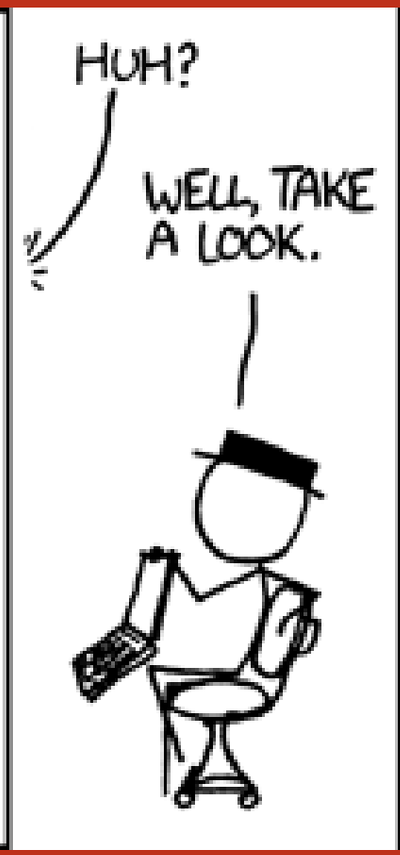
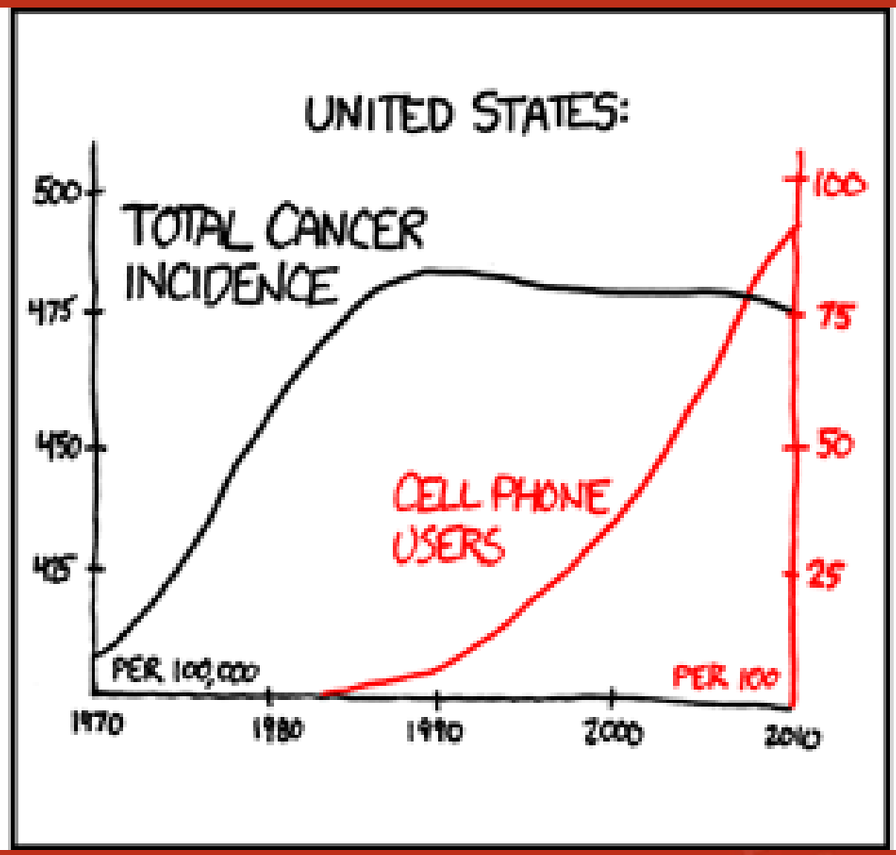
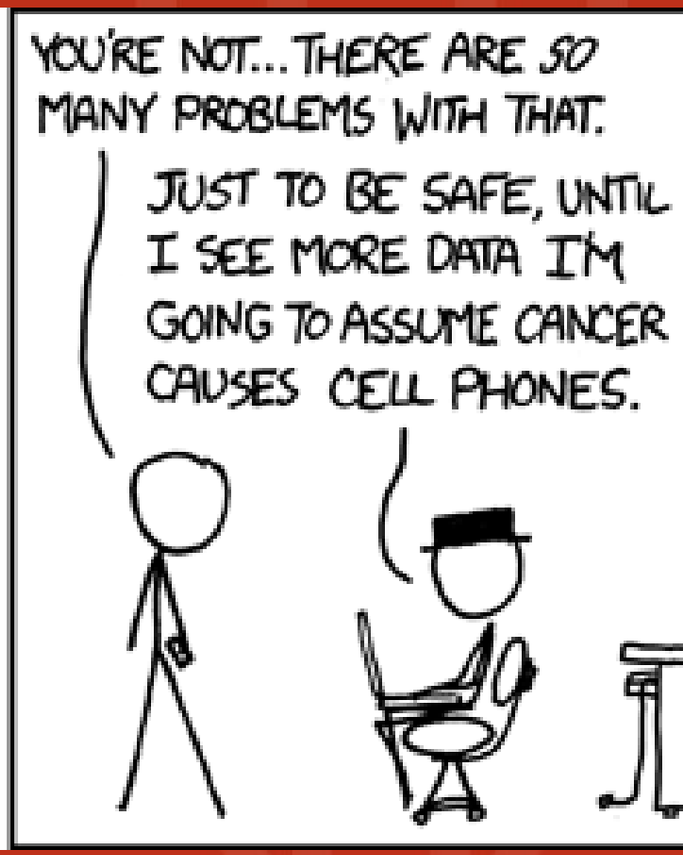

ANOTHER HUGE STUDY FOUND NO EVIDENCE THAT CELL PHONES CAUSE CANCER. WHAT WAS/ARE THE W.H.O. THINKING?
I THINK THEY JUST GOT IT BACKWARD.


HUH?
WELL, TAKE A LOOK.


UNITED STATES:
TOTAL CANCER INCIDENCE
CELL PHONE USERS
500
475
450
405
100
75
50
25
PER 100,000
PER 100
1970
1980
1990
2000
2010


YOU'RE NOT... THERE ARE SO MANY PROBLEMS WITH THAT.
JUST TO BE SAFE, UNTIL I SEE MORE DATA I'M GOING TO ASSUME CANCER CAUSES CELL PHONES.

XKCD

Data Mining : Introduction

Marco Loog
with a contribution by Roel Bouman

Radboud University Nijmegen

Very Briefly About Me…

- Math
- ML
- TUD

The Team

Marco Loog
marco.loog@ru.nl
lectures/general stuff

Tom Claassen
tomc@cs.ru.nl
lectures

Parisa Naseri
parisa.naseri@ru.nl
lecture

Roel Bouman
roel.bouman@ru.nl
lecture/practical sessions

Olivier Claessen
olivier.claessen@ru.nl
practical sessions

Student assistants [practical sessions, grading; in no particular order] :
Mikhail, Nadezhda, Tygo, Elina, Mats, Leah, Noah,
Fenna, Jochem, Martan, Cătălin, Ali, and Razvan

Outline

- Various organizational matters
- On to the actual introduction to data mining…

Course Basics

- 6 ECTS = 8 hours per week
- Lectures on Tuesday
- Practical sessions on Wednesday and Thursday;
  physically or on-line [through Discord]
- Midterm and endterm exams on campus
- “Learning tasks” : reading material + exercises for self study
- Lots of info on BrightSpace

Evaluation

- [Check the syllabus!]
- Two multiple-choice exams : one mid-term and one end-term*
- Project (more details later)
- Six homework assignments
- Mandatory: score ≥ 5.5 for at least 4 out of 6 homework assignments!
- Final grade: 0.35 × Midterm + 0.35 × Endterm + 0.3 × Project**

*Average of the exams needs to be at least 5.0 to pass the course

**If final grade ≥ 5.5 (pass), average of homework assignments replaces half of the average exam grade or the project grade if it helps to give you a higher grade

Lectures

- No streaming or recording
- Video lectures from some earlier years available through BrightSpace
- Note, however, one of the major changes :
  the chapter on association analysis has been removed form the course

Learning Tasks

- Course Content Learning Tasks
- To keep track and to study for exam
- Exercises are meant to practice
- Ask feedback when stuck on one of the exercises!

Background

Data mining is the art and science of extracting knowledge out of databases. This is a rather vague and general definition that will be made more specific. What kind of data? What type of problems? What kind of techniques are available? How does data mining relate to other fields?

Objectives

After completing this task you will be able to

- describe the objectives of data mining, its challenges, and its relationship with other fields of science;
- subdivide data mining tasks into different categories and give examples of problems for each of these.

Instructions

1. Read and study chapter 1 of TSK.
2. Make exercises 1 through 3 of TSK, section 1.7.
3. What is the definition of data mining in TSK? Find two other definitions for data mining and compare them.
4. Find at least two examples of data mining applications that appeared in the press (the more recent and the closer to home, the better...). Describe these. What data mining tasks are involved?
5. *Data mining is very closely related to machine learning. Check out this note to learn about its aims, success, and challenges.

Products

- Answers to the exercises.
- Three different definitions of data mining.
- Two "real-world" examples of data mining.

Reflection

- Can you explain the difference between data mining and statistics, knowledge discovery, machine learning, and so on?
- Given a particular problem, can you tell what data mining task it belongs to?
- Can you describe some challenges in machine learning/data mining?

ABOUT THE PRACTICALS

- Bridge the gap from theory to practice

- Assignments are in Python (Jupyter notebooks)
  - Libraries: NumPy, scikit-learn, Matplotlib, etc.

- Read the assignment guidelines on Brightspace! (Content → assignments → General Assignment guidelines)
  - How to present plots, figure descriptions, etc.

- Use the metadata_checker.py script! (Content → assignments → metadata_checker)

PRACTICAL ORGANISATION

- You work in pairs
  - Use the break, the practicals, Brightspace or Discord for finding a partner

- Practicals are at Wednesdays 15:30-17:15 (CS + other), and Thursdays (15:30-17:15)
  - You are welcome to join one, or both

- We use Discord as a teaching tool, you can join the channel through the Brightspace invite
  - Use Discord to ask questions during practicals
    - Reduced waiting times/TA distribution
  - Also online distance during practicals

PLANNING

- Q1:
  - First week: setup and starting A1
  - After that: 2 weeks per assignment
- Q2:
  - 2 assignments of 2 weeks, 1 assignment of 1 week
  - 1-2 weeks of project assistance practicals

PRACTICAL PREPARATION

- Before or during the first practical, you should:
  - Install a recent Anaconda distribution: https://www.anaconda.com/download

- The advised programming environment is directly in a self-run Jupyter Notebook
  - Online collaboration platforms tend to corrupt Notebook metadata needed for grading

- You are responsible for turning in a notebook which:
  - Can be run on a recent Anaconda installation
  - Passes the metadata_checker.py script!

WHERE DO THE PRACTICALS FIT?

- The practicals prepare you for the project
  - They are not representative for the midterm/final exam!

- The practicals should teach you the basic skills you need to do data science in practice
- And, more importantly, be critical of what not to do!

LARGE LANGUAGE MODELS (CHATGPT ETC.)

- LLMs are unavoidable nowadays, but you should be careful in using them!

- By thinking about the questions yourself, you will learn the lessons we are trying to teach you
  - Passing on the question might give you the wrong answer/gibberish (and you won't know why!)

- Ethical implications
  - What about the integrity/privacy of the training data?
  - How much energy do LLMs consume?

- In the end: if we think you used LLMs without understanding/being critical of the results, and/or have not contributed yourself → low(er) grade

RE-USING ASSIGNMENTS (FOR RESIT STUDENTS)

- You can re-use assignment grades from previous years
  - Assignment 5 has changed, so can't be re-used

- If you want to re-use any grades, e-mail Olivier (Olivier.Claessen@ru.nl) with the specific assignment(s), and the year you did the assignment in

QUESTIONS?

- If you have any questions regarding these few slides, ask me!

- For later questions regarding grading/practical content:
  - Olivier Claessen: olivier.claessen@ru.nl

The Book and Its Chapters

- We follow the **second edition** of the book “Introduction to Data Mining” by Tan, Steinbach, Karpatne, and Kumar

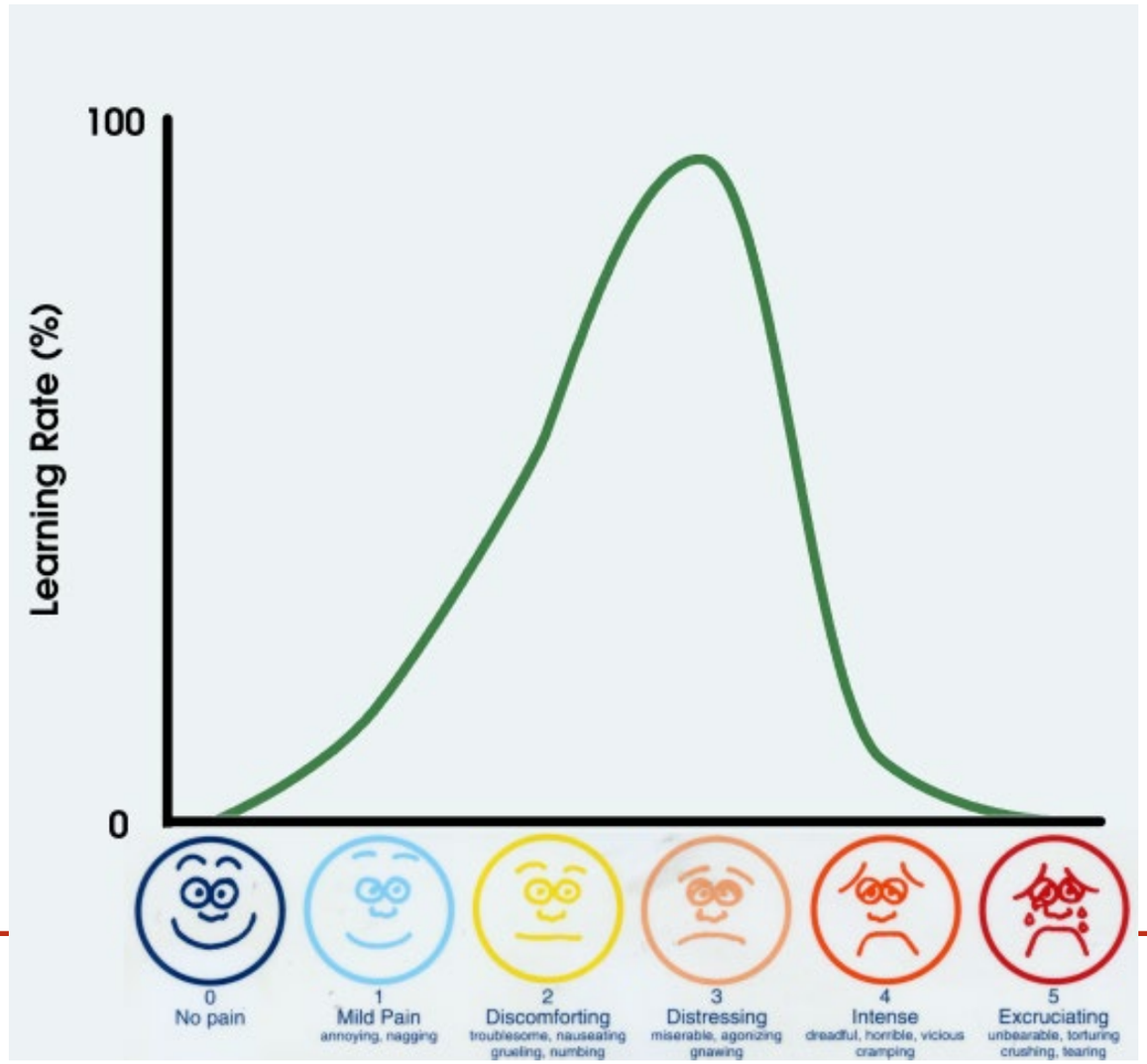
- See https://www-users.cse.umn.edu/~kumar001/dmbook/index.php
  where you can find
  - slides, errata, and some chapters
  - a “What is New in the Second Edition?”
  - …

- An overview of the chapters relevant to the course
  is provided under Lectures in BrightSpace

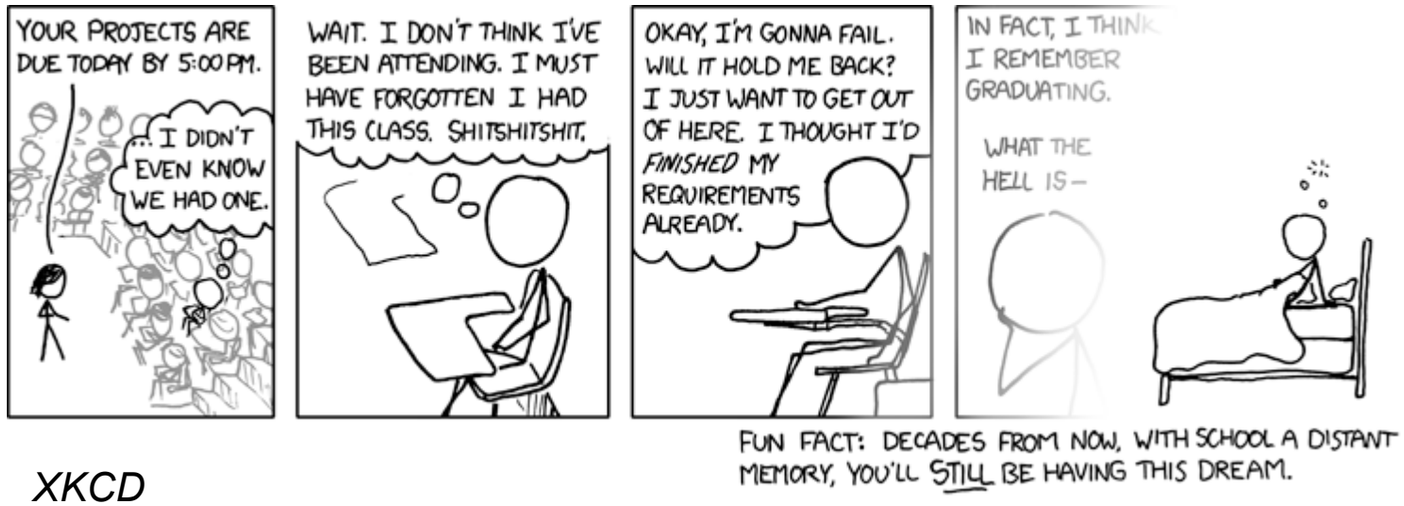
Advice
[From Somebody Who Taught This Course Many, Many Years]

- Keep track
- Follow the lectures
- Start looking at the homework at least 2 weeks before the deadline
- Practice with some of the exercises mentioned in the learning tasks
- If you're stuck, formulate why and ask!

