Introduction to Data Mining and Machine Learning

Dmitry I. Ignatov

National Research University Higher School of Economics
Computer Science Faculty
Dept. of Data Analysis and Artificial Intelligence

2020

Outline

Outline

Important Topics

1. Clustering
2. Classification
3. Support Vector Machines
4. Committee machines
5. Frequent itemsets and association rules
6. Recommender Systems and Algorithms
7. Multimodal Clustering*
8. Regression and Regularisation
9. Sequence Mining and Time Sereies*
10. Topic Modeling*
11. Artificial Neural Networks and Deep Learning*
12. Feature Selection and Dimensionality Reduction
13. Anomalies in Data and Outliers Detection. Missing Values.
14. Elements of Statistical Learning*
15. Big Data Tools and Techniques*

Final Grade

Scenario 1

Homeworks

Scenario 2

Homeworks + project (in groups or individual)

Outline

KDD and Data Mining

Knowledge discovery in Databases (KDD)

KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

Fayyad, Piatetsky-Shapiro, and Smyth 1996

Data Mining

Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data.

The same source.

KDD and Data Mining

KDD Scheme

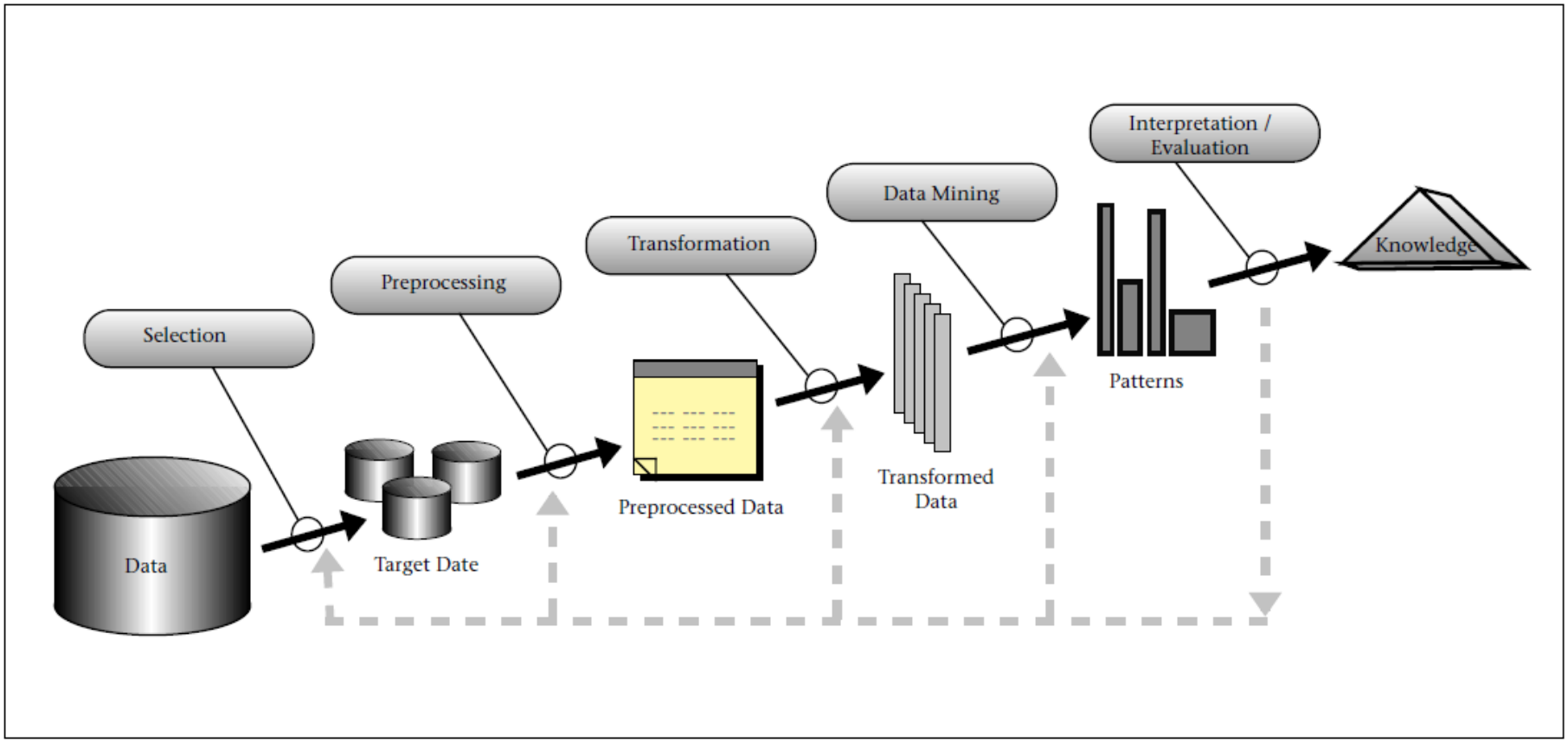


Figure 1. An Overview of the Steps That Compose the KDD Process.

(*Fayyad, Piatetsky-Shapiro, and Smyth 1996*)

KDD and Data Mining

[J. Han et al., Data Mining. Concepts and Techniques, 3rd Ed., 2012]

1. Data cleaning
2. Data integration
3. Data selection
4. Data transformation
5. Data mining (an essential process where intelligent methods are applied to extract data patterns)
6. Pattern evaluation
7. Knowledge presentation

Data Mining

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data.

Machine Learning (ML)

[T. Mitchell. The Discipline of Machine Learning, 2006]

The main question in ML

How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?

More detailed

To be more precise, we say that a machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E. Depending on how we specify T, P, and E, the learning task might also be called by names such as data mining, autonomous discovery, database updating, programming by example, etc.

Interdisciplinary Connections

Hypothesis

Data Mining $\stackrel{?}{=}$ Machine Learning

Related Disciplines

- Computer Science
- Artificial Intelligence
- Pattern Recognition
- Information Retrieval
- Social Network Analysis
- Probability Theory and Mathematical Statistics
- Discrete Mathematics (incl. Posets and Graphs)
- Optimisation

Applications of DM&ML

Application Fields

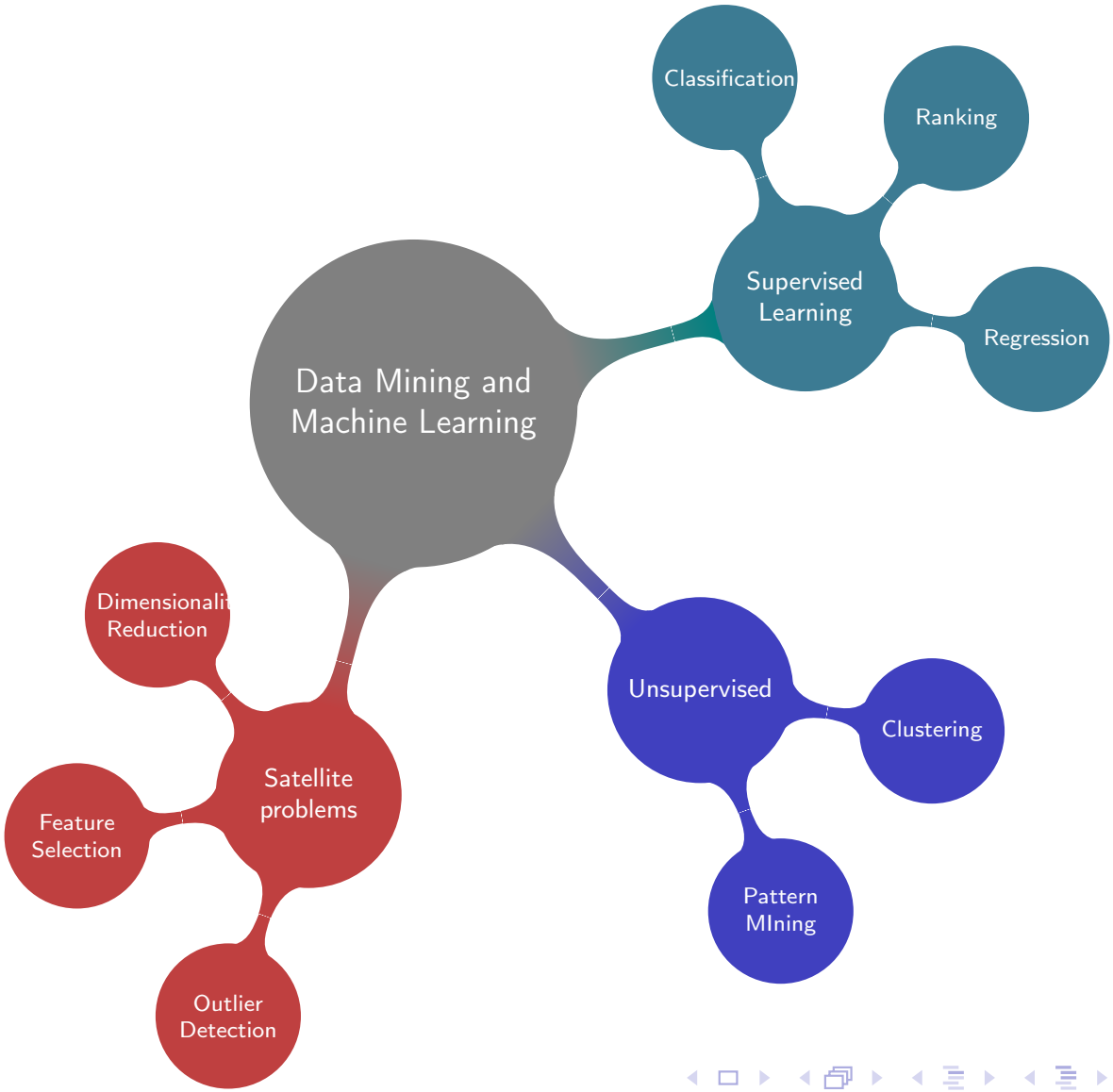
- Business
- Medicine
- Education
- Life Sciences
- Internet Data
- Banking and Finance
- Geosciences
- ...

