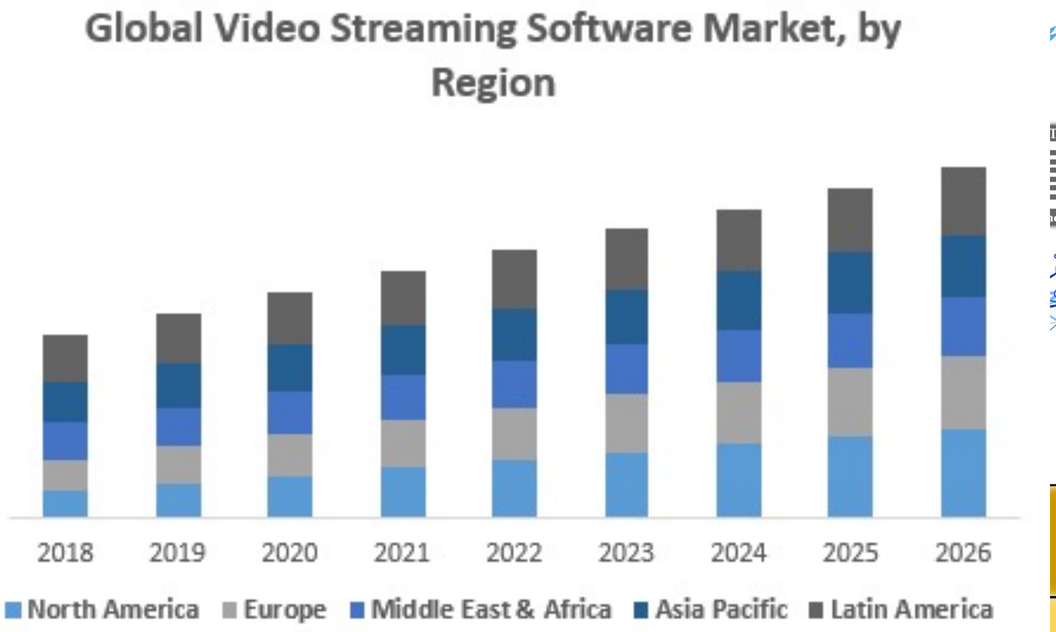
Variety

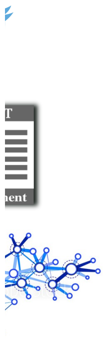
- Forms of data
- Unstructured challenges

Velocity

- Streaming data
- Data in motion

Big Data – Four Vs





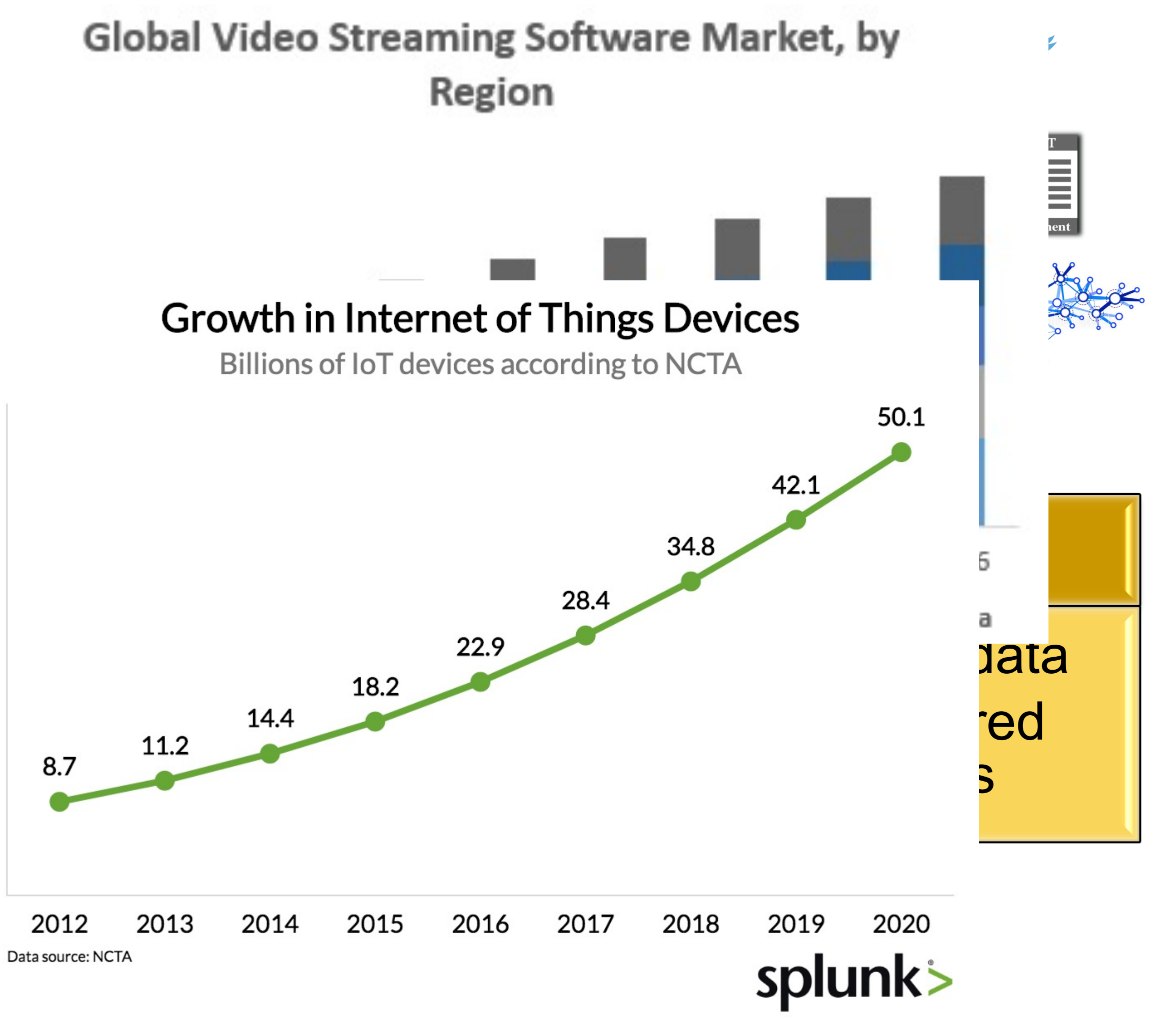
- Scale of data
- Data at rest

- Forms of data
- Unstructured challenges

Velocity

- Streaming data
- Data in motion

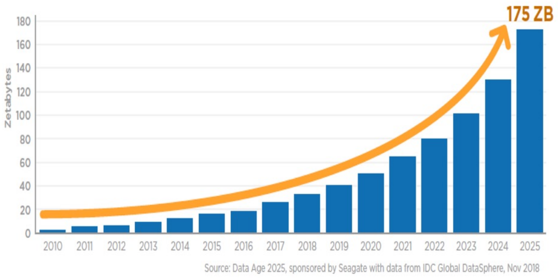
Big Data – Four Vs



Velocity

- Streaming data
- Data in motion

Big Data – Four Vs



Volume
• Scale of data • Data at rest

Variety
• Forms of data • Unstructured challenges

Velocity
• Streaming data • Data in motion

Veracity
• Uncertainty/ incorrectness in data • Data quality

Data Integration – Data Lakes

Data Quality in Big Data

89% of executives believe that data quality issues impact the quality of customer service they provide (2017)

experian

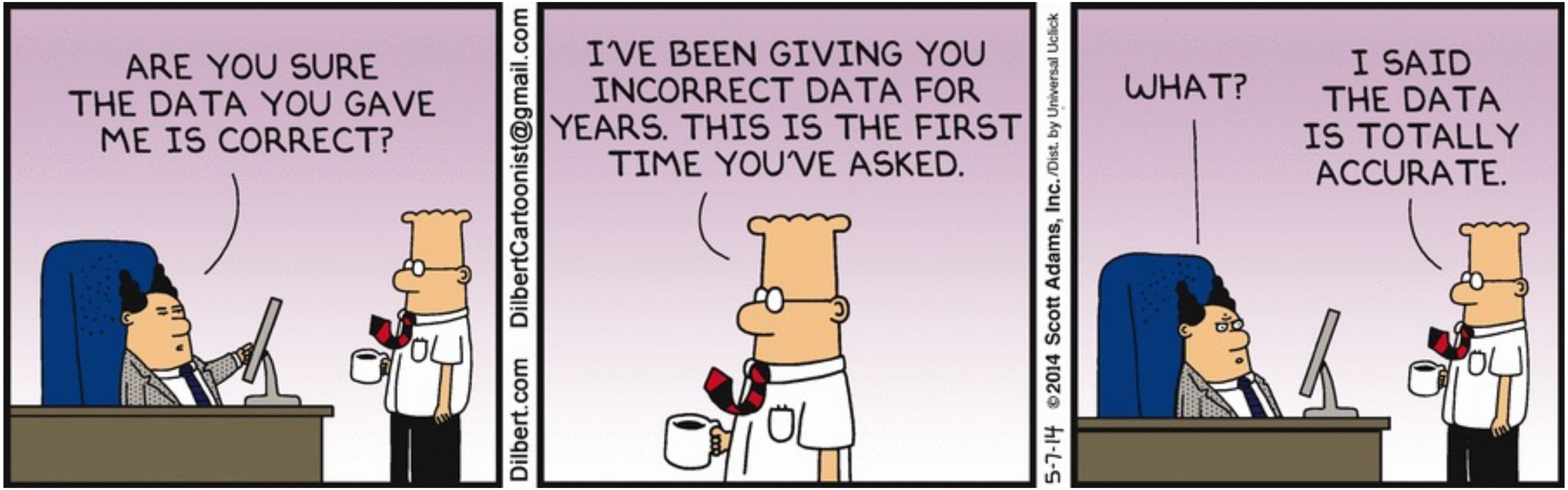
Only 33% of senior executives have a high level of trust in the accuracy of their big data analytics (2016)

KPMG

59% of executives do not believe their company has capabilities to generate business insights from their data (2016)

BAIN & COMPANY

Data Quality in Big Data



Data Quality Dimensions

DAMA UK Working Group, 2013

Data Quality Problems & Techniques

- Data unification
  - Schema mapping (if schemas exist)
  - Deduplicating records
  - Classification and mastering
- Data repair
  - Spotting errors and violations (e.g., outliers)
  - Repairing incorrect values
  - Missing value imputation

Data Analytics

The application of statistical and machine learning techniques to draw insights from data under study and to make predictions about the behaviour of the system under study

Data Analytics

The application of statistical and machine learning techniques to draw insights from data under study and to make predictions about the behaviour of the system under study

- Statistics
- Computer Science (DM/ML)

Data Analytics

The application of statistical and machine learning techniques to draw insights from data under study and to make predictions about the behaviour of the system under study

- Statistics
- Computer Science (DM/ML)

nature methods

Explore content ⌄ Journal information ⌄ Publish with us ⌄

nature › nature methods › this month › article

Published: 03 April 2018

Points of Significance

Statistics versus machine learning

Danilo Bzdok, Naomi Altman & Martin Krzywinski

Nature Methods **15**, 233–234 (2018) | Cite this article

50k Accesses | **192** Citations | **373** Altmetric | Metrics

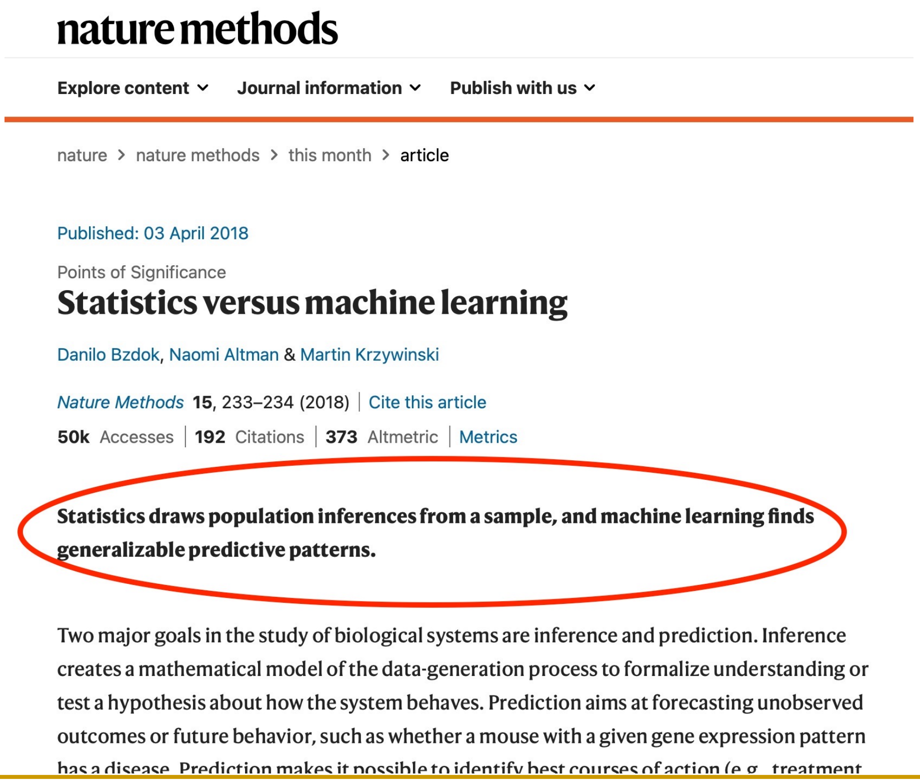
Statistics draws population inferences from a sample, and machine learning finds generalizable predictive patterns.

Two major goals in the study of biological systems are inference and prediction. Inference creates a mathematical model of the data-generation process to formalize understanding or test a hypothesis about how the system behaves. Prediction aims at forecasting unobserved outcomes or future behavior, such as whether a mouse with a given gene expression pattern has a disease. Prediction makes it possible to identify best courses of action (e.g., treatment

Data Analytics

The application of statistical and machine learning techniques to draw insights from data under study and to make predictions about the behaviour of the system under study

- Statistics
- Computer Science (DM/ML)
- The lines between the two disciplines have blurred



nature methods

Explore content ⌄ Journal information ⌄ Publish with us ⌄

nature › nature methods › this month › article

Published: 03 April 2018

Points of Significance

Statistics versus machine learning

Danilo Bzdok, Naomi Altman & Martin Krzywinski

Nature Methods **15**, 233–234 (2018) | Cite this article

50k Accesses | **192** Citations | **373** Altmetric | Metrics

Statistics draws population inferences from a sample, and machine learning finds generalizable predictive patterns.

