AIML427 Big Data

Week 1:What Can We Do with Big Data? (and what *should* we do)

Dr Bach Nguyen

School of Engineering and Computer Science

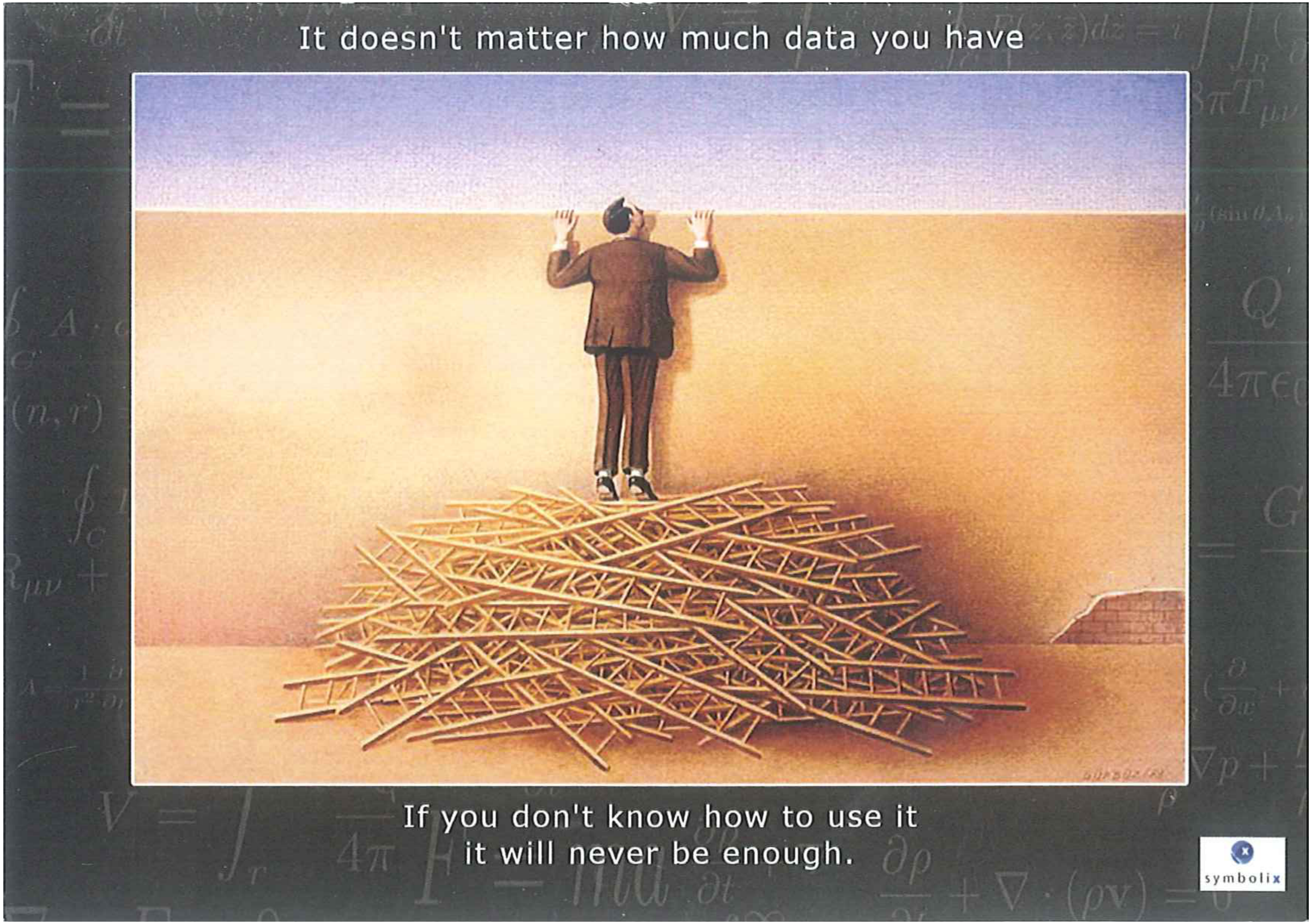
Victoria University of Wellington

Bach.Nguyen@vuw.ac.nz

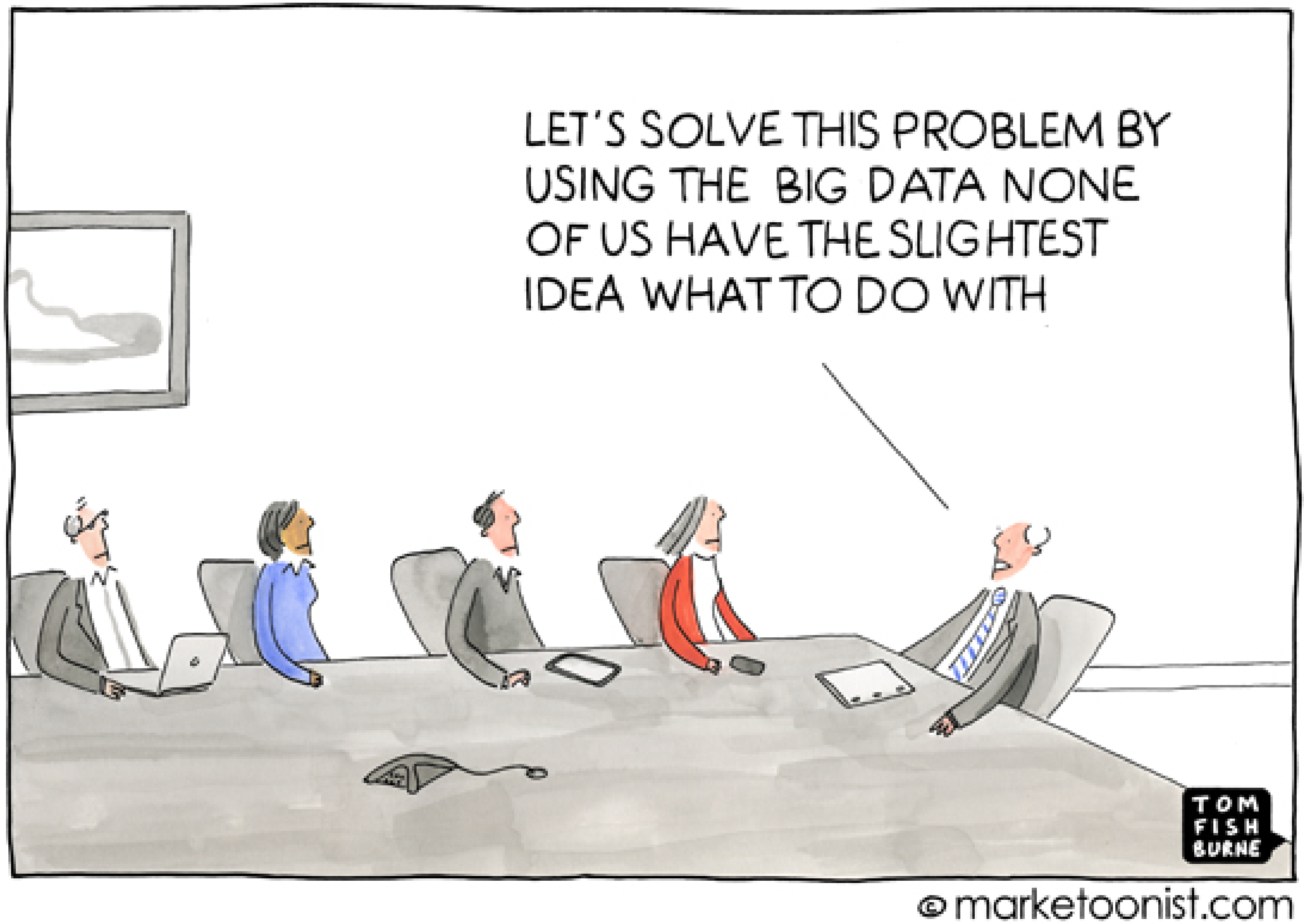
Outline

- Big Data, Data Mining, Machine Learning, and Statistical Learning
- 1936 U.S. Presidential Elections
- Google Flu Trends
- Tools for Assignment 1

What can we do with Big Data?


It doesn't matter how much data you have
If you don't know how to use it
it will never be enough.
symbolix

What should we do with Big Data?



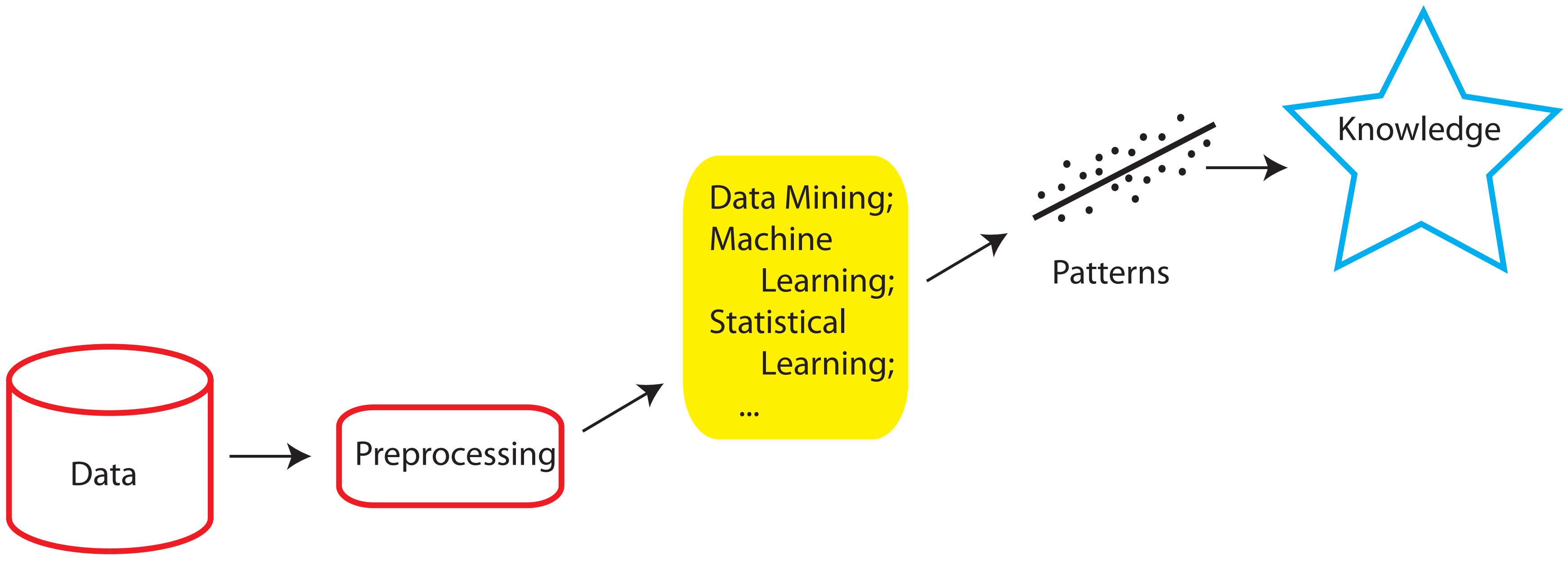
LET'S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH
TOM FISH BURNE

© marketoonist.com

What can we do with Big Data?

- Data Mining, Machine Learning, Statistical Learning, ... techniques can be used to discover interesting patterns, associations and knowledge from big data.

