Data Mining: Introduction

Tom Heskes

Radboud University Nijmegen

Instructors

Tom Heskes
tomh@cs.ru.nl
lectures/general stuff

Tom Claassen
tomc@cs.ru.nl
lectures

Roel Bouman
roel.bouman@ru.nl
coordination practical sessions/grading homework

student assistants (practical sessions, grading):

Lizzy Grootjen, David Leeftink, Steffen Ricklin, Ron Hommelsheim, Nienke Wessel, Janneke Verbeek, Linda Schmeitz, Frederik Stoel, Tamara Verbeek

Course Outline

- 6 ects = 8 hours per week
- All, except for the exams, will be on-line
- No new lectures: video lectures from last year
- Q&A lectures on Tuesdays through Webex/Zoom
- Practical sessions on Wednesdays/Thursdays through Discord
- “Learning tasks”: optional reading material + exercises for self study
- Lots of info on Brightspace

Evaluation

- 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 × Exam1 + 0.35 × Exam2 + 0.3 × Project**
- Re-exam: 0.7 × Re-exam + 0.3 × Project *
- If all else fails: full repeat next year

*On campus if corona permits. If not, we may consider dropping the midterm. Also: 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

- Video lectures from last year (and the year before) available in Brightspace
- Q&A sessions on Tuesday at the scheduled time
- Ask questions through the discussion board before or chat during the Q&A session
- NB: Q&A sessions will not be recorded

Homework assignments

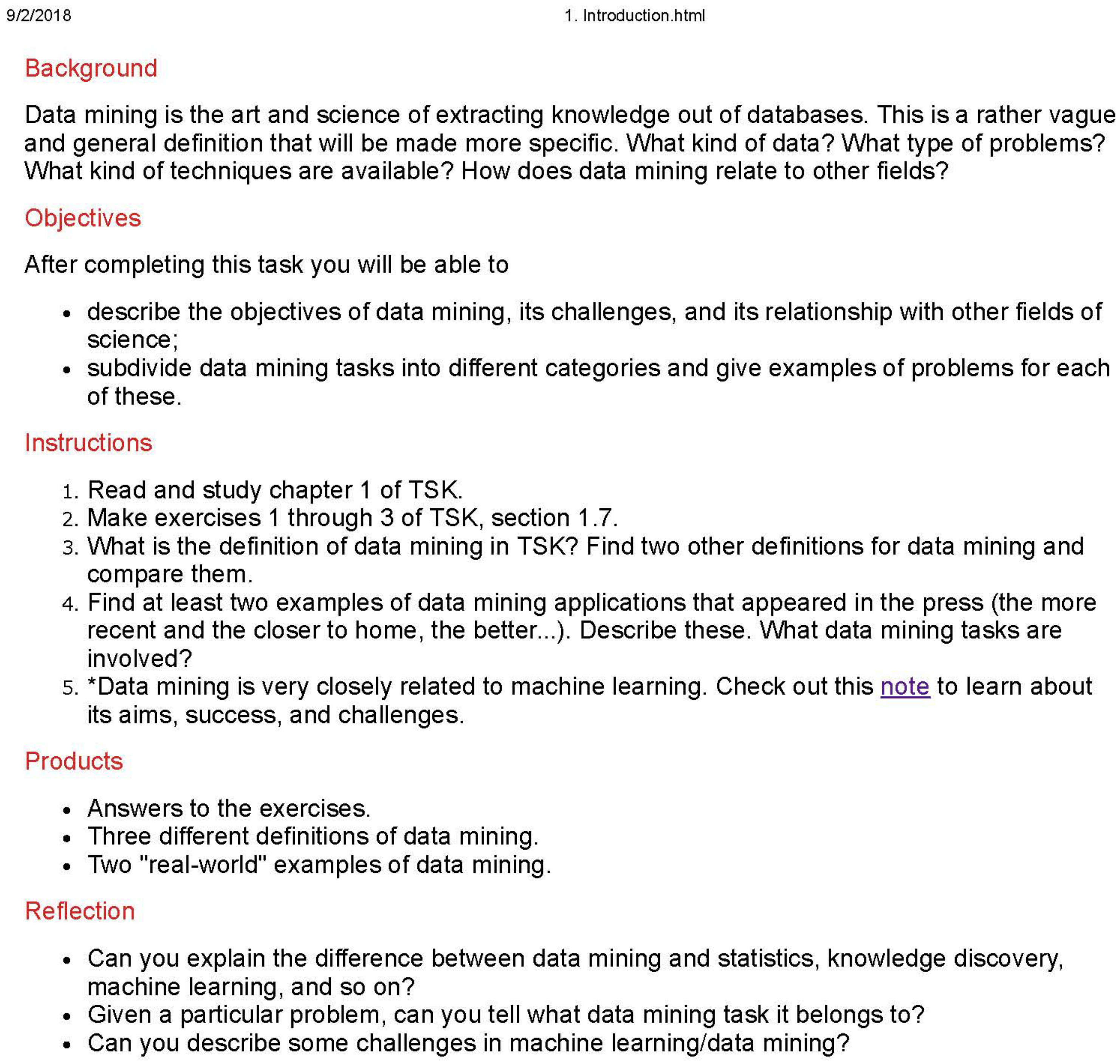
- Apply and understand data mining algorithms on (small) data sets
- Python:
  - open source, growing fast
  - many great data mining / machine learning packages
- Assistants at practical sessions are there to help you!
- Submit alone or as a pair; strictly follow guidelines on Brightspace!

Practical sessions

- Starting this week on Discord on Wednesday and Thursday
- The place to ask for help, with homework or other exercises
- Feel free to pick the slot that suits you best.
- This week: get used to Python
- In preparation: download and install the 3.8 version at https://www.anaconda.com/download/

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!



9/2/2018 1. Introduction.html

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?

Contents

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

- See

  https://www-users.cs.umn.edu/~kumar001/dmbook/firsted.php/

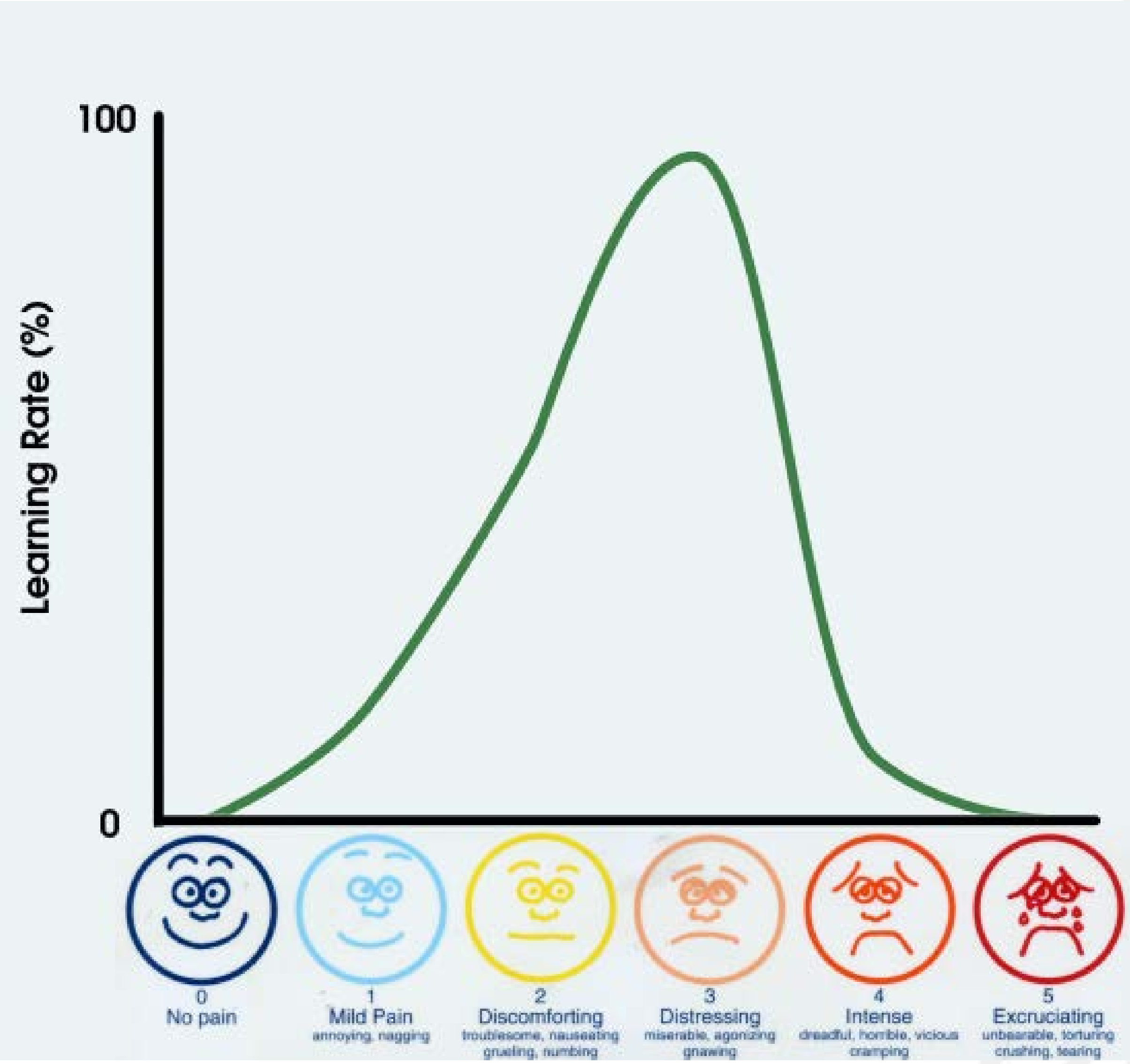
  where you can find slides, errata and some chapters

- Chapters 1, 2, 3, 4, 8, 10, 6, 5

Advice

- Keep track
- Go through the video lecture + accompying slides before the Q&A session
- 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

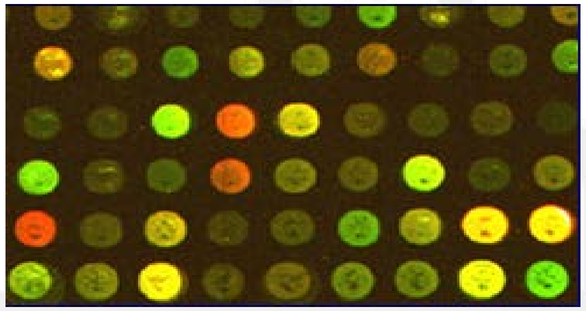
Data Mining: Introduction

- Motivation
- Examples
- Bit of history
- Challenges

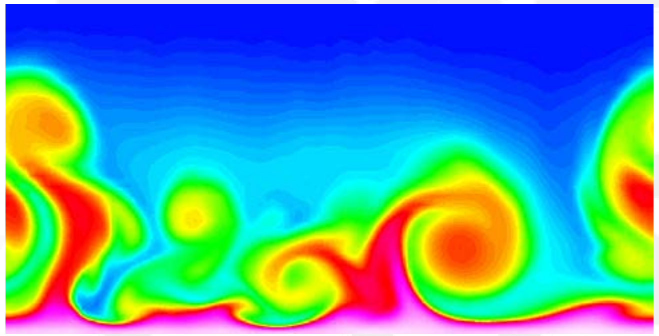
Why Mine Data? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - web data, e-commerce
  - purchases at department/ grocery stores
  - bank/credit card transactions
- Computers have become cheaper and more powerful
- Competitive pressure is strong
  - provide better, customized services for an *edge* (e.g. in customer relationship management)

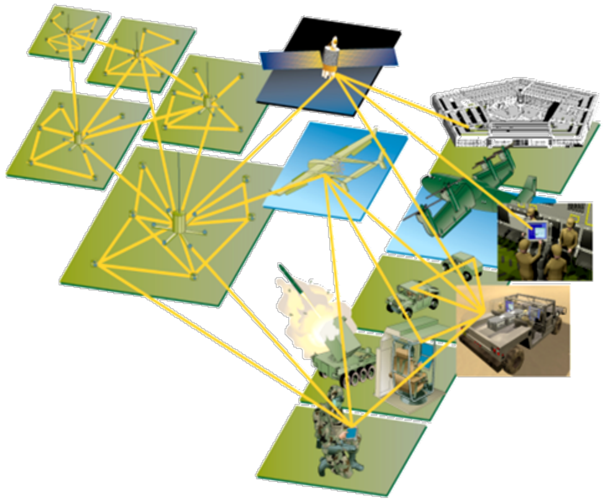
Why Mine Data? Scientific Viewpoint

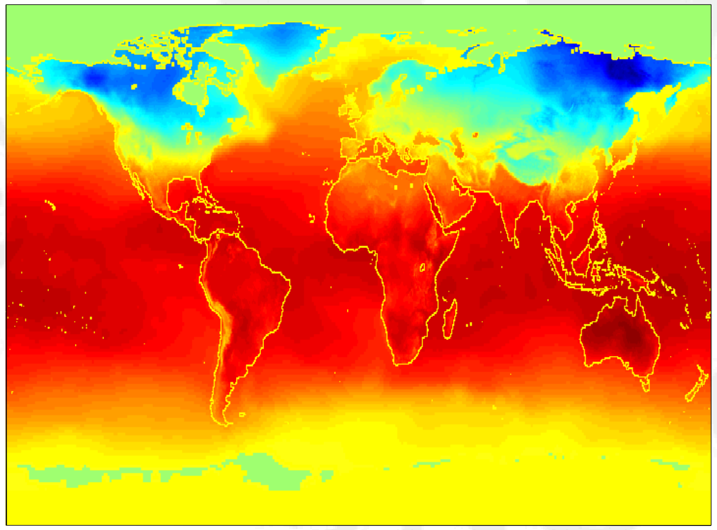


- Data collected and stored at enormous speeds (Gb/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data



- Traditional techniques infeasible for raw data
- Data mining may help scientists in classifying and segmenting data in hypothesis formation

