Doc. RNDr. Irena Holubová, Ph.D. & PROFINIT

DATA SCIENCE

NDBI048

Data Science Life Cycle, CRISP-DM Methodology

https://www.ksi.mff.cuni.cz/~holubova/NDBI048/

OUTLINE

- 10 questions to ask
- Data Science life cycle
- CRISP-DM methodology
- Other approaches
  - KDD, SEMMA

10 QUESTIONS TO ASK BEFORE STARTING A DATA SCIENCE PROJECT

1. **What is the business requesting?**
   - „Aggressively“ figure out their exact requests
     - Not just fuzzily
2. **What does the business need?**
   - Henry Ford: *“If I had asked people what they wanted, they would have said faster horses.”*
   - Don’t just build a faster horse
3. **Who are all of the stakeholders and what are their individual needs?**
   - Project’s impact likely extends beyond the requester
   - Pro-active stakeholder identification
     - Mitigates the risk of ignoring key stakeholders
     - Further value creation across the organization

https://www.datascience-pm.com/10-questions-to-ask-before-starting-a-data-science-project/

10 QUESTIONS TO ASK BEFORE STARTING A DATA SCIENCE PROJECT

4. Do the stakeholders have clear expectations?

- Not just another software project
- Set expectations for touch-points (e.g. review sessions), a highly visible project roadmap (will need to change!), …

5. What is the simplest solution that adds value to the stakeholders?

- Start small and deliver something of value as quickly as possible
  - E.g., an analysis that establishes the baseline, a mockup dashboard, …
- Opportunity to provide feedback => you know you're on the right path
- A "failed" deliverable adds value
  - "Fail fast", learn problems early
- A simple solution might even solve the problem

10 QUESTIONS TO ASK BEFORE STARTING A DATA SCIENCE PROJECT

6. What is the value of this project? How will it be measured?

- Helps prioritize projects
- Focus on maximizing/minimizing the target variable(s) that are most important

7. Why do this project?

- Just focusing only on the "what" is not sufficient
- A clear and common vision of the project's impact and its "why"
  - Motivation for executive sponsorship, data science development team, ...

8. What are the risks?

- Fundamental process in any project
- "What could go wrong?"
  - Various perspectives: technical, market, societal, legal, security, …
- Who is responsible for what

10 QUESTIONS TO ASK BEFORE STARTING A DATA SCIENCE PROJECT

9. What people and resources are needed?

- Who do you need to develop the solution? How much time they'll need?
- What data sources will you need? Where are they? Can you purchase them? Can you start collecting the data? What security / firewall requests will you need? Computing resources? Systems integrations?
- Bring together IT, business, and data science project team to avoid a disjointed approach

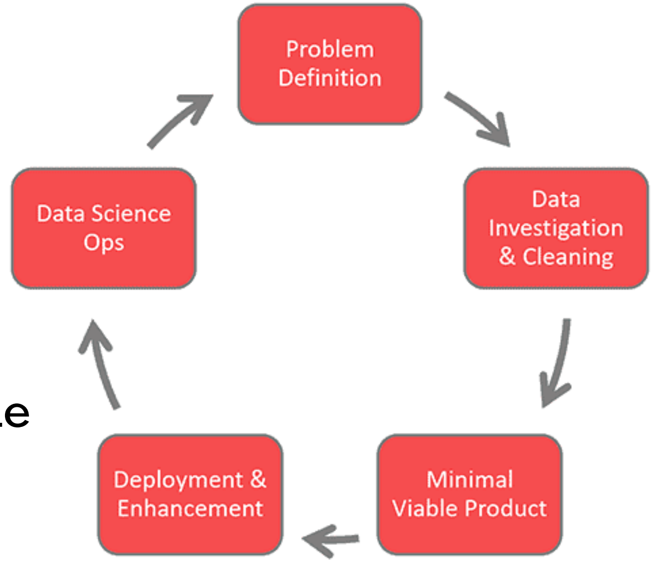
10. What other questions should be answered?

- The meta-question
- Ask yourself, your team, and your stakeholders some variant of: "What other key questions do we need to answer before committing to this proposed project?"