Theory of Pain and Learning



Struble, 2004



XKCD

Bill Watterson

Comments, Feedback, and Being Helpful…

- Be civil
- Be constructive
- Be timely
- …

- We take feedback into consideration absolutely seriously

"IT'S PAYBACK TIME! WE'RE BEING ASKED TO RATE OUR TEACHER!"

Be Critical!

- …yet constructive

- …towards
  - What we tell you
  - What you read
  - The results your computer spits out
  - What inputs algorithms take
  - What outputs methods provide
  - Code provided online
  - …

  - Last but not least, be critical constructive towards yourself

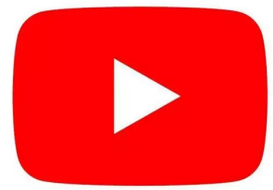
That Data Mining Introduction…

- Motivation
- Examples
- Bit of history [or rather, context?]
- Challenges

Why Mine Data? Commercial Viewpoint

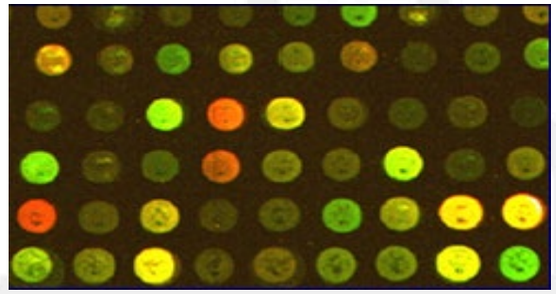
- Lots of data is being collected and warehoused
  - web data, e-commerce
  - purchases at department/ grocery stores
  - bank/credit card transactions

amazon

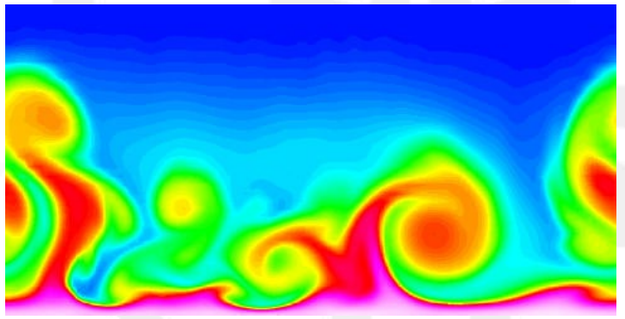


- Computers have become cheaper and more powerful
- Competitive pressure is strong
  - provide better, customized services for an edge [e.g. in customer relationship management]

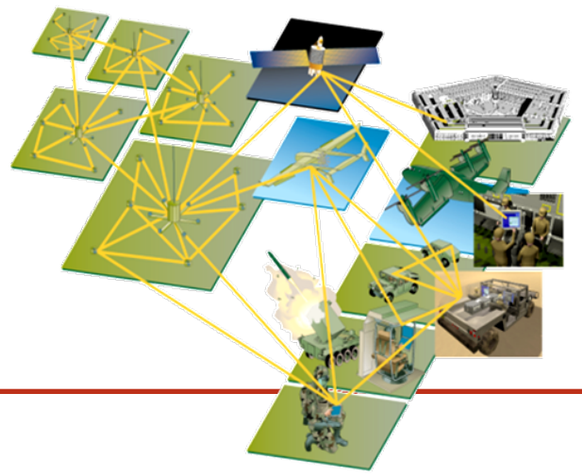
Why Mine Data? Scientific Viewpoint

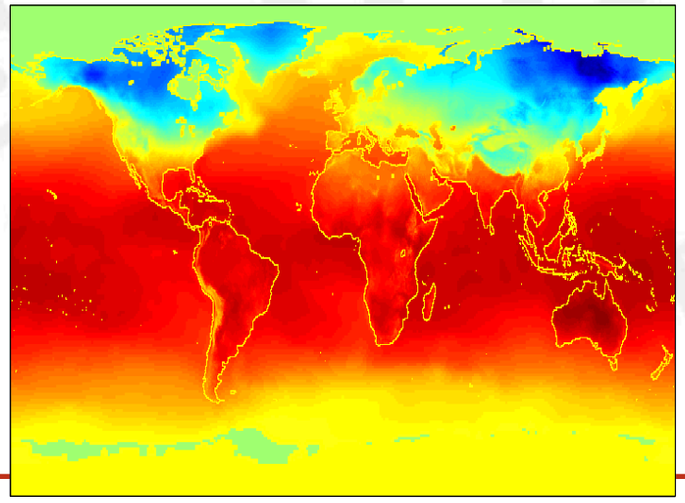


- Data collected and stored at enormous speeds [Gb/hour]
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data



- Traditional techniques infeasible for raw data
- Data mining can help scientists to analyse data and form hypotheses

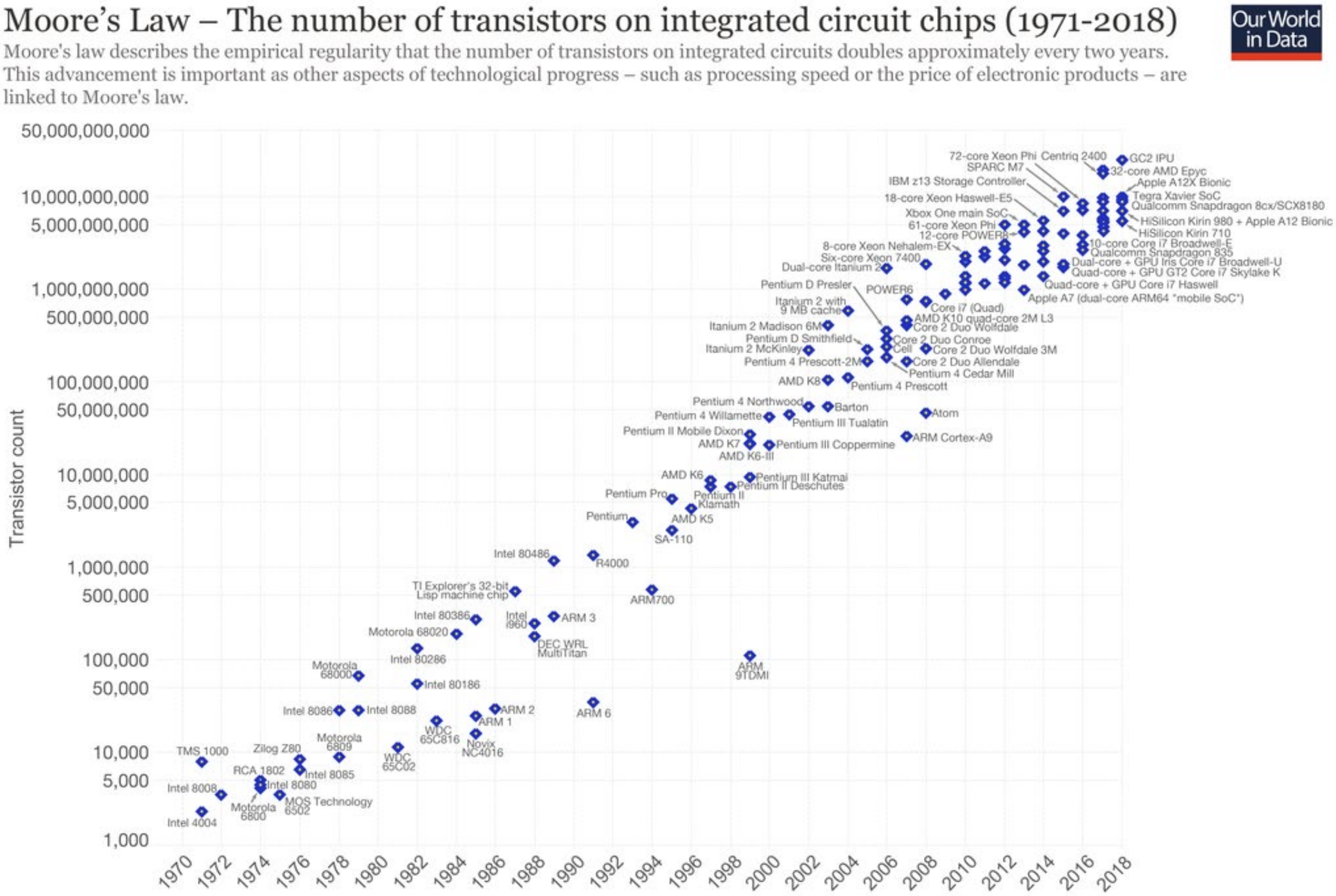




Another Scientific Viewpoint

- We can also investigate methods as such…
- Do we really understand the way the methods work on some data, how to interpret its outcome, etc.?
- Can we reason/understand why one method may be better than another method in specific cases?
- What new settings are there where data mining can also be useful and how to design methods for those settings?

A Known “Law”…



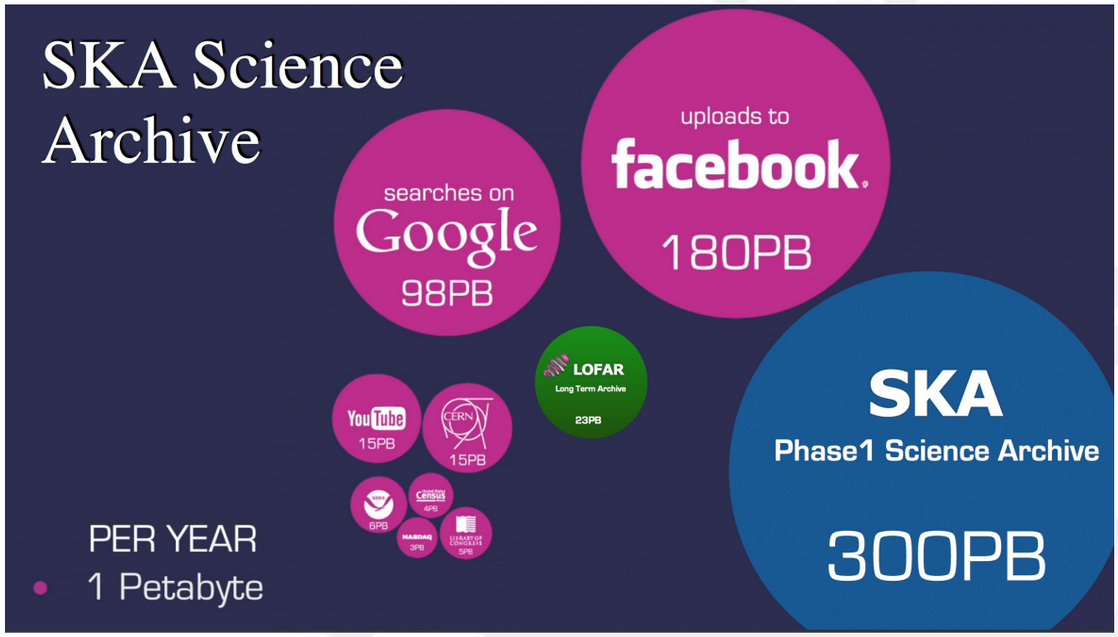
Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

Growth of Data Sets…

- Current data volume estimated at ~50 Zettabyte [10^{13} GB] and doubling every 1.5 years…
- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all

Examples of Massive Datasets

- Pubmed text database
  - Records for >30 million published articles
- Web search engines
  - 60 billion Web pages indexed
  - 100s of millions of site visitors per day
- CALTRANS loop sensor data [traffic]
  - Every 30 seconds, thousands of sensors, 2 Gbytes per day
- NASA MODIS satellite
  - Coverage at 250m resolution, 37 bands, whole earth, every day
- Retail transaction data
  - Ebay, Amazon, Walmart: >100 million transactions per day
  - Visa, Mastercard: similar or larger numbers

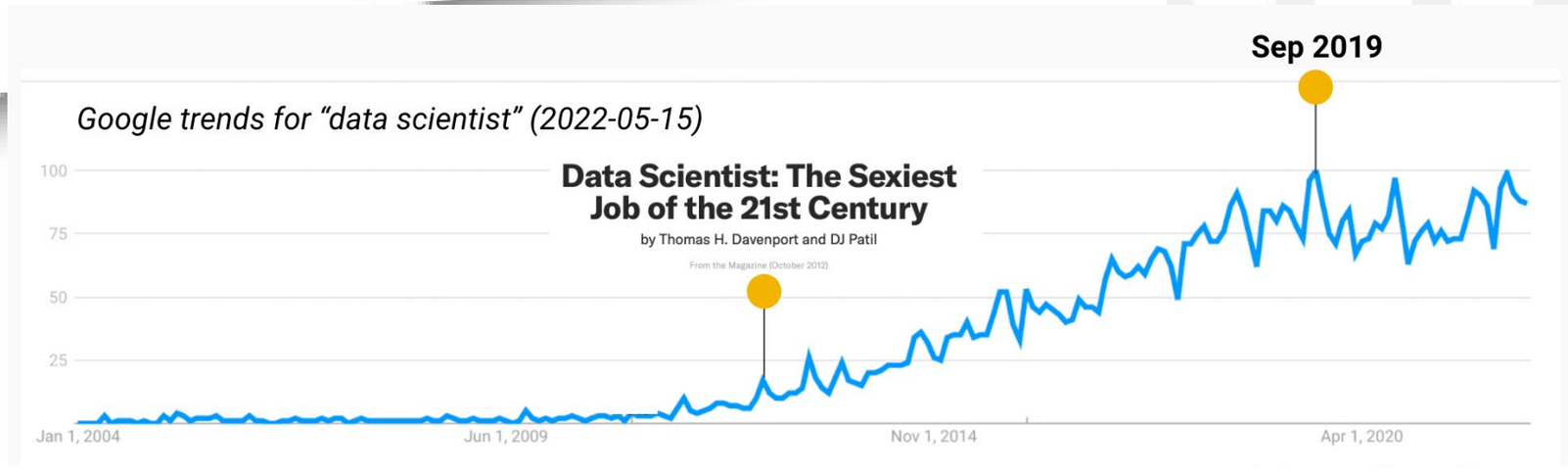


Harvard Business Review

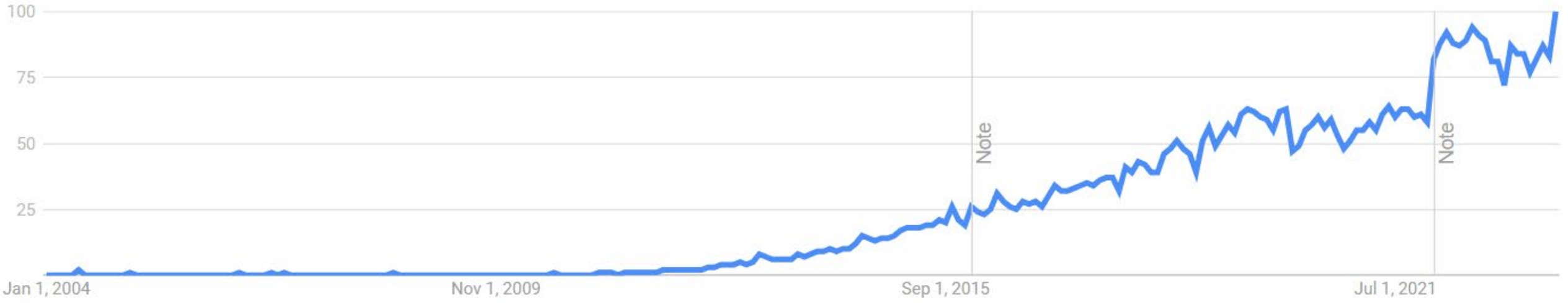
Data Scientist:
The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