DM&ML Application Trends

[J. Han et al., 2012]

- Application exploration: e.g., counter-terrorism and mobile (wireless) data mining
- Scalable and interactive data mining methods
- Integration of data mining with search engines, database systems, data warehouse systems, and cloud computing systems
- Mining social and information networks
- Mining spatiotemporal, moving-objects, and cyber-physical system
- Mining multimedia, text, and web data
- Mining biological and biomedical data
- Data mining with software engineering and system engineering
- Visual and audio data mining
- Distributed data mining and real-time data stream mining
- Privacy protection and information security in data mining

Taxonomy of DM&ML techniques

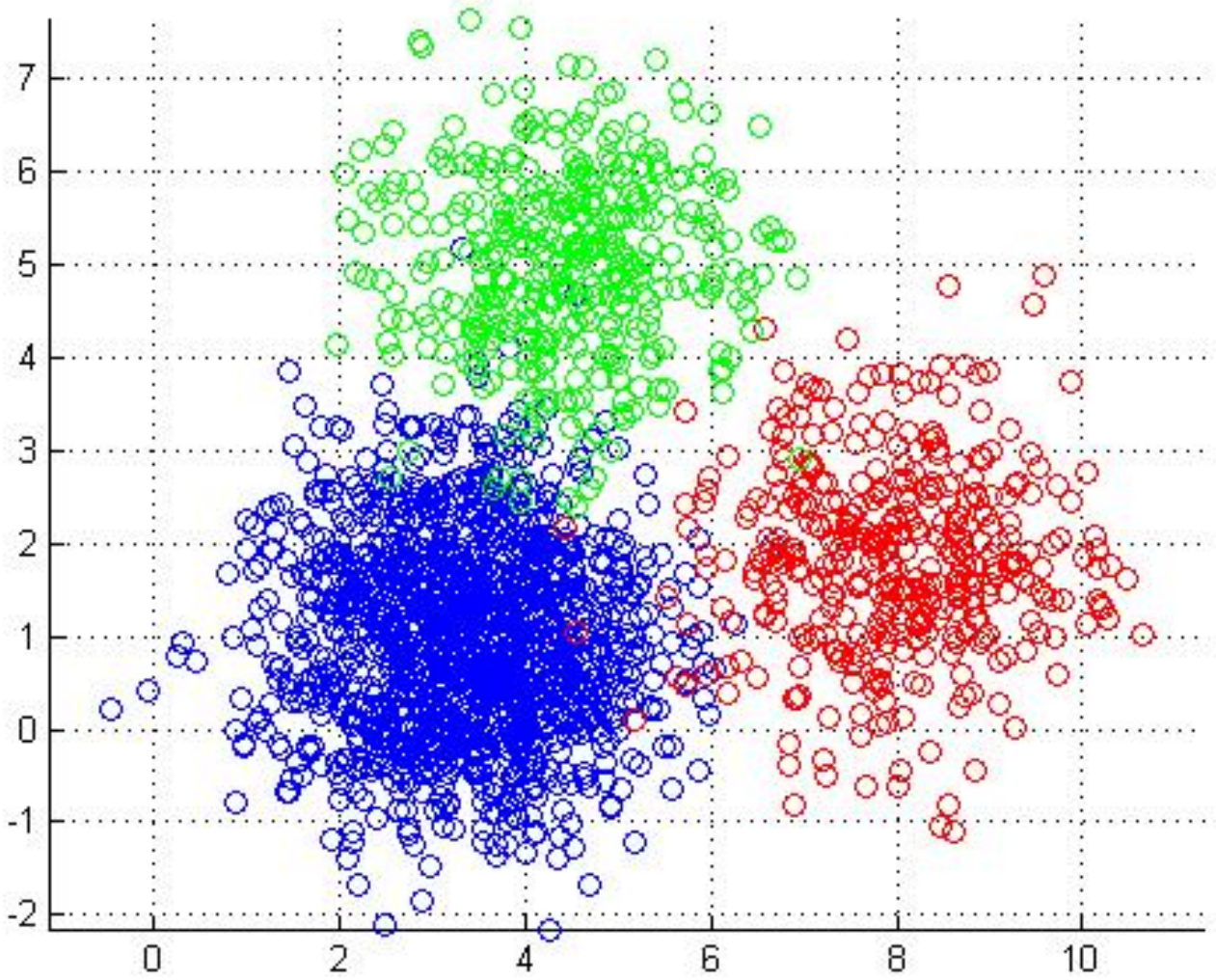


- Find a partition of an input set of objects into groups of objects (clusters) such that
- objects within one cluster are highly similar to each other.
- objects from different clusters are dissimilar.

- k-means
- Hierarchical clustering (agglomerative and divisive approaches)
- Spectral clustering
- Multimodal clustering: biclustering and triclsutering