Two major goals in the study of biological systems are inference and prediction. Inference creates a mathematical model of the data-generation process to formalize understanding or test a hypothesis about how the system behaves. Prediction aims at forecasting unobserved outcomes or future behavior, such as whether a mouse with a given gene expression pattern has a disease. Prediction makes it possible to identify best courses of action (e.g., treatment

Data Analytics Types

Descriptive

- What does the data reveals about what is happening?
- Exploratory analysis

Diagnostic

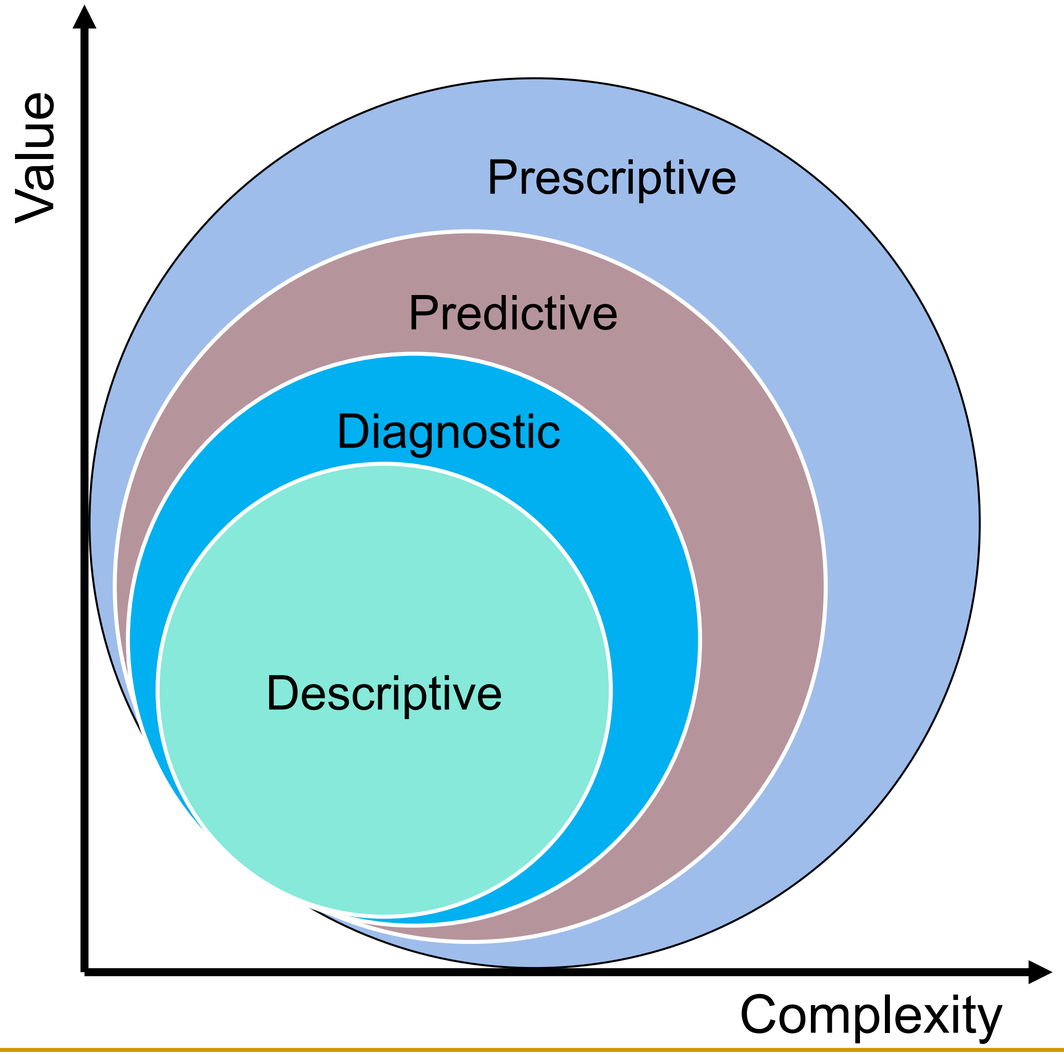
- Why is it happening?
- What does the data suggest about the reasons?

Predictive

- What is likely to happen?
- Decisions are affected
- Machine learning fits here

Prescriptive

- Recommended actions



https://www.kdnuggets.com/2017/07/4-types-data-analytics.html

Data Analytics Tasks

Clustering

- Grouping objects into clusters

Outlier detection

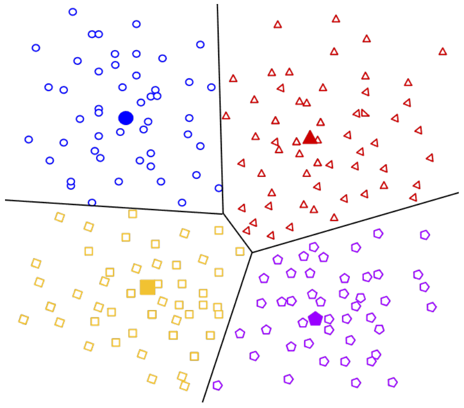
- Detection of anomalous (rare) data items

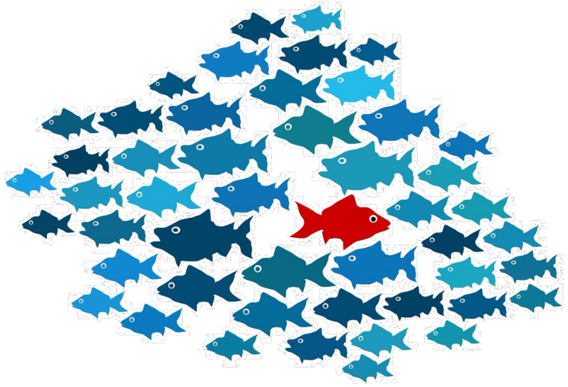
Association rule mining

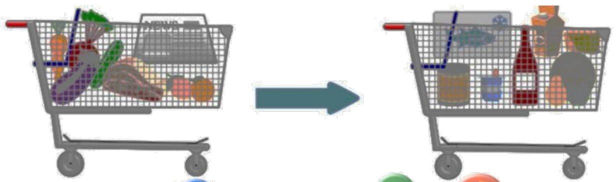
- Detecting relations between variables

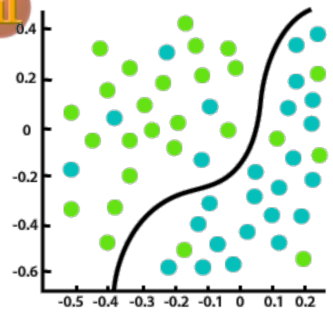
Prediction

- Classification and regression

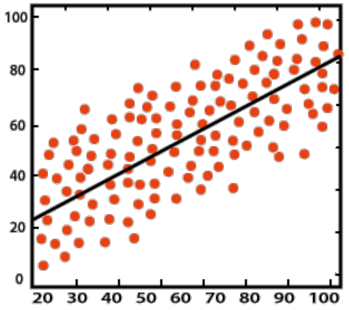








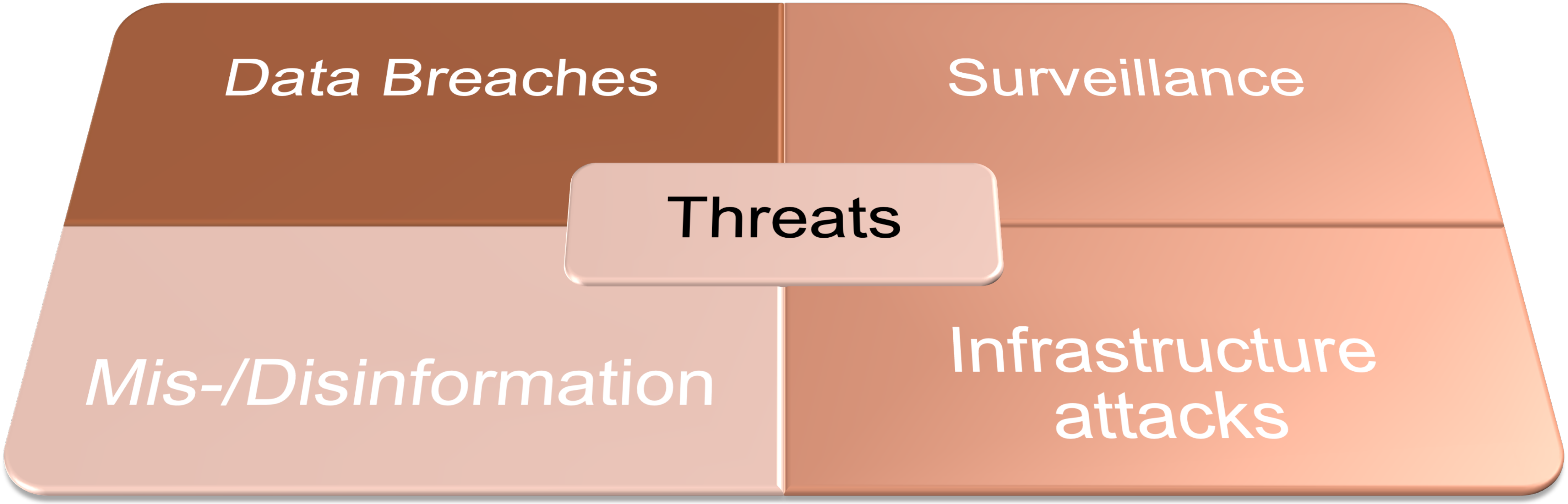
Classification



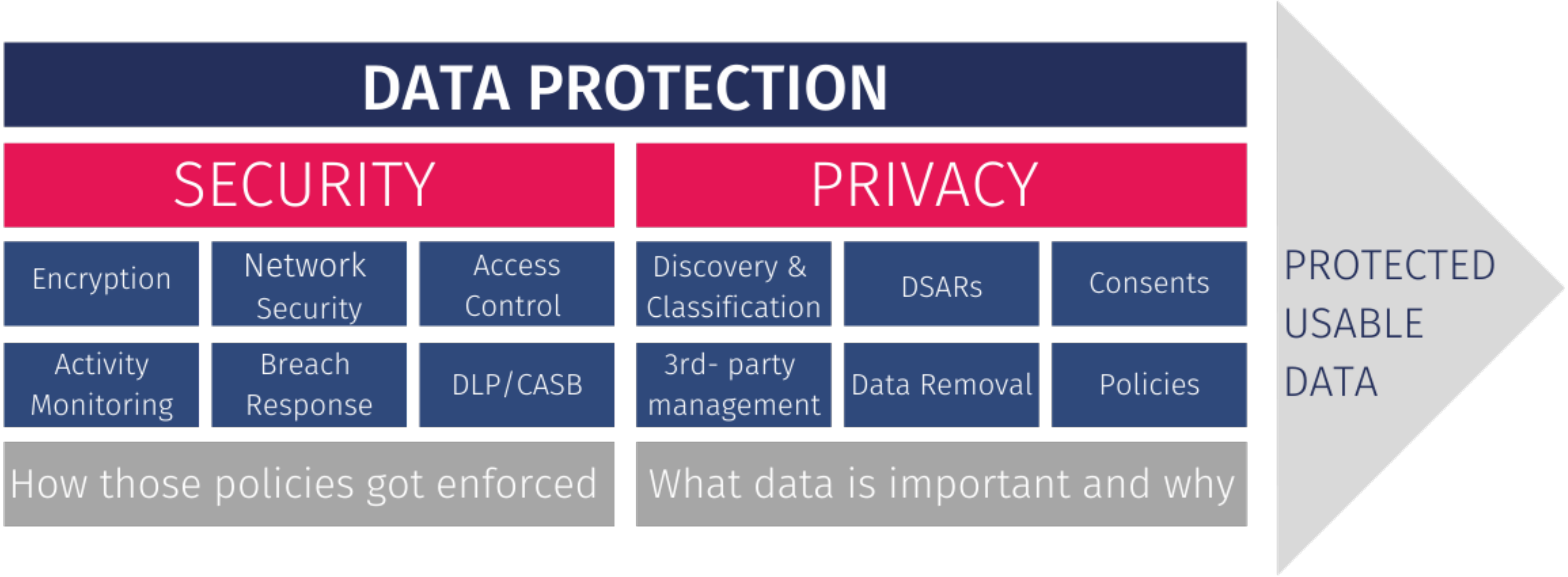
Regression

Data Security & Privacy

Big Data Privacy & Security Threats



Dimensions of Data Protection

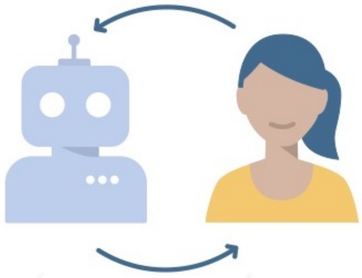


https://dataprivacymanager.net/security-vs-privacy/

Challenges

Human-in-the-loop

- Many data science processes involve humans, but controlling information in humans is different than in computer systems



Unintended side effects

- Traces of raw data persist into the latest steps of the data science process
- The combination of two data sources may reveal more than their “sum”

Distinct application requirements

- Aggregate data analysis is different from transaction analysis and different security and privacy mechanisms are needed

Inherent limitations

- Cannot have performance, accuracy (or utility) and security (or privacy) at the same time. At least one needs to go.

