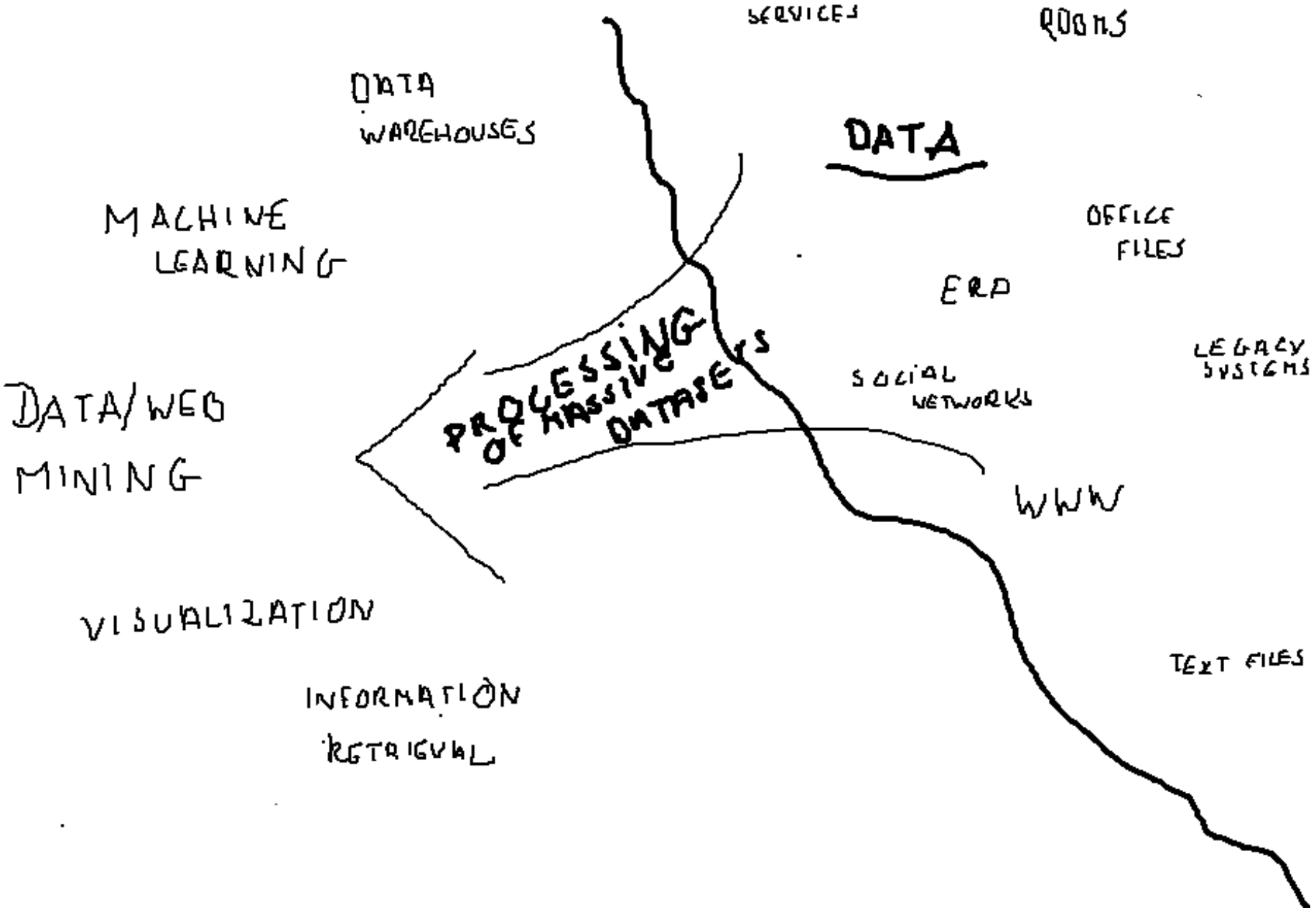
Mining of Massive Datasets

Krzysztof Dembczyński

Intelligent Decision Support Systems Laboratory (IDSS)
Poznań University of Technology, Poland

Software Development Technologies
Master studies, second semester
Academic year 2018/19 (winter course)

Goal: understanding data . . .


SERVICES
ROOMS
DATA WAREHOUSES
DATA
MACHINE LEARNING
OFFICE FILES
ERP
LEGACY SYSTEMS
SOCIAL NETWORKS
DATA/WEB MINING
PROCESSING OF MASSIVE DATABASES
WWW
VISUALIZATION
TEXT FILES
INFORMATION RETRIEVAL

Goal: . . . to make data analysis efficient.

- **Buzzwords**: Big Data, Data Science, Machine learning, NoSQL ...

- **Buzzwords**: Big Data, Data Science, Machine learning, NoSQL ...
- **How Big Data Changes Everything:**
  - Several books showing the impact of Big Data revolution (e.g., **Disruptive Possibilities: How Big Data Changes Everything** by Jeffrey Needham).

