Анализ данных

# Лекция 6
# Введение в Data Science

**Гедранович Ольга Брониславовна,**
старший преподаватель кафедры ИТ, МИУ
volha.b.k@gmail.com



## Вопросы лекции

- Понятие Data Science
- Отличие Data Science от Business Intelligence
- Процесс Data Science
- Технологии и инструментарий Data Science
- Machine Learning



## Понятие Data Science

- Interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, machine learning, data mining, and predictive analytics, similar to Knowledge Discovery in Databases (KDD).
- Раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме. Объединяет методы по обработке данных в условиях больших объёмов и высокого уровня параллелизма, статистические методы, методы интеллектуального анализа данных и приложения искусственного интеллекта для работы с данными, а также методы проектирования и разработки баз данных.

Нет однозначности



## Понятие Data Science

Анализ данных — область математики и информатики, занимающаяся построением и исследованием наиболее общих математических методов и вычислительных алгоритмов извлечения знаний из экспериментальных (в широком смысле) данных; процесс исследования, фильтрации, преобразования и моделирования данных с целью извлечения полезной информации и принятия решений.

Интеллектуальный анализ данных — это особый метод анализа данных, который фокусируется на моделировании и открытии данных, а не на их описании.



## Понятие Data Science

Data science — as a profession and as an academic discipline unto itself — is new, having been born in the first decade of the 21st century. It is a child born of the mature parental disciplines of scientific methods, data and software engineering, statistics, and visualization.



## Понятие Data Science

The term “Data Science” was coined at the beginning of the 21st Century. It is attributed to William S. Cleveland who, in 2001, wrote “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics”.

About a year later, the International Council for Science: Committee on Data for Science and Technology started publishing the CODATA Data Science Journal beginning April 2002.

Shortly thereafter, in January of 2003, Columbia University began publishing The Journal of Data Science.

## Понятие Data Science

Most organizations realize they have a wealth of data — but not all of them are able to realize its potential value because technological and cultural challenges often stand in the way.

Even though more lines of business are better at leveraging their data for their own purposes than they once were, the value of the data from an enterprise perspective may not yet be fully realized.

In addition, compliance, privacy, and security issues may limit the ways in which the data can be used.

Here are some ways to monetize data:

- Stop Revenue Leaks
- Embrace A New Revenue Model
- Embrace a New Business Model
- Rethink the Definition of Value
- Infer Customer Satisfaction
- Minimize Churn
- Improve Marketing ROI
- Detect Fraud And Piracy



## Понятие Data Science

Какие примеры решений/продуктов, созданных с помощью DS, вы знаете?



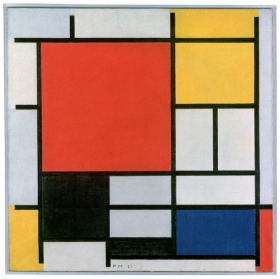
## Понятие Data Science

PRISMA

Prisma transforms your photos and videos into works of art using the styles of famous artists: Van Gogh, Picasso, Levitan, as well as world famous ornaments and patterns. A unique combination of neural networks and artificial intelligence helps you turn memorable moments into timeless art pieces.



## Piet Mondrian “Composition With Red Yellow And Blue”

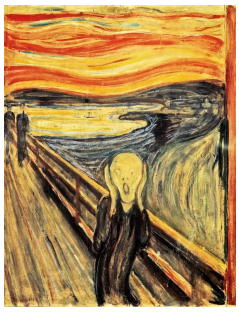


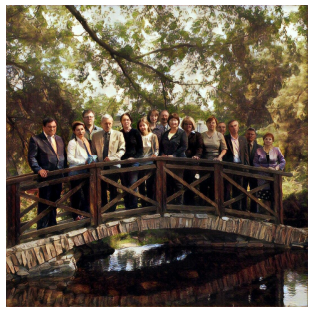


## Edgar Degas “Dancers in blue”



## Edvard Munch “The Scream”





15.12.2016 13

15.12.2016 14

## Понятие Data Science

**x.ai** – персональный ассистент, который планирует встречи

15.12.2016 15

## x.ai

**Scenario**

You and Mary reconnected during Adweek. She follows up a week later to chat more about a partnership opportunity.

1 A meeting request is made

From: Mary <mary@example.com>
To: Greg <greg@example.com>
Subject: Coffee?

