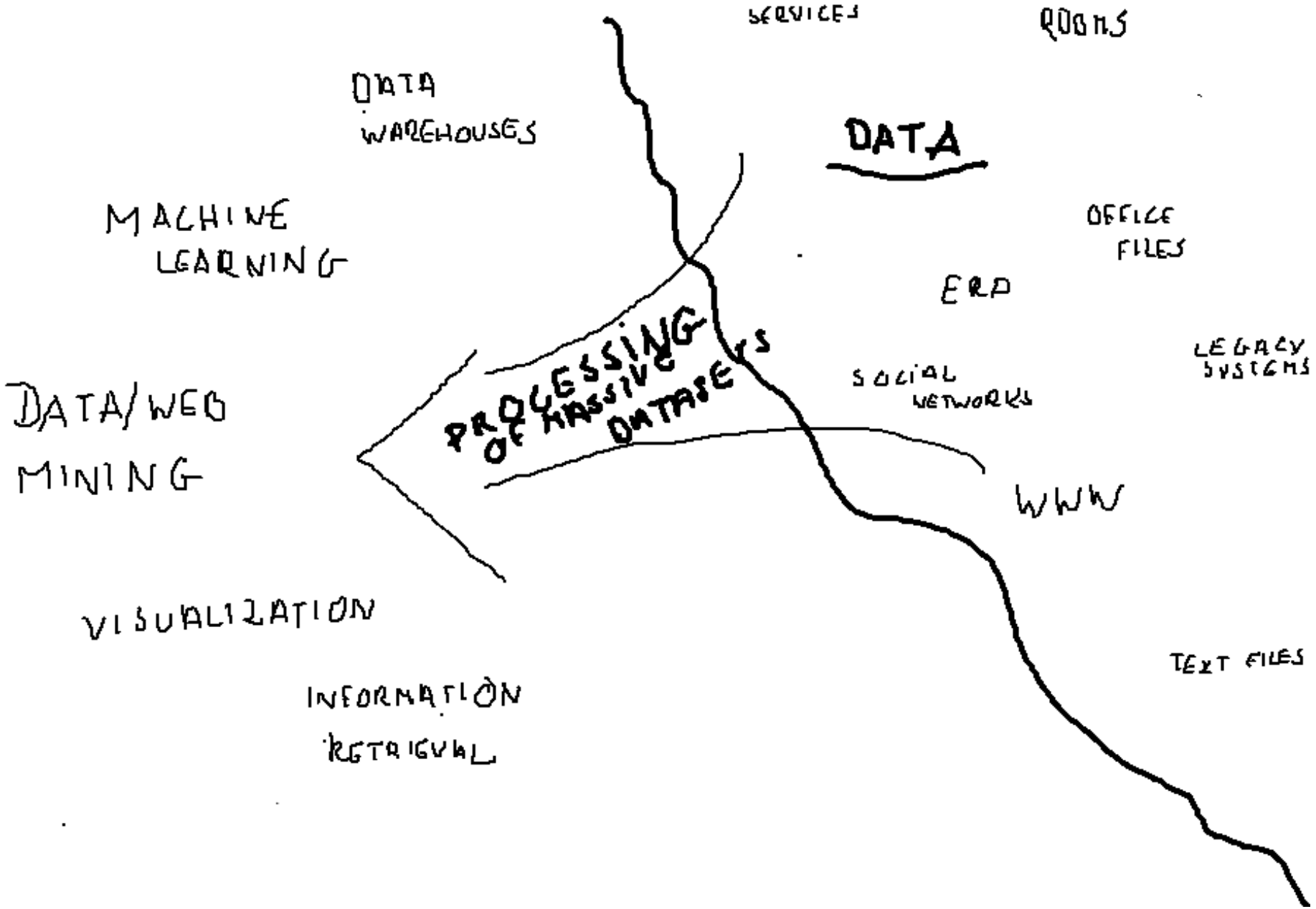
# Processing of Massive Datasets

Krzysztof Dembczyński

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


Bachelor studies, seventh semester
Academic year 2018/19 (winter semester)

**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 technologies and algorithms for processing massive datasets.

# The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:

# The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:
  - Write-once-read-many-times type of data stores,

# The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:
  - Write-once-read-many-times type of data stores,
  - Data warehouses and star schemas,

# The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:
  - Write-once-read-many-times type of data stores,
  - Data warehouses and star schemas,
  - Data structures and fast algorithms for processing massive datasets,