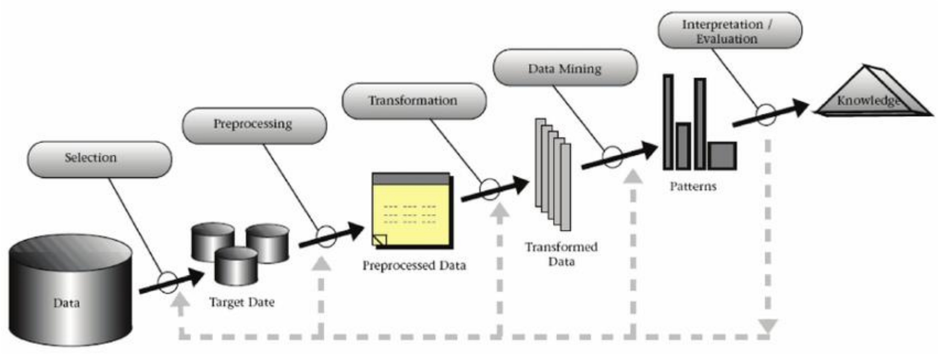
DATA SCIENCE LIFE CYCLE

https://www.datascience-pm.com/data-science-life-cycle/

DATA SCIENCE LIFE CYCLE



- An iterative set of steps to deliver a data science project
  - Different data science projects / teams = specific data science life cycle
    - E.g., just the data, modeling, and assessment steps; from business understanding to deployment; …
  - Most tend to flow through the same general life cycle
- Several steps
  - Typically not linear
  - The course depends on the particular DS project



CLASSICAL DATA SCIENCE LIFE CYCLES (FROM '90S)

- **CRISP-DM**: The CRoss Industry Structured Process for Data Mining
  - The most popular methodology
  - Broader-focused than the others
- Knowledge Discovery in Database (**KDD**) **Process**
  - General process of discovering knowledge in data through data mining, extraction of patterns, machine learning, statistics, and database systems.
- **SEMMA** (Sample, Explore, Modify, Model, and Assess)
  - Developed by SAS
  - To guide users through tools in SAS Enterprise Miner for data mining problems

OTHER DATA SCIENCE LIFE CYCLES

- **OSEMN** (Obtain, Scrub, Explore, Model, and iNterpret)
  - Steps: Business Understanding, Data Acquisition and Understanding, Modeling, Deployment, and Customer Acceptance
- **Microsoft TDSP** (the Team Data Science Process)
  - Combines many modern agile practices with a life cycle similar to CRISP-DM
- **Domino Data Labs Life Cycle**
  - Steps: Ideation, Data Acquisition and Exploration, Research and Development, Validation, Delivery, and Monitoring
- …

CRISP-DM

CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)

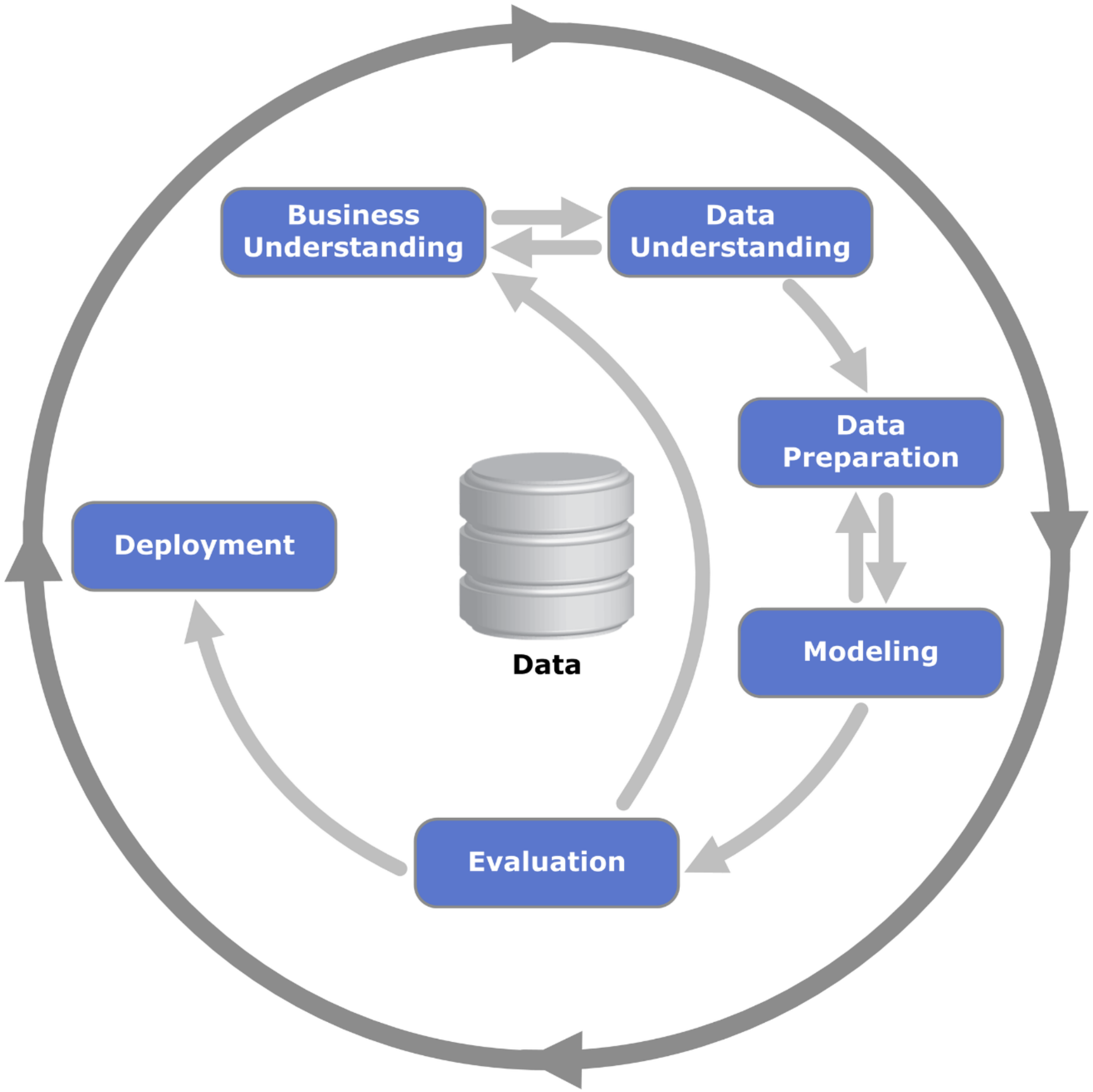
- The most widely used form of data-mining model
  - *"de facto standard for developing data mining and knowledge discovery projects"*
- Supported and promoted by
  - data mining software vendors
  - practitioners in data mining and in data warehousing
- Advantages:
  - Industry, tool, and application neutral
- Disadvantages:
  - Does not perform project management activities