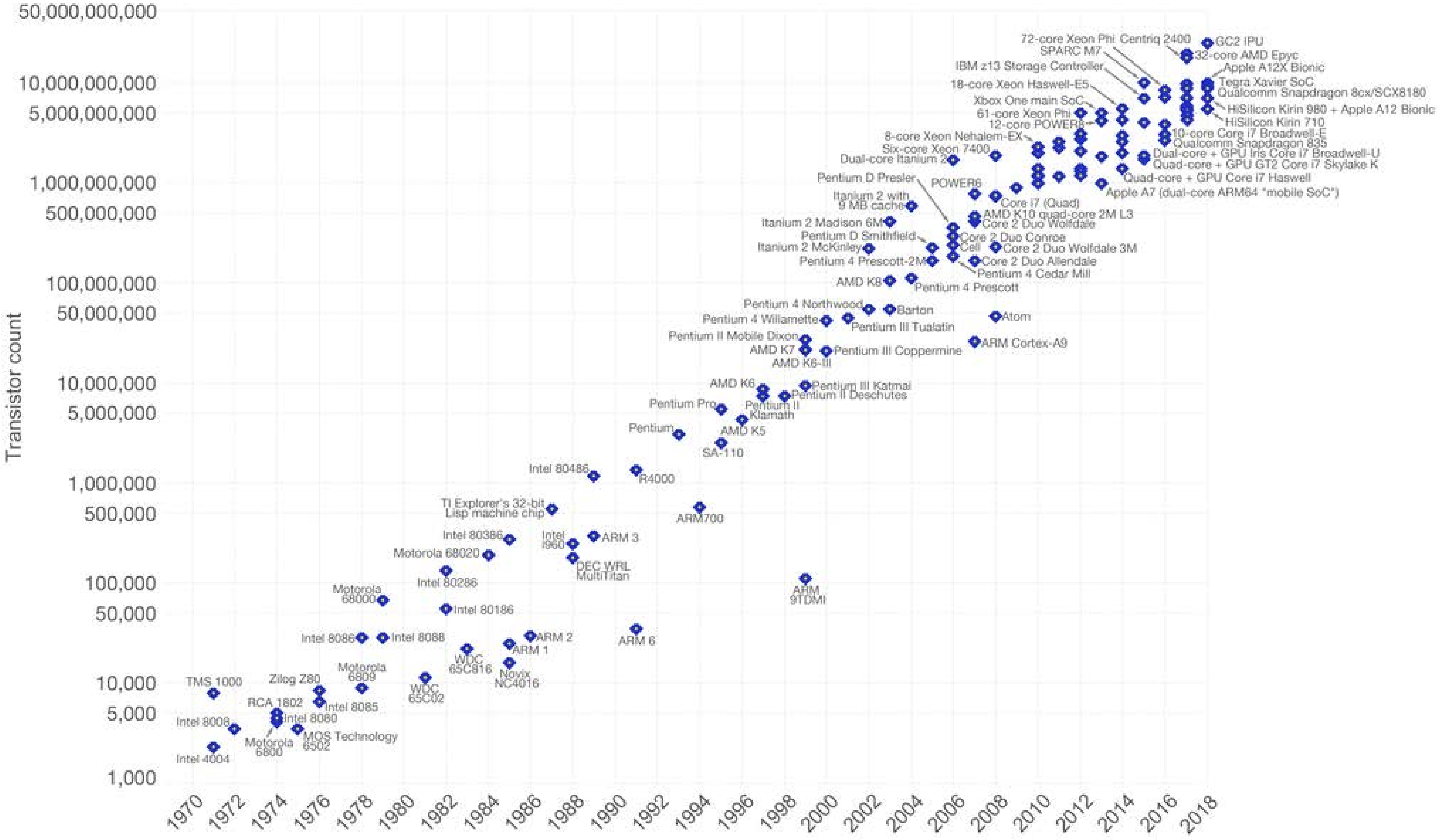




Moore’s law

Moore’s Law – The number of transistors on integrated circuit chips (1971-2018)

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are linked to Moore's 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.

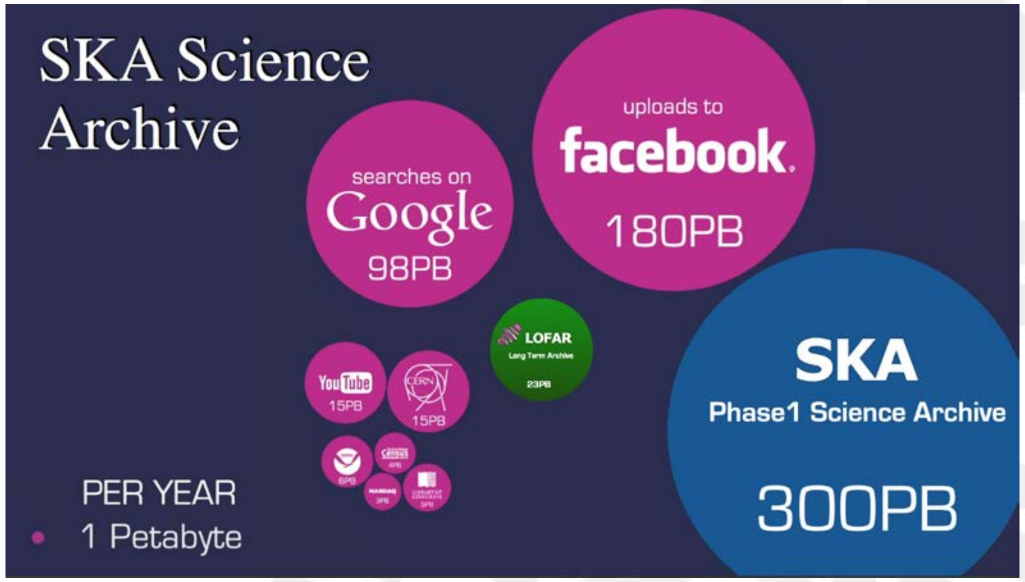
Licensed under CC-BY-SA by the author Max Roser.

Mining Large Data Sets - Motivation

- Current data volume estimated at ~40 Zettabyte (10^{13} GB), and doubling every 1.5 years to reach 44 ZB in 2020
- 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

FACEBOOK'S HADOOP CLUSTER CONTAINS OVER 100 PB OF DATA
NETFLIX STREAMS OVER 1 BILLION HOURS OF VIDEO MONTHLY
AMAZON OWNS THE 3 LARGEST LINUX DBs WITH DATA STORES UP TO 24.7 TB
DATA CENTER TO END USER TRAFFIC WILL BE 561 EB IN 2013
PayPal
PAYPAL PROCESSES OVER $315M IN PAYMENTS PER DAY
DATA
400,000,000 TWEETS PER DAY ON Twitter
350 MILLION IMAGES UPLOADED DAILY TO FACEBOOK
EBAY PROCESSES OVER 1 BILLION TRANSACTIONS DAILY
AMAZON CLOUD HANDLES 500,000 TRANSACTIONS PER SECOND
72 HOURS OF VIDEO ARE UPLOADED TO YOUTUBE EVERY MINUTE

Examples of Massive Datasets



- Pubmed text database
  - Records for >30 million published articles
- Web search engines
  - 60 billion Web pages indexed
  - 100’s 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

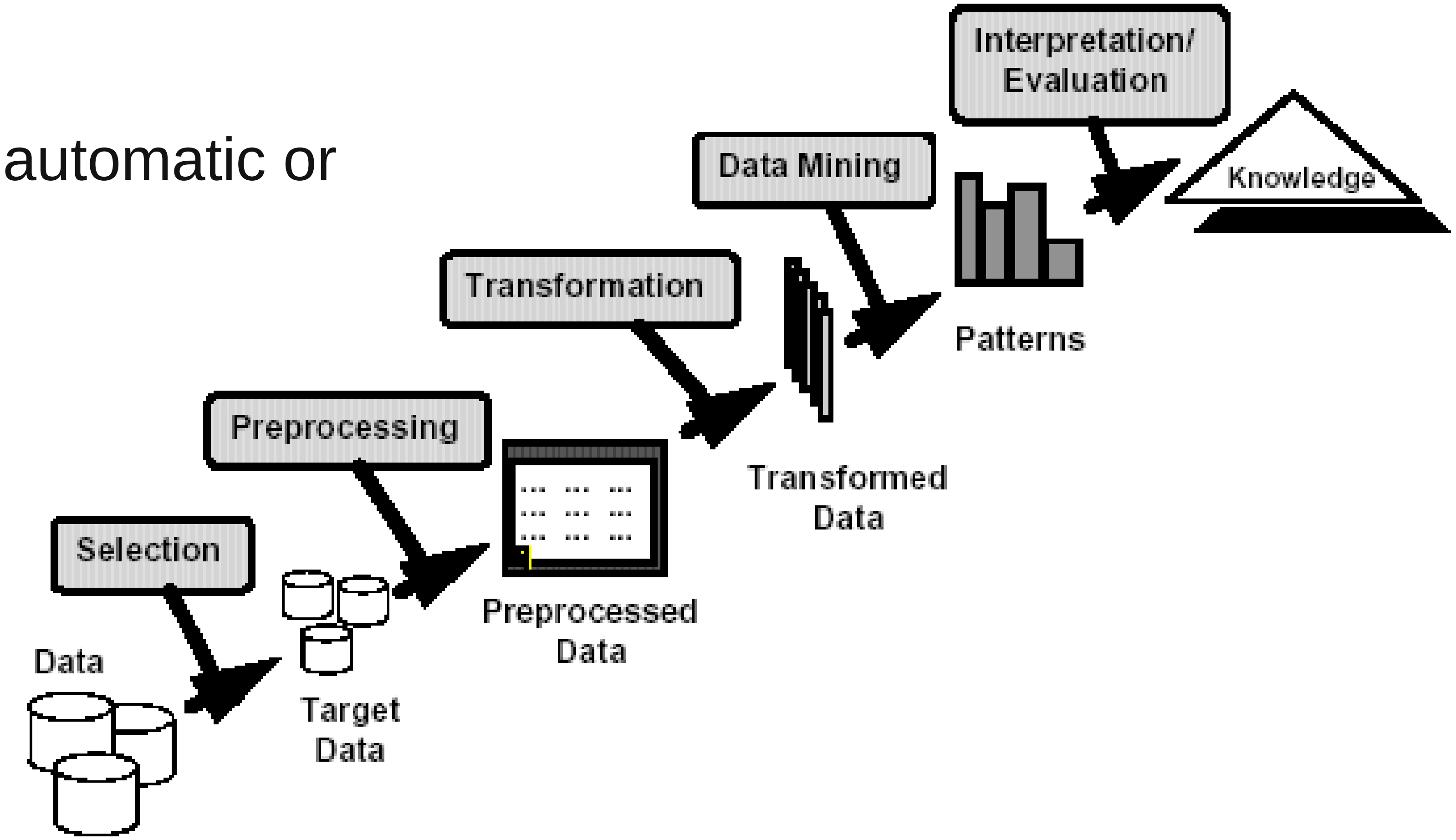
W[illegible]

What is Data Mining?

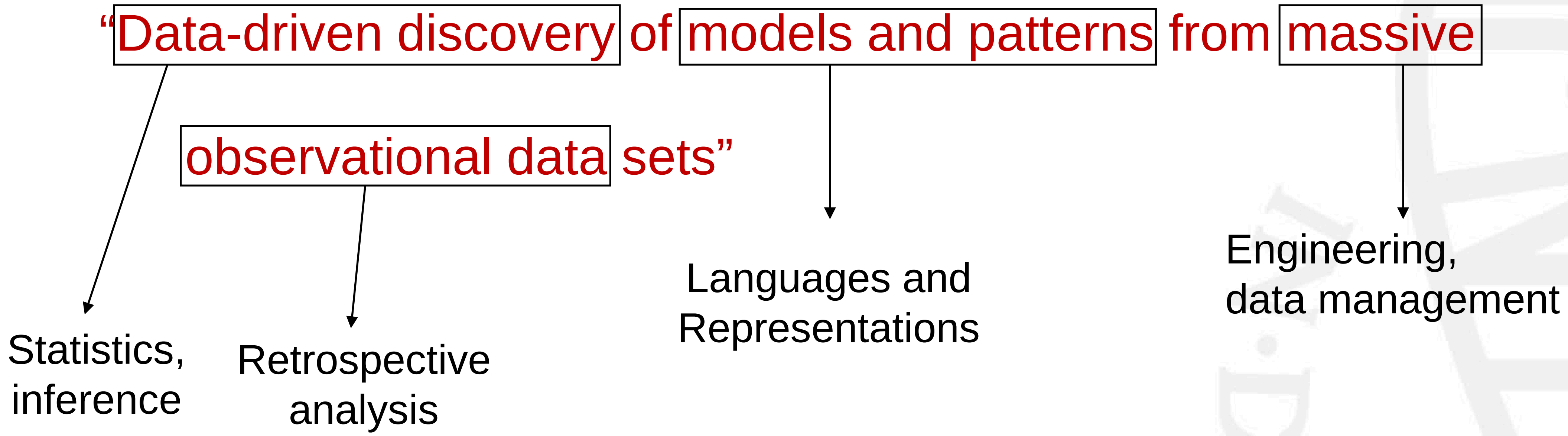
Many definitions:

- 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 in order 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?

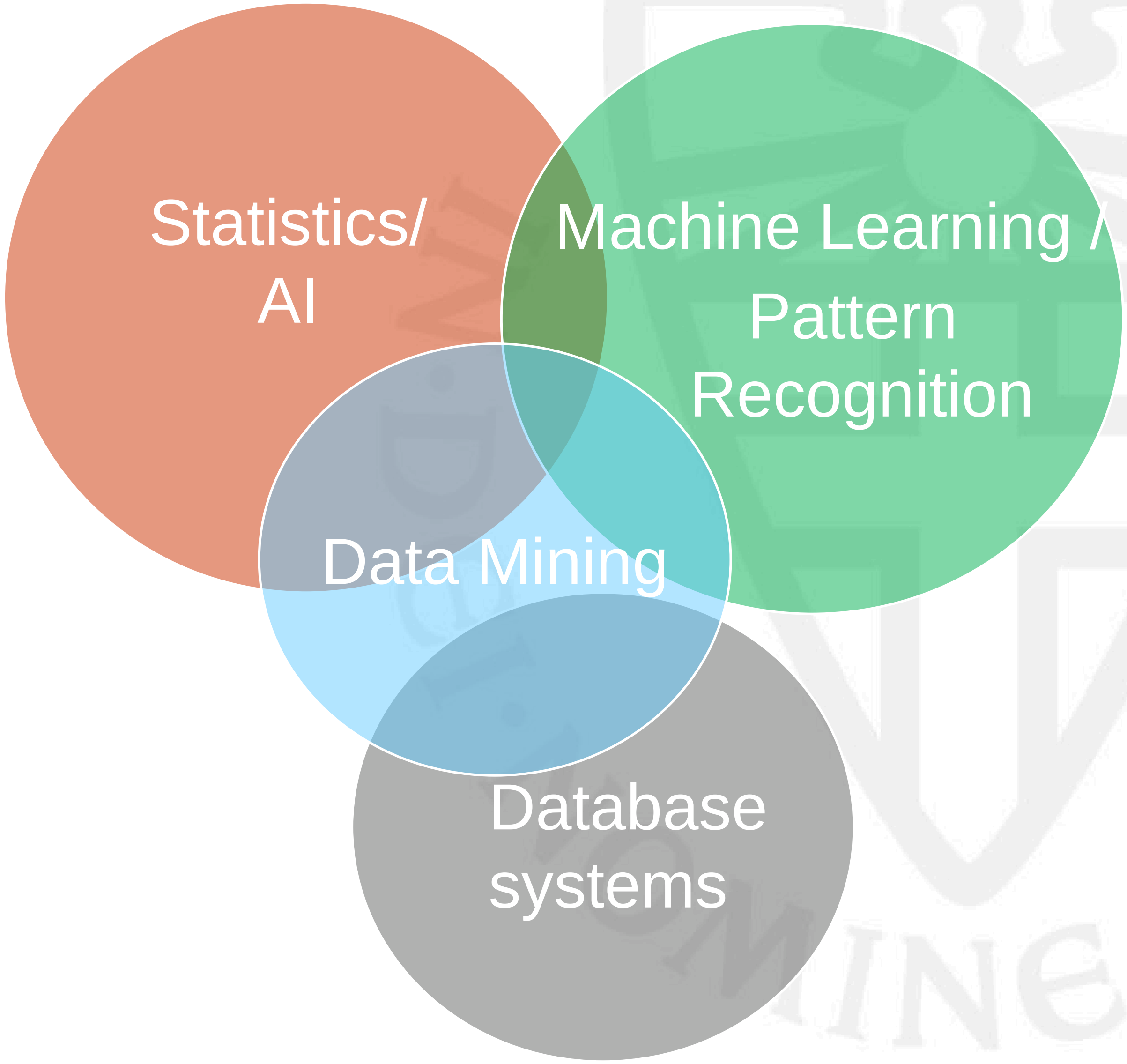
- What is not Data Mining?
  - Look up phone number in phone directory
  - Query a Web search engine for information about “Amazon”