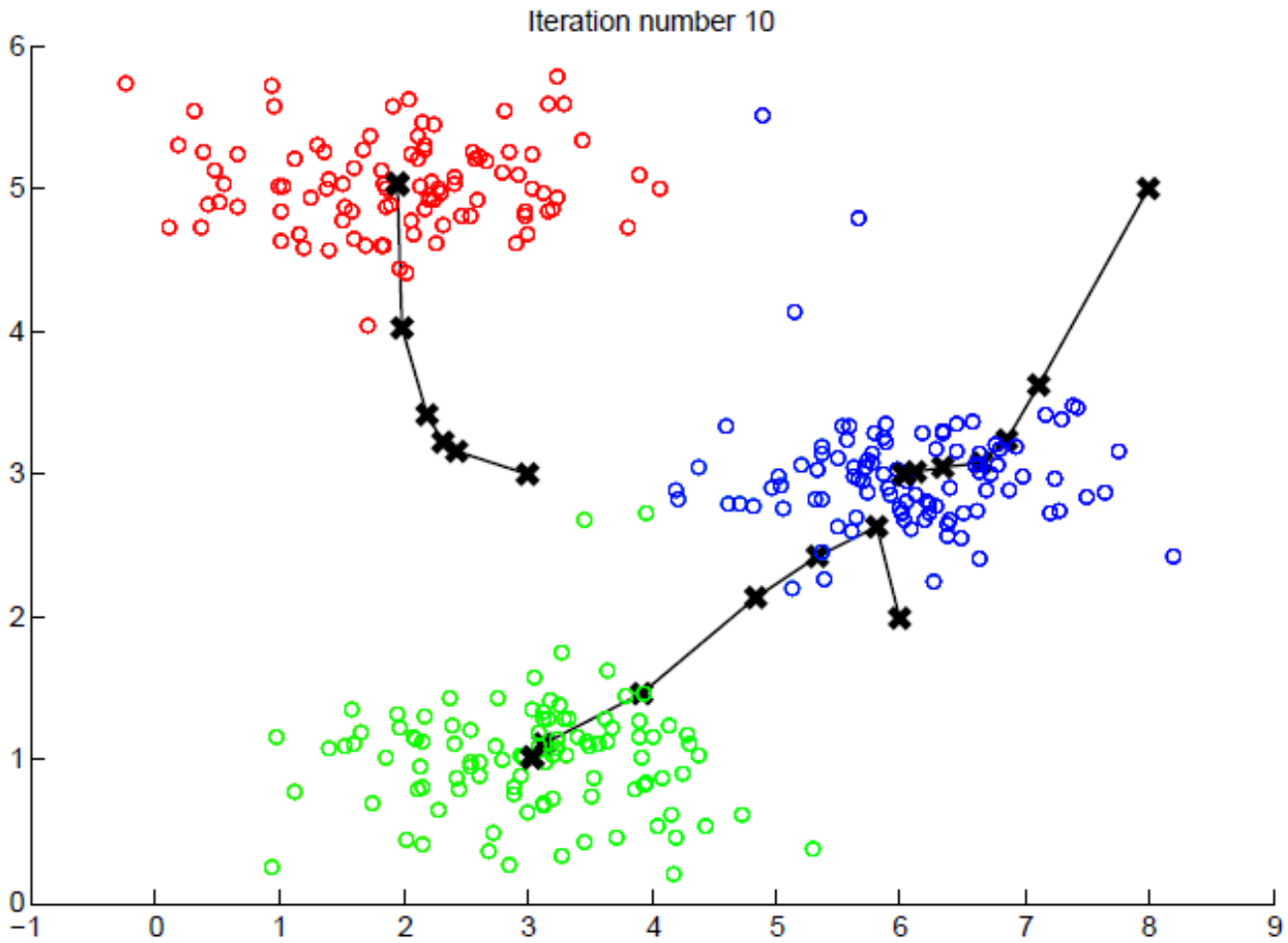
Clustering

k-means



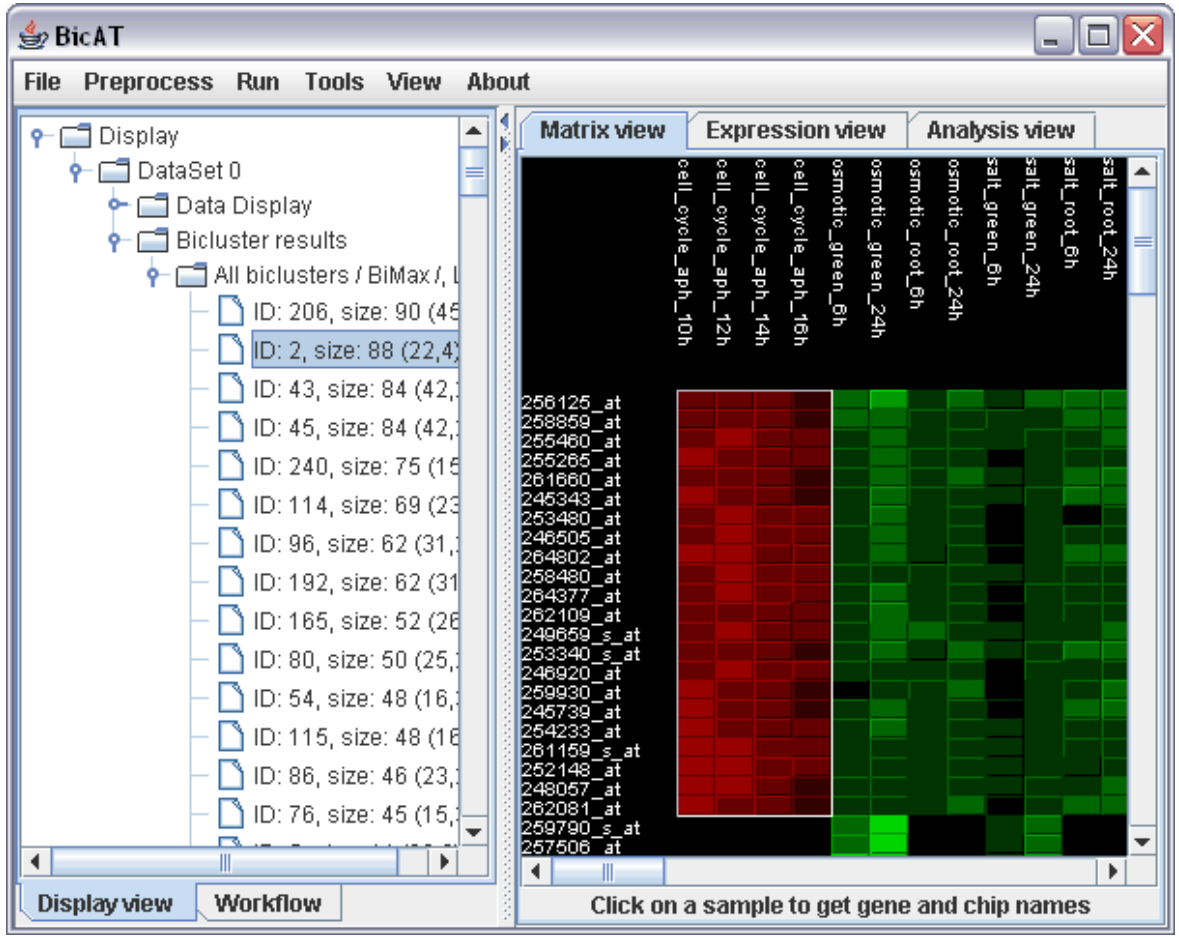
Clustering

k-means



Biclustering

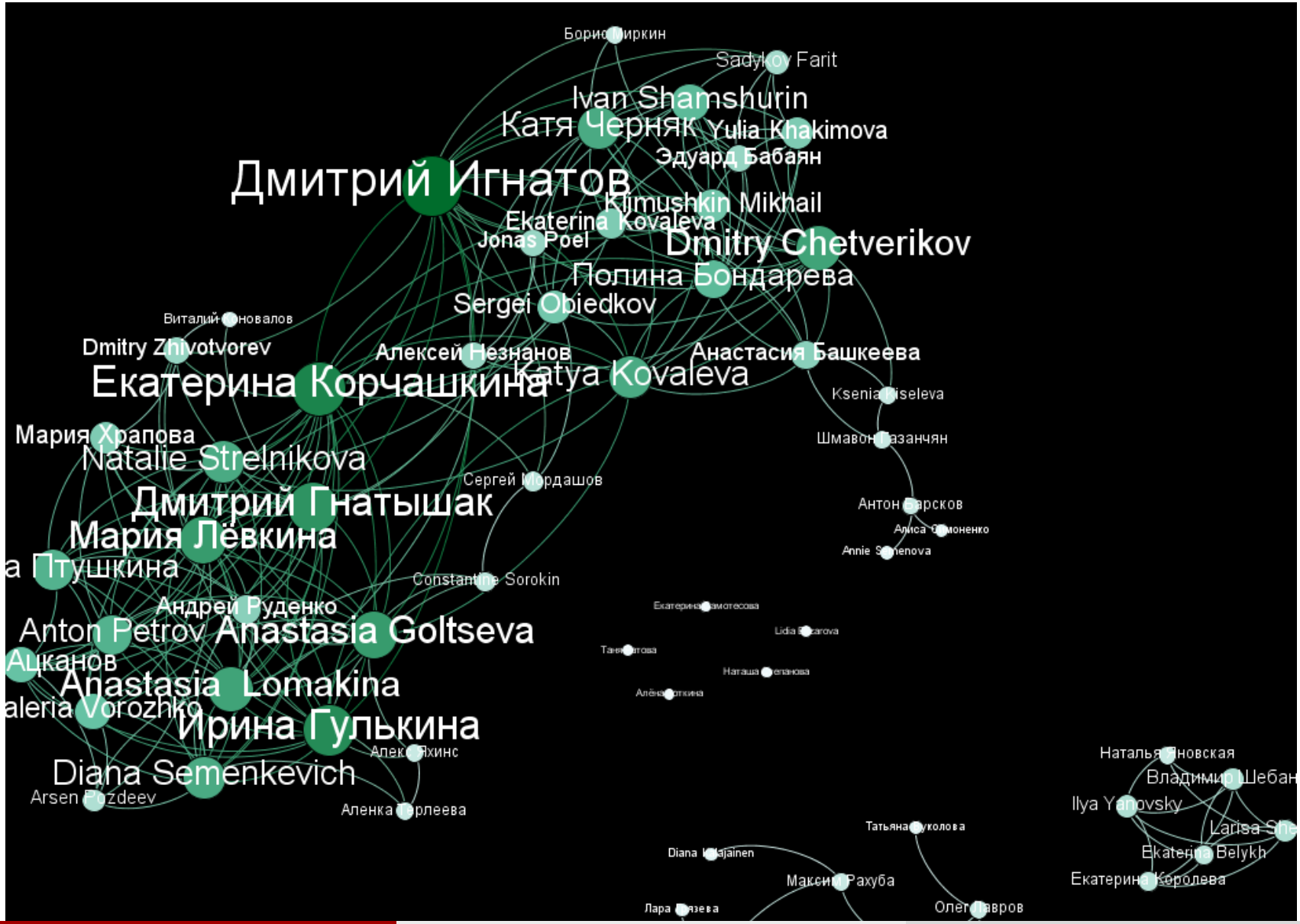
Gene expression analysis



Source: BicAT Tutorial

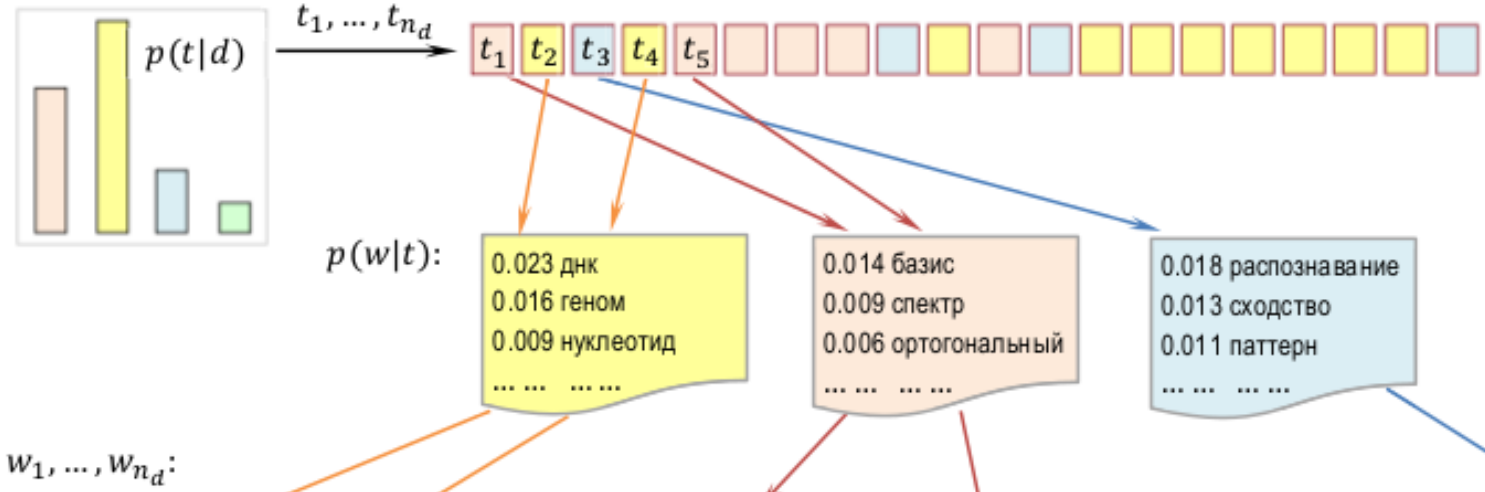
Social Network Analysis

Community detection as clustering technique



Topic Modeling

TM as text clustering

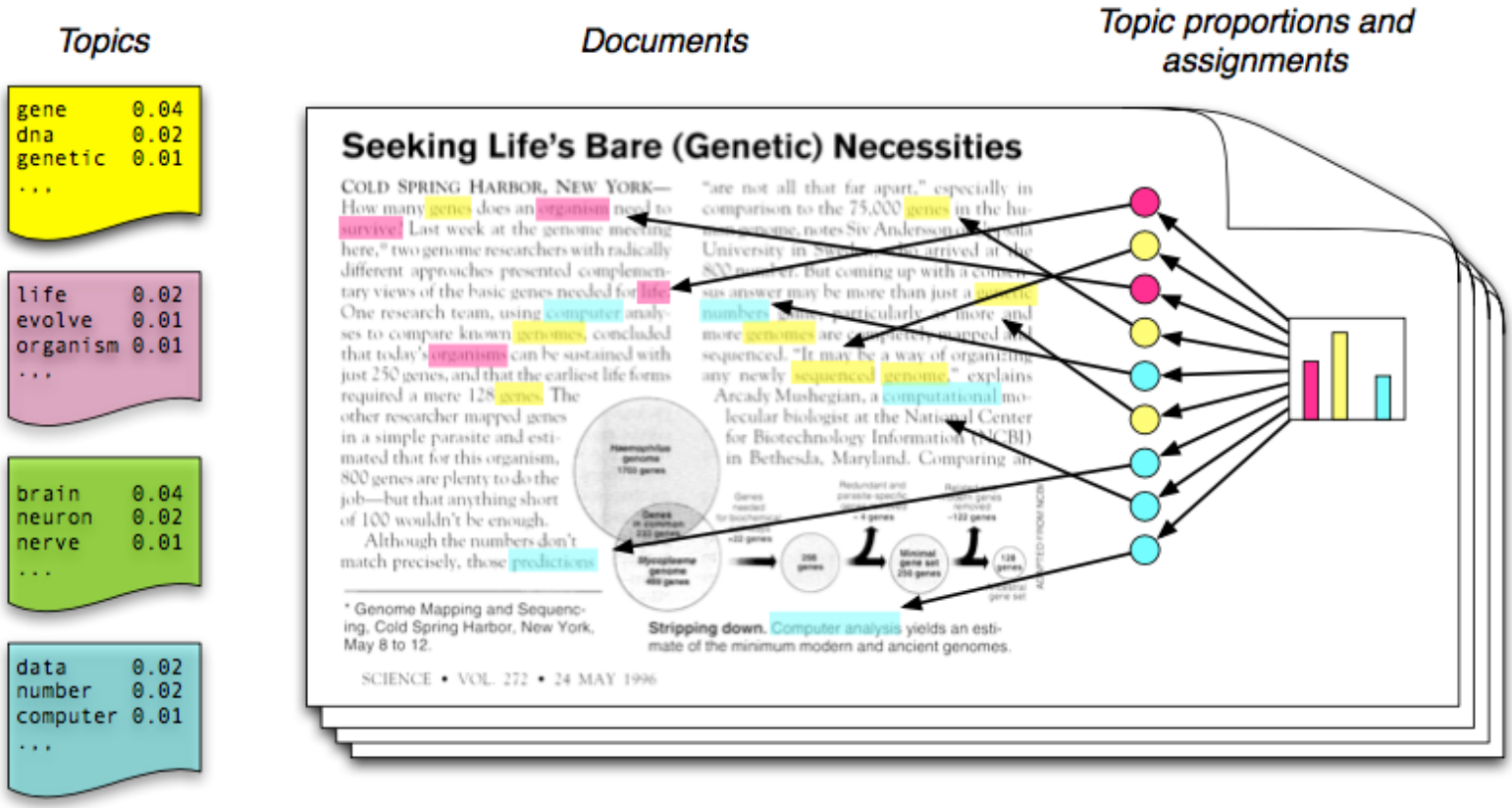


Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Source: **Project** http://bigartm.org/