Different Concepts of Security

Traditional Security & Privacy

- Confidentiality
  - Do not reveal data to unauthorized users
- Integrity
  - Unauthorized users should not be able to modify data

Data Security & Privacy in Data Science

- Privacy
  - Enable users to control their data usage by others
- Veracity
  - Data provided should be true and current

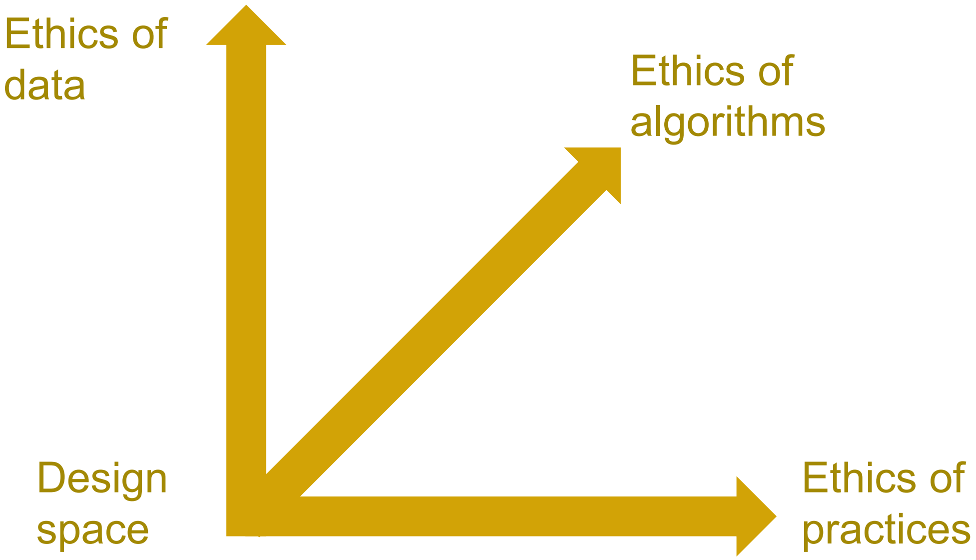
Data Ethics

"… the branch of ethics that studies and evaluates moral problems related to data, … algorithms, … and corresponding practices, in order to formulate and support morally good solutions."

L. Floridi & M. Taddeo, What is data ethics?, *Phil. Trans. R. Soc. A*, 2016.

Data Ethics

"… the branch of ethics that studies and evaluates moral problems related to data, … algorithms, … and corresponding practices, in order to formulate and support morally good solutions."



L. Floridi & M. Taddeo, What is data ethics?, *Phil. Trans. R. Soc. A*, 2016.

Ethics of Data

Ownership

- Who has ownership of data?
- Typically, individuals should have ownership

Transparency

- Subjects should know that data about them is being collected, stored and will be processed and how
- Consent

