LECTURE 1: COURSE INTRO & BASIC INTRODUCTION TO DATA SCIENCE

EKATERINA MURAVLEVA

OVERVIEW OF THE COURSE

Lecture 1: General course information, CRISP-DM methodology

Lecture 2: Supervised learning/unsupervised learning. Classification/regression problems. Accuracy metrics (precision, recall, ROC-AUC scores). Concept of loss functions, overfitting / underfitting.

Lecture 3: Classical ML: Linear regression, logistic regression, support vector machine

Lecture 4: Classical ML: Decision trees, random forests, boosting.

Lecture 5: Classical ML: Dimensionality reduction: linear, non-linear methods.

Lecture 6: Classical ML: Clustering methods

Lecture 7: Basic neural networks

Lecture 8: Scalable algorithms

GENERAL COURSE INFORMATION

Course instructor: Ekaterina Muravleva

TAs: Daniil Bershatskiy, Daria Frolova,

Daniil Merkulov, Simon Polyanskiy, Vlad Trifonov

2 Homeworks (the first homework is planned for the end of this week +

Team Project

Plagiarism: grade is divided on number of people with similar works

PLAN OF LECTURE 1

- Introduction to CRISP-DM Methodology
- Overview of CRISP-DM: 6 stages (Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment)

DATA ANALYSIS

Data analysis is about data gathering, prediction and decision making

Data Science is about finding patterns in data through analysis and making future predictions

Some decision that can be made:

1. Better decisions (should we choose A or B)
2. Predictive analysis (what will happen next)
3. Pattern discoveries (what is hidden in the data).

HOW WE DO DATA ANALYSIS

Data analysis is everywhere: in marketing, production, bioinformatics, optimization of processes, medical diagnosis, image classification/ generation/etc.

Can we do this in a systematic way?

There are several models of data mining.

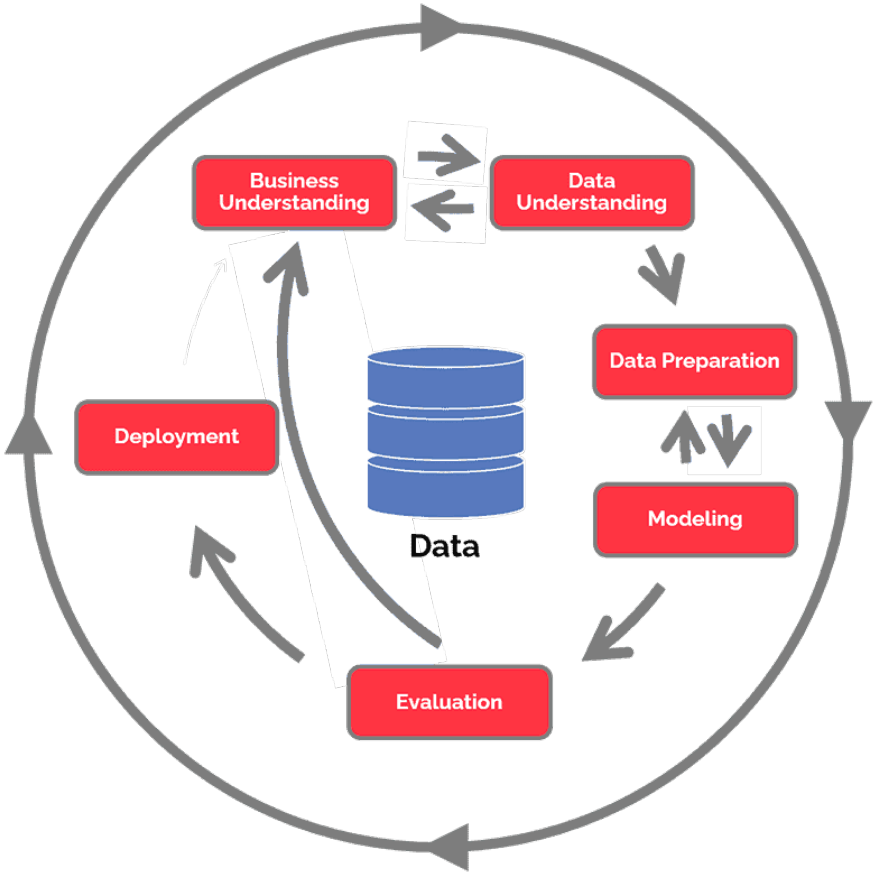
WHAT IS CRISP-DM

CRISP-DM stands for ‘The CRoss Industry Standard Process for Data Mining’

