Data Mining & Machine Learning

Introduction

Gergely Lukács

Pázmány Péter Catholic University

Faculty of Information Technology and Bionics

Budapest, Hungary

lukacs@itk.ppke.hu

Contents

- Course information
- Motivation
- What is it?
  - Definition
  - Related fields
- Application areas
- Task types
- Application areas
- Ethics

Course information

Course information

- Moodle: Slides, tests, additional information, supplementary material
- MS Teams: presentations, chat (communication)

- Free datacamp.com access: coming

Requirements, grading

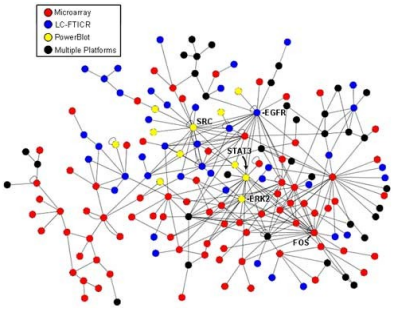
- **(Short) Moodle tests** every weak in the lab: ca 5 minutes, ca 3 questions.
- **Midterm test:**
  - During regular lab (lecture) time
    15th December?.
    If failed: replacement test (single)
  - Correction deadline: 17/12
  - Midterm replacement (21/12, 8 am?): only for those who failed regular midterm
- **Assigment !!!**
- Eligibility for „signature” (all needed)
  - Lab attendance > 80%
  - >50% of the short tests
  - >50 % midterm test
  - >50 % assignment

Requirements, grading 2

- Final exam
  - written (>50% !) + oral:
  - 40% is based on your performance during the semester (assignment: 20%, midterm test: 20%). 60% is based on your performance on the final exam.
  - last week in exam period: no first exam allowed
- Grading
  >= 50% satisfactory (2)
  >= 63% average (3)
  >= 75% good (4)
  >= 87% excellent (5)

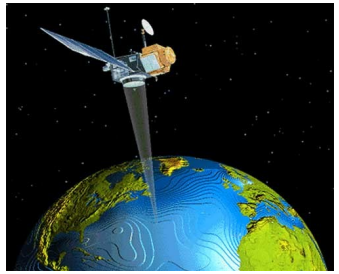
Data Mining: Motivation

Purchases



Biological data

Financial transactions



Satellites

Remote sensing

Telecommunication

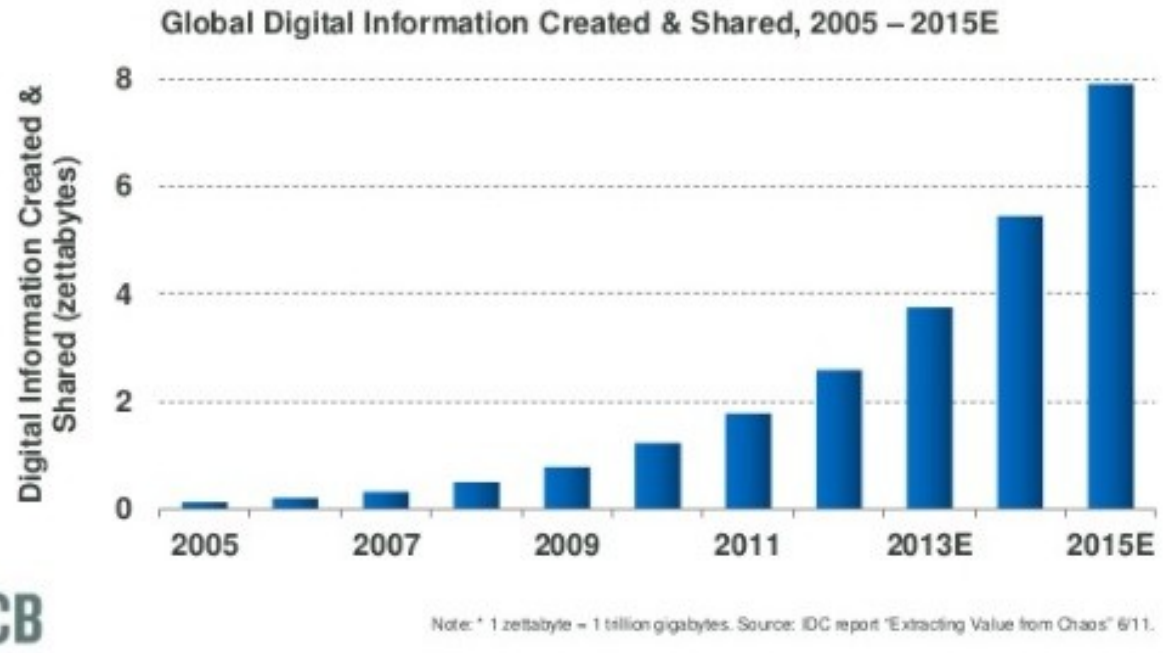
Sensors

WWW

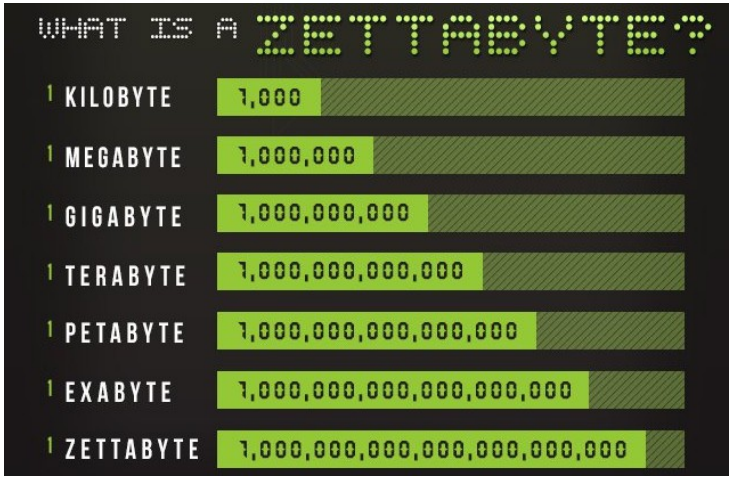
Media

Social media

Amount of digital data produced



Amount of data doubles every 20 months

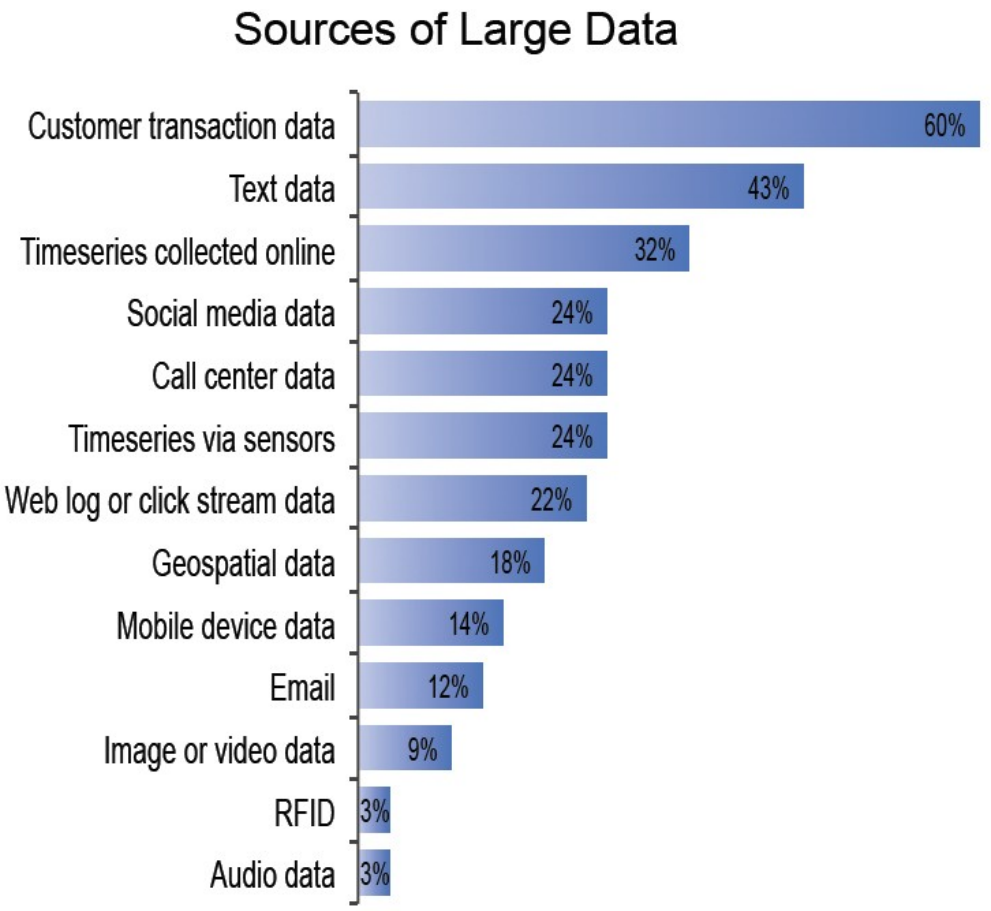


Data from: 2013

Customer Transactions: #1 Source of Large Data

Customer transactional data often affords the opportunity for a wide range of analytics due to the depth and scope of available data.

Among respondents who reported increases in data volume, 60% identified customer transaction data as a source of their large data sets.



Question: What are the sources of data for your large datasets? (select all that apply)

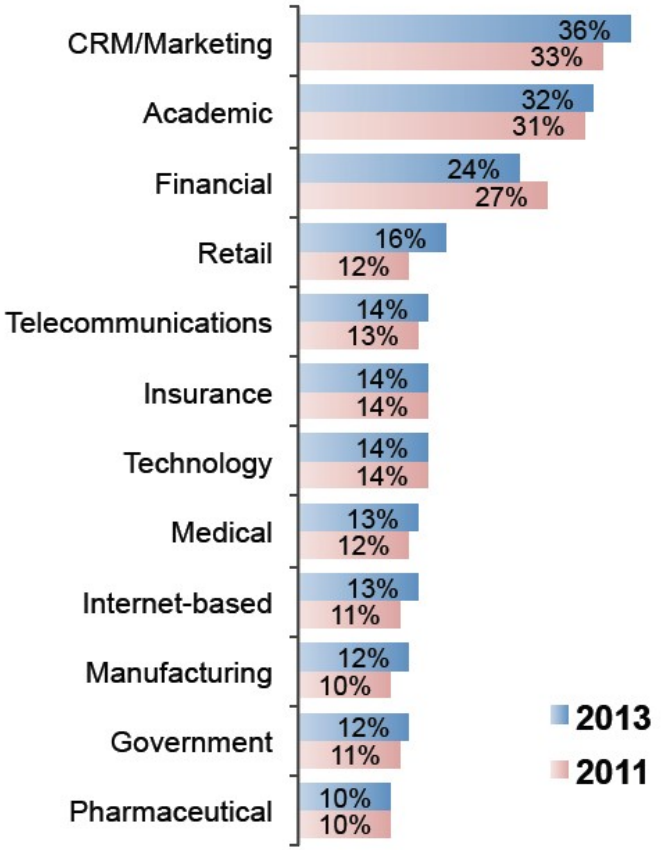
9

Data from: 2013

CRM / Marketing: #1 Place for Data Miners

CRM / Marketing remains the #1 area to which data mining is applied.