- Learning from big data is an **interdisciplinary** task:
  - Statistics; Computer Science; and Mathematics;



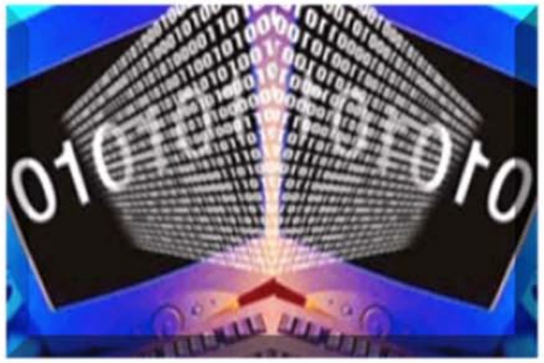
What can we do with Big Data?

Smarter Healthcare

Multi-channel

Finance

Log Analysis



Homeland Security

Traffic Control

Telecom

Search Quality

Manufacturing

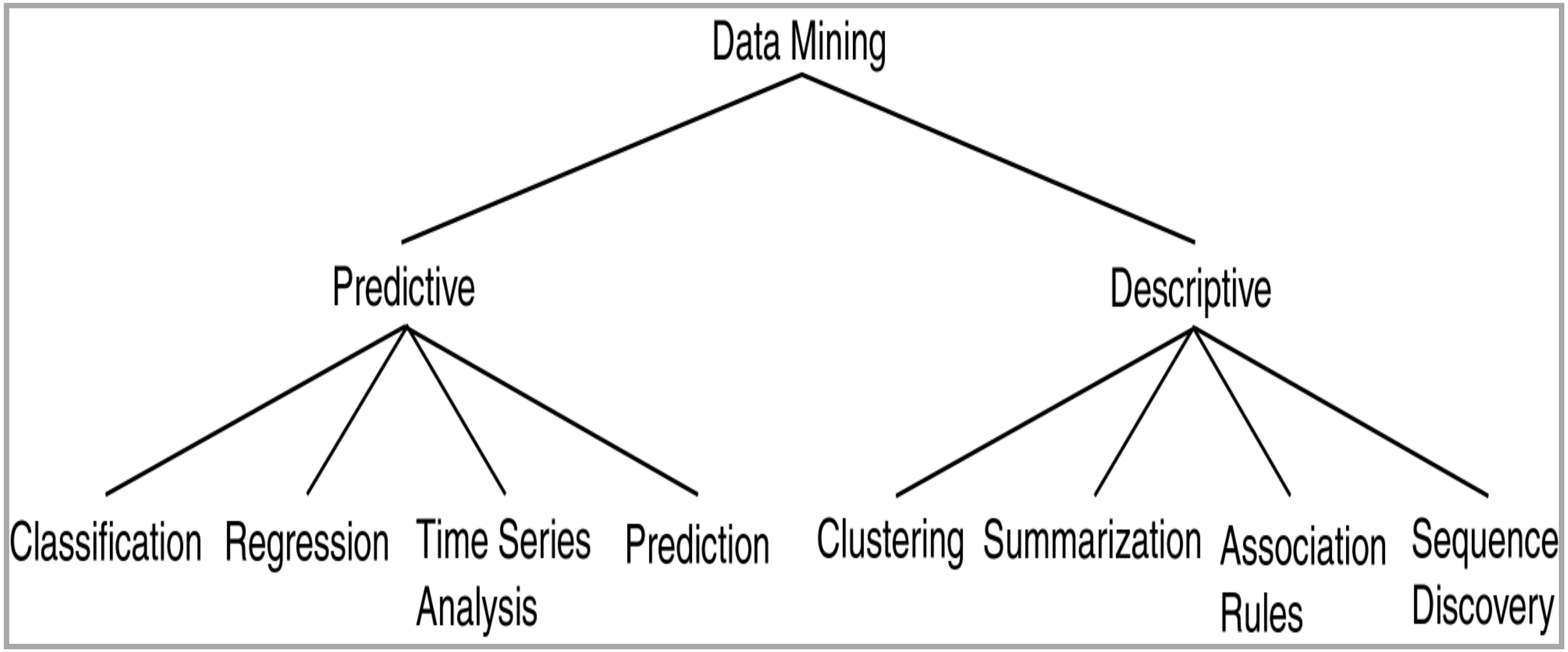
Trading Analytics

Fraud and Risk

Retail: Churn, NBO

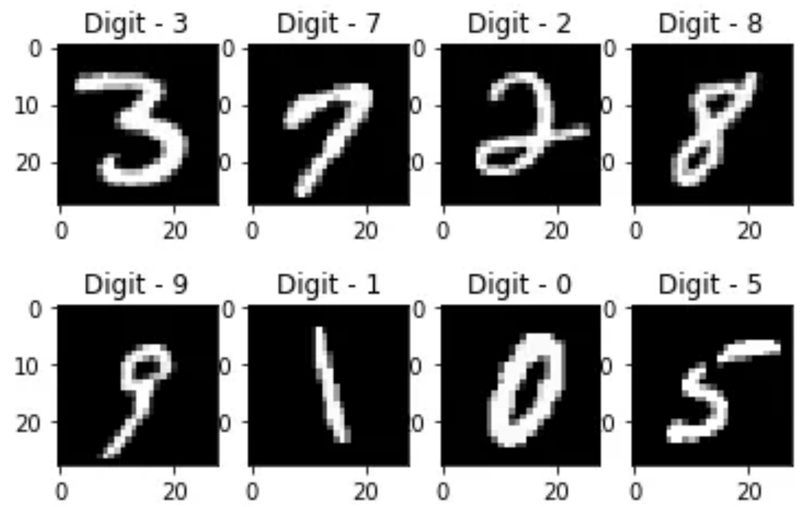
Data Mining

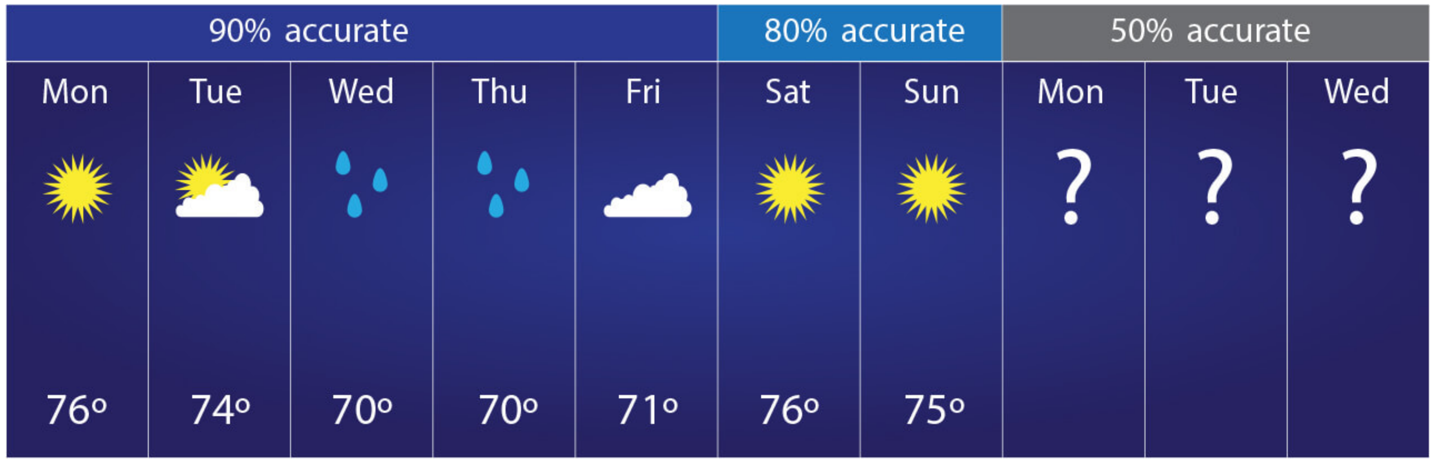
- Witten, Eibe and Hall (2011) say "Data mining is defined as the process of discovering patterns in data”
- Data mining is the process of extracting implicit, previously unknown, and potentially useful information from data.
- **How to form a learning problem is important!**



Classification

- *Examples*:
  - Handwritten digit classification
  - Classifying credit card transactions as legitimate or fraudulent
  - Does a given image contain a dog or not?
  - Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil
  - Categorizing news stories as finance, weather, entertainment, sports, etc.

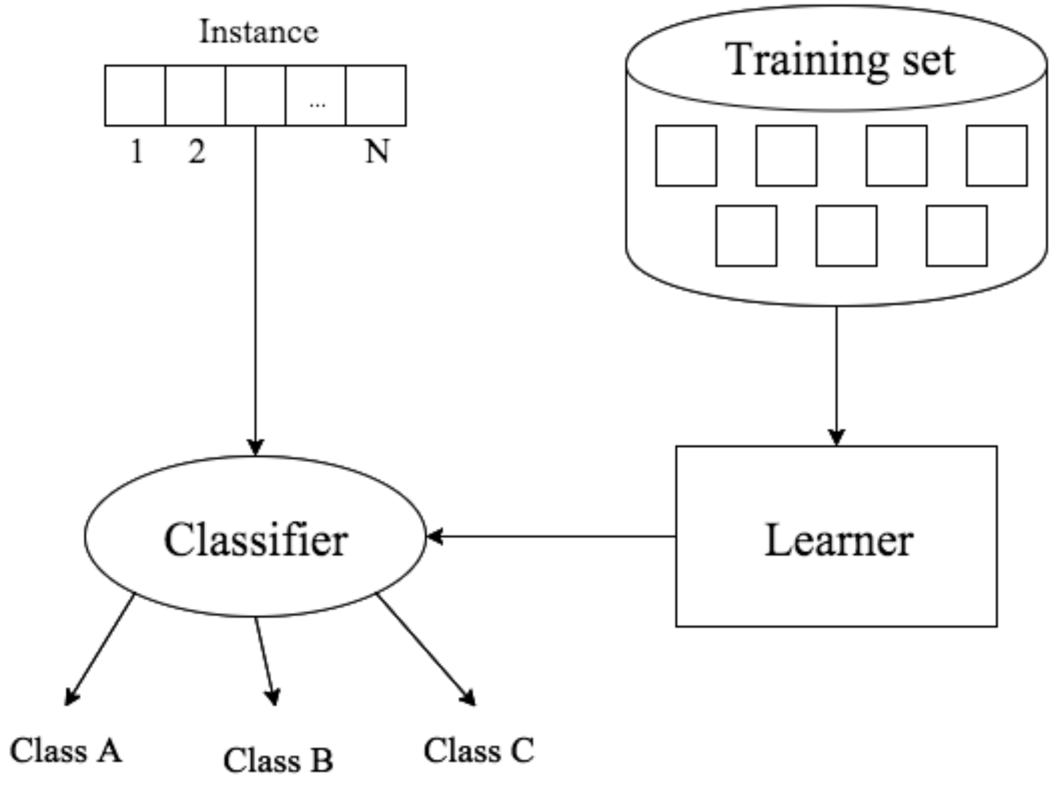




90% accurate					80% accurate		50% accurate		
Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed
☀	⛅	💧	💧	☁	☀	☀	?	?	?
76º	74º	70º	70º	71º	76º	75º			

Classification

- Y=f(x) is a predictive model learnt from the data set
  - D = {(X1,y1),…(Xn,y2}
- An important task in data mining:
  - Given a training set with a number of instances
    - Each instance is represented as a *feature vector* and a desired/target *class label*
  - Learn/train a classifier from the training set
  - The learned classifier determines the class label of a new, unseen (test) instance:
    - based on the feature values of that instance.
- Qualitative/ Discrete valued Y
- Measures: accuracy or error rate



