Data Science - Platforms

Lecture Notes

Prof. Dr. P. Erdelt

Berliner Hochschule für Technik

WiSe 21/22

Last change: October 5, 2021

Part I

Data Science

What is Data Science Workflow?

Workflow

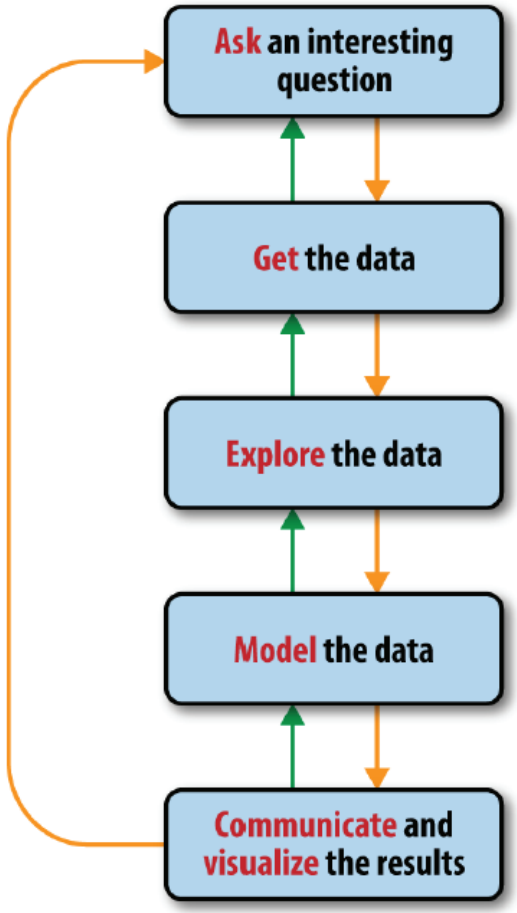
Definition (Workflow)

A **Workflow**[a] consists of

- an orchestrated and repeatable pattern of activity
- organization of resources into processes.

[a] Software AG, https://www.ftb.ca.gov/aboutFTB/Projects/ITSP/BPM_Glossary.pdf

Data Science Workflow



Source: J. Blitzstein, H. Pfister, Harvard data science course

What is Data Science?

Source: http://mes-consulting.org/

What is Data Science?

Definition (Data Science)

Data Science is about computer-based data analysis and generation of knowledge.

Something like

- Business Intelligence?
- Business Analytics?
- Information Retrieval?
- Information Theory?
- Knowledge Discovery?
- Data Mining?
- Statistics?
- Machine Learning?

Terminology

Business Intelligence

Definition (Business Intelligence)

Business Intelligence is about

- using IT
- using in-house data
- turning data into information
- analyze information
- to support management in making decisions.

Keywords: ETL, Data Warehouse, OLAP, Charts

Goal: **Answers to: What happened, when, how many?**

Business Analytics

Definition (Business Analytics)

Business Analytics enhances Business Intelligence by using

- Statistical Analysis
- Data Mining
- Predictive Modeling.

Goal: **Answers to: Why did it happen, what will happen?**

Business Analysis

Definition (Business Analysis)

Business Analysis is about understanding and improving your business processes.

Goal: **Answers to: Why did it happen, what will happen?**

Information Retrieval

Definition (Information Retrieval)

Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections [MRS08].

Goal: **Find and rank relevant information in data**

Information Theory

Definition (Information Theory)

Information Theory studies the quantification, storage, and communication of information [Wik20].

Goal: **Quantify portion of information in data**

Knowledge Discovery

Definition (Knowledge Discovery in Databases)

Knowledge Discovery in Databases (**KDD**) is[a] the process of discovering

- new
- useful and
- valid knowledge

from a collection of data.

[a] cf. [FPS96]