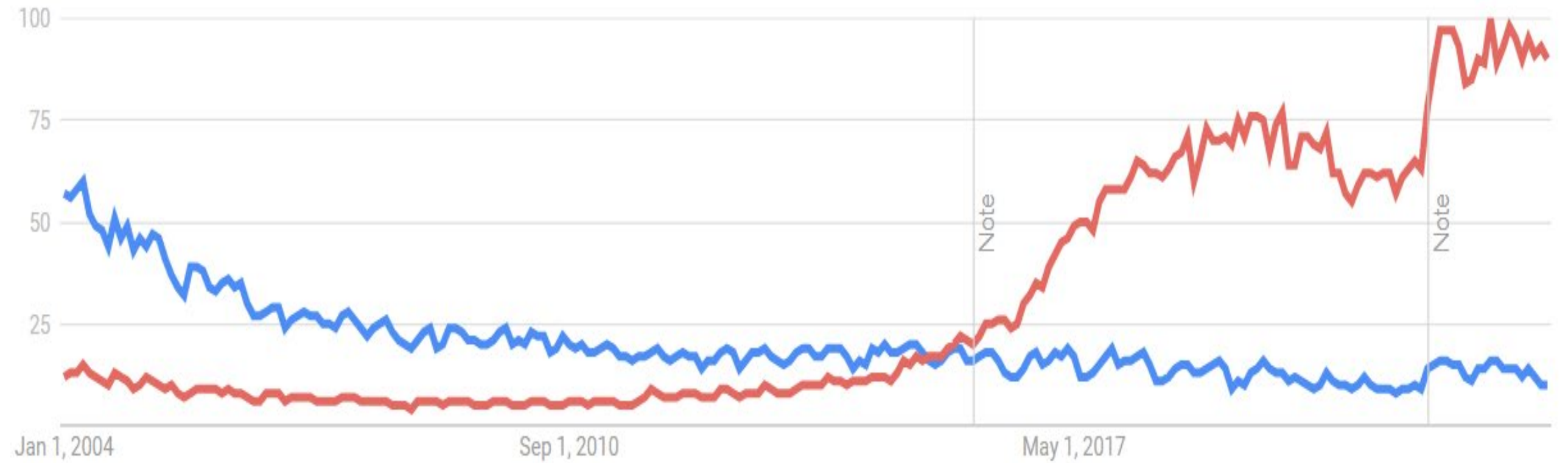


I Checked the Curve Yesterday…

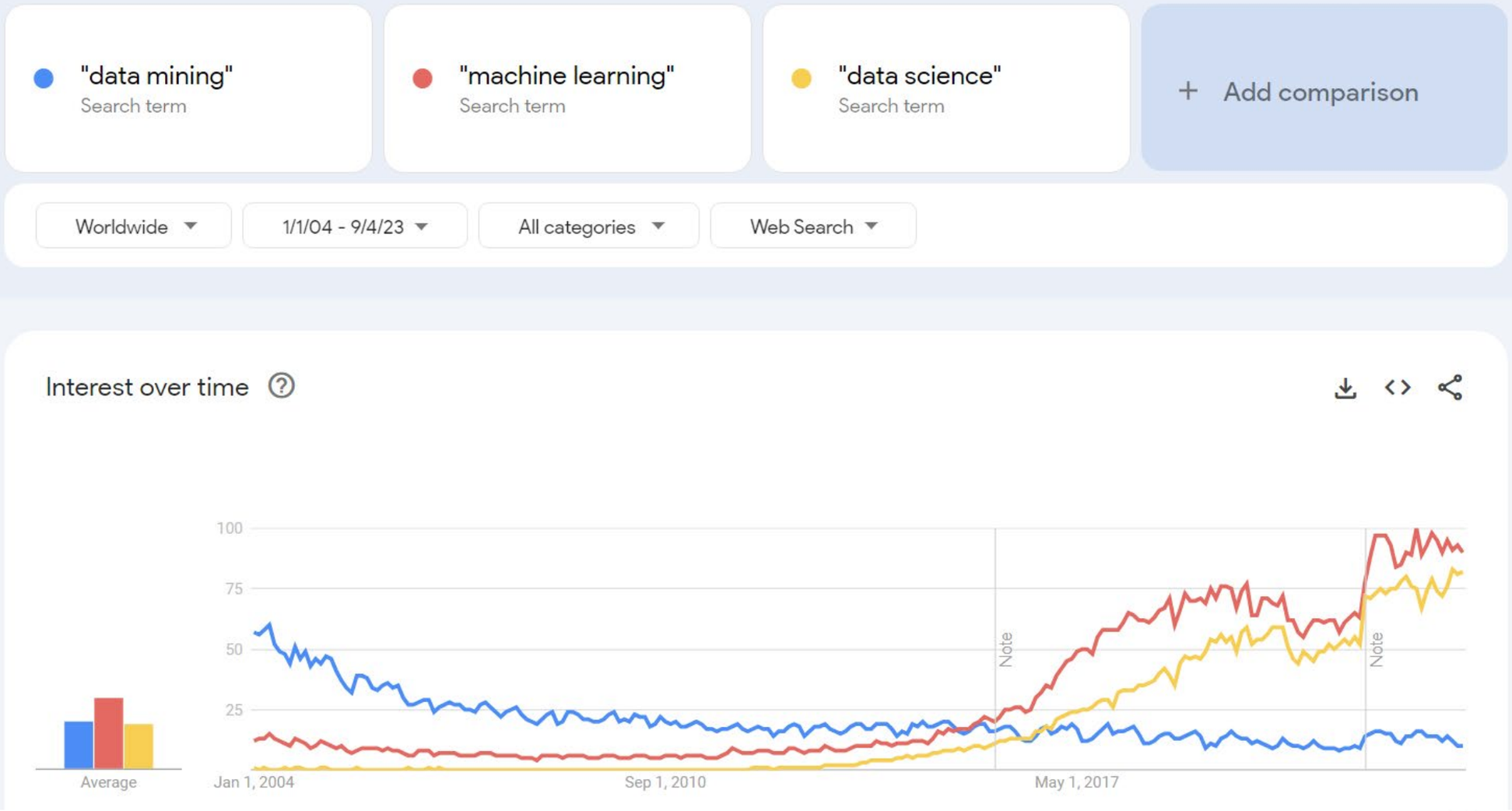


This, however, is Data Mining…

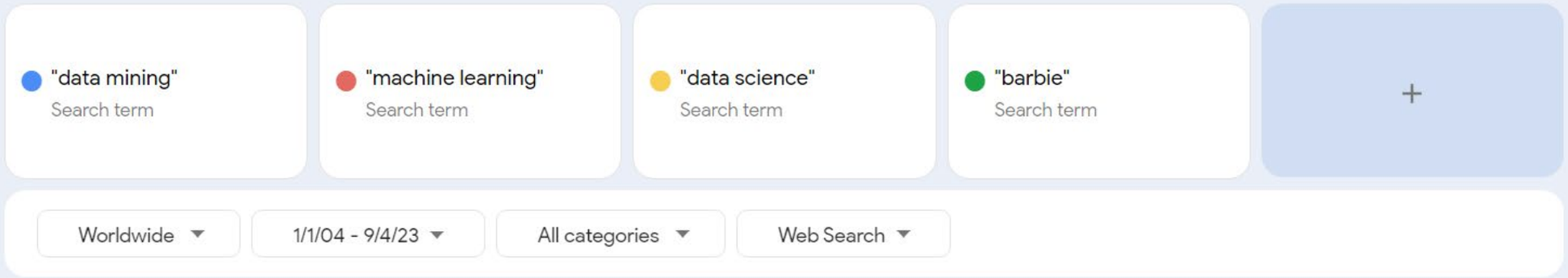
- What's in the name…
  Here's also what machine learning is doing…



Adding in Data Science…



And Putting Things in Perspective…

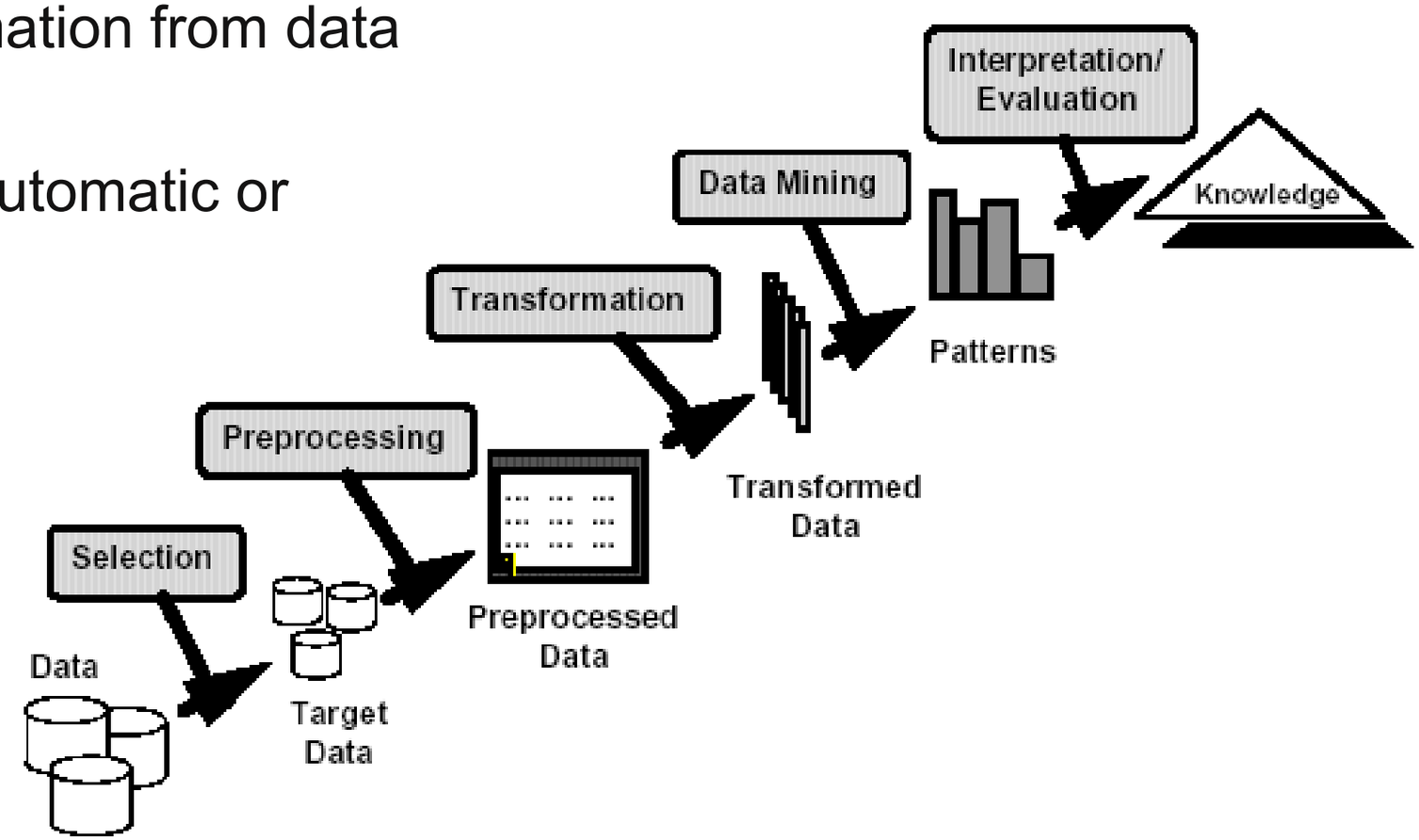


What is Data Mining?

Many definitions :

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data

- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data to discover meaningful patterns



Another Definition

"Data-driven discovery of models and patterns from massive observational data sets"

- Statistics, inference
- Retrospective analysis
- Languages and Representations
- Engineering, data management

Smyth, 2003

What is [Not] Data Mining?

- What is not Data Mining?
  - Look up phone number in phone directory
  - Query a Web search engine for information about “Amazon”

- What is Data Mining?
  - Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest vs. Amazon.com)

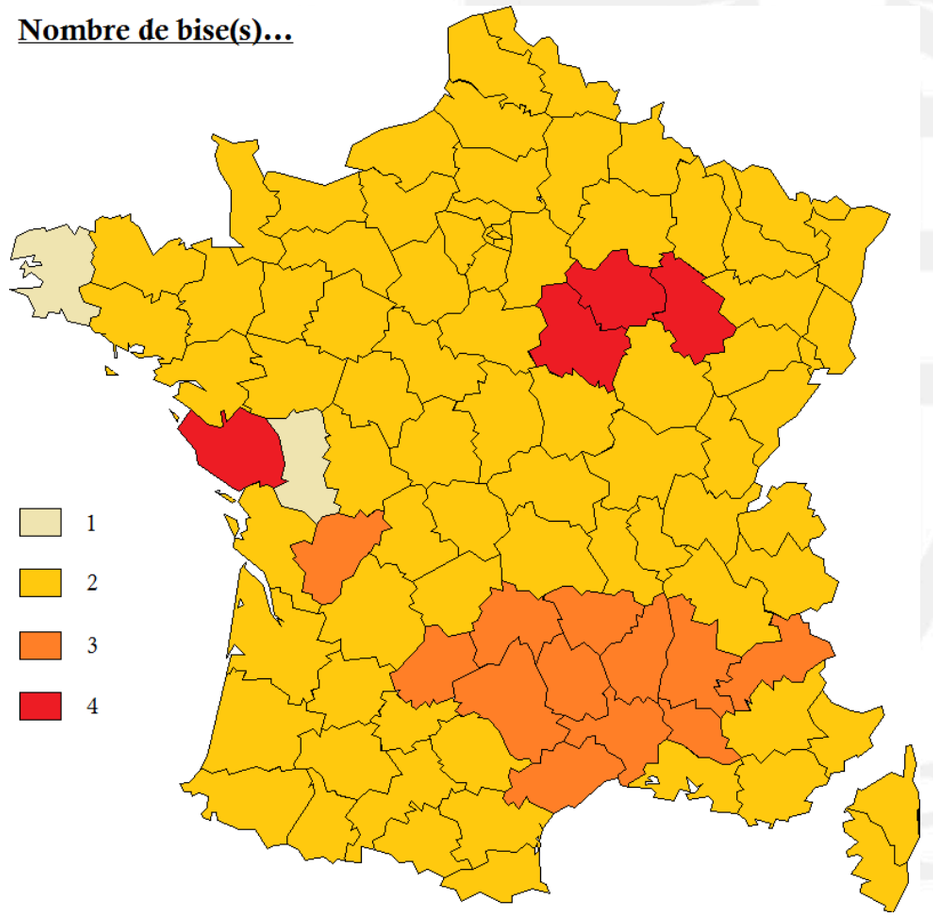
What is Data Mining?

- Any method that distills [actionable] information/knowledge from data?

A Lunch Discussion…

- One encounters [some form of] data mining in many settings…
- Bar lies very low for something to be called data mining?

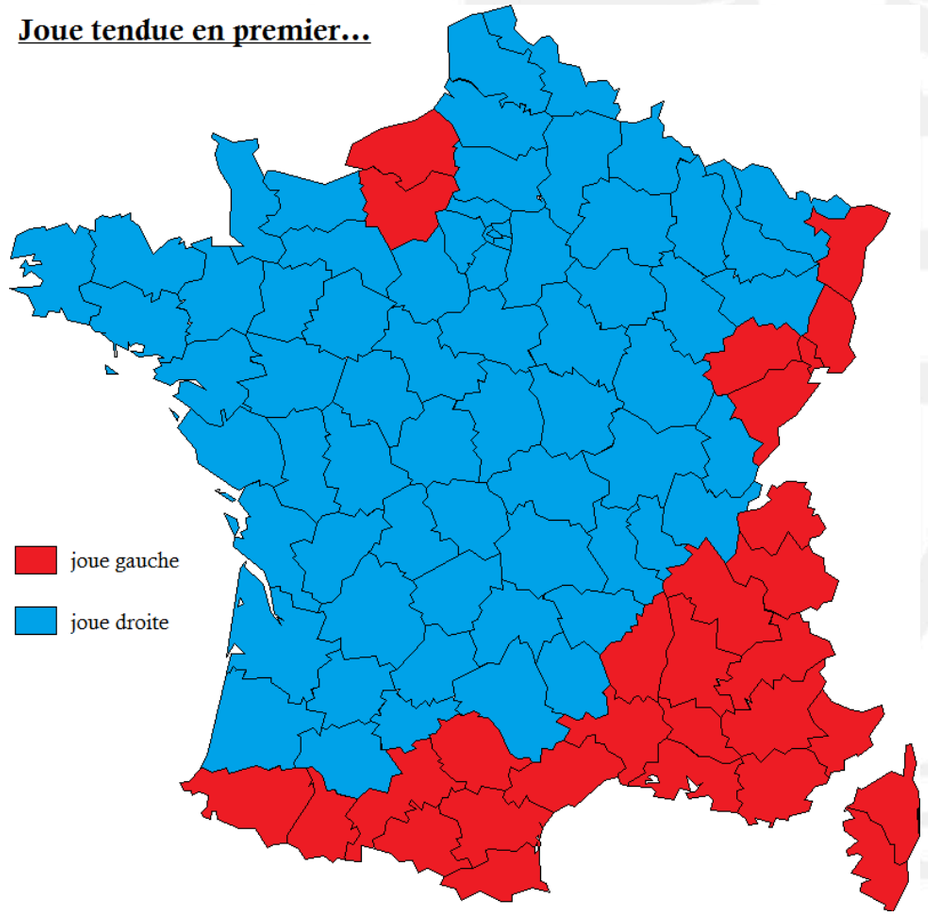
- Number of kisses when greeting



A Lunch Discussion…

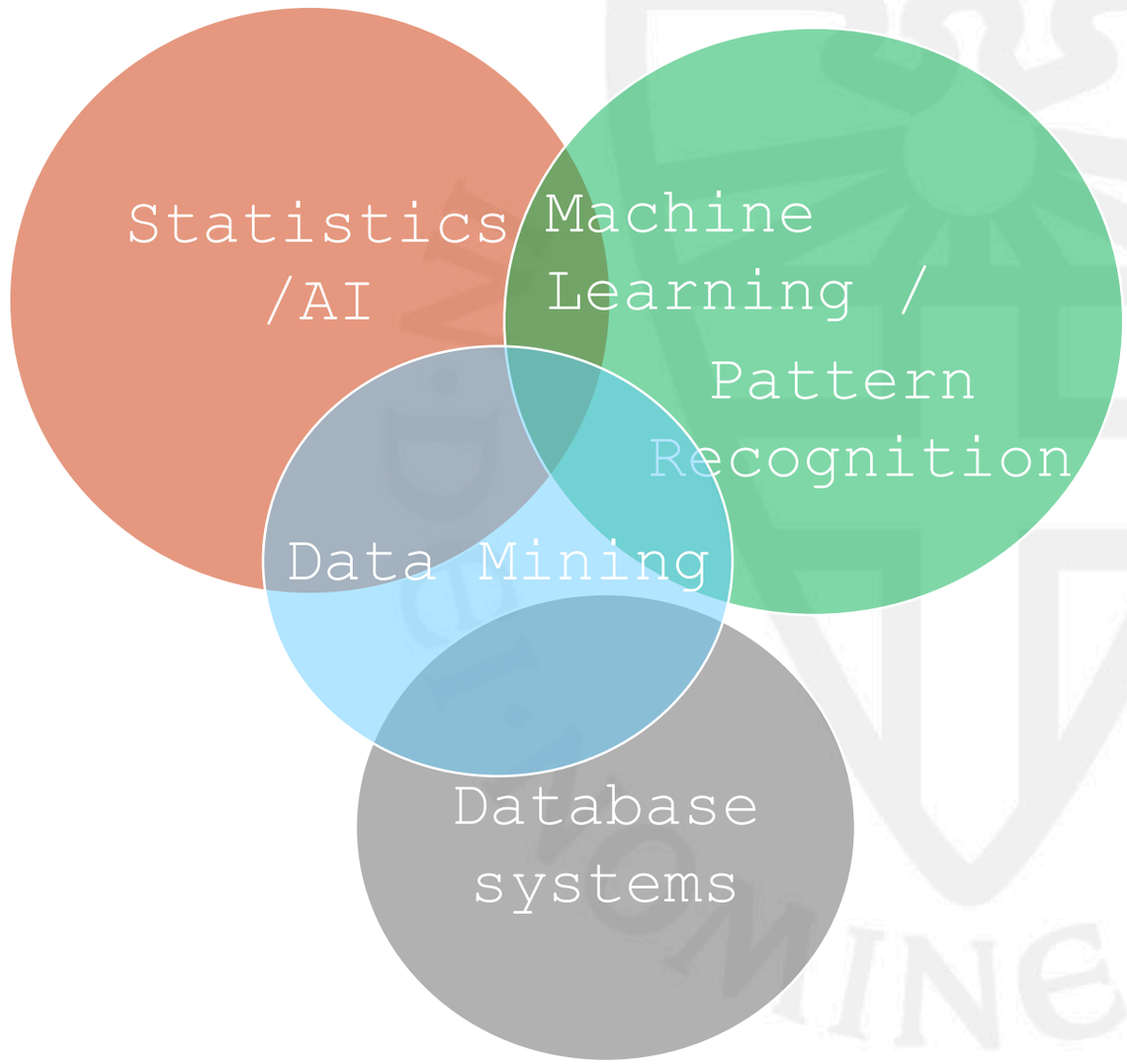
- One encounters [some form of] data mining in many settings…
- Bar lies very low for something to be called data mining?

