Medical Use Case Nephrology: Clinical Predictive Modeling

Borchert, Dr. Schapranow

Data Management for Digital Health
Winter 2023

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Agenda
Pillars of the Lecture

Hasso Plattner Institut

Medical Use Cases

- Biology Recap
- Oncology
- Nephrology
- Infectious Diseases

Technology Foundation

- Data Sources
- Data Formats
- Processing and Analysis
- Software Architectures

Machine Learning

Data

ML

Refine

Evaluate

Prediction + Probability

Agenda
Pillars of the Lecture

Medical Use Cases

- Biology Recap
- Oncology
- Nephrology
- Infectious Diseases

Technology Foundation

- Data Sources
- Data Formats
- Processing and Analysis
- Software Architectures

Machine Learning

Data

ML

Refine

Evaluate

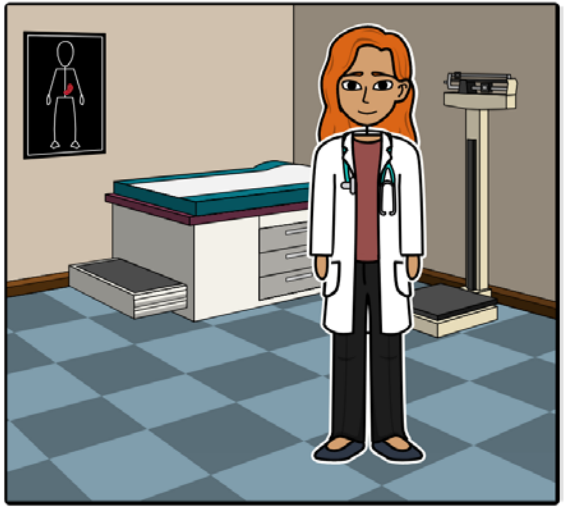
Prediction + Probability

Agenda

- Clinical Predictive Models (CPMs) build upon structured data and supervised le arning
- Requirements for CPMs
- Data for CPM development and how to prepare them
- Development and evaluation of CPMs
- Clinical deployment and monitoring of CPMs

Clinical Decision Making: Supporting Transplantation Nephrologists

- Persona: Susanne, nephrologist at transplantation center, 46yrs
- Consultation **before** and **after** transplantation
- Objectives:
  - Predict life expectancy and graft survival
  - Predict unplanned hospitalizations
  - Predict infections after transplantations
  - Analyze trends concerning kidney function
  - Identify similar patients for comparison
  - Assess whether the patient should wait for a "better" kidney



Clinical Decision Making: Predictive Analytics in Healthcare

- What could you do with this kind of information?

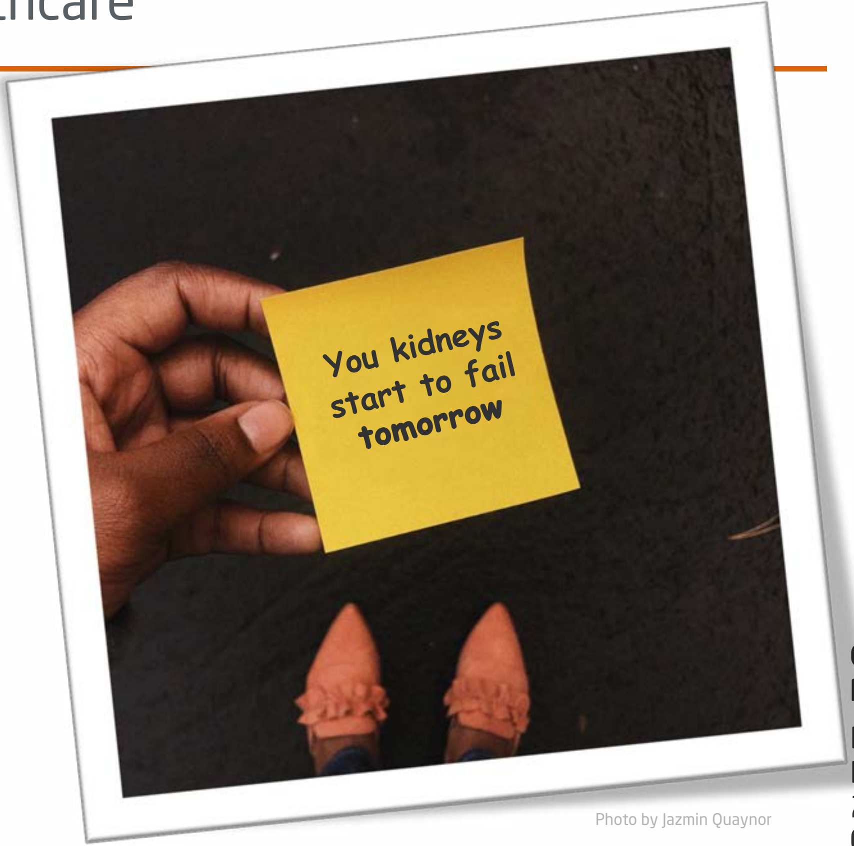


Photo by Jazmin Quaynor

Clinical Predictive Models in Intensive Care Units

Hasso Plattner Institut

- Aim: Risk prediction for critically ill patients in Intensive Care Units (ICU)
- Data: Routine physiological measurements, e.g. temperature, blood pressure, creatinine, white blood cell count, etc.
- Output: Maps to an individual numeric risk value for a specific clinical outcome

Source: Armed Forces Institute of Cardiology & National Institute of Heart Diseases (Pakistan)

Clinical Predictive Models in Intensive Care Units: Types of Scoring Systems

- First-day scoring systems
  - Acute Physiology and Chronic Health Evaluation (APACHE)
  - Simplified Acute Physiology Score (SAPS)
  - Mortality Prediction Model (MPM)
- Repetitive scoring systems
  - Organ System Failure (OSF)
  - Sequential Organ Failure Assessment (SOFA)
  - Organ Dysfunction and Infection System (ODIN)
  - Multiple Organ Dysfunction Score (MODS)
  - Logistic Organ Dysfunction (LOD)

http://scoringexpert.pl/2017/01/01/model-scoringowy-troche-teorii/

Clinical Predictive Models in Intensive Care Units: Comparison of First-day ICU Scores

System	Data collected	Physio-logical values	Other required data	Req. data items	Mortality pred. perf.
APACHE IV	First day on ICU (16-32h)	17	Age, six chronic health variables, ICU admission diagnosis, ICU admission source, LOS prior to ICU admission, emergency surgery, thrombolytic therapy, FiO_2, mechanical ventilation	32	AUC=88.0%, n=52,647
SAPS III	Prior to and within 1h of ICU admission	10	Age, six chronic health variables, ICU admission diagnosis, ICU admission source, LOS prior to ICU admission, emergency surgery, infection on admission, four variables for surgery type	26	AUC=84.8%, n=16,784
MPM III	Prior to and within 1h of ICU admission	3	Age, three chronic health variables, five acute diagnosis variables, admission type (e.g., medical-surgical) and emergency surgery, CPR within 1 h of ICU admission, mechanical ventilation, code status	16	AUC=82.3%, n=50,307

https://www.researchgate.net/figure/A-comparison-of-intensive-care-unit-ICU-scoring-systems-from-47-with-permission_tbl1_273059579

Clinical Predictive Models in Intensive Care Units: Acute Physiology & Chronic Health Evaluation II (APACHE II)

- AaDO2 or PaO2 (depending on FiO2)
- Potassium (serum)
- Creatinine
- Temperature (rectal)
- Sodium (serum)
- Hematocrit
- Mean arterial pressure
- Respiratory rate
- White blood cell count
- pH arterial
- Heart rate
- Glasgow Coma Scale

- Age Points + Chronic Health Points+ Acute Physiologic Score
- Try it out: https://www.mdcalc.com/apache-ii-score

APACHE II SCORE

AGE Points	
≤ 44y	0
45-54y	2
55-64y	3
65-74y	5
≥75y	6

CHRONIC HEALTH Points	
Non-operative, or emergency post-op & any conditions below*	5
Elective operation & any conditions below*	2

*Cirrhosis w/ portal Hypertension or encephalopathy; class IV angina, chronic hypoxia, ↑CO2 or polycytemia; chronic dialysis; immunocompromised

TOTAL APACHE SCORE = AP + CHP + APS

Sum Age Points (AP) + Chronic Health Points (CHP) + Acute Physiologic Score (APS) points.