Classification

- Classification algorithms:
  - K-nearest Neighbour method (KNN)
  - Decision tree (DT)
  - Support vector machines (SVMs)
  - Artificial Neural networks (ANNs)
  - Bayesian Classifier: Naive Bayes (NB)

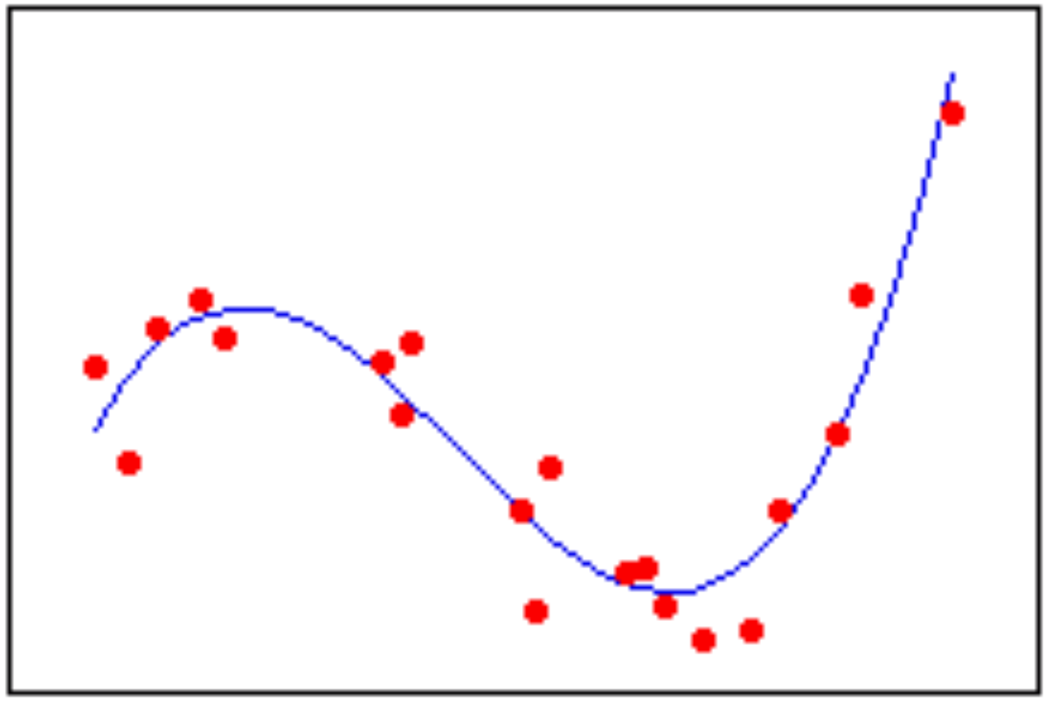
- COMP307 attachment!

Regression

- *Examples*:
  - How many students enroll AIML427 in 2025?
  - What is the average price of properties in Wellington?
  - How many of our customers will leave for a competitor this year?
  - How tall will a child be as an adult?

Regression

- Regression analysis is a form of predictive modelling technique, which investigates the relationship between a dependent (target) , **Y**, and independent variable (predictor), **X**.
  - Qualitative output (Y): Continues Values Y
  - Fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.

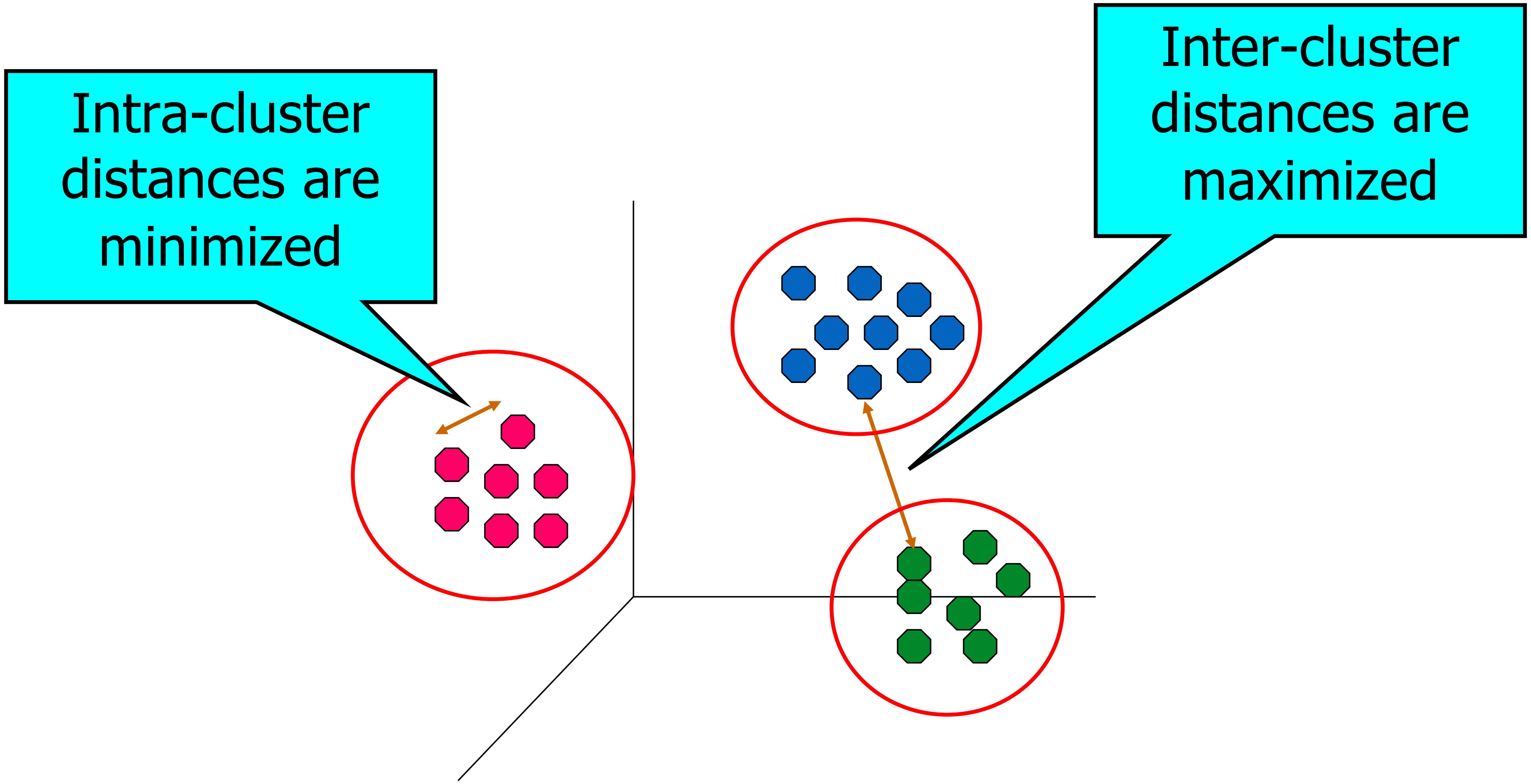


Regression

- *Methods:*
  - Linear regression
  - Support vector regression
  - K-nearest neighbor regression
  - Regression tree, CART

Clustering

- Clustering: Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Intra-cluster distances are minimized

Inter-cluster distances are maximized

Clustering

- *Examples*:
  - Marketing: Help marketers discover distinct groups in their customer bases
  - Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
  - City-planning: Identifying groups of houses according to their house type, value, and geographical location
  - Earth-quake studies: Observed earthquake epicentres should be clustered along continent faults
  - Land use: Identification of areas of similar land use in an earth observation database
- *Methods*:
  - K-means clustering
  - Hierarchical clustering
  - Fuzzy clustering
  - Density-based clustering

Machine Learning

- "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" (by Tom M. Mitchell)

- Machine learning: Definition Machine learning is a scientific discipline that is concerned with the design and development of algorithms that allow computers to learn based on data/ experience, such as from sensor data or databases.

Supervised, Unsupervised and Reinforcement Learning

Supervised learning

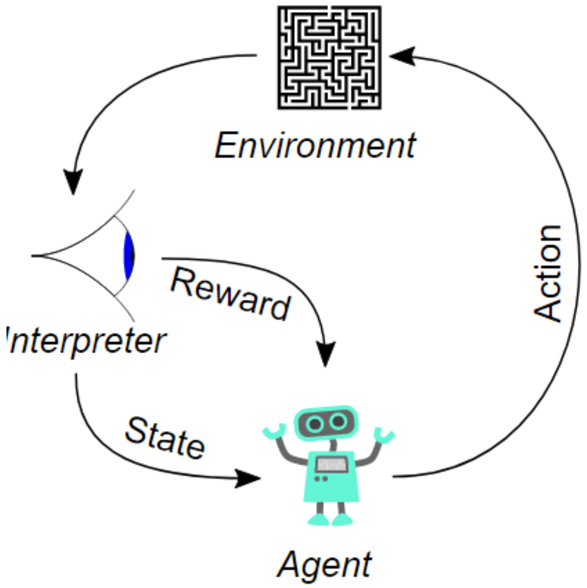
- Learn from labelled data: the data/examples, for a ML algorithm to learn from, are in correct input-output pairs,
  - Regression or classification
  - handwriting recognition

Unsupervised learning

- Learn from unlabeled data, working blind
  - Clustering
  - identify segments of customers with buying habits.

Reinforcement Learning

- No labelled data, but have a way to *measure* quantify its performance as a *reward signal*.
  - Games, AlphaGo



Classification Dataset — Example

- 14 days (examples/instances/observations/objects)
- 2 classes: Yes, No
- 4 features/variables/attributes
  (or 5 attributes including class label)

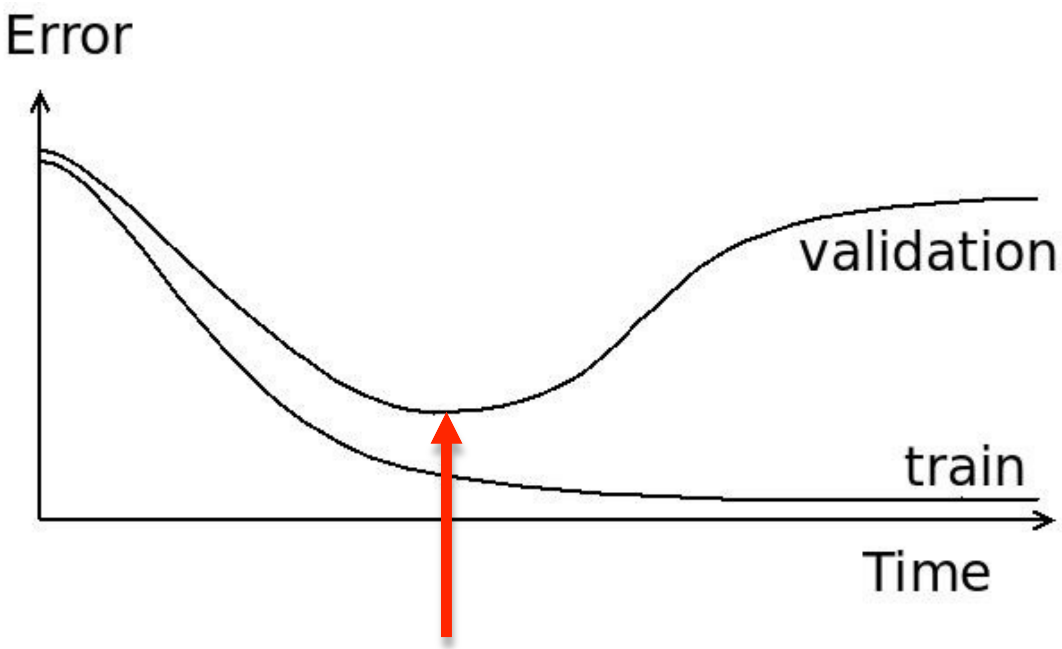
	Outlook	Humidity	Wind	Temperature	Play Tennis ?
Day 1	Sunny	85%	LIGHT	85	**No**
Day 2	Sunny	90%	STRONG	80	**No**
Day 3	Overcast	86%	LIGHT	83	**Yes**
Day 4	Rain	96%	LIGHT	70	**Yes**
Day 5	Rain	56%	LIGHT	64	**Yes**
Day 6	Rain	45%	STRONG	65	**No**
Day 7	Overcast	50%	STRONG	68	**Yes**
Day 8	Sunny	89%	LIGHT	71	**No**
Day 9	Sunny	50%	LIGHT	69	**Yes**
Day 10	Rain	52%	LIGHT	72	**Yes**
Day 11	Sunny	39%	STRONG	75	**Yes**
Day 12	Overcast	89%	STRONG	75	**Yes**
Day 13	Overcast	42%	LIGHT	81	**Yes**
Day 14	Rain	55%	STRONG	72	**No**