- Cheek turned for first kiss…



Origins of Data Mining

- Draws ideas from machine learning, AI,
  pattern recognition,
  statistics,
  database systems,
  optimization
  …

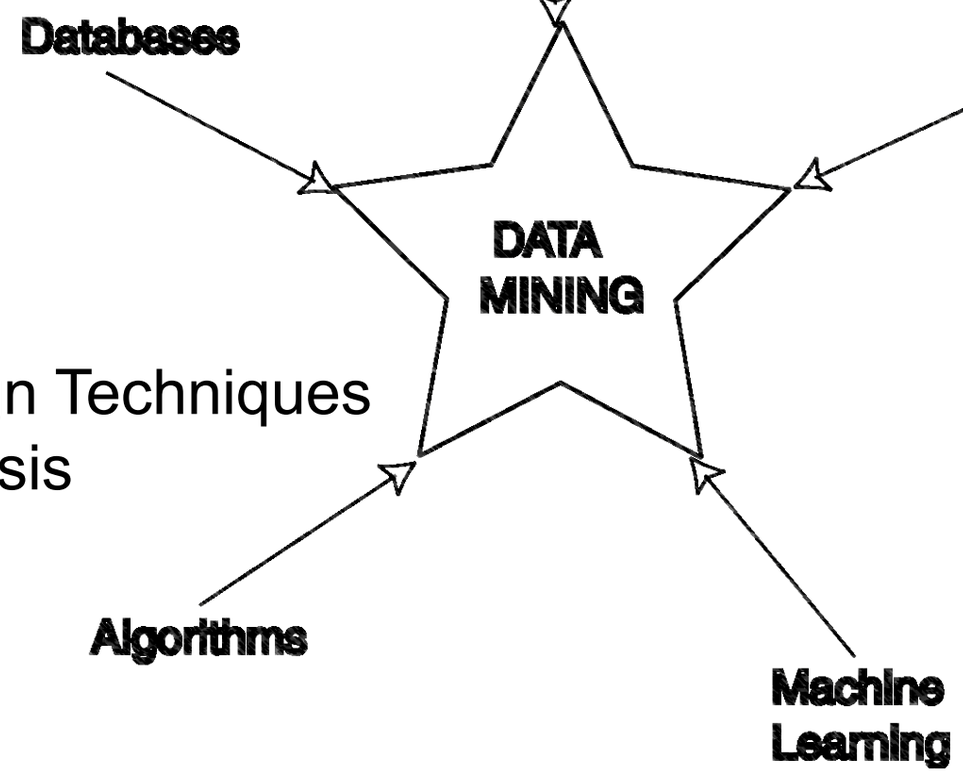


- Traditional techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data

Data Mining Development

- Relational data model
- SQL
- Association rule algorithms
- Data warehousing
- Scalability techniques

- Similarity measures
- Hierarchical clustering
- IR systems
- Imprecise queries
- Textual data
- Web search engines



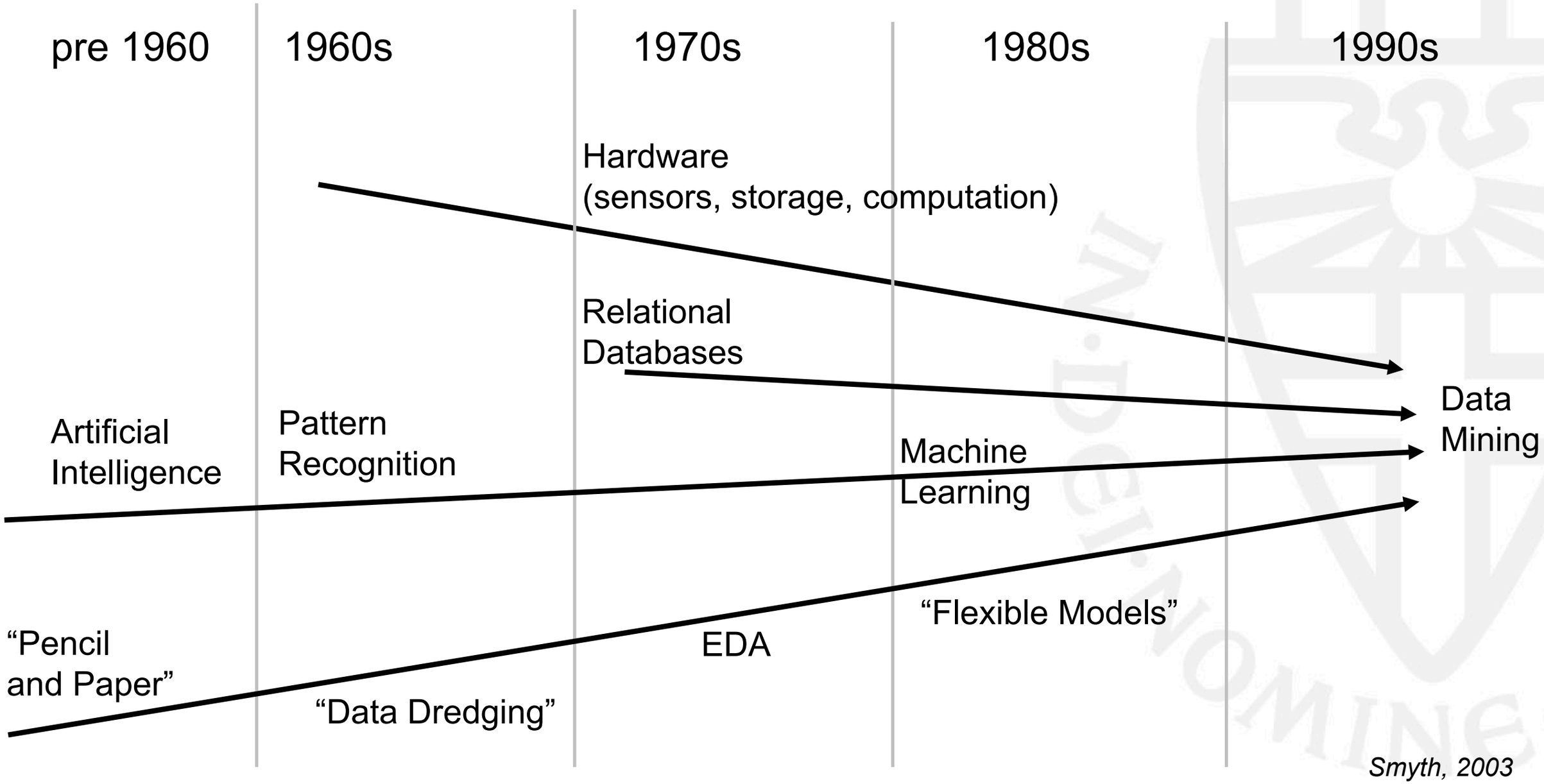
- Algorithm Design Techniques
- Algorithm Analysis
- Data Structures

- Bayes theorem
- Regression analysis
- EM algorithm
- K-means clustering
- Time series analysis

- Neural networks
- Decision tree algorithms

Dunham, 2003

Origins of Data Mining

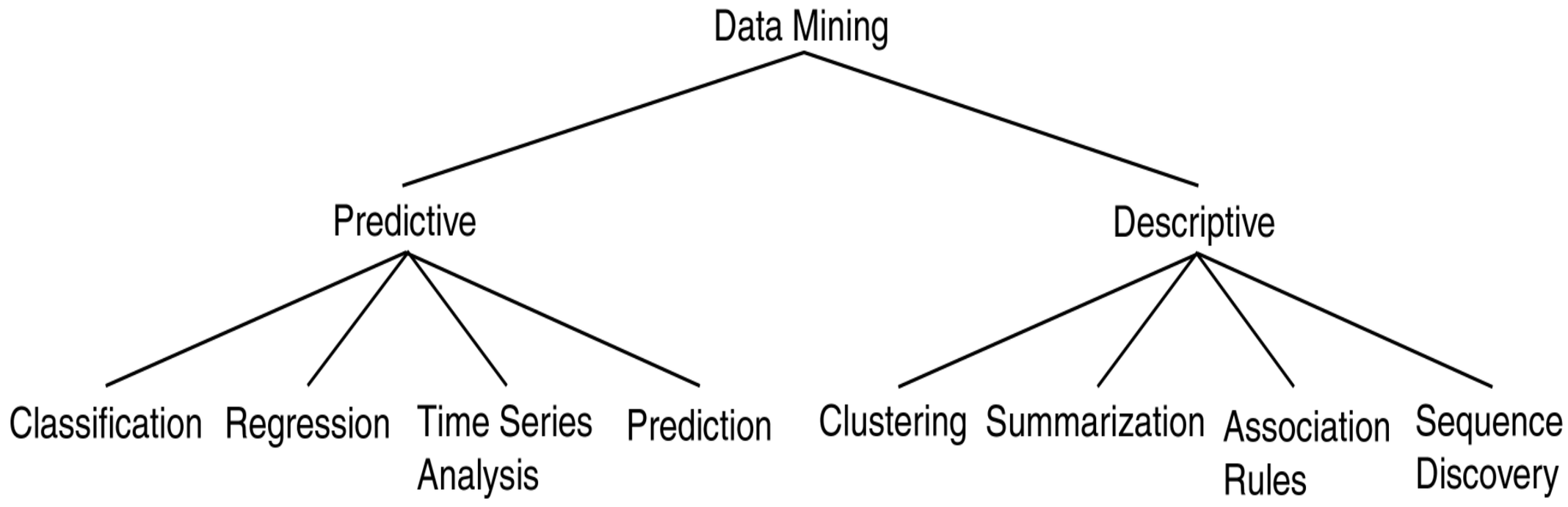


Smyth, 2003

Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.

Data Mining Tasks…



Dunham, 2003

Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association rule discovery [Descriptive]
- Regression [Predictive]
- Deviation detection [Predictive]

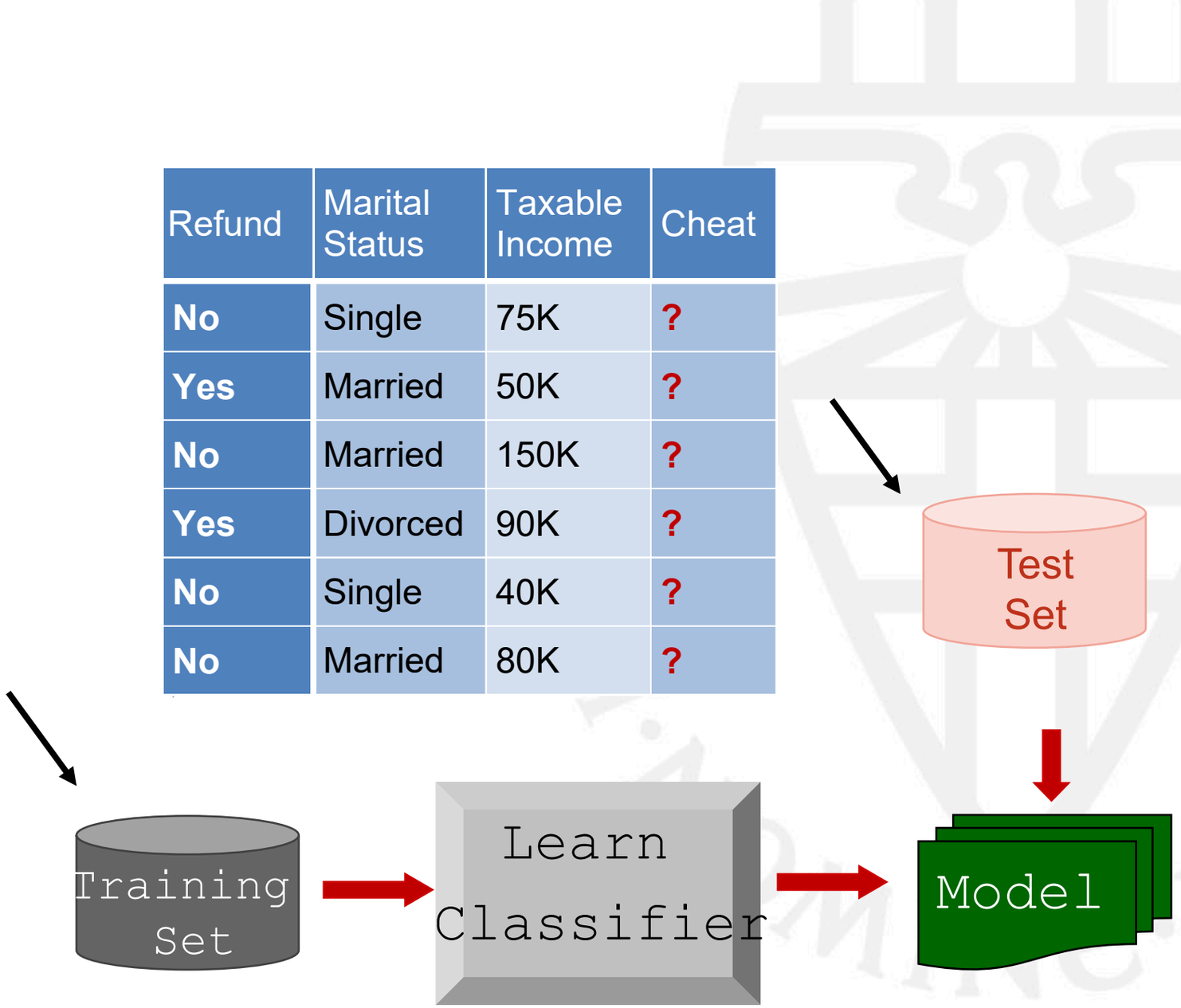
Classification…

- Given a collection of records [training set]
  - Each record contains a set of attributes, one of the attributes is the class.
- Find a model for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification…

	categorical	categorical	continuous	class
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	**?**
Yes	Married	50K	**?**
No	Married	150K	**?**
Yes	Divorced	90K	**?**
No	Single	40K	**?**
No	Married	80K	**?**



Classification : Example Application

Direct Marketing

- Goal: Reduce cost of mailing by targeting a set of consumers likely to buy a new cell-phone product.

- Approach:
  - Use the data for a similar product introduced before
  - We know which customers decided to buy and which decided otherwise
    This {buy, don't buy} decision forms the class attribute
  - Collect various demographic, lifestyle, and company-interaction related information about all such customers
  - Type of business, where they stay, how much they earn, etc.
  - Use this information as input attributes to learn a classifier model

Classification : Example Application

Fraud Detection

- Goal: Predict fraudulent cases in credit card transactions
- Approach:
  - Use credit card transactions and the information on its account-holder as attributes
  - When does a customer buy, what does he buy, how often he pays on time, etc.
  - Label past transactions as fraud or fair transactions. This forms the class attribute.
  - Learn a model for the class of the transactions
  - Use this model to detect fraud by observing credit card transactions on an account

Classification : Example Application

Customer Attrition/Churn

- Goal: To predict whether a customer is likely to be lost to a competitor.

- Approach:
  - Use detailed record of transactions with each of the past and present customers, to find attributes
  - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
  - Label the customers as loyal or disloyal
  - Find a model for loyalty

Classification : Example Application

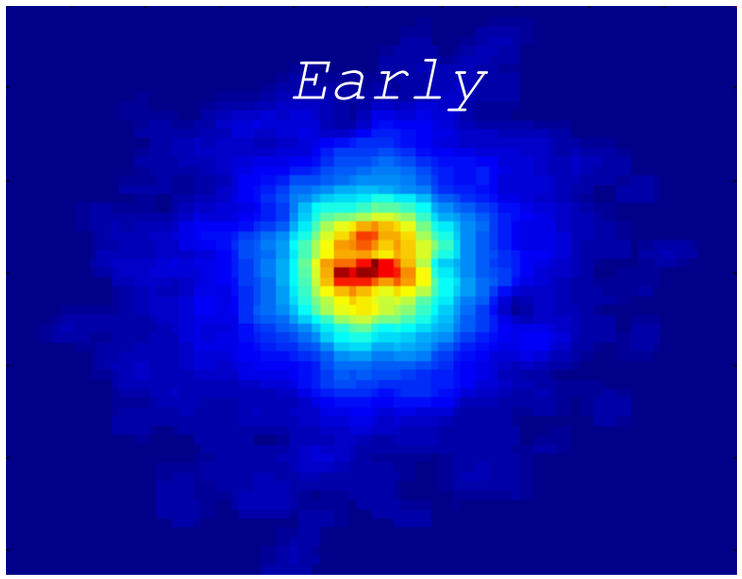
Sky Survey Cataloging

- Goal: To predict class [star or galaxy] of sky objects, especially visually faint ones, based on the telescopic survey images [from Palomar Observatory]
  - 3000 images with 23,040 x 23,040 pixels per image

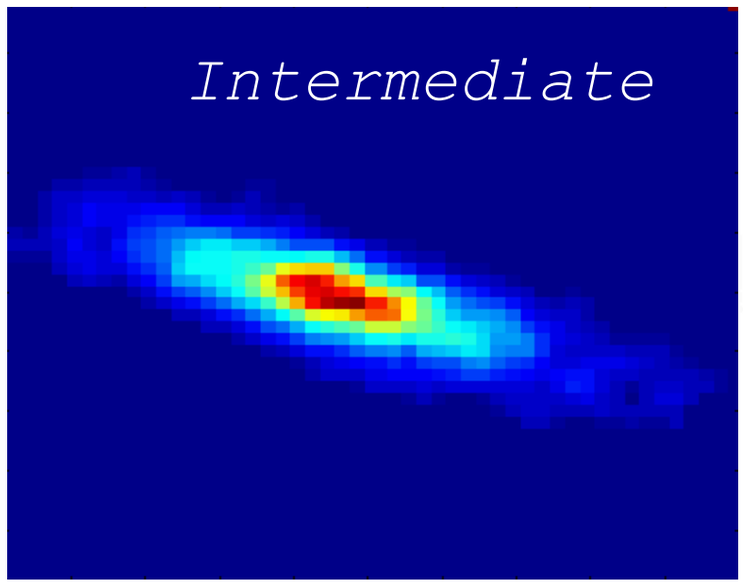
- Approach:
  - Segment the image
  - Measure image attributes (features) - 40 of them per object
  - Model the class based on these features
  - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

Classifying Galaxies

Courtesy: http://aps.umn.edu



Class :
- Stages of formation



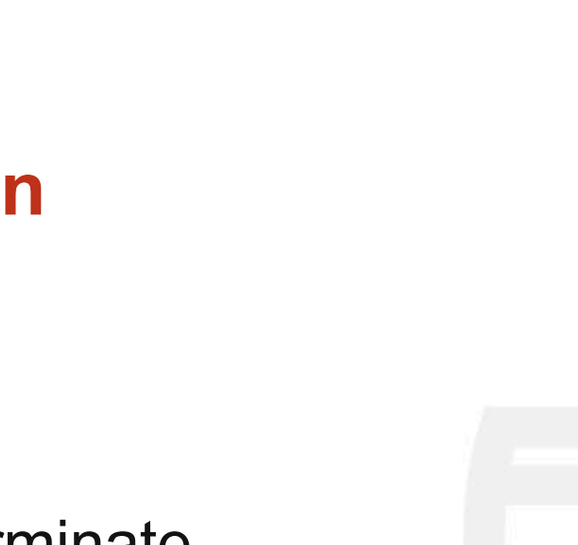
Attributes :
- Image features
- Characteristics of light waves received, etc.