*1 Sum all variables 1-12 for Acute Physiologic Score (APS) (use one variable each for 5 and 9).

Use the worst value from the preceding 24h.

APACHE II: a severity of disease classification system. Crit Care Med 1985;13:818-29.

ACUTE PHYSIOLOGIC SCORE*1 (APS)

	Physiologic Variable	Points 4	3	2	1	0	1	2	3	4
1	Temp °F	≤85.9	86.0-89.5	89.6-93.1	93.2-96.7	96.8-101.2	101.3-102.1		102.2-105.7	≥105.8
	°C	≤29.9	30-31.9	32-33.9	34-35.9	36 - 38.4	38.5-38.9		39-40.9	≥41
2	HR, bpm	≤39	40-54	55-69		70-109		110-139	140-179	≥180
3	MAP, mmHg	≤49		50-69		70-109		110-129	130-159	≥160
4	RR, bpm	≤5		6-9	10-11	12-24	25-34		35-49	≥50
5	Oxygenation: Use A-a Gradient (5a) if FiO_2 ≥0.5 or use PaO_2 (5b) if FiO_2 <0.5 (*see page 17*)									
5a	A-a Gradient					<200		200-349	350-499	≥500
5b	PaO_2	≤54	55-60		61-70	>70				
6	Na^+ (S, mmo/L)	≤110	111-119	120-129		130-139	150-154	155-159	160-179	≥180
7	K^+ (S, mmo/L)	≤2.4		2.5-2.9	3.0-3.4	3.5-5.4	5.5-5.9		6.0-6.9	≥7.0
8	Cr (S, mg/dL)			<0.6		0.6-1.4		1.5-1.9	2.0-3.4	≥3.5
9	Arterial pH is preferred. Use venous HCO3 if no ABGs.									
9a	pH (arterial)	≤7.14	7.15-7.24	7.25-7.32		7.33-7.49	7.5-7.59		7.6-7.69	≥7.7
9b	HCO_3 (venous)	≤14	15-17.9	18-21.9		22-31.9	32-40.9		41-51.9	≥52
10	WBC, cells/uL	≤1.0		1.0-2.9		3.0-14.9	15-19.9	20-39.9		≥40
11	Hct, %	≤20		20-29.9		30-45.9	46-49.9	50-59.9		≥60
12	GCS coma	Score = 15 – GCS Score (see below, Record e.g.: "GCS 9 = E2 V4 M3 at 17:35h".)								

Score	Mortality
0 - 4	4%
5 - 9	4%
10 - 14	15%
15 - 19	25%
20 - 24	40%
25 - 29	55%
30 - 34	75%
> 34	85%

GLASGOW COMA SCALE (GCS)

*Teasdale G, Jennett B. Lancet 1974,2:81-84.

EYE Opening	Best VERBAL	Best MOTOR	Points
		follows commands	6
	oriented	localizes pain	5
spontaneous	confused	withdraws to pain	4
to command	inappropriate words	flexor response	3
to painful stimuli	incomprehensible	extension (abnl)	2
no response	no response	no response	1

SCORE: *Sum Points (eye+verbal+motor categ).*

Severe ≤ 8. Mod = 9-12. Minor ≥ 13.

http://www.scymed.com/en/smnxpw/pwfbd770.htm

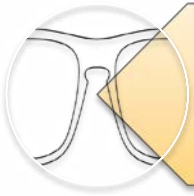
Clinical Predictive Models in Intensive Care Units: Glasgow Coma Score

- Neurological scale
- Give a reliable and objective way of recording the conscious
- Initially used to assess a person's level of consciousness after a head injury
- Now used by first responders, EMS, nurses, and doctors
- Part of several ICU scoring systems, including APACHE II, SAPS II, and SOFA

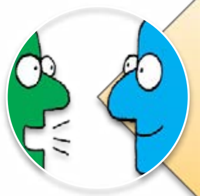
https://nurse.org/articles/glasgow-coma-scale/

Clinical Predictive Models in Intensive Care Units : Glasgow Coma Score (cont'd)

Behavior | Response



Eye Opening Response (E)
- 4 Spontaneously
- 3 To speech
- 2 To pain
- 1 No response



Verbal Response (V)
- 5 Oriented to time, person and place
- 4 Confused
- 3 Inappropriate words
- 2 Incomprehensible sounds
- 1 No response

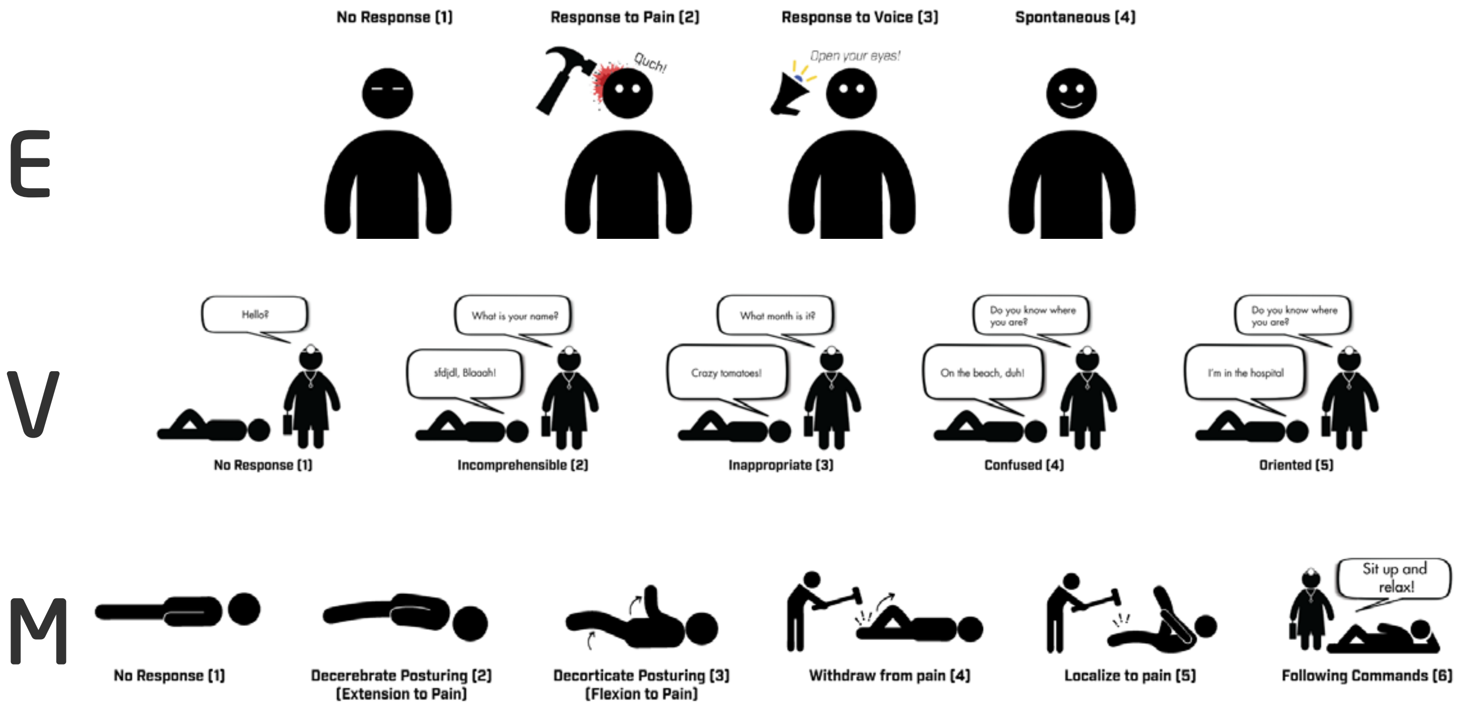
Motor Response (M)
- 6 Obeys command
- 5 Moves to localized pain
- 4 Flex to withdraw from pain
- 3 Abnormal flexion
- 2 Abnormal extension
- 1 No response