Goal: **Generate new knowledge for humans**

KDD

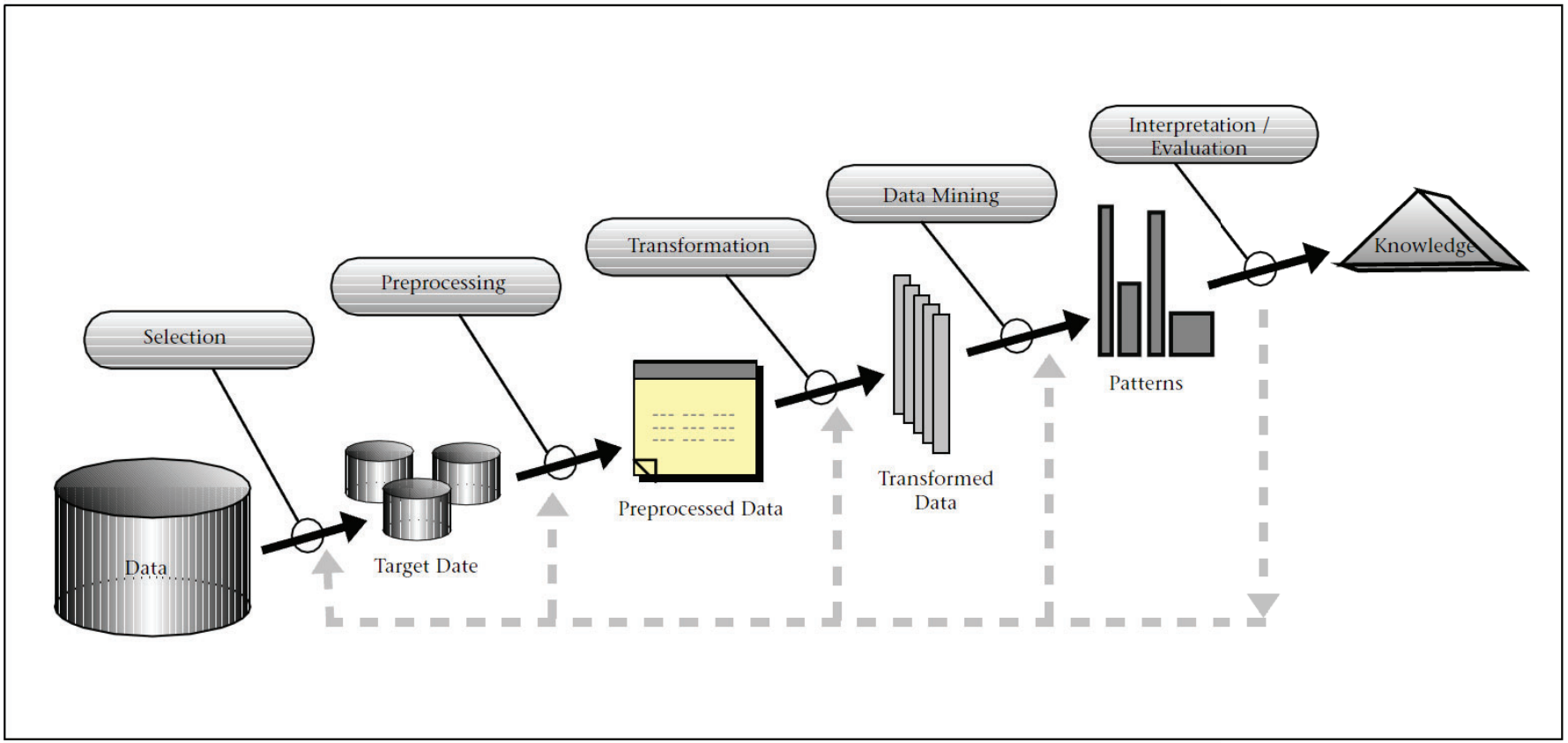


Figure: Knowledge Discovery in Databases

Source: [FPS96]

Knowledge Discovery

The Process [CL14]

1. Preparation
   - 1.1 Collect domain knowledge
   - 1.2 Collect data
   - 1.3 Define goals
2. Data selection
3. Data preprocessing
4. Data transformation / reduction
5. **Data Mining**
6. Interpretation

Data Mining

Definition (Data Mining)

Data Mining is[a] about efficient methods for (mostly automated) detection of non trivial patterns.

[a] cf. [FPS96]

Goal: **Find and explain relations and patterns**

Data Mining

The process also is about

- Explorative Analysis
- Descriptive Statistics
- Visualisation

and contains a lot of statistics.

Data Mining and Statistics

Definition (Statistics)

Statistics[a] is a mathematical and conceptual discipline that focuses on the relation between data and hypotheses.

Descriptive statistics summarizes features of data.
Inferential statistics deduces properties of unseen data.

[a] cf. [Rom18]

This is very, very rough!

Data Mining and Statistics

Inferential Statistics:

1. Formulate hypotheses
2. Plan experiments
3. Collect small, clean data
4. Validate hypotheses

based on theory.
Goal: **Generalize facts** to something you have not seen

Data Mining and Statistics

Data Mining: A lot unclean data already is there

1. Try
2. Validate
3. Try
4. Validate
5. ...

based on data.

CRISP-DM

Cross-industry standard process for data mining (**CRISP**)

Figure: CRISP

Source: Kenneth Jensen, https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

CRISP-DM

- **Business Understanding**
  Know the keywords, concepts and goals
- **Data Understanding**
  Know the schema and meaning, explore data
- **Data Preparation**
  Transform and clean data
- **Modeling**
  Apply algorithm
- **Evaluation**
  Validate if goal is reached
- **Deployment**
  Reporting for customer

Data Mining - Areas

Find and explain

- **Regression Analysis**: Real values
- **Classification**: Categorical belonging
- **Cluster Analysis**: Groups of data
- **Association Analysis**: Rules
- ...