Topic Modeling

TM as text clustering



- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

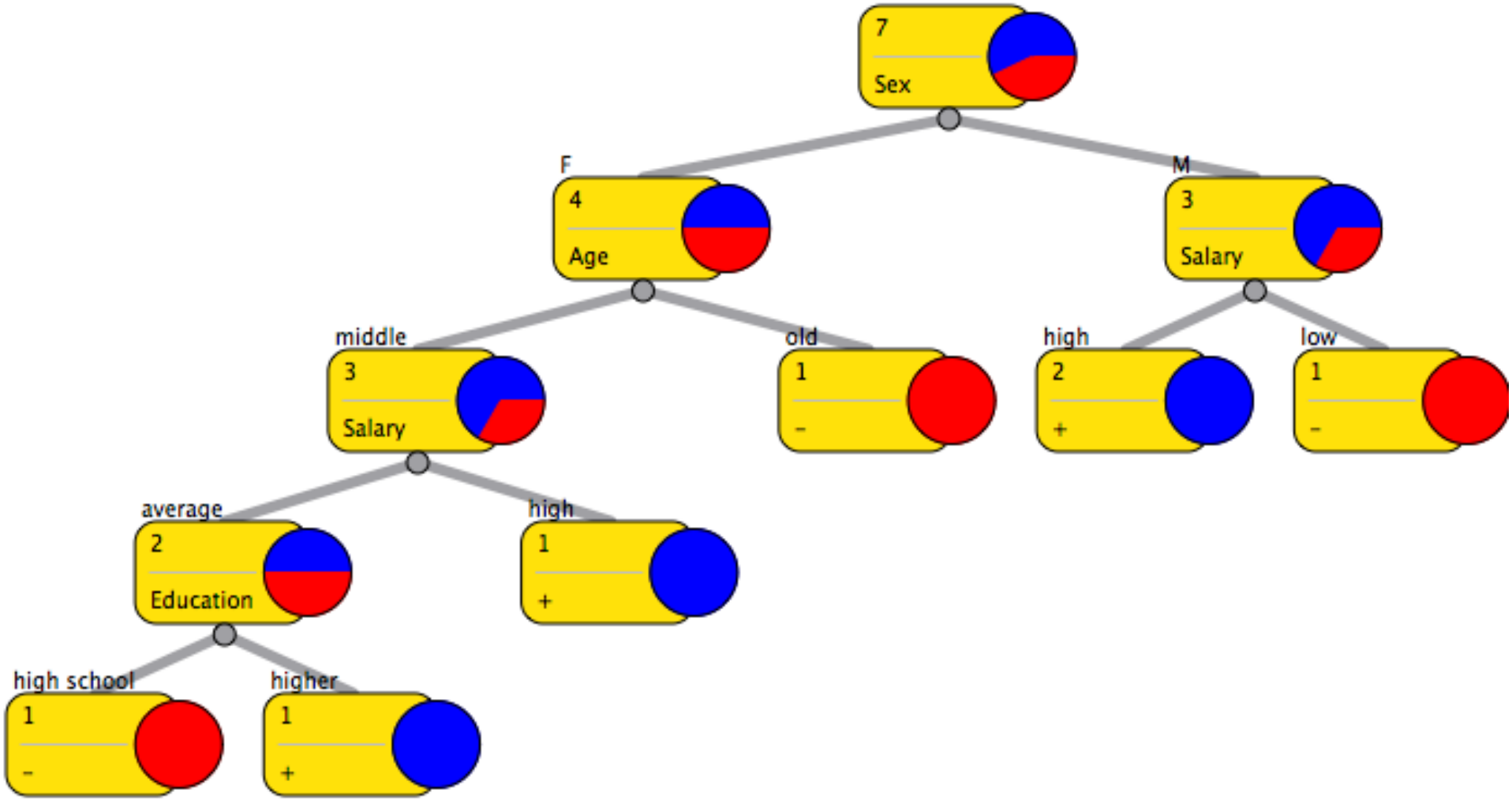
Classification

Problem Statement

- Having feature-based description of an input set of objects with known labels of classes, to predict the classes of objects of the same nature (in the same feature space) with unknown labels.

Classification

Decision trees in credit scoring



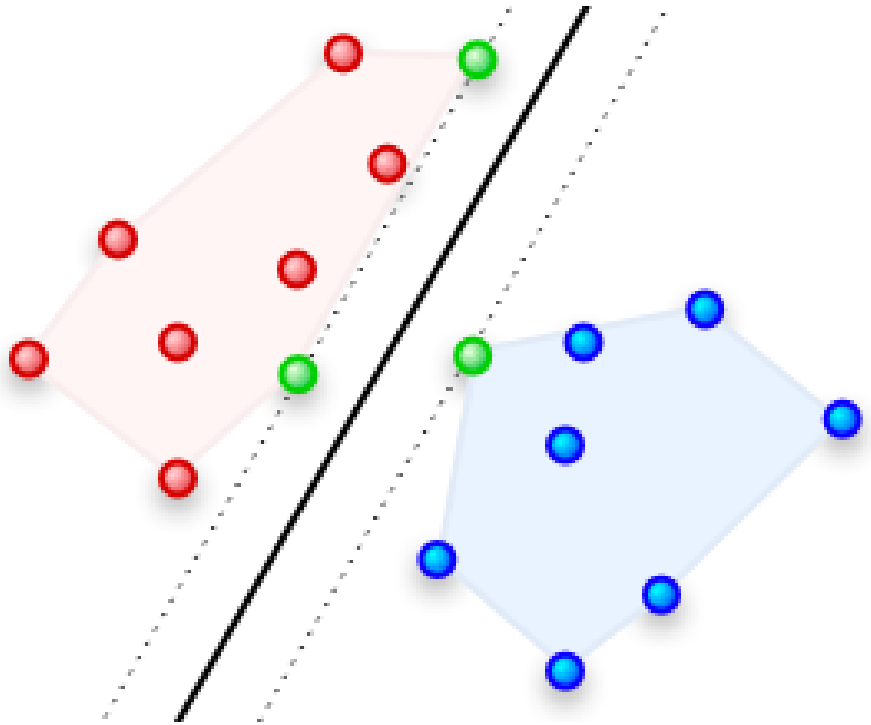
Classification

Classification technique

- 1-Rule
- kNN classifier (k nearest neighbours)
- Naïve Bayes classifier
- decision trees
- Support Vector Machines (SVM)

Classification

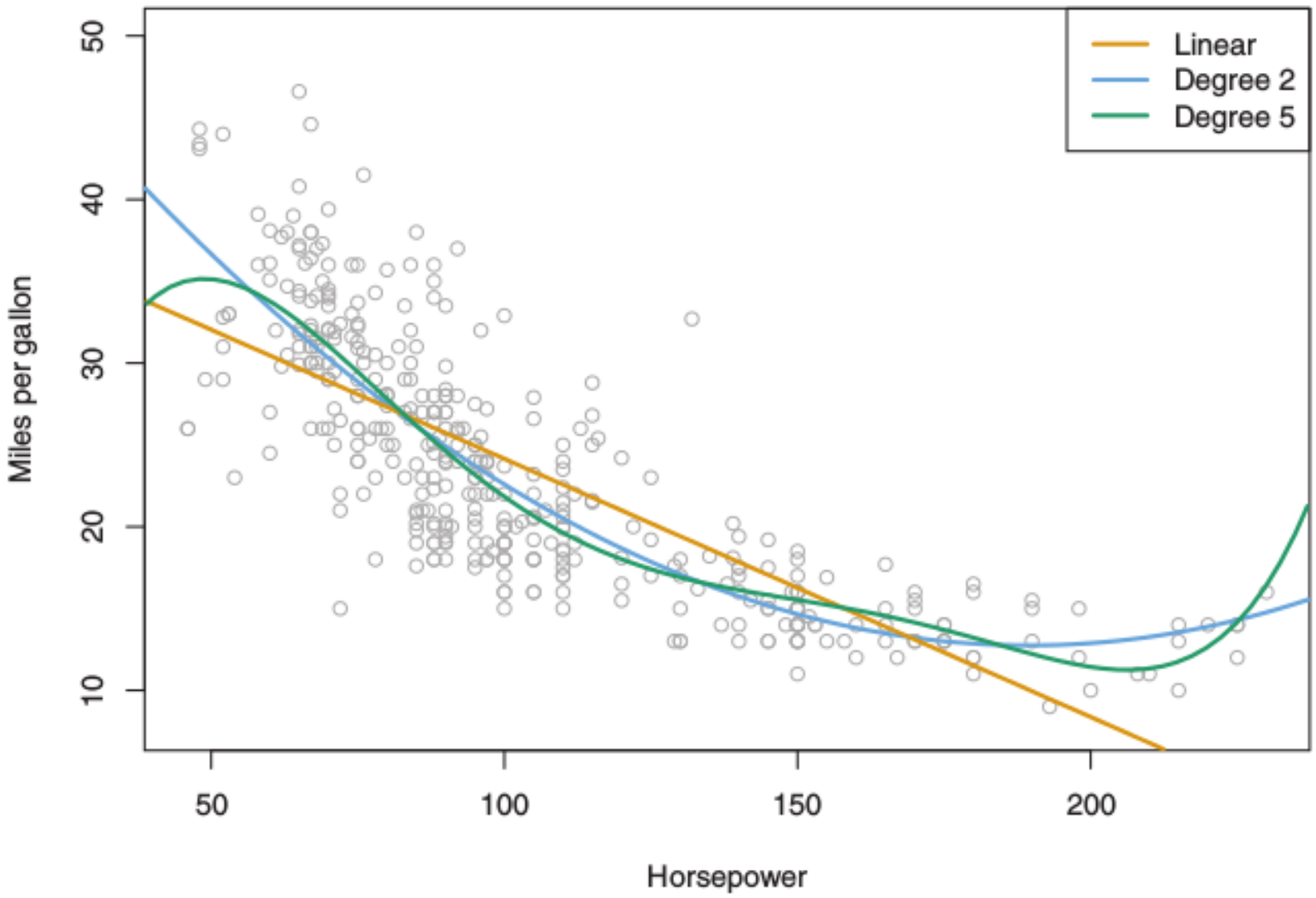
Support Vector Machines (SVM)