Total Score

Mild 13 - 15
Moderate 9 - 12
Severe 3 - 8

Icons from: https://gunnar.com, https://melibeeglobal.com/blog/2015/01/the-lack-of-non-verbal-communication-in-a-digital-world/, http://www.isfsports.org/gymnastics

Clinical Predictive Models in Intensive Care Units: Glasgow Coma Score Calculation

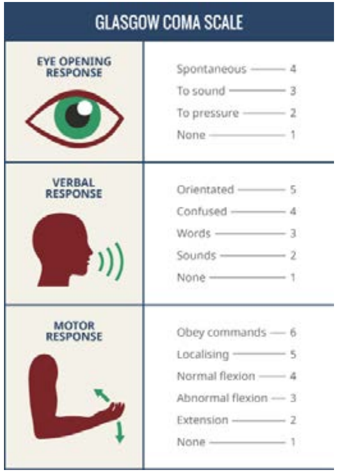


https://codehealth.io/library/article-1/glasgow-coma-scale/

Clinical Predictive Models in Intensive Care Units: Glasgow Coma Score Calculation (cont'd)

- Adult, moves the hand away when applying pressure on the nail bed. The patient can make words but not form sentences. The patient opens the eyes to pain, but not to speech.

E		V		M		
2	+	3	+	4	=	9



https://www.thompsons-scotland.co.uk/serious-head-and-brain-injury/brain-injury-solicitors-scotland/brain-injury-claims-and-the-glasgow-coma-scale

Recap: NephroCAGE: German-Canadian Consortium on AI for Improved Kidney Transplantation Outcome

Hasso Plattner Institut

- Applying AI technology for prediction of severe post-transplant risks
- Access to multi-national transplant data from 20+ years
- As first of its kind: Implements NephroCAGE federated learning infrastructure to keep sensitive data protected whilst allowing multi-site data analyses

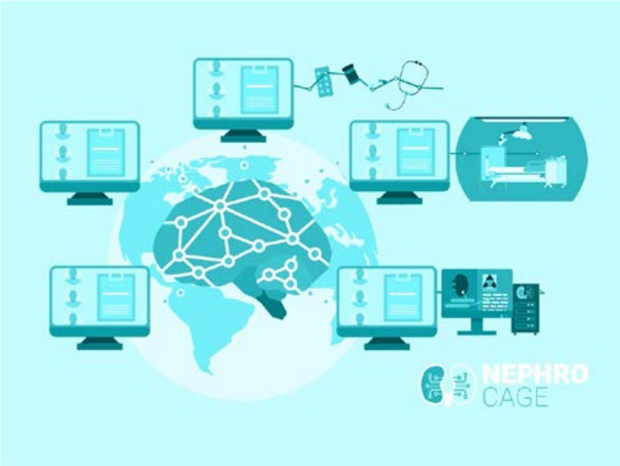
NEPHRO CAGE

NephroCAGE:
Nephrology Disease Cooperation between Canada and Germany for Applied AI

Real-world Demonstrator: CHARITÉ Universitätsmedizin Berlin, UBC, PIRCHE

Learning Systems and Federated Learning: HPI Hasso Plattner Institut, KIT Karlsruhe Institute of Technology

Data Providers and Clinical Experts: CHARITÉ Universitätsmedizin Berlin, Genome Québec, Genome Canada, McGill University Health Centre, CHUM, UBC, Genome BritishColumbia



Supported by:

Federal Ministry for Economic Affairs and Climate Action

on the basis of a decision by the German Bundestag

CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

KIT
Karlsruher Institut für Technologie

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

PIRCHE

Genome Québec

CHUM

Genome BritishColumbia

GenomeCanada

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

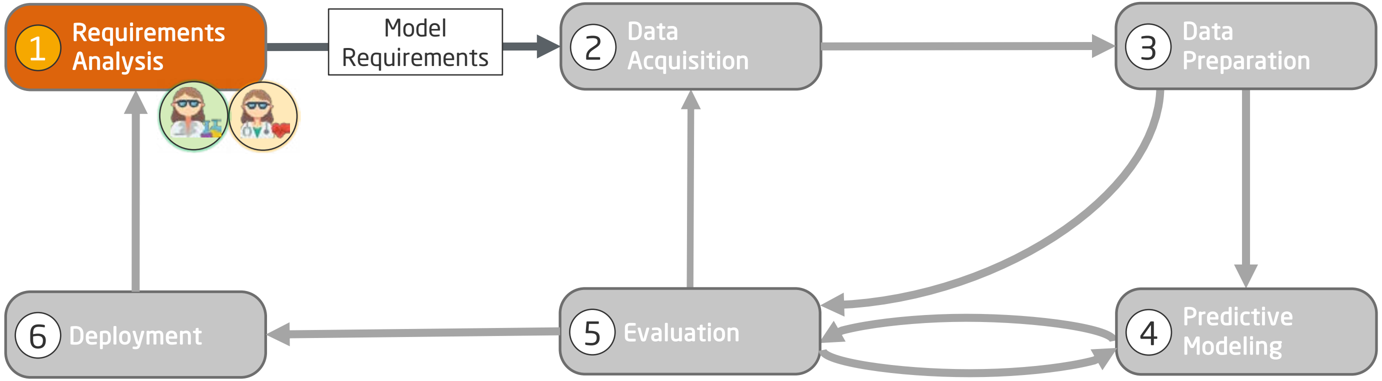
McGill UNIVERSITY

Université de Montréal

UBC THE UNIVERSITY OF BRITISH COLUMBIA

https://nephrocage.org/vision-objectives/

1. Requirements Analysis



Roles

- Data Scientist
- Domain Expert
- (Data) Engineer

1. Requirements Analysis: Use Case NephroCAGE

- Clinical predictive modeling of severe post-transplant endpoints
  I. Allograft failure
  II. Allograft rejection
  III. Patient death
- Time window: 1-5 years post-transplant
- Data:
  - Use of history of transplant data
  - From donors and recipients
  - Clinical, laboratory, transplant-related immunological, etc.
- Beware: Impact of model outcome on treatment process and future acquired data

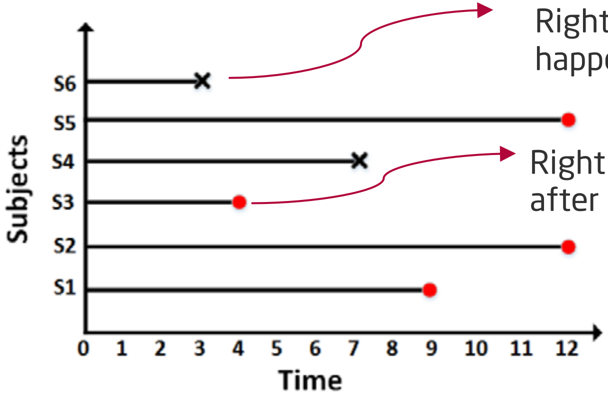
1 Requirements Analysis

- Business metrics
- Acceptance criteria
- Data protection
- Ethics
- Interpretability requirements