Hi Greg,
It was nice to meet you during Adweek. Do you have a little time for coffee to continue our conversation? I can swing by somewhere close to your office.
Thanks,
Mary

2 You reply and cc: Amy

From: Greg <greg@example.com>
To: Mary <mary@example.com>
CC: Amy <amy@x.ai>
Subject: Re: Coffee?

Sure, Mary,
Amy, can you find 30 minutes for coffee at my office?
Cheers,
Greg

15.12.2016 16

## x.ai

From: Amy <amy@x.ai>
To: Mary <mary@example.com>
Subject: Re: Coffee?

Hi Mary,
Happy to get something on Greg's calendar.
Does Tuesday, Mar 3 at 11:00 AM PST work? Alternatively, Greg is available Tuesday, Mar 3 at 4:00 PM PST or Tuesday, Mar 10 at 11:00 AM.
Greg's office is at 77 Geary Street, Suite 500, San Francisco, CA.
Amy

From: Mary <mary@example.com>
To: Amy <amy@x.ai>
Subject: Re: Coffee?

Amy,
Mornings are not so good. So 11am slots are out.
4pm I can do.
Thanks,
Mary

15.12.2016 17

## x.ai

From: Amy <amy@x.ai>
To: Mary <mary@example.com>
Subject: Re: Coffee?

Hi Mary,
Thanks for letting me know. I'll send out an invite.
Amy

4 Amy sends the meeting invite to both you and your guest(s)

**Mary, Greg | Coffee**

WHEN Tue Mar 03, 2015 4:00pm – 4:30pm PST
WHERE 77 Geary Street, Suite 500, San Francisco, CA | Greg's Office
WHO Greg Morgan Host
Mary Adler Guest
Amy Ingram Assistant to Greg

Invitation from x.ai — a personal assistant who schedules meetings for you

15.12.2016 18

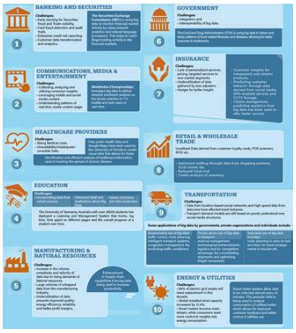
## Понятие Data Science

iCarbonX 碳云智能

**Optimize your life data**
**Know more about yourselves**

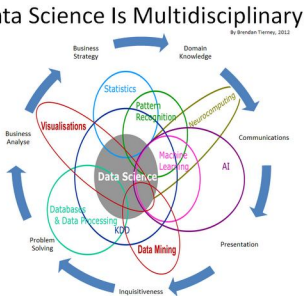
Complete your digital health records by storing genomic data, medical data and daily health data, so as to help you grasp your health condition whenever and wherever you want.

## Основные области применения



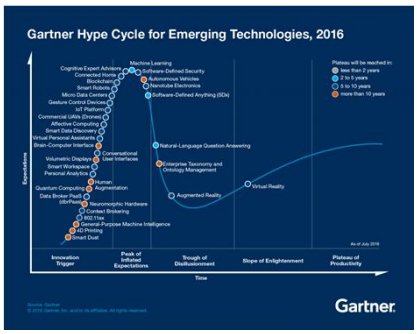



## Понятие Data Science






## Понятие Data Science

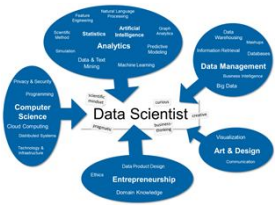


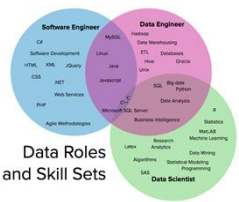



## Кто такой Data Scientist?




## Кто такой Data Scientist?

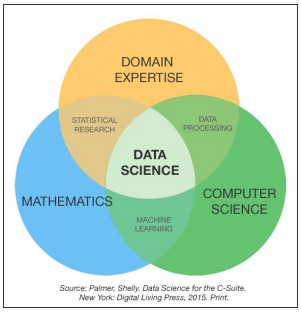



Data Roles and Skill Sets



## В чём отличие Data Science от смежных областей ИТ-индустрии?




Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

## Отличие Data Science от Business Intelligence

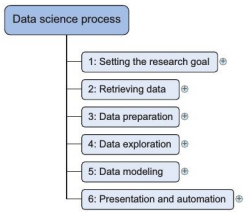
Data Science vs. Business Intelligence

| | Business Intelligence (BI) | Data Science |
|---|---|---|
| Data analysis | Yes | Yes |
| Statistics | Yes | Yes |
| Visualization | Yes | Yes |
| Data Sources | Usually SQL, often Data Warehouse | Less structured (logs, cloud data, SQL, noSQL, text) |
| Tools | Statistics, Visualization | Statistics, Machine Learning, Graph Analysis, NLP |
| Focus | Present and past | Future |
| Approach | Analytic | Scientific |
| Goal | Better strategic decisions | Advanced functionality |

The two fields are closely related. In some ways Data Science is an evolution of BI.