It is a process model that serves as a basis for data science process

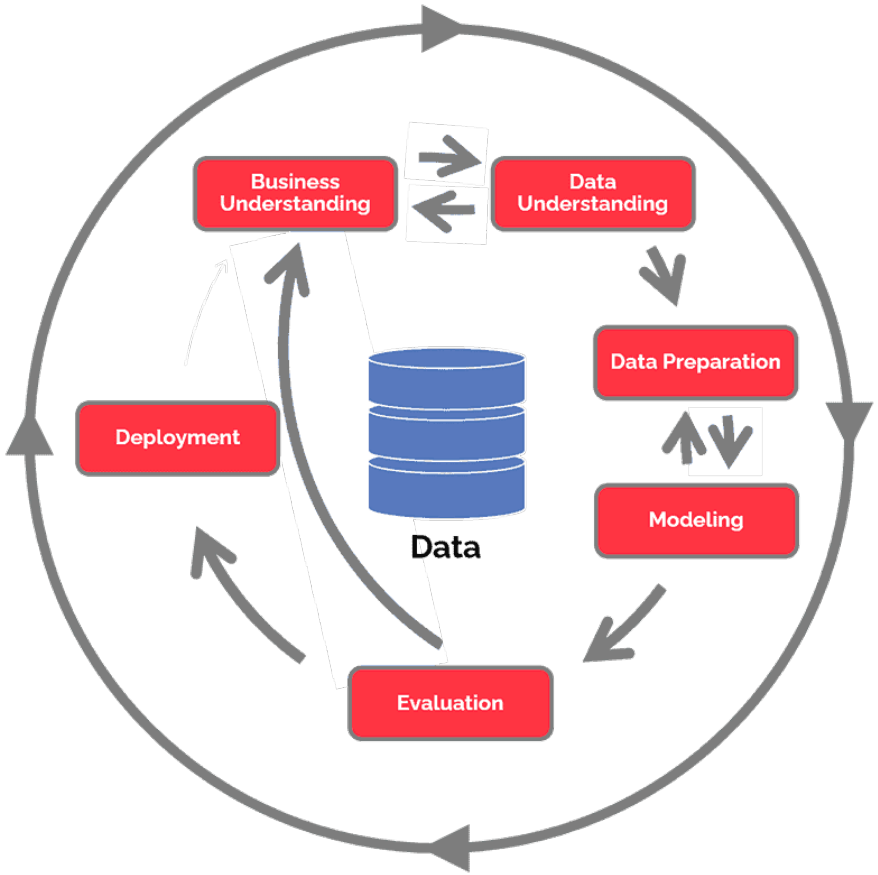
It has been introduced in 1999

Became the most common methodology for data mining, analytics and data science projects



WHAT ARE 6 CRISP-DM STAGES?

1. Business understanding
2. Data understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment



[PDF] **CRISP-DM**: Towards a standard process model for data mining

R Wirth, J Hipp - Proceedings of the 4th international conference on the …, 2000 - cs.unibo.it

… Basically, we used two resources for the development of the specialized process model, the **CRISP-DM** Generic User **Guide** and the documentation of three case studies, which …

☆ Save 99 Cite Cited by 1997 Related articles All 7 versions »

DISTRIBUTION BETWEEN STAGES IN KDDM (KNOWLEDGE DISCOVERY AND DATA MINING)