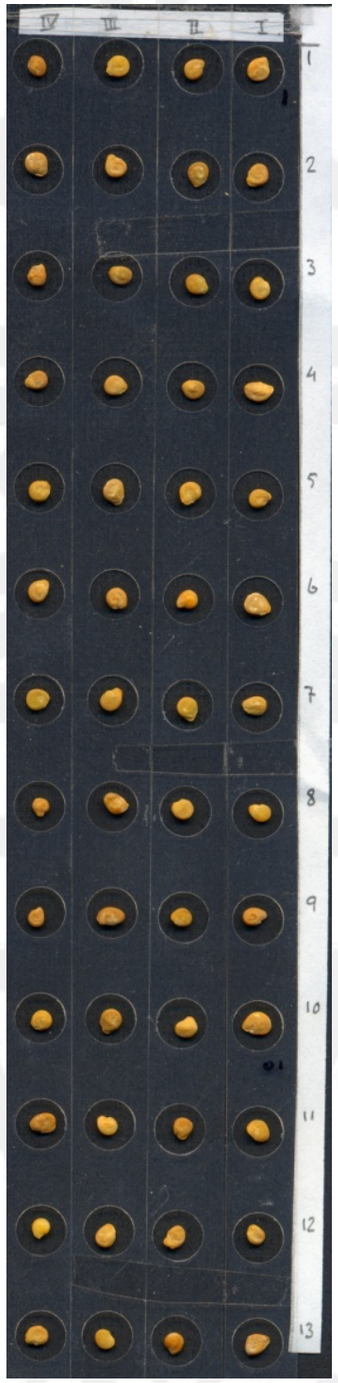
Data size :
- 72 million stars, 20 million galaxies
- Object catalog : 9 GB
- Image database : 150 GB

Classification : Example Application

Classification of tomato seeds

- Goal : to predict whether tomato seeds germinate
- Approach :
  - “scan” the seeds
  - extract features
  - build a classifier
  - use the classifier to blow away infertile seeds



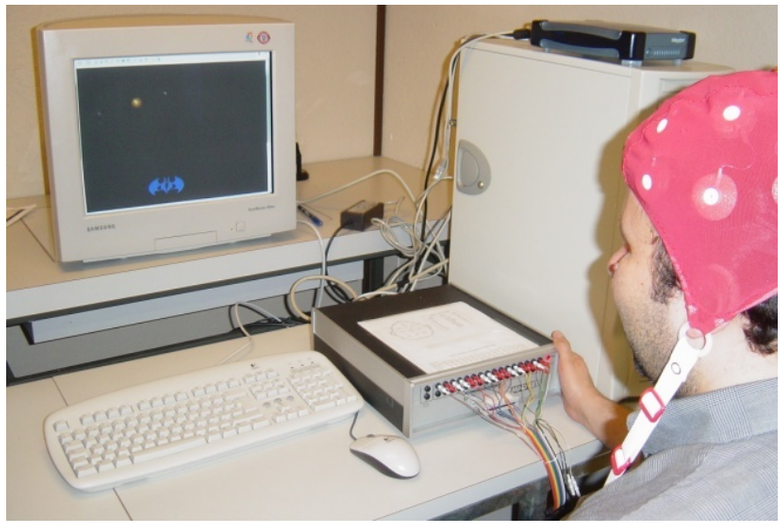


MACHINE2LEARN

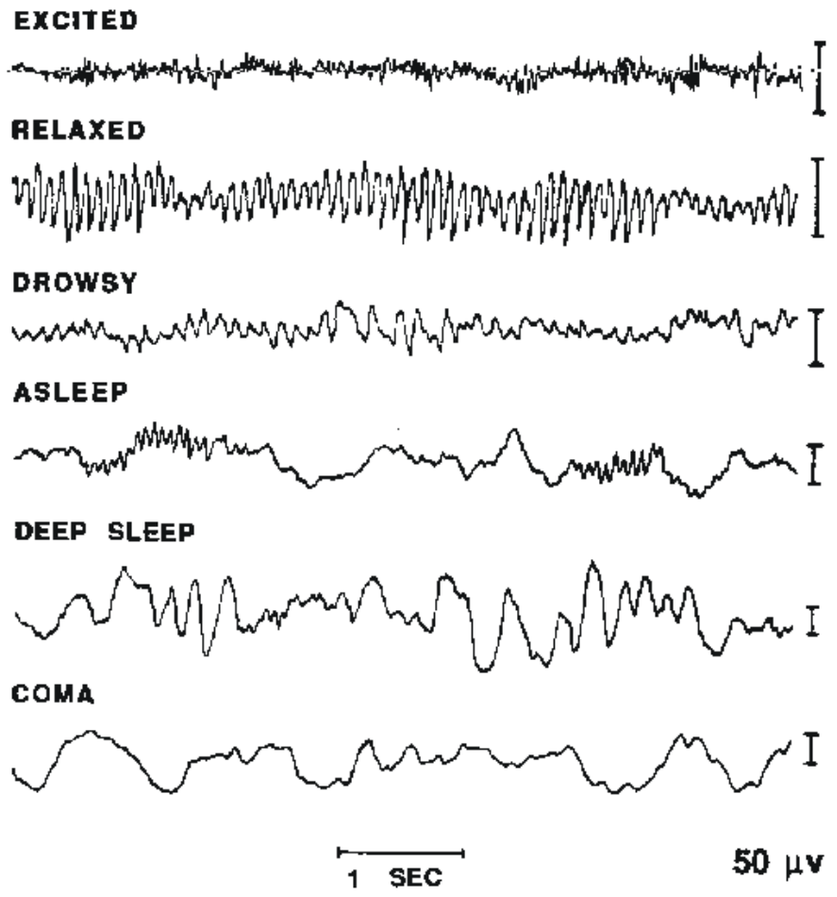
Classification : Example Application

Brain-computer interfacing

- Goal : read a person's mind
- Approach :
  - measure EEG signals
  - classify them



EEG
ElectroEncephaloGram

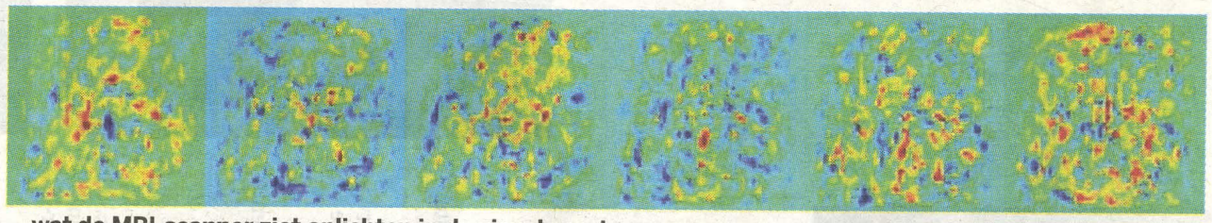


In *de Volkskrant*

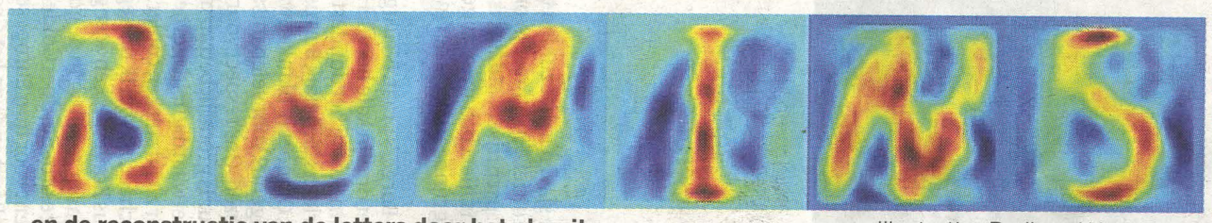
Welke letters las u zonet? De MRI-scanner weet het

Een team in Nijmegen is er voor het eerst in geslaagd om bij iemand die een woord ziet, te achterhalen welke letters hij heeft gelezen, gegeven welke stukjes hersenschors er oplichten. De crux zit hem in een wiskundig model.

Een selectie van de oorspronkelijke handgeschreven letters ...



... wat de MRI-scanner ziet oplichten in de visuele cortex ...



... en de reconstructie van de letters door het algoritme.

Illustraties Radboud Universiteit

Van onze verslaggever
Bard van de Weijer

AMSTERDAM Derek Ogilvie zal zijn vingers erbij aflikken: onderzoekers van de Radboud Universiteit hebben een methode ontwikkeld waarmee uit iemands hersenactiviteit afgeleid kan worden welke letters hij ziet.

Het aflezen gebeurt met behulp van een MRI-scanner die kijkt naar de visuele cortex, het hersengebied waar beeldinformatie wordt verwerkt. Daartoe worden kubusjes brein van 2×2×2 millimeter in de visuele cortex geanalyseerd. Deze kubusjes, zogenoemde voxels, lichten op als ze worden geactiveerd door visuele informatie.

Als een proefpersoon de letter G ziet, lichten andere voxels op dan bij de letter T. De MRI-scanner meet dus voor elke letter een ander activatiepatroon. Een algoritme kan uit deze patronen de letters reconstrueren die de proefpersoon in de scanner ziet. Het gaat om handgeschreven letters, in allerlei variaties, die alle door het systeem worden herkend.

'Het is geen gedachten lezen', zegt cognitief neurowetenschapper Marcel van Gerven van het Donders Instituut van de Radboud Universiteit. 'We reconstrueren perceptie, dus wat iemand ziet, niet wat hij denkt.' Een belangrijk verschil, omdat gedachten over 'alles' kunnen gaan en het analyseren van visuele informatie – letters in dit geval – het aantal mogelijkheden beperkt. Het algoritme is getraind op het herkennen van letters. Als een proefpersoon een afbeelding van een vliegtuig wordt voorgehouden, zal dat niet herkend worden.

We vermoeden dat het brein ook op deze manier werkt

Tot zover is er volgens Van Gerven nog niet veel nieuws onder de zon. 'We zijn niet de eersten die met MRI-scans beeldpatronen in de visuele cortex kunnen herkennen. Het is wel voor het eerst gelukt om met een wiskundig model het oorspronkelijke beeld met hoge kwaliteit te reconstrueren.'

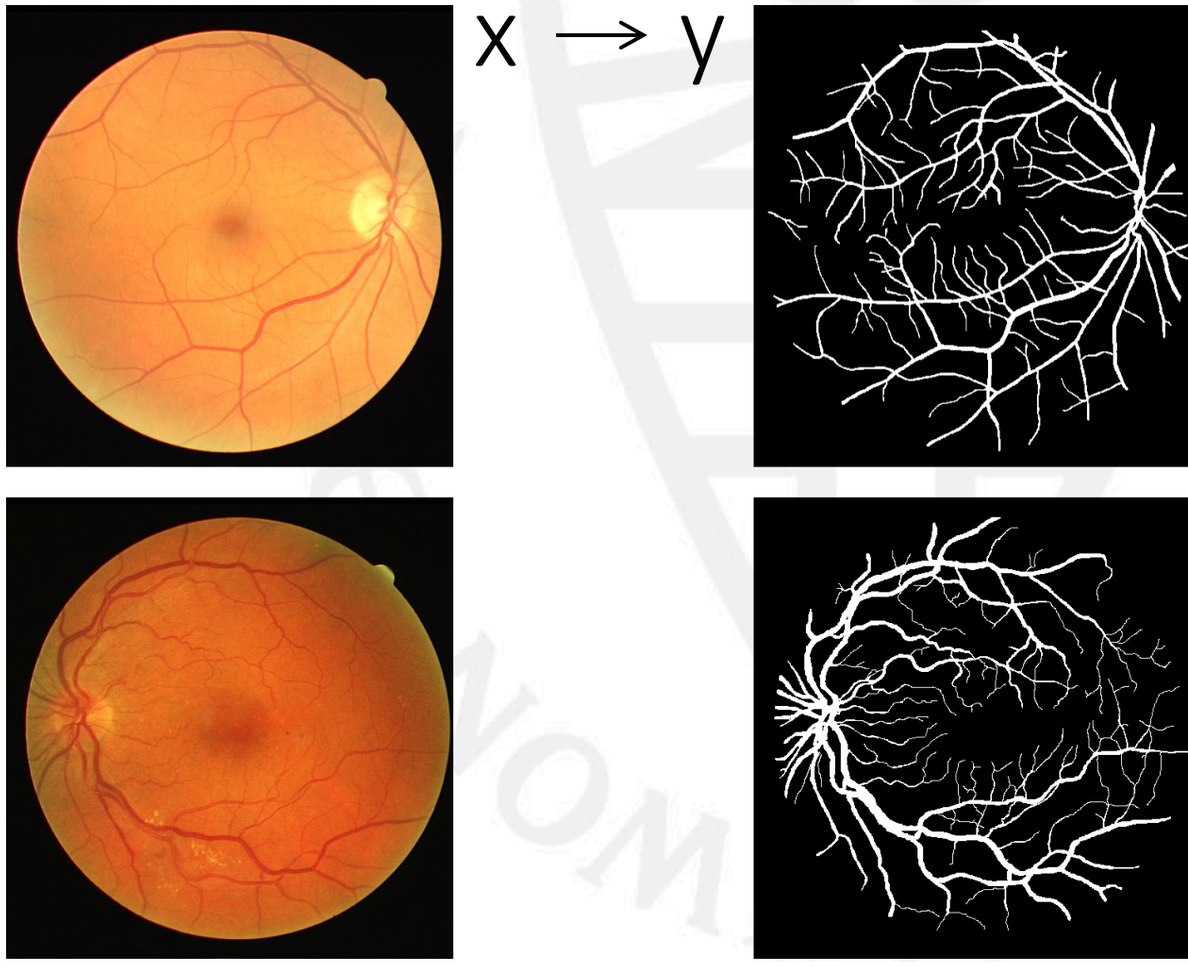
Dit gebeurt door twee bronnen te combineren: de onderzoekers kijken in een gebiedje van duizend voxels hoe deze reageren op externe stimuli. Deze gegevens – de wat gruizige afbeeldingen hierboven – worden gecombineerd met voorkennis over de eigenschappen van letters. Door de data van de MRI-scan te vergelijken met deze 'kennis' kan worden herleid welke letters de proefpersoon waarneemt.

'We vermoeden dat het brein ook op deze manier werkt', zegt Van Gerven. 'Je kunt al die lijntjes en bochtjes niet begrijpen voor je hebt leren lezen. Pas als sprake is van een zekere context kun je letters onderscheiden.' De onderzoekers hopen met hun onderzoek meer te weten te komen over de werking van het brein. Hoewel het bedenken van praktische toepassingen niet het eerste doel is, ziet de onderzoeker wel mogelijkheden. 'Er is een relatie tussen perceptie en verbeelding. Je zou wellicht een reconstructie kunnen maken van een beeld dat iemand zich in gedachten voorstelt. Denk aan een getuige die zich de verdachte inbeeldt en dat je dat beeld dan kunt visualiseren. Maar dat is echt de verre toekomst.'