- 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)

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems

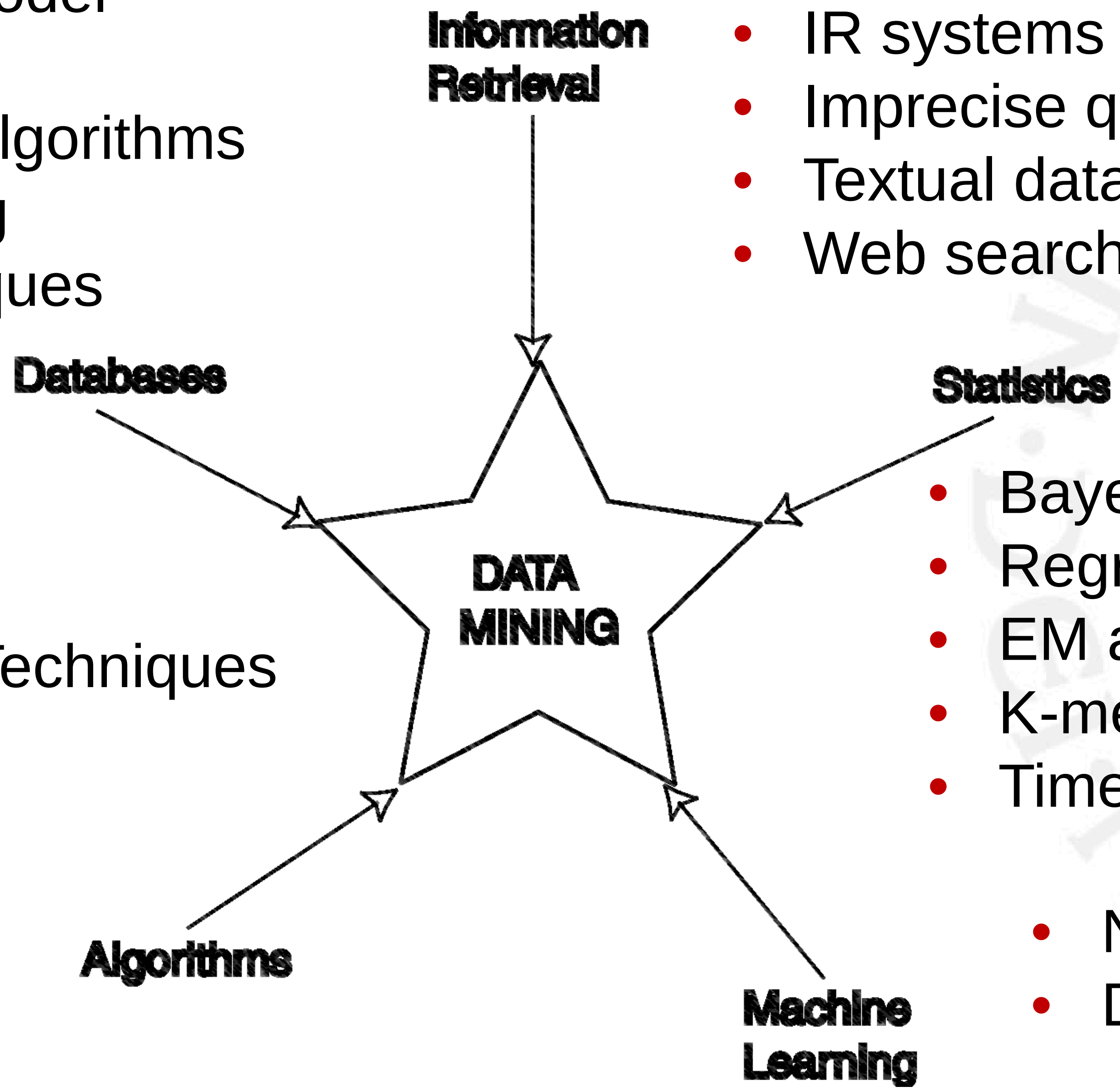
- Traditional techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



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



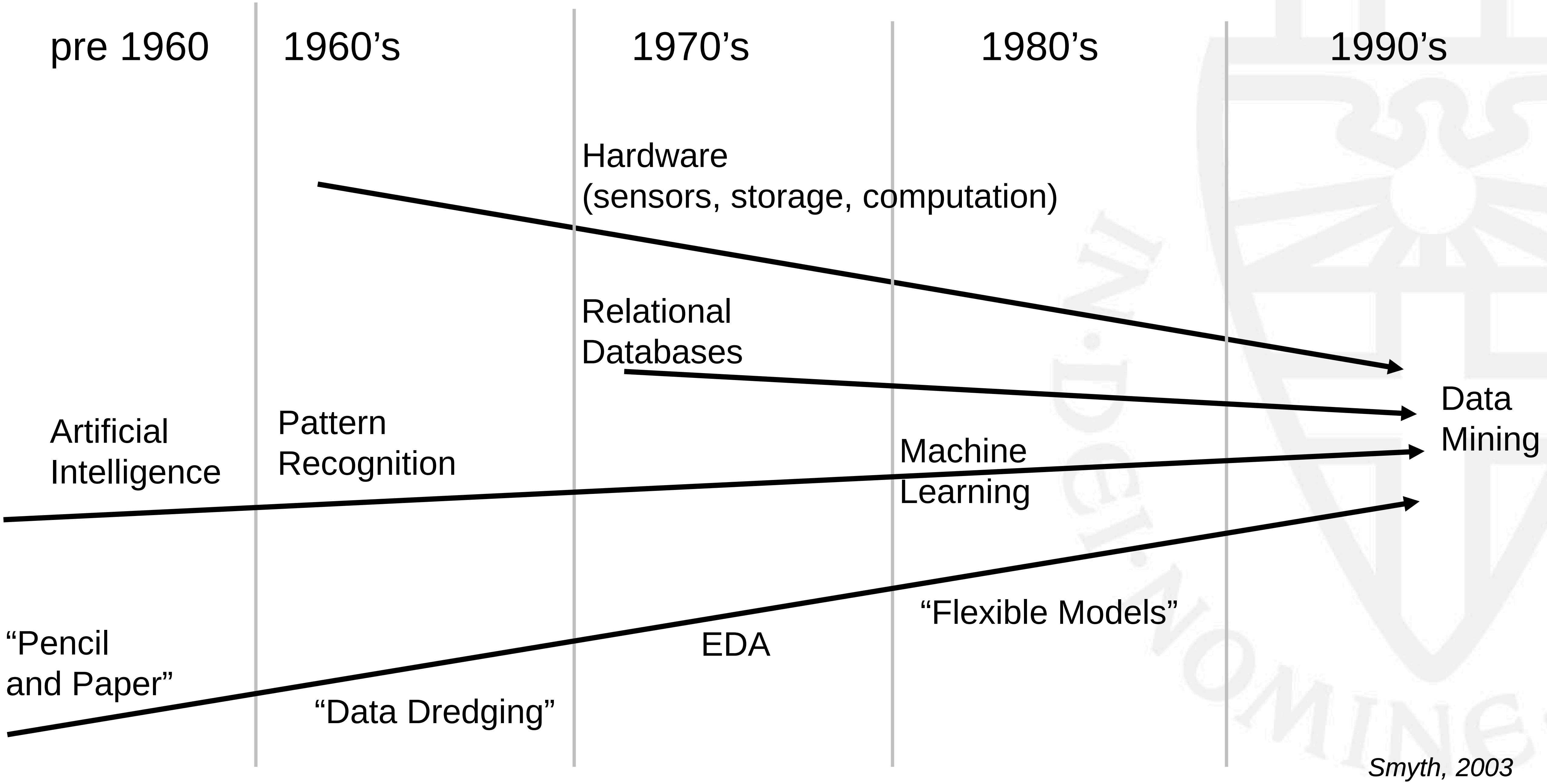
- Bayes theorem
- Regression analysis
- EM algorithm
- K-means clustering
- Time series analysis

- Algorithm Design Techniques
- Algorithm Analysis
- Data Structures

- Neural networks
- Decision tree algorithms

Dunham, 2003

Origins of Data Mining



pre 1960

1960’s

1970’s

1980’s

1990’s

Hardware
(sensors, storage, computation)

Relational
Databases

Artificial
Intelligence

Pattern
Recognition

Machine
Learning

Data
Mining

“Pencil
and Paper”

“Data Dredging”

EDA

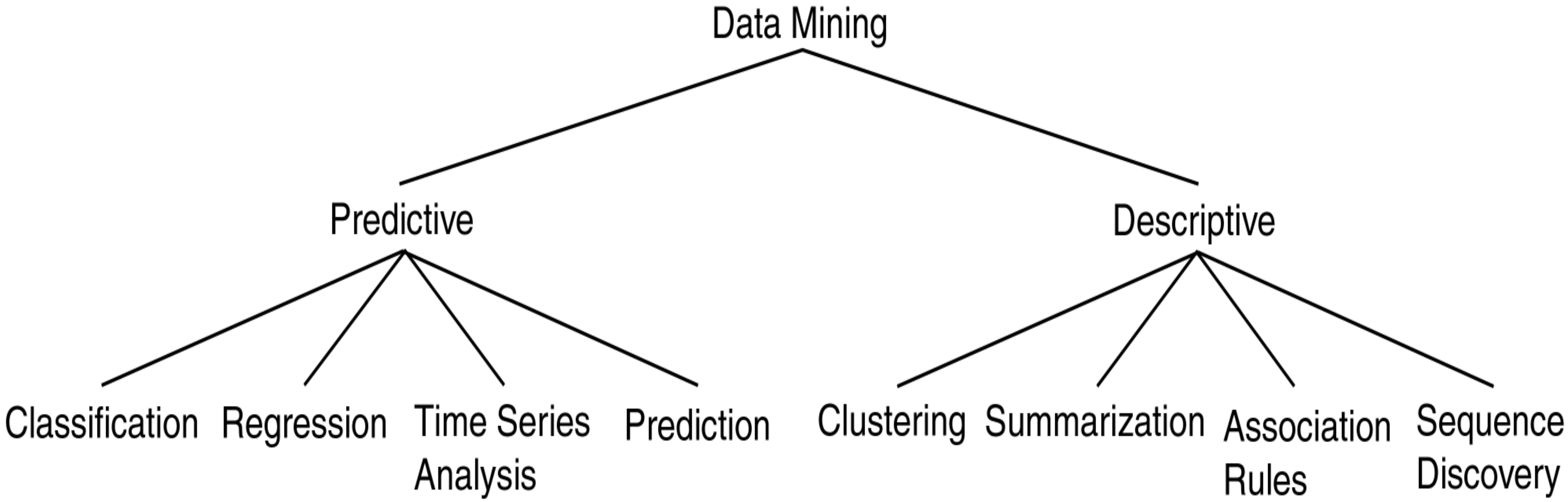
“Flexible Models”

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: Definition

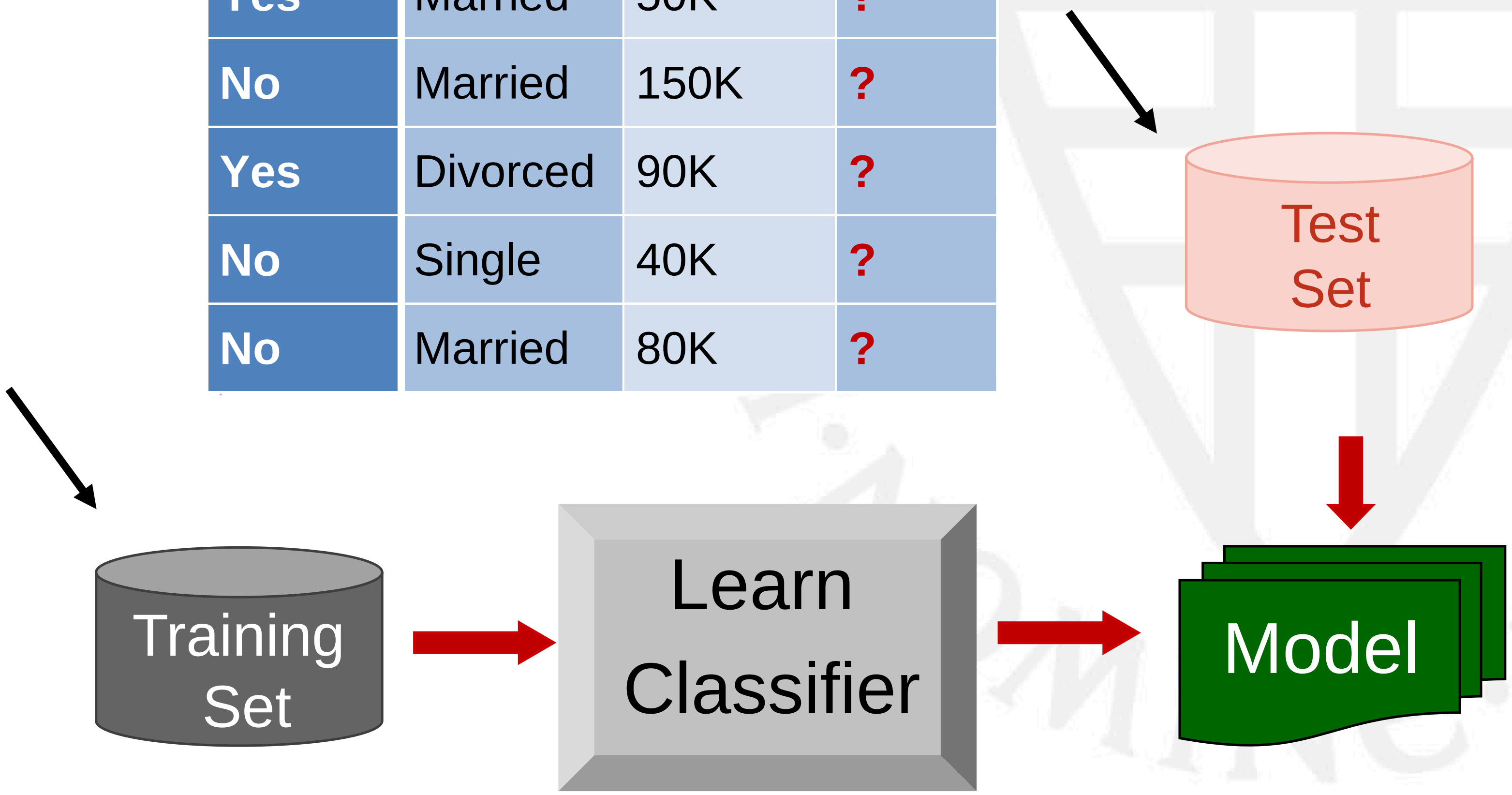
- Given a collection of records (training set).
  - Each record contains a set of attributes, one of the attributes is the class.
- Find a model for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - 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 Example