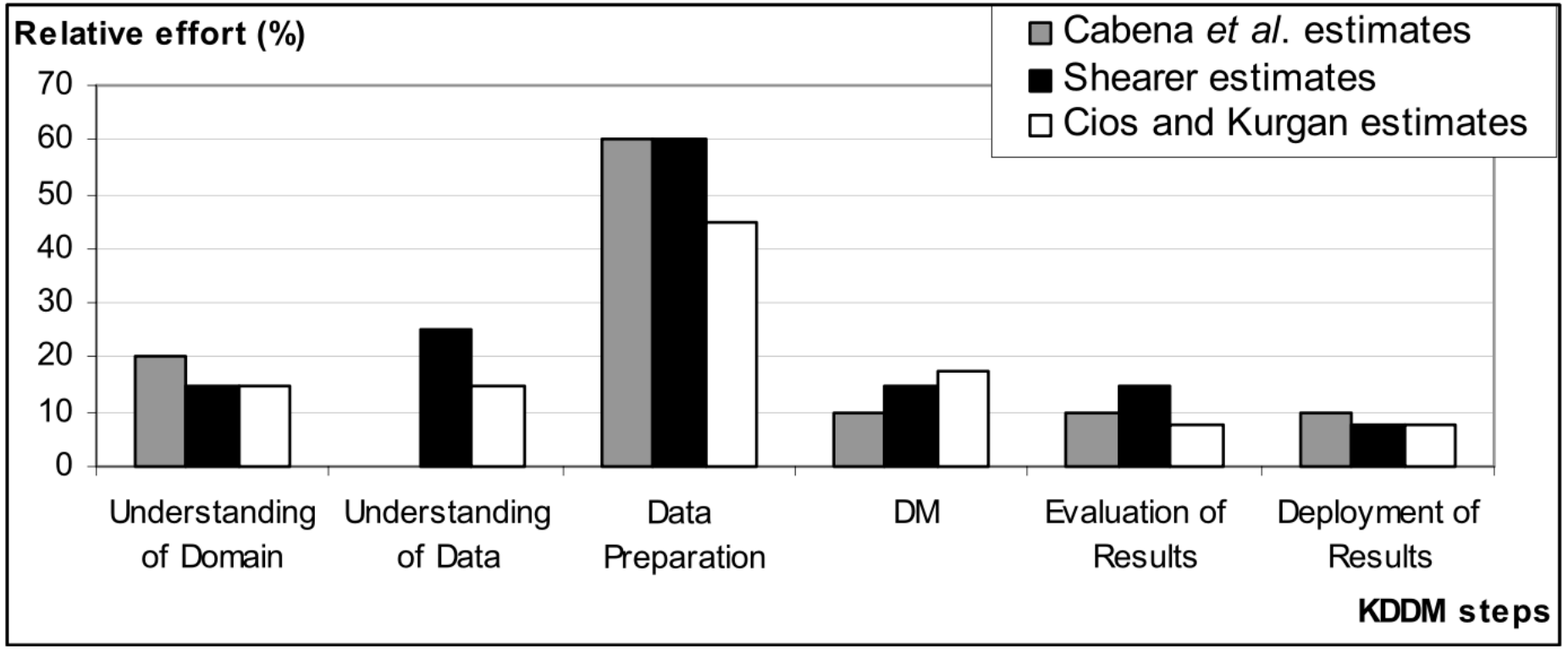


Figure 1 Relative effort spent on specific steps in the KDDM process

I. BUSINESS UNDERSTANDING

Any good projects starts from the understanding of customer needs. Data mining is no exception!

1. **Determine business objectives**. From the CRISP-DM guide: you should thoroughly understand from a business perspective, what the customer wants to accomplish, and define **business success criteria**
2. **Assess situation.** Determine resources availability, project requirements, risks and conduct cost-benefit analysis
3. **Determine data mining goals.** What success looks like from a technical data mining perspective
4. **Produce project plan.** Select technologies and tools and define detailed plans for each project phase.

II. DATA UNDERSTANDING

Any good projects starts from the understanding of customer needs.

Data mining is no exception!

1. **Collect initial data**. It is clear. Typical: how many images you need for your classifier (no exact answer, just random guess + experience from similar projects)
2. **Describe data.** Document the properties of data (type of data, data format, units of each data sample, especially from different sources — very important)
3. **Explore data.** Query data, visualize it, do some simple statistical analysis
4. **Verify data quality.** Don't believe that the data is correct, check it (and then check again!). Try to 'feel' your data! Hint: garbage in, garbage out.

III. DATA PREPARATION

A common rule of thumb is that 80% of the project is data preparation

1. **Select data**. Which datasets will be used for training and why
2. **Clean data.** Often, the lengthiest task (again, garbage in - garbage out). Wrong data can be removed/corrected
3. **Construct data (or feature engineering).** Build new features from the existing ones (this part is becoming more and more automated with deep learning / AutoML).
4. **Verify data quality.** Don't believe that the data is correct, check it (and then check again!). Try to 'feel' your data! Hint: garbage in, garbage out.