Data Mining - Areas

Find and explain

- **Regression Analysis**: Real values
  - Predict the numeric target label of a data point
- **Classification**: Categorical belonging
  - Predict if a data point belongs to one of the predefined classes
- **Cluster Analysis**: Groups of data
  - Identify natural clusters (groups) within the data set based on inherit properties within the data set
- **Association Analysis**: Rules
  - Identify relationships within an item set based on transaction data
- ...

It contains a lot of **Machine Learning**.

Machine Learning

Definition (Machine Learning)

Machine Learning is a part of artificial intelligence and is[a] about

- progressively improving performance
- on a specific task
- based on data
- without being explicitly programmed.

[a] cf. [Sam59]

Goal: **Generate new abilities for machines**

Traditional approach

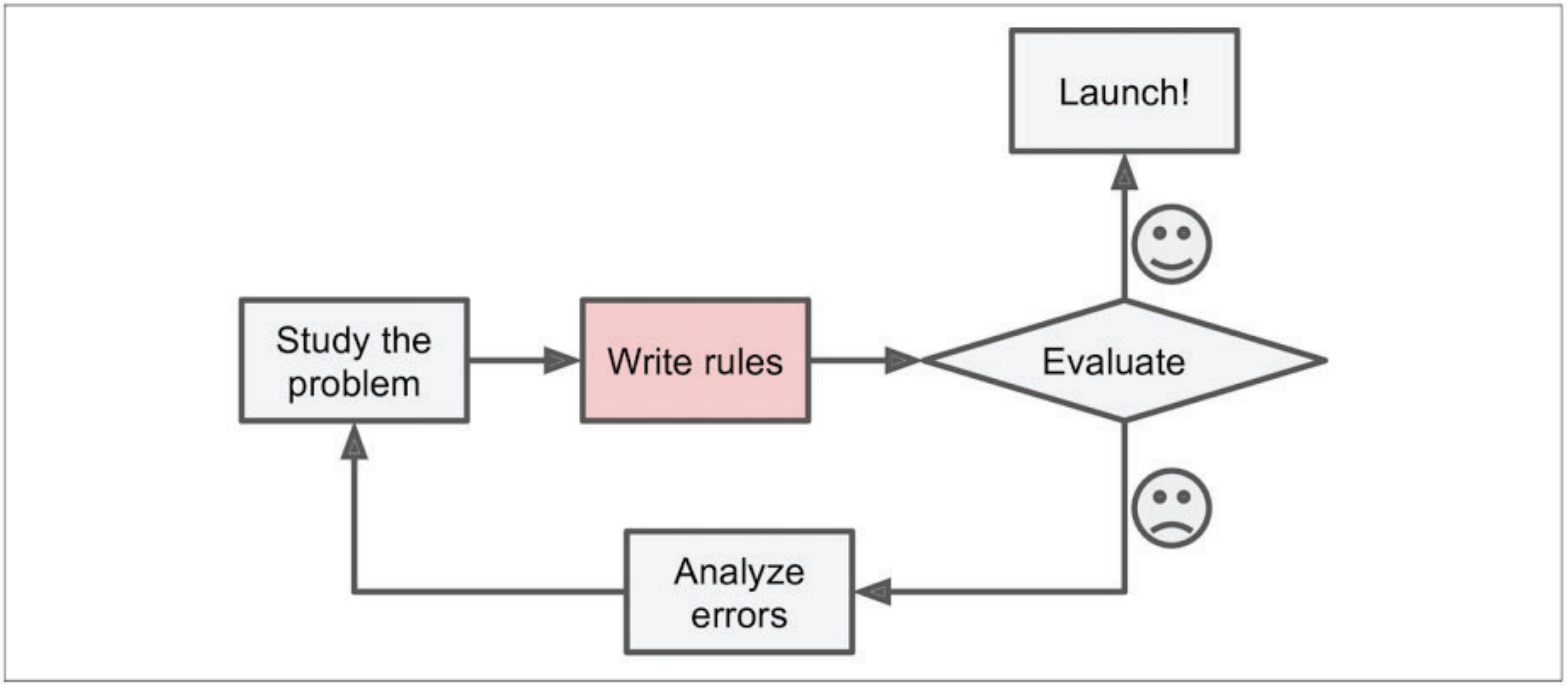


Figure: Write explicit rules

Image source: [Gér17]