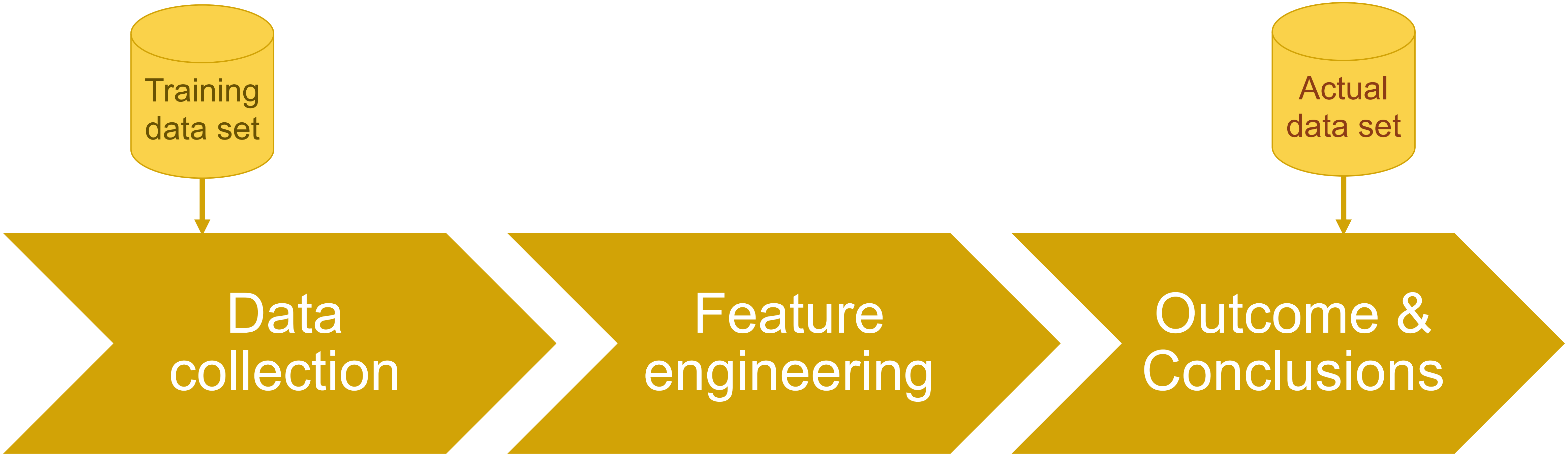
Privacy

- Personal identifiable information

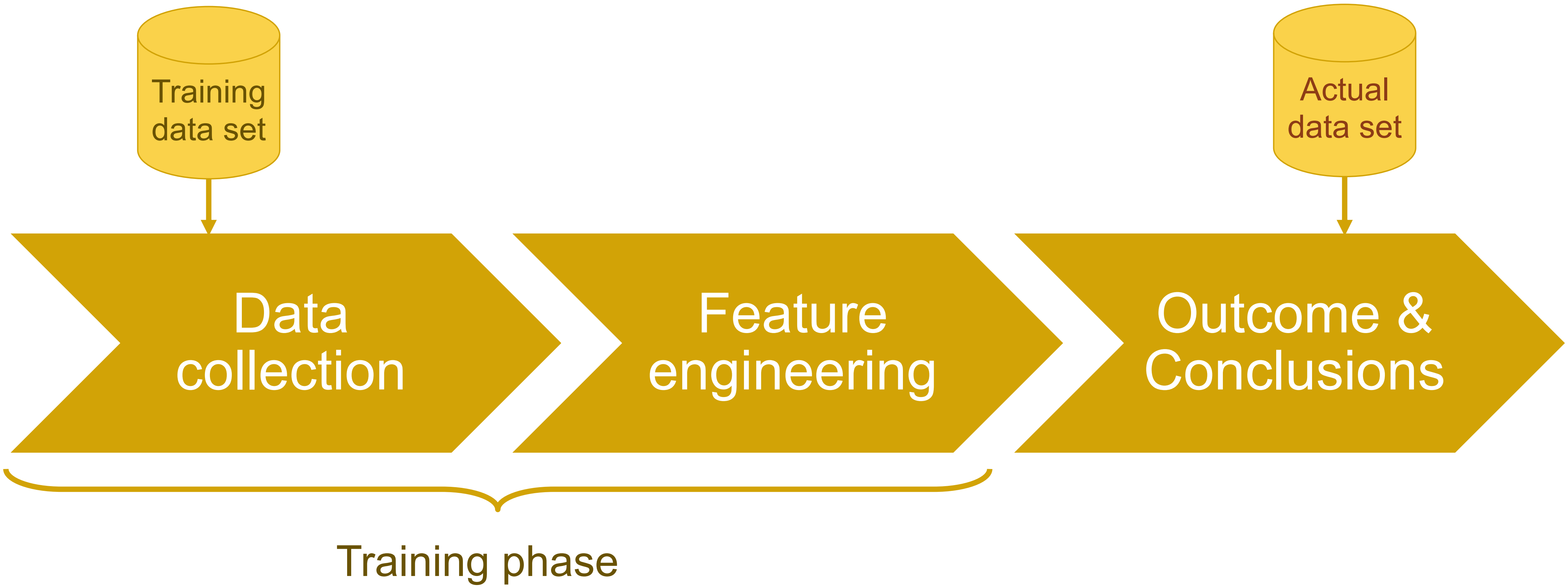
Intention

- What are you planning to do with the data?
- Secondary use

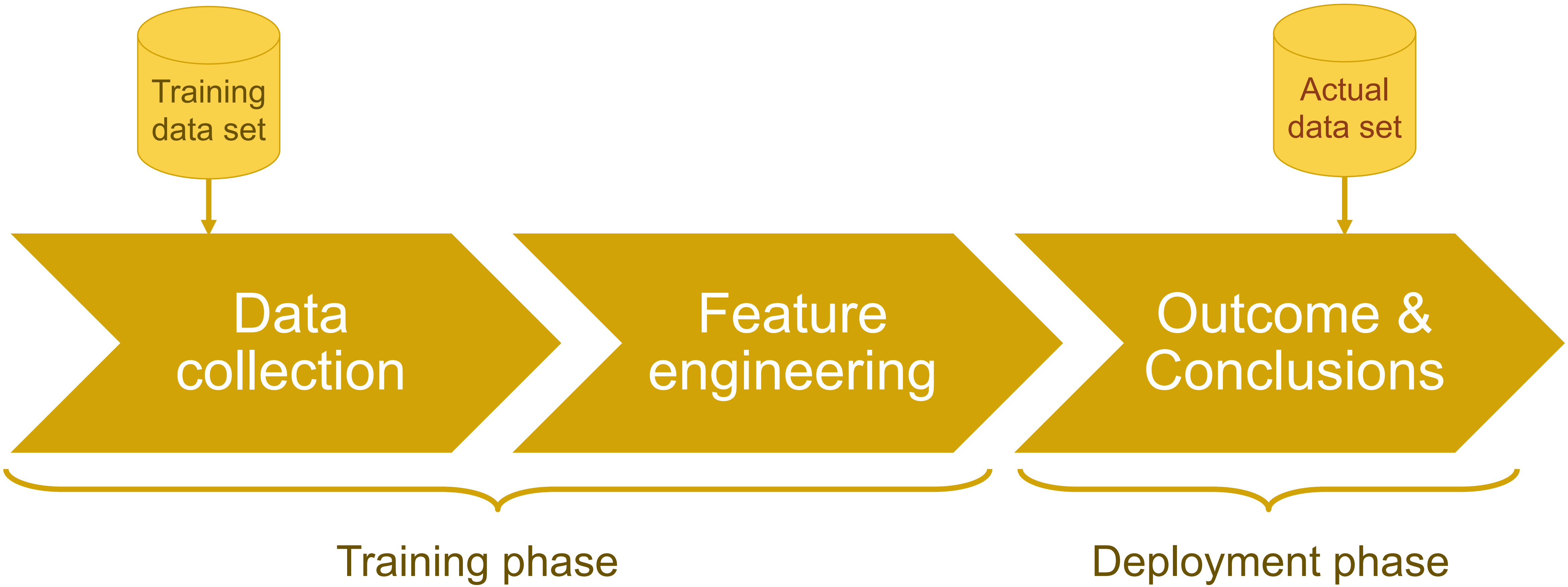
Ethics of Algorithms – Algorithmic Bias



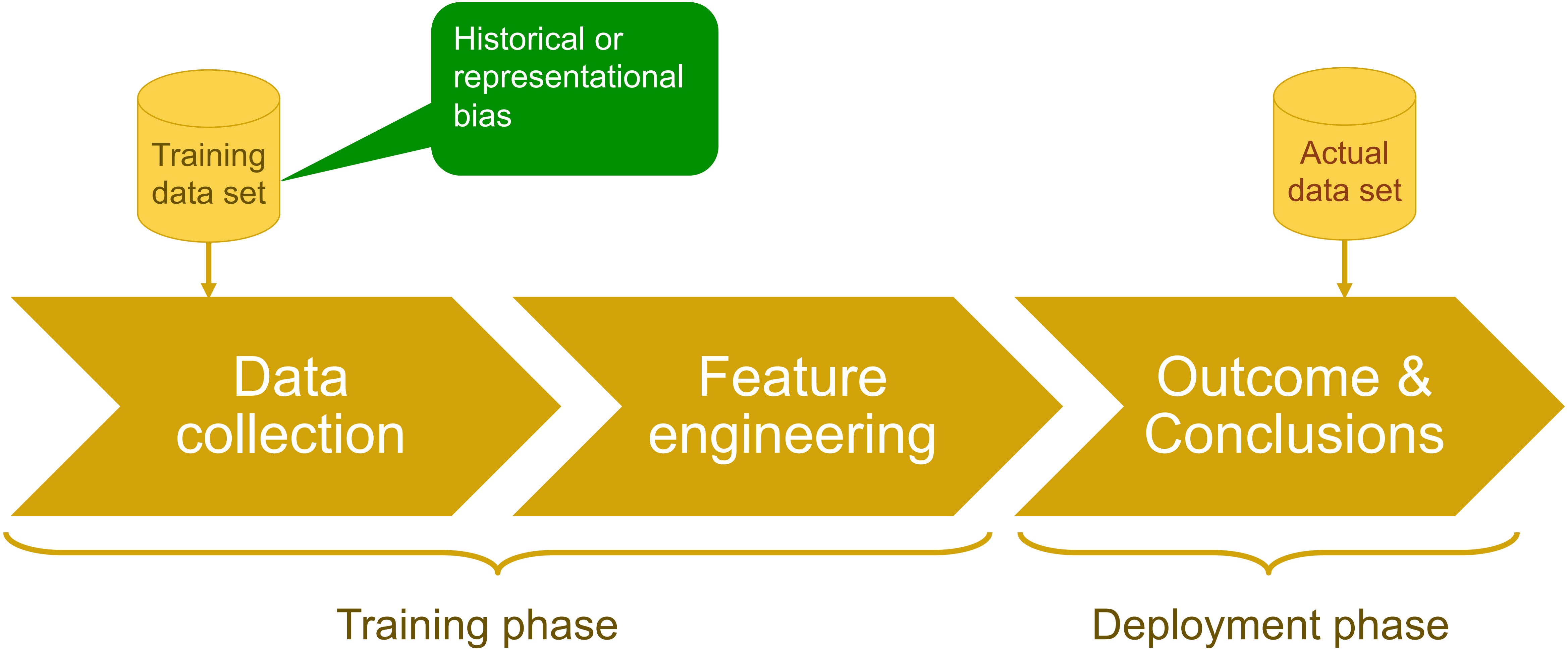
Ethics of Algorithms – Algorithmic Bias



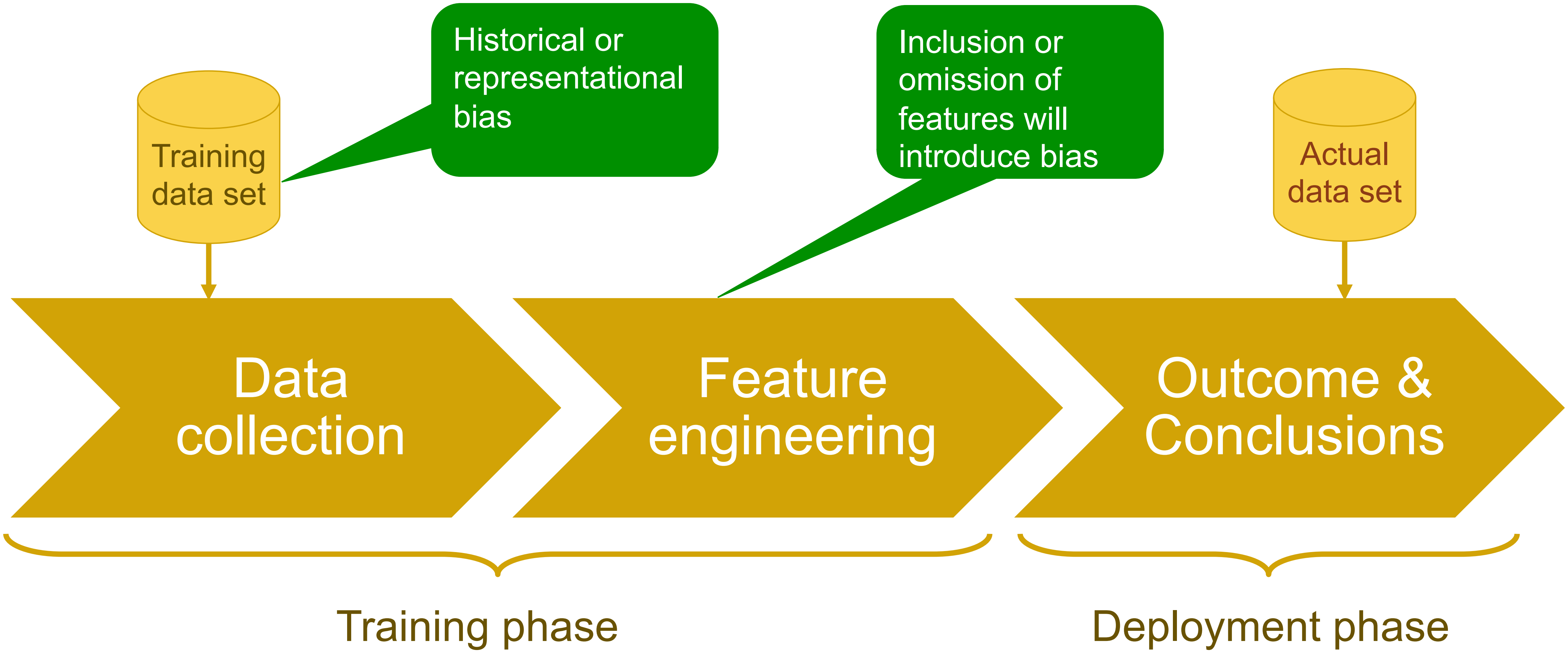
Ethics of Algorithms – Algorithmic Bias



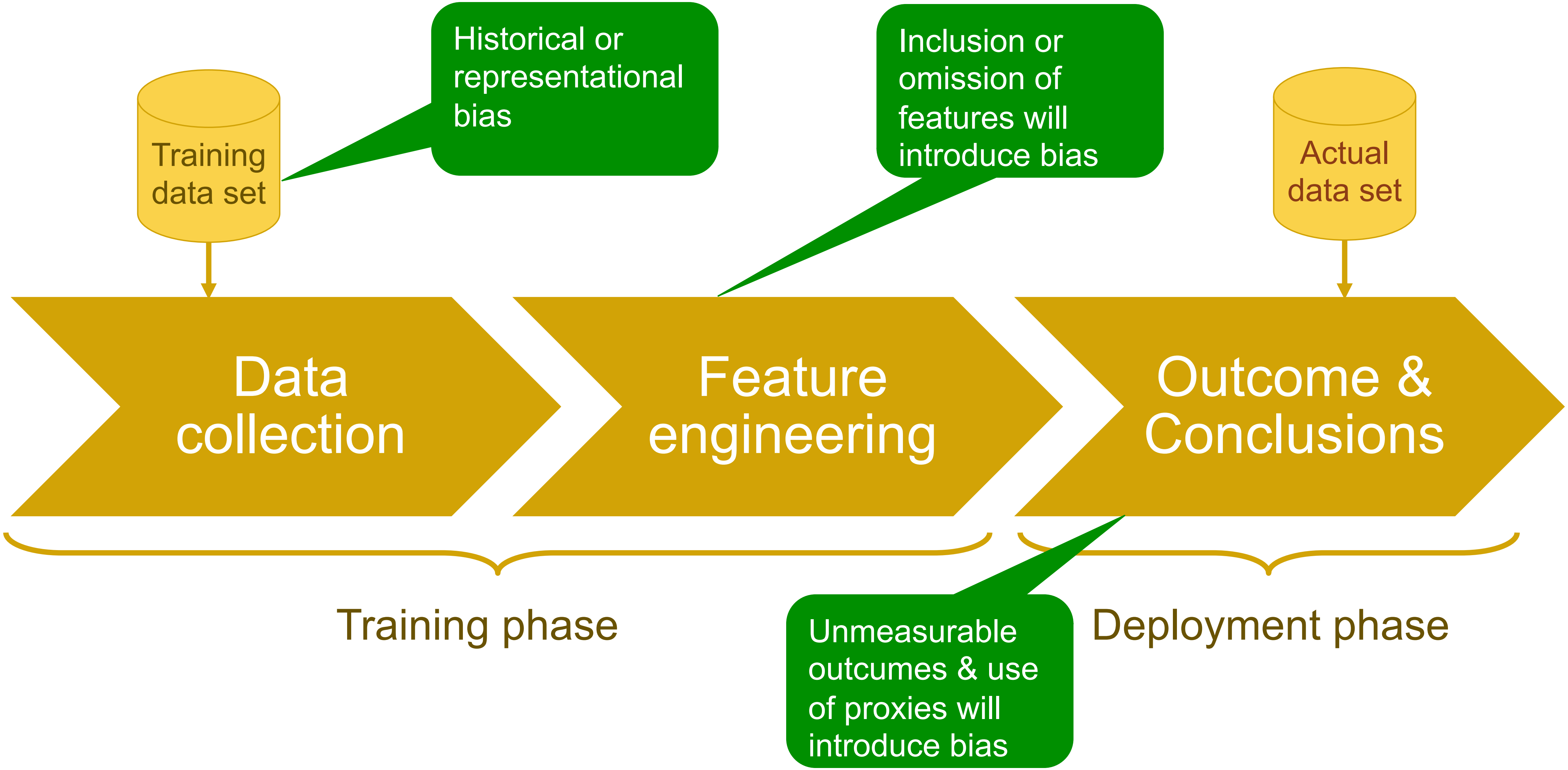
Ethics of Algorithms – Algorithmic Bias



Ethics of Algorithms – Algorithmic Bias



Ethics of Algorithms – Algorithmic Bias



Examples of Algorithmic Bias

Examples of Algorithmic Bias

SHARE

RESEARCH ARTICLE

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2,*], Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5,*,†]

+ See all authors and affiliations

Science 25 Oct 2019:
Vol. 366, Issue 6464, pp. 447-453
DOI: 10.1126/science.aax2342

Racial bias in health algorithms

The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer *et al.* find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half. Bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients. Reformulating the algorithm so that it no longer uses costs as a proxy for needs eliminates the racial bias in predicting who needs extra care.

Science, this issue p. 447; see also p. 421

Examples of Algorithmic Bias

Science | Contents | News | Careers | Journals

Read our COVID-19 research and news.

SHARE RESEARCH ARTICLE

PROPUBLICA

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low ris

Machine Bias

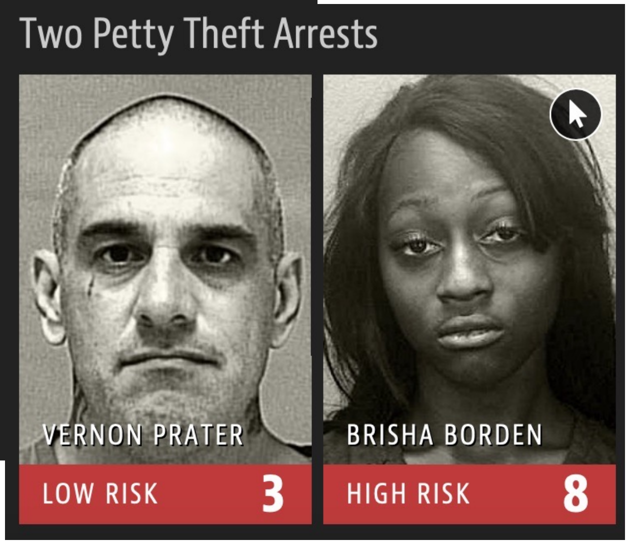
There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

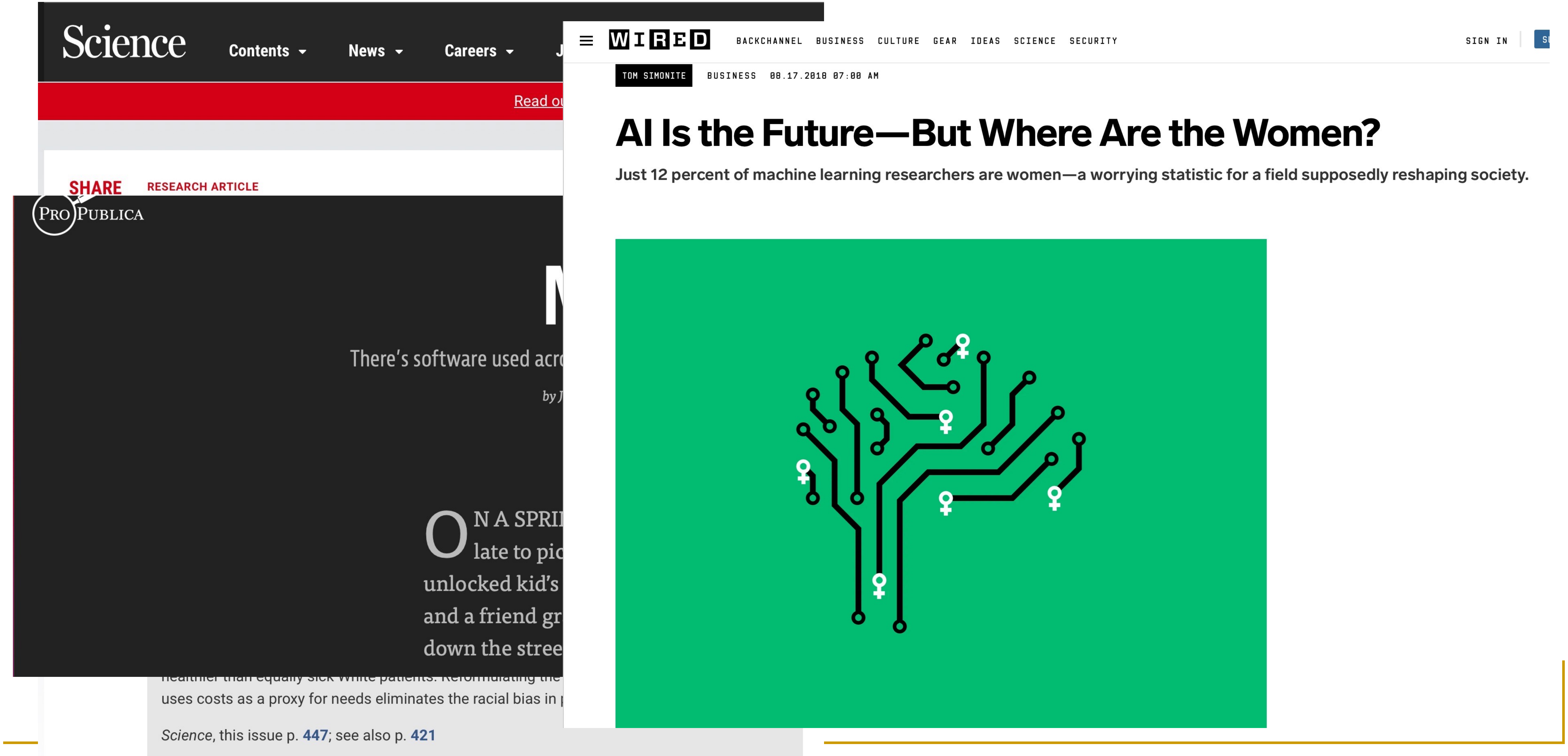
ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

uses costs as a proxy for needs eliminates the racial bias in predicting who needs extra care.