Training and Test Sets (Experiments)

- **Training** set: to learn/train a model/classifier
- **Test** set: to measure the performance of the classifier

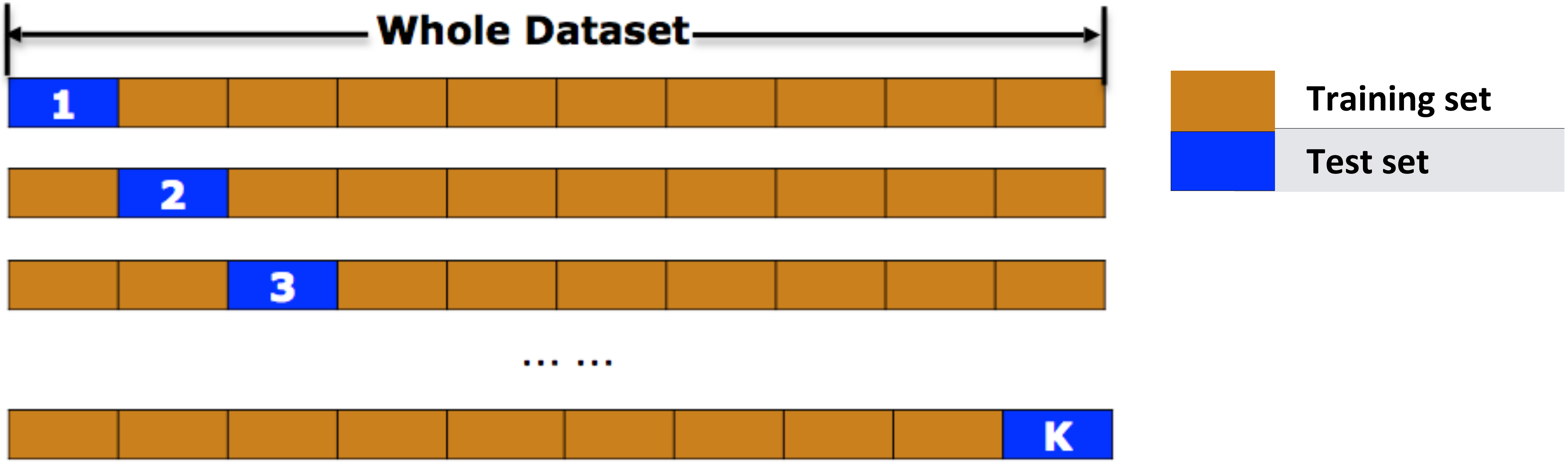
- Training—Test: 50%—50%; 2/3 — 1/3; 70%—30%
- Represent the original data
- Generalisation VS **overfitting**
  - **Validation** set: monitor the training process
  - Validation set VS Test set
  - Training—Test—Validation:

    1/3 — 1/3 — 1/3



K-fold Cross Validation (K-CV)

- Used when only a small number of instances are available
- Ideas:
  - Split the whole dataset to K folds with equal size
  - Use 1 fold as test set, and the other (K-1) folds as training set
  - Repeat K times to make sure each fold has a chance to be the test set
  - Average the K test performances (e.g. error rates)
- Leave-one-out cross validation:



Performance Evaluation

- Error:

  Mean Squared Error (regression)

  $$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

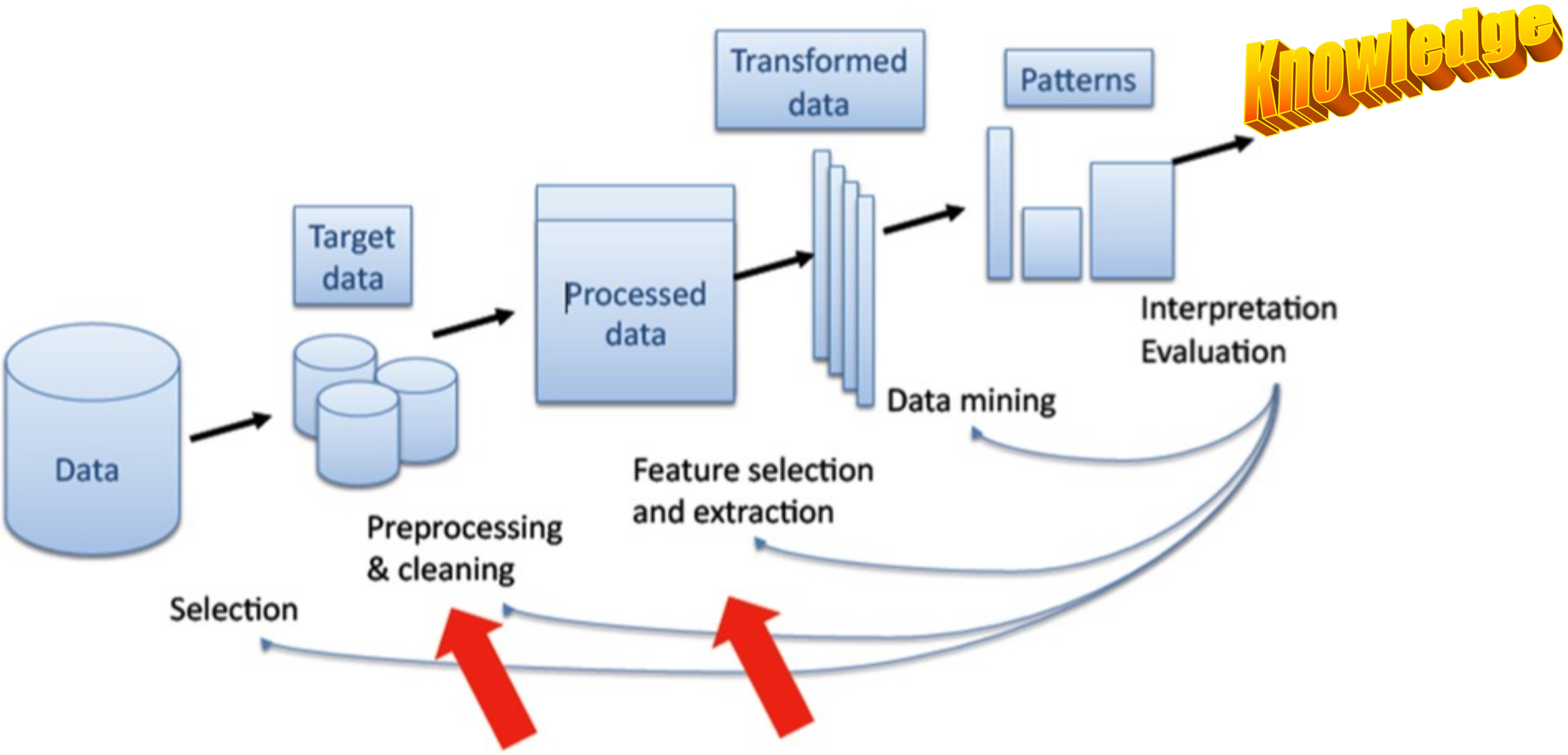
  Error Rate (classification)

  $$\mathrm{ER} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(y_i \neq \hat{y}_i).$$

- Training Error: the MSE (ER) computed from the data that was used to learn the model.
  - We generally don't care too much about training error (it's easy to construct a model with zero training error!).
- Testing Error: the MSE (ER) computed from test data that was not used to learn the model.

Big Data/DM in KDD

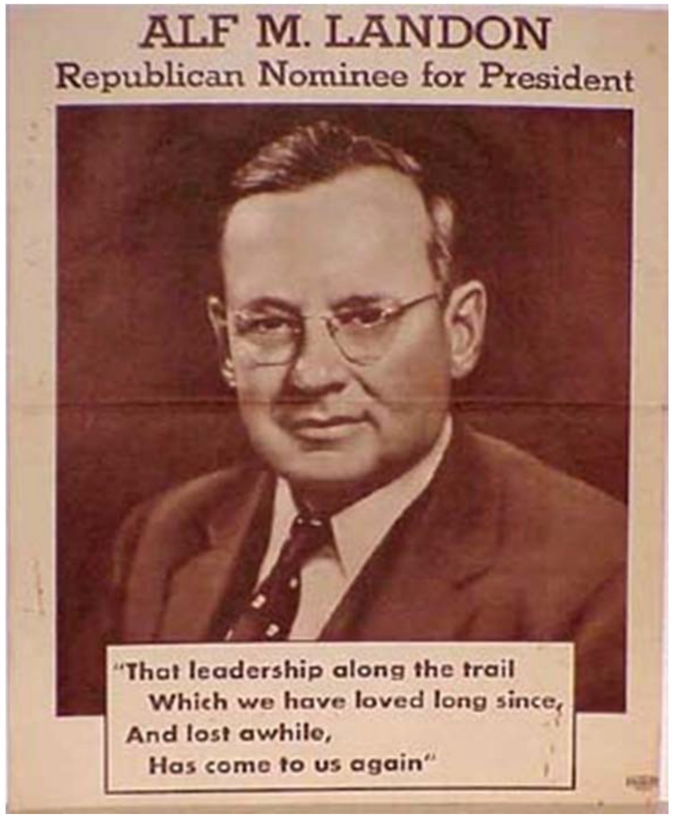
- Data mining plays an essential role in the knowledge discovery process



1936 U.S. Presidential Elections

The Literary Digest survey

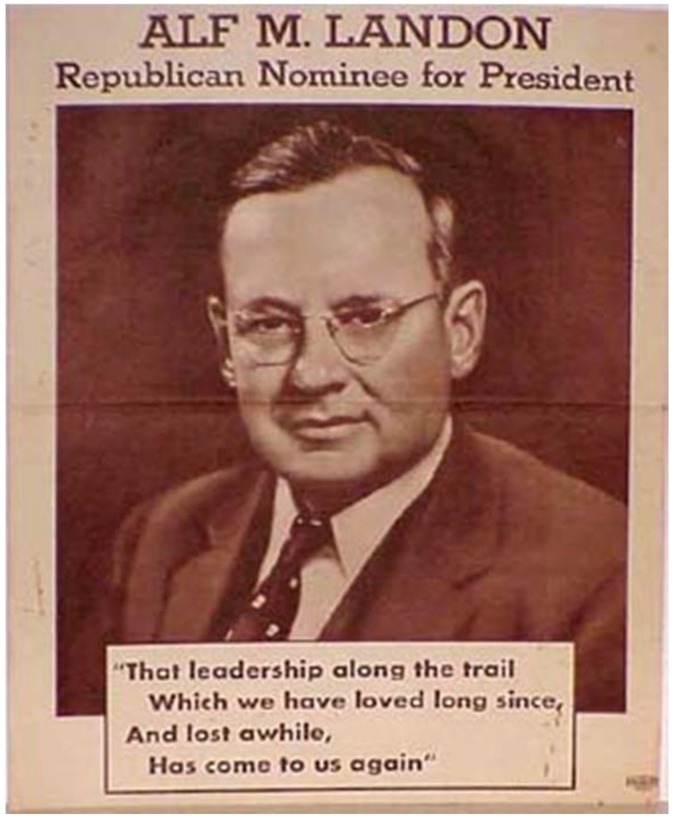
- Magazine had predicted **every** election (successfully) since 1916
- Sent out 10 million surveys
  - 2.4 million responded
- Prediction: Roosevelt 43%
- Actual: Roosevelt: 62%
- (Literary Digest went bankrupt soon after)


ALF M. LANDON
Republican Nominee for President
"That leadership along the trail
Which we have loved long since,
And lost awhile,
Has come to us again"

1936 U.S. presidential elections

George Gallup did it right

- Actual election result: 62%
- Literary Digest prediction: 43%
- Gallup's prediction of the Digest prediction (based on sample of 3,000): 44%
- Gallup's prediction of the election result (based on sample of 50,000): 56%


ALF M. LANDON
Republican Nominee for President
"That leadership along the trail
Which we have loved long since,
And lost awhile,
Has come to us again"

What Went Wrong?