Machine Learning approach

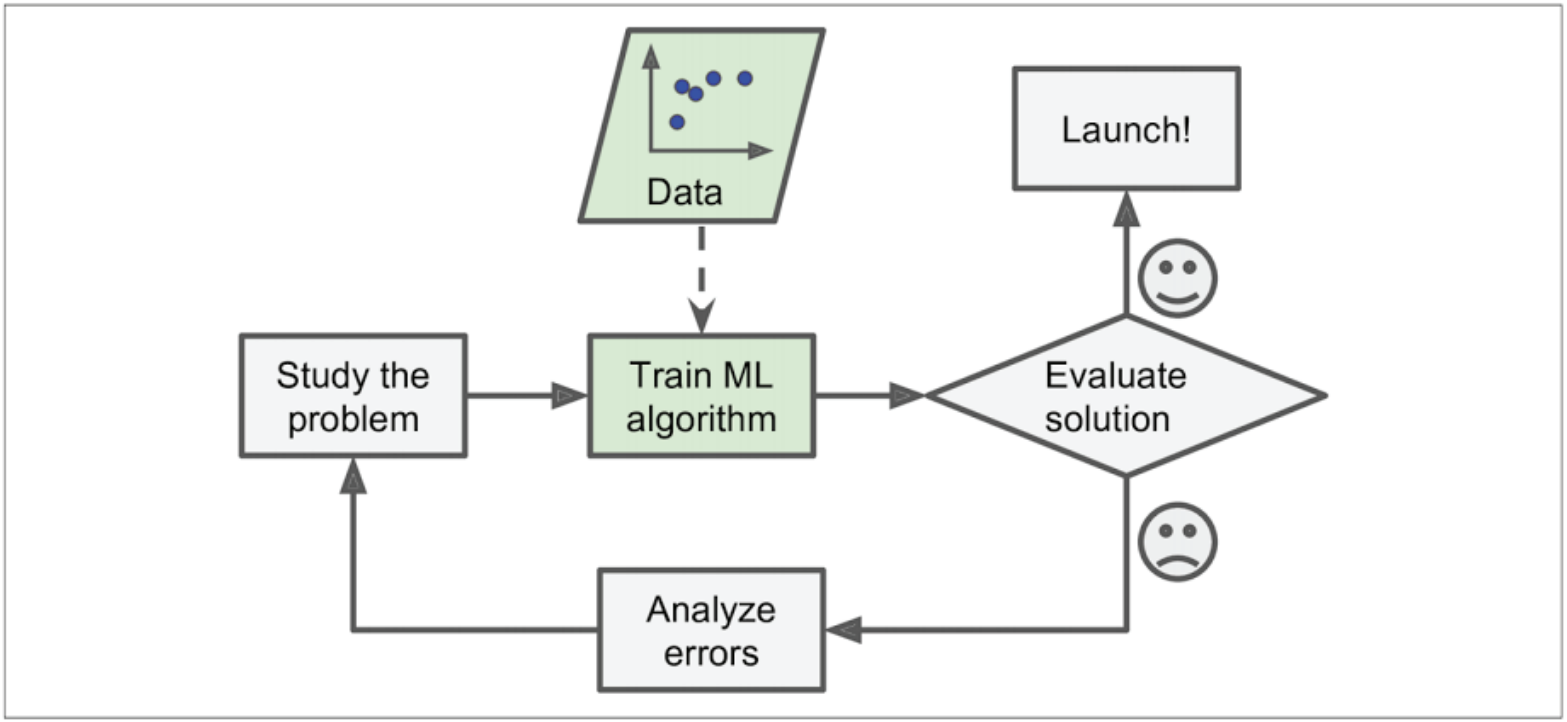


Figure: Train ML algorithm to learn from data

Image source: [Gér17]