# The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:
  - Write-once-read-many-times type of data stores,
  - Data warehouses and star schemas,
  - Data structures and fast algorithms for processing massive datasets,
  - Approximate query processing,

# The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:
  - Write-once-read-many-times type of data stores,
  - Data warehouses and star schemas,
  - Data structures and fast algorithms for processing massive datasets,
  - Approximate query processing,
  - Nearest neighbor search,

# The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:
  - Write-once-read-many-times type of data stores,
  - Data warehouses and star schemas,
  - Data structures and fast algorithms for processing massive datasets,
  - Approximate query processing,
  - Nearest neighbor search,
  - Data streams,

# The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:
  - Write-once-read-many-times type of data stores,
  - Data warehouses and star schemas,
  - Data structures and fast algorithms for processing massive datasets,
  - Approximate query processing,
  - Nearest neighbor search,
  - Data streams,
  - NoSQL and MapReduce technologies.

# The aim and the scope of the course

- **Aim**: To get to know technologies and algorithms for processing massive datasets.
- **Scope**: We will learn how to organize, store, access, and process massive datasets:
  - Write-once-read-many-times type of data stores,
  - Data warehouses and star schemas,
  - Data structures and fast algorithms for processing massive datasets,
  - Approximate query processing,
  - Nearest neighbor search,
  - Data streams,
  - NoSQL and MapReduce technologies.
- The course is based on parts of the **Mining of Massive Datasets** book: `http://www.mmds.org/`

# Main information about the course

- Instructors:
  - dr hab. inż. Krzysztof Dembczyński (kdembczynskicsputpoznanpl)
  - mgr inż. Kalina Jasinska (kjasinskacsputpoznanpl)
  - mgr inż. Marek Wydmuch (mwydmuchcsputpoznanpl)

# Main information about the course

- Instructors:
  - dr hab. inż. Krzysztof Dembczyński (kdembczynskicsputpoznanpl)
  - mgr inż. Kalina Jasinska (kjasinskacsputpoznanpl)
  - mgr inż. Marek Wydmuch (mwydmuchcsputpoznanpl)
- Website:
  - www.cs.put.poznan.pl/kdembczynski/lectures/pmds

# Main information about the course

- Instructors:
  - dr hab. inż. Krzysztof Dembczyński (kdembczynskicsputpoznanpl)
  - mgr inż. Kalina Jasinska (kjasinskacsputpoznanpl)
  - mgr inż. Marek Wydmuch (mwydmuchcsputpoznanpl)
- Website:
  - www.cs.put.poznan.pl/kdembczynski/lectures/pmds
- Time and place:
  - Lecture: Monday 13:30, lecture room 2 CW.
  - Labs:
    - Monday: 9:45 (lab 45, MW), 15:10 (lab 44, MW), 15:10 (lab 45, KD)
    - Tuesday: 9:45 (lab 143, KJ), 15:10 (lab 45, KJ)
    - Thursday: 15:10 (lab 45, KD)
  - Office hours: Thursday, 10:00-12:00, room 2 CW (Institute of Computing Science).

# Lectures

- Main topics of lectures:
  - Introduction
  - Evolution of database systems and data models
  - Data warehouses, star schema, dimensional modeling, ETL
  - Data structures and fast algorithms for processing massive datasets
  - Data streams and approximate query processing
  - Nearest neighbor search
  - NoSQL and MapReduce

# Labs

- Strong connection between lectures and labs.
- Software: programming language of your choice, bash, Spark (Python, Java, Scala).
- List of tasks and exercises for each week (also homeworks).
- Mainly mini programming projects and short exercises.
- Main topics:
  - Bonferroni's Principle
  - Data modeling/ETL case study
  - Multidimensional modeling
  - Exact and approximate query processing
  - Finding similar items
  - MapReduce in Spark

# Evaluation

- **Lecture**:

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

- **Labs**:

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

- **Scale**:

  | | | |
  |---|---|---|
  | 90 % of pts = 5.0 | 80 % of pts = 4.5 | 70 % of pts = 4.0 |
  | 60 % of pts = 3.5 | 50 % of pts = 3.0 | otherwise = 2.0 |

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

# Bibliography


- J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining of Massive Datasets*.
  Cambridge University Press, 2014
  `http://www.mmds.org`

- R. Kimball and M. Ross. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd Edition*.
  John Wiley & Sons, 2013

- 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/`