Regression

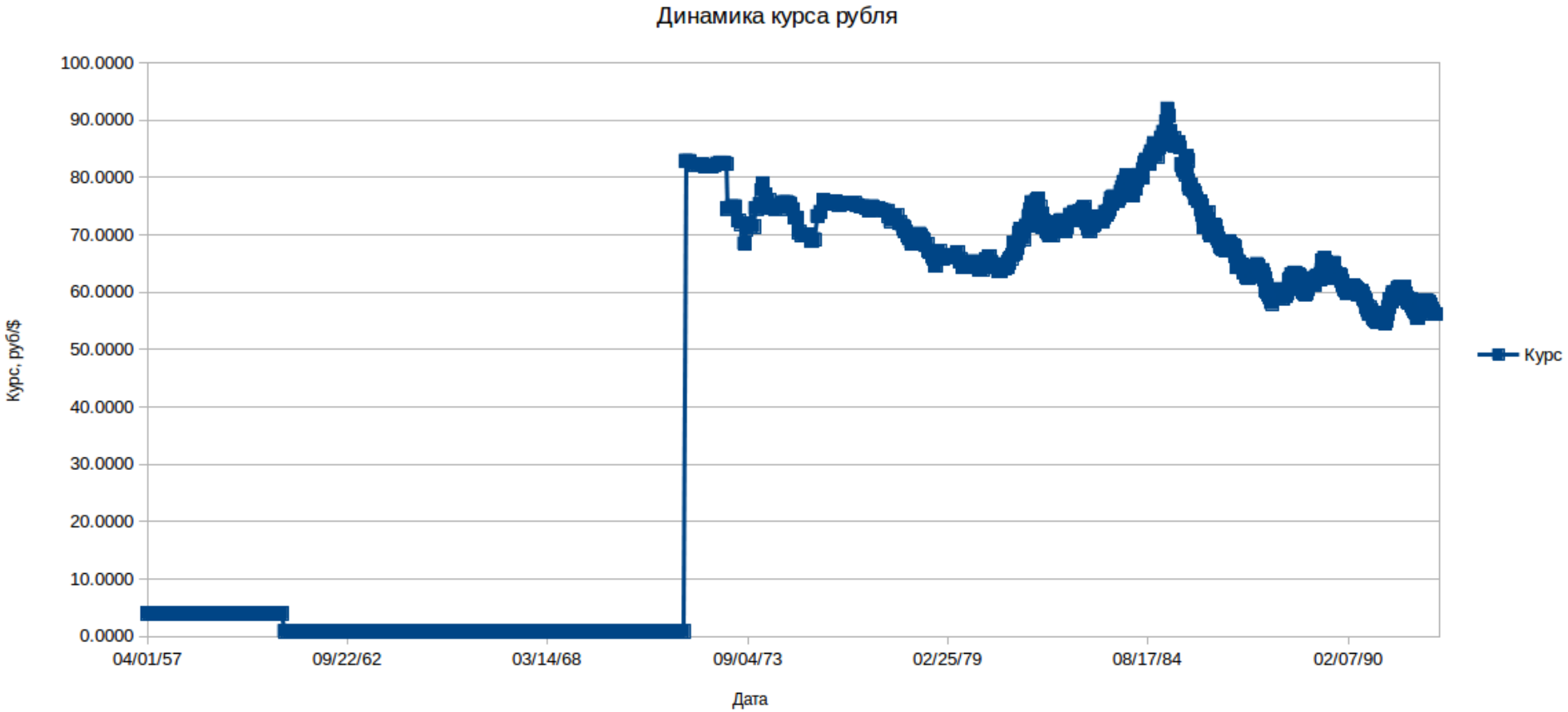
- Simple and multivariate linear regression
- Lasso regularisation (more features than objects)
- Logistic regression as classification technique

Regression. An Example



Source: An Introduction to Statistical Learning

Time Series Analysis



Source http://www.cbr.ru/currency_base/OldVal.aspx

Pattern mining

Problem Statement

- Searching for patterns of usage of certain resources. For example, frequently used resources.
- Example: $support(\{\text{bread, milk}\}) = 0.7$
- Often such pattern are represented as rules: $A \longrightarrow B$
- Example: $\{\text{Student, Age from 16 to 25}\} \longrightarrow \{iPhone, iPad\}$

Pattern Mining

Example

Customer/items	Beer	Cookies	Milk	Müsli	Chips
c_1	1	0	0	0	1
c_2	0	1	1	1	0
c_3	1	0	1	1	1
c_4	1	1	1	0	1
c_5	0	1	1	1	1

- $supp(\{\text{Beer, Chips}\}) = 3/5$
- $supp(\{\text{Cookies, Müsli }\} \rightarrow \{\text{ Milk }\}) = 2/5$
- $conf(\{\text{Cookies, Müsli }\} \rightarrow \{\text{ Milk }\}) = 1$
- $conf(\{\text{Milk}\} \rightarrow \{\text{ Cookies, Müsli }\}) = ?$

Pattern Mining

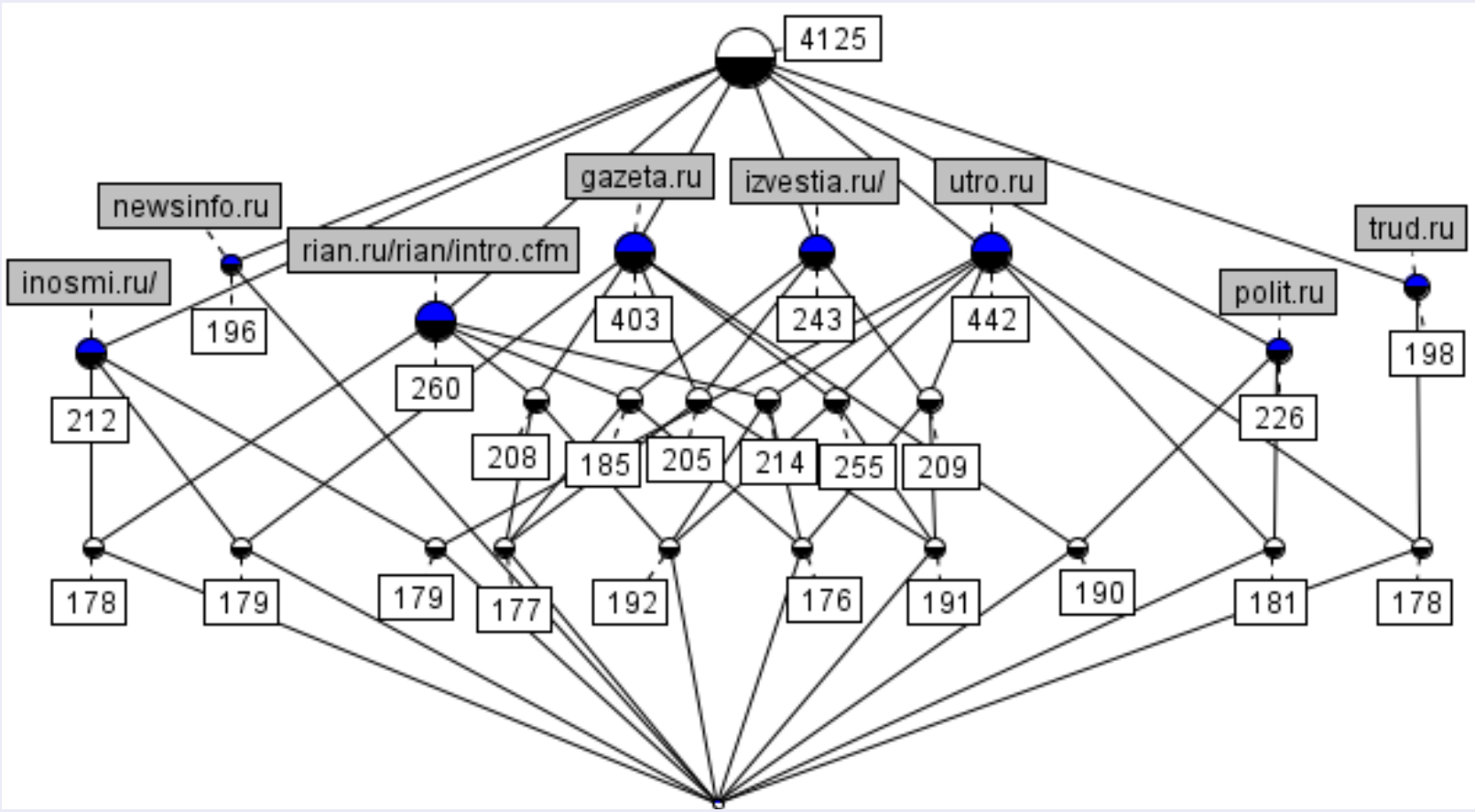
The FIMI'03 best implementation award was granted to Gosta Grahne and Jianfei Zhu (on the left). The award consisted of the most frequent itemset: $\{diapers, beer\}$.