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: Application 1

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: Application 2

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: Application 3

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: Application 4

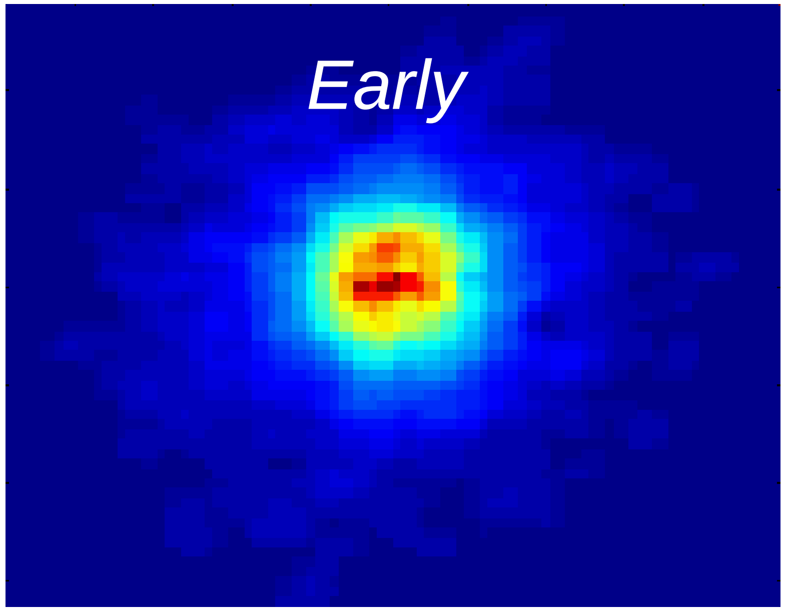
Sky Survey Cataloging

- 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!

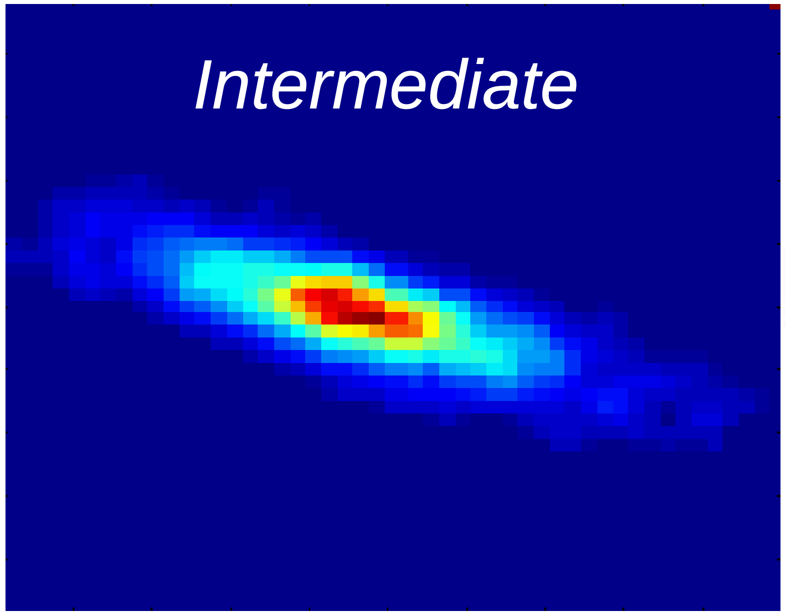
Classifying Galaxies

Courtesy: http://aps.umn.edu



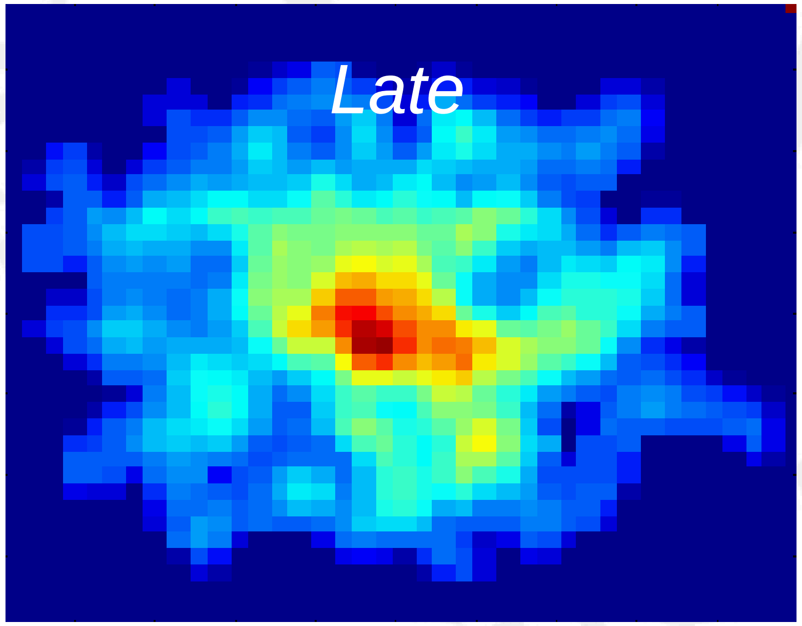
Class:

- Stages of formation



Attributes:

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



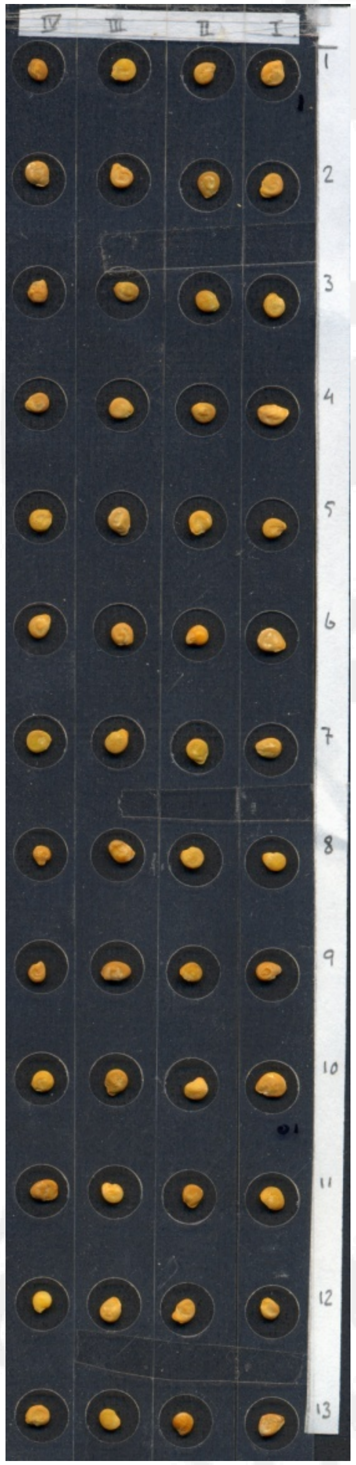
Data size:

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

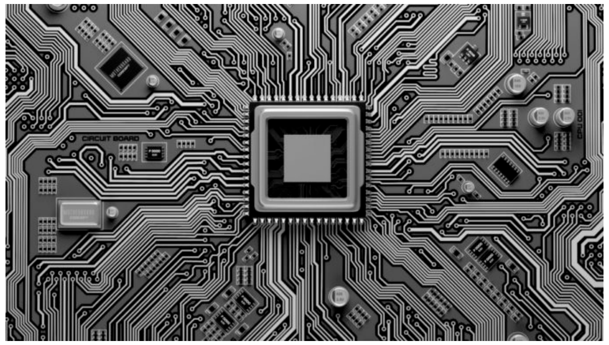
Classification: Application 5

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



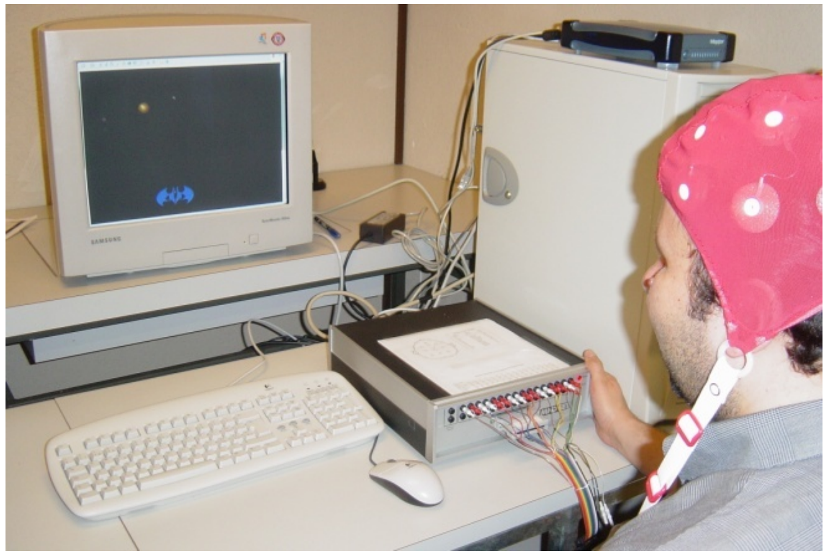
MACHINE2LEARN



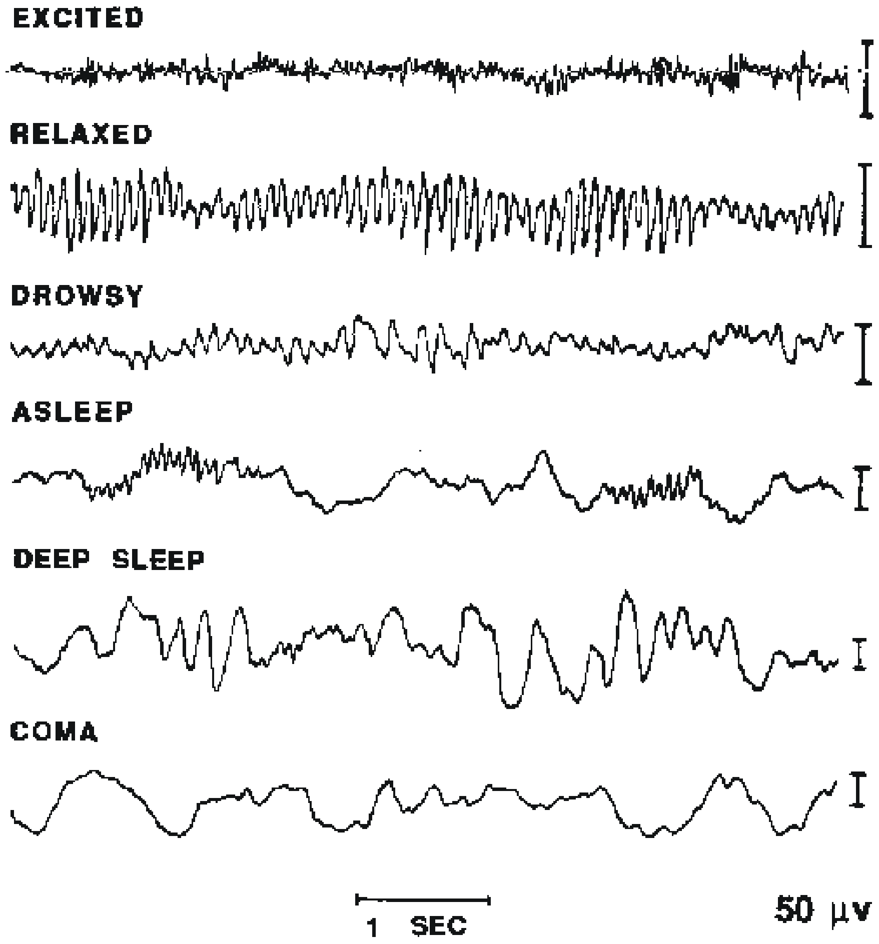
Classification: Example 6

Brain-computer interfacing

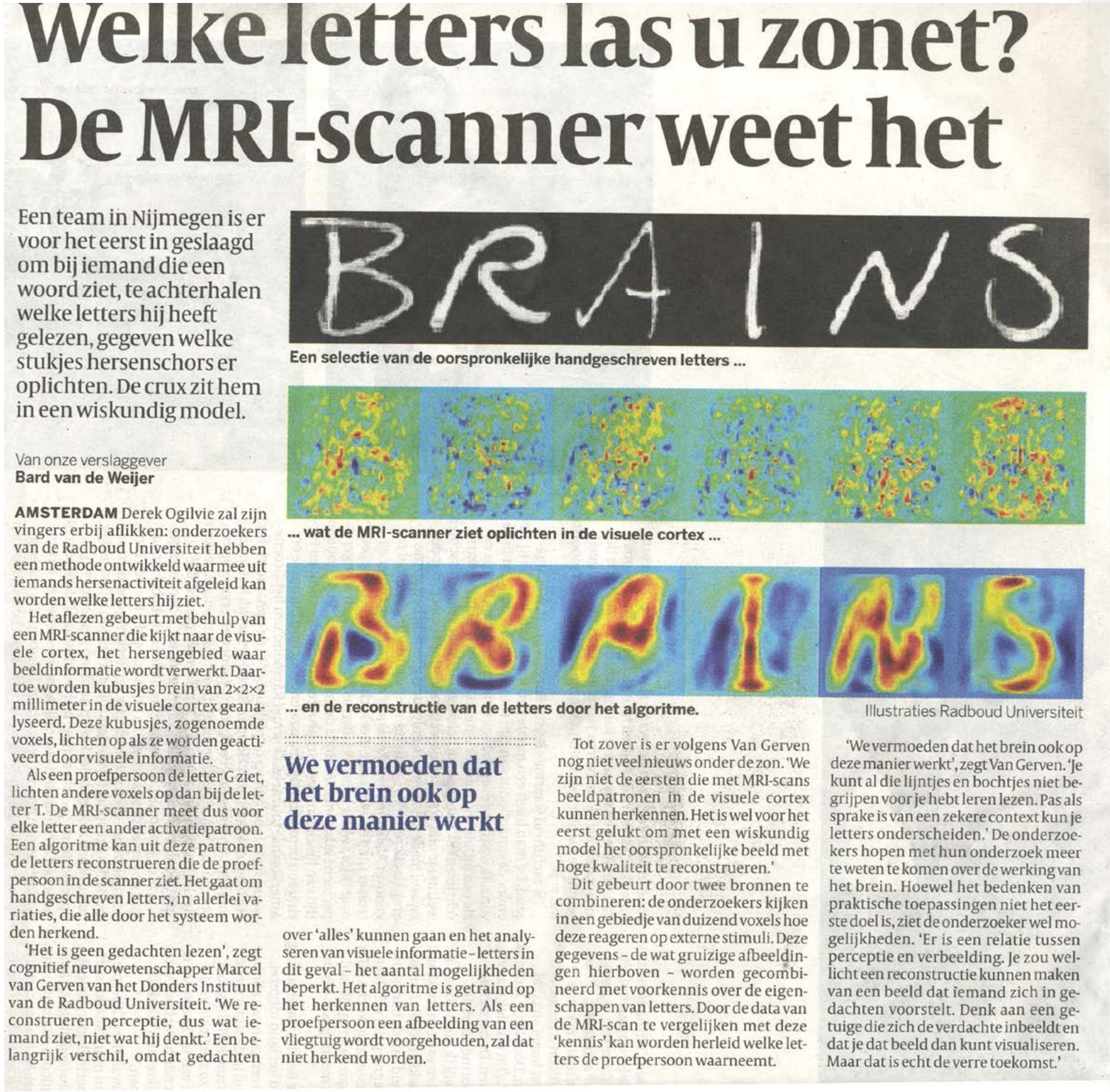
- Goal: read a person's mind
- Approach:
  - measure EEG signals
  - classify them



EEG
ElectroEncephaloGram



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.

