Data Mining and Machine Learning: Techniques and Algorithms

Eneldo Loza Mencía

eneldo@ke.tu-darmstadt.de

Knowledge Engineering Group, TU Darmstadt

International Week 2019, 21.1. – 24.1.
University of Economics, Prague

Myself

TECHNISCHE UNIVERSITÄT DARMSTADT

Graduated at Knowledge Engineering Group

- Leader: Prof. Johannes Fürnkranz
- currently around 8 colleagues
- Goals
  - acquisition of explicit, formalizable knowledge (e.g. rules, ontologies)
  - from sources that contain relevant information in implicit or not directly accessible form
- Methods
  - techniques from machine learning and data mining
  - knowledge acquisition by analysis of existing data or text collections, by interaction with human experts, or by experimentation and simulation

Myself

TECHNISCHE UNIVERSITÄT DARMSTADT

My interests

- multi-label classification
- human-interpretable machine learning models
- forecasting of epidemiological outbreaks
- automatic text summarization
- computer poker AI
- and many more

Goals of course

- You will learn about methods and techniques in Machine Learning
  - the main characteristics
  - their internals and functioning
  - advantages, disadvantages
  - in which (data) situations to employ
  - under which circumstances to employ
- Selection of algorithms
  - some in details (basic ones)
  - for some only an overview
  - many many others are not touched

What this course is not about

- Programming
- “Data Science”
- Deep Learning
- any many other aspects which are perhaps touched but there is no time :(

Organization

- 8 blocks á 90 minutes
  - Lectures, some exercises
- Grading
  - Exam of 60 minutes in the 9th block
    - Exercises about performing algorithms (no calculator needed)
    - Questions on the content
  - Grades until next week (?)
- Material
  - I will upload it during the course to
    https://www.ke.tu-darmstadt.de/staff/eneldo/IW19

Tentative Schedule

	Tue	Wed	Thu
8:30-10:00?	Introduction	Block 4	Block 7
10:30-12:00?	Block 2	Block 5	Block 8
13:30-15:xx?	Block 3 +Exercise	Block 6 + Exercise	Exam

Content (may change)

- Introduction
- Instance based learning
- Decision tree learning
- Evaluation
- Ensemble learning
- Semi-supervised and unsupervised methods
- Excursions
  - Neural networks
  - Text Mining and information retrieval
  - Recommender Systems
  - Reinforcement learning

Data Mining - Motivation

"Computers have promised us a fountain of wisdom but delivered a flood of data."

"It has been estimated that the amount of information in the world doubles every 20 months."

(Frawley, Piatetsky-Shapiro, Matheus, 1992)

„160,000,000 terabytes of data have been generated in 2006"

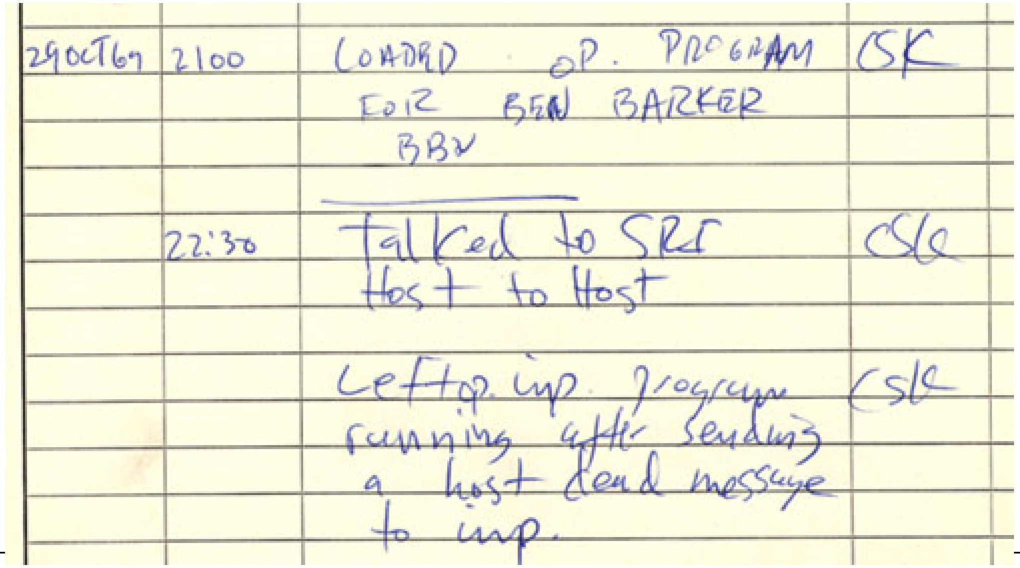
(Data Consortium)

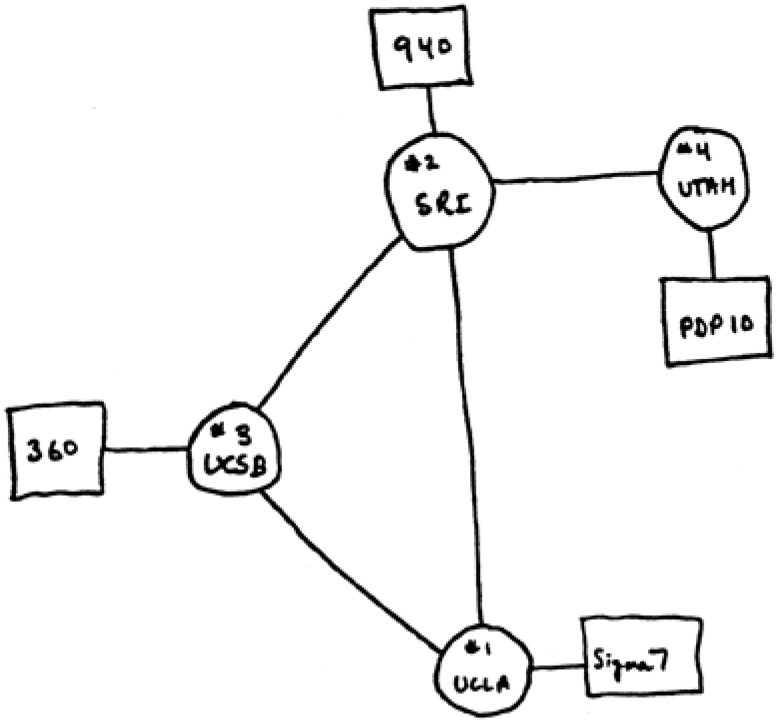
World-Wide Data Growth

- Science
  - satellite monitoring
  - human genome
  - CERN
- Business
  - OLTP (on-line transaction processing)
  - data warehouses
  - e-commerce
  - logistics
- Industry
  - process data
  - industry 4.0
- World-Wide Web

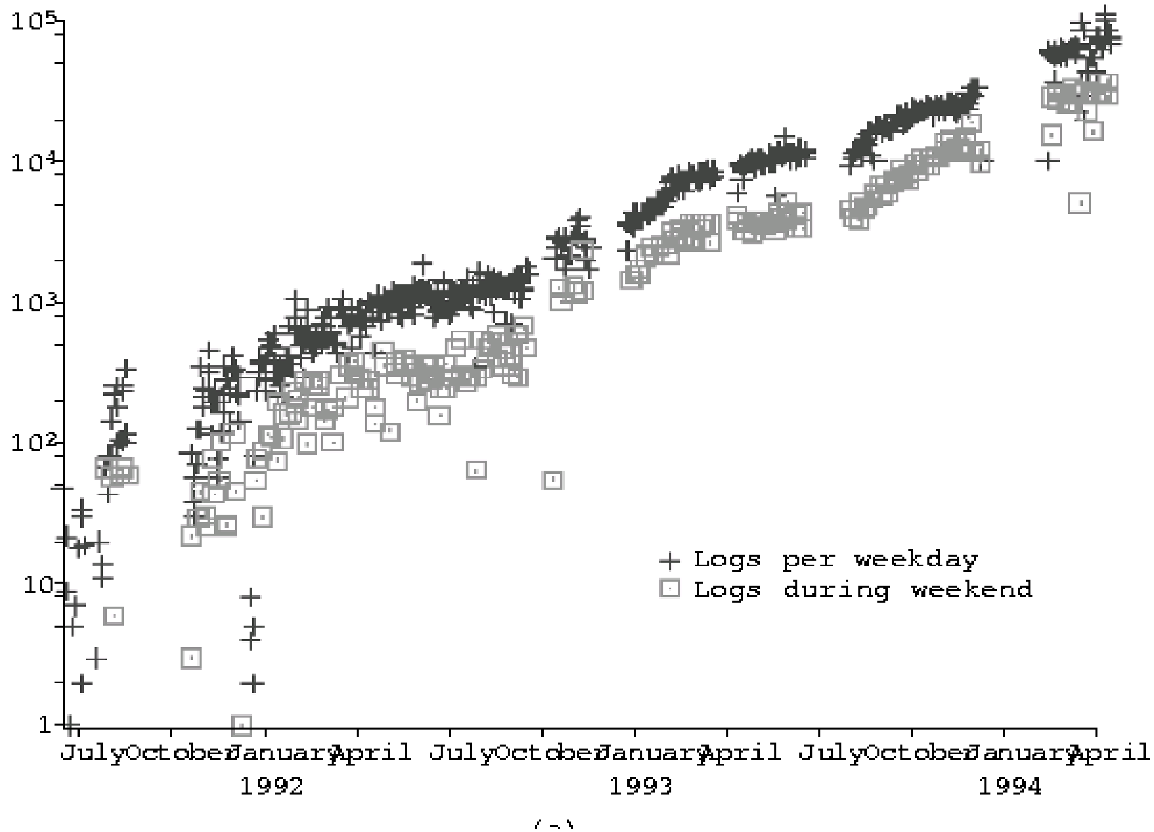
Size of the World Wide Web
The Birth of the Web

- ARPANET
  - started with 4 nodes at four universities
    - UCLA, UCSB, SRI, Utah
  - first message sent on October 29, 1969