CRISP-DM – HISTORY

- 1996 – created
- 1997 – became a European Union project
  - Led by five companies with different experiences in data mining
- **1999** – the first version was presented
- 2000 – published
- 2006 … 2008 - CRISP-DM 2.0 Special Interest Group was formed
  - Discussed updating of the CRISP-DM process model
  - Unknown status

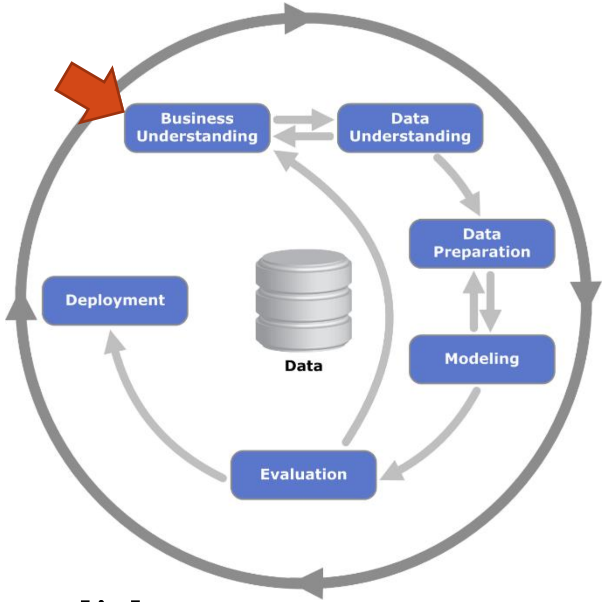
CRISP-DM PHASES

I. Business Understanding
II. Data Understanding
III. Data Preparation
IV. Modeling
V. Evaluation
VI. Deployment



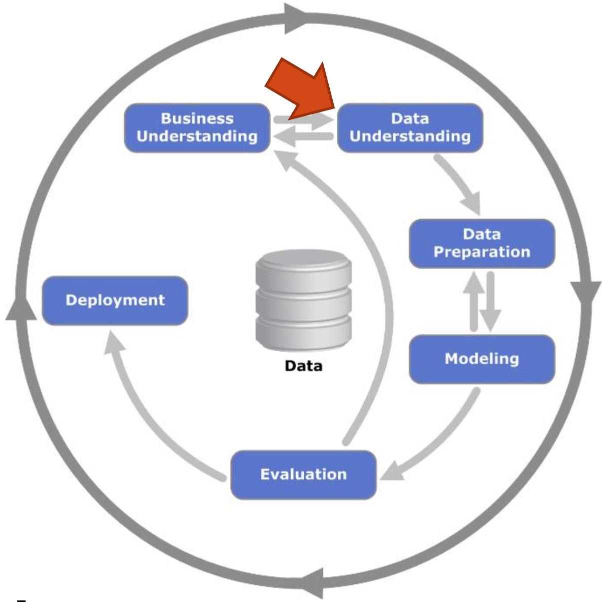
https://www.datascience-pm.com/crisp-dm-2/