IV. MODELING

Often, considered to be the most exiting part; often — the shortest (except for reaching SOTA/really challenging tasks).

1. **Select modeling techniques**. Algorithms: regression/neural networks
2. **Generate test design.** Split data intro training, test and validation sets
3. **Build a model** Sometimes exciting, sometimes reg = model.fit(X, y)
4. **Assess a model.** In real life (industry/academia) you always have several models to compare, so-called **baselines.** So, you need to compare them using different criteria, and the differences should be **statistically significant**

V. EVALUATION

Assess results of the Part IV focuses on technical model comparison;

EVALUATION is broader.

1. **Evaluate results**. Do the models meet the business success criteria? Which one(s) we approve for the business (or a research paper)?
2. **Review process.** Did we miss something? Did we do all correct?
3. **Determine the next steps.** Do we need to proceed to deployment? Do we need to iterate previous steps?

VI. DEPLOYMENT

Deployment can be easy (generate a report) or complex — implement in an enterprise. Model is not useful, if the customer can not access it results.

1. **Plan deployment**. Develop and document a plan for deploying the model.
2. **Plan monitoring and maintenance.** Develop a plan to avoid the issues during the operational phase (or post-operational phase) of the model.
3. **Final report.** The project team documents a summary of the project which might include a final presentation of data mining results
4. **Review project (again!)** Do a project retrospective: what gone well, what could be better, et

(IMAGINARY) CASE STUDY

The case study revolves around a retail company called "**Fashionista**", which specializes in selling trendy clothing and accessories for young adults.

The company has been struggling with declining sales and wants to improve their marketing strategy to attract more customers and increase revenue.

CASE STUDY: BUSINESS UNDERSTANDING

The first step in the CRISP-DM process is to understand the business objectives and requirements.

In this case, the objective is to increase sales and revenue for Fashionista. The company's management team has identified that their target market is primarily young adults aged 18-25, who are fashion-conscious and have a moderate to high disposable income.

They also want to identify the most effective marketing channels to reach their target audience.

CASE STUDY: DATA UNDERSTANDING:

In this phase, data is collected and analyzed to gain a better understanding of the current state of the business.

Fashionista's marketing team gathers data from various sources such as sales reports, customer surveys, and social media analytics.

They discover that the majority of their sales come from in-store purchases, but there is a growing trend of online shopping among their target audience.

CASE STUDY: DATA PREPARATION

Once the data is collected, it needs to be cleaned and prepared for analysis.

The marketing/data science team cleans up the data by removing any duplicates or irrelevant information.

They also merge data from different sources to create a comprehensive dataset for analysis.

CASE STUDY: DATA MODELING

In this phase, the data science team uses various data mining techniques to analyze the data and identify patterns and trends.

They use clustering analysis to group customers based on their buying behavior and preferences. They also use association analysis to identify which products are frequently purchased together (recommender systems).

CASE STUDY: EVALUATION

After analyzing the data, the data science team evaluates the results to determine which marketing channels are most effective in reaching their target audience.

They find that social media platforms such as Instagram and Snapchat have a high engagement rate among their target market, and online advertising is also gaining popularity.

CASE STUDY: DEPLOYMENT

Based on the results of the evaluation phase, a new marketing strategy for Fashionista is created.

They decide to increase their presence on social media platforms and invest more in online advertising. They also plan to launch a loyalty program to incentivize customers to make repeat purchases.

CASE STUDY: DEPLOYMENT

Based on the results of the evaluation phase, a new marketing strategy for Fashionista is created.

They decide to increase their presence on social media platforms and invest more in online advertising. They also plan to launch a loyalty program to incentivize customers to make repeat purchases.

RECAP OF LECTURE 1

- Introduction to CRISP-DM Methodology
- Overview of CRISP-DM: 6 stages (Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment)

NEXT LECTURE

Supervised learning/unsupervised learning.

Classification/regression problems.

Accuracy metrics (precision, recall, ROC-AUC scores).

Concept of loss functions, overfitting / underfitting.

QUESTIONS?