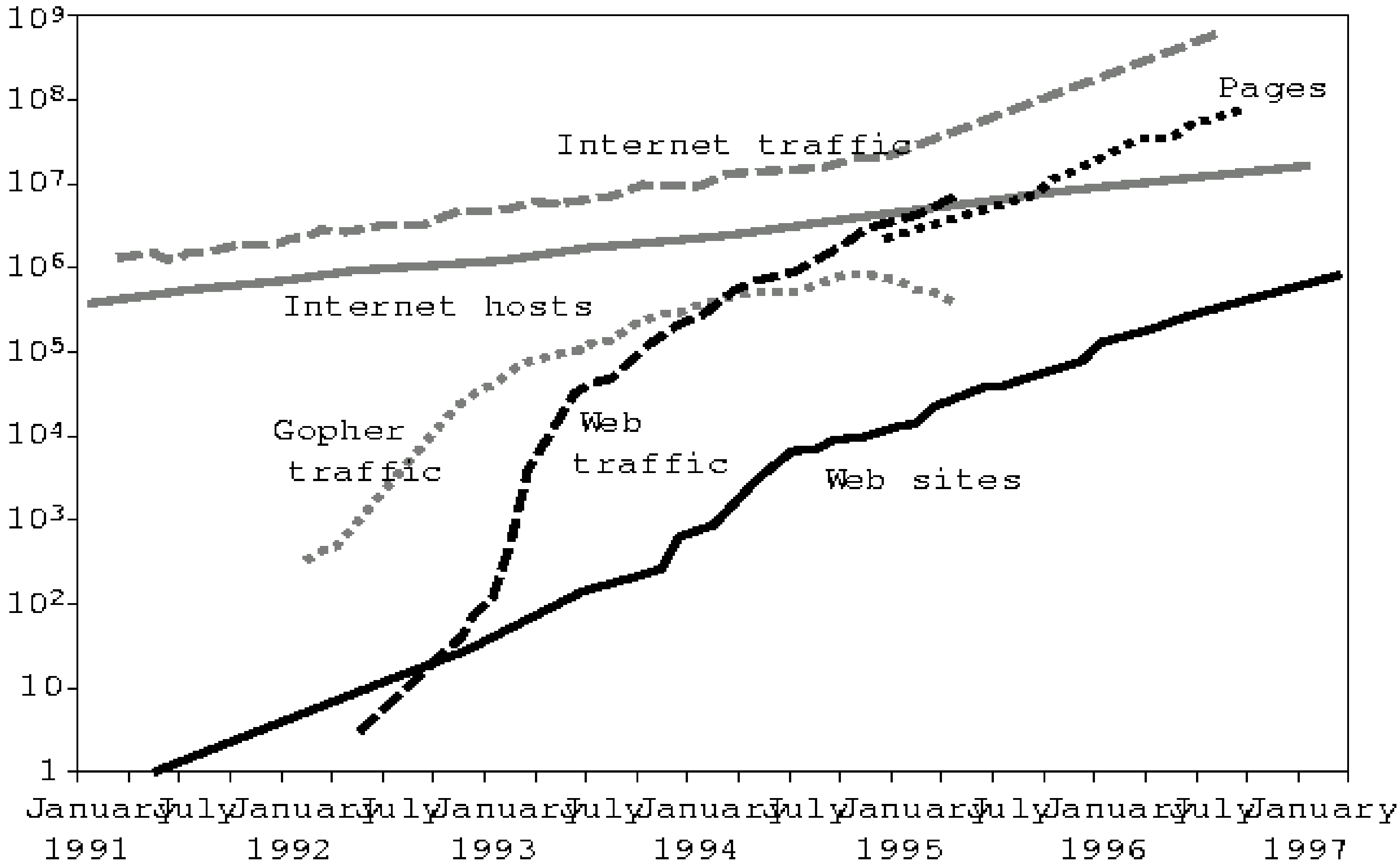
Size of the World Wide Web
The early days



CERN HTTP traffic grows by 1000 between 1991-1994 (image courtesy W3C)

Chakrabarti and Ramakrishnan

Size of the World Wide Web
The early days



The number of servers grows from a few hundred to a million between 1991 and 1997 (image courtesy Nielsen)

Chakrabarti and Ramakrishnan

Size of the World Wide Web
Recent development

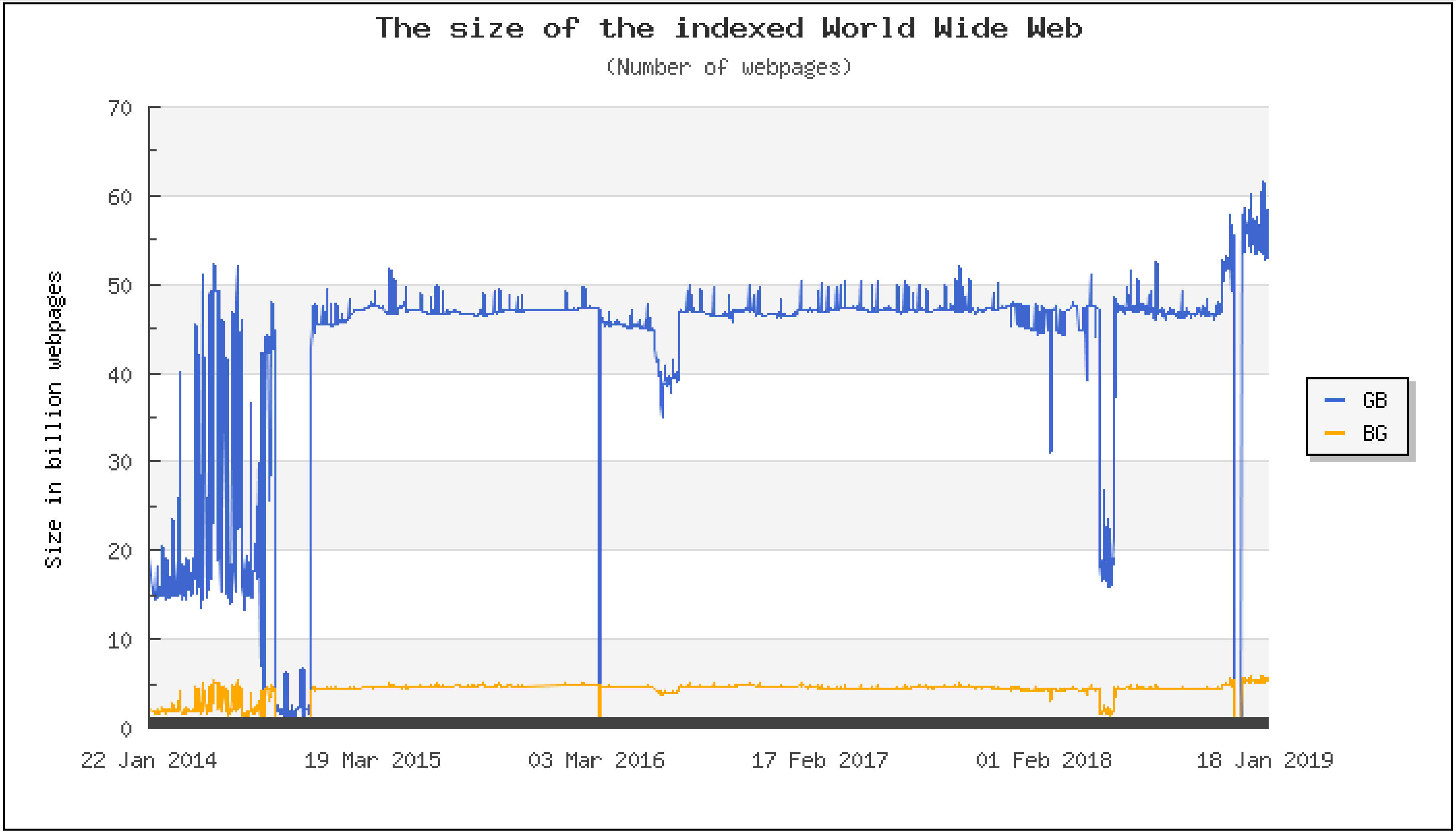
TECHNISCHE UNIVERSITÄT DARMSTADT

- Google:
  - early 2001: 1,346,966,000 web pages
  - 11.2.2002: 2,073,418,204
  - 2004: 4,285,199,774
  - 28.4.2005: 8,058,044,651

- Gulli & Signorini (2005)
  - estimate the size of the Web to 11.5 billion pages,
  - Coverage of search engines
    - Google=76.16%, Msn Beta=61.90%, Ask/Teoma=57.62%, Yahoo!=69.32%
- Hidden Web
  - Results from 1998 estimate that the best search engines index about 30% of the Web

Source: www.thestandard.com

Size of the World Wide Web Today



https://worldwidewebsize.com/

Definition

Data Mining is a non-trivial *process* of identifying

- valid
- novel
- potentially useful
- ultimately understandable

patterns in data.

(Fayyad et al. 1996)

It employs techniques from

- machine learning
- statistics
- databases

Or maybe:

- Data Mining is torturing your database until it confesses.

(Heikki Manilla (?) after Ronald Coase*)*

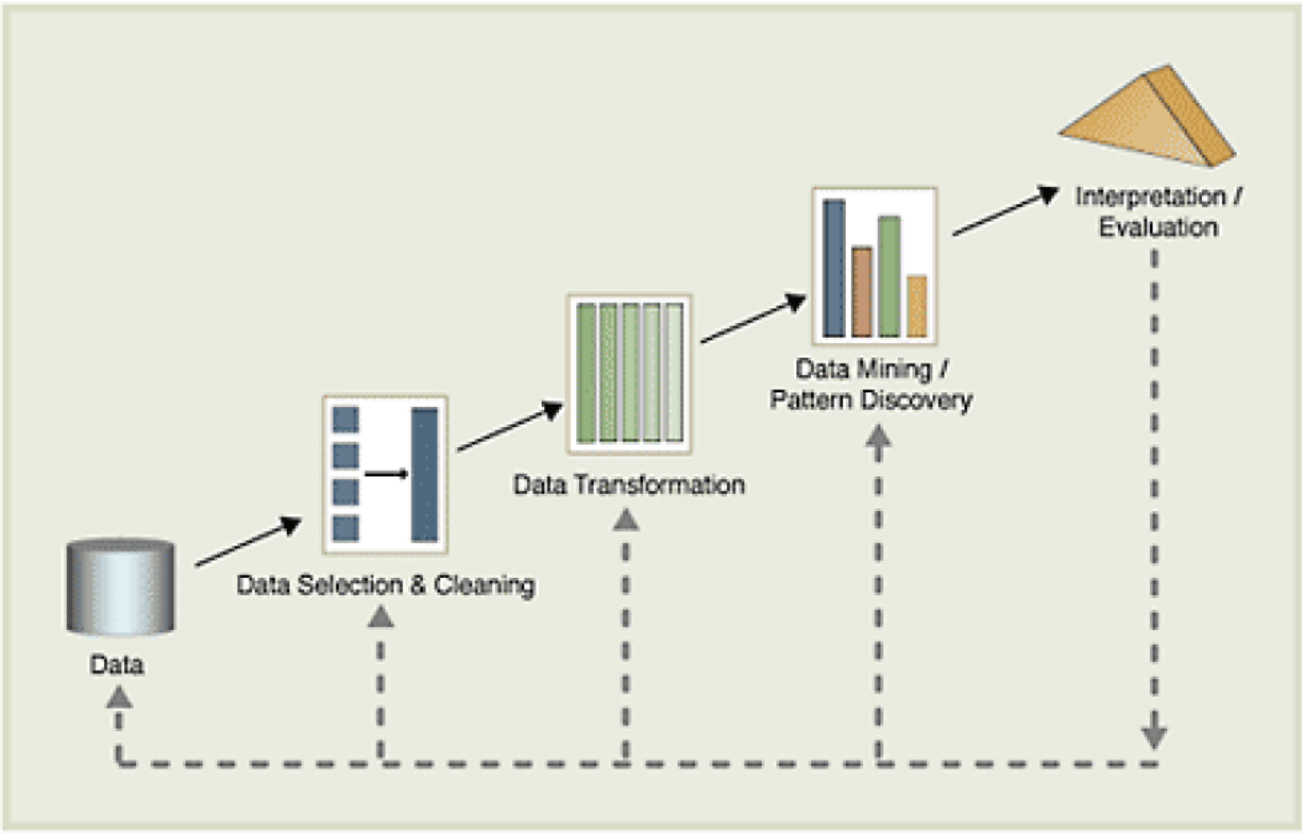
Knowledge Discovery in Databases: Key Steps

Key steps in the Knowledge Discovery cycle:

1. **Data Cleaning**: remove noise and incosistent data
2. **Data Integration**: combine multiple data sources
3. **Data Selection**: select the part of the data that are relevant for the problem
4. **Data Transformation**: transform the data into a suitable format (e.g., a single table, by summary or aggregation operations)
5. **Data Mining**: apply machine learning and machine discovery techniques
6. **Pattern Evaluation**: evaluate whether the found patterns meet the requirements (e.g., interestingness)
7. **Knowledge Presentation:** present the mined knowledge to the user (e.g., visualization)

Data Mining is a Process !