I. BUSINESS UNDERSTANDING



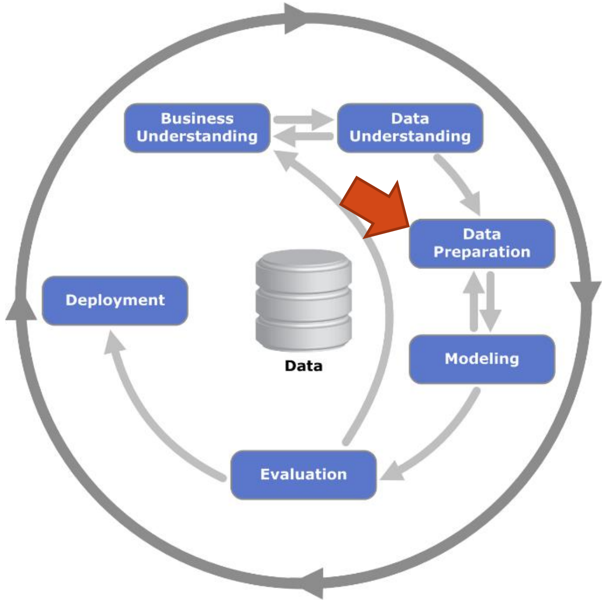
- Understanding the objectives and requirements of the project.
- **Determine business objectives**
  - Understand, from a business perspective, what the customer wants to accomplish
  - Define business success criteria
- **Assess situation**
  - Determine resources availability, project requirements, assess risks and contingencies
  - Conduct a cost-benefit analysis
- **Determine data mining goals**
  - Define what success looks like from a technical data mining perspective
- **Produce project plan**
  - Select technologies and tools
  - Define detailed plans for each project phase

II. DATA UNDERSTANDING



- To identify, collect, and analyze the data sets
- **Collect initial data**
  - Acquire the necessary data and (if necessary) load it into your analysis tool
- **Describe data**
  - Examine the data and document its surface properties
    - Data format, number of records, field identities, …
- **Explore data**
  - Dig deeper into the data
  - Query, visualize, identify relationships
- **Verify data quality**
  - How clean/dirty is the data?
  - Document any quality issues

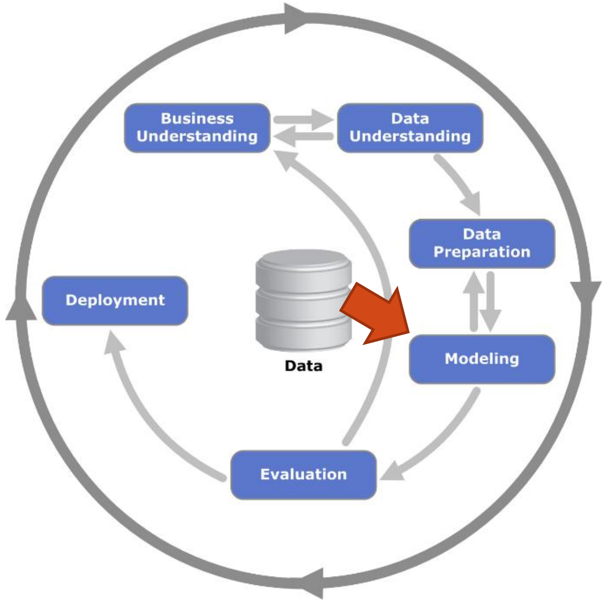
III. DATA PREPARATION



- Prepares the final data set(s) for modeling
- **Select data**
  - Which data sets will be used
  - Document reasons for inclusion/exclusion
- **Clean data**
  - Correct, impute, or remove erroneous values
- **Construct data**
  - Derive new attributes that will be helpful
    - E.g., derive someone's BMI from height and weight
- **Integrate data**
  - Create new data sets by combining data from multiple sources
- **Format data**
  - Re-format data as necessary.
    - E.g., Convert string values that store numbers to numeric values