The roots of data mining in customer focused analytics are strong. In each of the 6 Data Miner Surveys, more people report applying their analytics in the field of CRM / Marketing than any other field. In 2013, 36% of data miners indicated that they are commonly involved in CRM / Marketing data mining, up slightly from 2011. The number of data miners working in the overlapping area of Retail analytics is also increasing.



Data miners also report working in Non-profit (5%), Hospitality / Entertainment / Sports (4%), Military / Security (2%), and Other (10%).

Question: In what fields do you TYPICALLY apply data mining? (Select all that apply)

Data Mining: Knowledge Discovery in Databases

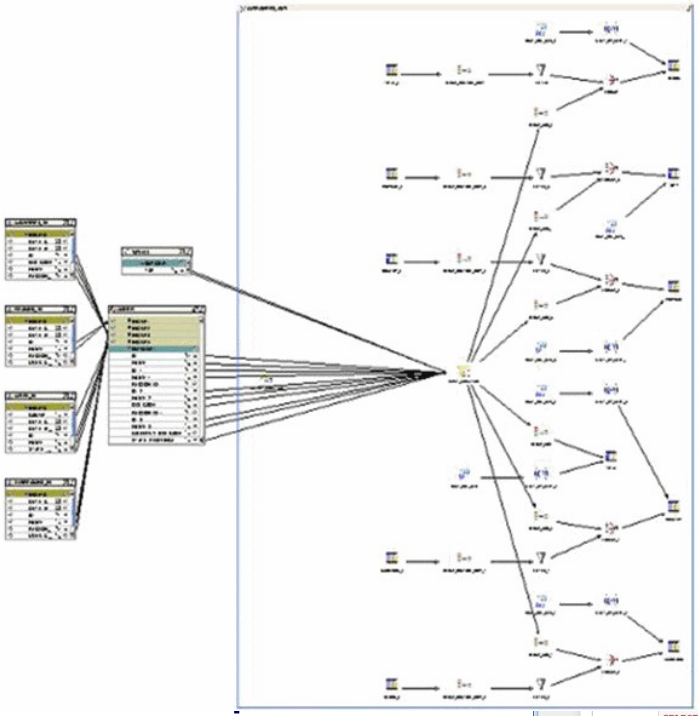
Knowledge Discovery is the non-trivial **process** of identifying **valid, novel, potentially useful**, and ultimately **understandable patterns** in **data**" [Osama Fayyad et al., 1996]

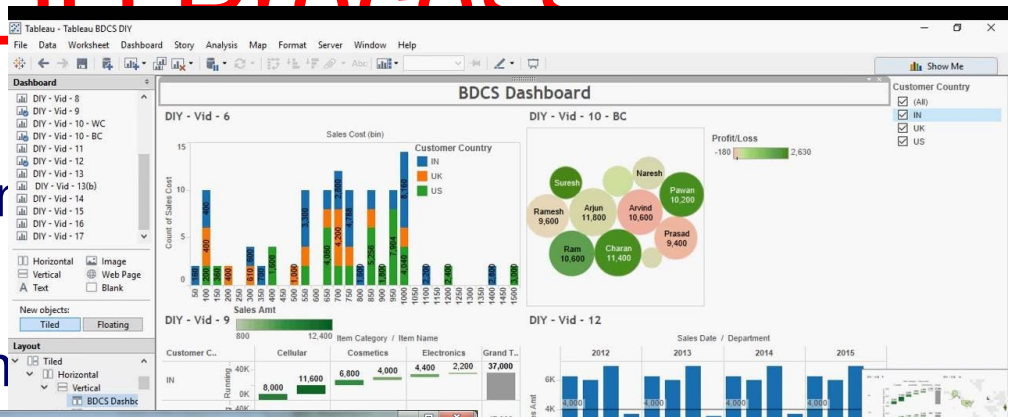
- **Data**: set of facts (e.g. instances in a DB).
- **Pattern**: An expression in a language describing facts in (a subset of) the data.
- **Process**: a multi-step process involving data preparation (selection, preprocessing, transformation), pattern searching, knowledge evaluation, and refinement with iteration after modification.
- **Valid**: Discovered patterns should be true on new data with some degree of certainty. Generalize to the future (other data).
- **Novel**: Patterns must be novel (should not be previously known).
- **Useful**: Actionable; patterns should potentially lead to some useful actions.
- **Understandable**: The process should lead to human insight. Patterns must be made understandable in order to facilitate a better understanding of the underlying data.