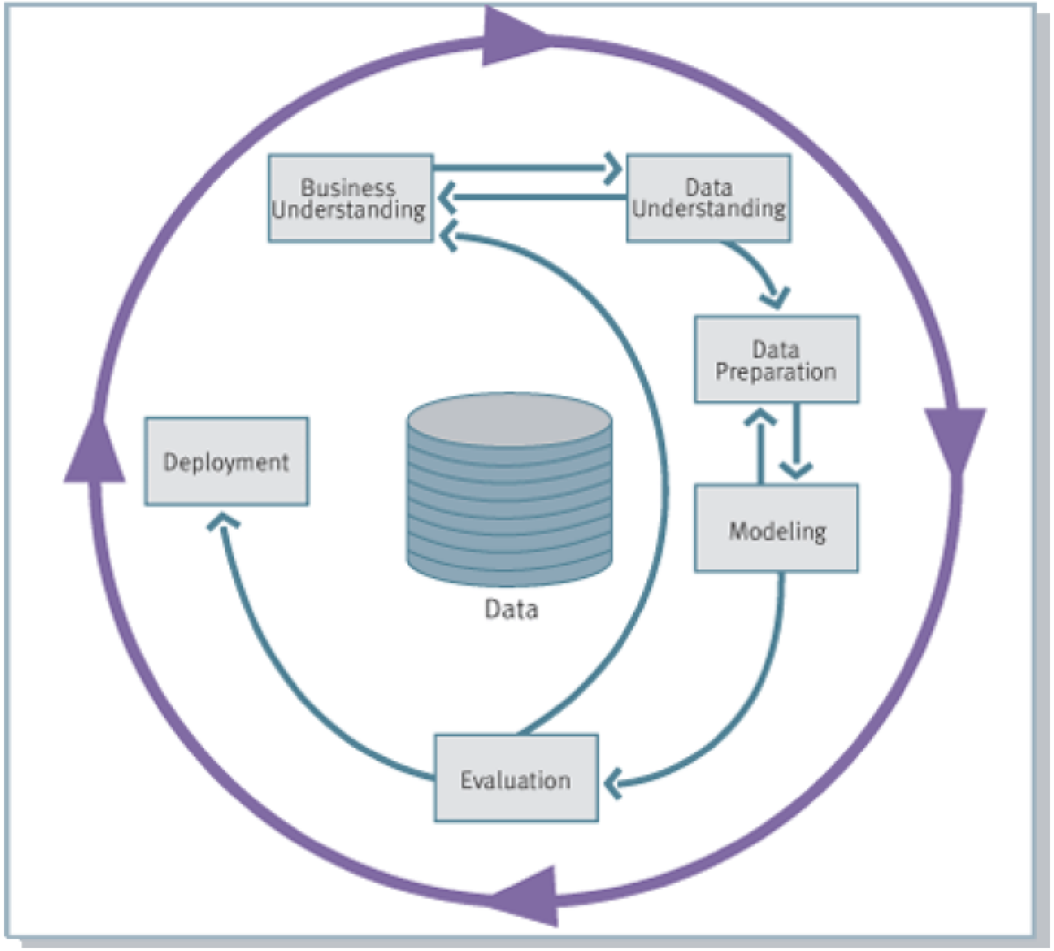
The steps are not followed linearly, but in an iterative process



http://alg.ncsa.uiuc.edu/tools/docs/d2k/manual/dataMining.html, after Fayyad, Piatetsky-Shapiro, Smyth, 1996

Data Mining is a Process !

The steps are not followed linearly, but in an iterative process



http://www.crisp-dm.org/

Research Issues

- Techniques for mining different types of knowledge
  - Predictions, Associations, Clusters, Outliers, ...
- Interactive Data Mining Techniques
  - A Human/Computer Team may be more efficient
- Incorporation of Background Knowledge
  - Knowledge about the task helps.
- Data Mining Query Languages
  - Querying patterns instead of querying database entries
- Presentation and Visualization of Results
  - How to explain the results to the CEO?
- Handling Noisy or Incomplete Data
  - Data are typically not neat and tidy, but noisy and messy.
- Pattern Evaluation
  - How can we define interestingness?

(A few) Data Mining Applications

- Business
  - predict credit rating
  - identify customer groups
  - direct marketing
  - market basket analysis
  - recommender systems
  - fraud detection
- Web Mining
  - categorize Web pages
  - classify E-mail (spam filters)
  - identify Web usage patterns (e.g. for identifying attacs, advertisements)
- Quality control
  - learn to assess quality of products
- Biological/Chemical
  - discover toxicological properties of chemicals
- Game Playing
  - identify common (winning) patterns in game databases

Machine Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

„Learning denotes changes in the system that ... enable the system to do the same task or tasks drawn from the same population more efficiently and more effectively the next time.''

[Simon,1983]

„Learning is making useful changes in our minds.''

[Minsky,1985]

„Learning is constructing or modifying representations of what is being experienced.''

[Michalski,1986]

Machine Learning Problem Definition

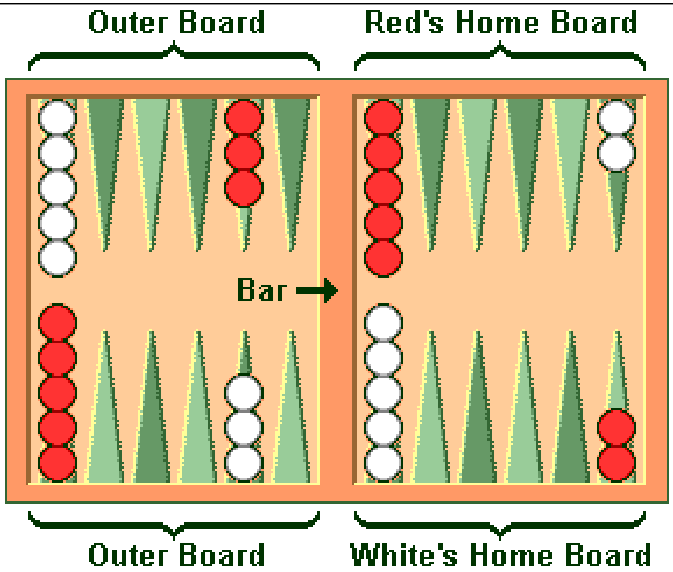
- Definition (Mitchell 1997)

  „A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.“

- Given:
  - a task T
  - a performance measure P
  - some experience E with the task
- Goal:
  - generalize the experience in a way that allows to improve your performance on the task

Learning to Play Backgammon

- Task:
  - play backgammon
- Performance Measure:
  - percentage of games won
- Experience:
  - previous games played



TD-Gammon:

- learned a neural network for evaluating backgammon boards
- from playing millions of games against itself
- successively improved to world-champion strength
- http://www.research.ibm.com/massive/tdl.html
  GNU Backgammon: http://www.gnu.org/software/gnubg/
- Current state of the art: systems self-learned to play Go and Chess beat humans

Recognizing Spam-Mail

- Task:
  - sort E-mails into categories (e.g., Regular / Spam)
- Performance Measure:
  - Weighted Sum of Mistakes (letting spam through is not so bad as misclassifying regular E-mail as spam)
- Experience:
  - Handsorted E-mail messages in your folder

In Practice:

- Many Spam-Filters (e.g., Mozilla) use Bayesian Learning for recognizing spam mails

Market Basket Analysis

- Task:
  - discover items that are frequently bought together
- Performance Measure:
  - ? (revenue by making use of the discovered patterns)
- Experience:
  - Supermarket check-out data

Myth:

- The most frequently cited result is:

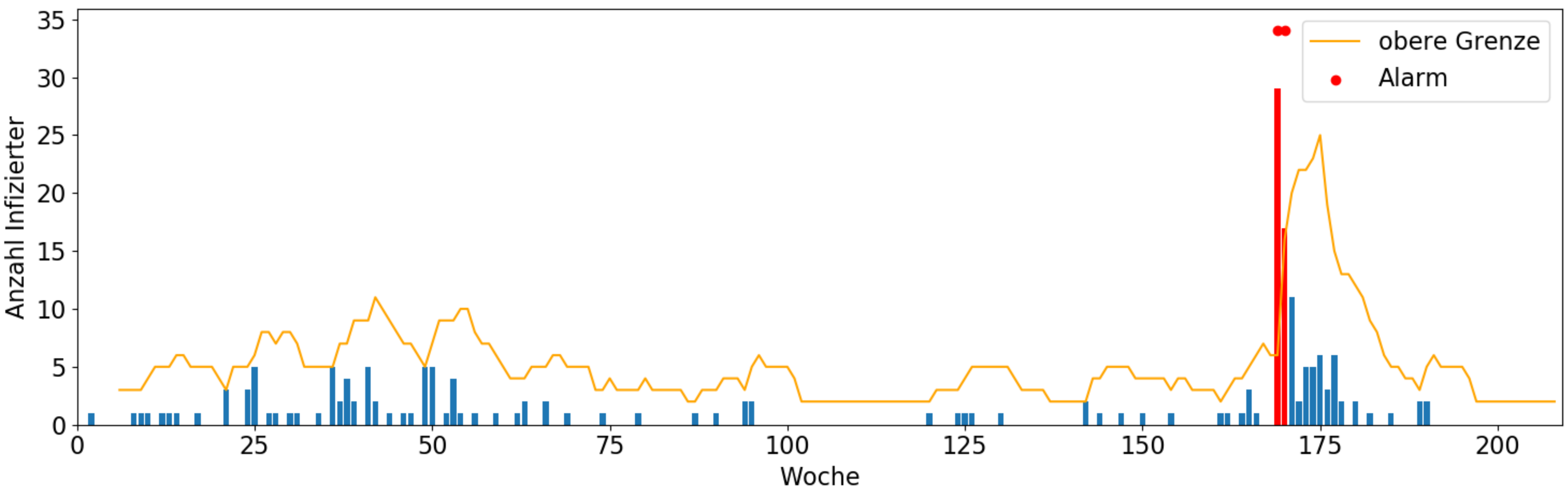
  diapers → beer

Projects at the Knowledge Engineering Group
PRORETA 4 – Security through learning

- Extension of driver assistance through adaptation to person and environment
- e.g.: adapting warnings and timings to person when turning left

Projects at the Knowledge Engineering Group
ESEG: prediction of epidemic outbreaks

- improved epidemic surveillance through methods from Machine Learning
  - anomaly detection
  - time series analysis
  - interpretable models



Dimensions of Learning Problems (1)

- Example Representation
  - attribute-value data vs. first-order logic
- Prediction Task
  - regression, binary, multi-class, multi-label, structured, ...
- Type of training information
  - supervised vs. unsupervised learning
- Availability of training examples
  - batch learning vs. on-line learning (incremental learning)
- Concept representation
  - IF-THEN rules, decision trees, neural networks…

Dimensions of Learning Problems (2)

- Purpose of modeling
  - characteristic vs. discriminative models, interpretable models
- Learning algorithm
  - divide-and-conquer, back-propagation,...
- Evaluation scenario
  - estimating predictive performance, cost-sensitive-learning,
- ...

A Sample Task

Attributes

TECHNISCHE UNIVERSITÄT DARMSTADT

Day	Temperature	Outlook	Humidity	Windy	Play Golf?
07-05	hot	sunny	high	false	no
07-06	hot	sunny	high	true	no
07-07	hot	overcast	high	false	yes
07-09	cool	rain	normal	false	yes
07-10	cool	overcast	normal	true	yes
07-12	mild	sunny	high	false	no
07-14	cool	sunny	normal	false	yes
07-15	mild	rain	normal	false	yes
07-20	mild	sunny	normal	true	yes
07-21	mild	overcast	high	true	yes
07-22	hot	overcast	normal	false	yes
07-23	mild	rain	high	true	no
07-26	cool	rain	normal	true	no
07-30	mild	rain	high	false	yes

today	cool	sunny	normal	false	?
tomorrow	mild	sunny	normal	false	?