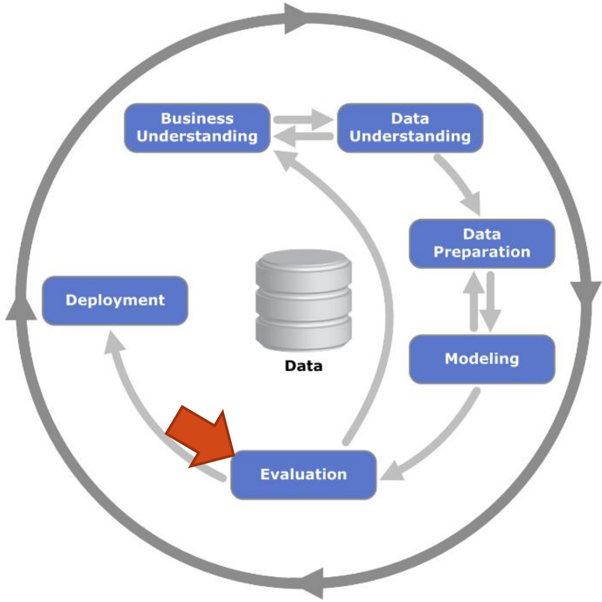
Often the lengthiest task

IV. MODELING

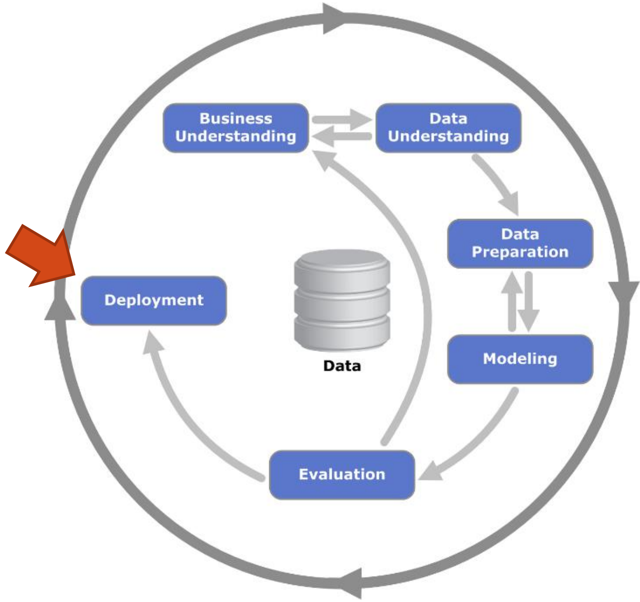


- Build and assess various models using different modeling techniques
- **Select modeling techniques**
  - Which algorithms to try (e.g. regression, neural network, …)
- **Generate test design**
  - E.g., split the data into training, test, and validation sets
- **Build model**
  - E.g. , reg = LinearRegression().fit(X, y)
- **Assess model**
  - Interpret the model results based on domain knowledge, pre-defined success criteria, and test design
- CRISP-DM guide: "iterate model building and assessment until you strongly believe that you have found the best model(s)"
- Practice: … until you find a "good enough" model, proceed through the CRISP-DM lifecycle, then further improve the model in future iterations

V. EVALUATION



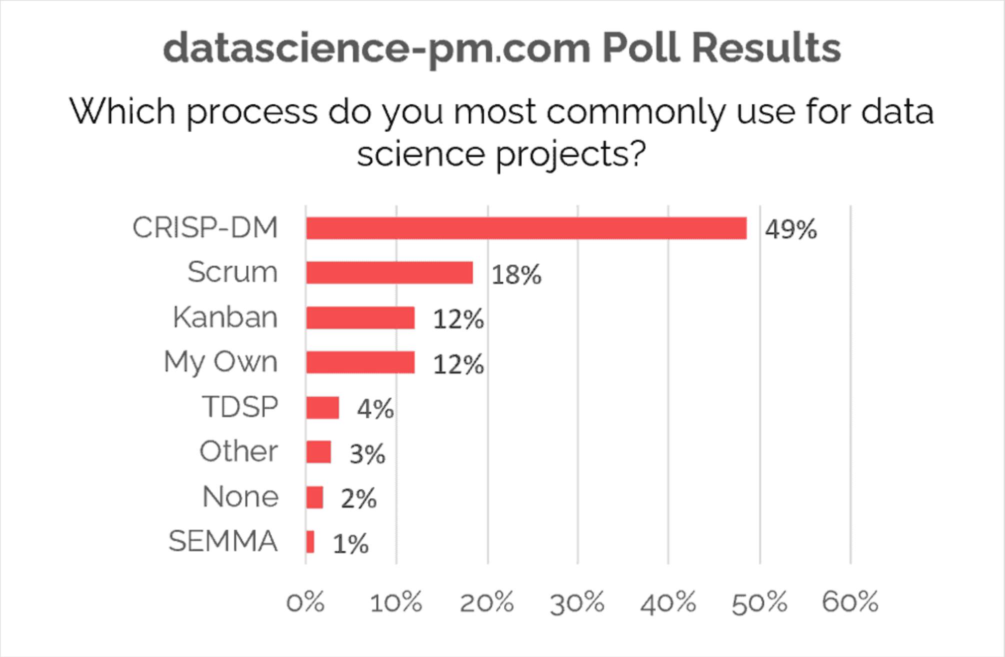
- Looks more broadly at which model best meets the business
- **Evaluate results**
  - Do the models meet the business success criteria?
  - Which one(s) should we approve for the business?
- **Review process**
  - Review the work accomplished
  - Was anything overlooked? Were all steps properly executed?
  - Summarize findings and correct anything if needed
- **Determine next steps**
  - Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects