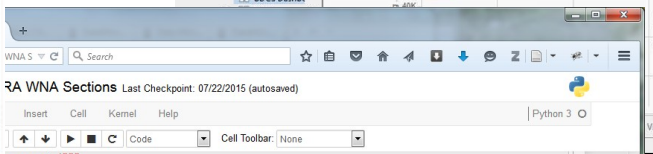
e KDD Process

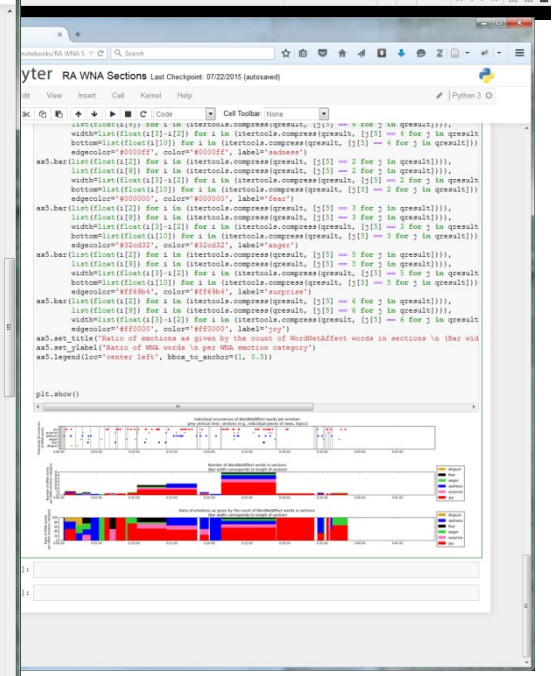
domair

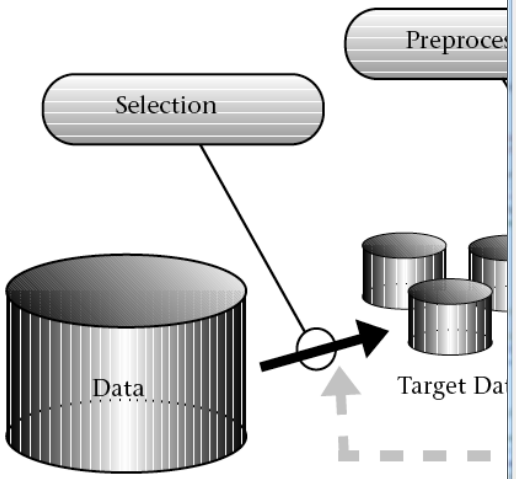
eansin





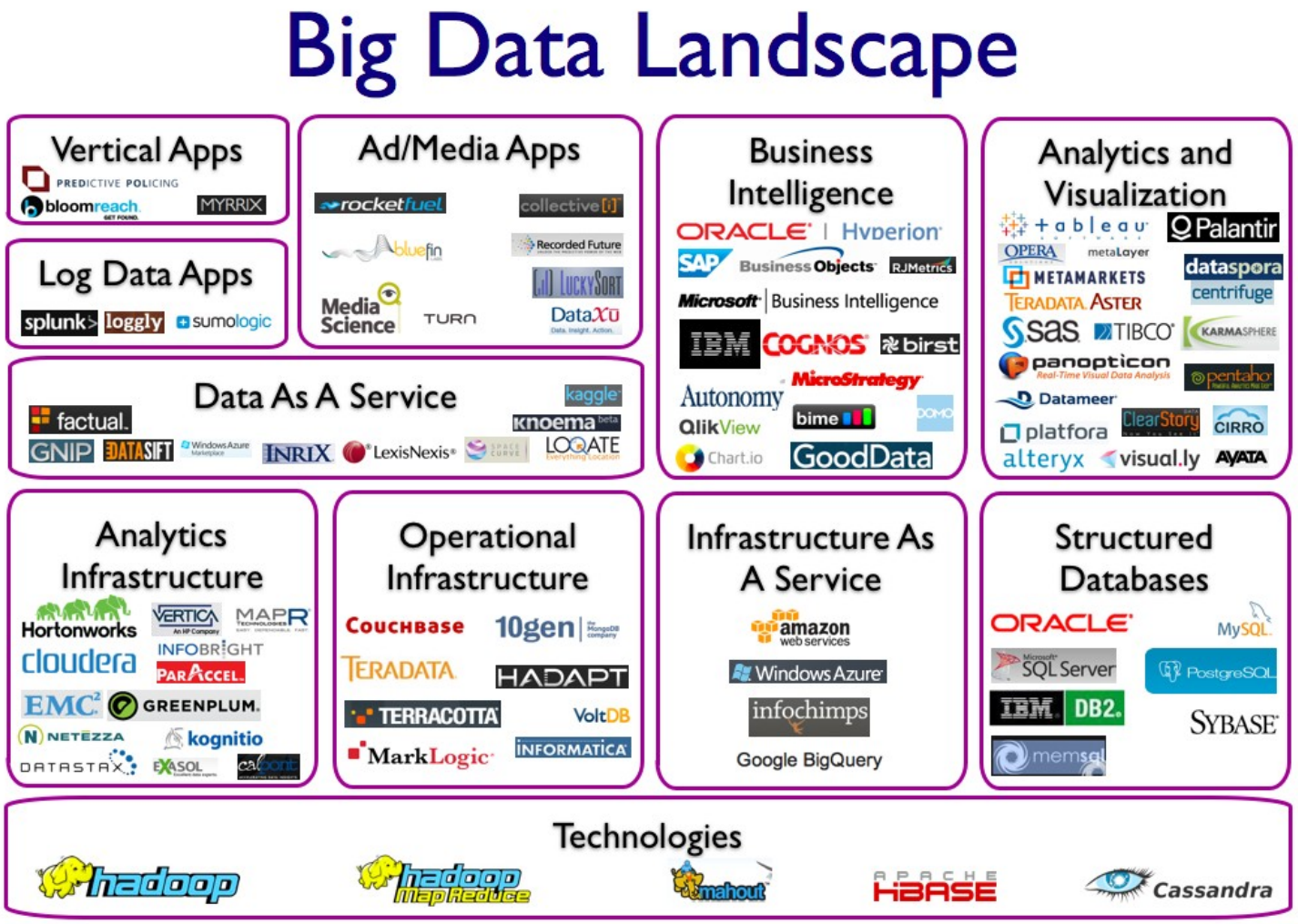







```
r.execute("""
SELECT section_id,
       block_id,
       section_beginning,
       section_ending,
       section_middle,
       emotion_code,
       count_emotion,
       Sum(count_emotion)
         OVER (
           partition BY section_id, block_id, section_beginning, section_ending,
         section_middle)                                                    AS
       count_allemotions,
       COALESCE(Sum(count_emotion)
                  OVER (
                    partition BY section_id, block_id, section_beginning,
                  section_ending,
                  section_middle
                    ORDER BY emotion_code DESC rows BETWEEN UNBOUNDED PRECEDING
                  AND 1
                  PRECEDING), 0)                                            AS
       countcum_emotion,
       Round(100 * count_emotion / Sum(count_emotion)
                                     OVER (
                                       partition BY section_id, block_id,
                                     section_beginning,
                                     section_ending,
                                     section_middle))                       AS
       ratio_emotions,
       COALESCE(Round(100 * ( Sum(count_emotion)
                                OVER (
                                  partition BY section_id, block_id,
                                section_beginning,
                                section_ending
                                ,
                                section_middle
                                  ORDER BY emotion_code DESC rows BETWEEN
                                UNBOUNDED
                                PRECEDING
                                AND 1
                                PRECEDING) ) / ( Sum(count_emotion)
                           OVER (
                             partition BY section_id, block_id,
                           section_beginning, section_ending,
                           section_middle) )), 0) AS
       ratiocum_emotions
FROM   (SELECT section_id,
               block_id,
               section_beginning,
               section_ending,
               section_middle,
               emotion_code,
               Count(*) AS count_emotion
        FROM   (SELECT section_id,
                       block_id,
                       section_beginning,
                       section_ending,
                       Round(( section_beginning + section_ending ) / 2, 2) AS
```

Data science, big data



dave@vcdave.com

blogs.forbes.com/davefeinleib

Goals of Data Mining

- **Prediction** future or unknown values supporting, automating, improving decision making
- **Description -- white box –**
  - human-interpretable patterns
  - gain „insights”

Machine learning

- Machine learning is the science of getting computers to learn, without being explicitely programmed (Arthur Samuel 1959)
- Well-posed learning problem: a computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

Machine learning„Software 2.0”

- Applications can’t program by hand
  - Voice recognition
  - Handwriting recognition
  - NLP
  - Computer vision

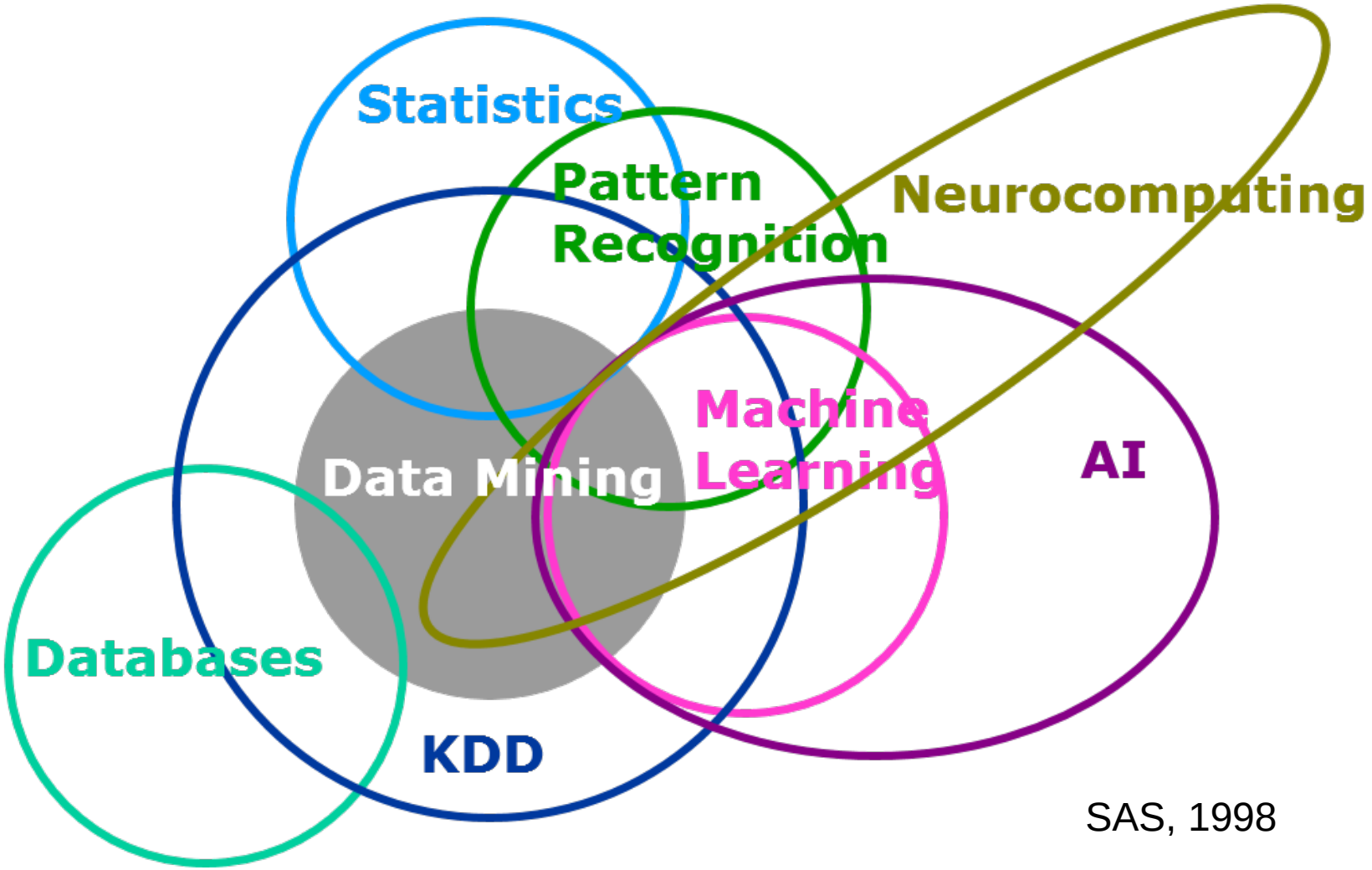
  - Software 1.0: explicitely write code
  - Software 2.0: get dataset (!) and train a ML model

Data Science Workflow

- „Data science”
- Data Science Workflow (different definitions; here: Blitzstein & Pfister)
  1. Ask an interesting question
  2. Get the data
  3. Explore the data
  4. Model the data
  5. Communicate and visualize the results

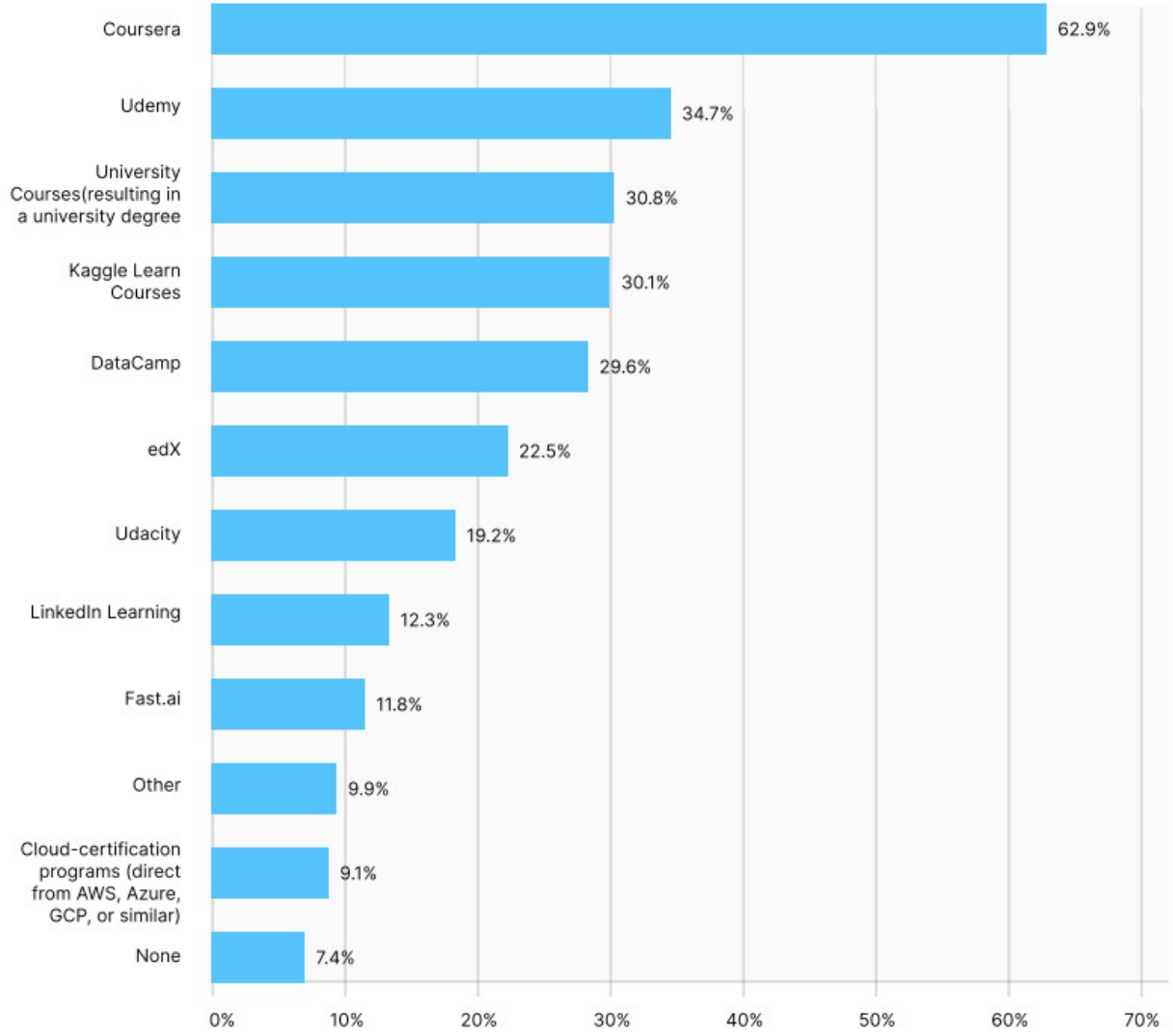
Origins, related disciplines?