Radboud University Nijmegen

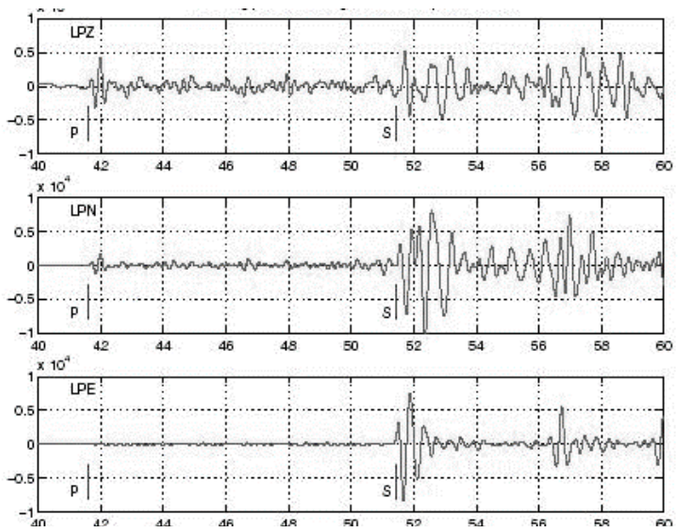
Classification : Example Application

- Image segmentation
  - E.g. retina images

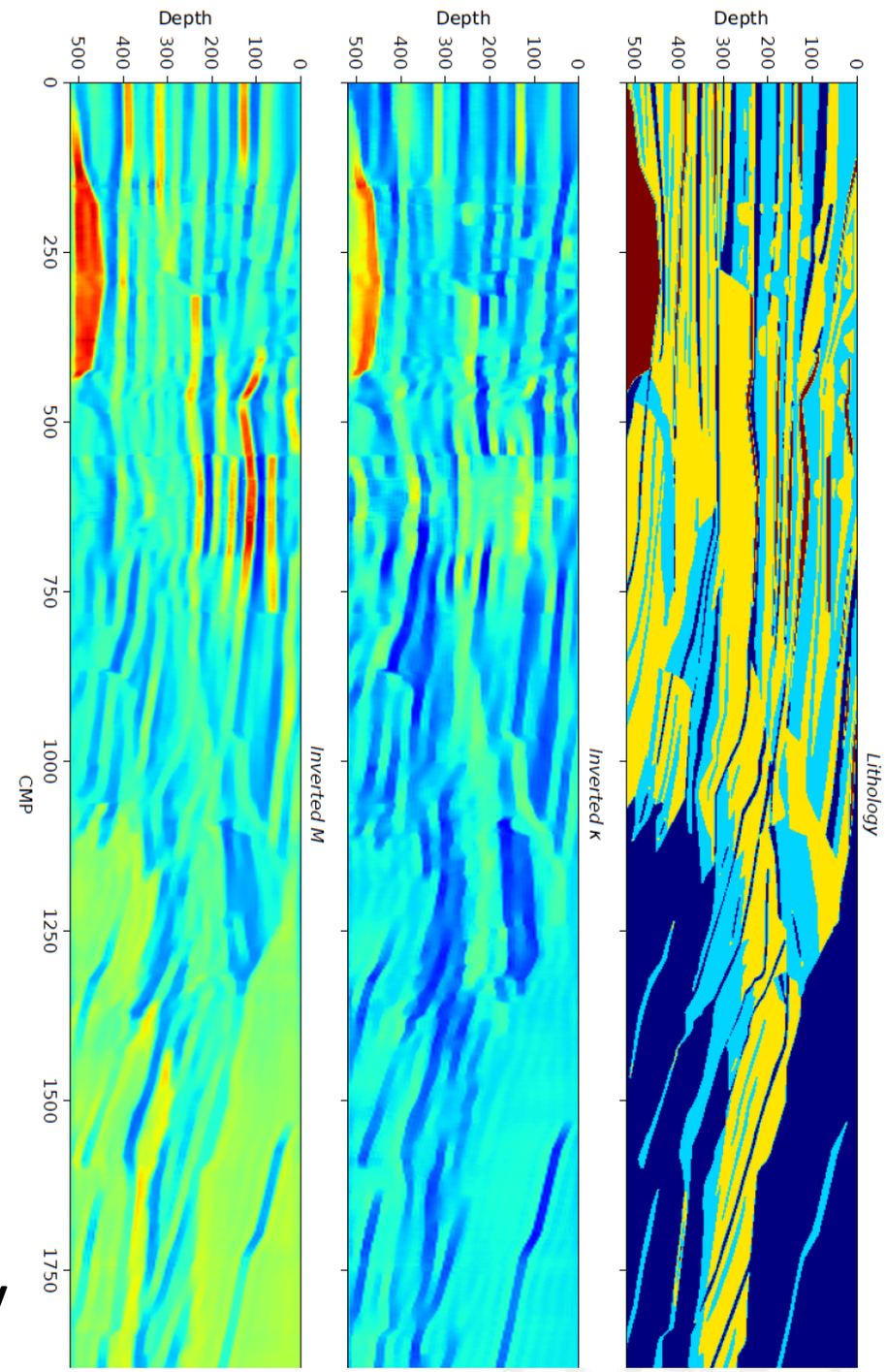


Classification : Example Application

- Seismic inversion
  and litho-type classification

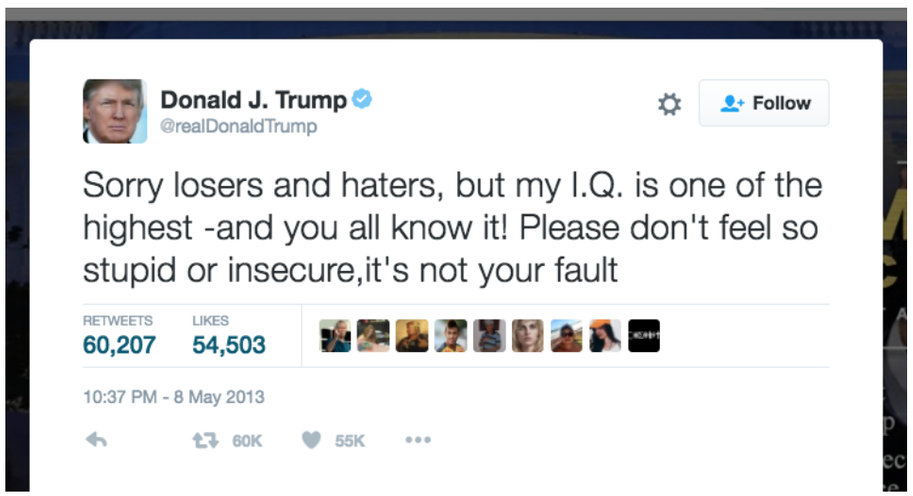


x ⟶ y



Classification : Example Application

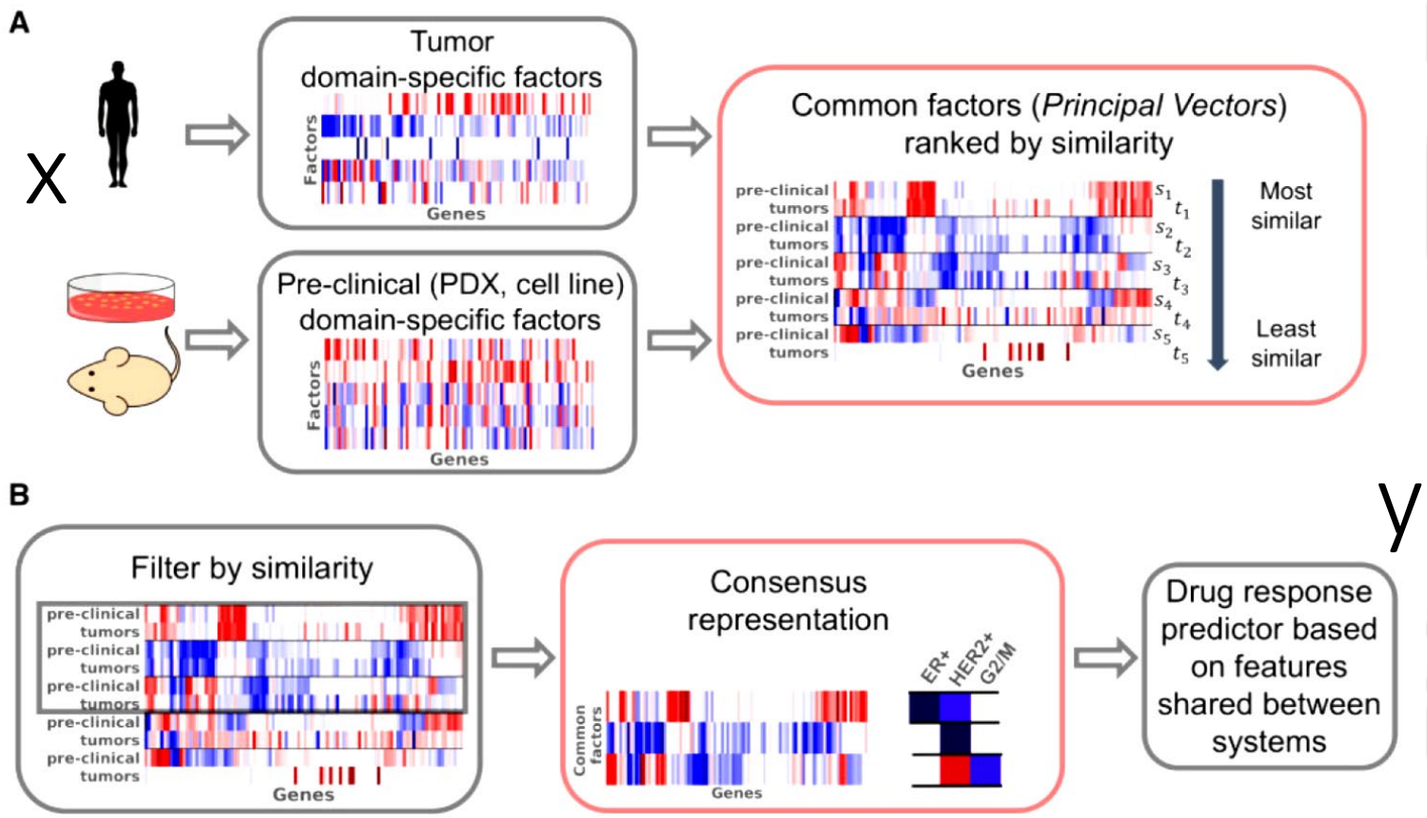
- Text sentiment classification



$x \rightarrow y$

Classification : Example Application

- Drug-response prediction



Clustering…

- Given a set of data points, each having the same set of attributes and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another
  - Data points in separate clusters are less similar to one another

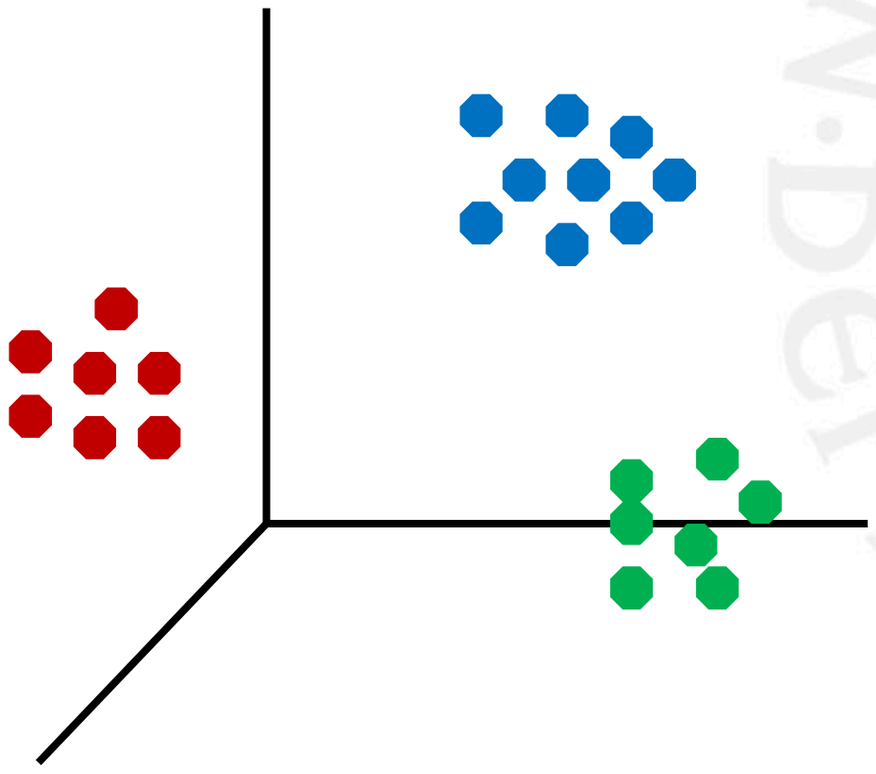
- Similarity measures :
  - Euclidean distance if attributes are continuous
  - Other problem-specific measures

Illustrating Clustering

- Euclidean distance-based clustering in 3D space

Intracluster distances are minimized

Intercluster distances are maximized



Clustering : Example Application

Market segmentation

- Goal : subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix

- Approach :
  - Collect different attributes of customers based on their geographical and lifestyle related information.
  - Find clusters of similar customers.
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering : Example Application

Document clustering

- Goal : To find groups of documents that are similar to each other based on the important terms appearing in them

- Approach : To identify frequently occurring terms in each document
  Form a similarity measure based on the frequencies of different terms
  Use it to cluster

- Gain : Information retrieval can utilize the clusters to relate a new document or search term to clustered documents

Illustrating Document Clustering

- Data points : 3204 articles of Los Angeles Times
- Similarity measure : How many words are common in these documents [after some word filtering]

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Clustering of S&P 500 Stock Data

- Observe daily stock movements
- Data points : time series of stock-{up/down}
- Similarity measure : Two points are more similar if the events described by them frequently happen together on the same day

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN,Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN,DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN,Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down,Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN,Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN,ADV-Micro-Device-DOWN,Andrew-Corp-DOWN,Computer-Assoc-DOWN,Circuit-City-DOWN,Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN,Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN,MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP,Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP,Schlumberger-UP	Oil-UP

Association Rule Discovery…

- Given a set of records each of which contain some number of items from a given collection: Produce dependency rules which will predict occurrence of an item based on occurrences of other items

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery : Example Application

Marketing and sales promotion

- Suppose the discovered rule is
  {Bagels, … } → {Potato Chips}

- Potato Chips as consequent : Can be used to determine what should be done to boost its sales

- Bagels in the antecedent : Can be used to see which products would be affected if the store discontinues selling bagels

- Bagels in antecedent *and* Potato chips in consequent : Can be used to see what products should be sold with Bagels to promote sale of Potato chips…

Association Rule Discovery : Example Application

Supermarket shelf management

- Goal : To identify items that are bought together by sufficiently many customers
- Approach : Process the point-of-sale data collected with barcode scanners to find dependencies among items
- A classic rule --
  - If a customer buys diaper and milk, then he is very likely to buy beer [on Thursday]

Regression…

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency

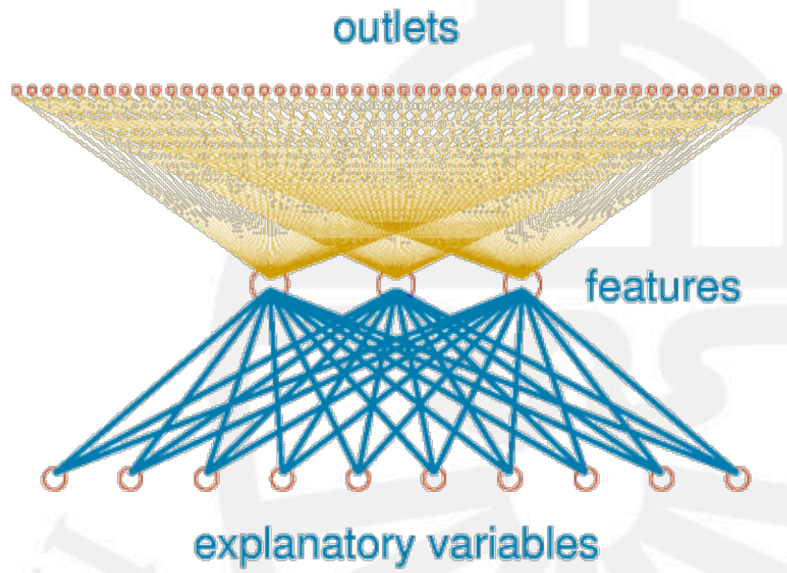
- Greatly studied in statistics, neural networks, etc.

- Examples :

  - Predicting sales amounts of new product based on advertising expenditure
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices

  - Predicting the next word based on foregoing words [or is this classification?]

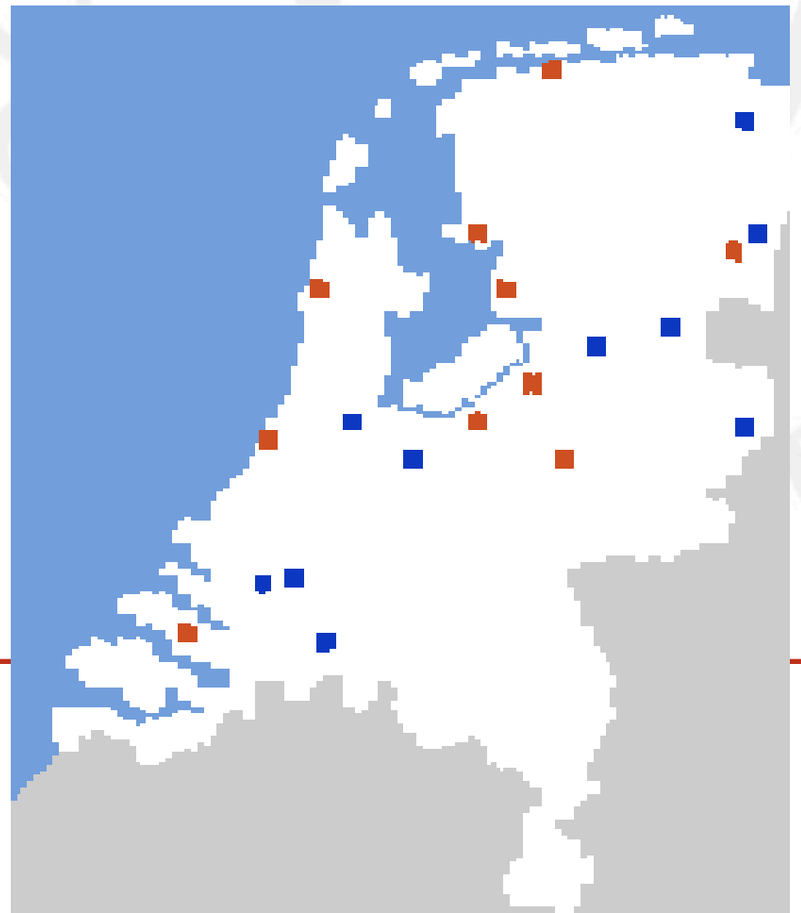
Regression : Example Application



Predicting newspaper sales

- Goal : optimize single-copy sales of De Telegraaf
- Approach :
  - learn from past sales
  - let outlets learn from each other

- better weather, more sales
- worse weather, more sales



Regression : Example Application

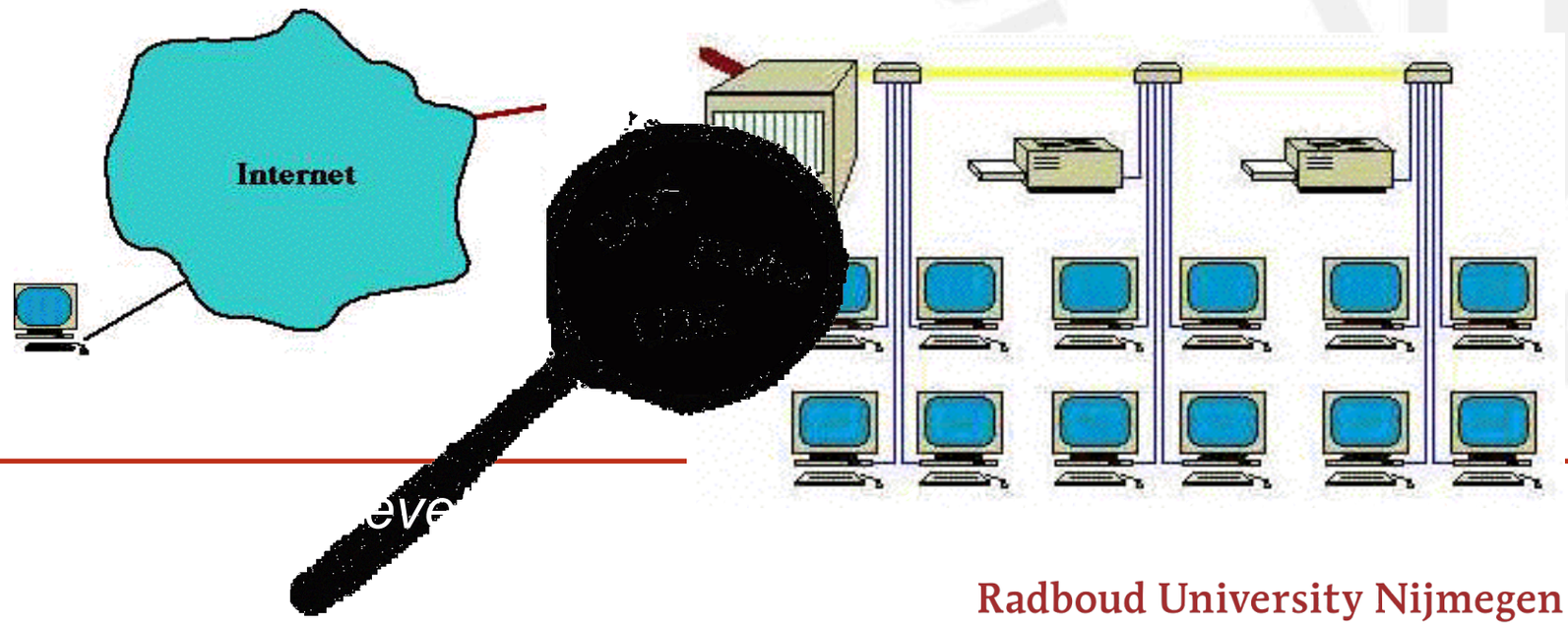
Reconstruction of faded Van Gogh drawings

x

y

Deviation / Anomaly Detection…

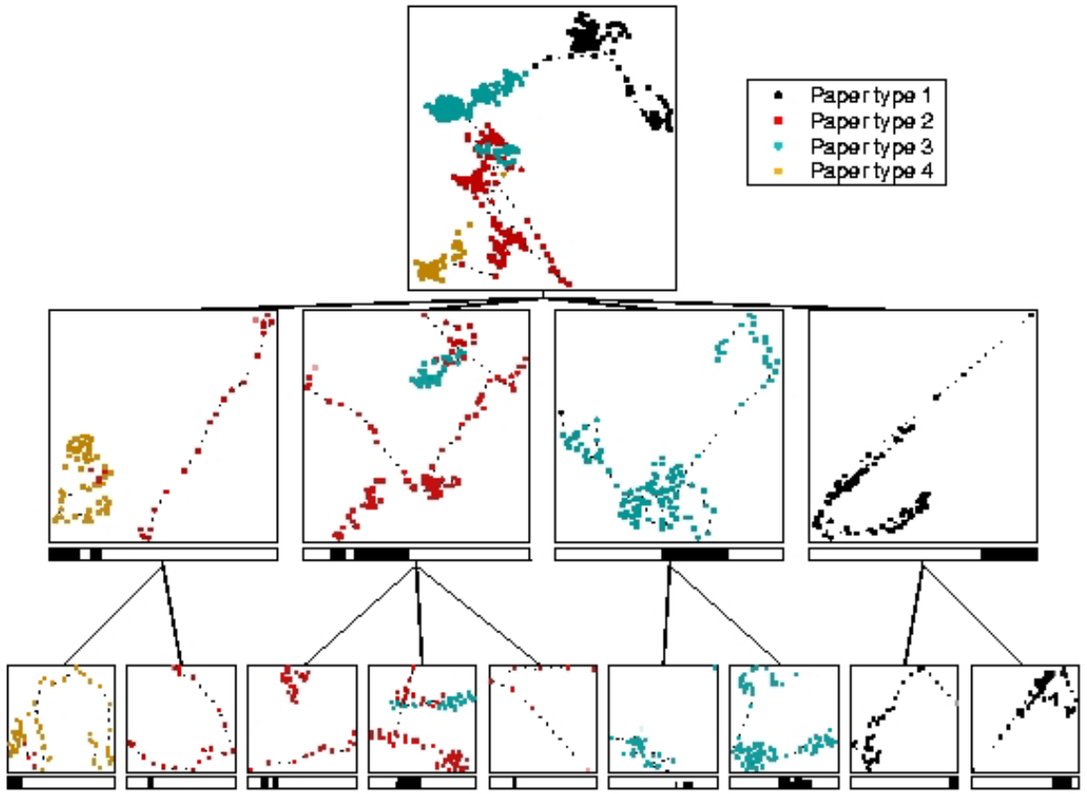
- Detect significant deviations from normal behavior
- Applications :
  - Credit card fraud detection
  - Network intrusion detection