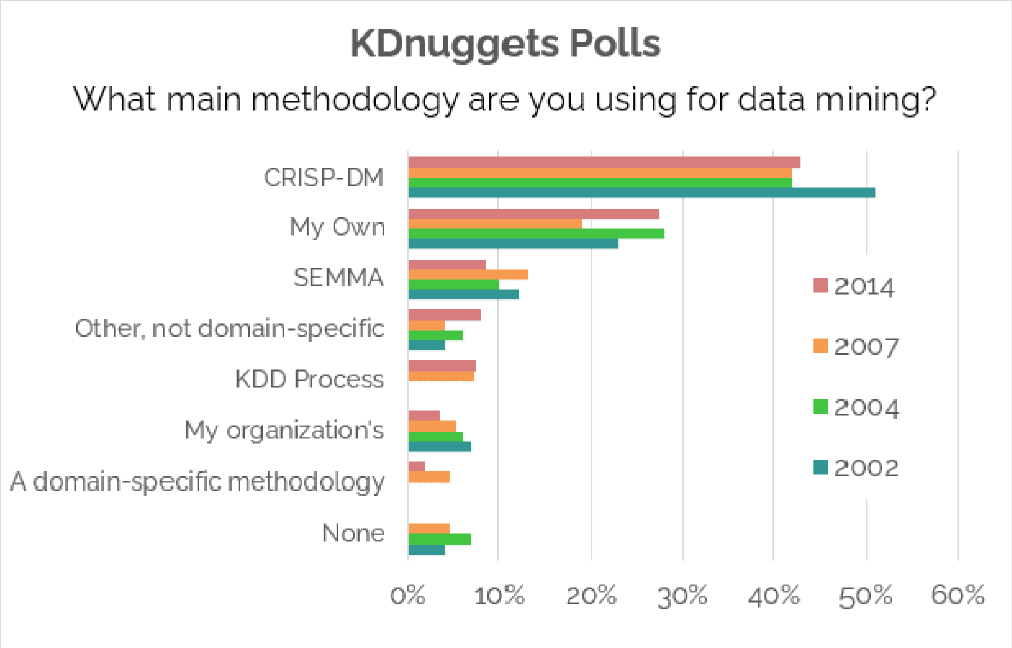


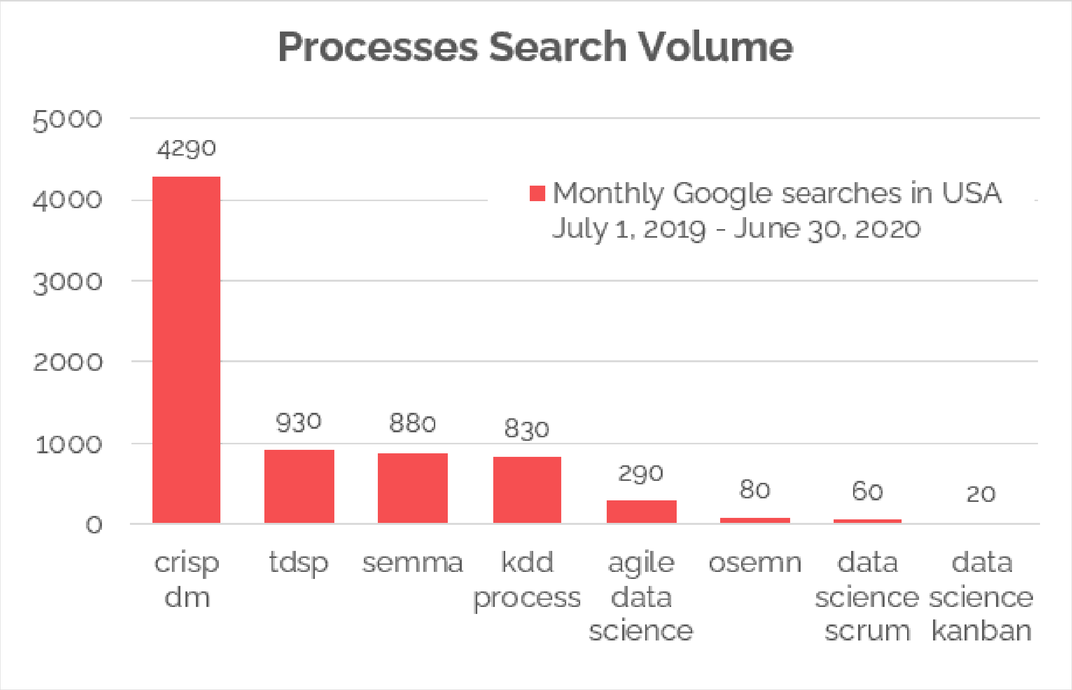
VI. DEPLOYMENT

- Complexity of this phase varies widely
- **Plan deployment**
  - Develop and document a plan for deploying the model
- **Plan monitoring and maintenance**
  - Develop a monitoring and maintenance plan
- **Produce final report**
  - A summary of the project which might include a final presentation of data mining results
- **Review project**
  - Conduct a project retrospective about how to improve in the future
- CRISP-DM does not outline what to do after the project ("operations")