Binary target, batch learning

Nominal Attribute

Example Representation

- Attribute-Value data:
  - Each example is described with values for a fixed number of attributes/features/variables
    - **Nominal/Categorical/Discrete Attributes:**
      - store an unordered list of symbols (e.g., *color*)
    - **Numeric Attributes:**
      - store a number (e.g., *income*)
    - **Other Types:**
      - ordered values
      - hierarchical attributes
      - set-valued attributes
  - the data corresponds to a single relation (spreadsheed)
- Multi-Relational data:
  - The relevant information is distributed over multiple relations
  - Inductive Logic Programming

Type of Training Information

- **Supervised Learning:**
  - A teacher provides the value for the target function for all training examples (labeled examples)
  - concept learning, classification, regression
- **Semi-supervised Learning:**
  - Only a subset of the training examples are labeled (labeling examples is expensive!)
- **Reinforcement Learning:**
  - A teacher provides feedback about the values of the target function chosen by the learner
- **Unsupervised Learning:**
  - There is no information except the training examples
  - clustering, subgroup discovery, association rule discovery

Example Availability

- Batch Learning
  - The learner is provided with a set of training examples
- Incremental Learning / On-line Learning
  - There is constant stream of training examples
- Active Learning
  - The learner may choose an example and ask the teacher for the relevant training information

Prediction Targets

TECHNISCHE UNIVERSITÄT DARMSTADT

Binary Classification

- binary targets
- e.g.: event happening, presence of a property,..

i	x_1	x_2	x_3	...	x_a	y
1	A	1	0	...	0.1	**0**
2	B	2	1	...	0.3	**1**
3	C	3	0	...	0.5	**1**
4	D	4	1	...	0.6	**0**
...						

Multiclass Classification:

- nominal targets, finite set of possible classes (>2)
- e.g.: categorization

i	x_1	x_2	x_3	...	x_a	y
1	A	1	0	...	0.1	**A**
2	B	2	1	...	0.3	**B**
3	C	3	0	...	0.5	**A**
4	D	4	1	...	0.6	**C**
...						

Prediction Targets

TECHNISCHE UNIVERSITÄT DARMSTADT

Regression

- numeric values as targets
- e.g.: rating, some measurable property

i	x_1	x_2	x_3	...	x_a	y
1	A	1	0	...	0.1	**0.23**
2	B	2	1	...	0.3	**1.876**
3	C	3	0	...	0.5	**9.3**
4	D	4	1	...	0.6	**-1.4**
...						

Prediction Targets

Multi-label Classification

- multiple class labels possible, subset of labels
- e.g.: keyword tagging, object recognition in scenes

i	x_1	x_2	x_3	...	x_a	**y**
1	A	1	0	...	0.1	$\{\lambda_1,\lambda_n\}$
2	B	2	1	...	0.3	$\{\lambda_2\}$
3	C	3	0	...	0.5	$\{\}$
4	D	4	1	...	0.6	$\{\lambda_1\}$
...						

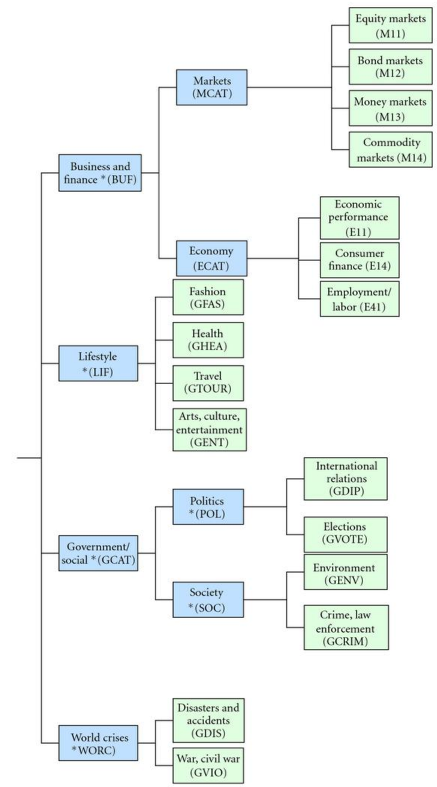
i	x_1	x_2	x_3	...	x_a	y_1	y_2	...	y_n
1	A	1	0	...	0.1	**1**	**0**	**...**	**1**
2	B	2	1	...	0.3	**0**	**1**	**...**	**0**
3	C	3	0	...	0.5	**0**	**0**	**...**	**0**
4	D	4	1	...	0.6	**1**	**0**	**...**	**0**
...									

Prediction Targets

TECHNISCHE UNIVERSITÄT DARMSTADT

Hierarchical Multilabel Classification

- labels organized in hierarchies or graphs



Label Ranking

- learn from and predict rankings on labels

$$\{\lambda_1\} \succ_{\mathcal{T}} \{\lambda_2\} \succ_{\mathcal{T}} \{\lambda_3\} \succ_{\mathcal{T}} \{\lambda_4\}$$

(a) total label ranking

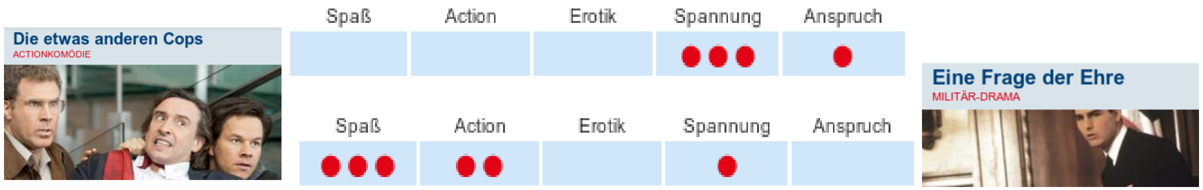
$$\{\lambda_1\} \succ_{\mathcal{T}} \{\lambda_2, \lambda_3, \lambda_4\}$$

(b) bipartite

Prediction Targets

Ordered Classification

- labels can have (ordered) degrees



Collaborative Filtering

- only some output variables are missing, usually no input data

	Book 1	Book 2	Book 3	Book 4	Book 5	Book 6
Customer A	X			X		
Customer B		X	X		X	
Customer C	?	X	X	?	?	?
Customer D		X				X
Customer E	X				X	

Multivariate regression

- likewise several outputs, but real valued instead of binary

X_1	X_2	X_3	X_4
0.34	0	10	174
1.45	0	32	277
1.22	1	46	421
0.74	1	25	165
0.95	1	72	273
1.04	0	33	158
0.92	1	81	382

Y_1	Y_2	Y_3	Y_4
14	0.3	10	10
15	1.4	30	50
23	0.7	20	17
19	1.2	40	60
12	0.6	60	48
17	0.9	61	29
16	1.1	71	54

Multi-target prediction

- general concept of learning multiple targets in parallel

Multi-task learning

- general concept of learning multiple tasks in parallel

Concept Representation

- Most Learners generalize the training examples into an explicit representation
  (called a model, function, hypothesis, concept...)
  - mathematical functions (e.g., polynomial of 3^{rd} degree)
  - logical formulas (e.g., propositional IF-THEN rules)
  - decision trees
  - neural networks
  - ....

- Lazy Learning
  - do not compute an explicit model
  - generalize „on demand“ for a given training example
  - example: nearest neighbor classification

Purpose of modeling

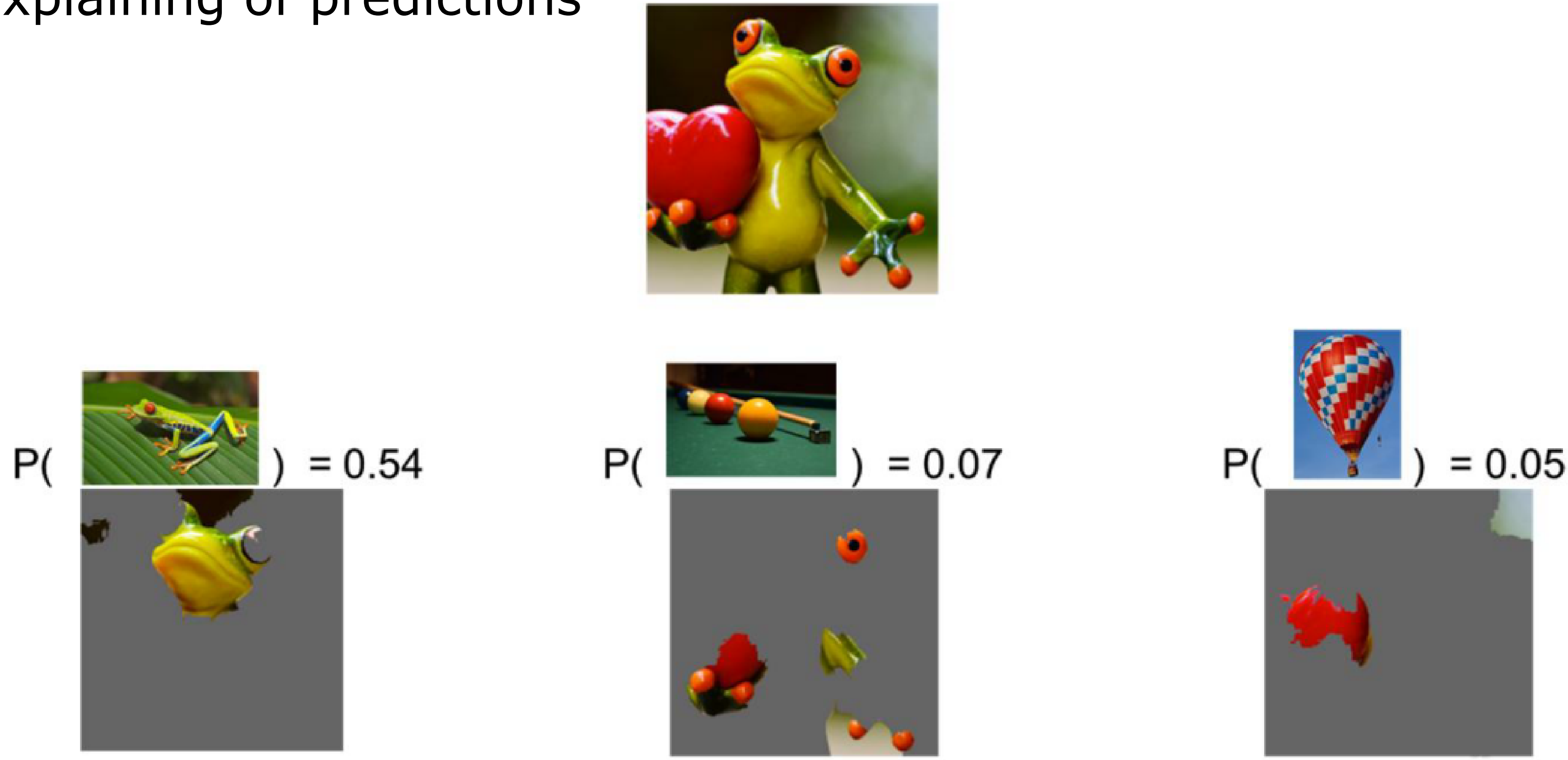
Discriminative

vs.