Pattern Mining

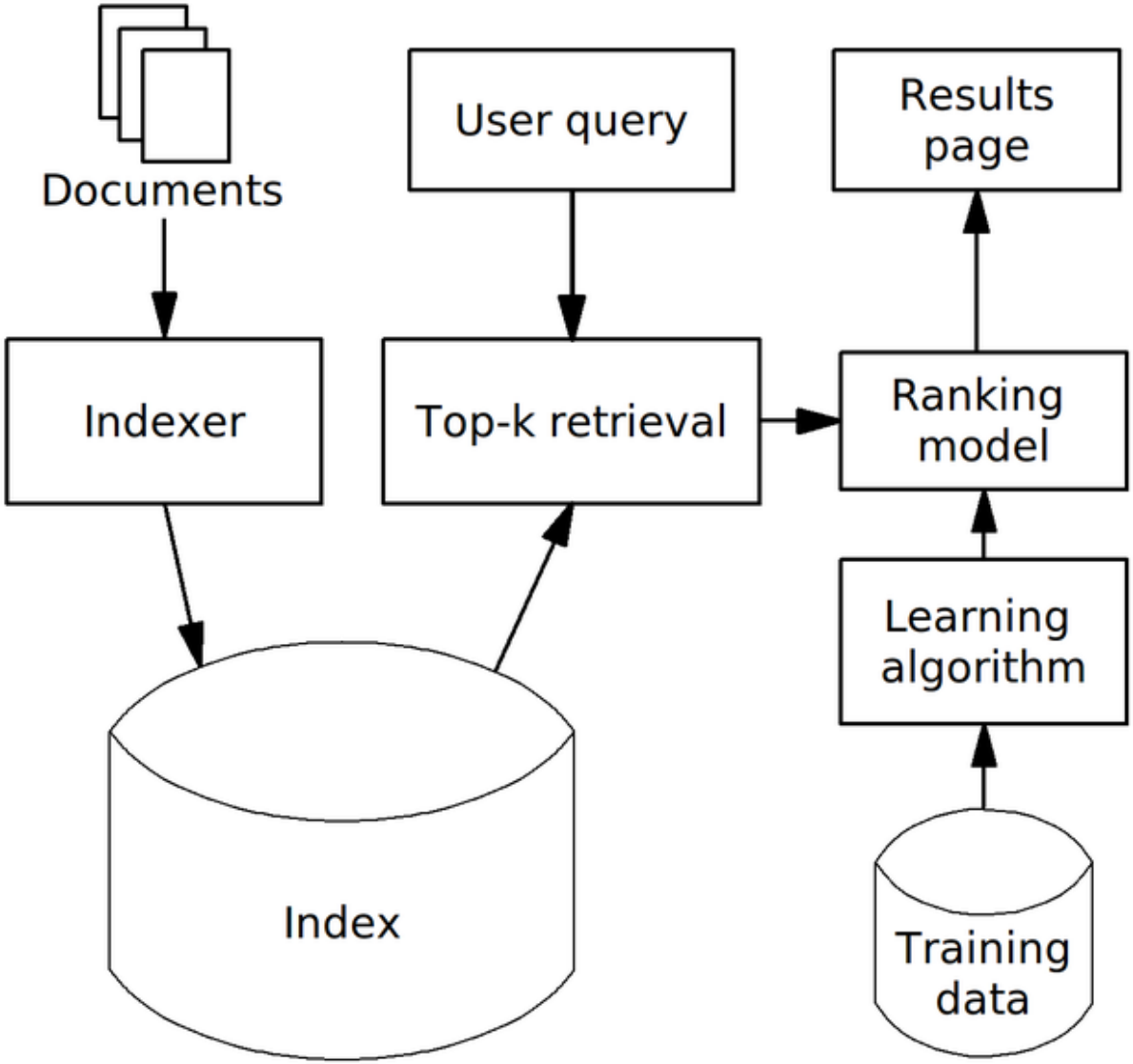
Analysis of web site audience

HSE website in September 2006 w.r.t. news web sites' visits.

Iceberg-lattice for 25 largest concepts



Ranking


Documents
User query
Results page
Indexer
Top-k retrieval
Ranking model
Learning algorithm
Index
Training data

Recommender Systems

http://Amazon.com

Frequently Bought Together

Price For All Three: $86.01

Add all three to Cart | Add all three to Wish List

Show availability and shipping details

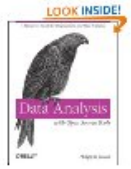
- **This item:** Machine Learning for Hackers by Drew Conway Paperback $33.87
- Machine Learning in Action by Peter Harrington Paperback $25.75
- Programming Collective Intelligence: Building Smart Web 2.0 Applications by Toby Segaran Paperback $26.39

Customers Who Bought This Item Also Bought

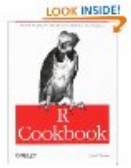
Programming Collective Intelligence: Building ...
› Toby Segaran
(84)
Paperback
$26.39

Machine Learning in Action
› Peter Harrington
(10)
Paperback
$25.75

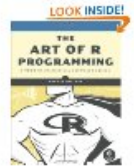
Mining the Social Web: Analyzing Data from ...
› Matthew A. Russell
(19)
Paperback
$26.36



Data Analysis with Open Source Tools
› Philipp K. Janert
(29)
Paperback
$24.05



R Cookbook (O'Reilly Cookbooks)
› Paul Teetor
(18)
Paperback
$32.43



The Art of R Programming: Tour of Statistical ...
Norman Matloff
(29)
Paperback
$25.06

Are any of these items inappropriate for this page? Let us know

Recommender Systems

http://Imhonet.ru

Оценки фильма Любопытное стечение обстоятельств

Een Bizarre Samenloop Van Omstandigheden, A Curious Conjunction of Coincidences

Фильмы / Комедии / обстоятельств» /

Да, Вам стоит смотреть фильм «Любопытное стечение обстоятельств»

Людям, с оценками, похожими на Ваши, этот фильм нравится

А ещё они рекомендуют Вам 31 фильм

Ваша прогнозируемая оценка фильма после его просмотра
8.2

Смотрели? Оцените

Не рекомендовать

Про фильм | Онлайн | Скачать | Отзывы | Персоны | Кадры | Оценки | Похожие

Распределение оценок

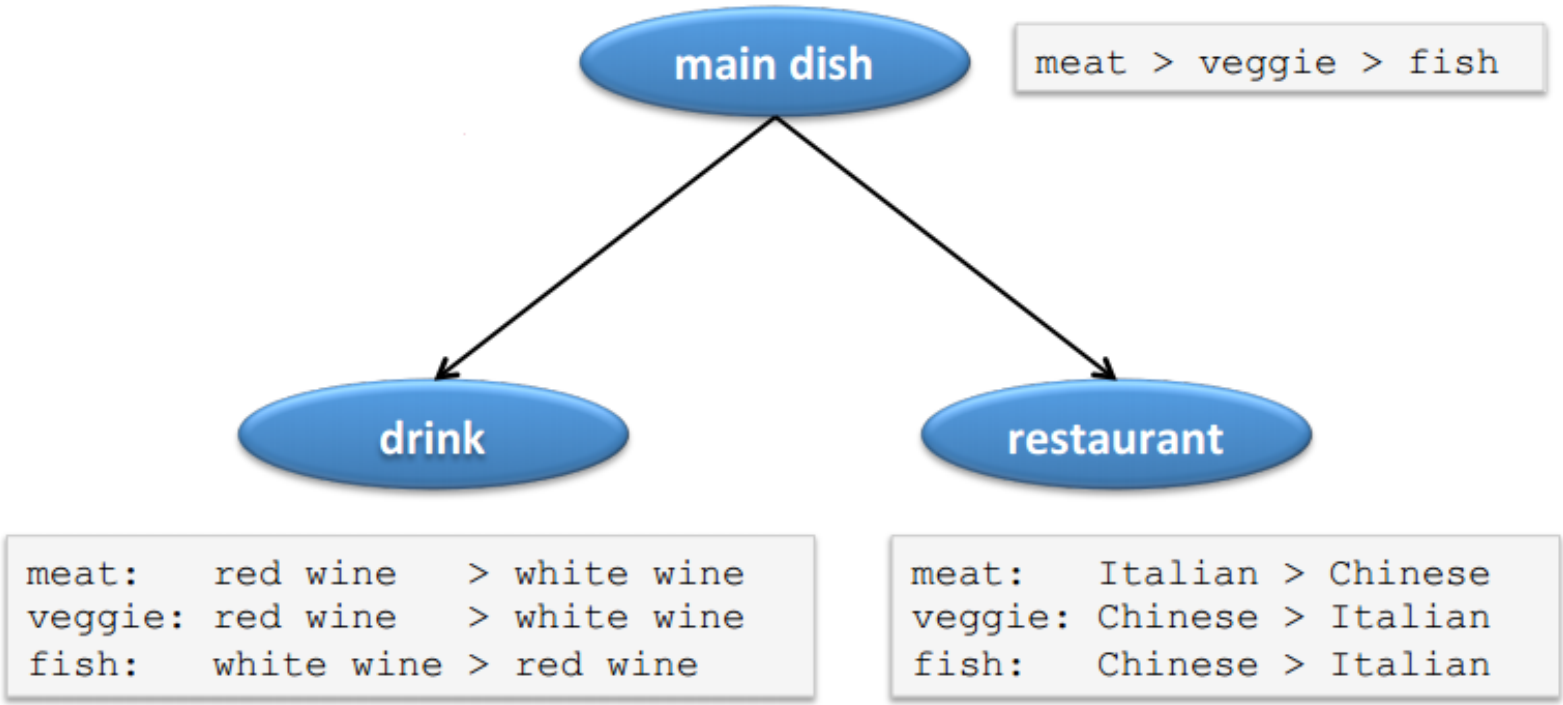
1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	0	1	4	4