## Процесс Data Science






## Процесс Data Science

- **Setting the research goal.** Data science is mostly applied in the context of an organization. When the business asks you to perform a data science project, you'll first prepare a project charter. This charter contains information such as what you're going to research, how the company benefits from that, what data and resources you need, a timetable, and deliverables.
- **Retrieving data.** The second step is to collect data. You've stated in the project charter which data you need and where you can find it. In this step you ensure that you can use the data in your program, which means checking the existence of, quality, and access to the data. Data can also be delivered by third-party companies and takes many forms ranging from Excel spreadsheets to different types of databases.



## Процесс Data Science

- **Data preparation.** Data collection is an error-prone process; in this phase you enhance the quality of the data and prepare it for use in subsequent steps. This phase consists of three sub-phases: data cleansing removes false values from a data source and inconsistencies across data sources, data integration enriches data sources by combining information from multiple data sources, and data transformation ensures that the data is in a suitable format for use in your models.
- **Data exploration.** Data exploration is concerned with building a deeper understanding of your data. You try to understand how variables interact with each other, the distribution of the data, and whether there are outliers. To achieve this you mainly use descriptive statistics, visual techniques, and simple modeling. This step often goes by the abbreviation EDA, for Exploratory Data Analysis.



## Процесс Data Science

- **Data modeling or model building.** In this phase you use models, domain knowledge, and insights about the data you found in the previous steps to answer the research question. You select a technique from the fields of statistics, machine learning, operations research, etc. Building a model is an iterative process that involves selecting the variables for the model, executing the model, and model diagnostics.
- **Presentation and automation.** Finally, you present the results to your business. These results can take many forms, ranging from presentations to research reports. Sometimes you'll need to automate the execution of the process because the business will want to use the insights you gained in another project or enable an operational process to use the outcome from your model.

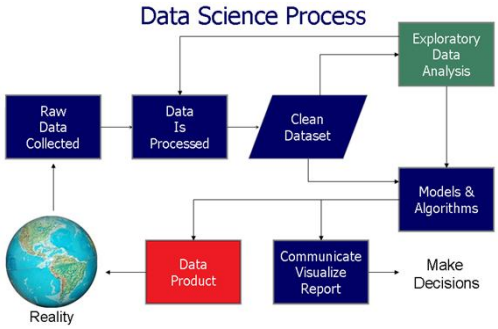


## Процесс Data Science

The previous description of the data science process gives you the impression that you walk through this process in a linear way, but in reality you often have to step back and rework certain findings.

For instance, you might find outliers in the data exploration phase that point to data import errors. As part of the data science process you gain incremental insights, which may lead to new questions.

## Процесс Data Science



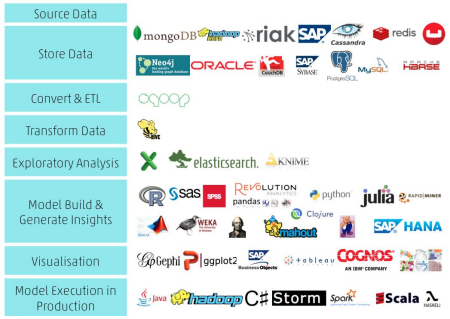



## Технологии и инструментарий

- Linear Regression
- Logistic Regression
- Jackknife Regression
- Density Estimation
- Confidence Interval
- Test of Hypotheses
- Pattern Recognition
- Clustering - (aka Unsupervised Learning)
- Supervised Learning
- Time Series
- Decision Trees
- Random Numbers
- Monte-Carlo Simulation
- Bayesian Statistics
- Naive Bayes
- Principal Component Analysis - (PCA)
- Ensembles
- Neural Networks
- Support Vector Machine - (SVM)
- Nearest Neighbors - (k-NN)
- Feature Selection - (aka Variable Reduction)
- Indexation / Cataloguing
- (Geo-) Spatial Modeling
- Recommendation Engine
- Search Engine
- Attribution Modeling
- Collaborative Filtering
- Rule System
- Linkage Analysis
- Association Rules
- Scoring Engine
- Segmentation
- Predictive Modeling
- Graphs
- Deep Learning
- Game Theory
- Imputation
- Survival Analysis
- Arbitrage
- Lift Modeling
- Yield Optimization
- Cross-Validation
- Model Fitting
- Relevancy Algorithm
- Experimental Design