Science, this issue p. 447; see also p. 421



Examples of Algorithmic Bias



Examples of Algorithmic Bias – Gender Shades

Data Ethics Checklist

- Have we listed how this technology can be attacked or abused? [SECURITY]
- Have we tested our training data to ensure it is fair and representative? [FAIRNESS]
- Have we studied and understood possible sources of bias in our data? [FAIRNESS]
- Does our team reflect diversity of opinions, backgrounds, and kinds of thought? [FAIRNESS]
- What kind of user consent do we need to collect to use the data? [PRIVACY/TRANSPARENCY]
- Do we have a mechanism for gathering consent from users? [TRANSPARENCY]
- Have we explained clearly what users are consenting to? [TRANSPARENCY]
- Do we have a mechanism for redress if people are harmed by the results? [TRANSPARENCY]
- Can we shut down this software in production if it is behaving badly?
- Have we tested for fairness with respect to different user groups? [FAIRNESS]
- Have we tested for disparate error rates among different user groups? [FAIRNESS]
- Do we test and monitor for model drift to ensure our software remains fair over time? [FAIRNESS]
- Do we have a plan to protect and secure user data? [SECURITY]

Issues at the Intersections

- Data science components should not be siloed
- Many important problems at the intersections remain to be solved
- Examples
  - Data visualization – Visual analytics
  - Data management – Machine Learning
    - DM for ML
    - ML for DM
  - Privacy & security – Ethics