Machine Learning automation

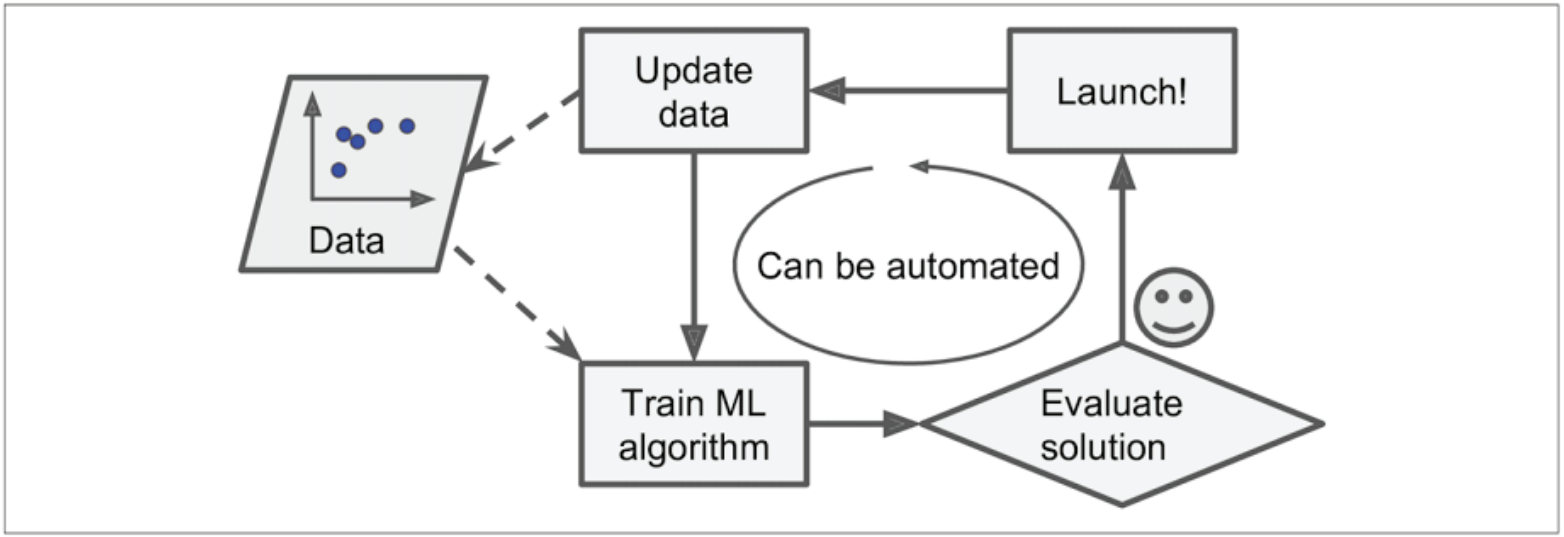


Figure: Learning is iterative: Start, evaluate, improve

Image source: [Gér17]

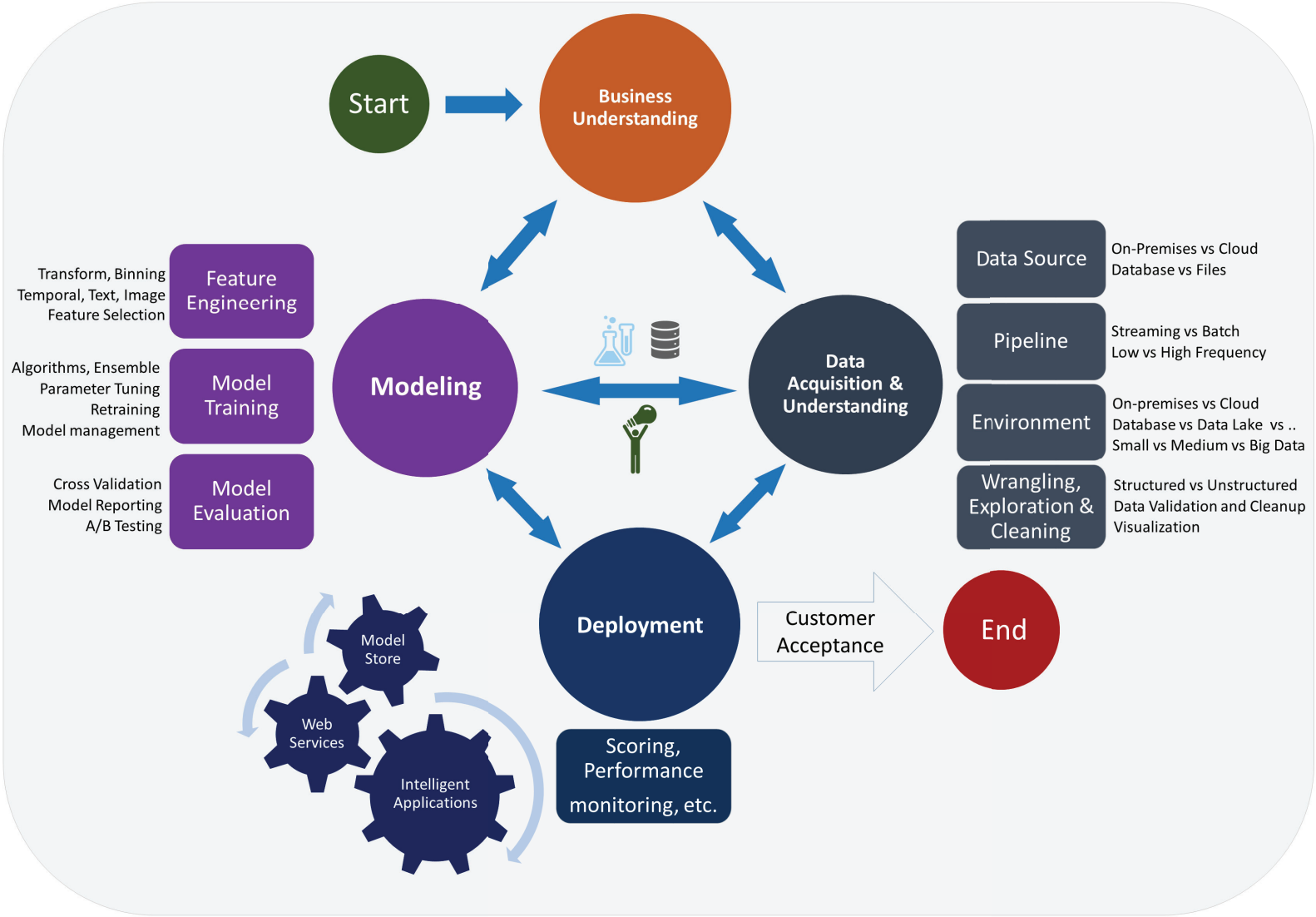
Predictive Analytics

One very important ability is prediction:

Definition (Predictive Analytics)

Predictive Analytics is about analyzing current or past events in order to predict future (unknown) events.