- Perspectives converging…
- DM (compared to ML)
  - Basis: Large (structured, relational) database
  - Efficiency and scalability are important
  - Often focuses on incomplete / dirty real world data
  - Domain knowledge may be given
  - Description – understandable patterns
- Data mining vs. **statistics**
  - Statistics: mathematically well founded, traditional discipline,
  - DM focuses on the goal;
    Even incorrect statistical assumptions can be OK, if it works!


Statistics
Pattern Recognition
Neurocomputing
Machine Learning
AI
Data Mining
Databases
KDD

SAS, 1998

POPULAR ONGOING LEARNING RESOURCES



kaggle

kaggle

Home
Compete
Data
Notebooks
Discuss
Courses
Jobs
More

Search

Our Titanic Competition is a great first challenge to get started.

Titanic: Machine Learning from Disaster — Knowledge
Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started • Ongoing • 19088 Teams

All Competitions

Active | Completed | InClass | All Categories | Default Sort

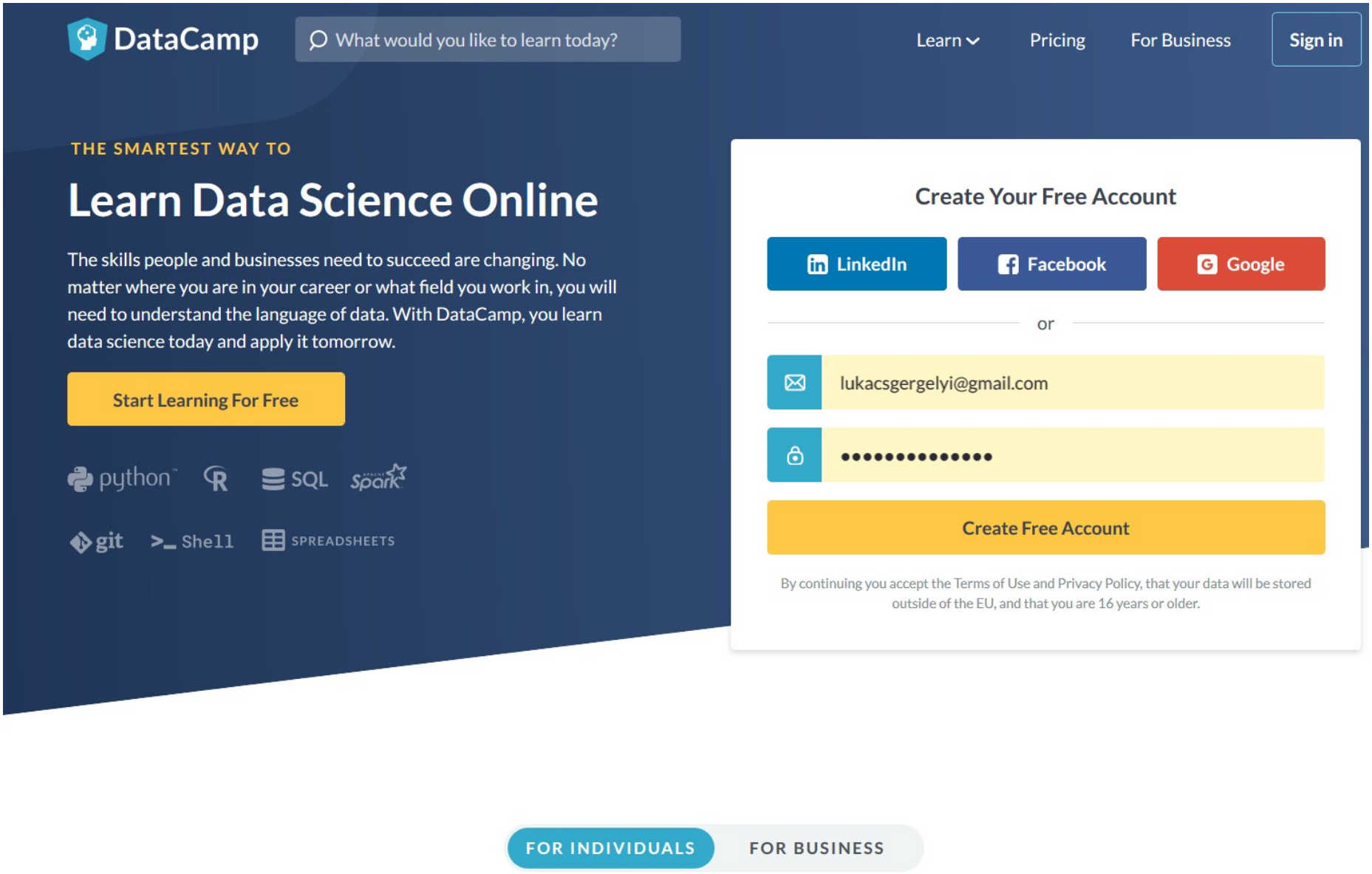
OSIC Pulmonary Fibrosis Progression	Predict lung function decline Featured • a month to go • Code Competition • 1532 Teams	$55,000
Lyft Motion Prediction for Autonomous Vehicles	Build motion prediction models for self-driving vehicles Featured • 3 months to go • Code Competition • 320 Teams	$30,000
Mechanisms of Action (MoA) Prediction	Can you improve the algorithm that classifies drugs based on their biological activity? Research • 3 months to go • Code Competition • 654 Teams	$30,000
Cornell Birdcall Identification	Build tools for bird population monitoring Research • 8 days to go • Code Competition • 1302 Teams	$25,000
Google Landmark Recognition 2020	Label famous (and not-so-famous) landmarks in images Research • 22 days to go • Code Competition • 555 Teams	$25,000
Halite by Two Sigma	Collect the most halite during your match in space Featured • 8 days to go • Simulation Competition • 1124 Teams	Swag
Conway's Reverse Game of Life 2020	Reverse the arrow of time in the Game of Life Playground • 3 months to go • Code Competition • 41 Teams	Swag
Predict Future Sales	Final project for "How to win a data science competition" Coursera course Playground • 4 months to go • 8592 Teams	Kudos
Titanic: Machine Learning from Disaster	Start here! Predict survival on the Titanic and get familiar with ML basics Getting Started • Ongoing • 19088 Teams	Knowledge