Van onze verslaggever
Bard van de Weijer

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

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.'

Clustering Definition

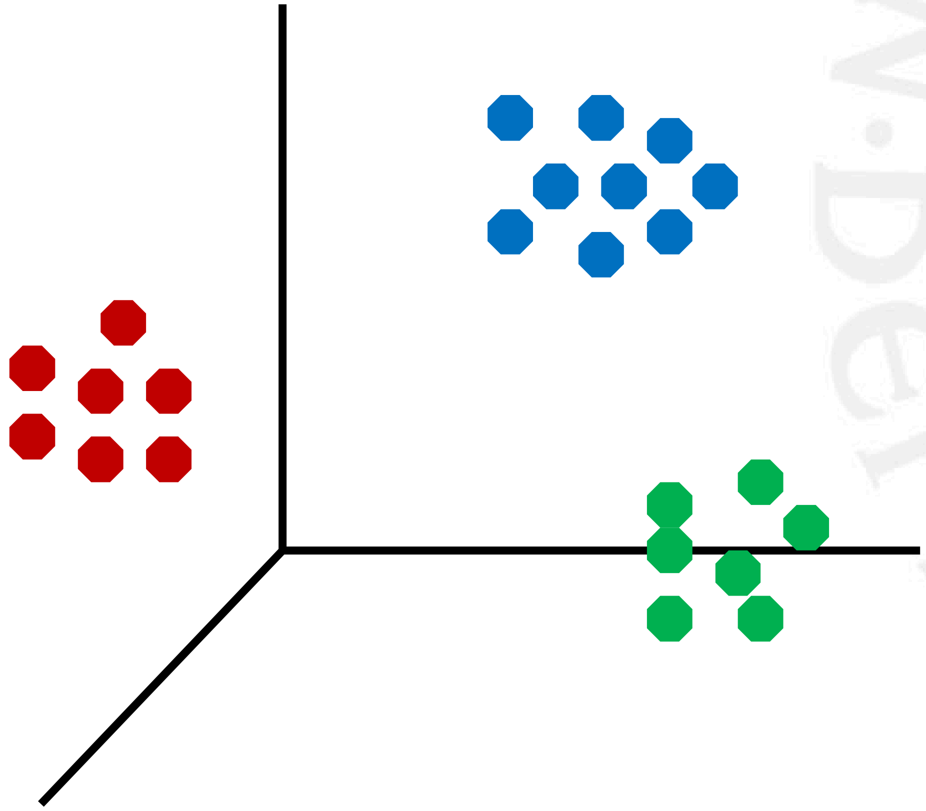
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity measures:
  - Euclidean distance if attributes are continuous.
  - Other problem-specific measures.

Illustrating Clustering

- Euclidean distance based clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Clustering: Application 1

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: Application 2

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 stock movements on a daily basis.
- 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: Definition

- 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: Application 1

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: Application 2

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)

Association Rule Discovery: Application 3

Market basket analysis at Schuitema (now Jumbo)

- Goal: find and visualize clusters of products that are “similar”, i.e., are typically bought together with the same products
- Approach: self-organizing map using specific similarity measure based on co-occurrence

Self-organizing Map

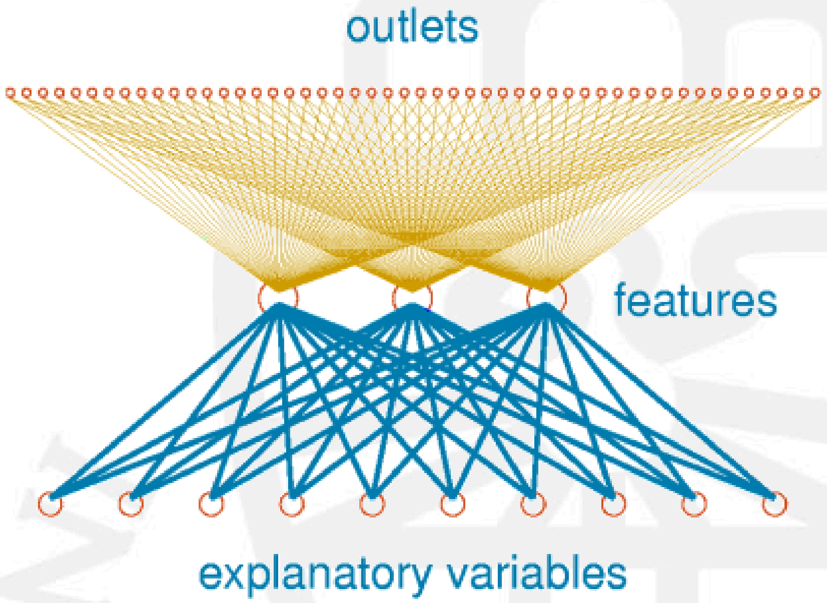
Regression

- 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.
- 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.

Regression: Application 1

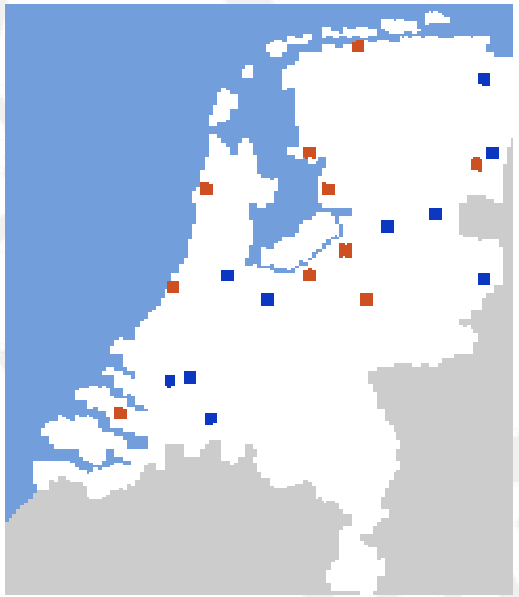
Predicting newspaper sales

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



🟥 better weather, more sales

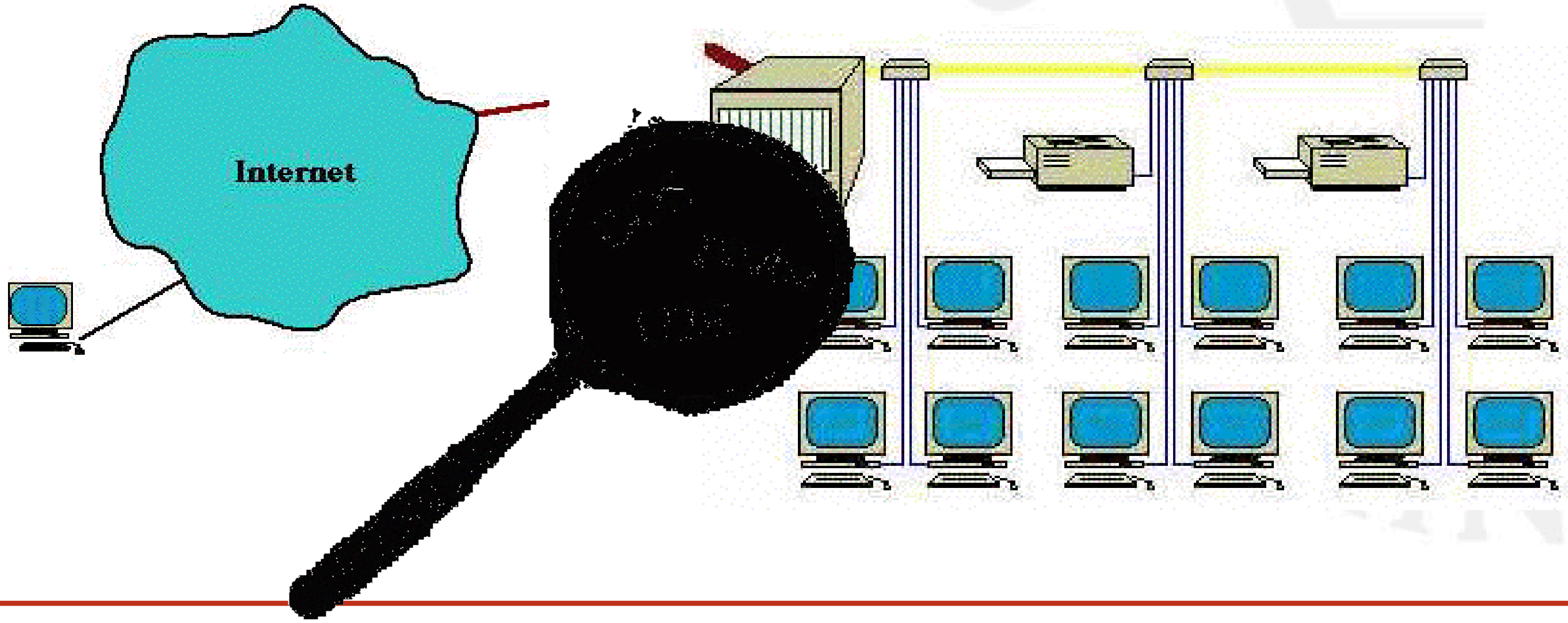
🟦 worse weather, more sales



Deviation / Anomaly Detection

- Detect significant deviations from normal behavior

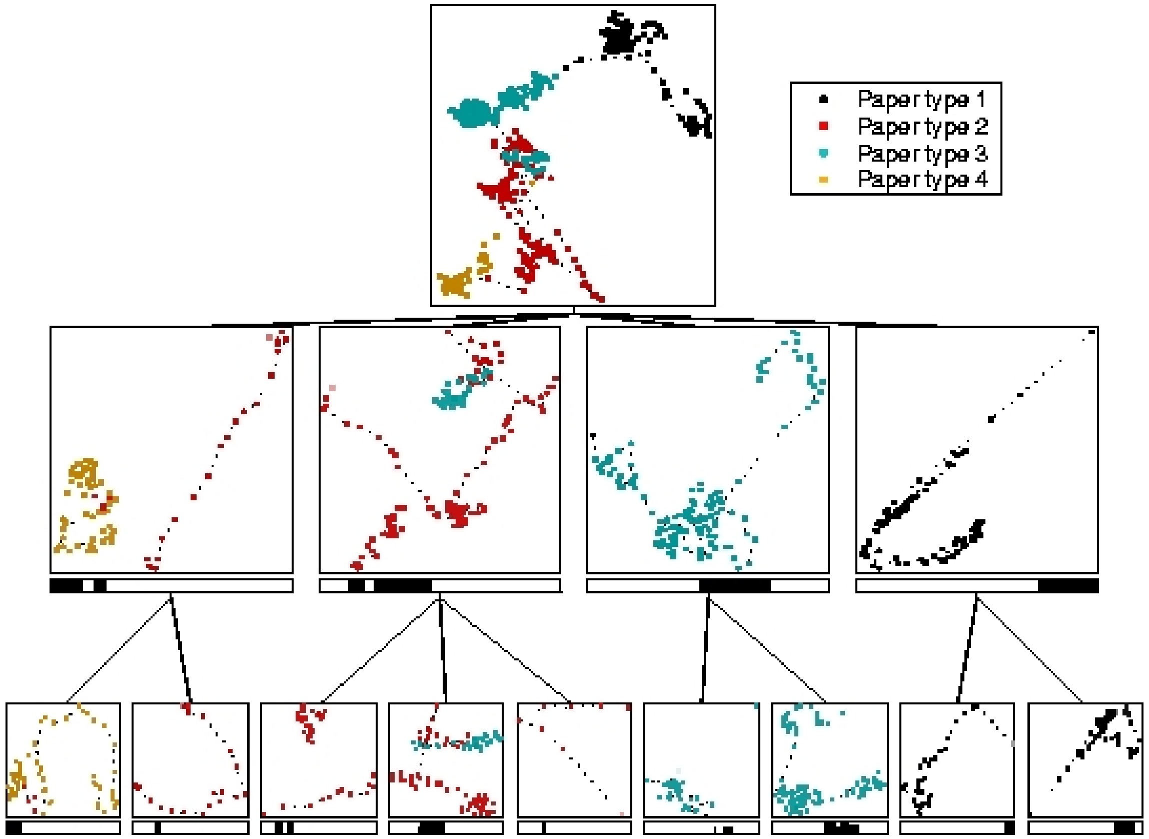
- Applications:
  - Credit card fraud detection
  - Network intrusion detection



Deviation / Anomaly Detection: Application 1

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



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.

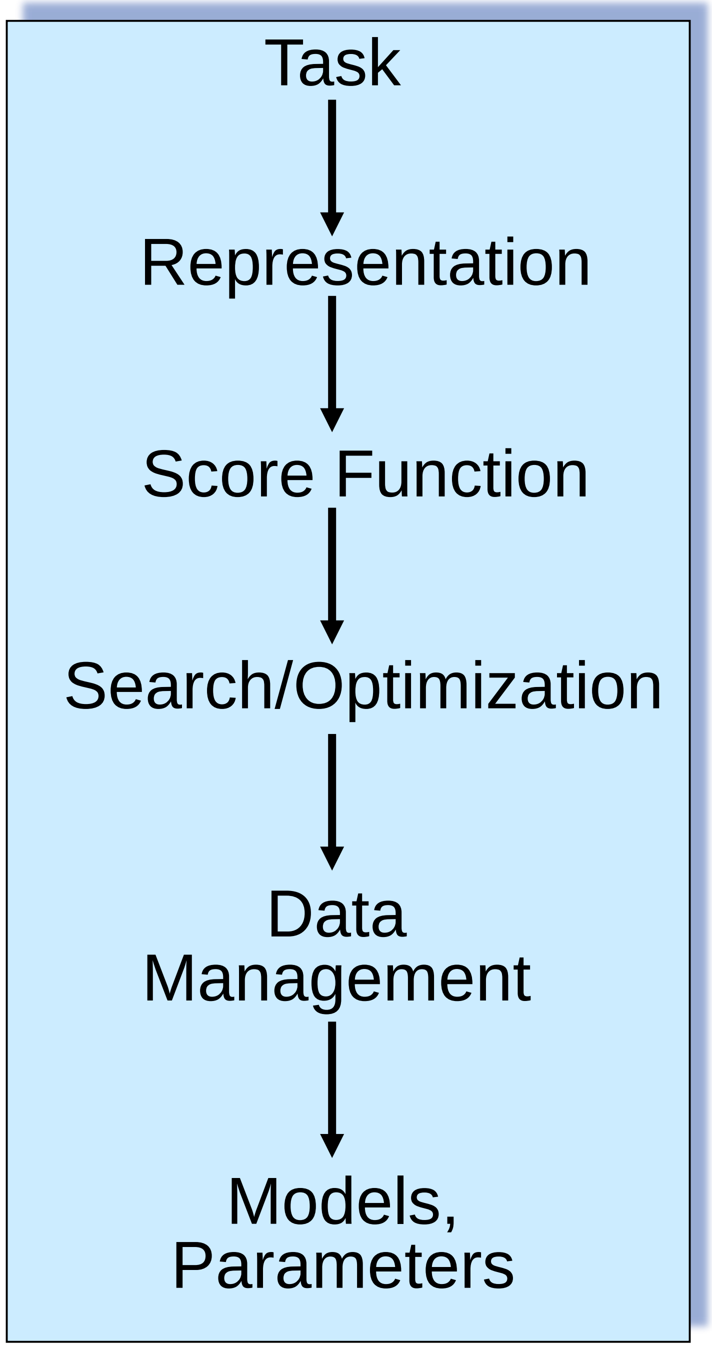
Data mining?