Characteristic Models

Purpose of modeling

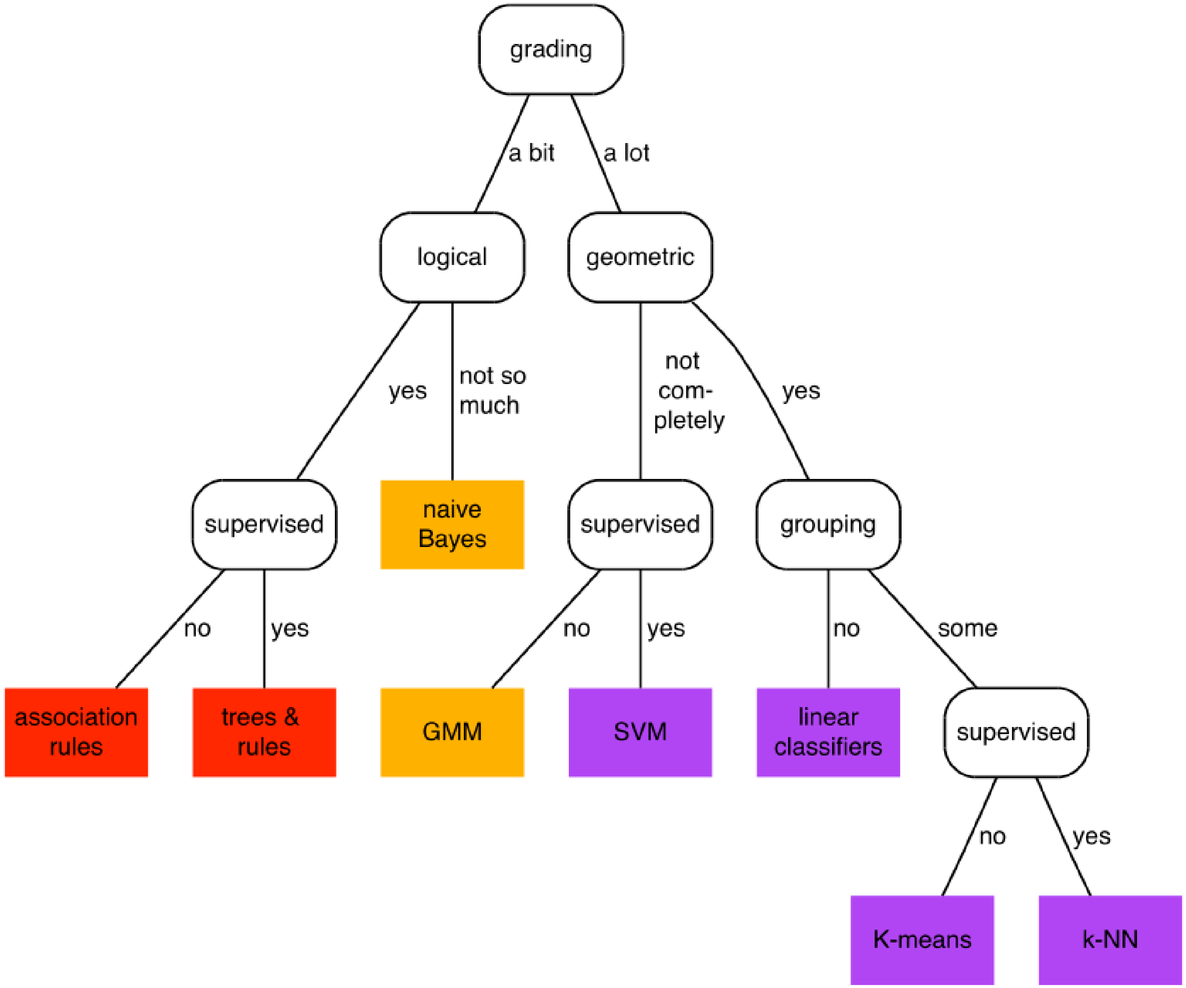
- Interpretable models
- Explaining of predictions



A Selection of Learning Techniques

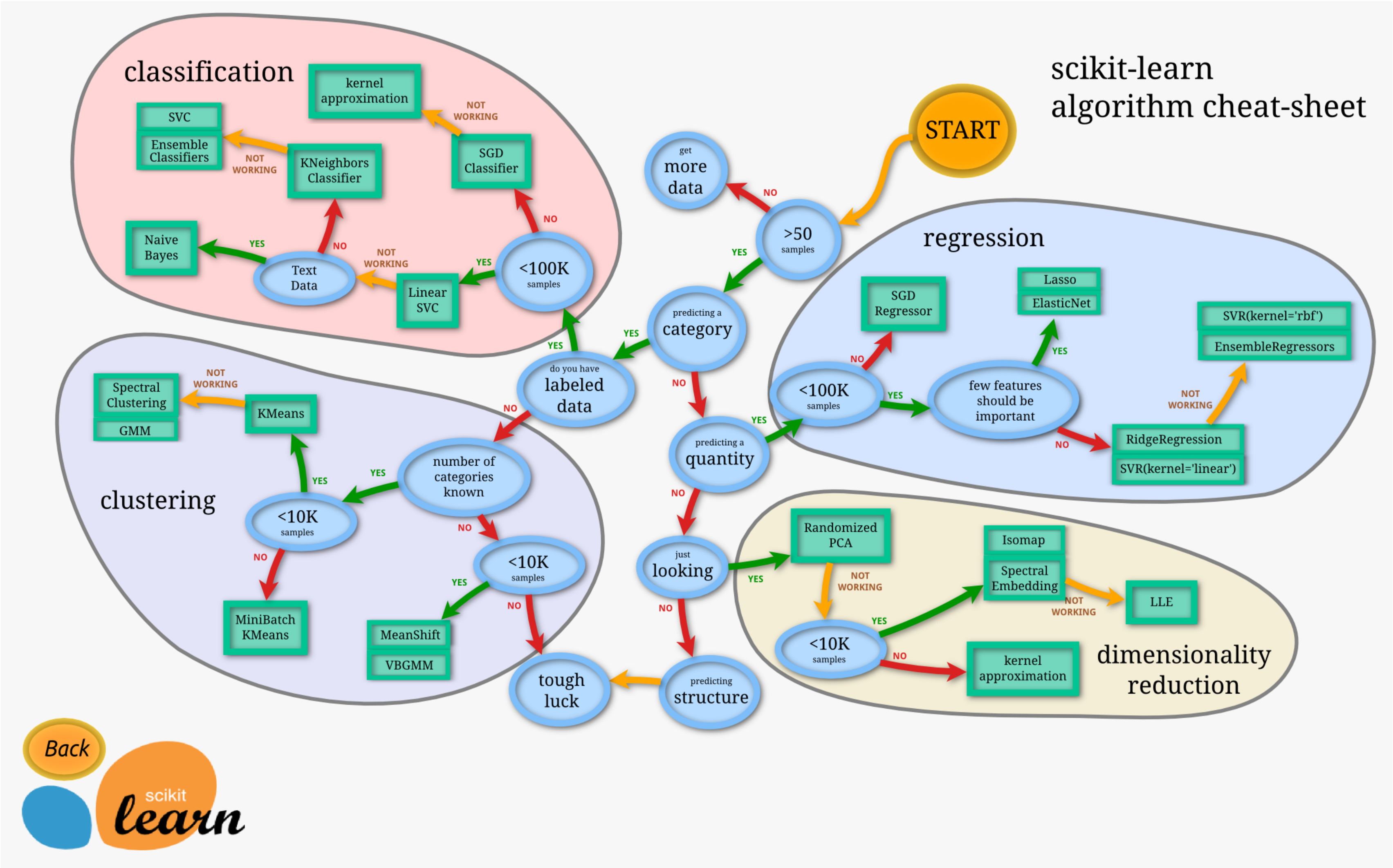
- Decision and Regression Trees
- Classification Rules
- Association Rules
- Inductive Logic Programming
- Neural Networks
- Support Vector Machines
- Statistical Modeling
- Clustering Techniques
- Case-Based Reasoning
- Genetic Algorithms
- ...

Taxonomies of Machine Learning



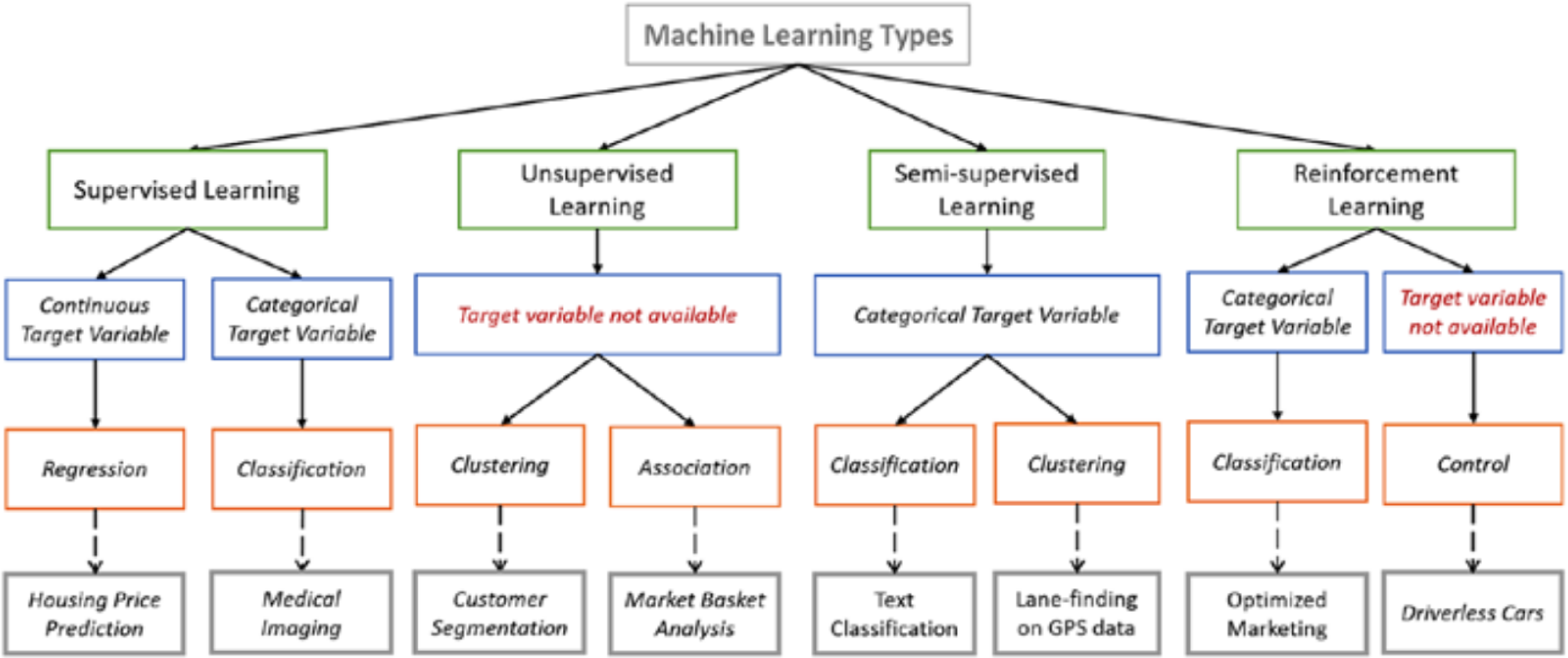
Peter Flach: Machine Learning – The Art and Science of Algorithms that Make Sense of Data