HOW POPULAR IS CRISP-DM?







https://www.datascience-pm.com/crisp-dm-2/

RECOMMENDATIONS

- Iterate quickly
  - Don’t fall into a waterfall trap by working thoroughly across layers of the project
  - Deliver thin vertical slices of end-to-end value.
- Document enough…but not too much
- Don’t forget modern technologies
  - E.g., Add steps to leverage cloud architectures, git version control, ...
- Set expectations
  - CRISP-DM lacks communication strategies with stakeholders
- Combine with a project management approach
  - CRISP-DM is not truly a project management approach

KDD PROCESS

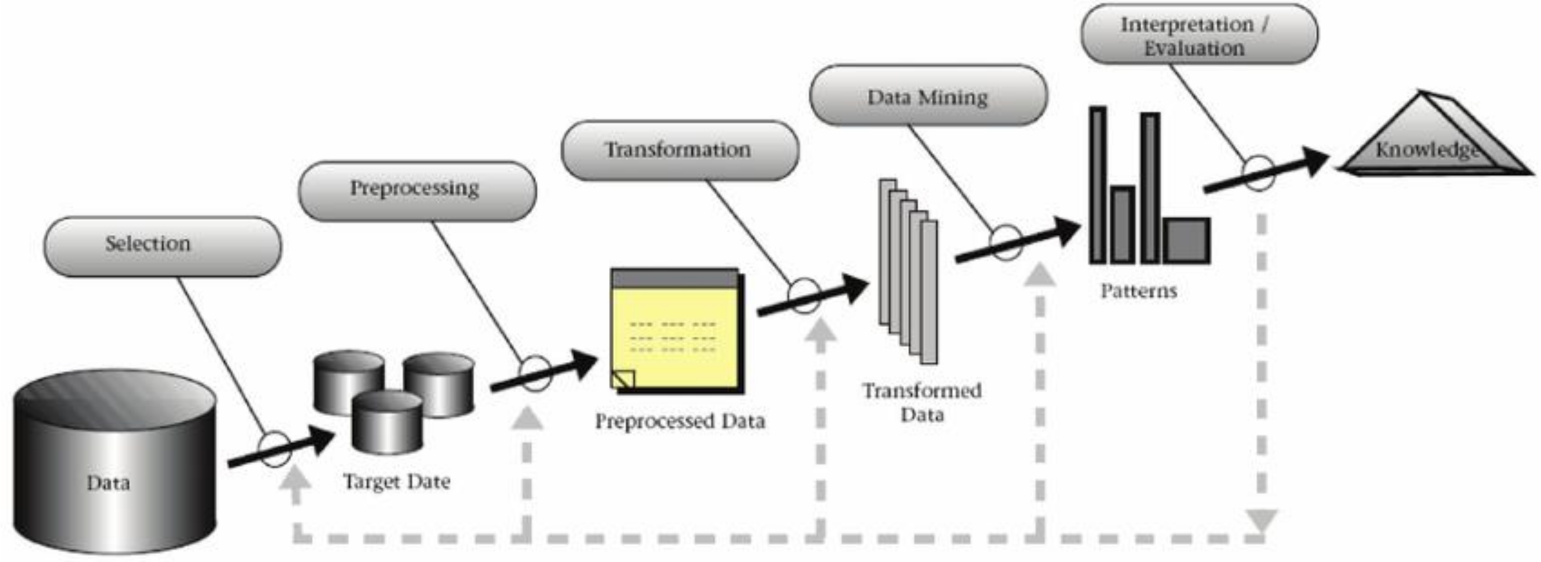
KDD (KNOWLEDGE DISCOVERY IN DATABASES) PROCESS

- **1989**
- Overall process of collecting data and methodically refining it
- Term "data mining" is often interchanged with KDD
- Use cases:
  - Market forecasting – consumer trends, product focus, …
  - Anomaly identification – „holes" in a process, security vulnerabilities, …
- Cons:
  - Not a full project management approach
  - Outdated – does not address modern realities of data science projects
    - Big Data, ethics, …

Fayyad, U. M. et al. 1996. From data mining to knowledge discovery: an overview. In Advances in knowledge discovery and data mining. AAAI Press / The MIT Press.

https://www.datascience-pm.com/kdd-and-data-mining/

KDD PROCESS



- **Selection**: Targeted data is determined, variables for knowledge discovery are determined
- **Pre-processing**: Improving the data being worked (cleaning)
  - Predictive models are established to predict similarly faulty, missing, attributional mismatched data to remove
- **Transformation**: Converting the pre-processed data to the fully utilizable kind
  - Narrowing the variety, establishing data attributes for forthcoming evaluation, organization (sorting) of the information
- **Data Mining**: Sifting through the transformed data to seek out patterns of interest
  - Patterns are graphed, trended, and charted
  - Involves grouping, clustering, and regression
- **Interpretation/Evaluation**: Data is handed off for interpretation and documentation
  - Cleaned, converted, picked apart based on relevant attributes, and framed into visual representations