kaggle Search kaggle

Kaggle is the place to do dat
science projects

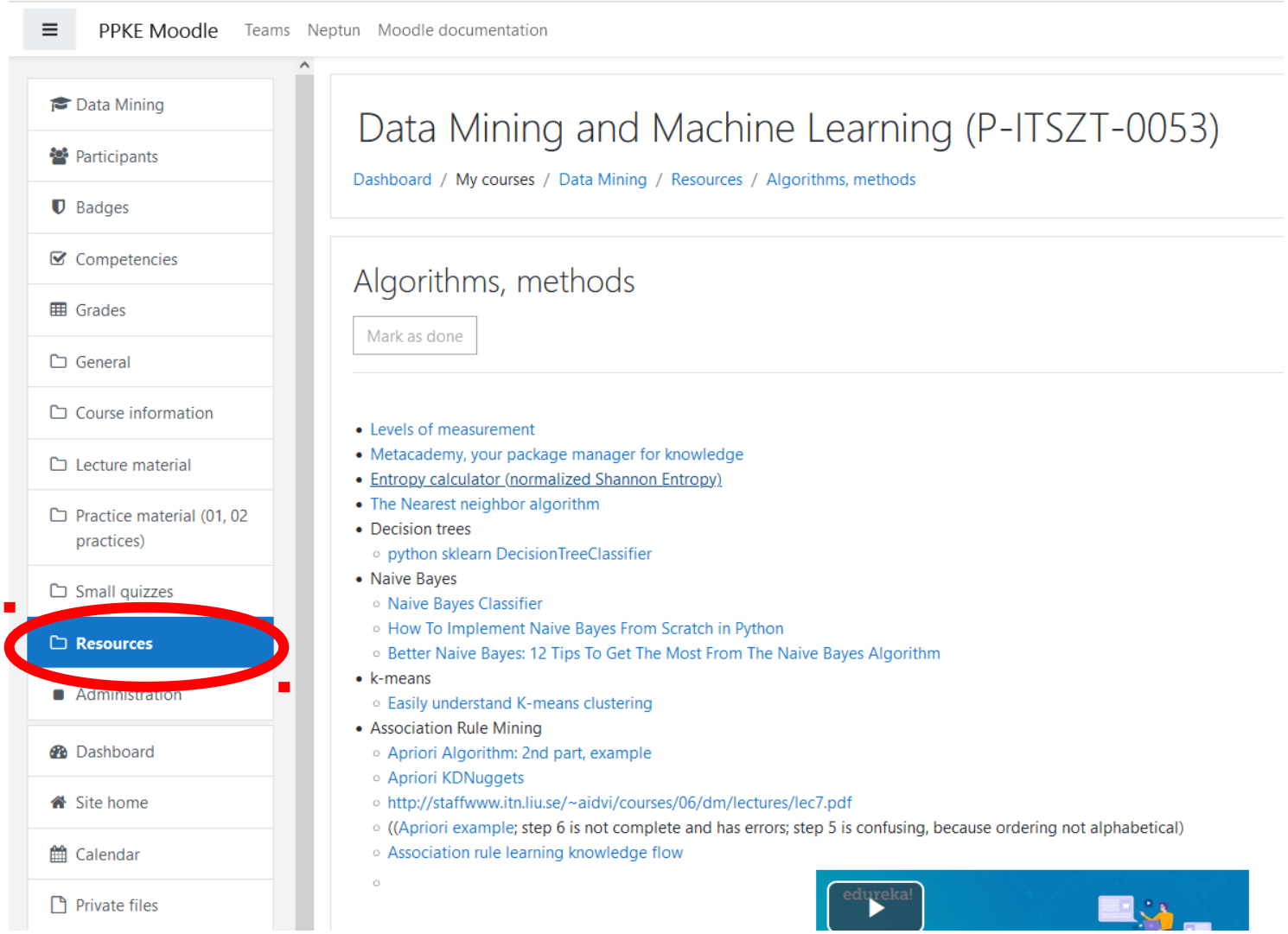
See how it works

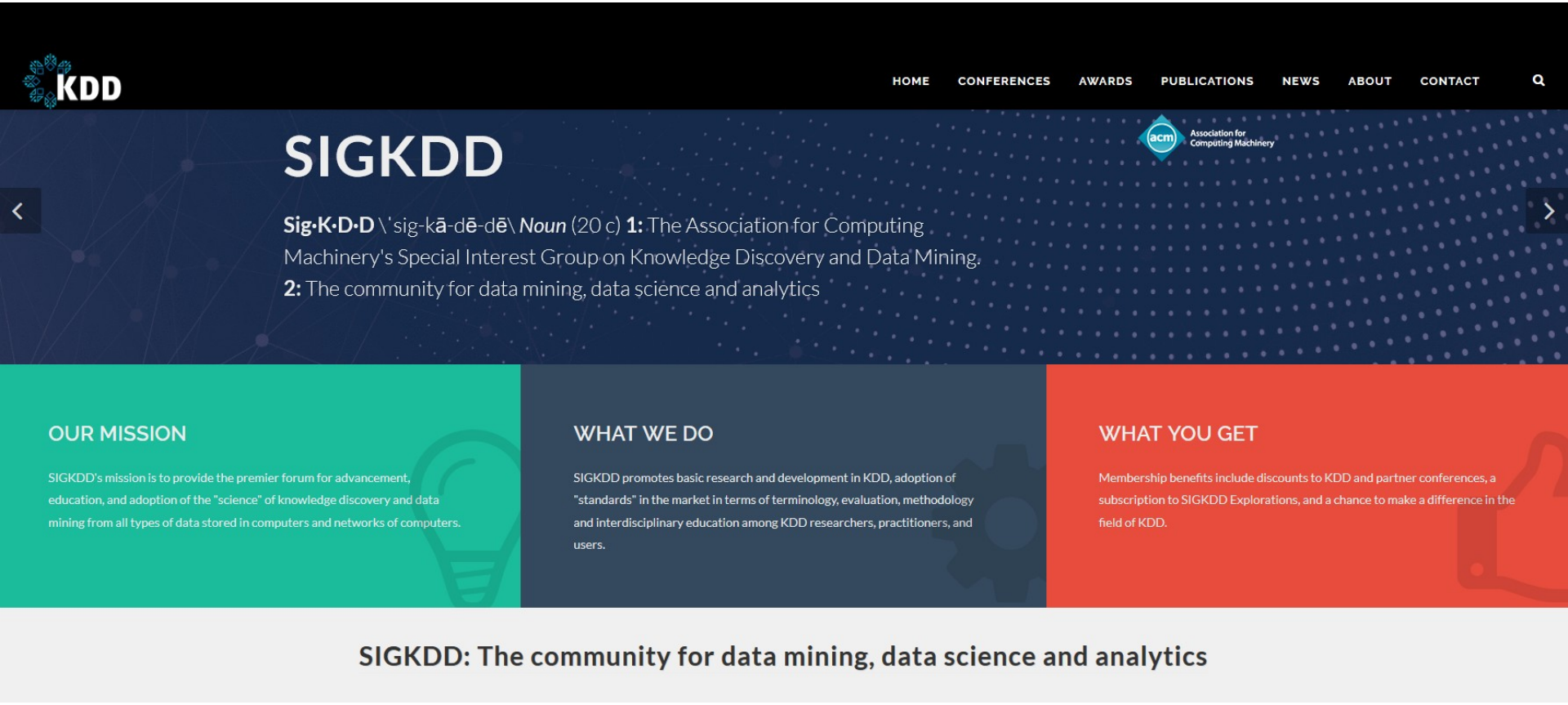
Start a new project


DataCamp
What would you like to learn today?
Learn
Pricing
For Business
Sign in
THE SMARTEST WAY TO
Learn Data Science Online
The skills people and businesses need to succeed are changing. No matter where you are in your career or what field you work in, you will need to understand the language of data. With DataCamp, you learn data science today and apply it tomorrow.
Start Learning For Free
python
SQL
Spark
git
Shell
SPREADSHEETS
Create Your Free Account
LinkedIn
Facebook
Google
or
lukacsgergelyi@gmail.com
Create Free Account
By continuing you accept the Terms of Use and Privacy Policy, that your data will be stored outside of the EU, and that you are 16 years or older.
FOR INDIVIDUALS
FOR BUSINESS

Moodle, resources

(currently on old homepage, ITK wiki;
will be copied iteratively to moodle))

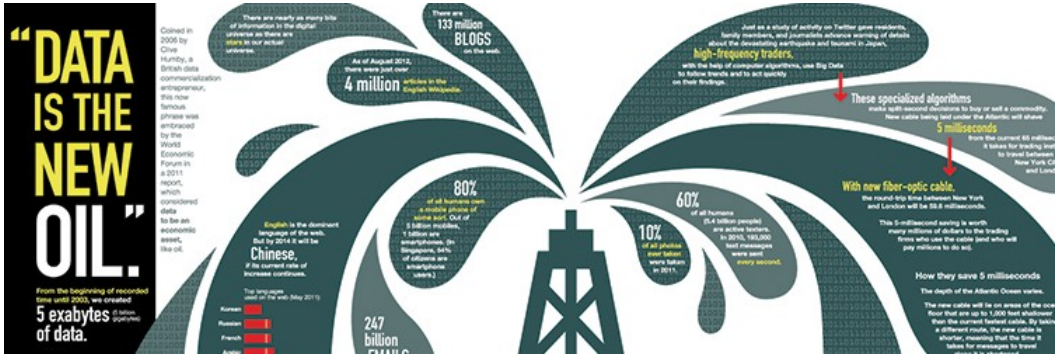


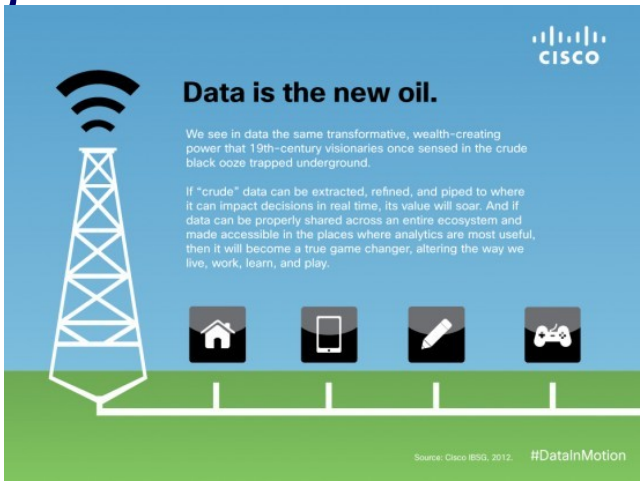

KDD
HOME
CONFERENCES
AWARDS
PUBLICATIONS
NEWS
ABOUT
CONTACT
SIGKDD
Association for Computing Machinery
Sig·K·D·D \ˈsig-kā-dē-dē\ Noun (20 c) 1: The Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining.
2: The community for data mining, data science and analytics
OUR MISSION
SIGKDD's mission is to provide the premier forum for advancement, education, and adoption of the "science" of knowledge discovery and data mining from all types of data stored in computers and networks of computers.
WHAT WE DO
SIGKDD promotes basic research and development in KDD, adoption of "standards" in the market in terms of terminology, evaluation, methodology and interdisciplinary education among KDD researchers, practitioners, and users.
WHAT YOU GET
Membership benefits include discounts to KDD and partner conferences, a subscription to SIGKDD Explorations, and a chance to make a difference in the field of KDD.
SIGKDD: The community for data mining, data science and analytics

Application areas (examples!)