- 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?



Task

Representation

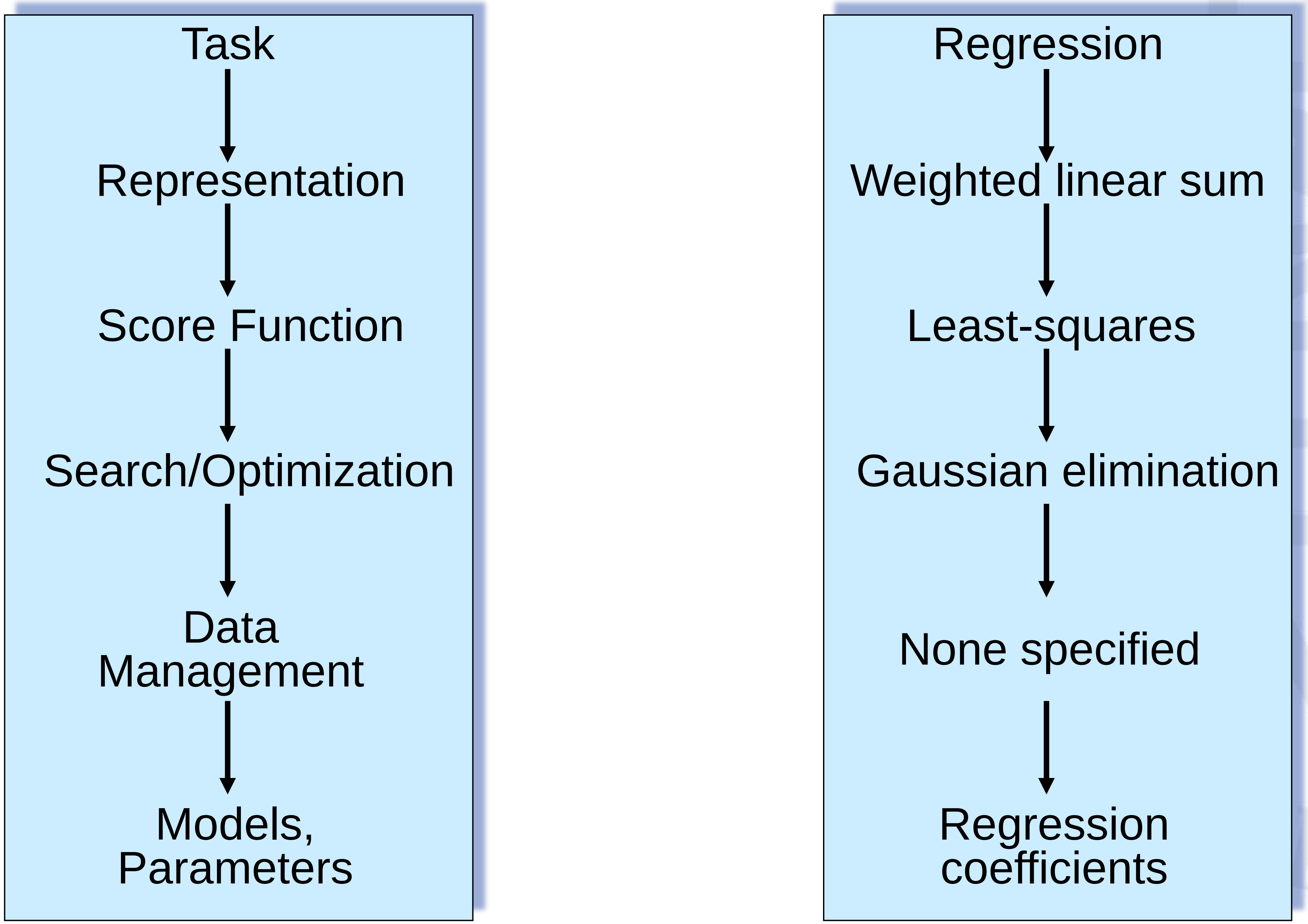
Score Function

Search/Optimization

Data Management

Models, Parameters

Multivariate Linear Regression



Task

Representation

Score Function

Search/Optimization

Data Management

Models, Parameters

Regression

Weighted linear sum

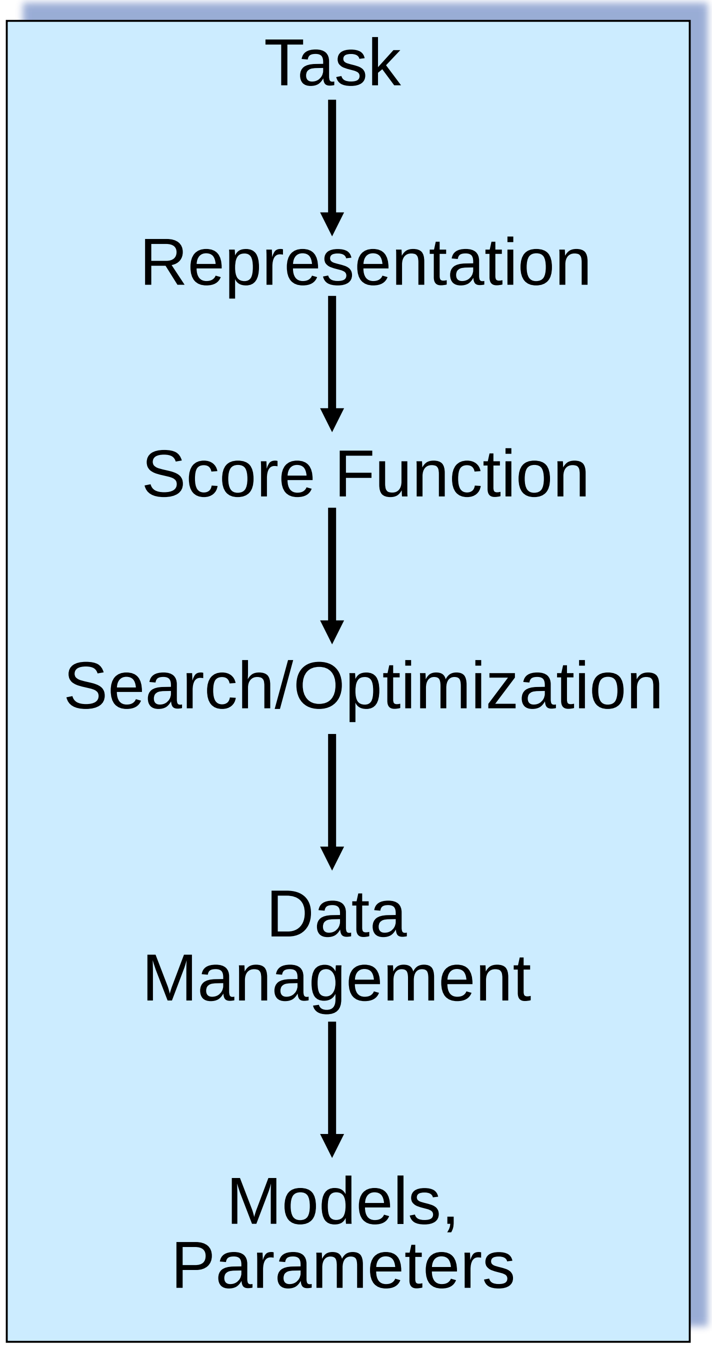
Least-squares

Gaussian elimination

None specified

Regression coefficients

Decision trees (CART, ID3, …)