SEMMA

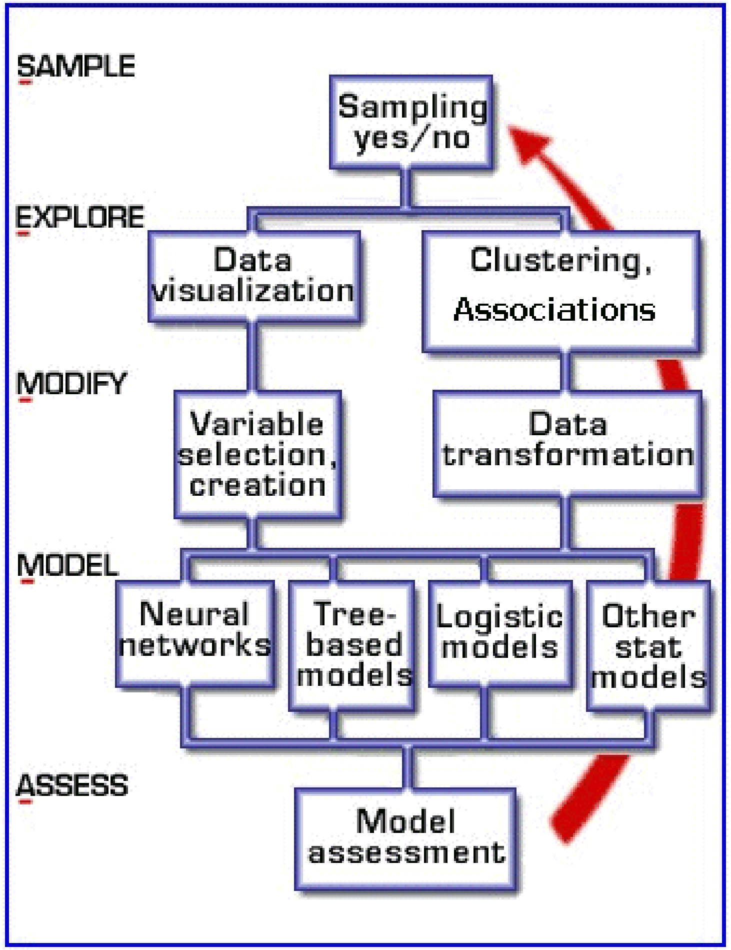
SEMMA (SAMPLE, EXPLORE, MODIFY, MODEL, AND ASSESS)

- Developed by the SAS Institute as the process of data mining
  - *„to uncover previously unknown patterns which can be utilized as a business advantage“*
  - Logical organization of the functional tool set of SAS Enterprise Miner
    - Enables to carry out the core tasks of data mining
- SAS Institute – producer of statistics and business intelligence software
- Use cases: fraud identification, customer retention and turnover, database marketing, customer loyalty, bankruptcy forecasting, market segmentation, risk, affinity, and portfolio analysis

https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm

SEMMA

- **Sample**: Vast input dataset => choose a subset of the appropriate volume
  - Large enough to contain the significant information, small enough to process
  - Identify variables or factors (both dependent and independent) influencing the process
- **Explore**: Study relationships between data elements, identify gaps in the data
  - Multivariate analysis – studies the relationships between variables
  - Univariate analysis – looks at each factor individually to understand its part in the overall scheme
- **Modify**: Data is parsed and cleaned
- **Model**: Applies a variety of data mining techniques in order to produce a projected model
- **Access**: Model is evaluated for how useful and reliable it is for the studied topic



SUMMARY OF THE CORRESPONDENCES BETWEEN KDD, SEMMA AND CRISP-DM

KDD	SEMMA	CRISP-DM
Pre KDD	-------------	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	-------------	Deployment

Ana Azevedo, Manuel Filipe Santos. KDD, SEMMA and CRISP-DM: a parallel overview. IADIS European Conference on Data Mining 2008, Amsterdam, 2008.

REFERENCES

- https://www.datascience-pm.com/
- Ana Azevedo, Manuel Filipe Santos. KDD, SEMMA and CRISP-DM: a parallel overview. IADIS European Conference on Data Mining 2008, Amsterdam, 2008.
- CRISP-DM 1.0 Step-by-step data mining guide https://www.the-modeling-agency.com/crisp-dm.pdf
- https://www.datascience-pm.com/kdd-and-data-mining/
- https://www.datascience-pm.com/semma/
- https://mbi.vse.cz/public/cs/obj/METHOD-113