- Context: Great Depression
  - 9 million unemployed; real income down 33%
  - Landon: “Cut spending” versus Roosevelt: “Balance peoples’ budgets before government’s budget”

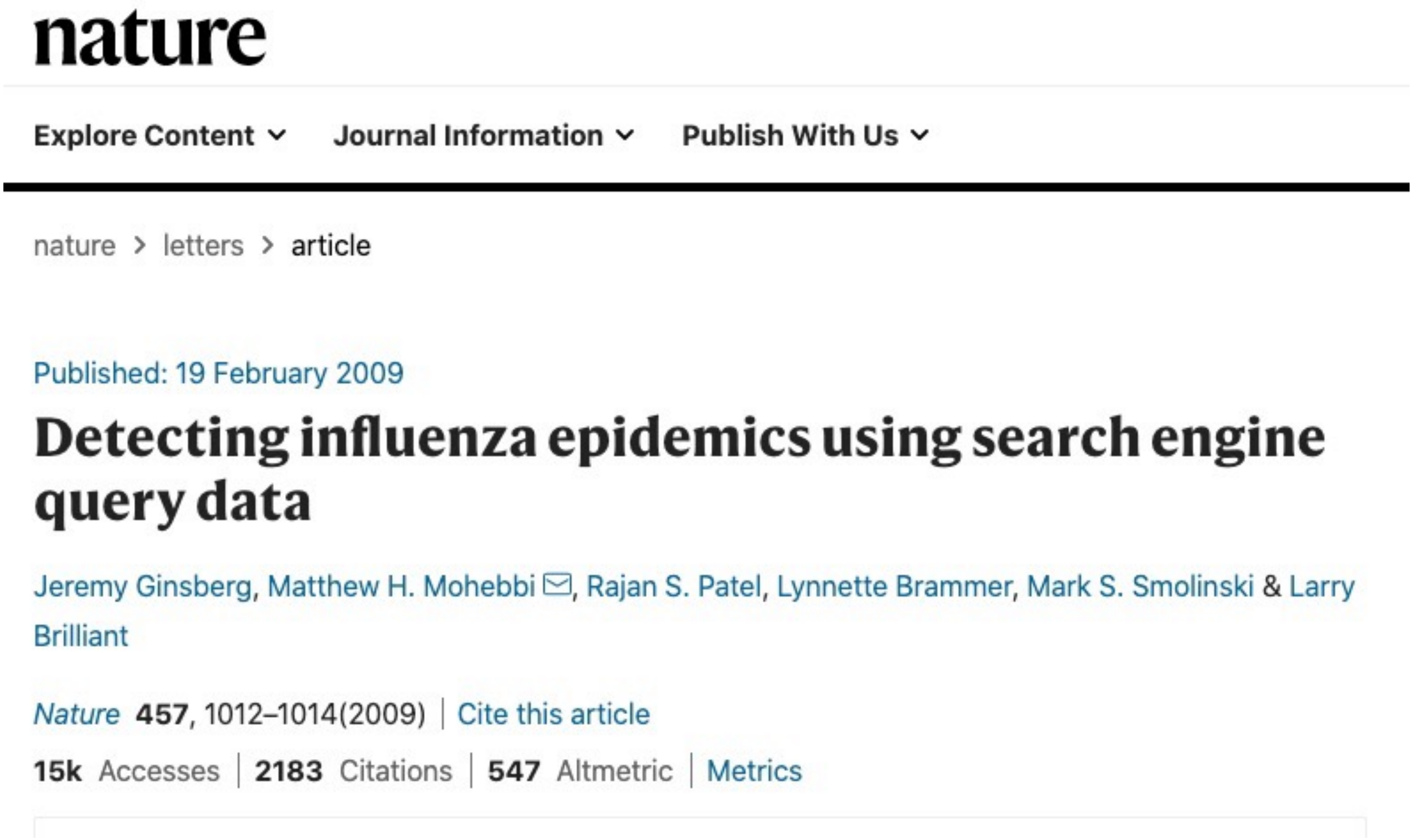
- Polling
  - Survey sent out to 10 million people from subscription list, telephone directories, phone books and club memberships lists (only 1 in 4 households had a phone!).
  - 2.4 million responded

What Went Wrong?

- Sampling Bias, Sampling not representative
    - *The poor tended to vote Roosevelt.*
    - The sample of 2.4 million voters was biased!
- Non-response bias:
    - Only 1 in 4 responded.
    - The anti-Roosevelt forces were angry --- and had a higher response rate!
- Gallup used Random sampling
    - Every combination of people has equal chance to be selected

When a procedure is biased, taking a larger sample does not help. It just repeats the basic mistake on a larger scale!

Google Flu Trends



nature

Explore Content ⌄ Journal Information ⌄ Publish With Us ⌄

nature > letters > article

Published: 19 February 2009

Detecting influenza epidemics using search engine query data

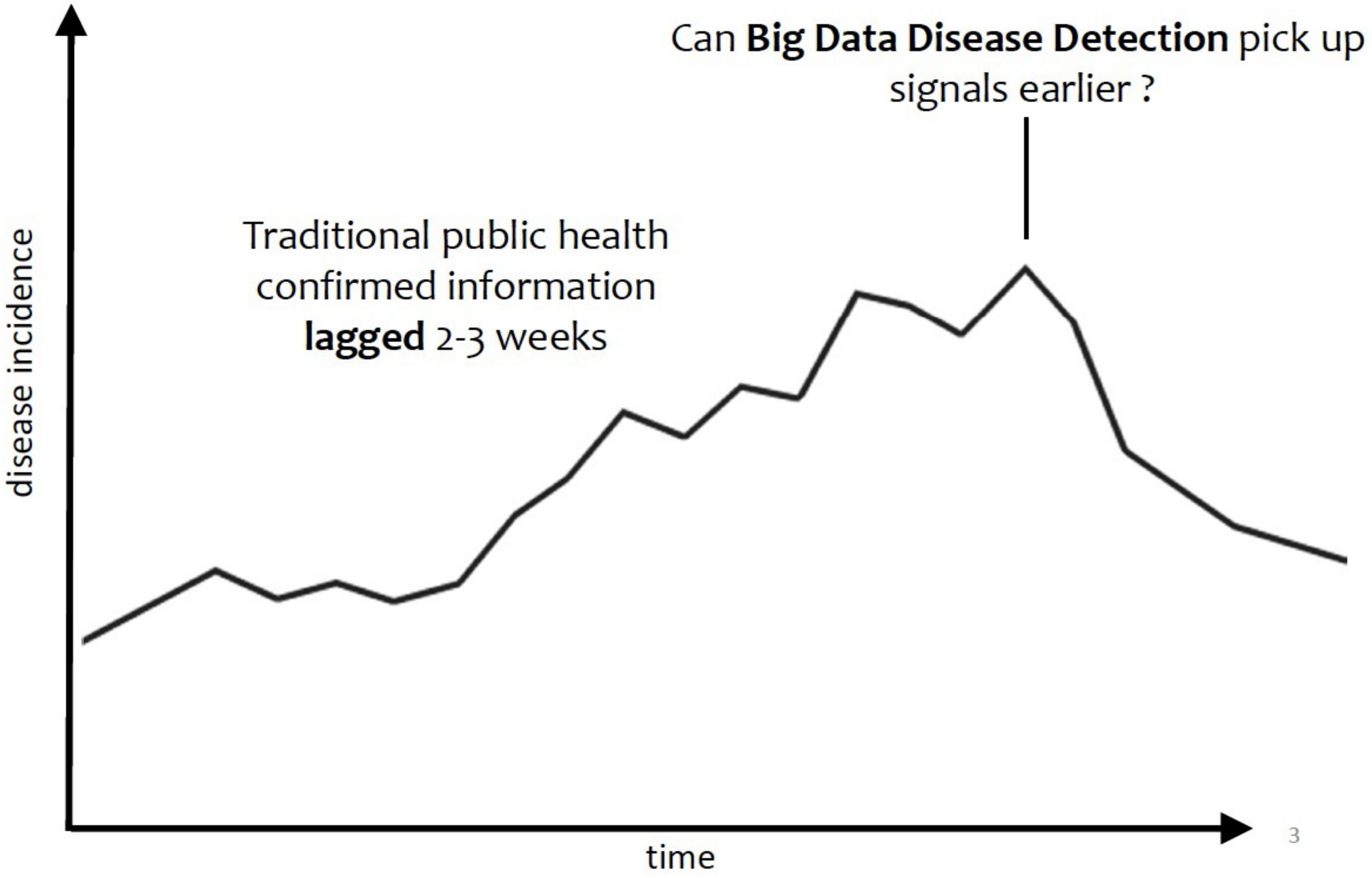
Jeremy Ginsberg, Matthew H. Mohebbi ✉, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski & Larry Brilliant

Nature **457**, 1012–1014(2009) | Cite this article

15k Accesses | **2183** Citations | **547** Altmetric | Metrics

Using Big Data To Predict Flu Trends

- Importance: tens of millions of respiratory illnesses, 250,000 - 500,000 deaths from respiratory illnesses worldwide each year[1]
- During flu season, more people enter search queries concerning flu
- Each year 90 million American adults search web for info about specific illnesses = **LOTS OF DATA**



[1] World Health Organization. Influenza fact sheet. ⟨http://www.who.int/mediacentre/factsheets/2003/fs211/en/⟩ (2003)

Google Flu Trends

- “We want to estimate flu activity based on more than just a few queries”
- “We‘ve found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate current flu activity around the world in near real-time.”
  http://www.google.org/flutrends/ (no longer available)

- Predictions by Google Flu are 1-2 weeks ahead of CDC’s ILI (Influenza-like illness) surveillance reports
  - The Centers for Disease Control and Prevention (CDC) took one or two weeks to determine flu trends based on information from doctors’ clinics. GFT took one to two days!