What is Data Science

Data Science Applications

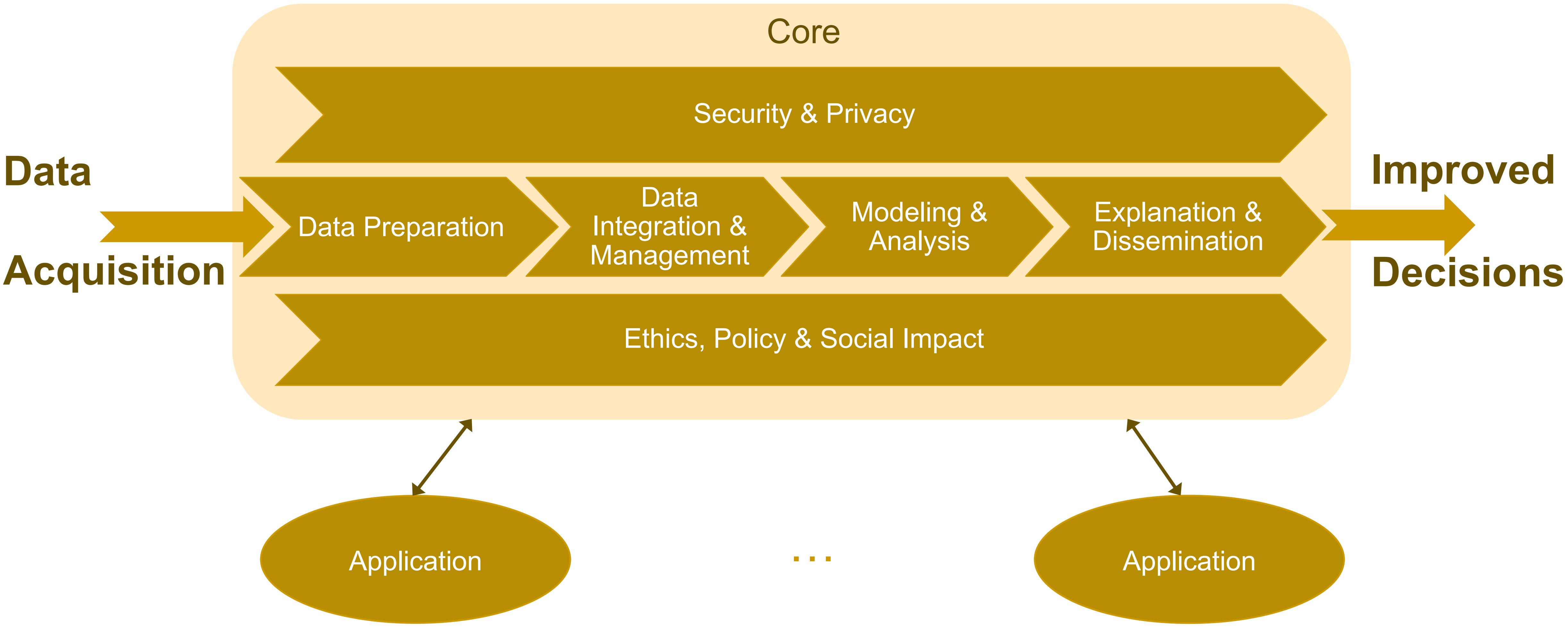
Data Science Ecosystem

Data Science Lifecycle

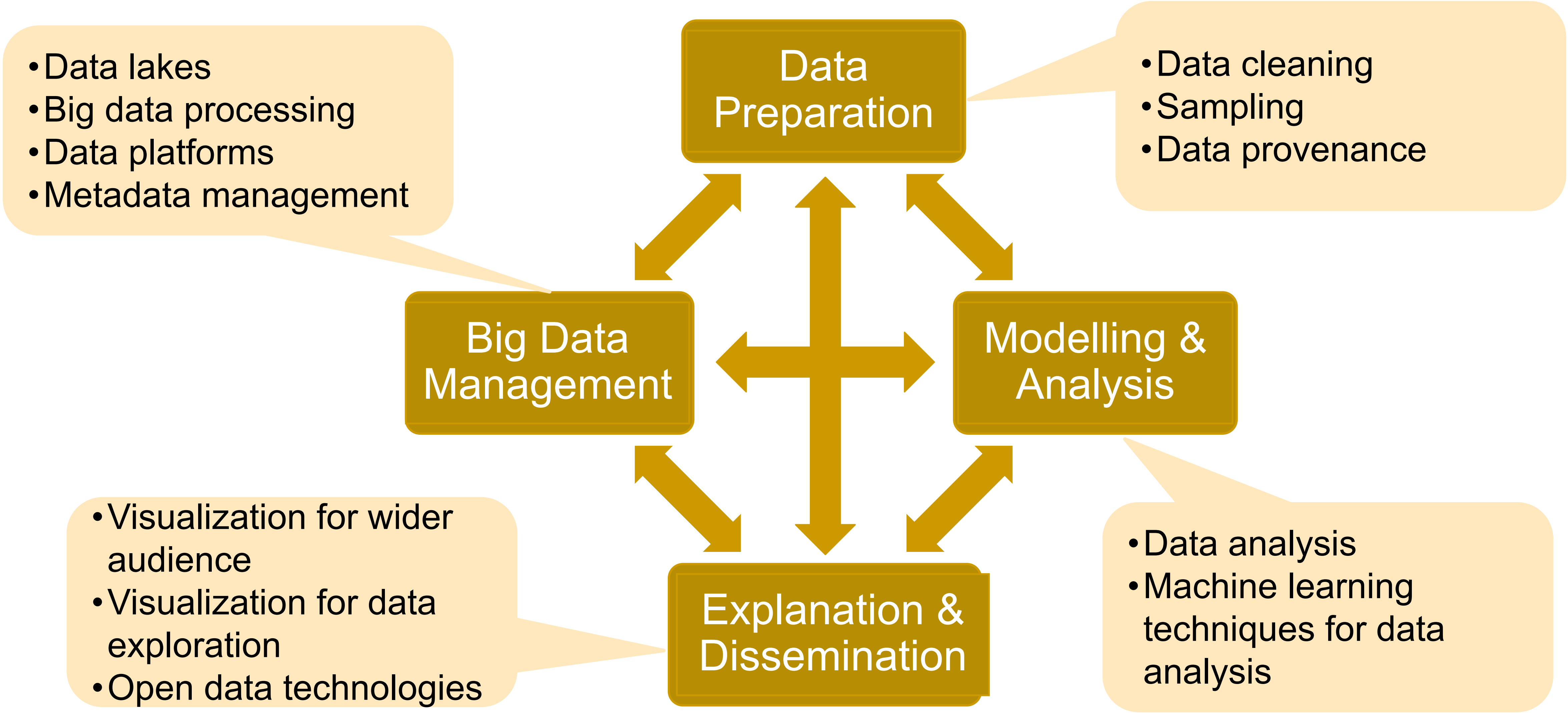
Data Science System Architecture

Who Owns Data Science

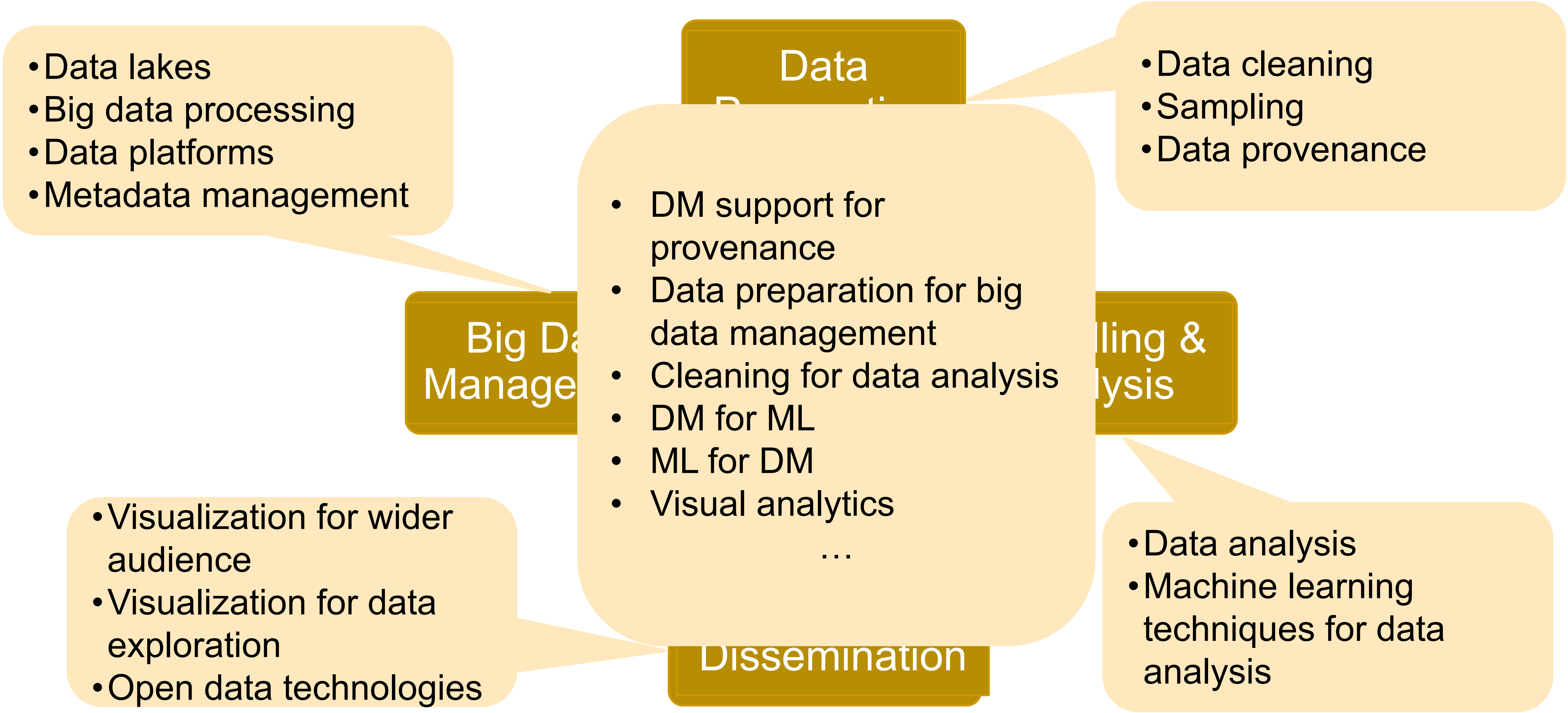
Data Science Lifecycle



Core Research Issues and Interactions



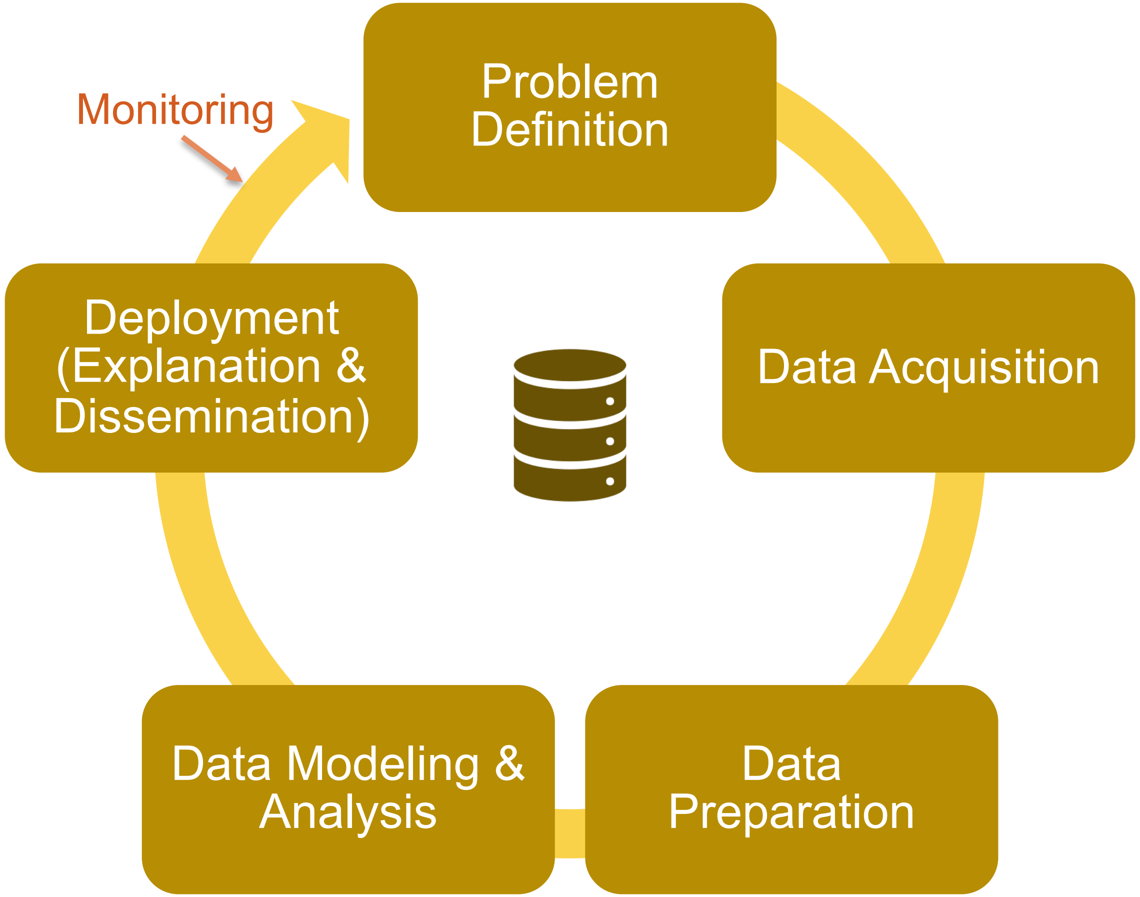
Core Research Issues and Interactions



Data Science Lifecycle – Alternative

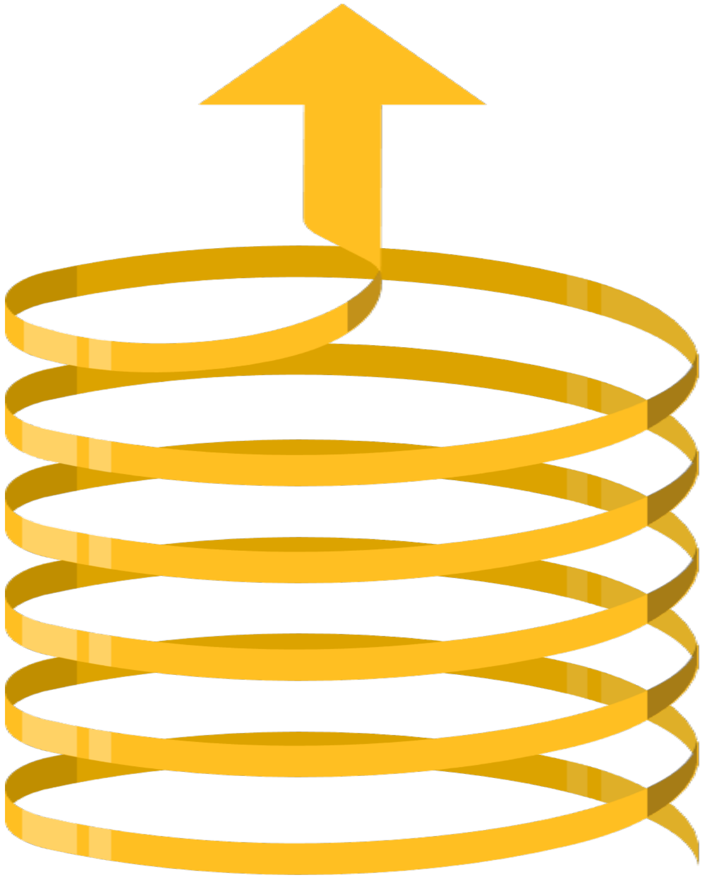
Data Science Lifecycle – Alternative

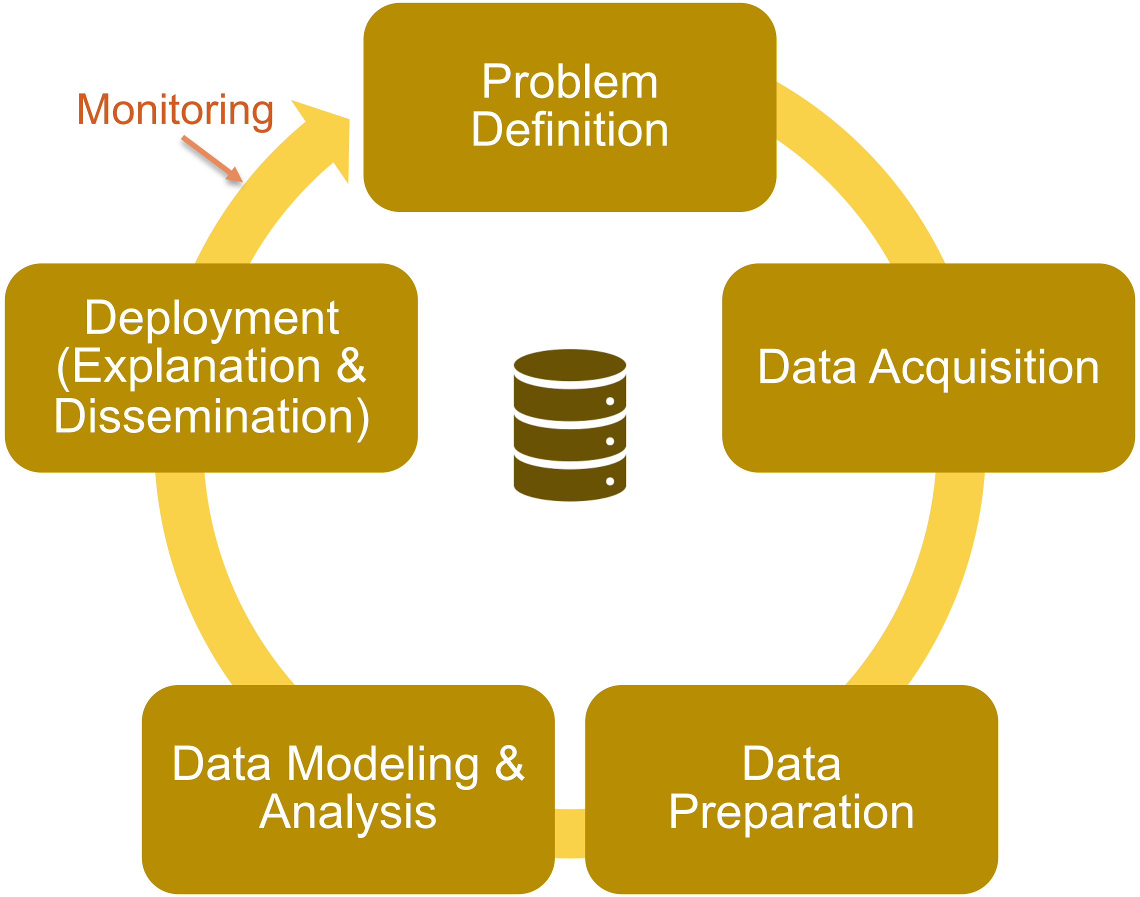
Data Science Lifecycle – Alternative



C. Shearer, The CRISP-DM Model, *J. Data Warehousing*, 2000

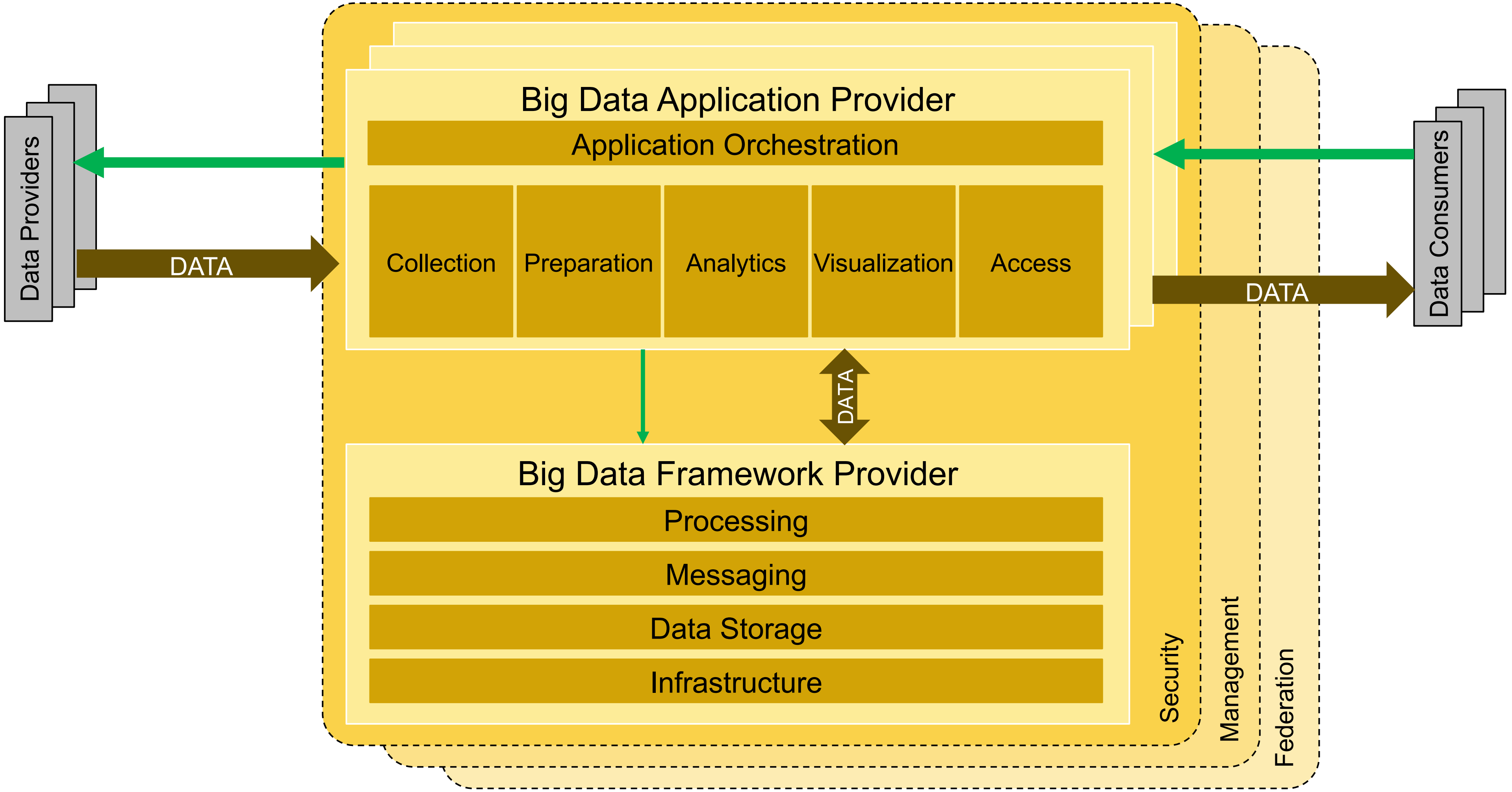
Data Science Lifecycle – Alternative





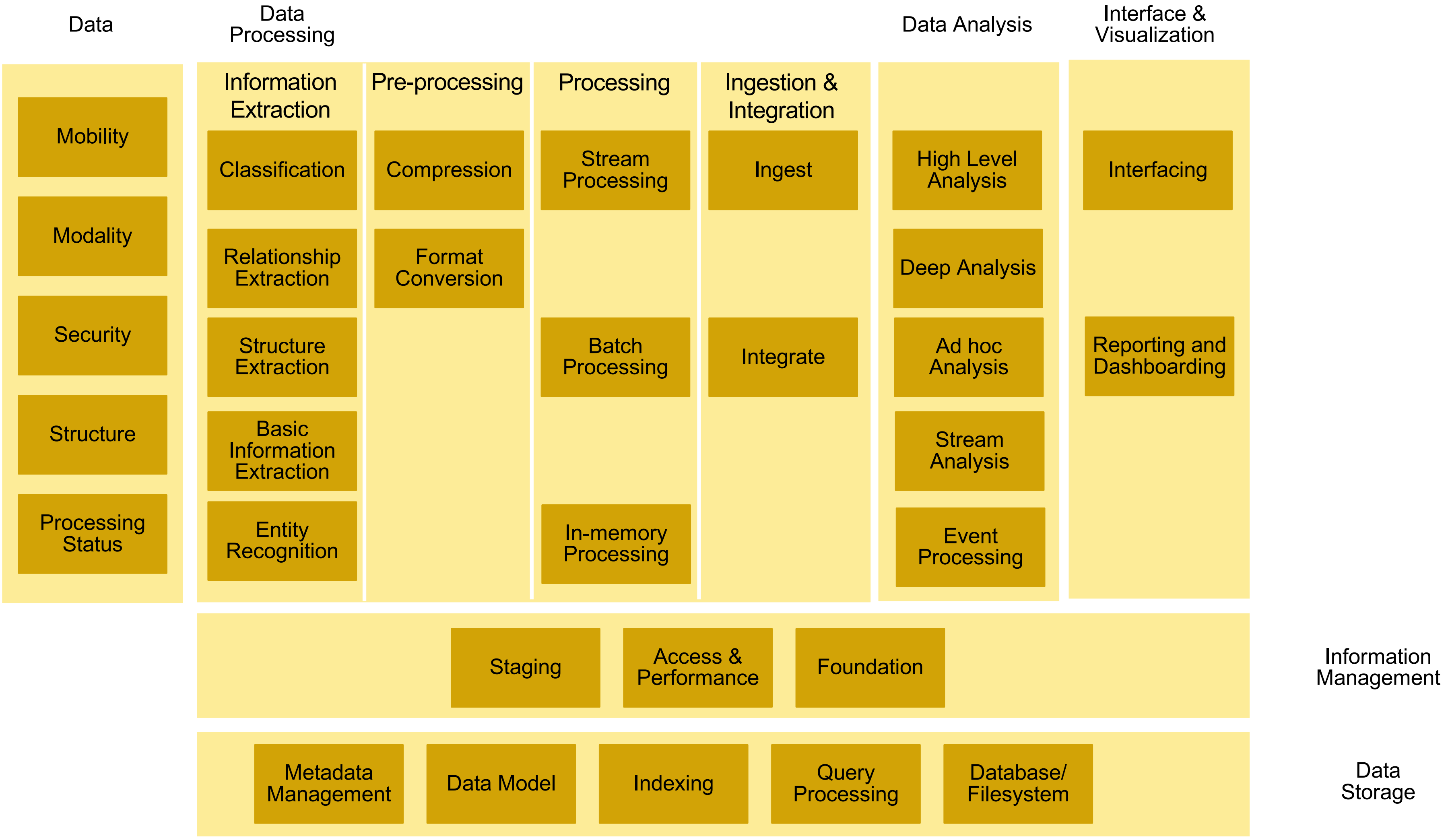
- What is Data Science
- Data Science Applications
- Data Science Ecosystem
- Data Science Lifecycle
- Data Science System Architecture
- Who Owns Data Science