Кому больше нравится

Кому нравится: 99% 1%

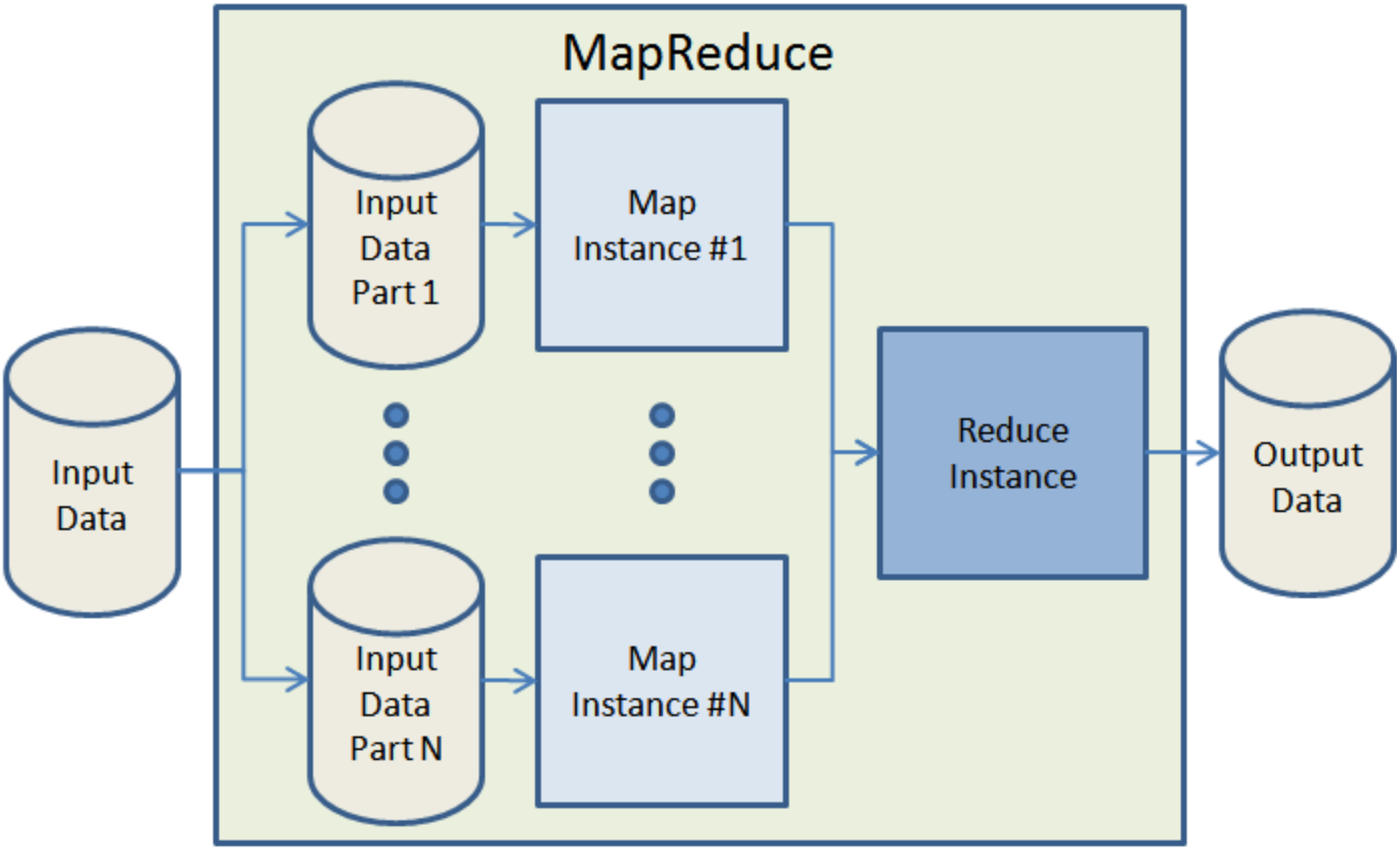
Preference Learning

http://www.preference-learning.org/



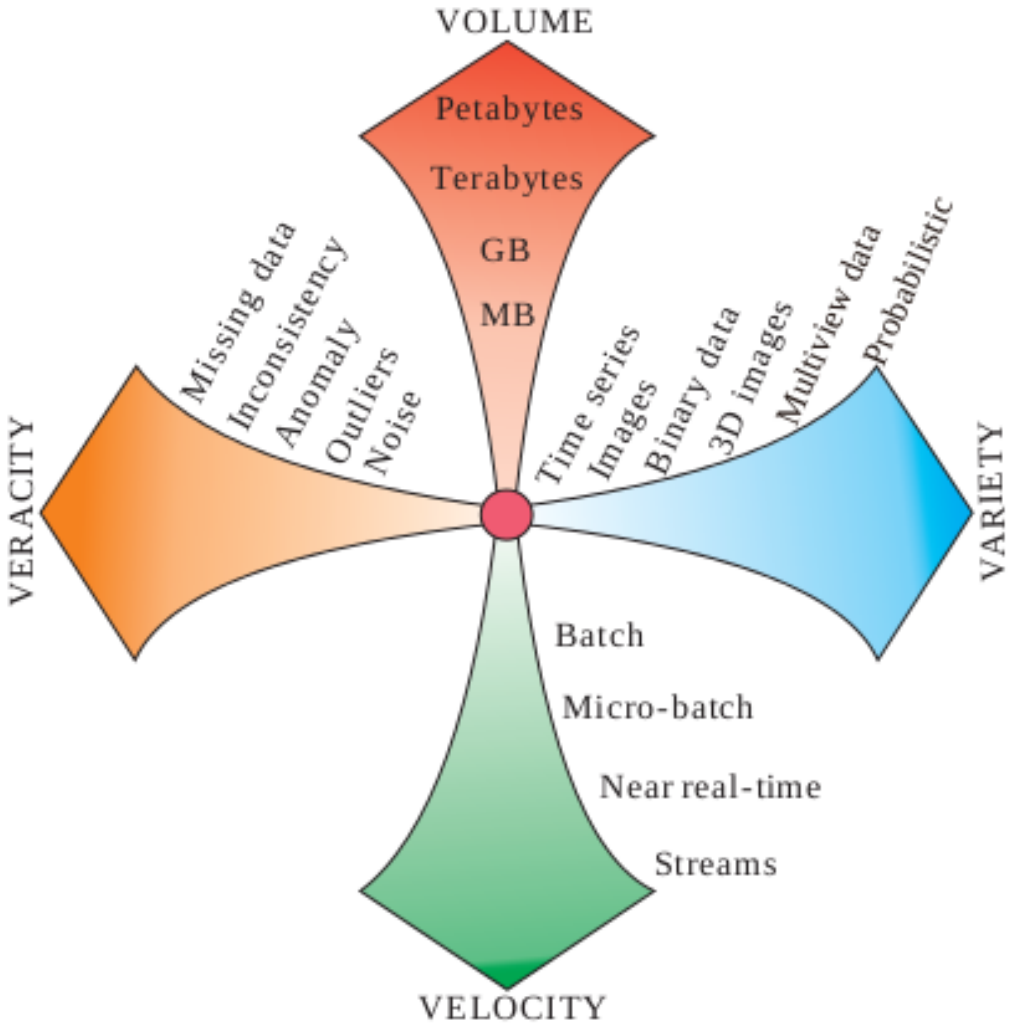
Big Data

MapReduce Technology

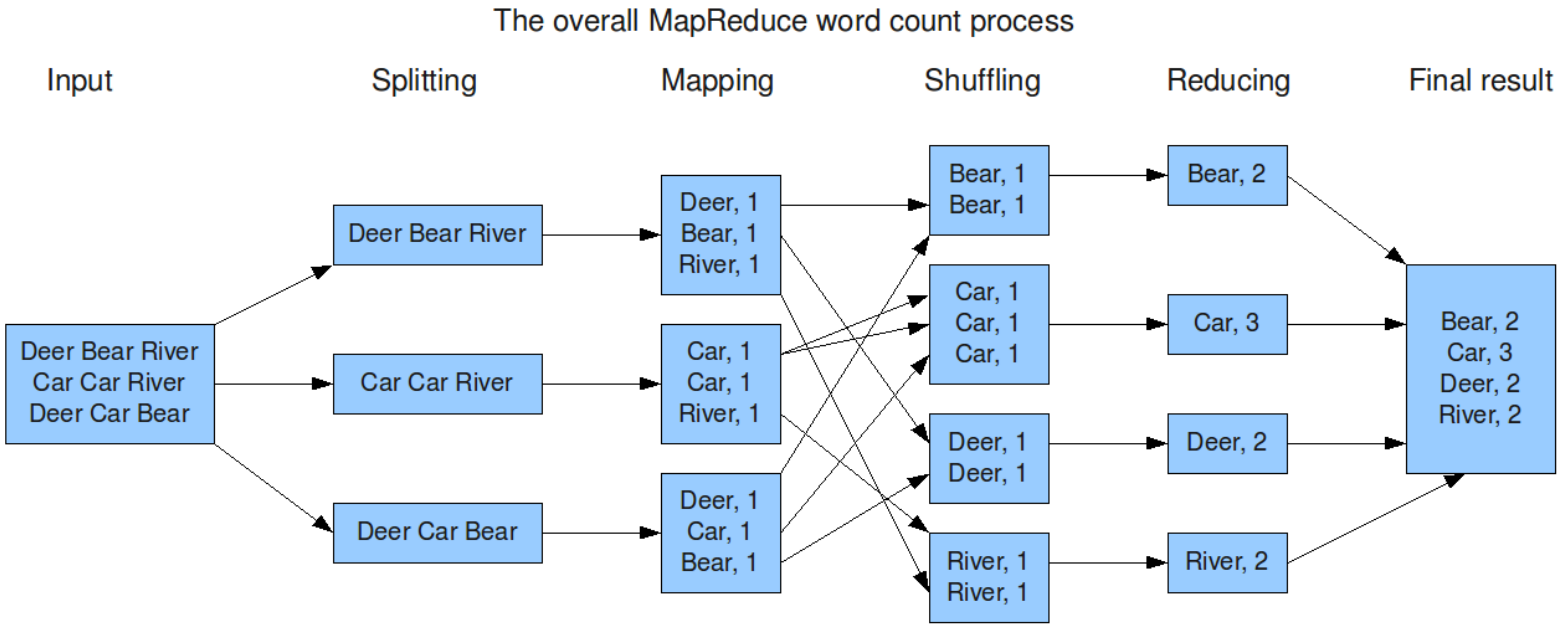


Big Data

4 main challenges



Source: Foundation and Trends in Machine Learning


The overall MapReduce word count process
Input
Splitting
Mapping
Shuffling
Reducing
Final result
Deer Bear River
Car Car River
Deer Car Bear
Deer Bear River
Car Car River
Deer Car Bear
Deer, 1
Bear, 1
River, 1
Car, 1
Car, 1
River, 1
Deer, 1
Car, 1
Bear, 1
Bear, 1
Bear, 1
Car, 1
Car, 1
Car, 1
Deer, 1
Deer, 1
River, 1
River, 1
Bear, 2
Car, 3
Deer, 2
River, 2
Bear, 2
Car, 3
Deer, 2
River, 2

What is Apache Mahout?

The Apache Mahout^{TM} project's goal is to build an environment for quickly creating scalable performant machine learning applications.

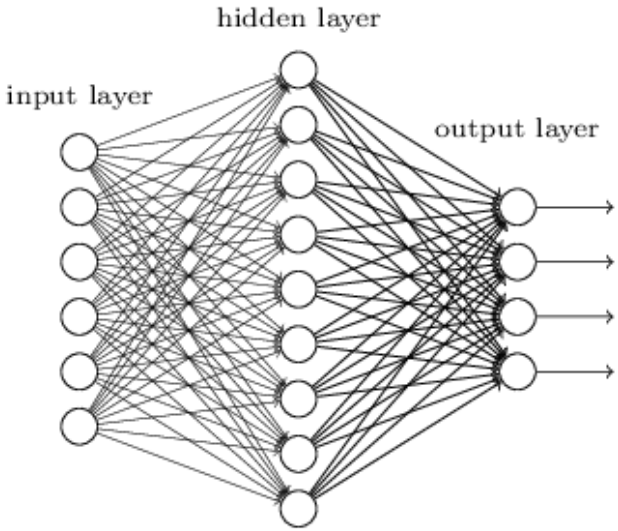
What is Apache Spark?