Taxonomies of Machine Learning

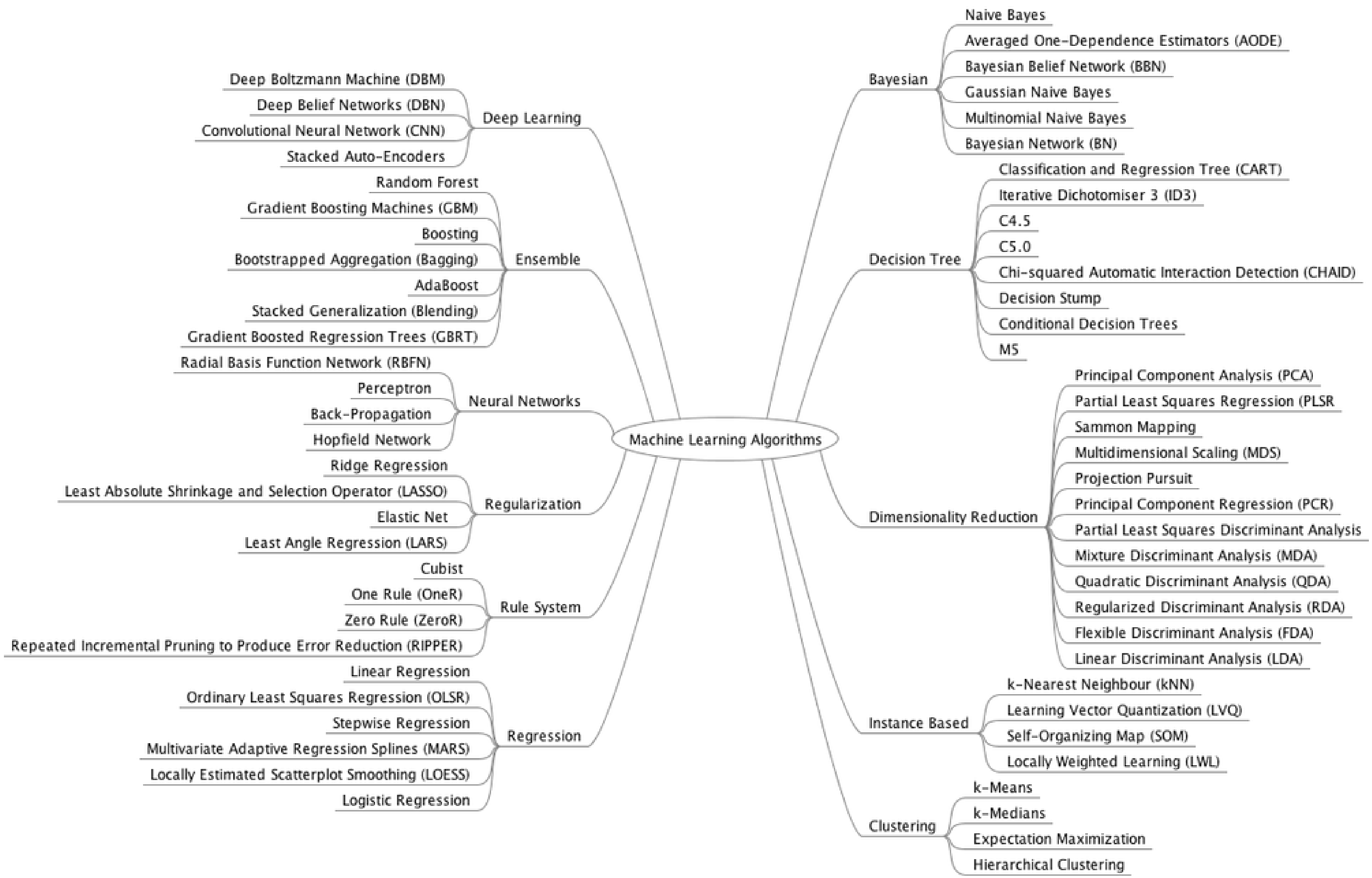


http://scikit-learn.org/stable/_static/ml_map.png

Taxonomies of Machine Learning



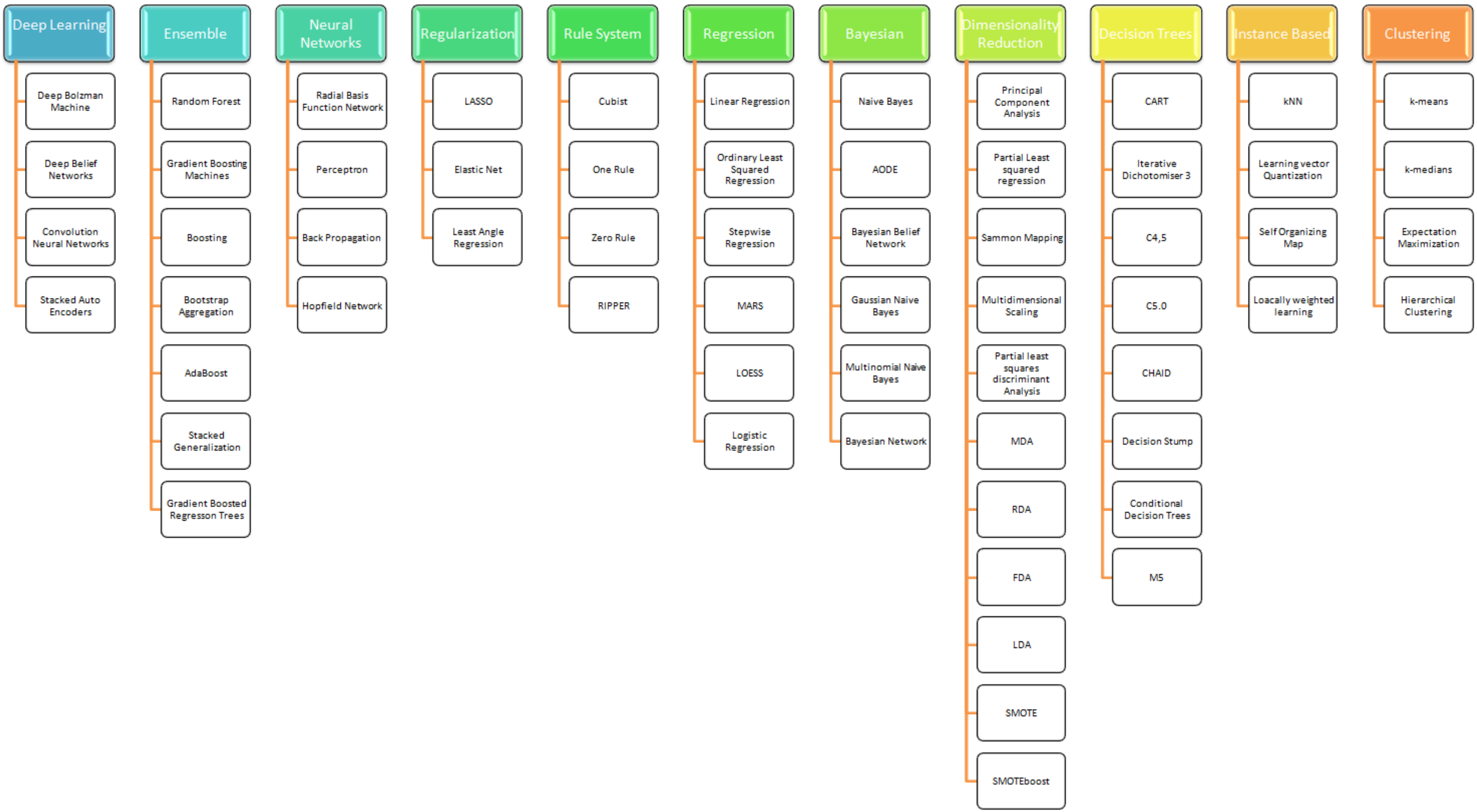
Taxonomies of Machine Learning



https://www.deepmarketer.com/blog/2017/1/30/machine-learning-algorithm-taxonomy

Taxonomies of Machine Learning

TECHNISCHE UNIVERSITÄT DARMSTADT



http://trymachinelearning.com/wp-content/uploads/2016/07/Machine-Learning-Classification.png

Induction of Classifiers

The most „popular“ learning problem:

- Task:
  - learn a model that predicts the outcome of a dependent variable for a given instance
- Experience:
  - experience is given in the form of a data base of examples
  - an example describes a single previous observation
    - *instance:* a set of measurements that characterize a situation
    - *label:* the outcome that was observed in this situation
- Performance Measure:
  - compare the predicted outcome to the observed outcome
  - estimate the probability of predicting the right outcome in new situation

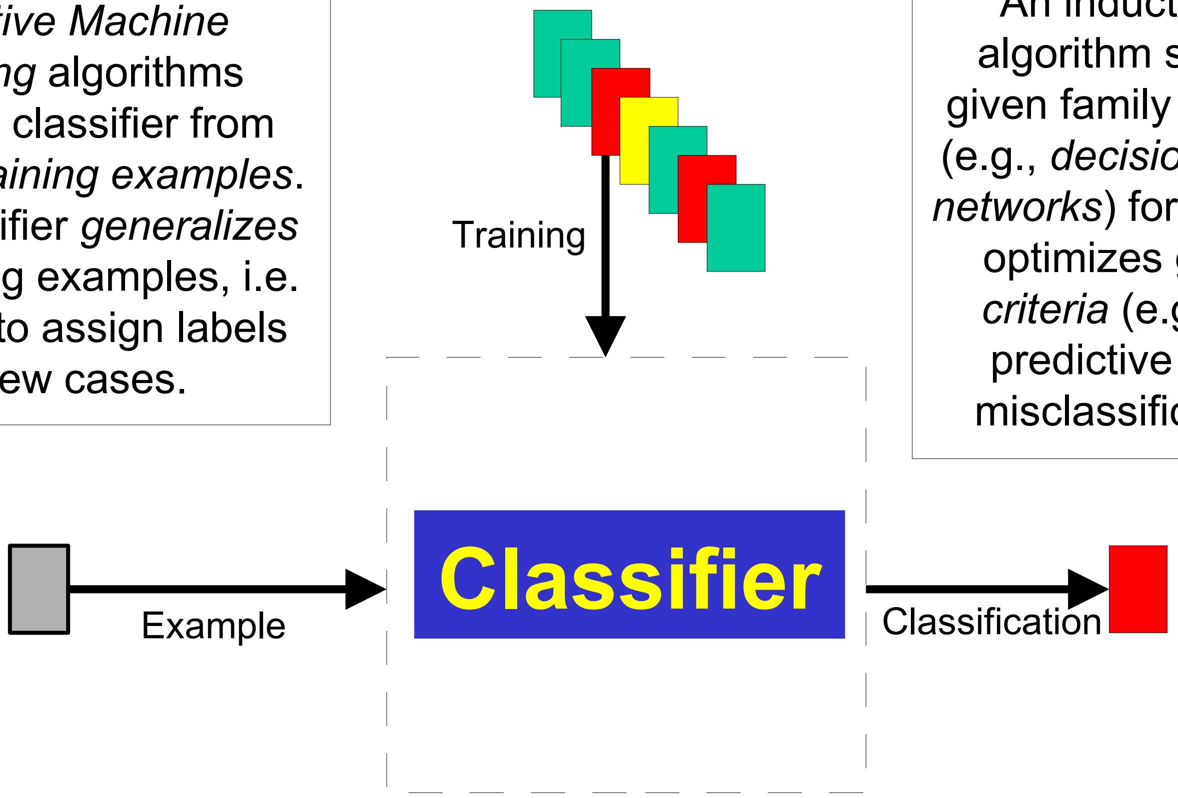
Induction of Classifiers

Typical Characteristics

- attribute-value representation (single relation)
- batch learning from off-line data (data are available from external sources)
- supervised learning (examples are pre-classified)
- numerous learning algorithms for practically all concept representations (decision trees, rules, neural networks, SVMs, statistical models,...)
- often greedy algorithms (may not find optimal solution, but fast processing of large datasets)
- evaluation by estimating predictive accuracy (on a portion of the available data)

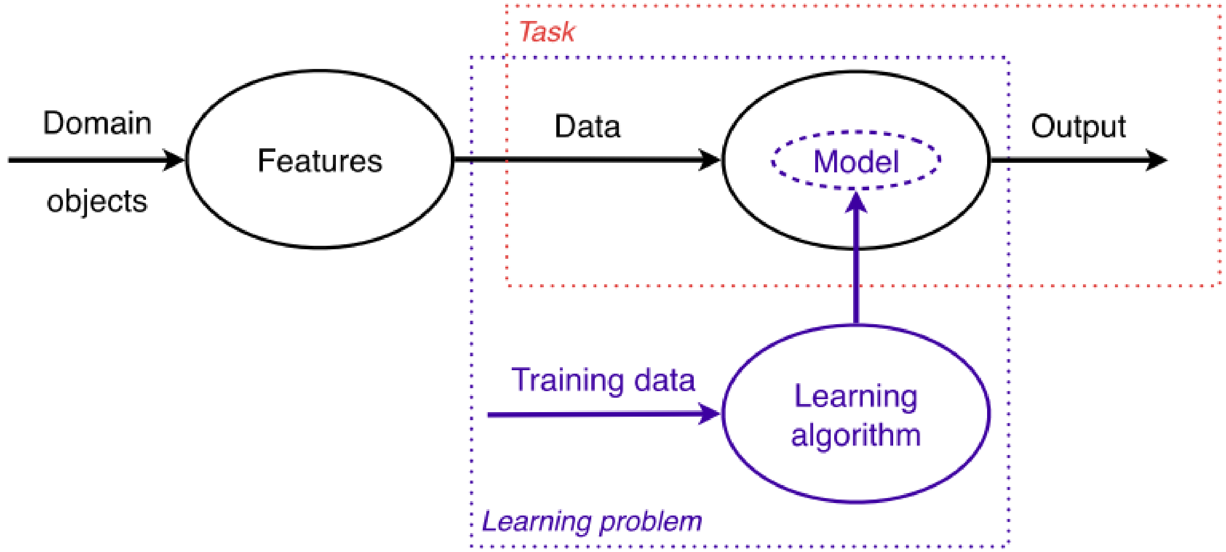
Induction of Classifiers

Inductive Machine Learning algorithms induce a classifier from *labeled training examples*. The classifier *generalizes* the training examples, i.e. it is able to assign labels to new cases.