[Ginsberg Nature ‘09]

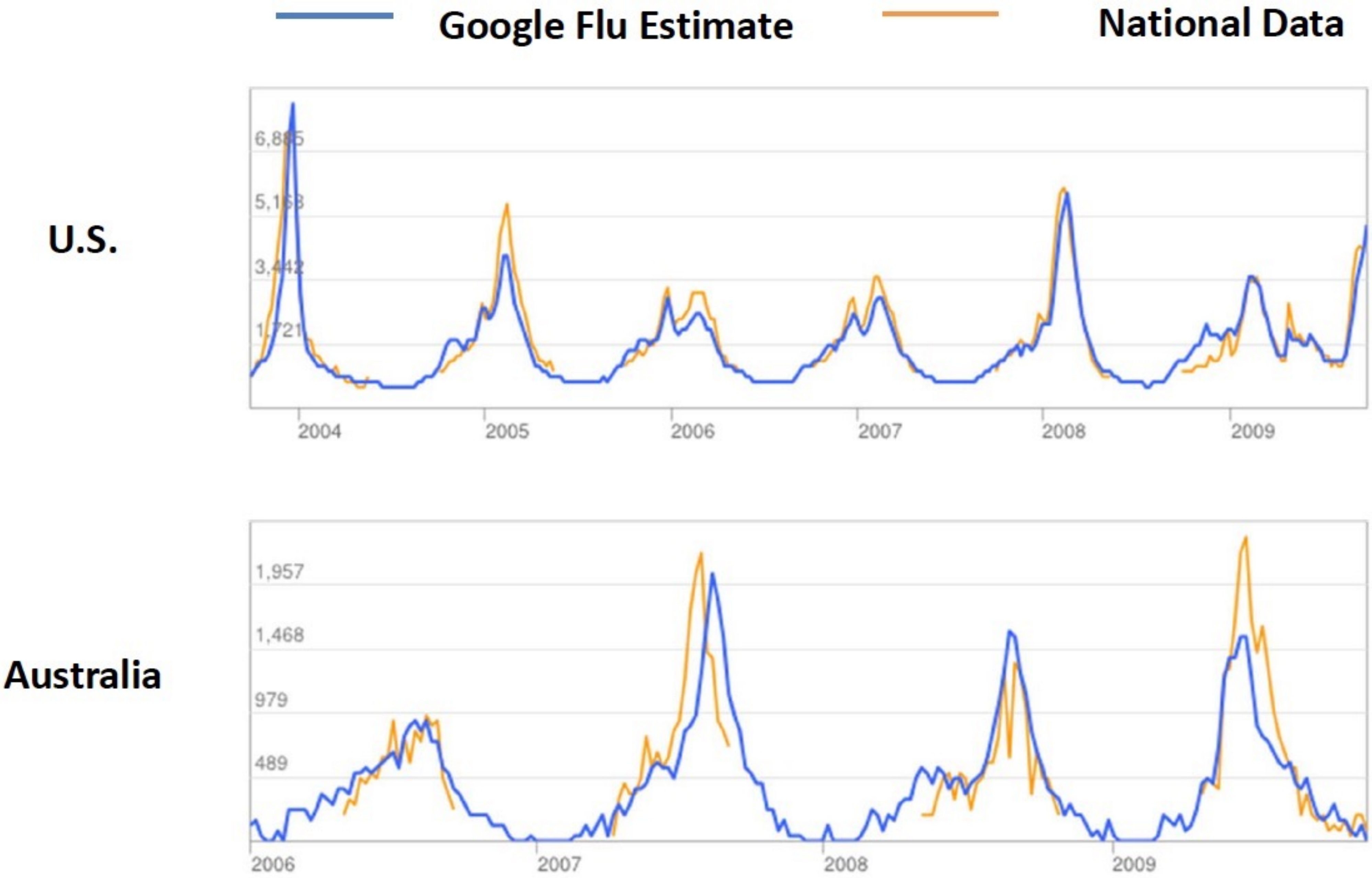
Google Flu Trends

- Took ***50 million*** of the most common search queries between 2003-2008 and did *a weekly count for each state*:
  - Each query is tested for *correlation* with CDC data
  - *Normalized* data by dividing count by total searches for the week (thereby getting a percentage)
  - *Ranked* – from most to least correlated

- Google added *top ranked queries* together to see what number would yield the most accurate results
  - The magic number is **45** (Google has not released them !)

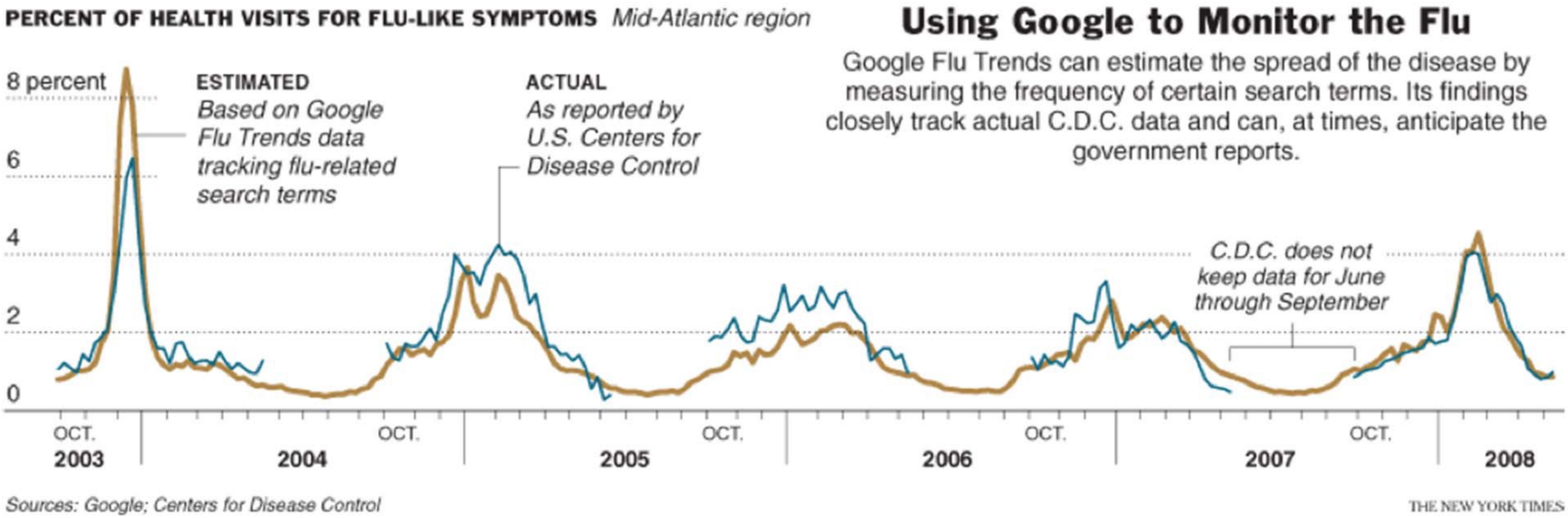
Google Flu Trends

- Generate accurate estimates faster than CDC
- It is theory-free



The New York Times

- Very promising retrospective comparison!



PERCENT OF HEALTH VISITS FOR FLU-LIKE SYMPTOMS *Mid-Atlantic region*

Using Google to Monitor the Flu

Google Flu Trends can estimate the spread of the disease by measuring the frequency of certain search terms. Its findings closely track actual C.D.C. data and can, at times, anticipate the government reports.

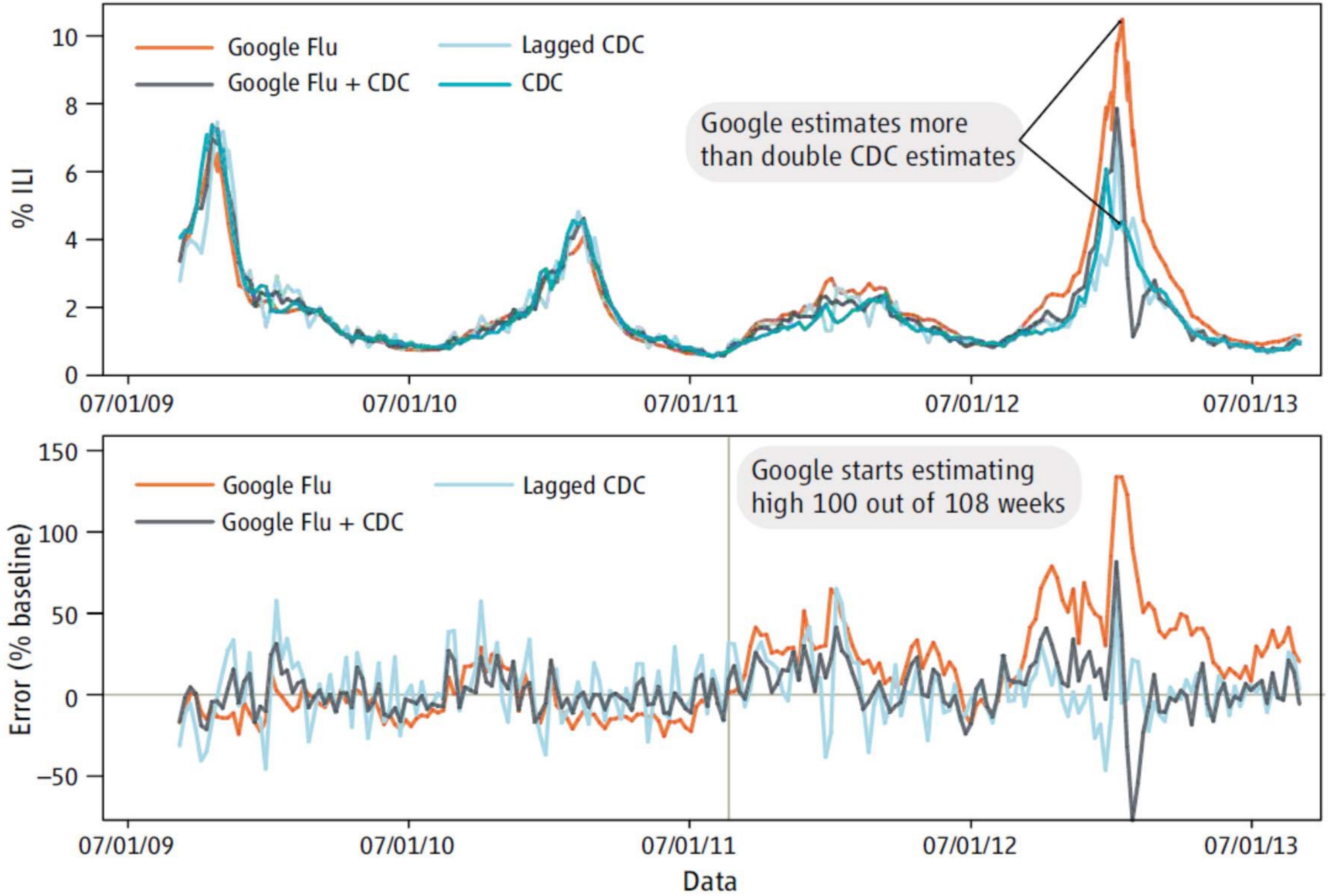
Sources: Google; Centers for Disease Control

THE NEW YORK TIMES

> "In April 2009, Dr. Brilliant said it epitomized the power of Google's vaunted engineering prowess to make the world a better place, and he predicted that it would save untold numbers of lives."

As time goes by … GFT's prediction was way off

- Starting from August 2011, GFT overestimated flu trends for 100 out of 108 weeks. In January 2013, the GFT estimate was twice as high as the real data.