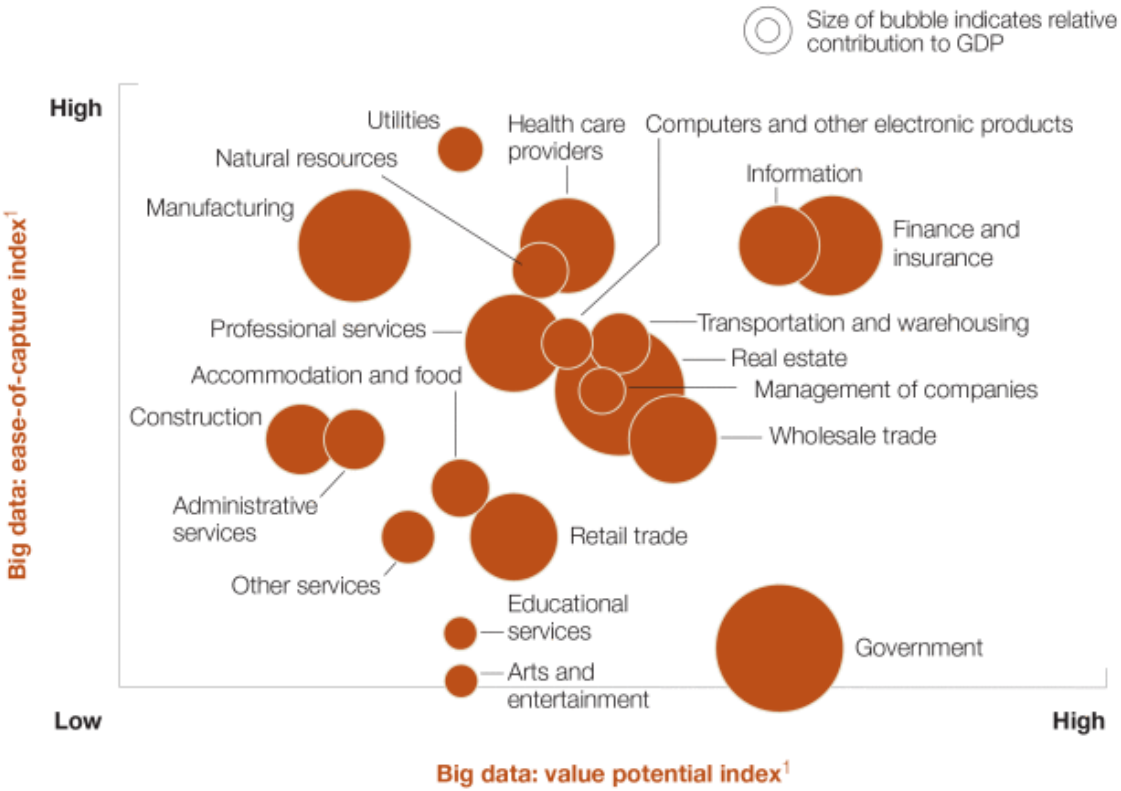
Why Data Is The New Oil, Fortune, 11.07.2016.

„Artificial intelligence is only as good as the data it crunches.”





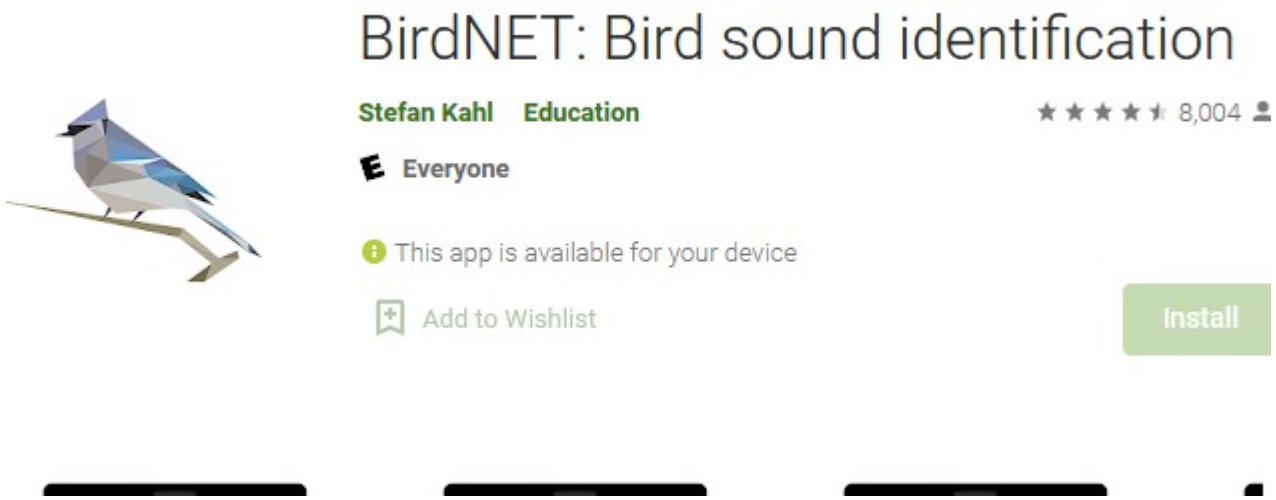
Potential in Big Data



[1]For detailed explication of metrics, see appendix in McKinsey Global Institute full report *Big data: The next frontier for innovation, competition, and productivity*, available free of charge online at mckinsey.com/mgi.

Source: US Bureau of Labor Statistics; McKinsey Global Institute analysis

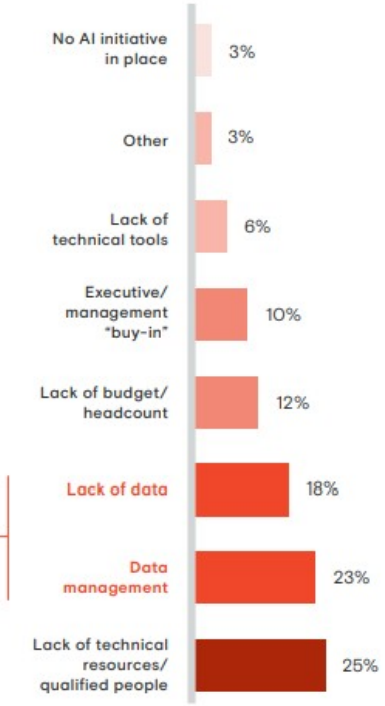
Precision agriculture


BirdNET: Bird sound identification
Stefan Kahl
Education
8,004
Everyone
This app is available for your device
Add to Wishlist
Install

Data is the Bottleneck!

Companies are now reporting that they are updating their models more frequently. We found that, of those who update at least quarterly, 40% report that lack of data or data management were of the biggest roadblocks to AI success.

Figure 11: What do you consider the biggest bottleneck to any of your AI initiatives or project?



Participants surveyed: 173 - update models quarterly or more often
Single select

appen

Most Time-Consuming Tasks for Data Scientists

When asked to cite up to two of their most time-consuming tasks, the highest percentage of respondents (66.7 percent) said "cleaning and organizing data." This finding correlates strongly with the most common obstacle respondents reported earlier in the survey—"too much time spent cleaning data," cited by 57.5 percent of respondents.

"Collecting data sets" was cited by 52.9 percent of respondents as one of their most time-consuming tasks, coming in a close second to cleaning and organizing data. Offloading these time-devouring tasks from data scientists' plates represents a significant opportunity for companies to gain efficiencies and give data scientists more time for the strategic work they also actually enjoy doing.

66.7% said cleaning and organizing data is one of their most time-consuming tasks

52.9% said collecting data sets is one of their most time-consuming tasks

30.7 % listed mining for patterns in data among their most time-consuming tasks

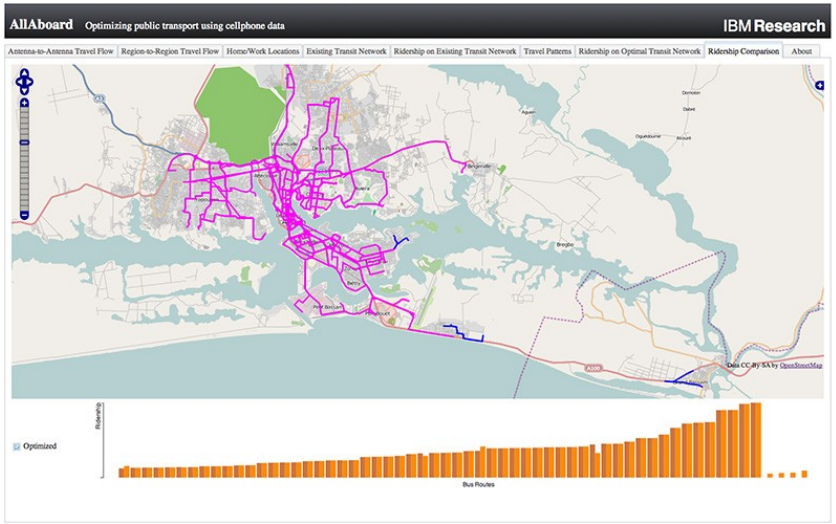
1 Accenture, "The Team Solution to the Data Scientist Shortage," 2013
2 McKinsey Global Institute, "Big Data: The Next Frontier for Innovation, Competition, and Productivity," 2011

CrowdFlower
http://www.crowdflower.com

7

Data for Development Challenge

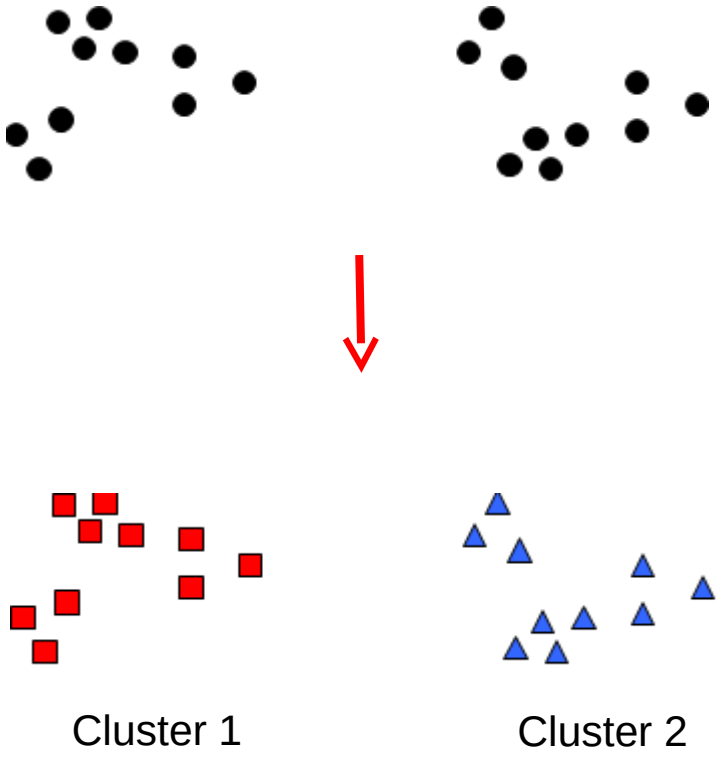
- Orange (mobile phone), Ivory Coast
- Four datasets:
  - Aggregate communication between cell towers;
  - Mobility traces: fine resolution dataset;
  - Mobility traces: coarse resolution dataset;
  - Communication sub-graphs
- Results
  - anticipating epidemics
  - reacting during times of crisis
  - optimizing the use of certain infrastructures
  - designing new services to meet the needs of populations



Classes of Tasks

Clustering

- Divide a set of objects (data instances) into subsets – clusters -- , based on similarity
  - Meaning of subsets not defined
  - Number of subsets?
  - Similarity ?
- **Unsupervised learning**
  - **Data not labelled**
- **Focuses on data**



Many possible solutions…

How many clusters?