NEPHRO CAGE

1. Requirements Analysis: Censored Data

- **Right censoring:** Subject leaves study before an event occurs or the study ends before the event has occurred
- **Left censoring:** event of interest has already occurred before enrolment



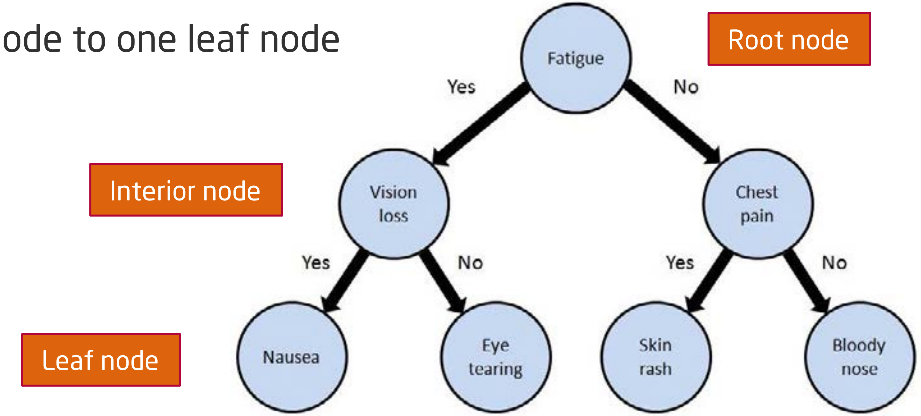
Right censored: event would have happened, but patient died before

Right censored: Event happened after patient left study

Application field	Start of study event	End of study event	Censoring example
Medical research on kidney failure	Time the patient received the new kidney	Time the patient experienced graft failure	Patient died due to a cardiovascular disease

1. Requirements Analysis: Interpretability of Models: Decision Trees

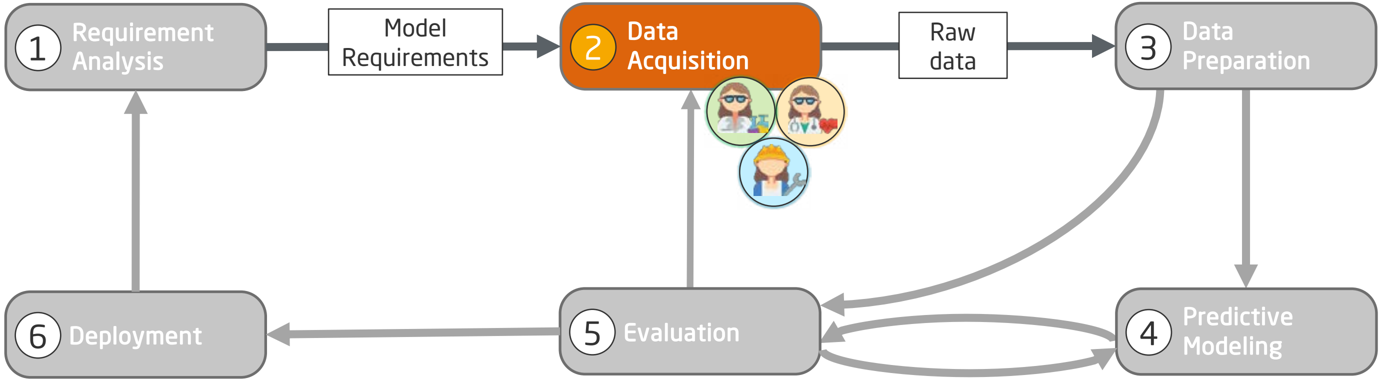
- Decision trees are human-readable from the root node to one leaf node
- Decision rules are often derived from data
- Advantages:
  - High interpretability
  - Can be combined with other algorithms
  - Requires little data preparation
- Disadvantages:
  - With an increasing number of dimensions, the decision trees becomes complex
  - May lack generalization, prone to overfitting
  - Creates bias if classes are unbalanced



http://web.eecs.umich.edu/~cscott/research/decision_tree.jpg

http://scikit-learn.org/stable/modules/tree.html

2. Data Acquisition



Roles

- Data Scientist
- Domain Expert
- (Data) Engineer

2. Data Acquisition
NephroCAGE Data Set

- Transplant data for 10+ yrs from multiple transplant centers in DE & CA
- Public reference: Scientific Registry of Transplant Recipients (SRTR)

- **NephroCAGE Data Set I**:
  - Available at all centers
  - Ex.: Recipient and donor, recipient biomarkers
- **NephroCAGE Data Set II:**
  - Involves acquisition of additional data or extraction from additional clinical systems
  - Ex.: Biopsy, HLA data, medication and hospitalization

2 Data Acquisition

- Data collection
- ETL
- Data integration
- De-identification

2. Data Acquisition:
NephroCAGE Data Set (cont'd)

	NephroCAGE Data Set	CHA	UBC	MUHC	CHUM
Period	**1998-2020**	1998-2020	2008-2018	2012-2019	2011-2019
Duration (yrs)	**23**	23	11	8	9
Patients	**8,067**	4,742	2,510	415	400
Male vs. female [n] (%)	**5,081 (63%): 2,986 (37%)**	2,940 (62%): 1,802 (38%)	1,606 (64%): 904 (36%)	279 (67%): 136 (33%)	256 (64%): 144 (36%)
Age (yrs), mean (SD)	**51.7 (14.3)**	51.3 (14.0)	51.9 (15.3)	55.6 (12.4)	52.0 (12.8)

Hasso Plattner Institut

2 Data Acquisition

- Data collection
- ETL
- Data integration
- De-identification

NEPHRO CAGE

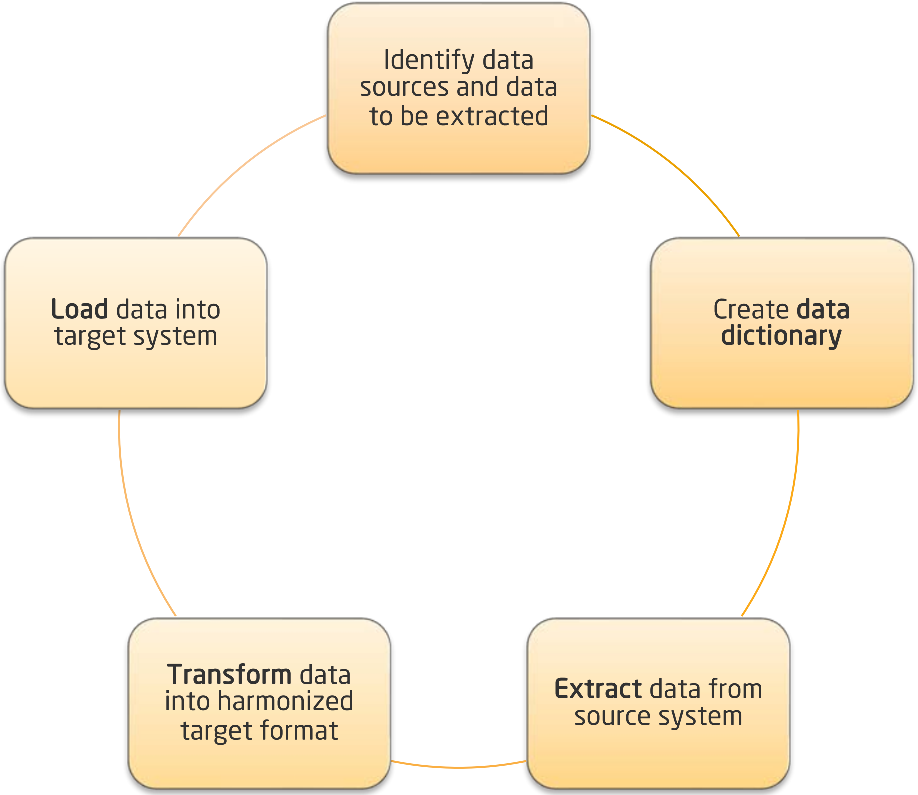
2. Data Acquisition: NephroCAGE Data Set (cont'd)