Task

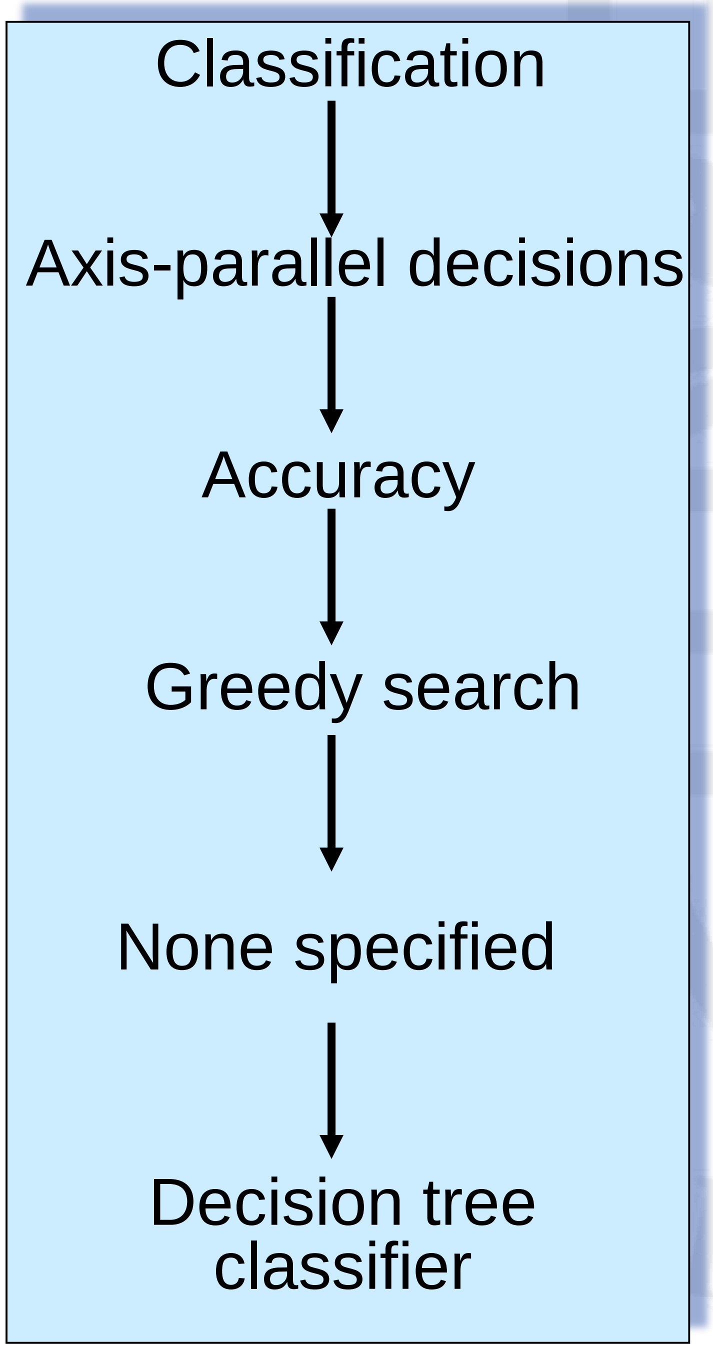
Representation

Score Function

Search/Optimization

Data Management

Models, Parameters



Classification

Axis-parallel decisions

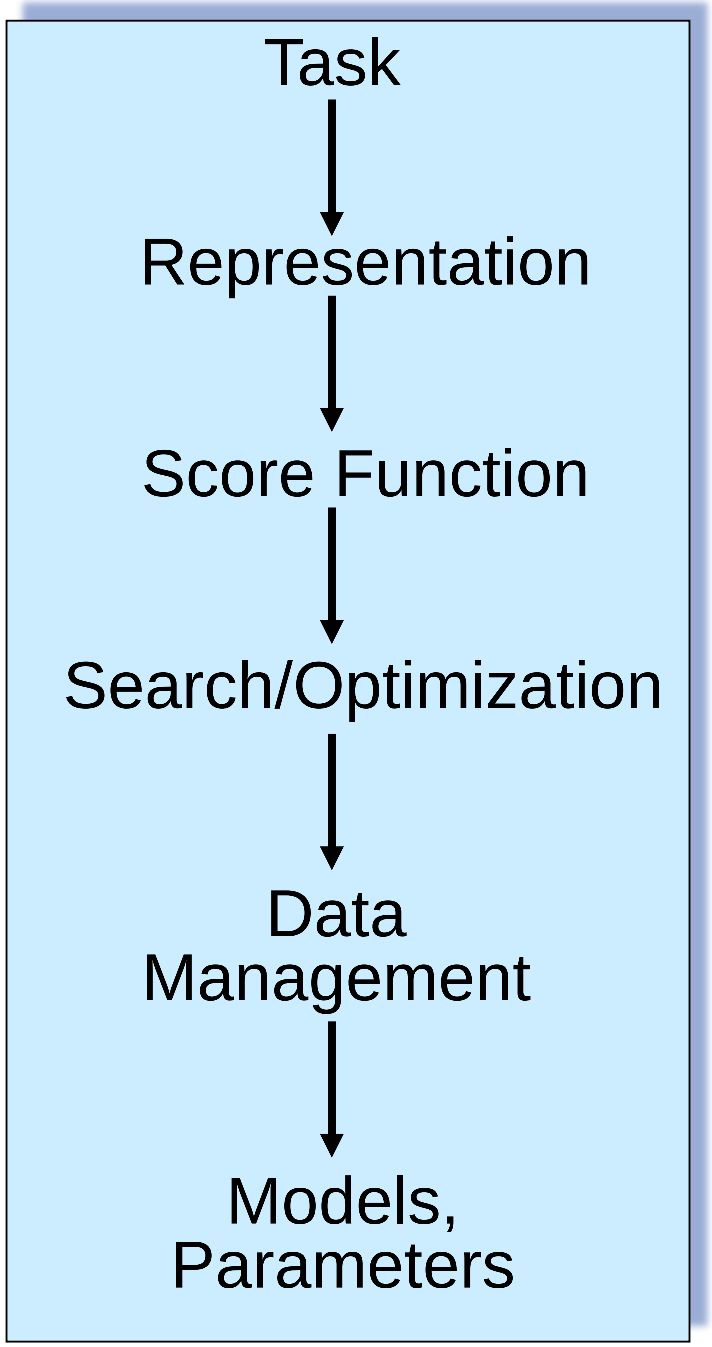
Accuracy

Greedy search

None specified

Decision tree classifier

Hierarchical Clustering



Task

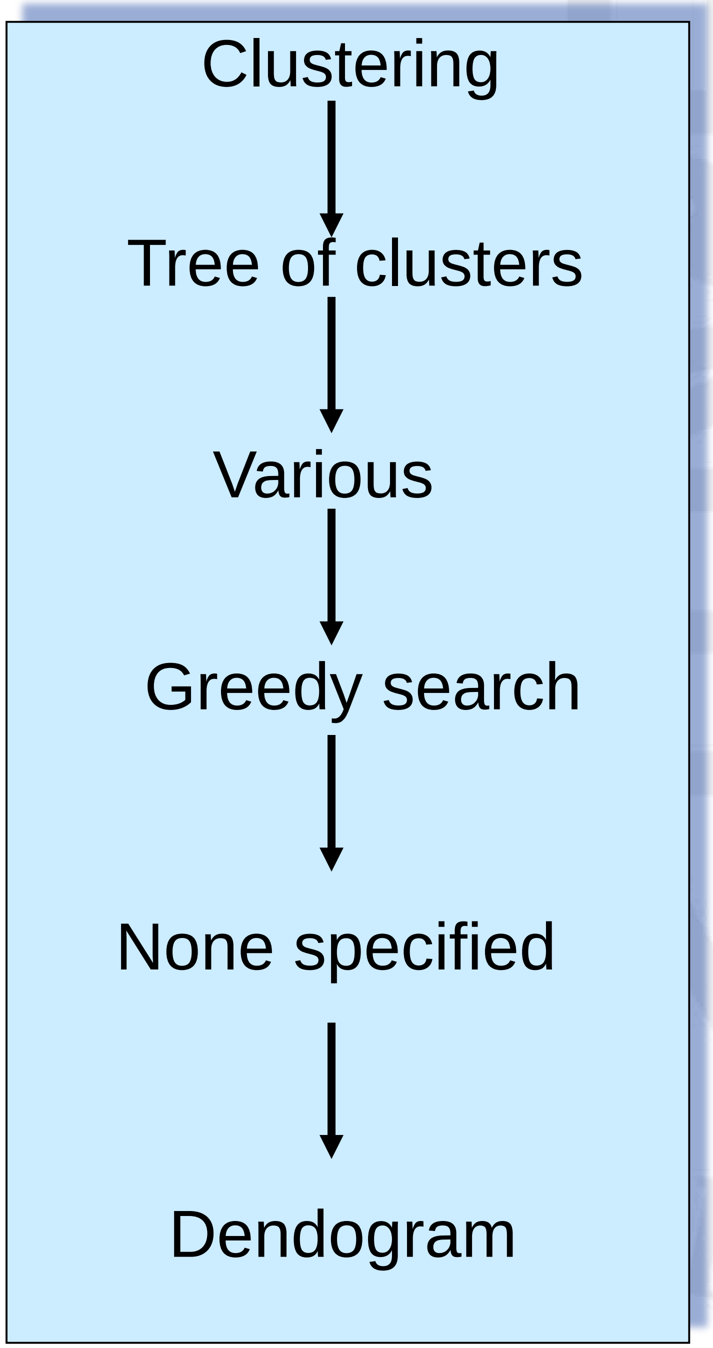
Representation

Score Function

Search/Optimization

Data Management

Models, Parameters



Clustering

Tree of clusters

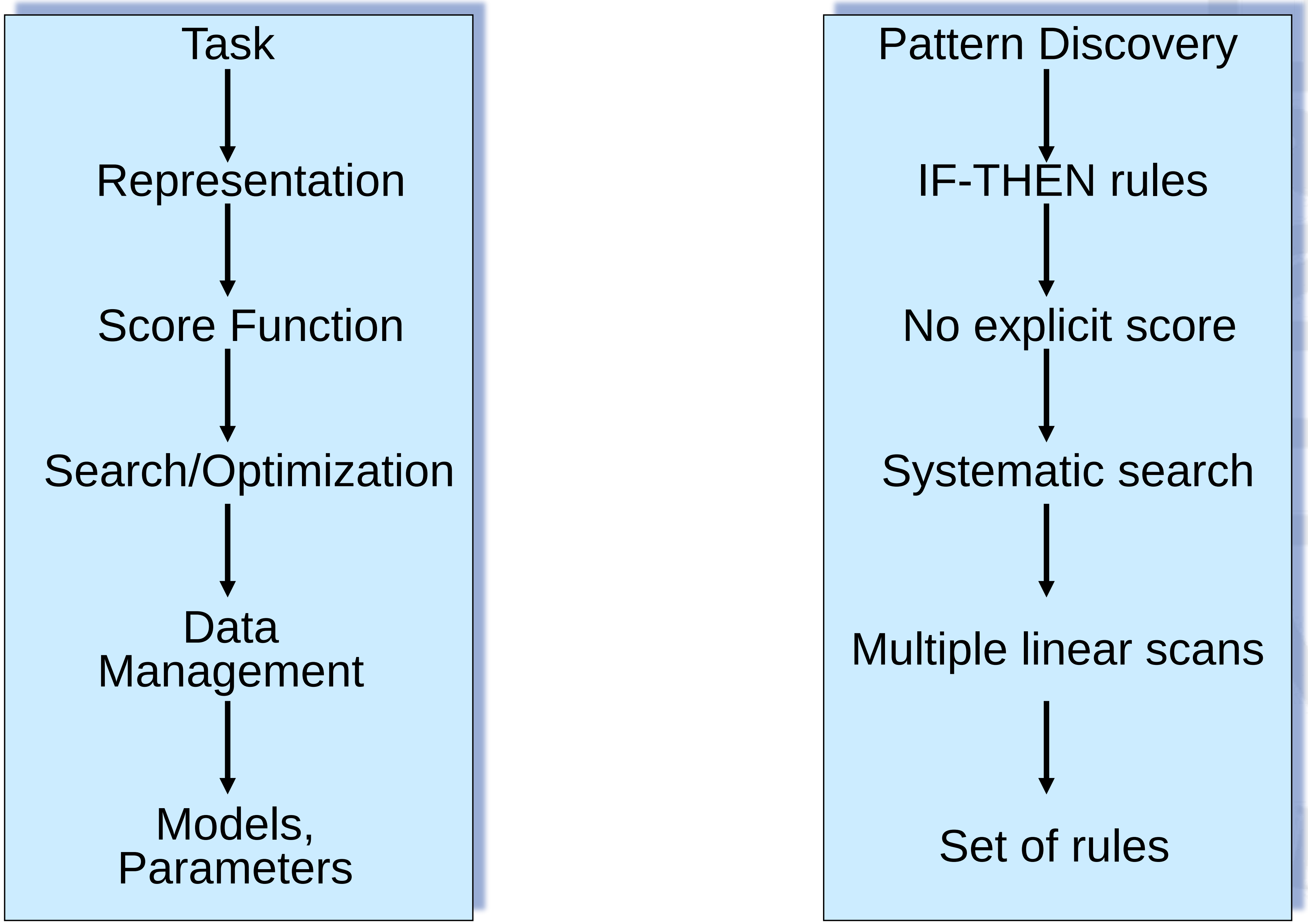
Various

Greedy search

None specified

Dendogram

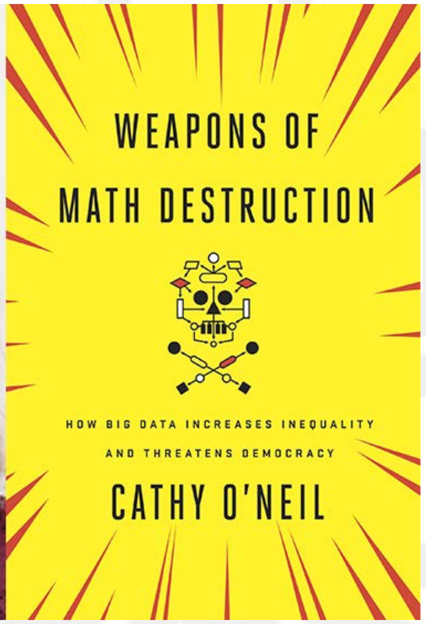
Association Rules



Task → Representation → Score Function → Search/Optimization → Data Management → Models, Parameters

Pattern Discovery → IF-THEN rules → No explicit score → Systematic search → Multiple linear scans → Set of rules

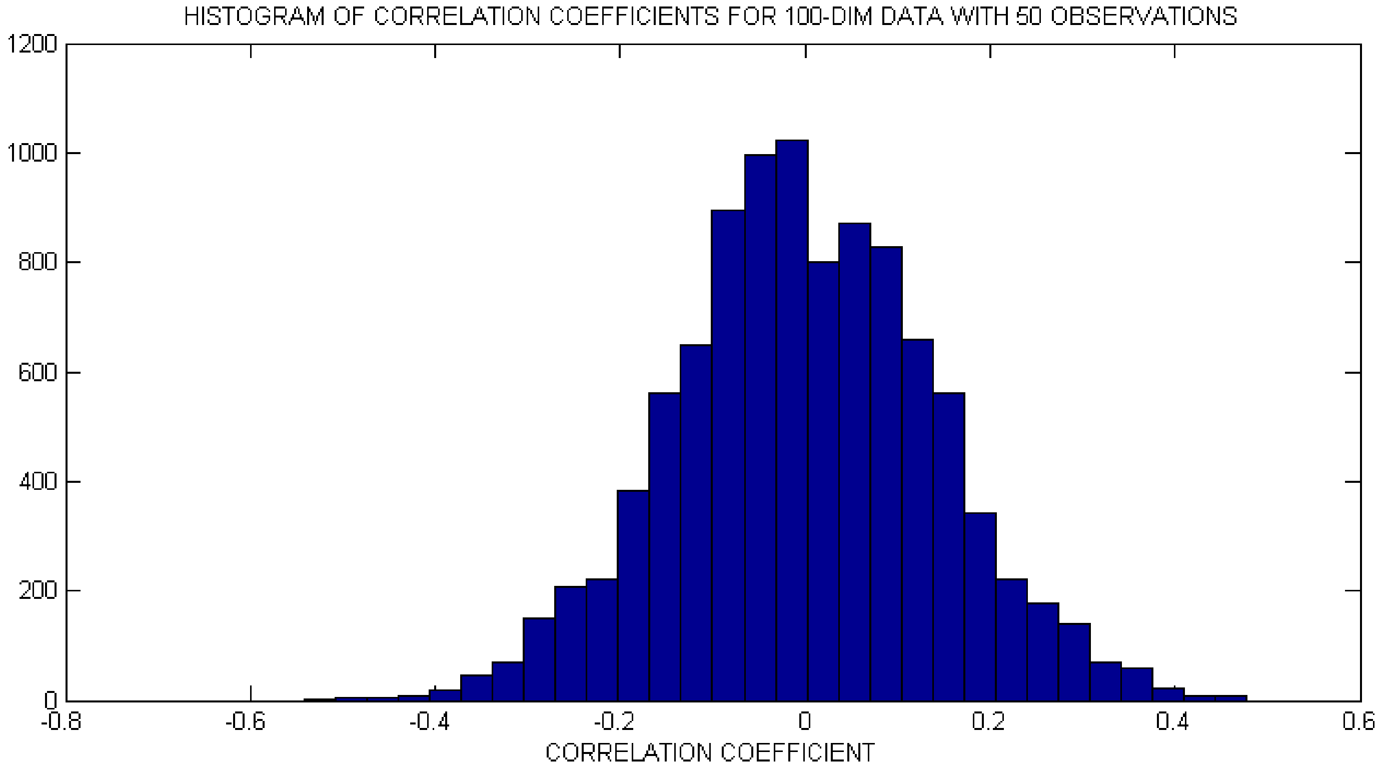
Data Mining: the Downside



- Hype
- One of the “weapons of math destruction”
- 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

Example of “Data Fishing”

- Example: 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.



PYTHON Code for Correlations

```
import numpy as np
import matplotlib.pyplot as plt

nObjects = 50
nVariables = 100
# Generate matrix with standard normal random variables
mu = 0.0
sigma = 1.0
x = np.random.normal(mu, sigma, (nVariables, nObjects))
# Compute correlations between variables
corrvector = np.array([])
for i in range(1, nVariables):
    for j in range(i+1, nVariables):
      # Numpy return the CC’s in matrix format
      corMtrx = np.corrcoef(x[:,i], x[:,j])
        corrvector = np.append(corrvector, corMtrx[0,1])

# Plot the histogram
plt.hist(corrvector, 20)
plt.show()
```

PYTHON Code for Correlations

```python
import numpy as np
import matplotlib.pyplot as plt

nObjects = 50
nVariables = 100

# Generate matrix with standard normal random variables
mu = 0.0
sigma = 1.0
x = np.random.normal(mu, sigma, (nVariables, nObjects))

# Compute correlations between variables
correlations = np.corrcoef(x)
dummy = np.triu(correlations, 1);
corrvector = dummy[dummy != 0];

# Plot the histogram
plt.hist(corrvector, 20)
plt.show()
```