- "Apache Spark™ is a fast and general engine for large-scale data processing."
- Including libraries:
  - Spark SQL
  - GraphX (graph processing)
  - Spark Streaming (stream processing)
  - MLlib (for ML)
- It works with(out) Hadoop.

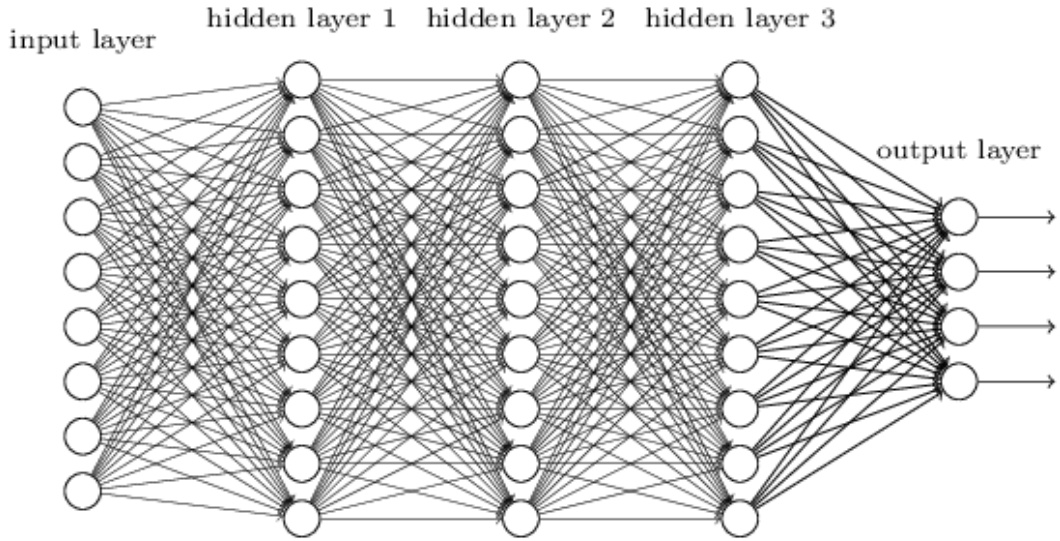
Deep Learning

Intro by Grigory Sapunov in Russian

"Non-deep" feedforward neural network



Deep neural network



Outline

ML&DM Systems and Software Tools

1. Orange (freely available)
2. Weka (freely available)
3. Knime (community edition for free)
4. RapidMiner (community edition for free)
5. Deductor (freely available for educational purposes)
6. QuDA (freely available)

ML&DM libraries

1. scikit-learn (freely available Machine Learning in Python)
2. H_2O (freely available Machine Learning in Python)
3. MALLET — MAchine Learning for LanguagE Toolkit (freely available)
4. Accord.NET Framework (.NET machine learning framework combined with audio and image processing libraries completely written in C#)
5. Infer.NET (framework for running Bayesian inference in graphical models)
6. ML.NET (An open source and cross-platform machine learning framework)
7. R (free software environment for statistical computing and graphics+many packages for ML&DM) & MLR

ML&DM Standards

http://www.dmg.org

PMML

(Predictive Model Markup Language — PMML) is developed by Data Mining Group (DMG). It is one of the leading standards for statistical and data mining models supported by over 20 vendors and organizations. With PMML, one can develop a model on one system using one application and deploy the model on another system using another application, simply by transmitting an XML configuration file.

Outline

Books

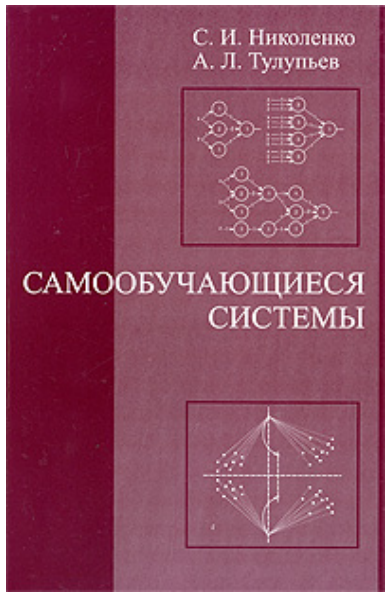
- P. Flach Machine Learning: The Art and Science of Algorithms that Make Sense of Data, 2012
- M. Zaki et al. *Data Mining and Analysis: Fundamental Concepts and Algorithms*, 2014 (free)
- J. Leskovec et al. Mining of Massive Datasets, 2014 (free)
- C.M. Bishop *Pattern Recognition and Machine Learning*, 2006
- D. Barber Bayesian Reasoning and Machine Learning, 2012 (free)
- K.P. Murphy Machine Learning: a Probabilistic Perspective, 2012
- T. Hastie et al. *Elements of Statistical Learning*, 2009 (free)
- G. James et al. *An Introduction to Statistical Learning with Applications in R*, 2013 (free)
- J. Han et al. *Data Mining. Concepts and Techniques*, 2012
- T. Mitchell *Machine Learning*, 1997
- T. Segaran *Programming Collective Intelligence*, 2007 (in English)
- Барсегян А. et al. *Analysis of data and processes (Анализ данных и процессов)*, 2009 (in Russian))

- Lectures of K. Vorontsov. *Математические методы обучения по прецедентам (Mathematical Methods of Learning by Examples)*
- Lectures of D. Vetrov, D. Kropotov Байесовские методы машинного обучения (Bayesian Machine Learning), 2014
- Texbook by A. Dyakonov. *Анализ данных, обучение по прецедентам, логические игры, системы WEKA, RapidMiner и MatLab (Data Analysis, Learning by Examples, and Logic Games, WEKA, RapidMiner and MatLab*, 2010

Slides and book of Sergey Nikolenko

http://logic.pdmi.ras.ru/~sergey/

- Player of Chto? Gde? Kogda? (What? Where? When?)
- S.Nikolenko, A. Tulupiev. *Самообучающиеся системы (Self-learning systems)* 2009 (In Russian)
- Николенко С.И., Кадурин А.А., Архангельская Е.О. Глубокое обучение (Deep Learning) 2018 (in Russian)



Coursera: courses and specialisations

http://www.coursera.org/

- Andrew Ng. *Machine Learning*
- Jiawei Han Pattern Discovery in Data Mining
- Jure Leskovec et al. Mining Massive Datasets
- Hastie & Tibshirani Statistical Learning

Specialisations (paid certificates) consist of separate courses (participation for free)

- Data Mining
- Data Science

Deep Learning (Глубинное обучение или глубокое обучение)

- Deep Learning Spec. by Coursera
- Deep Learning by Udacity
- Deep Learning Course by NVIDIA
- Geoffrey Hinton. *Neural Networks for Machine Learning* (since 2012)

- Internet University of Information Technologies (Интернет-университет информационных технологий) since 10.04.2003
- K. Vorontsov Machine Learning, 2015 (Videos at the website School of Data Analysis Yandex)
- I. Chubukova. Data Mining, 2006

Community

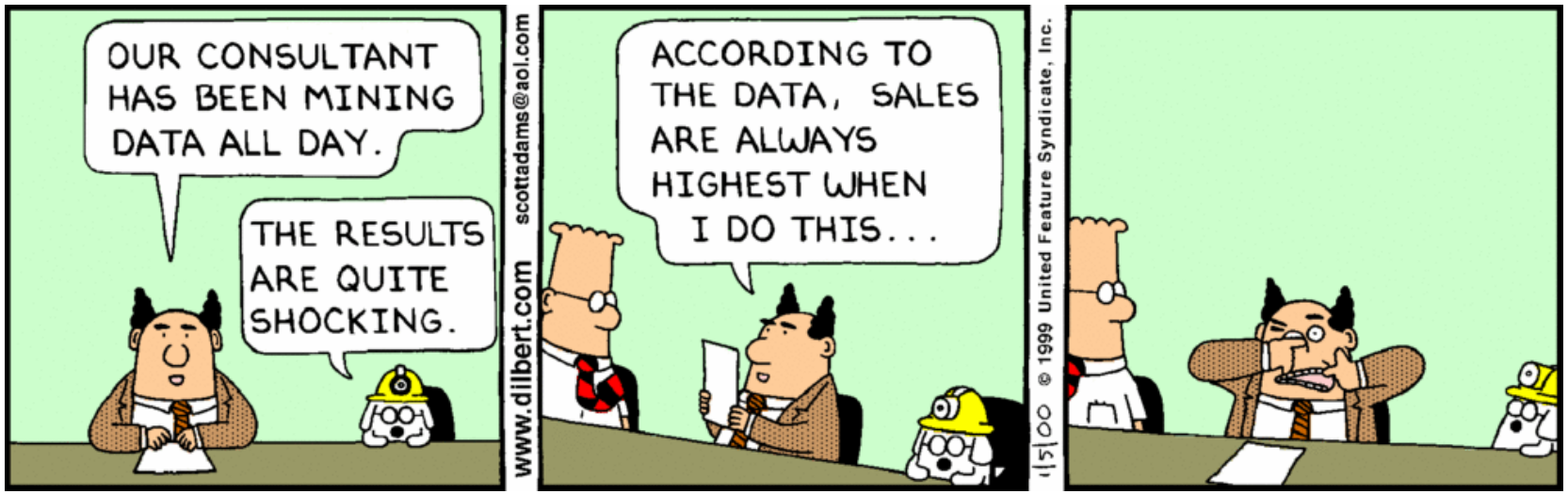
- IMLS – The International Machine Learning Society
- Kaggle – a world-wide platform for data mining competitions
- KDD Nuggets – Data Mining Community Top Resource
- Open ML – Machine Learning community portal
- UCI Machine Learning Repository – Datasets for ML

Conferences

- ICML – International Conference on Machine Learning
- IEEE ICDM – IEEE International Conference on Data Mining
- KDD – ACM SIGKDD Conference on Knowledge Discovery and Data Mining
- ECML & PKDD – European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases
- NIPS – Neural Information Processing Systems
- RecSys – The ACM conference series on Recommender Systems
- ИОИ & ММРО – A conference series on Intelligent Data Processing(«Интеллектуализация обработки информации»)/Mathematical Methods of Pattern Recognition («Математические методы распознавания образов»)
- AIST (АИСТ means "stork" in Russian) – International conference on Analysis of Images, Social Networks, and Texts

Just for fun или шутки ради

http://dilbert.com



Questions and contacts

www.hse.ru/staff/dima

Thank you!

dmitrii.ignatov[at]gmail.com