NEPHRO CAGE

Hasso Plattner Institut

- Type of dialysis
- Trans-plantation date
- Donor-specific antibody
- Graft loss date
- Age recipient
- Date of death
- Sex of recipient
- HLA mis-matches
- Type of donation
- Age of donor
- Sex of donor
- Delayed graft function
- Cold ischemia time
- Creatinine
- Hospital-ization
- Medication
- Protein
- Biopsy

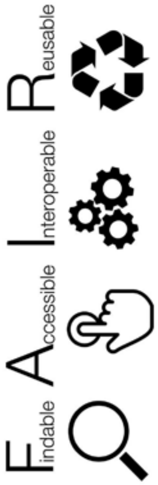
2. Data Acquisition: Extract, Transform, Load (ETL)



2 Data Acquisition

- Data collection
- ETL
- Data integration
- De-identification

2. Data Acquisition: Data Integration: NephroCAGE Data Dictionary



1	original_column_names	dtypes	new_column_names	extended_datase	minimum_datase	description	categorie	note	RangeValue
2	PatientID	int64	REC_ID	1	1				
3	TransplantationID	int64	TX_ID	1	1				
4	SpenderID	numeric	DON_ID	1	1				1177.0 - 8136.0
5	Date of Death	datetime64[ns]	DEATH_DATE	1	1				
6	Gender	category	REC_SEX	1	1		['m' 'w']		
7	ESRD	category	REC_ESRD	1	0				
8	ESRD2	category	REC_ESRD_2	0	0				
9	type of dialysis	category	REC_DIAL_TYPE	1	0		['HD' 'CAPD' 'keine']		
10	number of transplantation	numeric	REC_NUM_TX	1	0				1.0 - 4.0
11	age_recipient	numeric	REC_AGE	1	0				18.13 - 86.0
12	Center	category	CENTER	1	0		['CCM' 'CVK']		
13	p_bloodgroup	category	REC_BLOOD_GROUP	1	0				
14	p_height	numeric	REC_HEIGHT	1	0				110.0 - 205.0
15	CMV_AK	category	REC_CMV	1	0		['positiv' 'negativ']		
16	HbS_AG	category	REC_HBS_AG	0	0		['positiv' 'negativ']		
17	HCV_AK	category	REC_HCV	0	0		['negativ' 'positiv']		
18	anti_HIV	category	REC_HIV	0	0		['negativ' 'positiv']		
19	EBV_IgG	category	REC_EBV	0	0		['positiv' 'negativ']		
20	delayed graft function_inverse	category	DGF	1	0		['ja' 'nein']		
21	number_dialysis	numeric	DGF_DIAL_NUM	1	0	Dialysis after TX			
22	cold ischemia time	numeric	CIT_HOUR	1	1	number (if filled)			0.5 - 34.0
23	MMA_broad	category	MMA	0	1				0.0 - 2.0
24	MMB_broad	category	MMB	0	1				0.0 - 3.0
25	MMDR_broad	category	MMDR	0	1				0.0 - 3.0
26	MM_broad	category	MM	1	1				0.0 - 6.0
27	Loss cause	category	GF_CAUSE	1	1	we have 1 Loss cause			
28	Loss cause	category	GF_CAUSE_2	0	0				
29	Programm	category	TX_PROGRAMM	0	0	ETKAS_AM_ESP_HU	['ETKAS' 'AM' 'ESP']	skip	

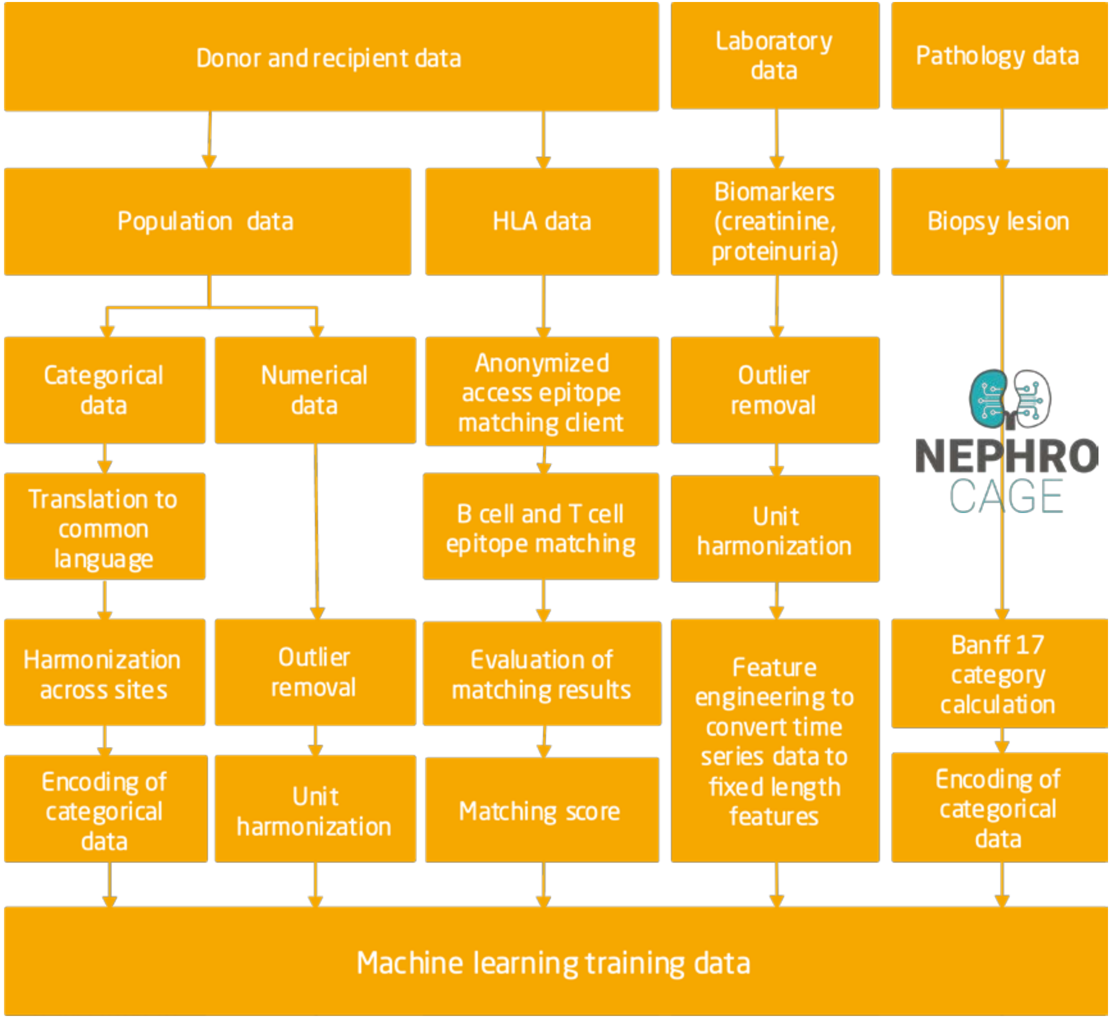
2 Data Acquisition

- Data collection
- ETL
- Data integration
- De-identification

NEPHRO CAGE

3. Data Preparation in NephroCAGE (cont'd)

- How to obtain training data for development of CPMs?
- Data sources
  - Donor and recipient data
  - Lab data
  - Pathology data
- Every data item requires extraction, harmonization and pre-processing aligned across sites and countries