An inductive learning algorithm searches in a given family of hypotheses (e.g., *decision trees, neural networks*) for a member that optimizes given *quality criteria* (e.g., estimated predictive accuracy or misclassification costs).

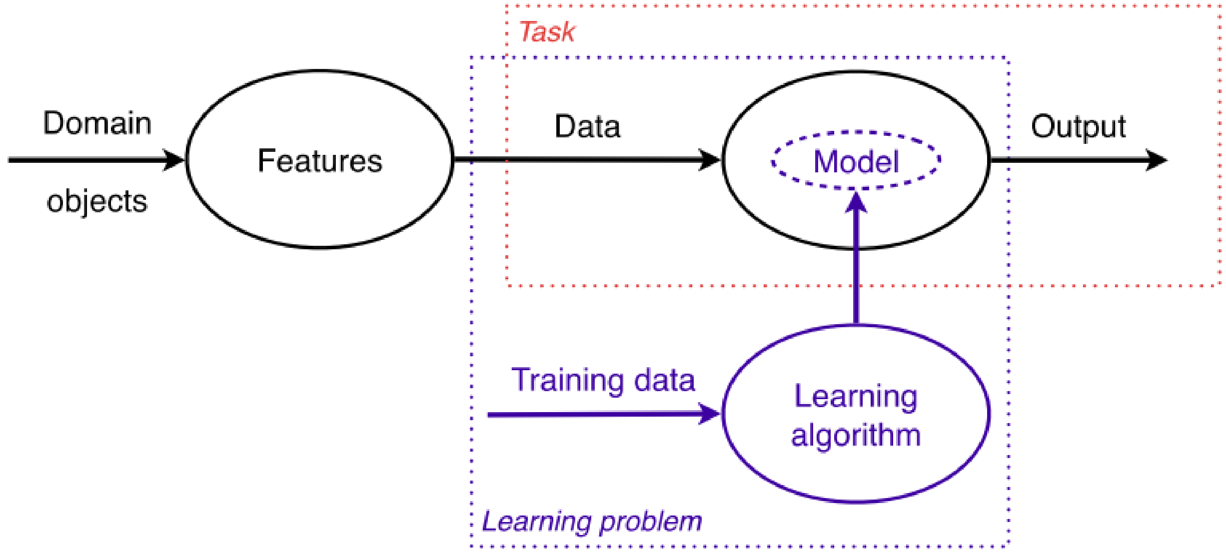
Yet a different view



- A task requires an appropriate mapping – a model – from data described by features to outputs.
- Obtaining such a mapping from training data is what constitutes a learning problem

Peter Flach: Machine Learning – The Art and Science of Algorithms that Make Sense of Data

Yet a different view



- "Tasks are addressed by models, whereas learning problems are solved by learning algorithms that produce models"
- "Machine learning is concerned with using the right features to build the right models that achieve the right tasks."

Peter Flach: Machine Learning – The Art and Science of Algorithms that Make Sense of Data

Yet a different view (2)

Given a (potentially unknown) mapping function

$$f(\mathrm{x}) = \mathrm{y},\ (\mathrm{x},\mathrm{y}) \in \mathrm{X} \times \mathrm{Y}$$

learn a function

$$\overline{f}(\mathrm{x}) \approx f(\mathrm{x})$$

on known $\mathrm{x} \in \mathrm{X}^t$ (training set), $\mathrm{X}^t \subset \mathrm{X}$, so that

$$\overline{f}(\mathrm{x}) \approx f(\mathrm{y}) \text{ for all } \mathrm{x} \in \mathrm{X} \setminus \mathrm{X}^t$$

Theorems and Concepts in Machine Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

- Bias and Generalization
  - Occam’s Razor
  - Overfitting
  - Bias and Variance
- No Free Lunch Theorem
- Curse of Dimensionality
- …

Bias and Generalization

Bias: (Machine Learning Definition)
Any criterion that prefers one concept over another except for completeness/consistency on the training data.

- **Language Bias:**
  Choose a hypothesis representation language
- **Selection Bias:**
  Which hypotheses will be preferred during the search?
- **Overfitting Avoidance Bias:**
  Avoid too close approximations to training data

- Bias is necessary for <u>generalization</u>
  - without bias all complete and consistent hypotheses (those that correctly explain all training examples) are equally likely
  - for any example, half of them will predict one class, the other half the opposite class (*no free lunch theorems*)

“No Free Lunch” Theorem

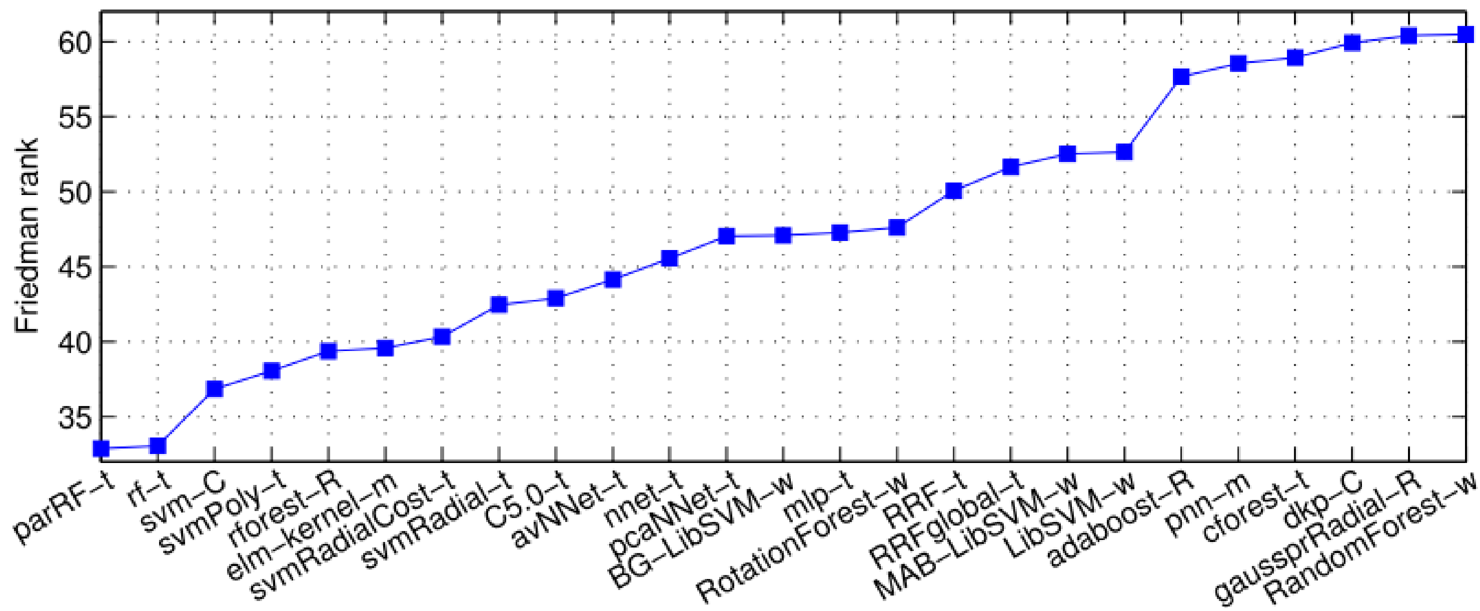
- In a nutshell: no one algorithm works best for every problem

  → try many different algorithms for your problem
- but also: do not waste time and make some preparatory analysis
  - data characteristics (≈inputs)
  - task characteristics (≈targets)
  - appropriate models
  - appropriate learning algorithms
- still, some general trends can be seen...

https://elitedatascience.com/machine-learning-algorithms

"No Free Lunch" Theorem

Comparison of 179 algorithms from 17 algorithm families on 121 UCI datasets

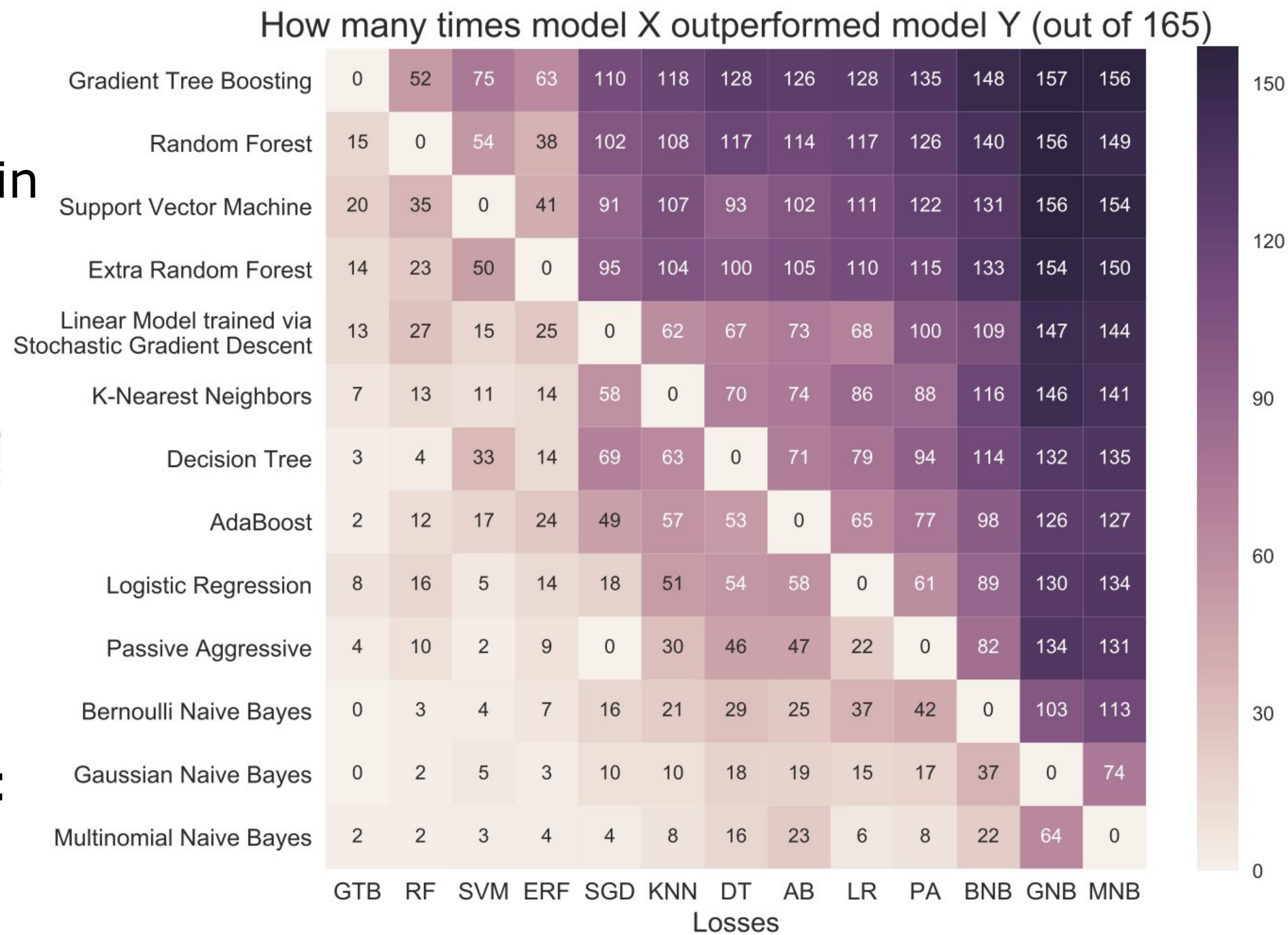


one algorithm family missing: neural networks!