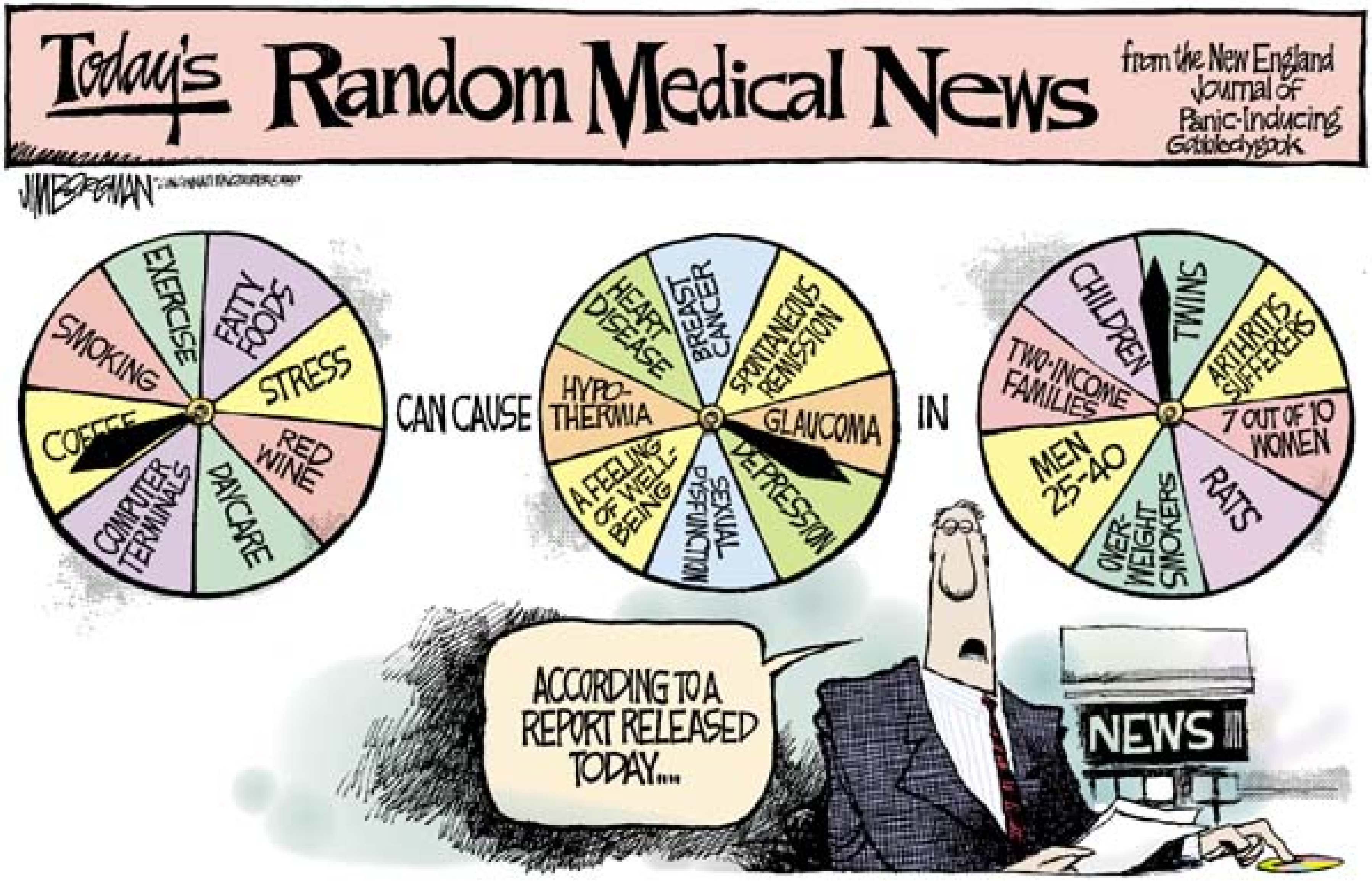
Possible Pitfalls

Let the data speak…

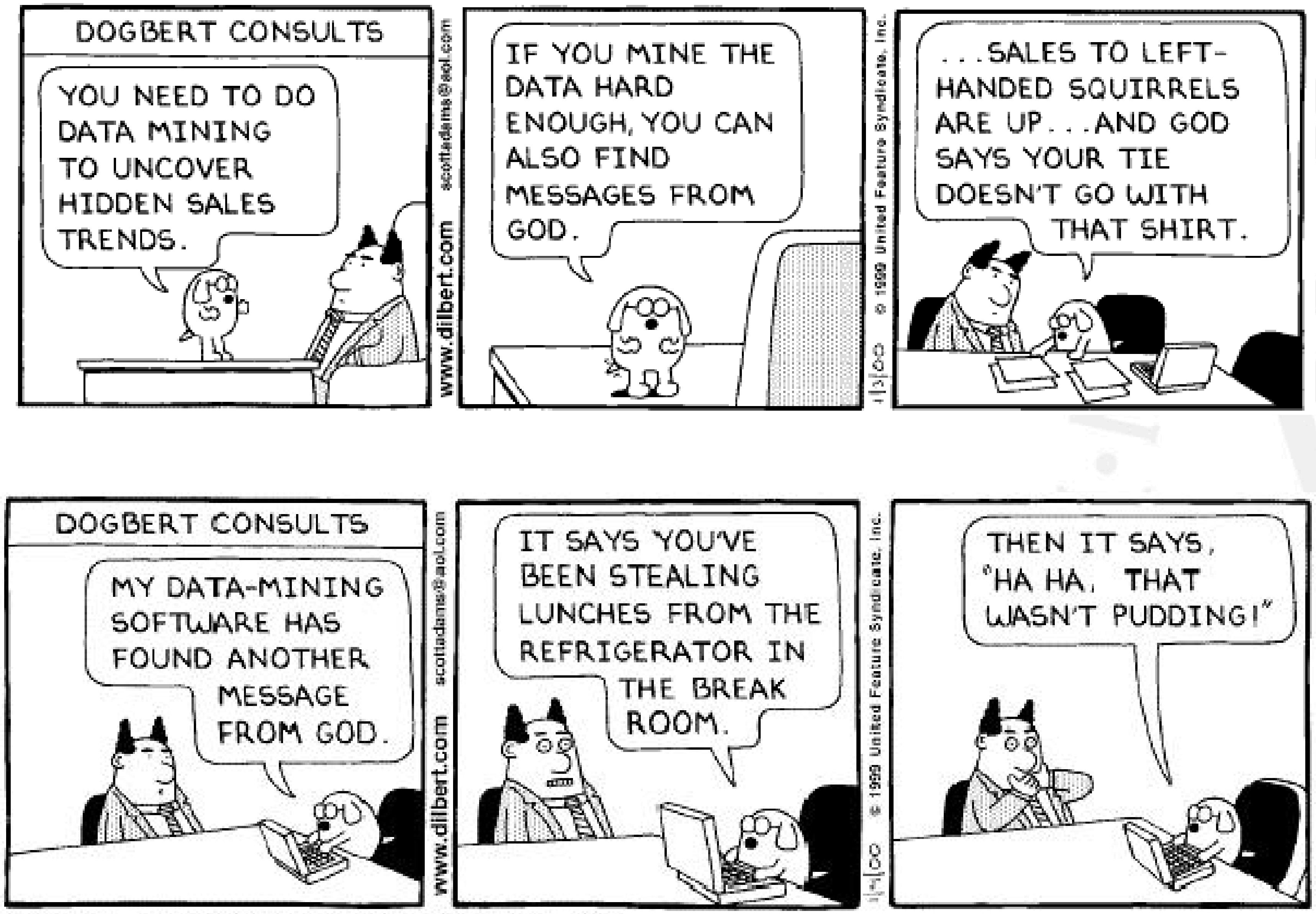
The data may have quite a lot to say…..
but it may just be noise!

Smyth, 2003

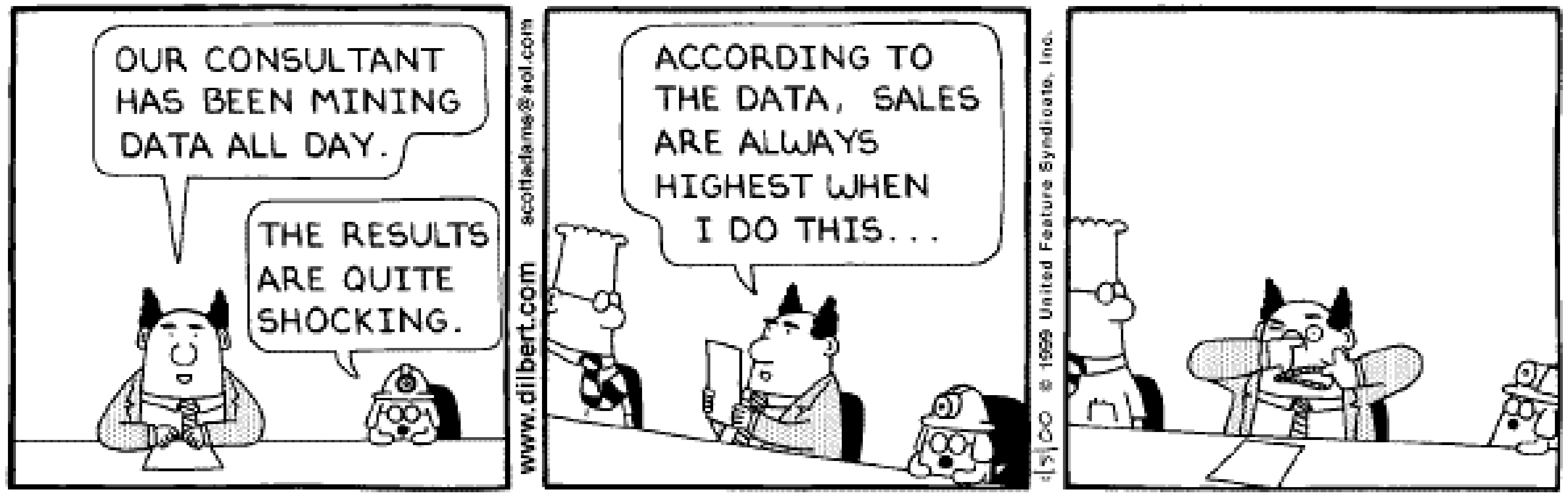
Data Issues in Health Science

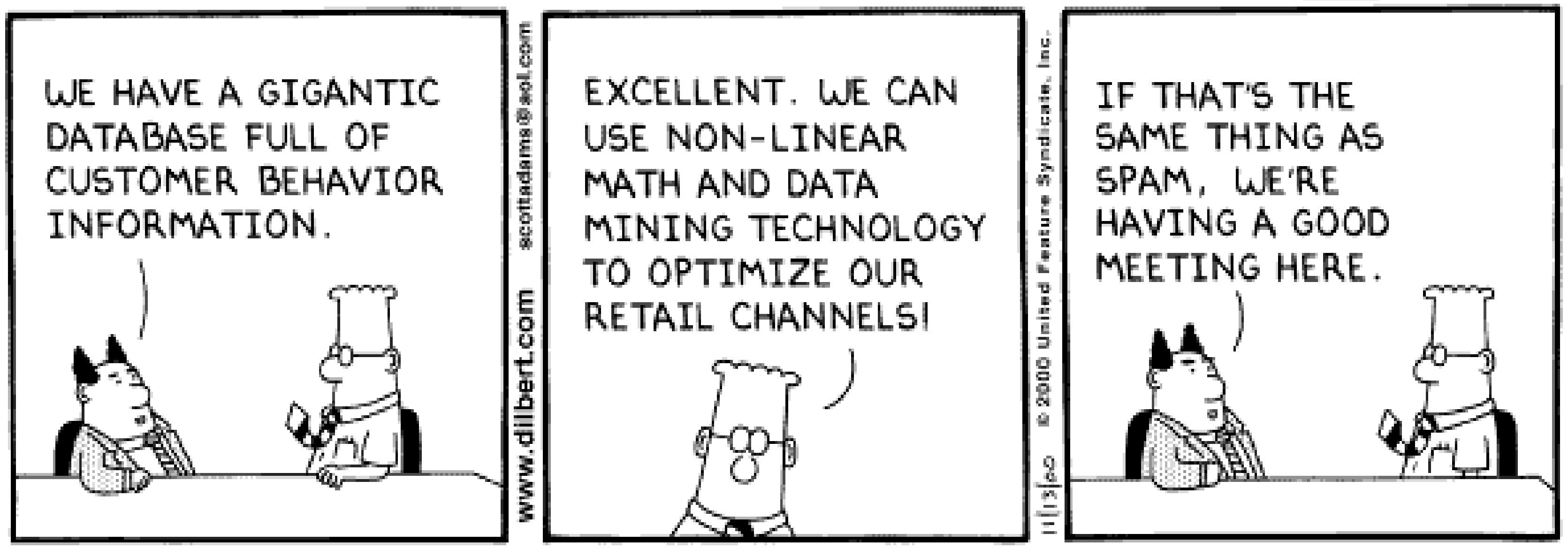


Dilbert (1)


DOGBERT CONSULTS
YOU NEED TO DO DATA MINING TO UNCOVER HIDDEN SALES TRENDS.
scottadams@aol.com
www.dilbert.com
IF YOU MINE THE DATA HARD ENOUGH, YOU CAN ALSO FIND MESSAGES FROM GOD.
© 1999 United Feature Syndicate, Inc.
1/3/00
...SALES TO LEFT-HANDED SQUIRRELS ARE UP...AND GOD SAYS YOUR TIE DOESN'T GO WITH THAT SHIRT.
DOGBERT CONSULTS
MY DATA-MINING SOFTWARE HAS FOUND ANOTHER MESSAGE FROM GOD.
scottadams@aol.com
www.dilbert.com
IT SAYS YOU'VE BEEN STEALING LUNCHES FROM THE REFRIGERATOR IN THE BREAK ROOM.
© 1999 United Feature Syndicate, Inc.
1/4/00
THEN IT SAYS, "HA HA, THAT WASN'T PUDDING!"

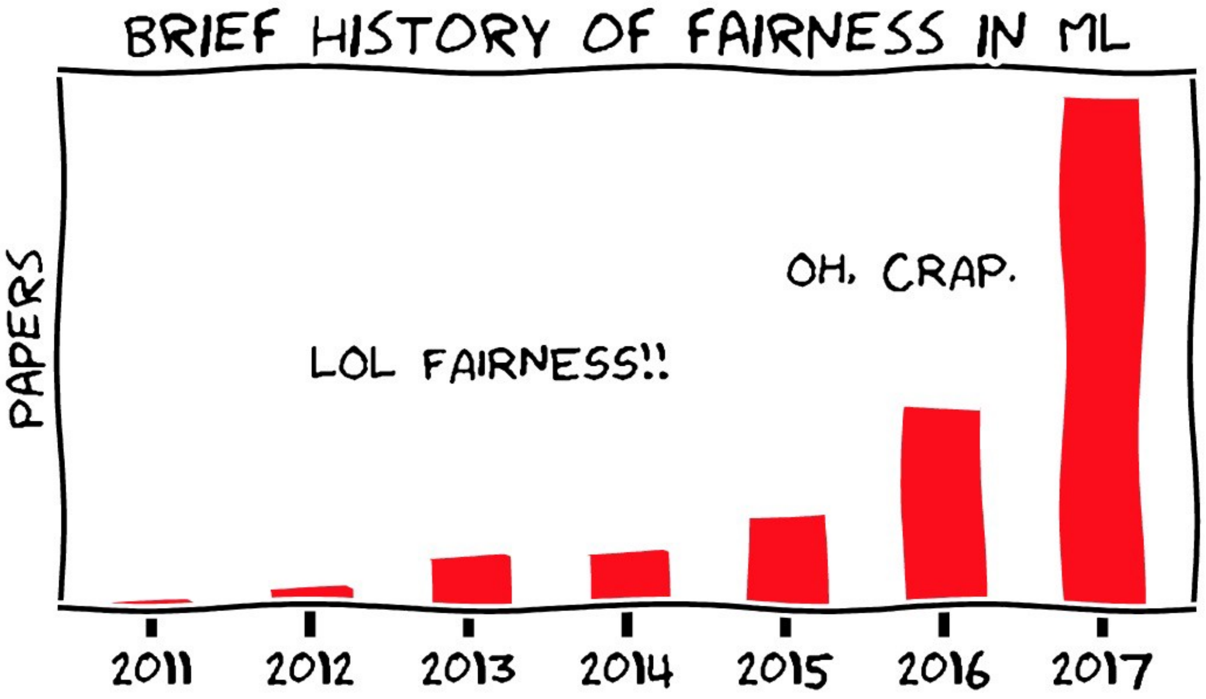
Dilbert (2)


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...
1/9/00
© 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!
11/13/00
© 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 preservation
- Fairness
- Streaming data (e.g., intrusion detection)


BRIEF HISTORY OF FAIRNESS IN ML
PAPERS
LOL FAIRNESS!!
OH. CRAP.