Google estimates more than double CDC estimates

Google starts estimating high 100 out of 108 weeks

What Happen?

- No one outside Google really knows!
- Some reasonable explanations
  - Panic strikes?
  - Media driven flu hysteria?
  - N ≠ ALL?
  - Algorithm dynamics, tweaks?
  - Correlation does not imply causation?
- Is This All Bad News?
  - NO! Combining two week old CDC data with GFT provided a better model than GFT alone (*still fast* — one day for GFT compared to two weeks for CDC!).

It is not the size that matters!
Big data present great opportunities,
BUT an improper use may lead to erroneous predictions.

Applications of Big Data Across Industries

According to Research and Market reports, in 2017 the global Big Data market was worth $32 billion and by 2026 it is expected to reach by $156 billion.

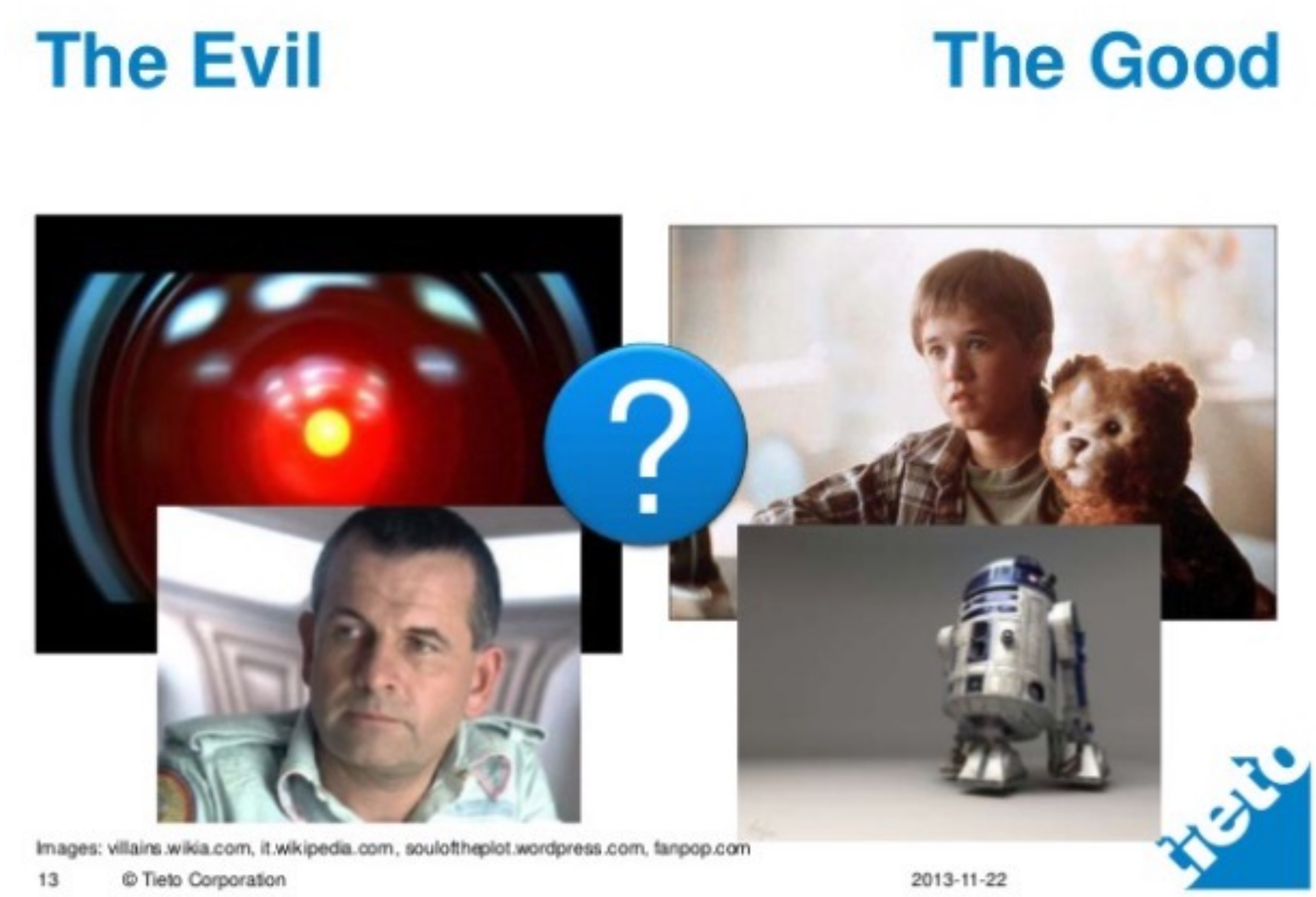
- Banking and Securities
  - The Securities Exchange Commission (SEC) is using big data to monitor financial market activity
  - heavily relies on big data for risk analytics, including; anti-money laundering, demand enterprise risk management, "Know Your Customer," and fraud mitigation

Big Data Applications

Communications, Media and Entertainment

- To
  - Create content for different target audiences
  - Recommend content on demand
  - Measure content performance
- Challenges:
  - Collecting, analysing, and utilizing consumer insights
  - Leveraging mobile and social media content
  - Understanding patterns of real-time, media content usage
  - Interpretable/explainable models

Big Data Ethics



The Evil

The Good

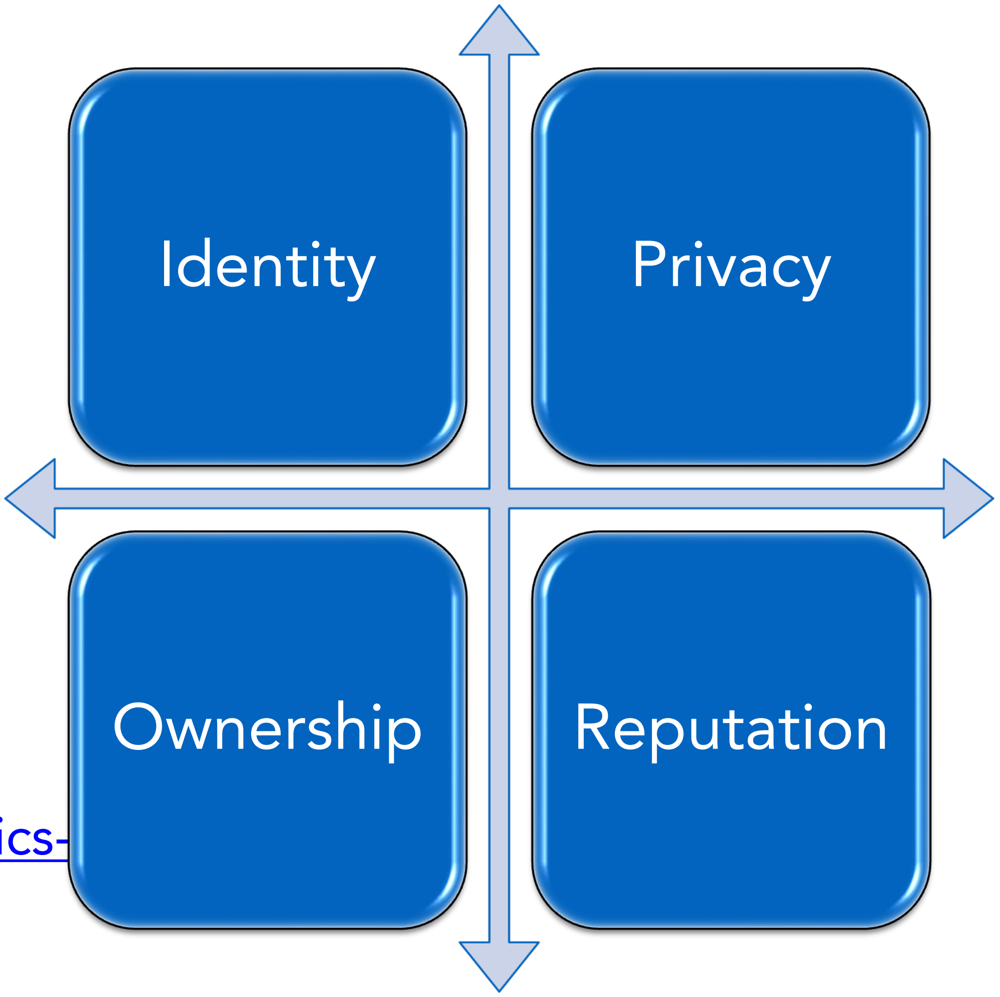
?

Images: villains.wikia.com, it.wikipedia.com, souloftheplot.wordpress.com, fanpop.com

13 © Tieto Corporation 2013-11-22

Four Aspects of Big Data Ethics

- Identity
  - Is offline existence identical to online existence?
- Privacy
  - Who should control access to data about you?
- Ownership
  - What does it mean to own data about ourselves?
- Reputation
  - How can we determine what is trustworthy?
  - How are we perceived and judged by using data – be fair



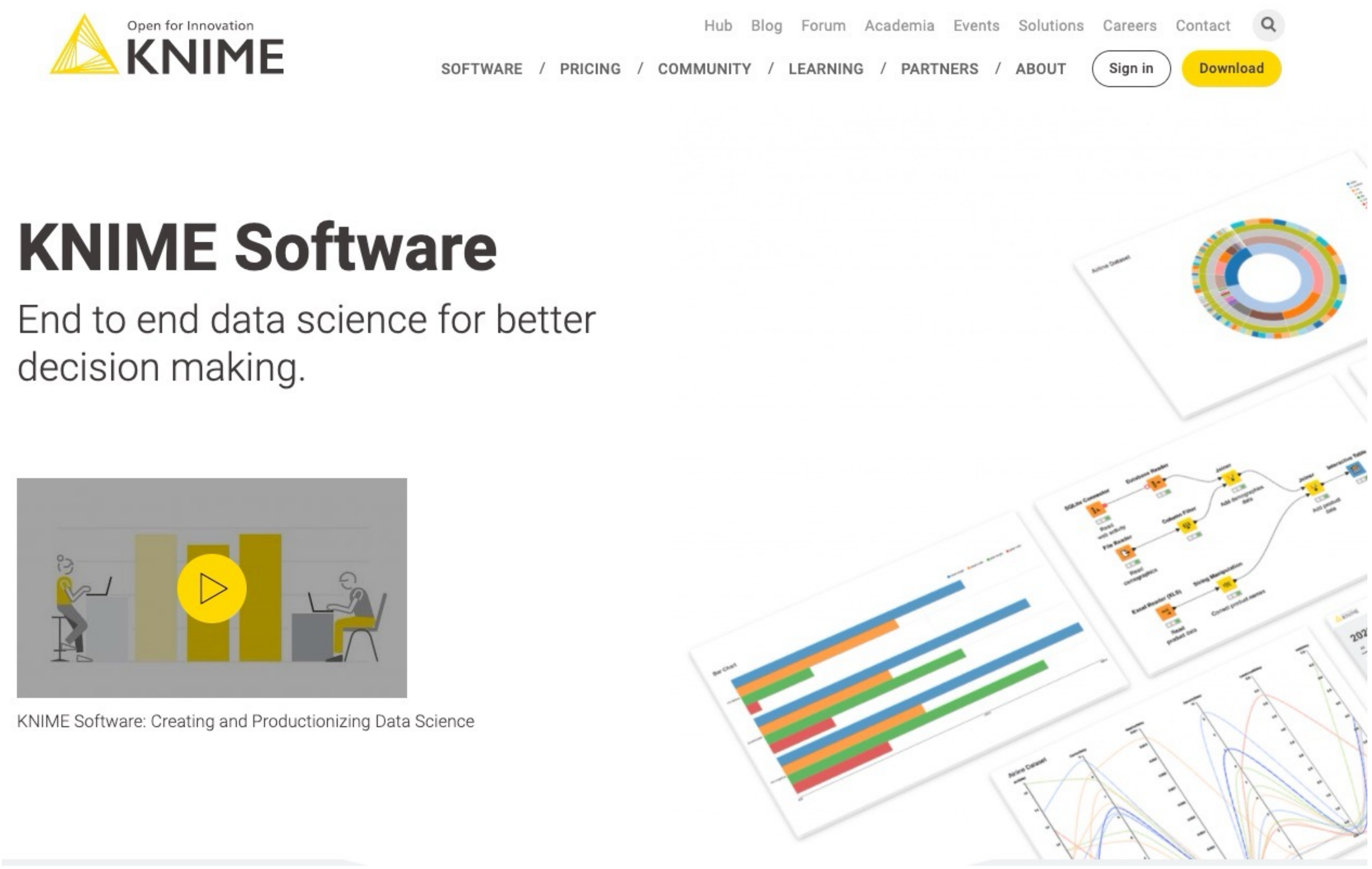
https://towardsdatascience.com/5-principles-for-big-data-ethics-b5df1d105cd3