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? Journal of Machine Learning Research, 15, 3133–3181.

"No Free Lunch" Theorem

Comparison of algorithms (implemented in SKlearn) on bioinformatics problems



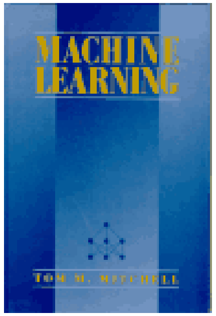
one algorithm family missing: neural networks!

Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A., & Moore, J. H. (2017). Data-driven Advice for Applying Machine Learning to Bioinformatics Problems. arXiv preprint Link: https://arxiv.org/abs/1708.05070

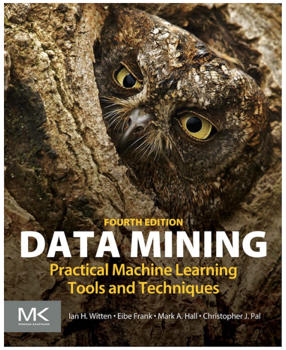
Recommended Readings

Textbooks

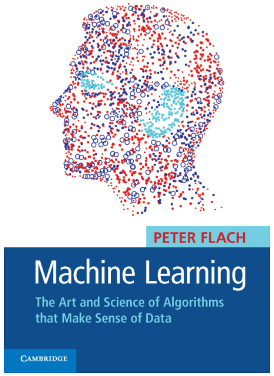
- Tom Mitchell, Machine Learning, McGraw Hill 1997. http://www-2.cs.cmu.edu/~tom/mlbook.html



- Ian H. Witten, Eibe Frank, Mark Hall, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 3nd edition, 2011. https://www.cs.waikato.ac.nz/ml/weka/book.html



- Peter Flach, Machine Learning: The Art and Sci ence of Algorithms that Make Sense of Data, Cam bridge University Press, 2012. http://www.cs.bris.ac.uk/~flach/mlbook/



Recommended Readings

TECHNISCHE UNIVERSITÄT DARMSTADT

Papers

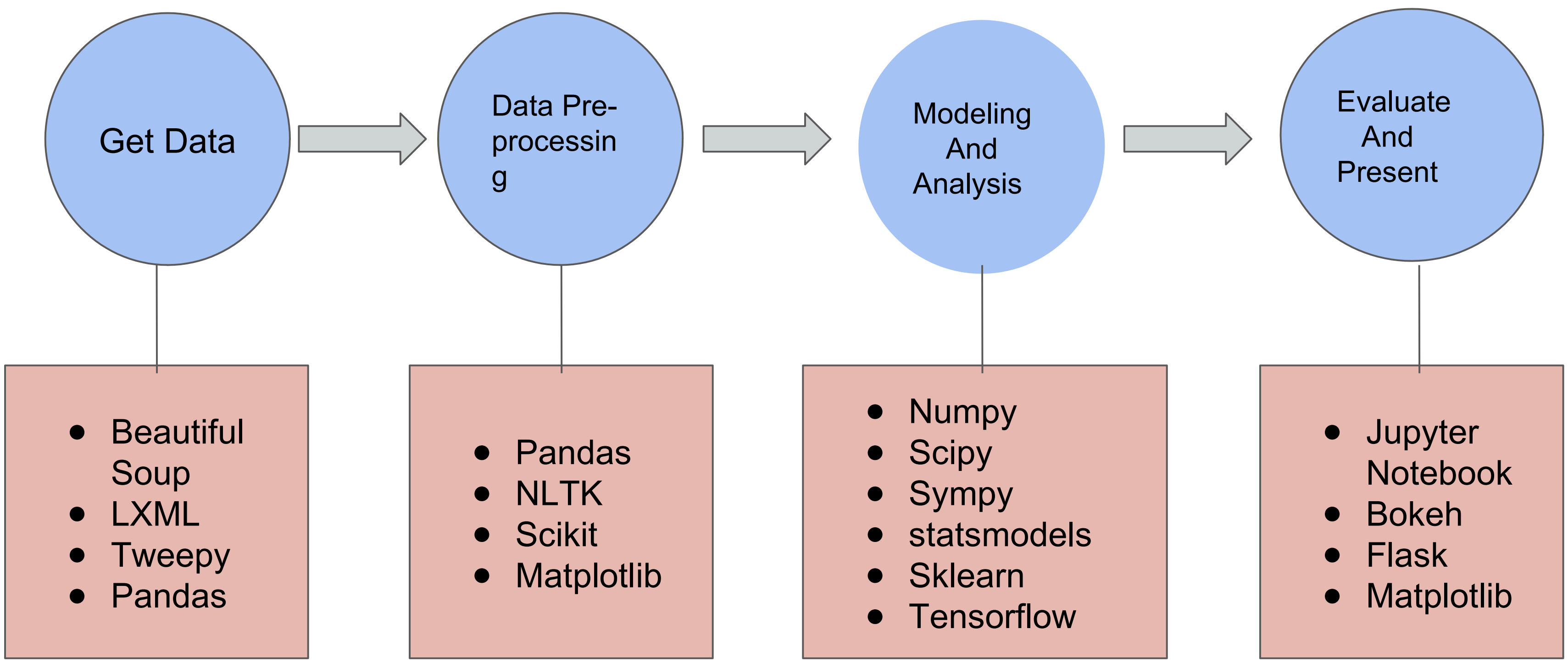
- Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87. https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf
- Zinkevich, M. (2017). Rules of Machine Learning: Best Practices for ML Engineering

Lectures

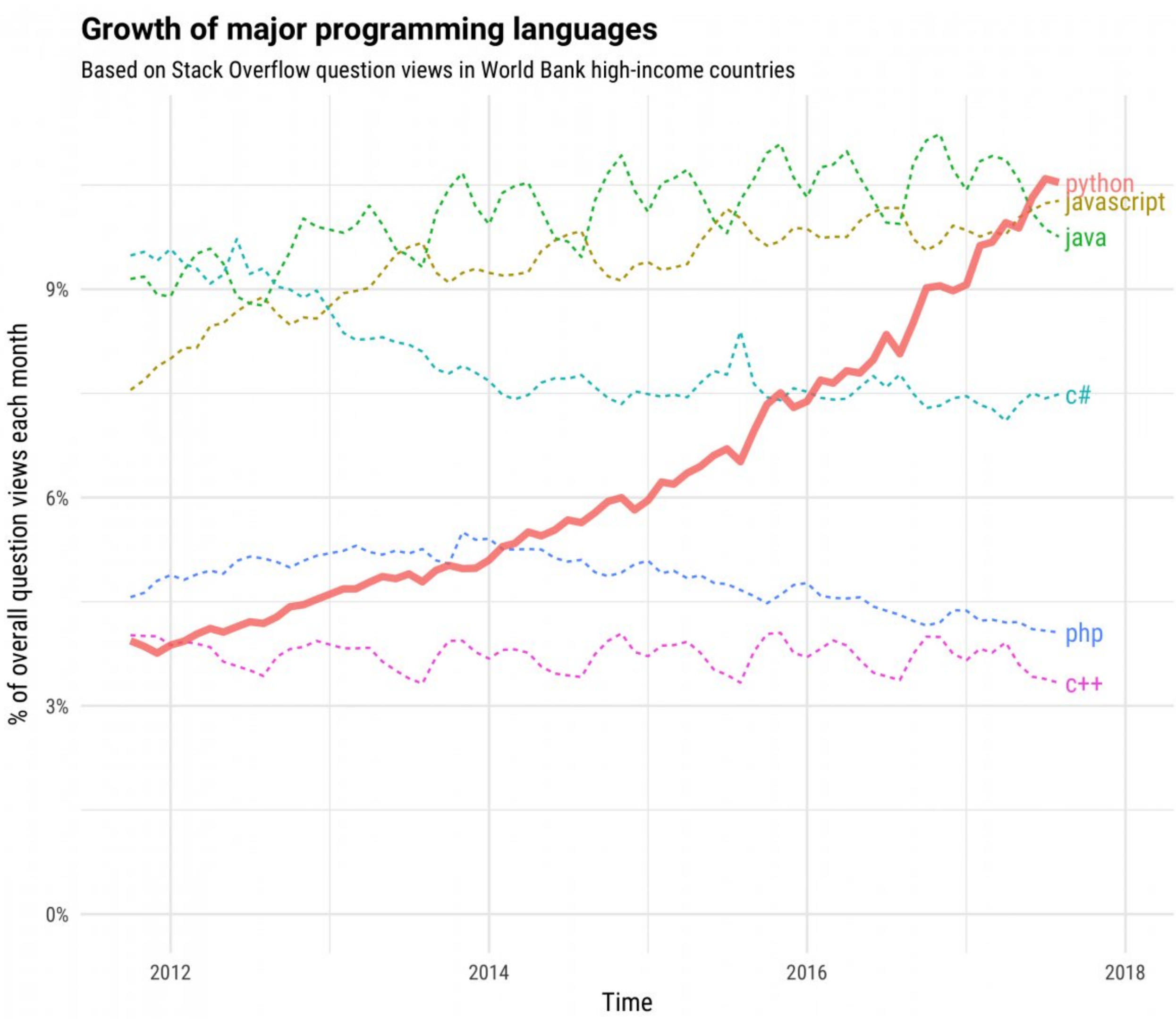
- Nathan Sprague: CS 444 Artificial Intelligence https://w3.cs.jmu.edu/spragunr/CS444/
- Nicholas Ruozzi: CS 6375 Machine Learning https://www.utdallas.edu/~nrr150130/cs6375/2017fa/index.html
- Steven Skiena: CSE 519 Data Science https://www3.cs.stonybrook.edu/~skiena/519/
- Andreas Mueller: COMS W4995 Applied Machine Learning http://www.cs.columbia.edu/~amueller/comsw4995s18/

Software Tools

- Java
  - Weka, RapidMiner, DeepLearning4j, Spark MLlib
- Python
  - Orange Suite
  - and ...
- R
  - R-Studio, CRAN
- ...



Software Tools



What is missing?

- Dimensionality Reduction
  - Feature Subset Selection
- Visualizations
- Hyperparameter Optimization
  - Auto Machine Learning
- Learning Theory
- Anomaly Detection
- Time Series Analysis
- Transfer Learning and Domain Adaptation
- Optimization
- Matrix Factorization
- Feature Learning
- Generative Learning etc. etc.

Content (may change)

- Introduction
- Instance based learning
- Decision tree learning
- Evaluation
- Ensemble learning
- Semi-supervised and unsupervised methods
- Excursions
  - Neural networks
  - Text Mining and information retrieval
  - Recommender Systems
  - Reinforcement learning