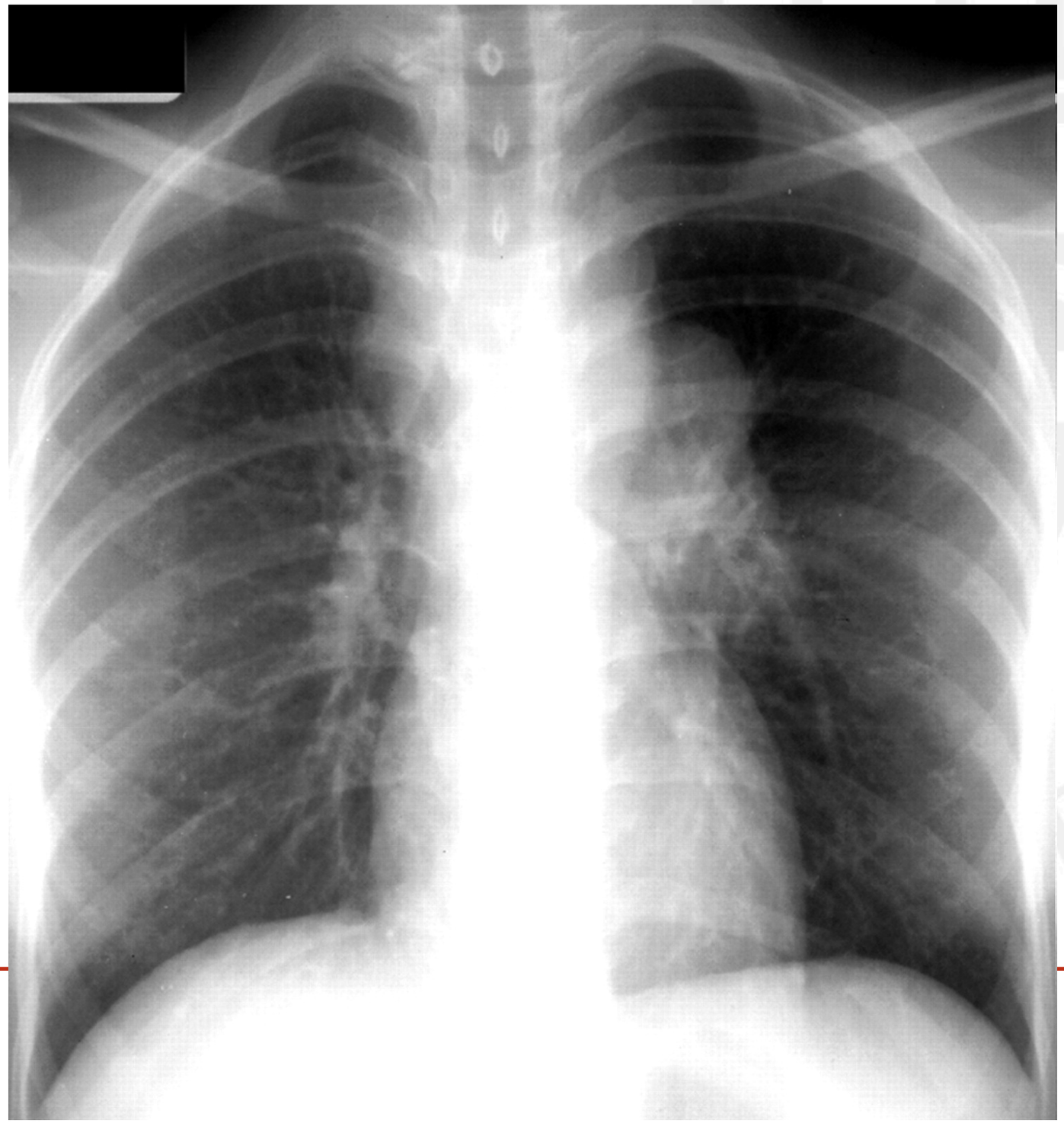
Deviation / Anomaly Detection : Example Application

Monitoring paper mills

- Goal : alert operators when the paper mill starts behaving “weirdly”
- Approach : visualize the dynamics by cleverly projecting the measurements of hundreds of sensors



Deviation / Anomaly Detection : Example Application



- Anomaly detection in chest X-rays

Data Mining?

- Dividing the customers of a company according to their gender
- Predicting the profitability of customers
- Computing the total sales of a company
- Sorting a student database based on student identification numbers
- Predicting the outcomes of tossing a fair pair of dice
- Predicting the outcomes of tossing a possibly unfair pair of dice after having seen some amount of tosses

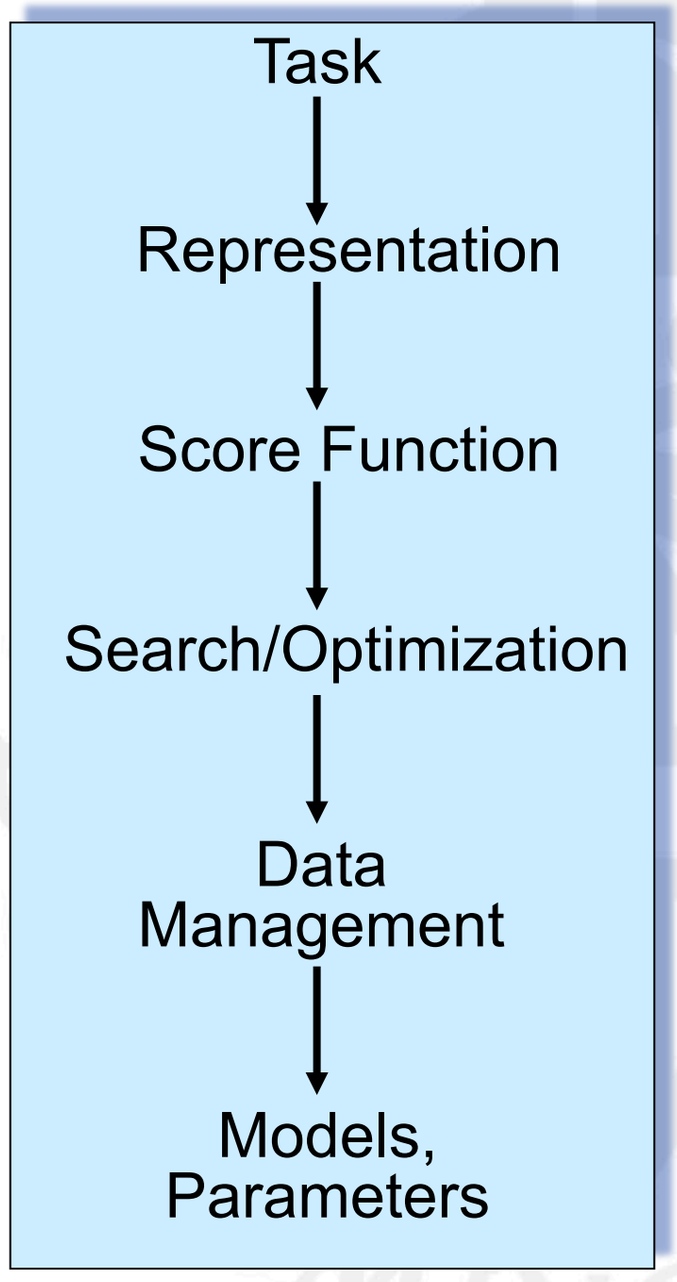
Data Mining?

- Predicting the future stock price of a company using historical records
- Monitoring the heart rate of a patient for abnormalities
  given observations of both abnormal and normal behavior
- Monitoring the heart rate of a patient for abnormalities
  given observations of only normal behavior
- Monitoring seismic waves for earthquake activities
- Extracting the frequencies of a sound wave

Components of Data Mining Algorithms

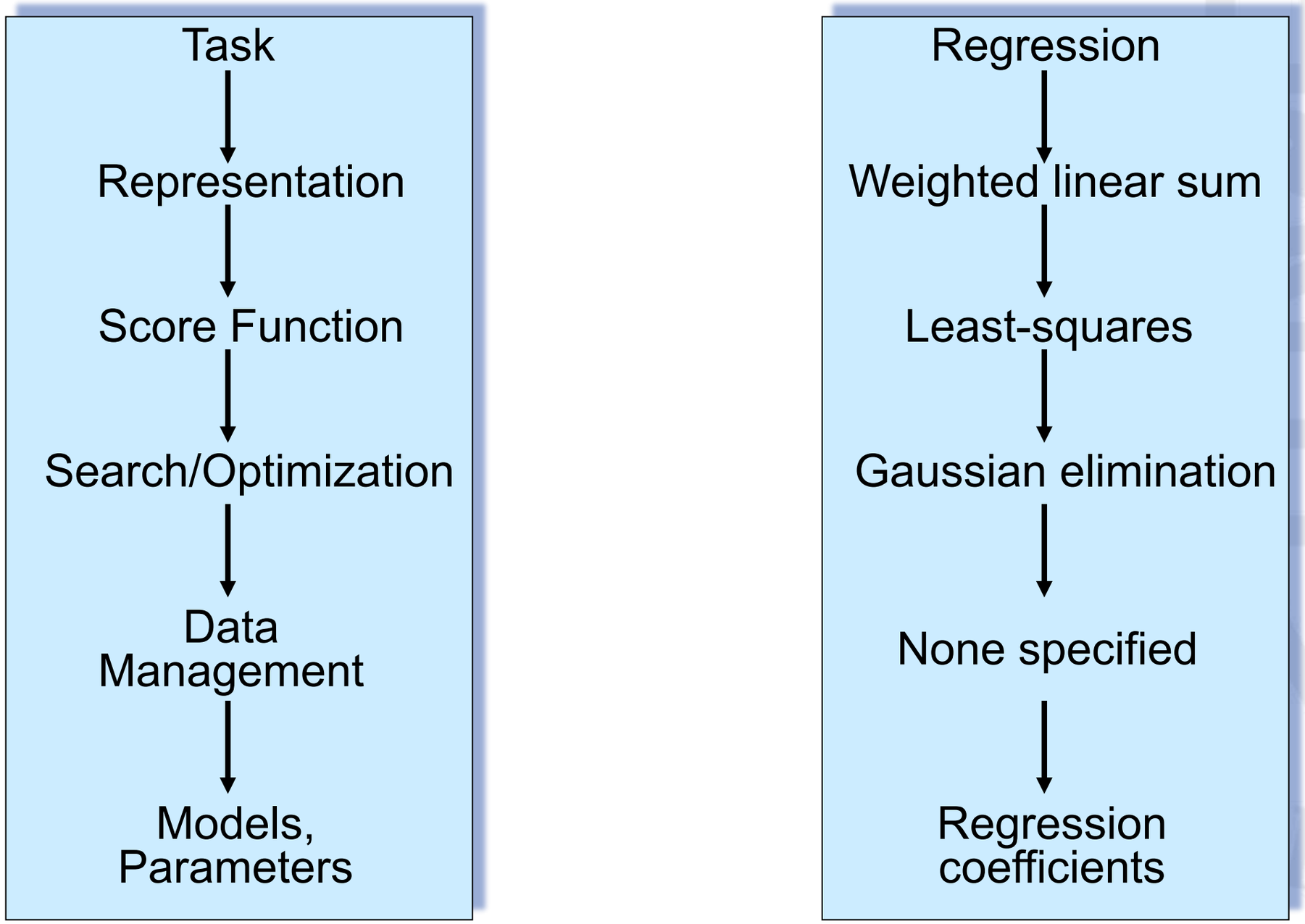
- Representation:
  - Determining the nature and structure of the representation to be used
- Score function:
  - quantifying and comparing how well different representations fit the data
- Search/Optimization method:
  - Choosing an algorithmic process to optimize the score function
- Data Management:
  - Deciding what principles of data management are required to implement the algorithms efficiently

What's in a Data Mining Algorithm?