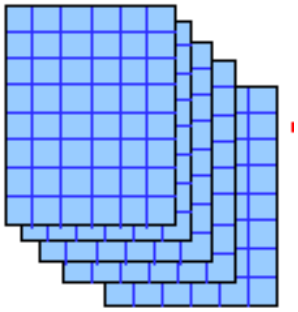
Two Clusters

Four Clusters

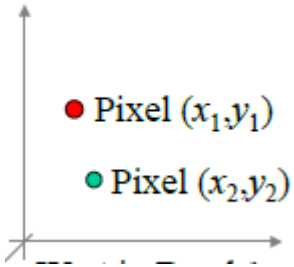
Six Clusters

Example: Satellite data processing

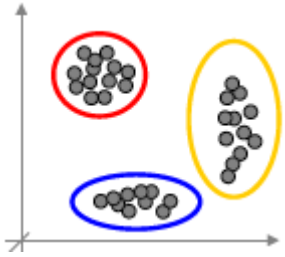
Photos in 5
different bands



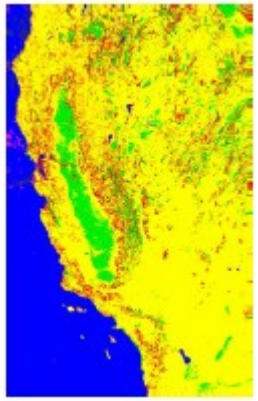
Pixels in the space
of the bands (dimensions)



Clustering

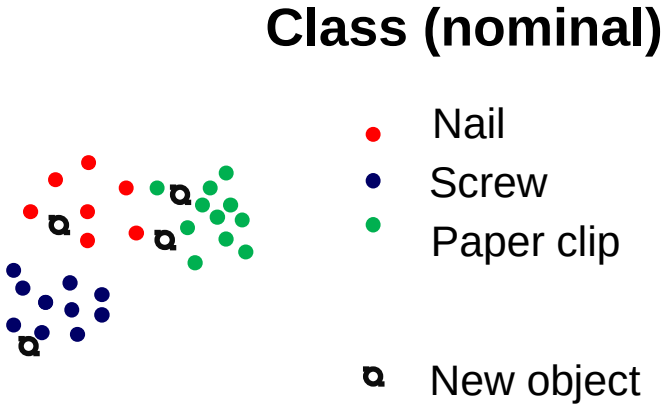


Transformation to original coordinates
Coloring by clusters



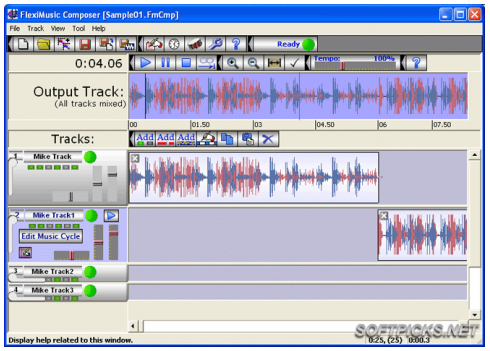
Classification

- Given: set of classified -- labelled -- data;
- Class attribute: nominal (categories!)
- Goal: learn patterns allowing to classify (label) unclassified data
- Binary / multiclass
- **Supervised learning**
- **Labelled data used**
  - **Labelled data used**
- **Prediction (task) driven**

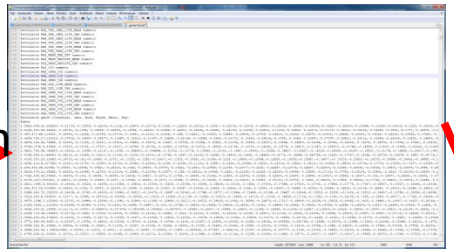


Music instrument recognition

Labelled training data
(known instrument)



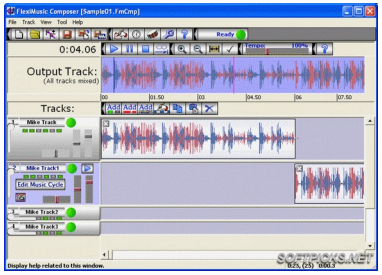
Feature extraction

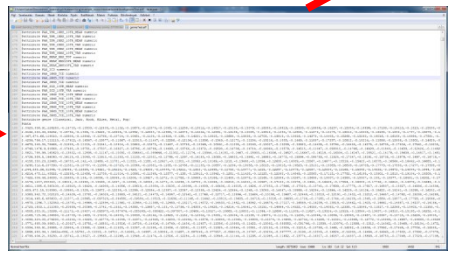


Learning

„Patterns”

Unlabelled data
(instrument unknown)





Predicting

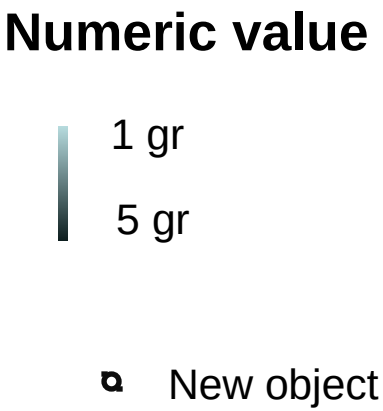
Predictions

Diagnosis of machine faults

- Diagnosis is a classical domain of expert systems
- Given: Fourier analysis of vibrations measured at various points of a device's mounting
- Problem: which fault is present?
- Preventative maintenance of electromechanical motors and generators
- Information very noisy
- So far: diagnosis by expert/hand-crafted rules
- Available: 600 faults with expert's diagnosis
  - Learned rules outperformed hand-crafted ones

Numeric prediction („Regression”)

- Similar to classification;
- Class attribute: **numeric**, instead of nominal
- Goal: learn patterns allowing to order a numeric value to unclassified data
- **Supervised learning**

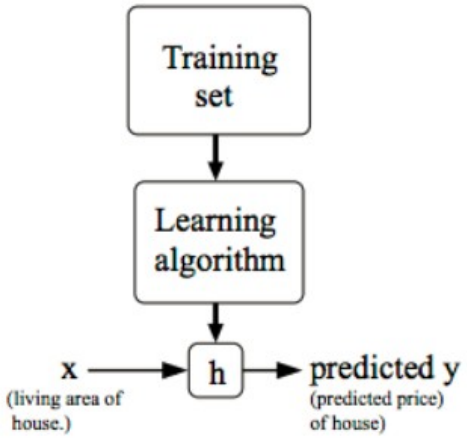


Load forecasting

- Electricity supply companies require forecast of future demand for power
  - Accurate forecasts of minimum and maximum load for each hour result in significant savings
- Given: manually constructed static load model
  - base load for the year
  - load periodicity over the year
  - effect of holidays
- Problem: adjusting for weather conditions
- Prediction corrected using “most similar” days
  - temperature,
  - humidity,
  - wind speed
  - cloud cover

Supervised learning (classification, regression)

- Input values: independent variables
- Output value: dependent variable
- Learning/training: Inferring a function from training data
- Predicting: for unknown data



Association rule mining

a,**b,c,d,**e
b,c,d
a,**b,c,d**
a,**b,c,d,**e
a,c,e,f
c,d,e,f
a,**b,c,d,**f

b,c,d co-occur

in 71% of the cases

If b and c then d
in 100% of the cases

a	b	c	d	e	f
1	1	1	1	1	0
0	1	1	1	0	0
1	1	1	1	0	0
1	1	1	1	1	0
1	0	1	0	1	1
0	0	1	1	1	1
1	1	1	1	0	1

Find all rules in the form:

If a and b and c occurs in the set,
then t is also part of the set with a probability of > x %

Data unlabelled (no selected class attribute)

- Typically: Boolean attributes (true/false)

Unsupervised learning

Application example

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

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

Reinforcement learning (not covered in this course)

- (software) agent learns to perform certain actions in an environment which lead it to maximum reward
  - Sequence of actions
  - Trial and error
  - Maximizes reward
- Applications
  - Self driving cars, trajectory optimization
  - Trading, finance
  - ……

Types of data mining analysis

- Unsupervised
  - Clustering
    - discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data
    - similar to classification, but classes are not known ahead of time
  - Association rule mining
    - Searches for relationships between variables.
    - e.g., which products are frequently bought together (72% of customers who bought cookies also bought milk…)
- Supervised
  - Classification
    - e.g., Is a new customer applying for a loan a good investment or not?

    if STATUS = married and INCOME > 50K and HOUSE_OWNER = yes

    then INVESTMENT_TYPE = good
    - e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
  - Numeric prediction/„regression")
  - Attempts to find a function which models the data with the least error
- Reinforcement Learning