Microsoft - Data Science Lifecycle



Source: https://docs.microsoft.com/de-de/azure/machine-learning/team-data-science-process/overview

Data Science

Definition (Data Science)

Data Science is about computer-based data analysis and generation of knowledge.
It contains

- Domain Knowledge (Business aspects)
- Statistics (Math / theory)
- Data Mining (Process of data treatment)
- Machine Learning (Clever algorithms)
- Computer Science (Modern SW / HW architecture)

Data Science means the whole thing!

Data Science includes modern Computer Science

Computer Science

- Big Data (3V-5V)
  - A lot of data
  - Also unstructured data
    - Natural language
    - Images
    - ...
    - NoSQL
- Distributed (everything)
- Coding
  - SQL
  - R
  - Python
- Deployment to machines
  - Docker
  - Kubernetes
  - Clouds

DA = KDD = DM = ML (= DS)

☺ Everything in this lecture is correct!

☹ Not everything in this lecture is the perfect truth[1]!

[1] You can always find somebody having a slightly different but sound opinion about this terminology. Not really *definitions*, sorry. ☹

Software

Gartner - Magic Quadrant for Data Science and Machine Learning Platforms

Source: https://www.gartner.com/doc/reprints?id=1-25DIVGDE&ct=210303&st=sb

Data Science and Machine Learning Platforms

- We will use **RapidMiner**
- There are other products
- In particular interesting: Visual Workflow Designer

RapidMiner

RapidMiner[2]

- is a Visionary in Gartner's Magic Quadrant [Gar21]
- Origin: Technische Universität Dortmund
- Strong presence in many industries
  - but especially manufacturing, life sciences, banking, insurance, energy, business services, government and education
- Strong presence in the academic world
- Strengths
  - Multipersona collaboration
  - Clear vision and delivery of aligned features
  - Explainable, governed and secured AI

[2] https://rapidminer.com/

RapidMiner

RapidMiner Studio[3]

- Visual Workflow Designer
- Java-based
- More than 1500 operators
- Origin: Technische Universität Dortmund
- Windows / Mac / Linux
- Educational License
  - https://my.rapidminer.com/nexus/account/index.html#licenses/request
- Deployment: Kubernetes Cluster

[3] https://rapidminer.com/products/studio/

RapidMiner: Documentation

RapidMiner Studio

- Documentation
  - Docs:
    - https://docs.rapidminer.com/latest/studio/getting-started/
    - https://docs.rapidminer.com/latest/studio/operators/rapidminer-studio-operator-reference.pdf
    - https://rapidminer.com/wp-content/uploads/2013/10/RapidMiner-5.2-Advanced-Charts-english-v1.0.pdf
  - Videos:
    - https://rapidminer.com/training/videos/
  - Youtube Channel:
    - https://www.youtube.com/channel/UCxneJBWWNLs-A6ckls1Rrug
  - Books: [KD14], [Chi13]
- Model Filter: https://mod.rapidminer.com/

RapidMiner: Feature List

RapidMiner Studio

- Feature List[4]
  - **Data Access**
    Access, load and analyze data
  - **Data Exploration**
    Extract statistics and key information
  - **Data Prep**
    Cleanse data for predictive analytics
  - **Modeling**
    Build and deliver models
  - **Validation**
    Estimate model performance

[4] https://rapidminer.com/products/studio/feature-list/

RapidMiner: Data Access

RapidMiner Studio

- Data Access
  - Structured
    - CSV
  - Semi-Structured
    - HTML
  - Unstructured
    - Text
    - Audio
    - Video
  - Cloud Storage
    - Dropbox
    - AWS S3

RapidMiner: Data Access

RapidMiner Studio

- Data Access
  - Databases
    - JDBC
  - NoSQL
    - MongoDB
    - Cassandra

Literature

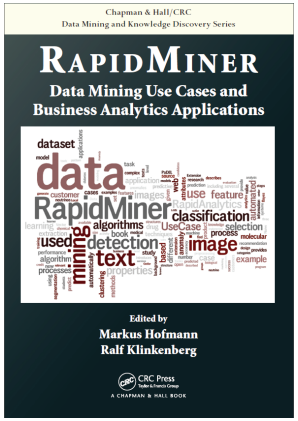


Figure: [HK16]

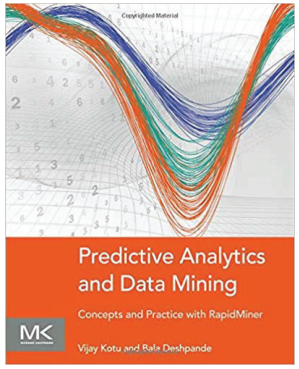


Figure: [KD14]