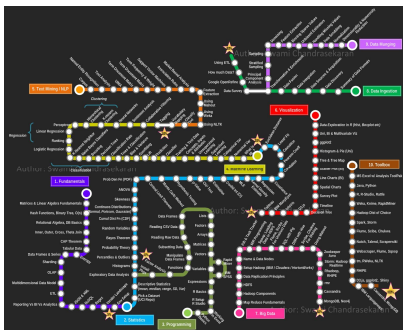


## Технологии и инструментарий






## Технологии и инструментарий





## Machine Learning

Machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959).



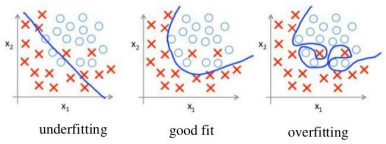
## Machine Learning

Basic tasks:

- In classification, inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more (multi-label classification) of these classes. This is typically tackled in a supervised way. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are "spam" and "not spam".
- In regression, also a supervised problem, the outputs are continuous rather than discrete.
- In clustering, a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.
- Collaborative filtering is a technique used by recommender systems. It is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating).

# Classification

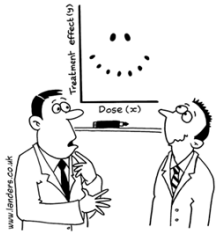
Overfitting and underfitting



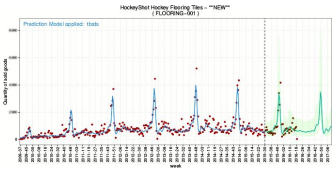
underfitting good fit overfitting



# Regression



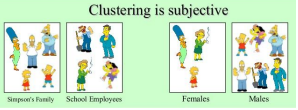

"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."



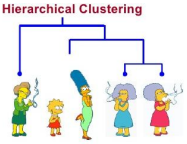


# Clustering

Clustering is subjective



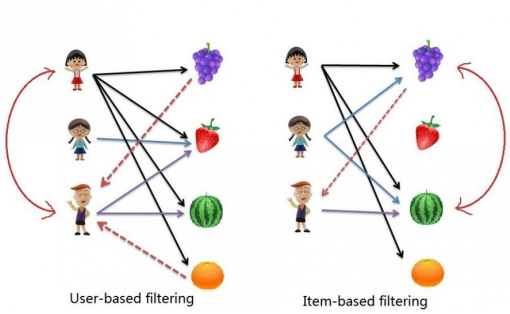

Hierarchical Clustering



Partition-based Clustering



# Collaborative filtering






# Основная литература


- IData Science: an introduction to. Jeffrey Stanton, Syracuse University (https://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf)
- Introducing Data Science: Big data, machine learning, and more, using Python tools. Davy Cielen, Arno D. B. Meysman, and Mohamed Ali. May 2016. ISBN 9781633430037. 320 pages (https://www.manning.com/books/introducing-data-science)
- 8 Ways To Monetize Data. Lisa Morgan. InformationWeek: Connecting The Business Technology Community (http://www.informationweek.com/big-data/big-data-analytics/8-ways-to-monetize-data/d/d-id/1323932)
- Data Science: an introduction. Wikibooks (https://en.wikibooks.org/wiki/Data_Science:_An_Introduction)
- 16 analytic disciplines compared to data science. Vincent Granville. Data Science Central: the online resource for big data practioners (http://www.datasciencecentral.com/profiles/blogs/17-analytic-disciplines-compared)
- How Top 10 Industries Use Big Data Applications. Data Science Association (http://www.datascienceassn.org/content/how-top-10-industries-use-big-data-applications)
- Data science. Wikipedia (https://en.wikipedia.org/wiki/Data_science)
- 40 Techniques Used by Data Scientists (http://www.datasciencecentral.com/profiles/blogs/40-techniques-used-by-data-scientists)