Data mining and ethics

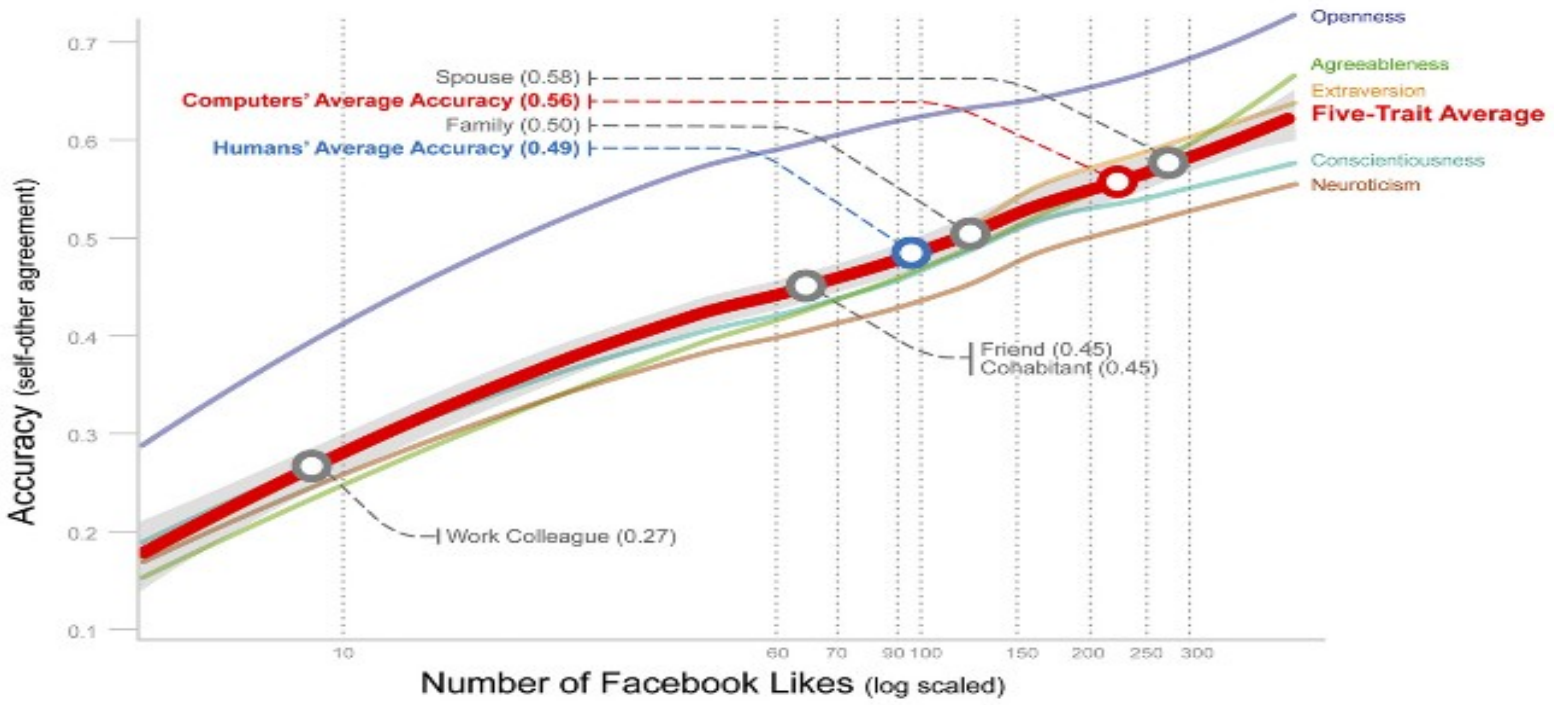
Data mining and ethics

- Important questions in practical applications:
  - For what purpose was the data collected?
  - Who is permitted access to the data?
  - What kind of conclusions can be legitimately drawn from it?
- Ethical situation depends on application
  - E.g. same information OK in medical application, not OK for discrimination (e.g., loan applications … sex, religion, race)

Data mining and ethics 2

- Privacy, data ownership, access to data
- GDPR vs USA
- Reidentificitaion
  - postal code, sex, date of birth:
    85% of population in USA can be identified!
  - human mobility
    - location of an individual is specified hourly
    - spatial resolution equal to that given by the carrier's antennas,
    - four spatio-temporal points
    - 95% of the individuals can be uniquely identified

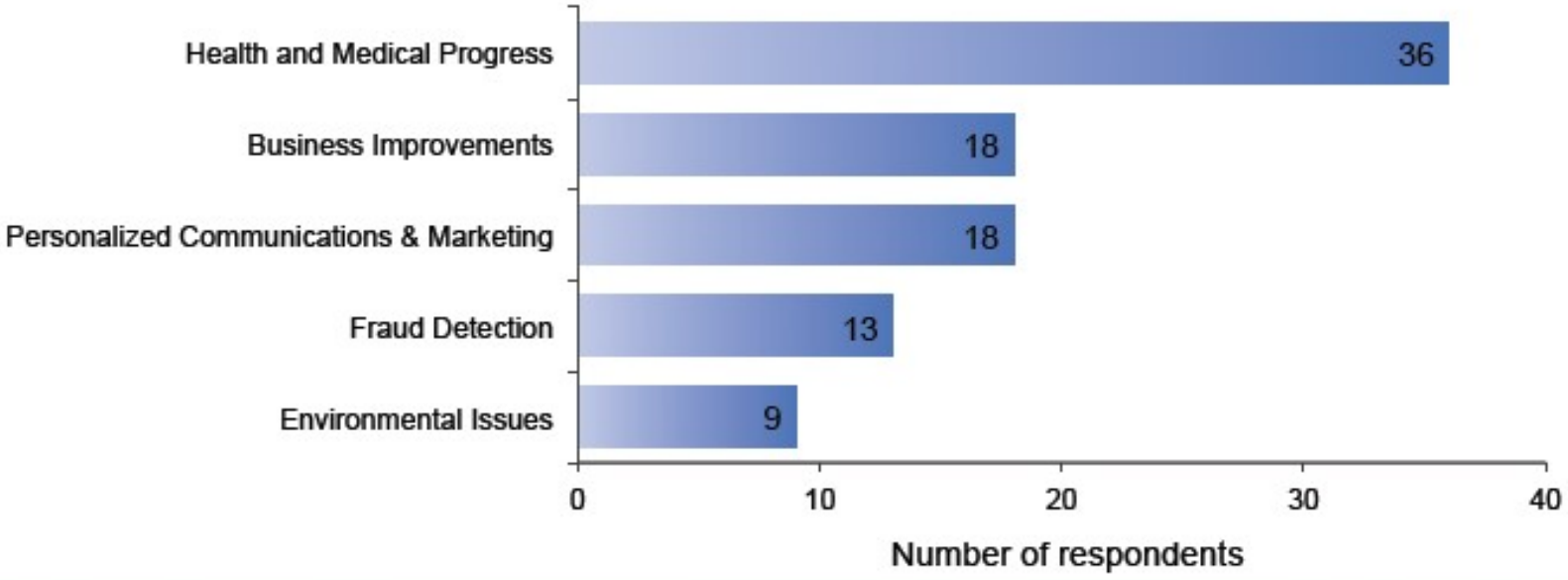
Computer-based personality judgments are more accurate than those made by humans http://www.pnas.org/content/112/4/1036.full.pdf+html




NATE SILVER ON WHAT OBAMA SHOULD DO NEXT, P. 44
DON'T MOCK THE ARTISANAL PICKLE MAKERS, P. 14
A NANNY'S VIEW OF THE WORLD, P. 47
MANAGEMENT TIPS FROM 'DOWNTON ABBEY,' P. 52
GREECE CONFRONTS ITS SPARTAN FUTURE, P. 38
The New York Times Magazine
February 19, 2012
How your shopping habits reveal even the most personal information. By Charles Duhigg

Positive Impact of Data Mining

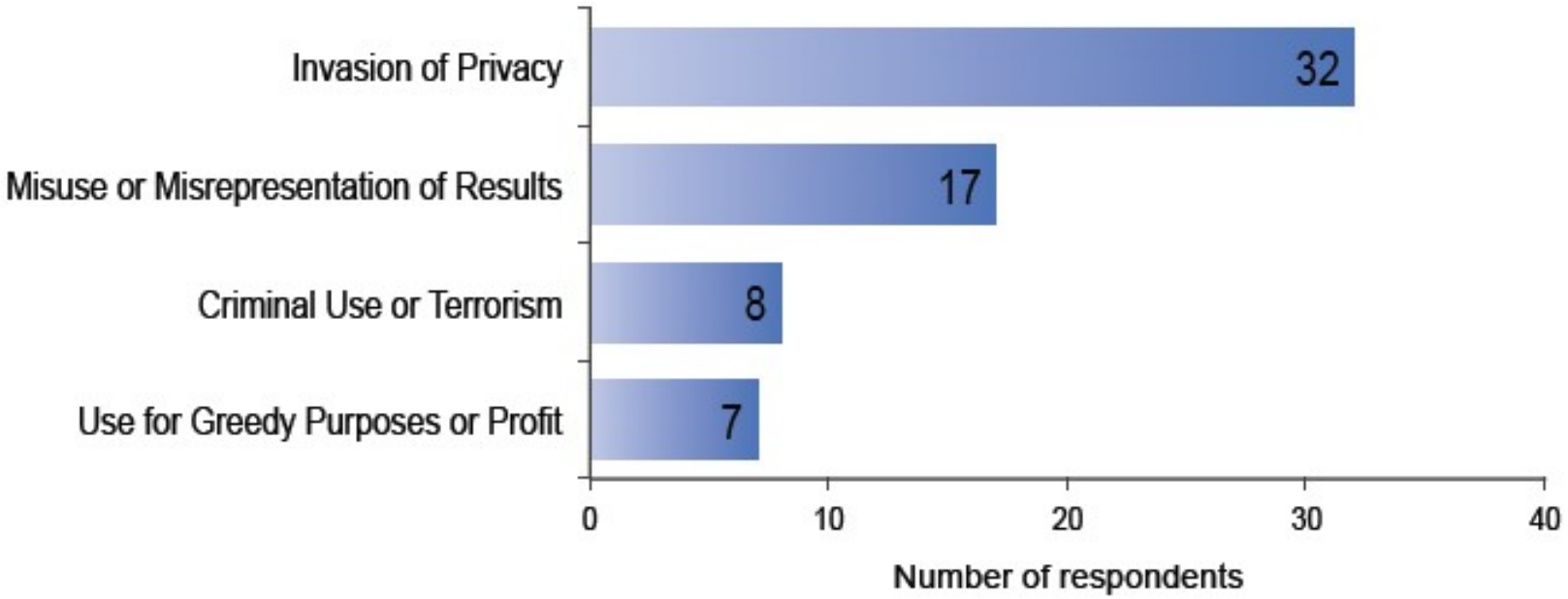
- Survey respondents shared their ideas about the positive impact of data mining on society (an open-ended survey question).
- The largest number of respondents identified positive impacts on our health and progress in medical fields.
- For a complete list of respondents' ideas about the positive impact of data mining, see www.RexerAnalytics.com/DMSurvey2011_PositiveImpact.



Question: Please share with us the best examples you know of that highlight the positive impact that data mining can have to benefit society, health, the world, etc. (text box provided for response)

Negative Impact of Data Mining

- Survey respondents also shared their ideas about the negative impact of data mining on society (an open-ended survey question).
- The largest number of respondents were concerned about the invasion of privacy that can sometimes accompany data mining.



Question: Please share with us the worst examples you know of that highlight a negative use of data mining. (text box provided for response)

Summary

Summary

- Definition of DM, ML
- KDD process
- Related fields
- Main goals of DM
- Major classes of DM tasks
  - Unsupervised
    - Clustering
    - Association rule mining
  - Supervised
    - Classification
    - Numeric prediction
  - Reinforcement learning