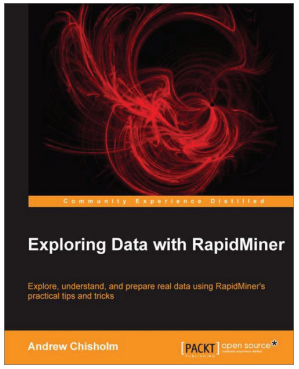


Figure: [Chi13]

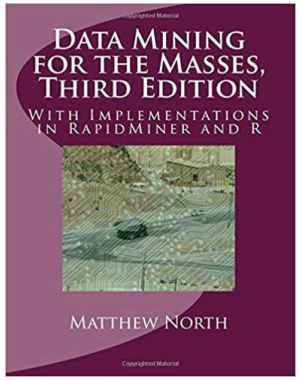


Figure: [KD18]

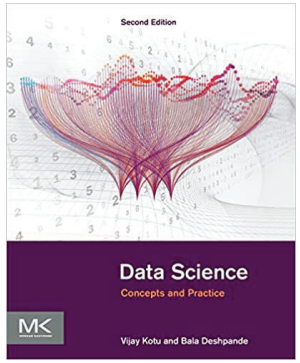


Figure: [Nor18]

KNIME

KNIME Analytics Platform[5]

- is a Visionary in Gartner's Magic Quadrant
- Visual Workflow Designer
- Java-based
- Origin: Universität Konstanz
- Windows / Mac / Linux
- Client base spans all industries and company sizes
- Strengths
  - Breadth and depth of capabilities
  - Open-Source platform
  - Coherence of visual workflow

[5] https://www.knime.com/

Orange

Orange[6]

- Visual Workflow Designer
- C++ / Python-based
- Origin: University of Ljubljana
- Windows / Mac / Linux
- Part of Anaconda

[6] https://orange.biolab.si/

Alteryx Designer

alteryx

Alteryx[7]

- is a Challenger in Gartner's Magic Quadrant
- Alteryx Designer[8]: Visual Workflow Designer
- Irvine, California, United States
- Cloud-based
- Educational License: https://www.alteryx.com/sparked
- Clients in most domains and industries
  - but especially manufacturing, financial services, consumer packaged goods, retail, healthcare and government
- Strengths
  - Ease of use
  - Go-to-market strategy
  - Customer and operational support

[7] https://www.alteryx.com/

[8] https://www.alteryx.com/products/alteryx-platform/alteryx-designer

Dataiku

Dataiku[9]

- is a Leader in Gartner's Magic Quadrant
- End-to-End AI Platform
- Visual Workflow Designer
- New York City, United States
- Cloud-based, Windows / Mac / Linux
- Clients spans many industries and business functions
- Strengths
  - Also for beginning data scientists
  - Focus on business value
  - Increasing market traction
- Deployment: Kubernetes Cluster

[9] https://www.dataiku.com/

Azure Machine Learning

Azure Machine Learning[10]

- is a Visionary in Gartner's Magic Quadrant
- Microsoft Corporation
- Redmond, Washington, United States
- Designer[11]: Visual Workflow Designer
- Cloud-based
- Clients spans many industries and business functions
- Strengths
  - Enterprise data science
  - Multipersona
  - Openness and partnerships

[10] https://azure.microsoft.com/en-us/services/machine-learning/

[11] https://azure.microsoft.com/en-us/services/machine-learning/designer/

Exercise: RapidMiner Basics

Exercise

Please work on exercise 1

Literature

Literature I

- DS: [Pie15], [Gru16], [Gru15], [CMA16], [PF13]
- KDD: [Fay96], [FPS96]
- ML: [Sam59], [BRF16], [Gér17], [Ert16]
- DM: [Run10], [WFH11], [Tor10], [CL14], [Liu07], [Bro14], [BCJ14]
- RM: [HK16], [Chi13], [KD14], [KD18], [Nor18]

[BCJ14] A. Bari, M. Chaouchi, and T. Jung. *Predictive Analytics For Dummies*. –For dummies. Wiley, 2014. ISBN: 9781118729410. URL: https://books.google.de/books?id=IjMKAwAAQBAJ.

[BRF16] H. Brink, J.W. Richards, and M. Fetherolf. *Real-world Machine Learning*. Manning, 2016. ISBN: 9781617291920. URL: https://books.google.de/books?id=DoQAswEACAAJ.

[Bro14] M.S. Brown. *Data Mining For Dummies*. –For dummies. Wiley, 2014. ISBN: 9781118893173. URL: https://books.google.de/books?id=zcDlBQAAQBAJ.

[Chi13] A. Chisholm. *Exploring Data with RapidMiner*. Community experience distilled. Packt Publishing, 2013. ISBN: 9781782169345. URL: https://books.google.de/books?id=FustAgAAQBAJ.