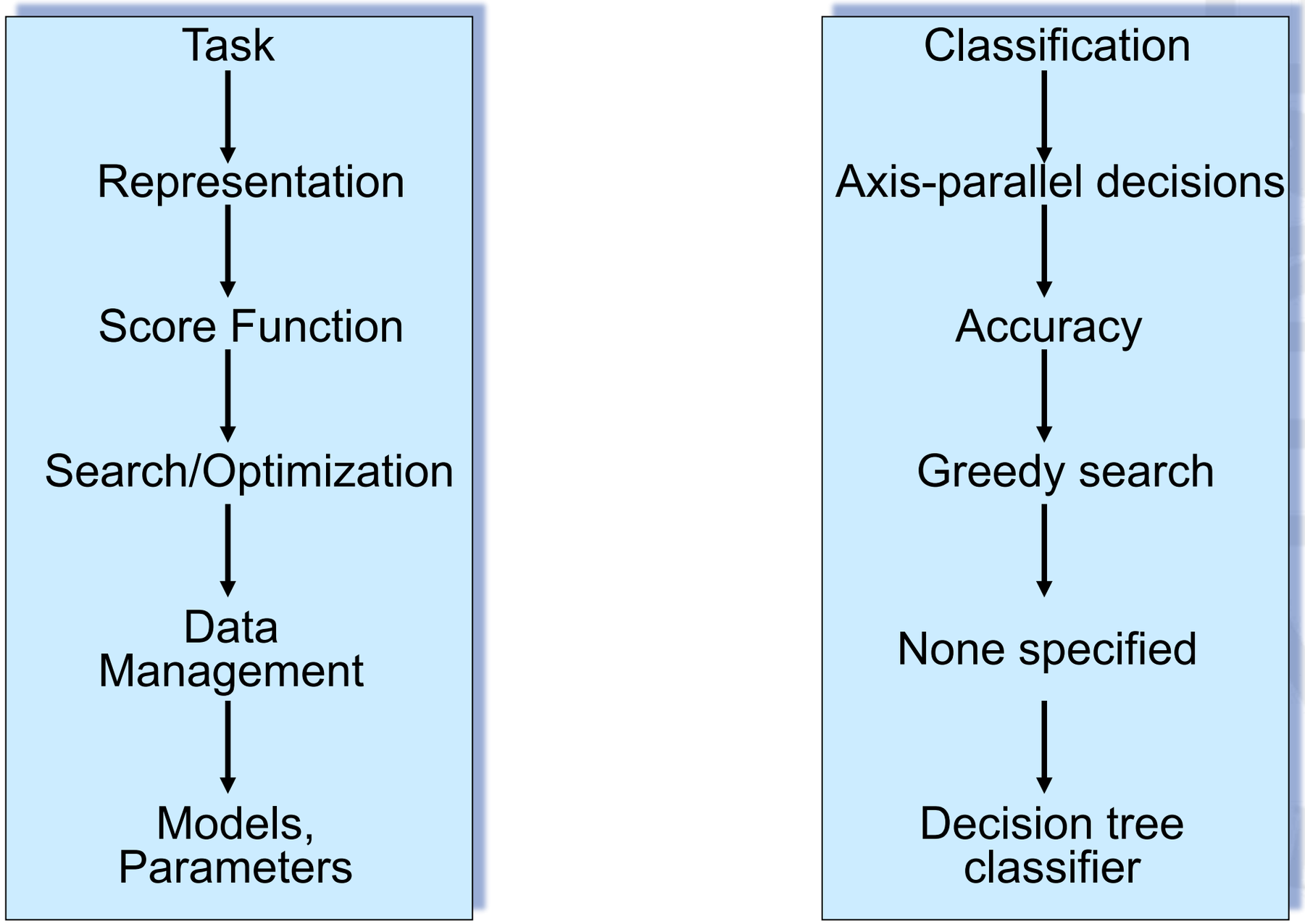


Smyth 2003

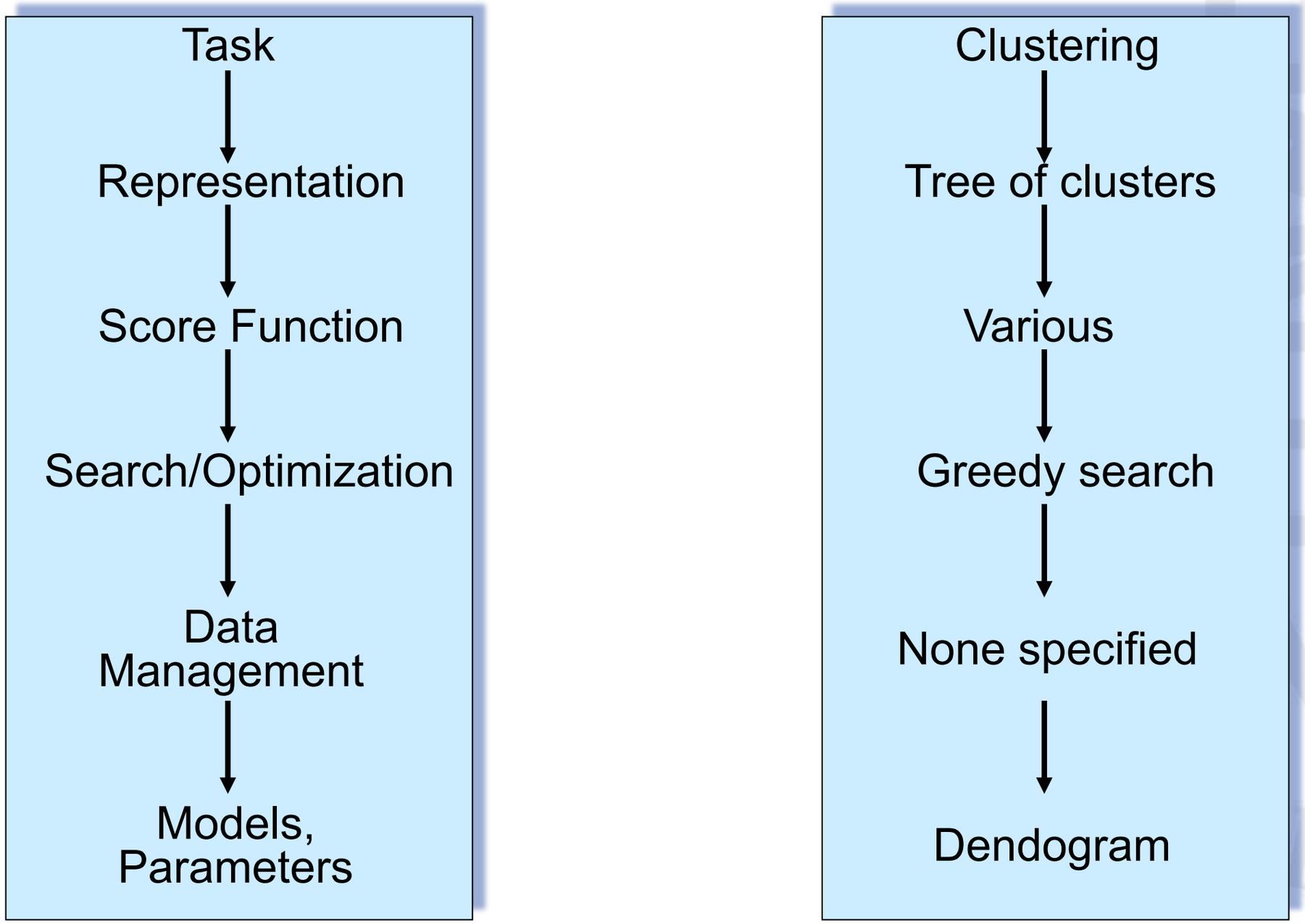
Multivariate Linear Regression



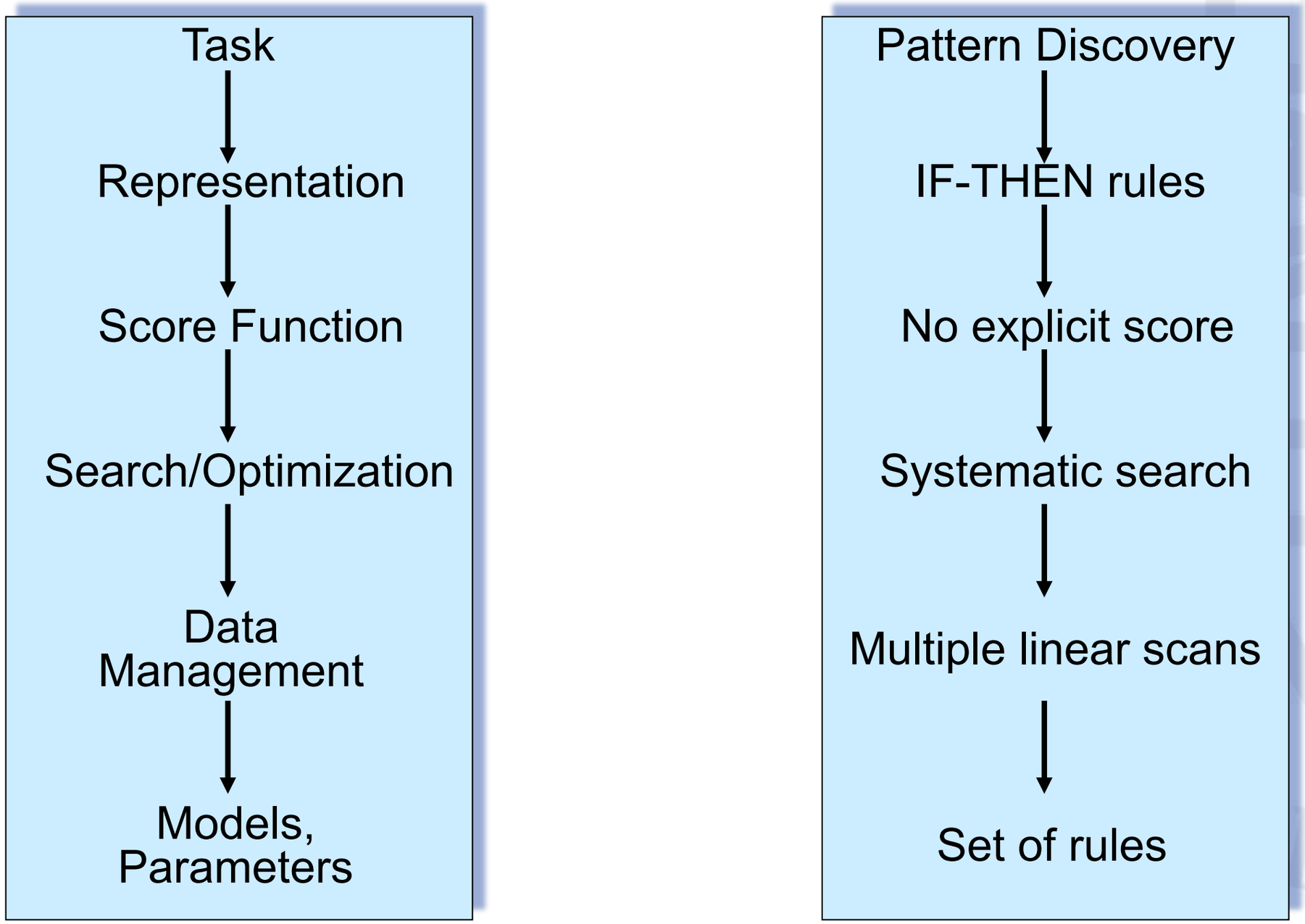
Decision trees (CART, ID3, …)



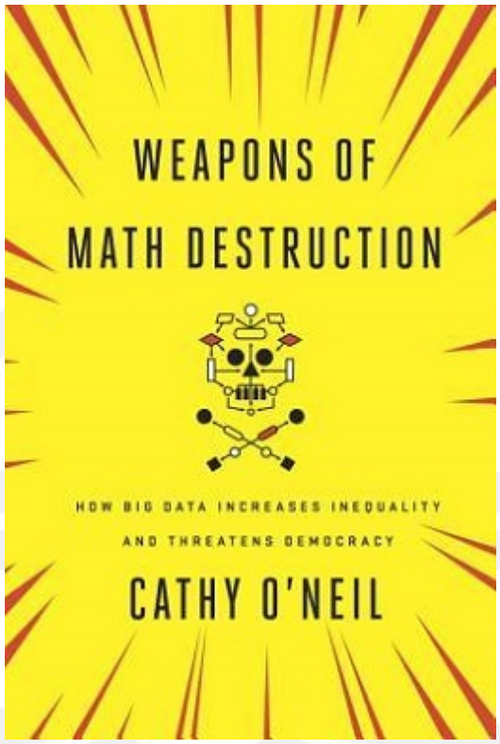
Hierarchical Clustering



Association Rules

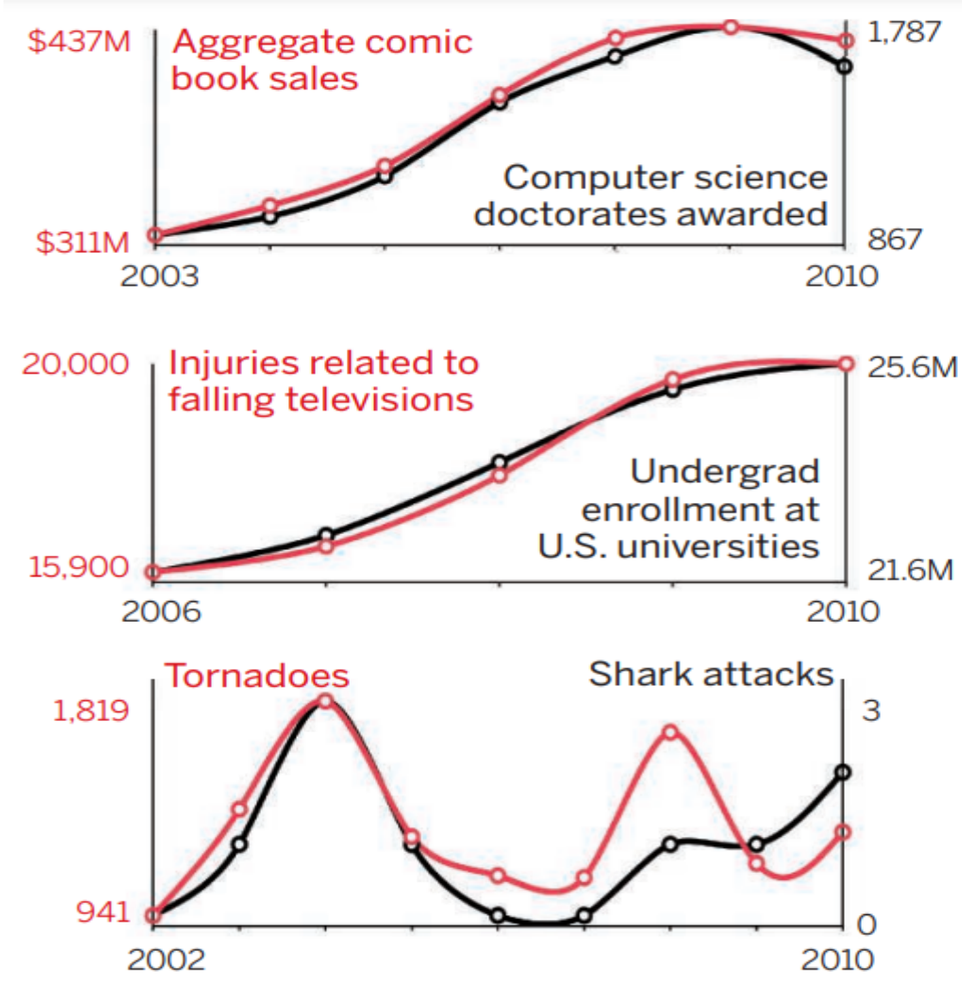


Data Mining : the Downside



- Hype [at least it was?]
- One of the “weapons of math destruction”
  - Tools can reinforce discrimination and unfairness
- Data dredging, snooping and fishing
  - Finding spurious structure in data that is not real
- Historically, ‘data mining’ was
  a derogatory term in the statistics community
  - The Super Bowl fallacy
  - Bangladesh butter prices and the US stock market
- The challenges of being interdisciplinary
  - computer science, statistics, domain discipline

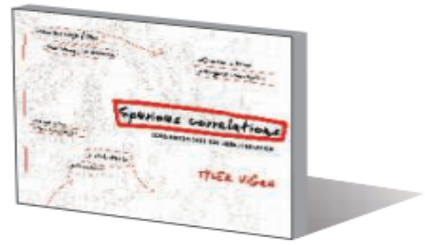
Spurious Correlations



Correlations are critical in scientific analysis, but given enough data, it is possible to find things that correlate, even when they shouldn't. More examples of how not to use statistics can be found on the author's website, tylervigen.com.

STATISTICS

Spurious Correlations
Tyler Vigen
Hachette Books, 2015, 207 pp.

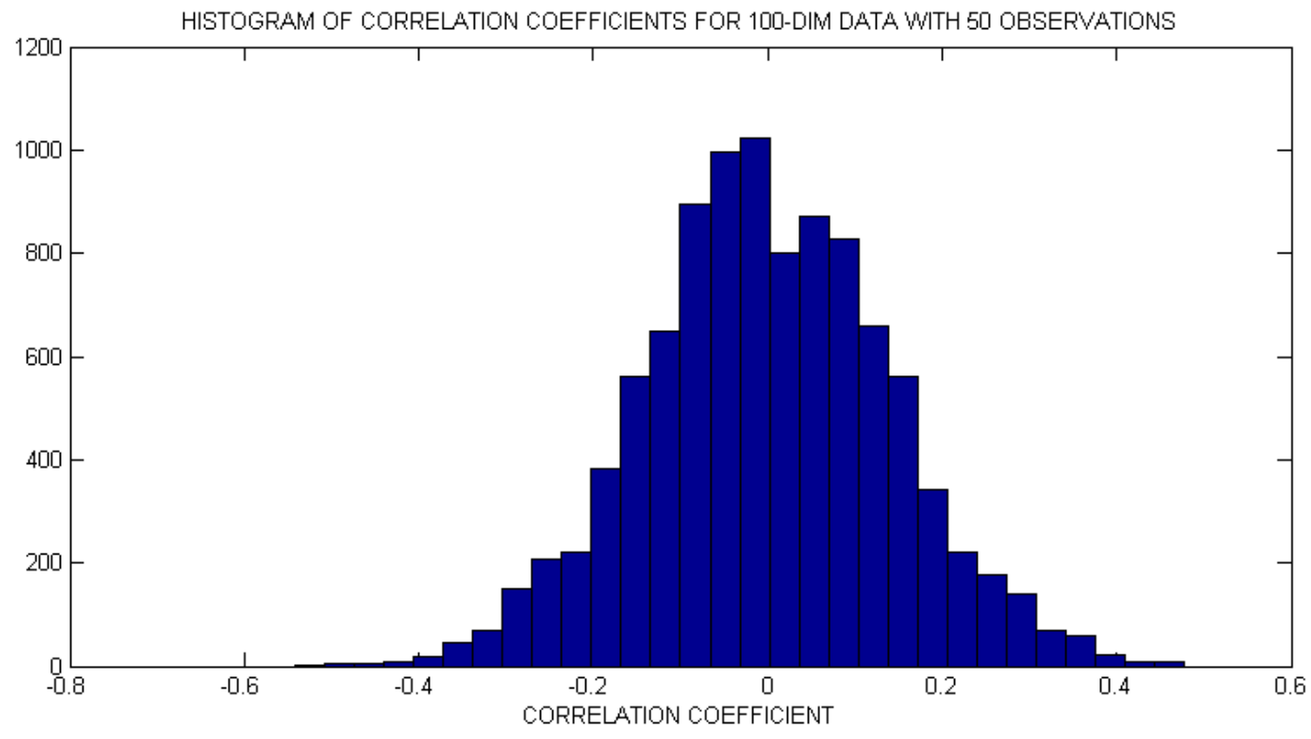


The number of civil engineering doctorates awarded in the United States between 2000 and 2009 was strongly correlated (95.9%) with mozzarella cheese consumption during the same period. Does that mean aspiring engineers should start stockpiling this delicious dairy staple? Of course not—the similarity in variance is purely a coincidence, identified by a technique known as "data dredging," in which one data set is blindly compared to hundreds of others until a correlation is identified. Presented as a series of graphs prepared from real data sets, *Spurious Correlations* serves as a hilarious reminder that correlation most certainly does not equal causation.

10.1126/science.aac5518

Rough Eplanation of “Data Fishing”

- Data set with
  - 50 data vectors
  - 100 variables
  - Even if data are entirely random [no dependence] there is a very high probability some variables will appear dependent just by chance.



Possible Pitfalls

Let the data speak…

The data may have quite a lot to say…..
but it may just be noise!

Smyth 2003

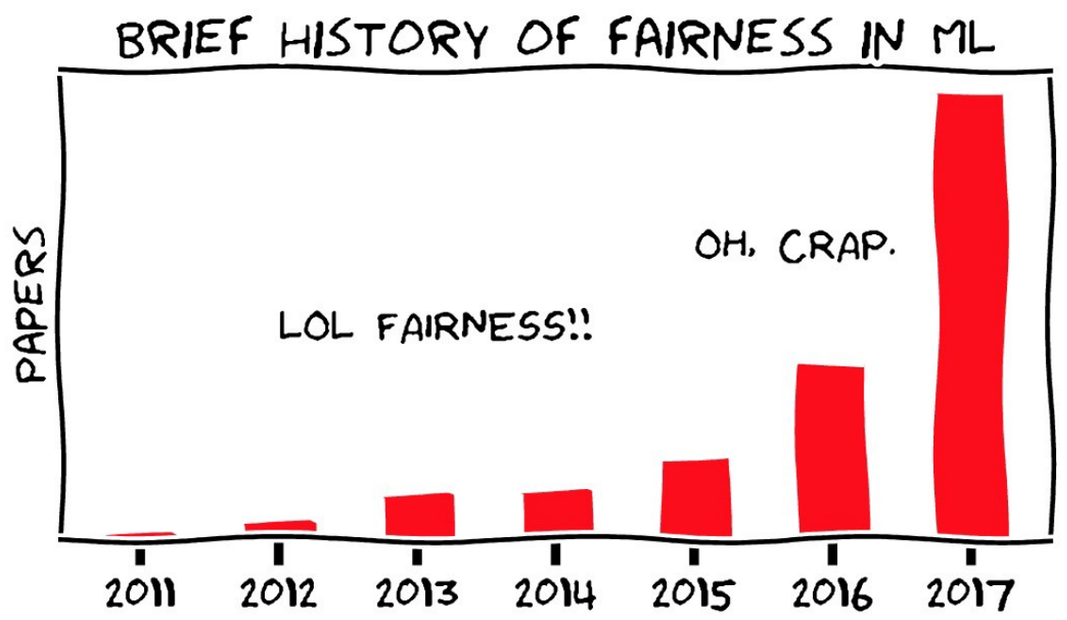
Dilbert…


OUR CONSULTANT HAS BEEN MINING DATA ALL DAY.
THE RESULTS ARE QUITE SHOCKING.
www.dilbert.com scottadams@aol.com
ACCORDING TO THE DATA, SALES ARE ALWAYS HIGHEST WHEN I DO THIS...
© 1999 United Feature Syndicate, Inc.


WE HAVE A GIGANTIC DATABASE FULL OF CUSTOMER BEHAVIOR INFORMATION.
www.dilbert.com scottadams@aol.com
EXCELLENT. WE CAN USE NON-LINEAR MATH AND DATA MINING TECHNOLOGY TO OPTIMIZE OUR RETAIL CHANNELS!
© 2000 United Feature Syndicate, Inc.
IF THAT'S THE SAME THING AS SPAM, WE'RE HAVING A GOOD MEETING HERE.

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and heterogeneous data
- Data quality
- Data ownership and distribution
- **Privacy**
- **Accountability**
- **Fairness**



QS?

Radboud University Nijmegen