Reference Architecture I



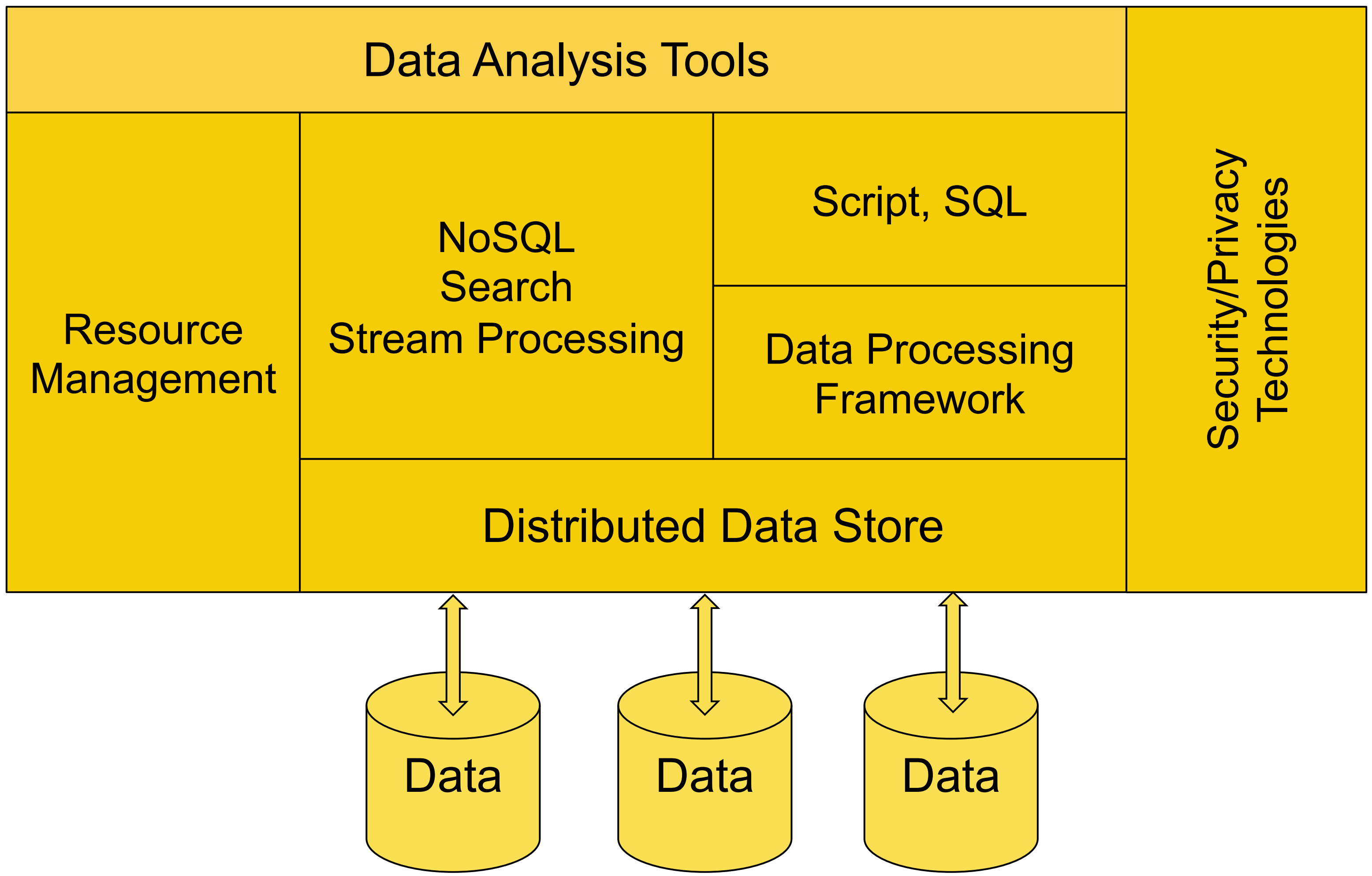
J. Klein et al., A Reference Architecture for Big Data Systems in the National Security Domain, *Proc. 2nd Int. Workshop on Big Data Soft. Eng.*, 2016

Reference Architecture II

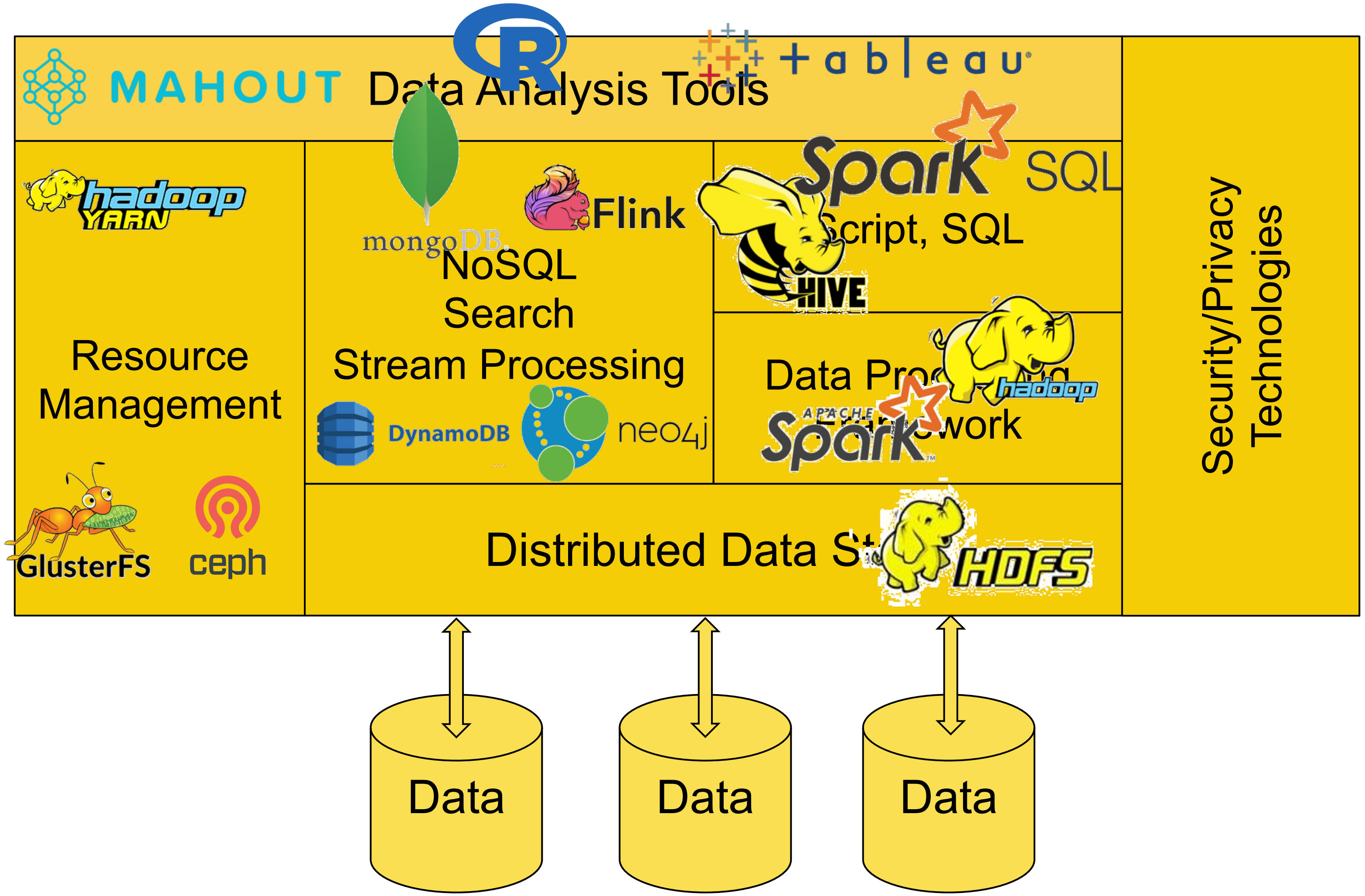


C. Avci Salma et al., Domain-Driven Design of Big Data Systems Based on a Reference Architecture, *Software Architecture for Big Data and the Cloud*, 2017

Concrete Architecture – Data Science Software Stack



Concrete Architecture – Data Science Software Stack

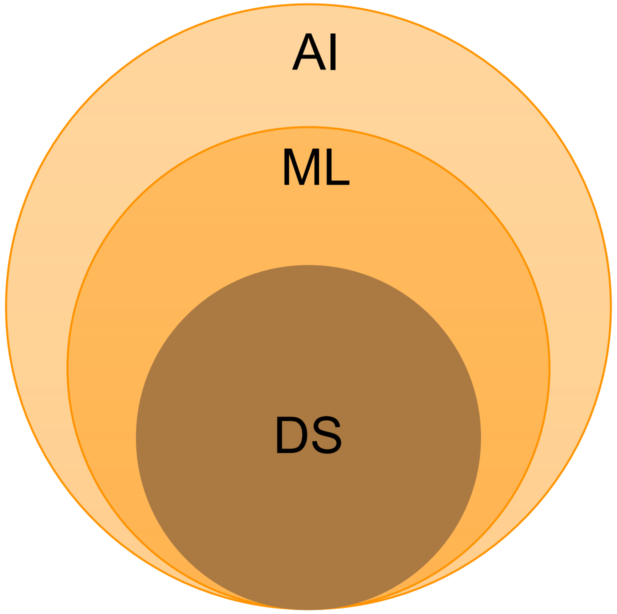


- What is Data Science
- Data Science Applications
- Data Science Ecosystem
- Data Science Lifecycle
- Data Science System Architecture
- Who Owns Data Science

Who Owns Data Science?

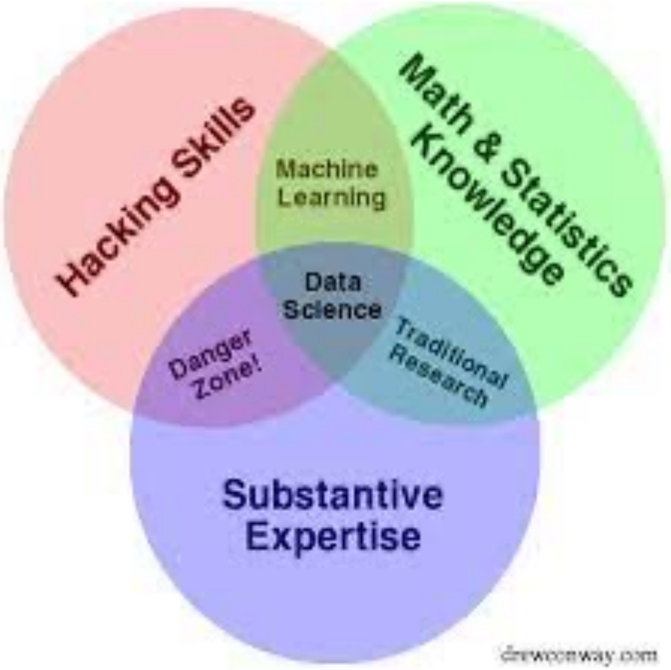
Computer Science

- It is all AI



Statistics – Conway Diagram

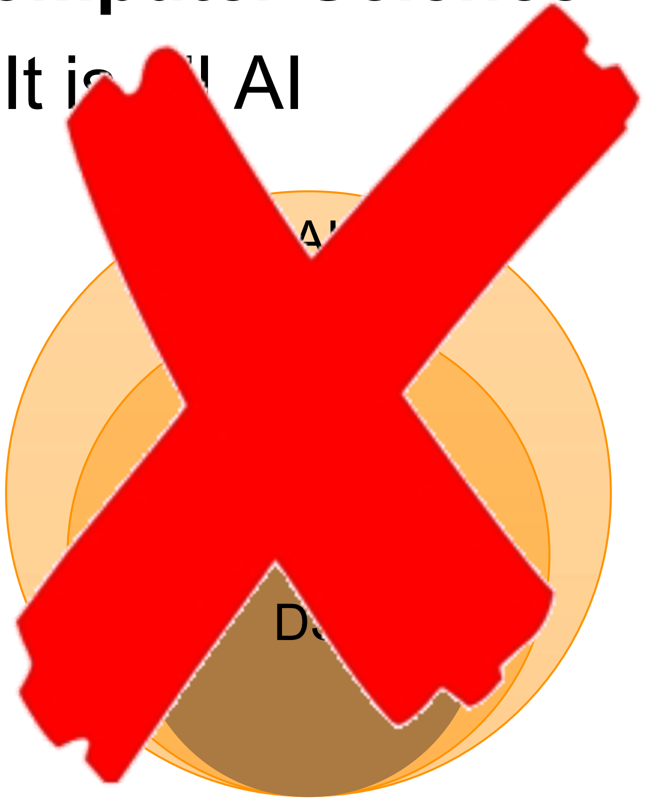
- CS part is just hacking



Who Owns Data Science?

Computer Science

- It i… ll AI

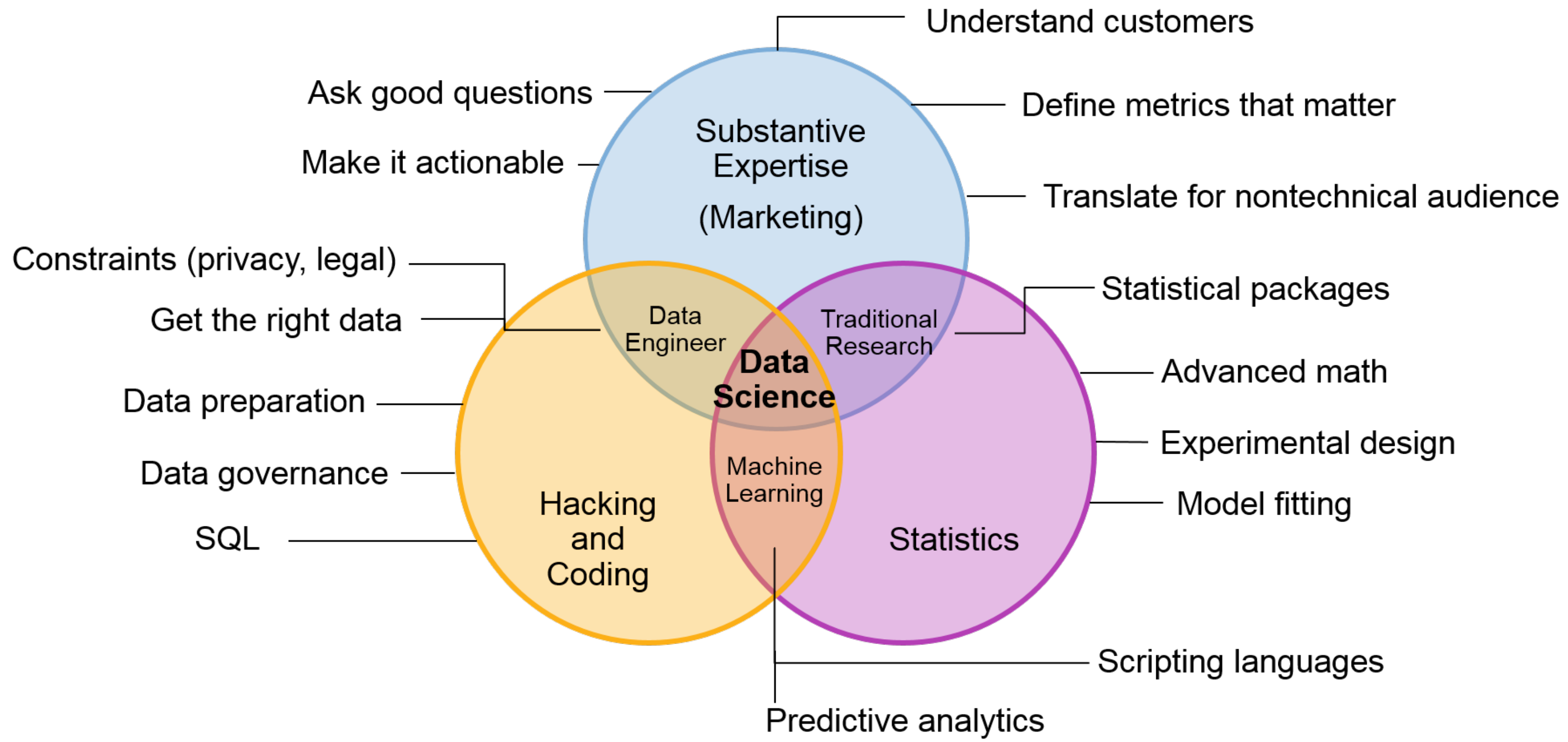


Statistics – Conway Diagram

- CS p… s just h… g

Who Owns Data Science

There seems to be great interest in this argument & in these diagrams



https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html/2

Who are the Constituents?