- **Buzzwords**: Big Data, Data Science, Machine learning, NoSQL ...
- **How Big Data Changes Everything:**
  - Several books showing the impact of Big Data revolution (e.g., **Disruptive Possibilities: How Big Data Changes Everything** by Jeffrey Needham).
- **Computerworld (Jul 11, 2007)**:
  - *12 IT skills that employers can't say no to:*
    1) Machine learning
    ...

- **Buzzwords**: Big Data, Data Science, Machine learning, NoSQL ...
- **How Big Data Changes Everything:**
  - Several books showing the impact of Big Data revolution (e.g., **Disruptive Possibilities: How Big Data Changes Everything** by Jeffrey Needham).
- **Computerworld (Jul 11, 2007)**:
  - *12 IT skills that employers can't say no to:*
    1) Machine learning
    ...
- **Three priorities of Google announced at BoxDev 2015**:
  - Machine learning – speech recognition
  - Machine learning – image understanding
  - Machine learning – preference learning/personalization

- **Buzzwords**: Big Data, Data Science, Machine learning, NoSQL ...
- **How Big Data Changes Everything:**
  - Several books showing the impact of Big Data revolution (e.g., **Disruptive Possibilities: How Big Data Changes Everything** by Jeffrey Needham).
- **Computerworld (Jul 11, 2007)**:
  - *12 IT skills that employers can't say no to:*
    1) Machine learning
    ...
- **Three priorities of Google announced at BoxDev 2015**:
  - Machine learning – speech recognition
  - Machine learning – image understanding
  - Machine learning – preference learning/personalization
- **OpenAI** founded in 2015 as a non-profit artificial intelligence research company.

Data mining

- Data mining is the discovery of **models** for data, ...
- But what is a model?

if all you have is a hammer, everything looks like a nail

How to characterize your data?

- **Database programmer** usually writes:

```
select avg(column), std(column) from data
```

How to characterize your data?

- **Database programmer** usually writes:

```
select avg(column), std(column) from data
```

- **Statistician** might decide that the data comes from a Gaussian distribution and use a formula to compute the most likely parameters of this Gaussian: the mean and standard deviation.

How to characterize your data?

- **Database programmer** usually writes:

```
select avg(column), std(column) from data
```

- **Statistician** might decide that the data comes from a Gaussian distribution and use a formula to compute the most likely parameters of this Gaussian: the mean and standard deviation.
- **Machine learner** will use the data as training examples and apply a learning algorithm to get a model that predicts future data.

How to characterize your data?

- **Database programmer** usually writes:

```
select avg(column), std(column) from data
```

- **Statistician** might decide that the data comes from a Gaussian distribution and use a formula to compute the most likely parameters of this Gaussian: the mean and standard deviation.
- **Machine learner** will use the data as training examples and apply a learning algorithm to get a model that predicts future data.
- **Data miner** will discover the most frequent patterns.

They all want to understand data and use this knowledge for making better decisions

Data+ideas vs. statistics+algorithms

- About the Amazon's recommender system:

  It's often more important to creatively invent new data sources than to implement the latest academic variations on an algorithm.

Data+ideas vs. statistics+algorithms

- About the Amazon's recommender system:

  > *It's often more important to creatively invent new data sources than to implement the latest academic variations on an algorithm.*

- WhizBang! Labs tried to use machine learning to locate people's resumes on the Web: the algorithm was not able to do better than procedures designed by hand, since a resume has a quite standard shape and sentences.

Data+computational power

- Object recognition in computer vision:

Data+computational power

- Object recognition in computer vision:
  - Scanning large databases **can perform better** than the best computer vision algorithms!

Data+computational power

- Object recognition in computer vision:
  - Scanning large databases **can perform better** than the best computer vision algorithms!
- Automatic translation

Data+computational power

- Object recognition in computer vision:
  - Scanning large databases **can perform better** than the best computer vision algorithms!
- Automatic translation
  - Statistical translation based on large corpora **outperforms** linguistic models!

Human computation

- CAPTCHA and reCAPTCHA
- ESP game
- Check a lecture given by Luis von Ahn:
  `http://videolectures.net/iaai09_vonahn_hc/`
- Amazon Mechanical Turk

Data+ideas vs. statistics+algorithms

Those who ignore Statistics are condemned to reinvent it.

Brad Efron

- In Statistics, a term **data mining** was originally referring to attempts to extract information that was not supported by the data.
- Bonferroni's Principle: "if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap".
- Rhine paradox.

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).
  - Pattern finding: association rules.

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).
  - Pattern finding: association rules.
  - Netflix: recommender system.

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).
  - Pattern finding: association rules.
  - Netflix: recommender system.
  - Google and PageRank.

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).
  - Pattern finding: association rules.
  - Netflix: recommender system.
  - Google and PageRank.
  - Clustering of Cholera cases in 1854.