[CL14] J. Cleve and U. Lämmel. *Data Mining*. De Gruyter Studium Series. De Gruyter Oldenbourg, 2014. ISBN: 9783486713916. URL: https://books.google.de/books?id=4i2nngEACAAJ.

Literature II

[CMA16] D. Cielen, A. Meysman, and M. Ali. *Introducing Data Science: Big Data, Machine Learning, and More, Using Python Tools*. Manning Publications, 2016. ISBN: 9781633430037. URL: https://books.google.de/books?id=zYbisgEACAAJ.

[Ert16] W. Ertel. *Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung*. Computational Intelligence. Springer Fachmedien Wiesbaden, 2016. ISBN: 9783658135492. URL: https://books.google.de/books?id=ecD3DAAAQBAJ.

[Fay96] U.M. Fayyad. *Advances in Knowledge Discovery and Data Mining*. AAAI Press Series. AAAI Press, 1996. ISBN: 9780262560979. URL: https://books.google.de/books?id=XqVQAAAAMAAJ.

[FPS96] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "Advances in Knowledge Discovery and Data Mining". In: ed. by Usama M. Fayyad et al. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996. Chap. From Data Mining to Knowledge Discovery: An Overview, pp. 1–34. ISBN: 0-262-56097-6. URL: http://dl.acm.org/citation.cfm?id=257938.257942.

[Gar21] Gartner, Inc. *Magic Quadrant for Data Science and Machine Learning Platforms*. [Online; accessed 5. Mar. 2021]. Mar. 2021. URL: https://www.gartner.com/doc/reprints?id=1-25DIVGDE&ct=210303&st=sb.

[Gér17] A. Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2017. ISBN: 9781491962244. URL: https://books.google.de/books?id=bRpYDgAAQBAJ.

[Gru15] J. Grus. *Data Science from Scratch: First Principles with Python*. O'Reilly Media, 2015. ISBN: 9781491904398. URL: https://books.google.de/books?id=24kdCAAAQBAJ.

[Gru16] J. Grus. *Einführung in Data Science: Grundprinzipien der Datenanalyse mit Python*. O'Reilly, 2016. ISBN: 9783960100256. URL: https://books.google.de/books?id=g-RNDAAAQBAJ.

Literature III

[HK16] M. Hofmann and R. Klinkenberg. *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, 2016. ISBN: 9781482205503.

[KD14] V. Kotu and B. Deshpande. *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Elsevier Science, 2014. ISBN: 9780128016503. URL: https://books.google.de/books?id=dRHoAwAAQBAJ.

[KD18] V. Kotu and B. Deshpande. *Data Science: Concepts and Practice*. Elsevier Science, 2018. ISBN: 9780128147627. URL: https://books.google.de/books?id=-nt8DwAAQBAJ.

[Liu07] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Data-centric systems and applications. Springer, 2007. ISBN: 9783540378815. URL: https://books.google.de/books?id=6Mh50Uaq6AIC.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN: 0521865719, 9780521865715.

[Nor18] M. North. *Data Mining for the Masses, Third Edition: With Implementations in RapidMiner and R*. CreateSpace Independent Publishing Platform, 2018. ISBN: 9781727102475. URL: https://books.google.de/books?id=stwbvAEACAAJ.

[PF13] F. Provost and T. Fawcett. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media, 2013. ISBN: 9781449374280. URL: https://books.google.de/books?id=4ZctAAAAQBAJ.

[Pie15] L. Pierson. *Data Science For Dummies*. –For dummies. Wiley, 2015. ISBN: 9781118841457. URL: https://books.google.de/books?id=Jx%5C_JBgAAQBAJ.

[Rom18] Jan-Willem Romeijn. *Philosophy of Statistics*. [Online; accessed 5. Apr. 2018]. Apr. 2018. URL: https://plato.stanford.edu/entries/statistics.

Literature IV

[Run10] T.A. Runkler. *Data Mining: Methoden und Algorithmen intelligenter Datenanalyse*. Computational Intelligence. Vieweg+Teubner Verlag, 2010. ISBN: 9783834893536. URL: https://books.google.de/books?id=cnYMys3V2t4C.

[Sam59] A. L. Samuel. "Some Studies in Machine Learning Using the Game of Checkers". In: *IBM Journal of Research and Development* 3.3 (July 1959), pp. 210–229. ISSN: 0018-8646. DOI: 10.1147/rd.33.0210.

[Tor10] L. Torgo. *Data Mining with R: Learning with Case Studies*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Taylor & Francis, 2010. ISBN: 9781439810187. URL: https://books.google.de/books?id=EaNQPgAACAAJ.

[WFH11] I.H. Witten, E. Frank, and M.A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011. ISBN: 9780080890364. URL: https://books.google.de/books?id=bDtLM8CODsQC.

[Wik20] Wikipedia. *Information theory - Wikipedia*. [Online; accessed 24. Feb. 2020]. Feb. 2020. URL: https://en.wikipedia.org/wiki/Information_theory.