Who are the Constituents?

STEM – Core

People who are involved in developing the core technologies

Who are the Constituents?

STEM – Core

People who are involved in developing the core technologies

STEM – Application

People who are involved in data science applications in some domain

Who are the Constituents?

STEM – Core

People who are involved in developing the core technologies

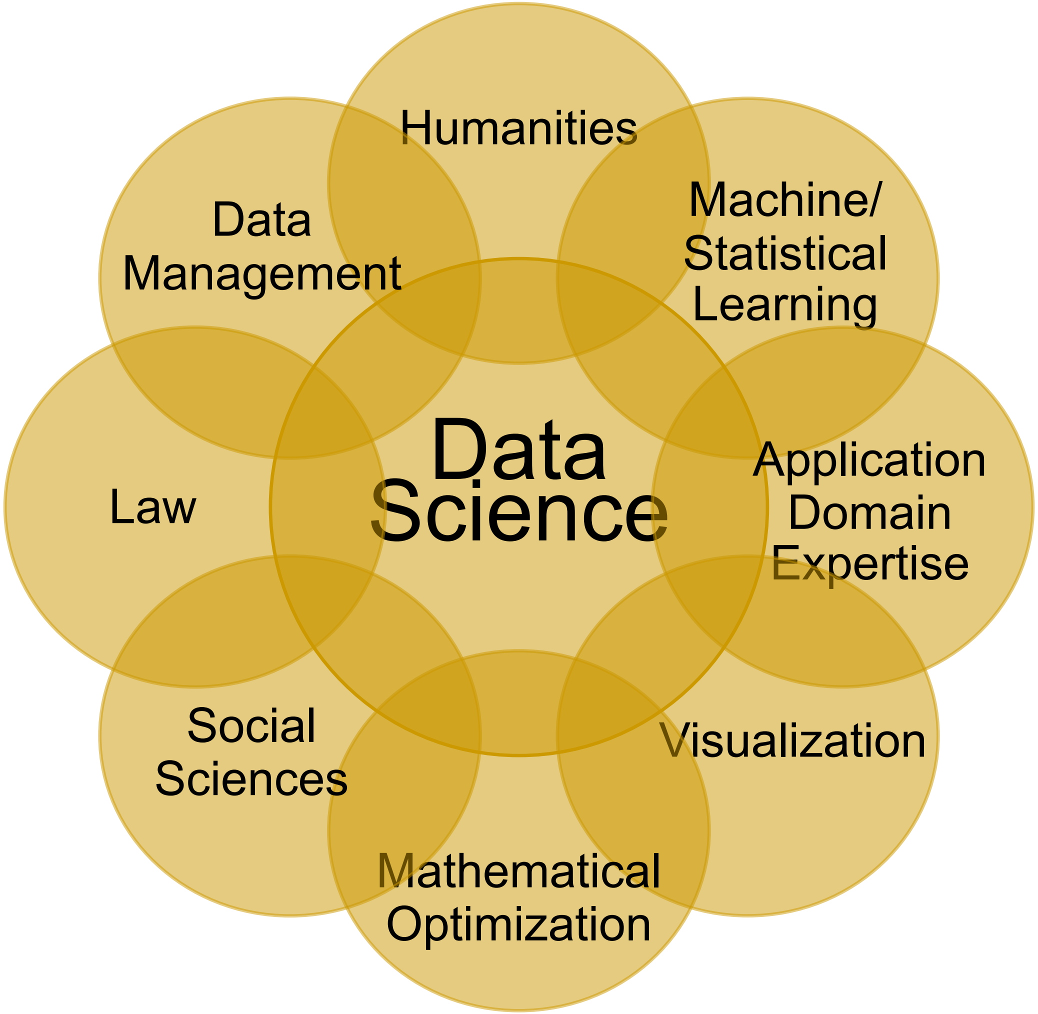
STEM – Application

People who are involved in data science applications in some domain

Non-STEM

People in social sciences and humanities who might be involved in applications or data ethics or social aspects or policy issues

Who are the Constituents?

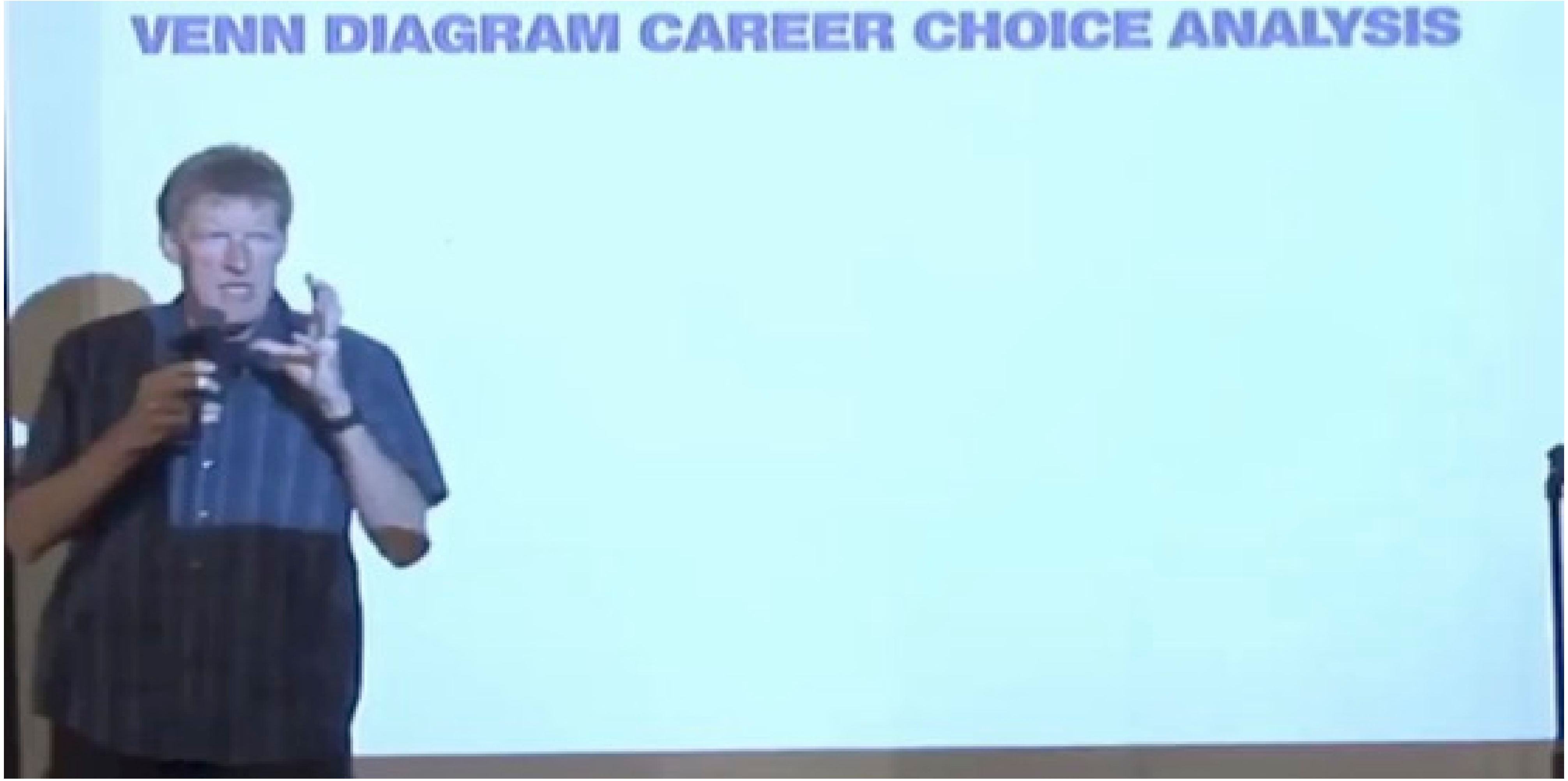


Who is a Data Scientist?

Core competencies

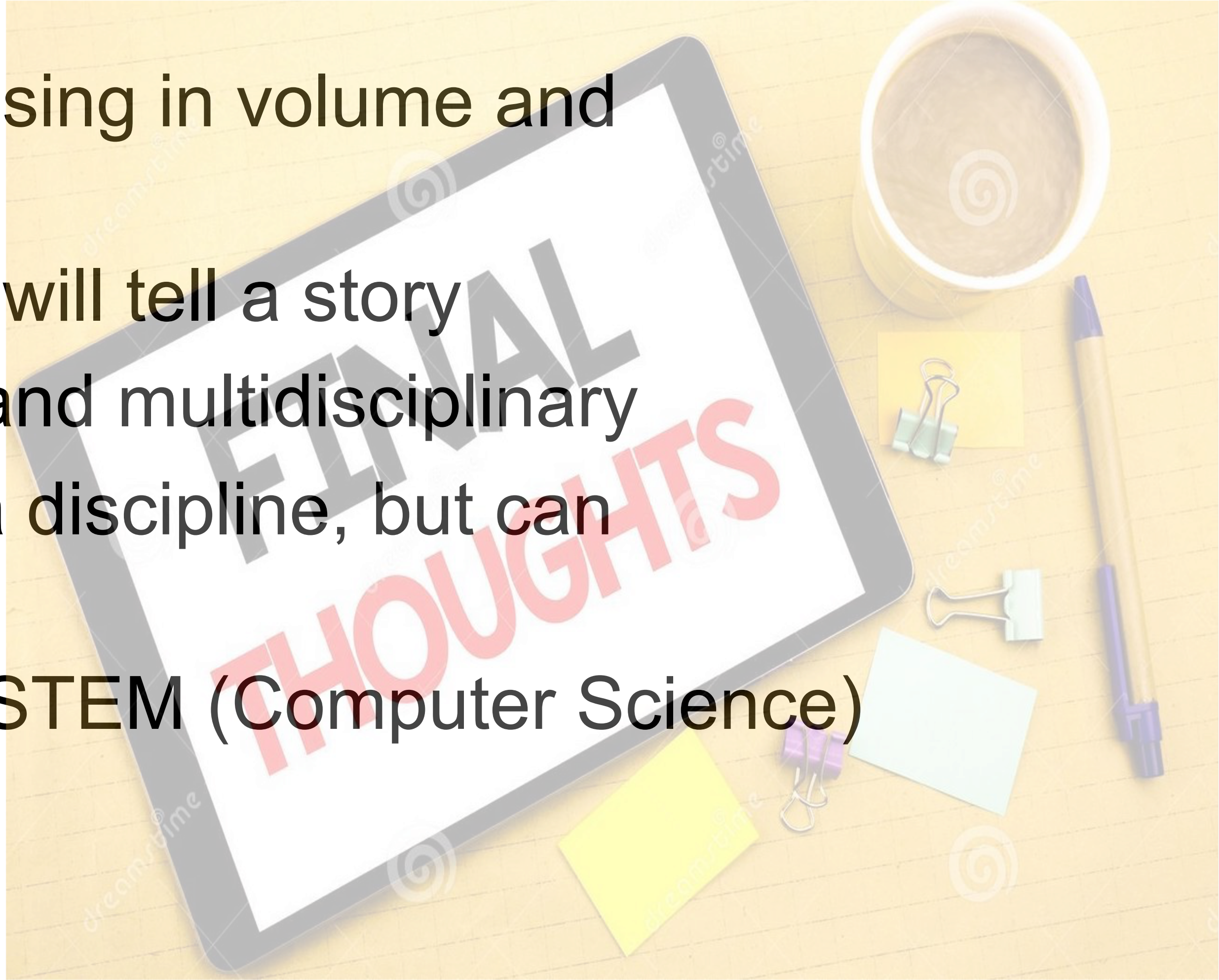
- In-depth knowledge of at least one of data engineering or data analytics pillars (expert level)
- Knowledge of the other two pillars of data security & privacy and data ethics (acquaintance)
- In-depth knowledge of at least one, preferably two, application areas (almost expert level)
- Ability to work in a team & communicate

Battle of the Venn Diagrams



Final Thoughts

- Data is central and it is increasing in volume and complexity
- Treat the data properly and it will tell a story
- Data science is multifaceted and multidisciplinary
- Data science may not yet bea discipline, but can become one
- The view I presented is from STEM (Computer Science) perspective
  - There is much more



Thank you