2. Data Acquisition: De-Identification of Dates in NephroCAGE Data Set

Hasso Plattner Institut

- Set to the start of the month
- Relative days to date of the transplant
- Implicit information is accessible after de-identification
- Original data set:

TX_DATE	CREA	LAB_DATE	ID
09.01.2020	2.7	10.01.2022	23
09.01.2020	1.7	11.01.2022	23
09.01.2020	1.5	12.01.2022	23

- De-identified data set:

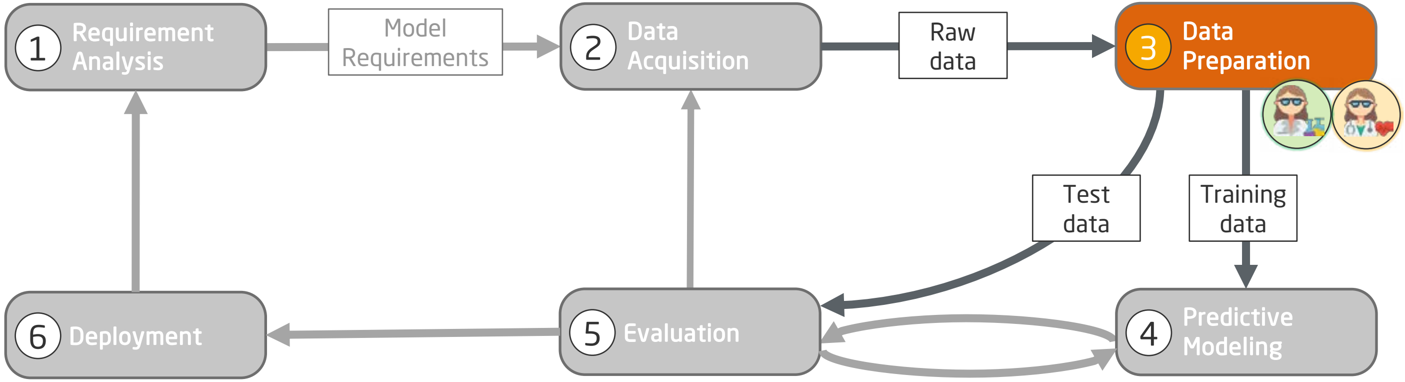
TX_DATE	CREA	LAB_DATE (d)	ID
01.01.2020	2.7	+1	23
01.01.2020	1.7	+2	23
01.01.2020	1.5	+3	23

2 Data Acquisition

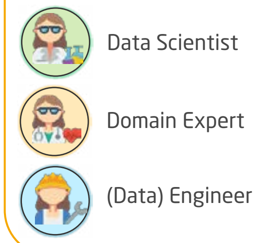
- Data collection
- ETL
- Data integration
- De-identification

NEPHRO CAGE

3. Data Preparation



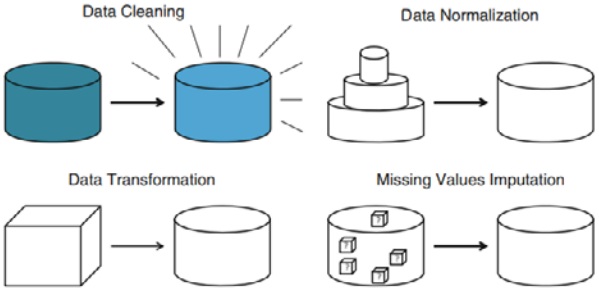
Roles



3. Data Preparation: Involved Aspects

1. How to understand available data / gain insights?
   → Data Exploration
2. How to harmonize data? → Data Cleansing, Transformation
3. How to combine data from different departments, devices, units → Data Normalization
4. How to handle missing data? → Imputation
5. How to derive input for the model development? → Feature engineering

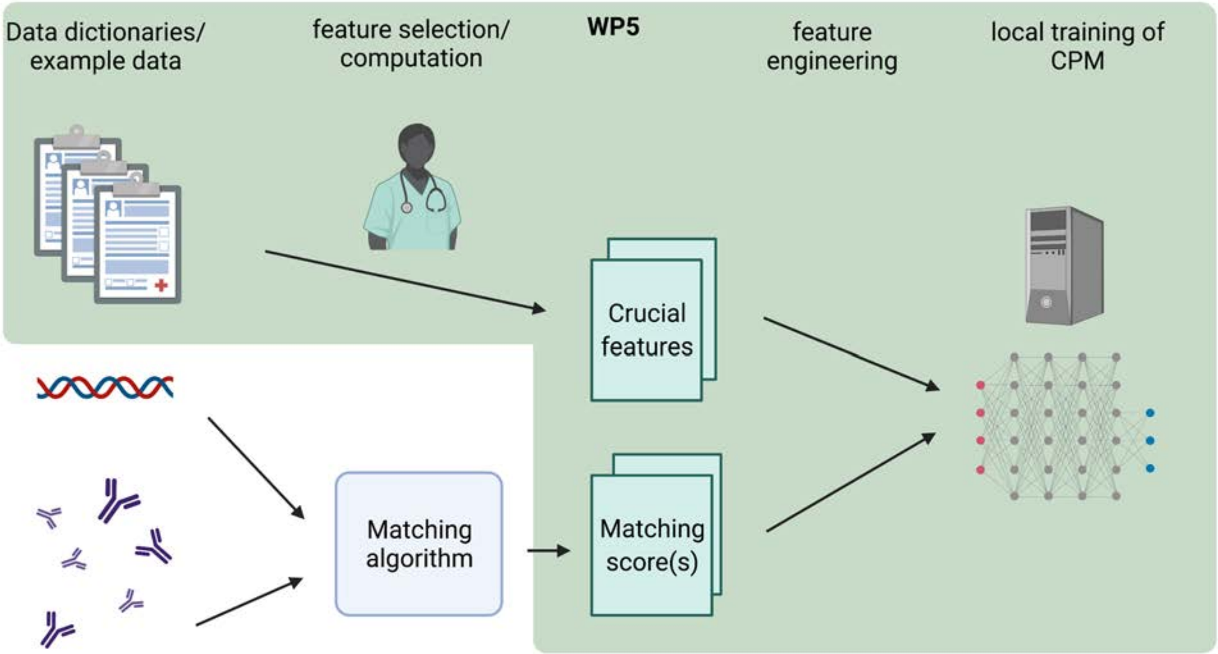
- Bear in mind: Consider tool support for the above steps



3 Data Preparation

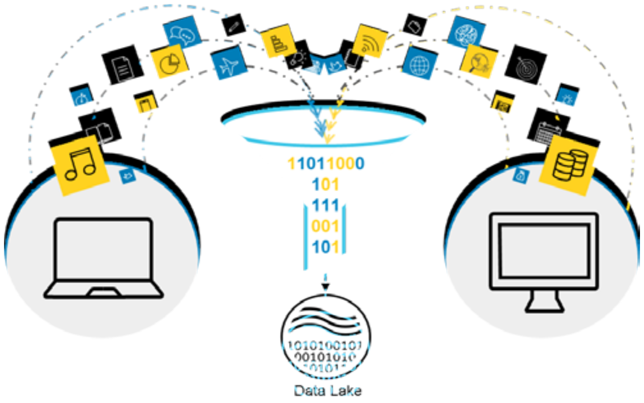
- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

3. Data Preparation in NephroCAGE



3. Data Preparation: Data Transformation

- Scaling: Data may contain attributes with a mixtures of scales, but ML methods require data attributes to have the same scale
- Decomposition: Features may represent a complex concept that may be more useful to a ML method when split into its parts, e.g. data, zip code, etc.
- Aggregation: Features that might be aggregated into a single feature