Big Data Ethics

- Private customer data and identity should remain private
- Shared private information should be treated confidentially
- Customers should have a transparent view
- Big Data should not interfere with human will
- Big data should not institutionalize unfair biases

https://towardsdatascience.com/5-principles-for-big-data-ethics-b5df1d105cd3

Tools for Assignment 1



KNIME Software

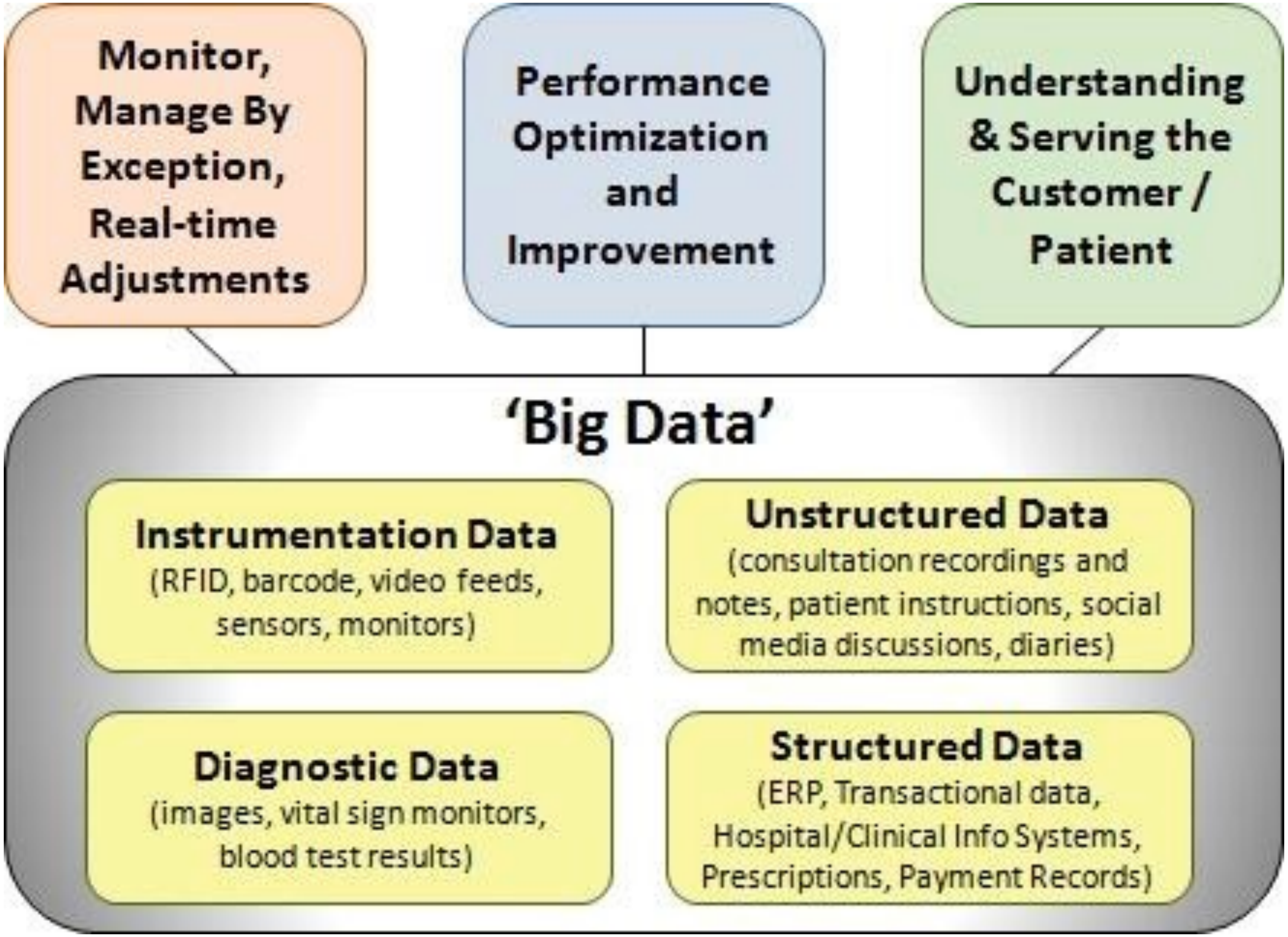
End to end data science for better decision making.

KNIME Software: Creating and Productionizing Data Science

https://www.knime.com/software-overview

Healthcare Providers

- Challenges
  - electronic data is unavailable, inadequate, or unusable
  - difficult to link data that can show patterns useful in the medical field.
-


Monitor, Manage By Exception, Real-time Adjustments
Performance Optimization and Improvement
Understanding & Serving the Customer / Patient
'Big Data'
Instrumentation Data (RFID, barcode, video feeds, sensors, monitors)
Unstructured Data (consultation recordings and notes, patient instructions, social media discussions, diaries)
Diagnostic Data (images, vital sign monitors, blood test results)
Structured Data (ERP, Transactional data, Hospital/Clinical Info Systems, Prescriptions, Payment Records)

[Source: Big Data in the Healthcare Sector Revolutionizing the Management of Laborious Tasks]

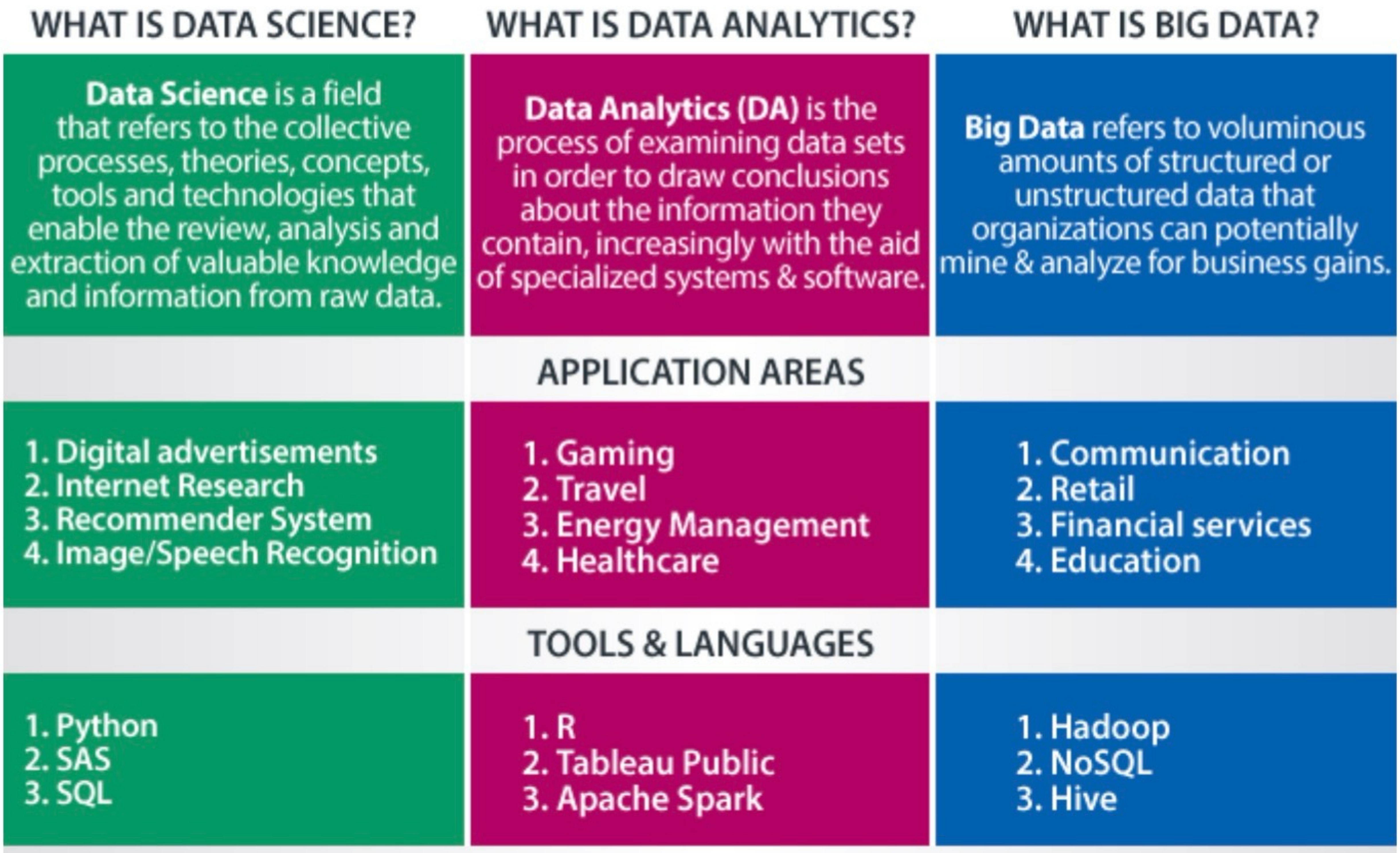
Education

- Challenges:
  - integrating data from different sources on different platforms and from different vendors that were not designed to work with one another.

- The University of Tasmania. An Australian university with over 26000 students has deployed a Learning and Management System that tracks, among other things, when a student logs onto the system, how much time is spent on different pages in the system, as well as the overall progress of a student over time.

Data Science vs Big Data vs Data Analytics

- You do not have to agree with ☺



WHAT IS DATA SCIENCE?	WHAT IS DATA ANALYTICS?	WHAT IS BIG DATA?
Data Science is a field that refers to the collective processes, theories, concepts, tools and technologies that enable the review, analysis and extraction of valuable knowledge and information from raw data.	**Data Analytics (DA)** is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems & software.	**Big Data** refers to voluminous amounts of structured or unstructured data that organizations can potentially mine & analyze for business gains.
	APPLICATION AREAS	
1. Digital advertisements 2. Internet Research 3. Recommender System 4. Image/Speech Recognition	1. Gaming 2. Travel 3. Energy Management 4. Healthcare	1. Communication 2. Retail 3. Financial services 4. Education
	TOOLS & LANGUAGES	
1. Python 2. SAS 3. SQL	1. R 2. Tableau Public 3. Apache Spark	1. Hadoop 2. NoSQL 3. Hive

https://www.whizlabs.com/blog/data-science-vs-big-data-vs-data-analytics/