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).
  - Pattern finding: association rules.
  - Netflix: recommender system.
  - Google and PageRank.
  - Clustering of Cholera cases in 1854.
  - Win one of the Kaggle's competitions!!! `http://www.kaggle.com/`.

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).
  - Pattern finding: association rules.
  - Netflix: recommender system.
  - Google and PageRank.
  - Clustering of Cholera cases in 1854.
  - Win one of the Kaggle's competitions!!! `http://www.kaggle.com/`.
  - Autonomous cars.

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).
  - Pattern finding: association rules.
  - Netflix: recommender system.
  - Google and PageRank.
  - Clustering of Cholera cases in 1854.
  - Win one of the Kaggle's competitions!!! `http://www.kaggle.com/`.
  - Autonomous cars.
  - Deep learning.

Data+ideas vs. statistics+algorithms

- The data mining algorithms can perform quite well!!!
  - XBox Kinect: object tracking vs. pattern recognition (check: `http://videolectures.net/ecmlpkdd2011_bishop_embracing/`).
  - Pattern finding: association rules.
  - Netflix: recommender system.
  - Google and PageRank.
  - Clustering of Cholera cases in 1854.
  - Win one of the Kaggle's competitions!!! `http://www.kaggle.com/`.
  - Autonomous cars.
  - Deep learning.
  - And many others.

Data+ideas+computational power+statistics+algorithms

To be learned in the upcoming semester ...

The aim and the scope of the course

- **Aim**: To get to know the latest technologies and algorithms for mining of massive datasets.

The aim and the scope of the course

- **Aim**: To get to know the latest technologies and algorithms for mining of massive datasets.
- **Scope**: We will learn about scalable algorithms for:

The aim and the scope of the course

- **Aim**: To get to know the latest technologies and algorithms for mining of massive datasets.
- **Scope**: We will learn about scalable algorithms for:
  - Classification and regression,

The aim and the scope of the course

- **Aim**: To get to know the latest technologies and algorithms for mining of massive datasets.
- **Scope**: We will learn about scalable algorithms for:
  - Classification and regression,
  - Searching for similar items,

The aim and the scope of the course

- **Aim**: To get to know the latest technologies and algorithms for mining of massive datasets.
- **Scope**: We will learn about scalable algorithms for:
  - Classification and regression,
  - Searching for similar items,
  - And recommender systems.

The aim and the scope of the course

- **Aim**: To get to know the latest technologies and algorithms for mining of massive datasets.
- **Scope**: We will learn about scalable algorithms for:
  - Classification and regression,
  - Searching for similar items,
  - And recommender systems.
- The course is mainly based on the **Mining of Massive Datasets** book: `http://www.mmds.org/`

Main information about the course

- Instructors:
  - dr hab. inż. Krzysztof Dembczyński (kdembczynskicsputpoznanpl)
- Website:
  - www.cs.put.poznan.pl/kdembczynski/lectures/mmds
- Time and place:
  - Lecture: Thursday 13:30, room L125 BT.
  - Labs: Wednesday, 13:30 and 16:50, room 45 CW.
  - Office hours: Thursday, 10:00-12:00, room 2 CW.

Lectures

- Main topics of lectures:
  - Introduction
  - Classification and regression (x6)
  - Evolution of database systems and Spark (x3)
  - Finding similar items (x3)

Labs

- Strong connection between lectures and labs.
- Software: Python and Spark.
- List of tasks and exercises for each week (also homeworks).
- Mainly mini programming projects and short exercises.
- Main topics:
  - Bonferroni's principle (x2)
  - Classification and Regression (Python) (x6)
  - MapReduce (Spark) (x2)
  - Finding similar items (x2)

Evaluation

- **Lecture**:

  Test: 75 % (min. 50%)
  Labs: 25 % (min. 50%)

- **Labs**:

  Regular exercises and home works: 100 % (min. 50%)

- **Scale**:

90 %	–	5.0	80 %	–	4.5	70 %	–	4.0
60 %	–	3.5	50 %	–	3.0	$<$ 50 %	–	2.0

- **Bonus points for all**: up to 10 percent points.

Bibliography

- A. Rajaraman and J. D. Ullman. *Mining of Massive Datasets*.
  Cambridge University Press, 2011
  `http://www.mmds.org`

- H. Garcia-Molina, J. D. Ullman, and J. Widom. *Database Systems: The Complete Book. Second Edition*.
  Pearson Prentice Hall, 2009

- J.Lin and Ch. Dyer. *Data-Intensive Text Processing with MapReduce*.
  Morgan and Claypool Publishers, 2010
  `http://lintool.github.com/MapReduceAlgorithms/`

- T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning: Second Edition*.
  Springer, 2009
  `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`

- Ch. Lam. *Hadoop in Action*.
  Manning Publications Co., 2011