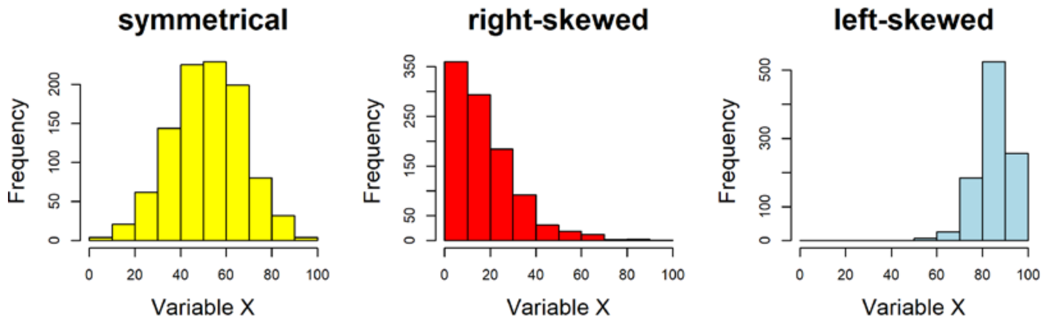
https://blog.dellemc.com/en-us/digital-transformation-just-got-easier-with-analytic-insights/

3 Data Preparation

- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

3. Data Preparation: Data Transformation

- Box-Cox transformation: transform non-normal dependent variables to normal symmetrical shape
- Log transformation: for strongly right-skewed data
- Sqrt transformation: for slightly right-skewed data
- Power transformation: for left-skewed data



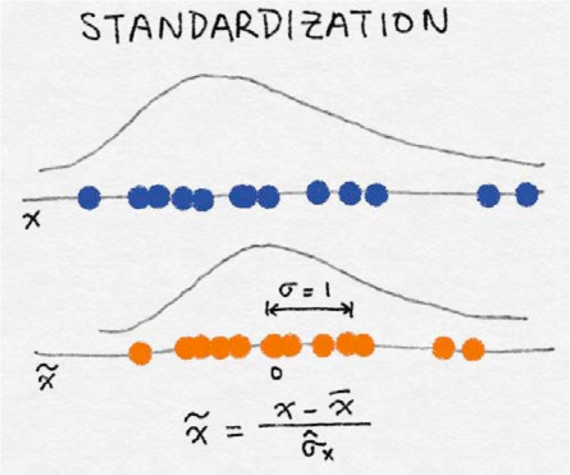
https://www.davidzeleny.net/anadat-r/doku.php/en:data_preparation

3 Data Preparation

- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

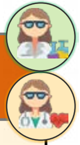
3. Data Preparation:
Variance Scaling / Standardization

- $\tilde{x} = \frac{x - \text{mean}(x)}{\text{sqrt}(\text{var}(x))}$
- Let x be an individual feature value
- Variance scaling:
  - Subtract the mean of the feature from x, and
  - Divide by std. dev.
- Result: Standardized feature has a mean of 0 and a variance of 1
- If the original feature showed a Gaussian distribution, the scaled feature will keep this property.



Feature Engineering for Machine Learning
Principles and Techniques for Data Scientists
Alice Zheng and Amanda Casari, O'Reilly, 2018

Hasso Plattner Institut

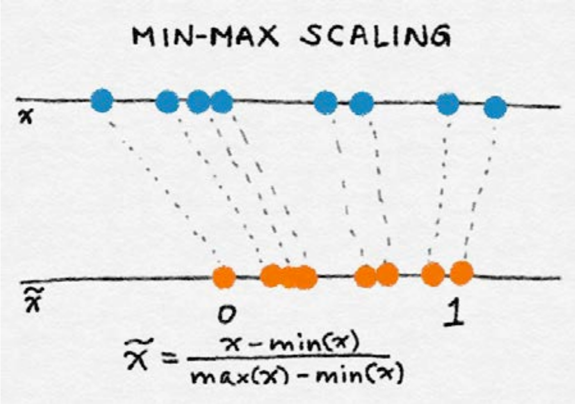


3 Data Preparation

- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

3. Data Preparation: Min-Max Scaling

- $\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)}$
- Let $\min(x)$ and $\max(x)$ be the minimum and maximum values of this feature across the entire dataset
- Result: Min-max scaling squeezes/stretches all values into the interval [0, 1]



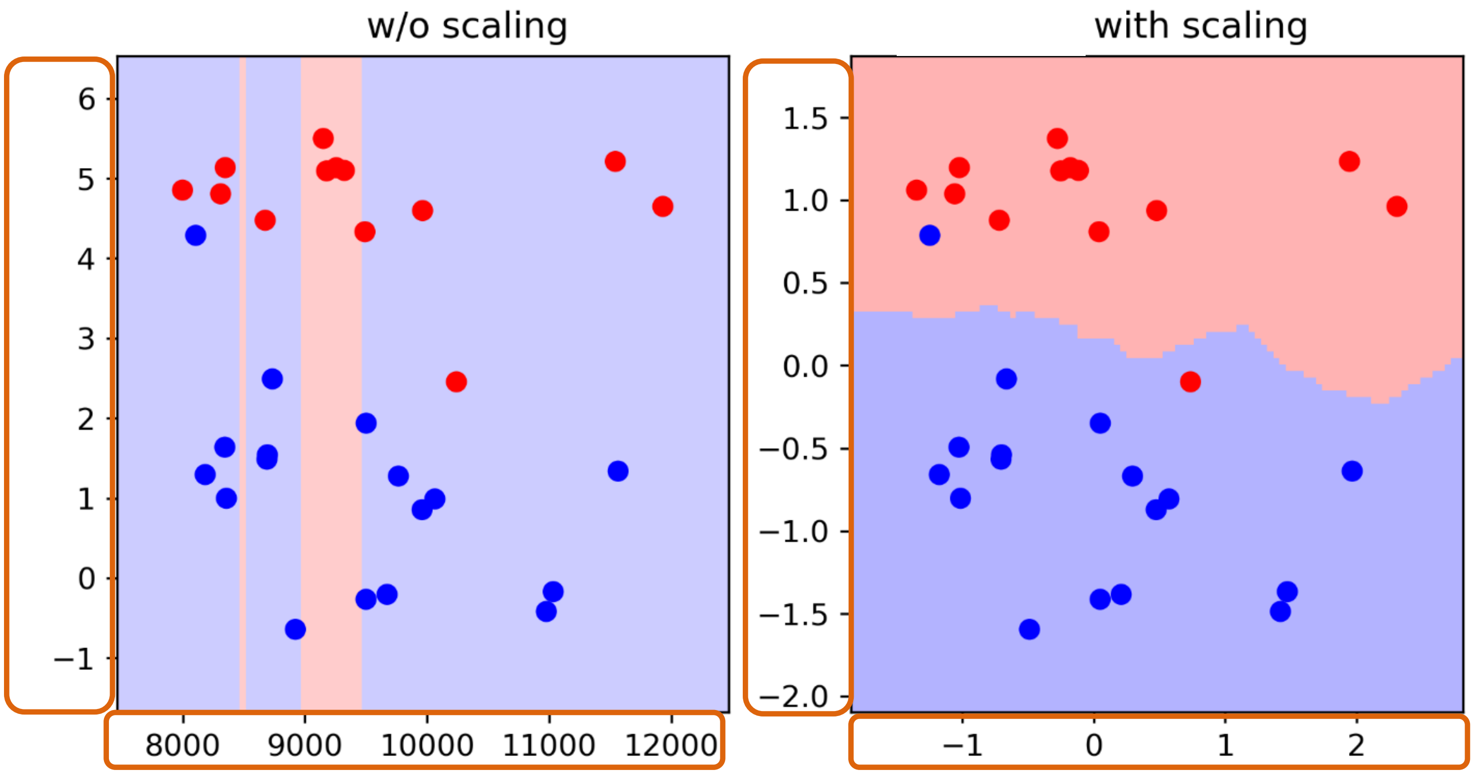
Feature Engineering for Machine Learning
Principles and Techniques for Data Scientists
Alice Zheng and Amanda Casari, O'Reilly, 2018

Hasso Plattner Institut

3 Data Preparation

- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

3. Data Preparation: Benefits from Scaling?



https://blog.dellemc.com/en-us/digital-transformation-just-got-easier-with-analytic-insights/

Hasso Plattner Institut

3 Data Preparation

- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

3. Data Preparation: Principal Component Analysis (PCA)

- Aim: Data reduction of high-dimensional data sets
- Transformation of data to a lower number of dimensions without losing information

Pros	Cons
Removes correlated features	Independent variables become less interpretable
Reduces chance for overfitting	Data standardization is must before PCA
Improves visualization	Loss of information

3. Data Preparation:
Example: Eating and Drinking in the UK (cont'd)

1. High-dimensional dataset

	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

http://setosa.io/ev/principal-component-analysis/

Hasso Plattner Institut

3 Data Preparation

- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

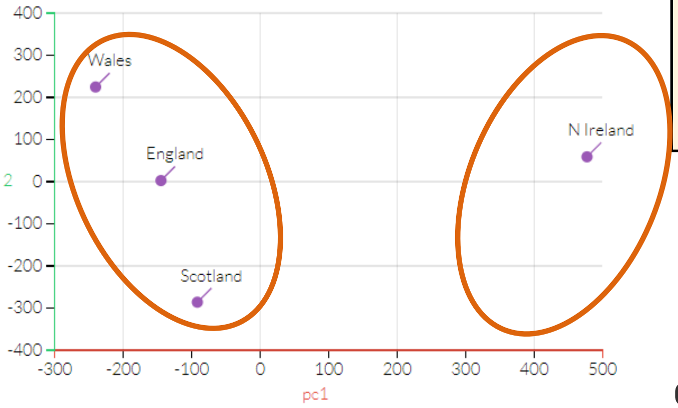
3. Data Preparation: Example: Eating and Drinking in the UK (cont'd)

1. High-dimensional dataset

	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

http://setosa.io/ev/principal-component-analysis/

2. Mapping n dim to PCA (2D)



3 Data Preparation

- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

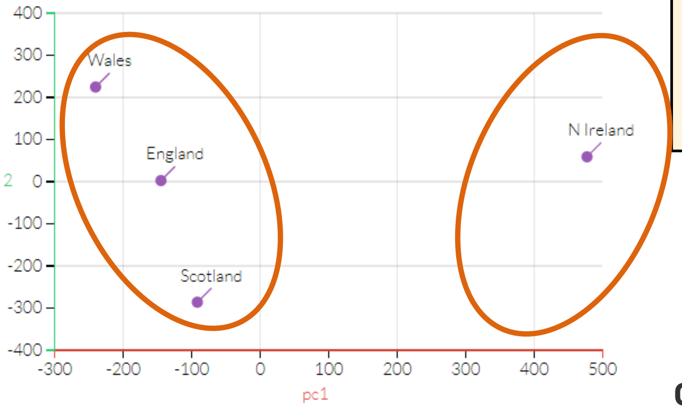
3. Data Preparation: Example: Eating and Drinking in the UK (cont'd)

1. High-dimensional dataset

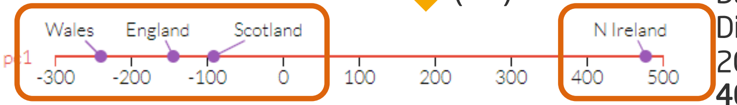
	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

http://setosa.io/ev/principal-component-analysis/

2. Mapping n dim to PCA (2D)



3. Mapping to pc1 (1D)



3 Data Preparation

- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

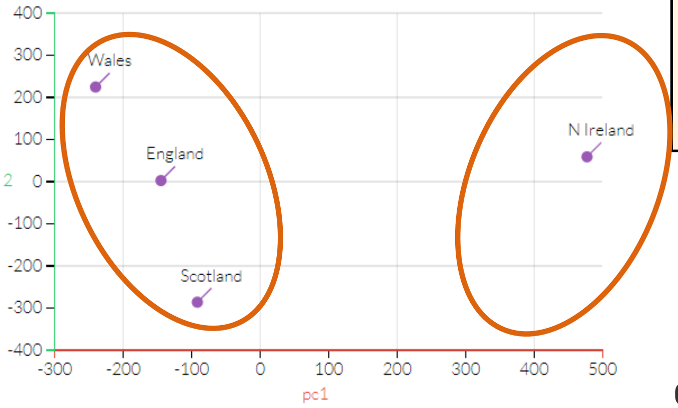
3. Data Preparation: Example: Eating and Drinking in the UK (cont'd)

1. High-dimensional dataset

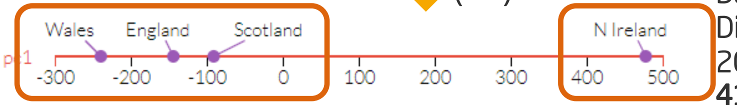
	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

http://setosa.io/ev/principal-component-analysis/

2. Mapping n dim to PCA (2D)



3. Mapping to pc1 (1D)



3 Data Preparation

- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

3. Data Preparation:
Data Imputation: Mean/Median Values

Hasso Plattner Institut

- Calculating the mean/median of the non-missing values in a column

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

Pros	Cons
Easy and fast	Correlations between features are ignored
Works well in small numerical datasets	Poor results on encoded categorical features
	Not very accurate

3 Data Preparation

- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

3. Data Preparation:
Data Imputation: Most Frequent or Zero/Constant Values

- Statistical strategy to impute missing values using most frequent values
- Zero or Constant imputation replaces the missing values with either zero or any constant value you specify

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

df.fillna(0) →

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	0.0
1	9	0.0	9.0	0	7.0
2	19	17.0	0.0	9	0.0

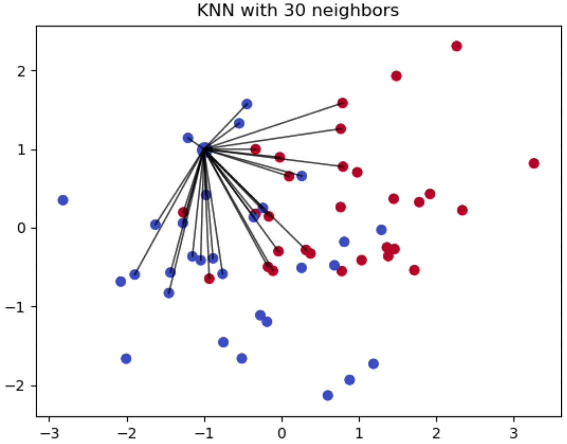
Pros	Cons
Works also with categorical features	Correlations between features are ignored
	Might introduce bias in the data

Hasso Plattner Institut

3 Data Preparation
- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

3. Data Preparation: Data Imputation: k-nearest Neighbors

- *k*-nearest neighbors is classification algorithm
- Algorithm uses feature similarity to predict the values of new data points
- Imputed data point is assigned to the class according the class with the most of its k neighbors



Pros	Cons
Can be much more accurate than the mean, median or most frequent imputation methods (It depends on the dataset)	Computationally expensive. KNN works by storing the whole training dataset in memory
	K-NN is quite sensitive to outliers in the data (unlike SVM)

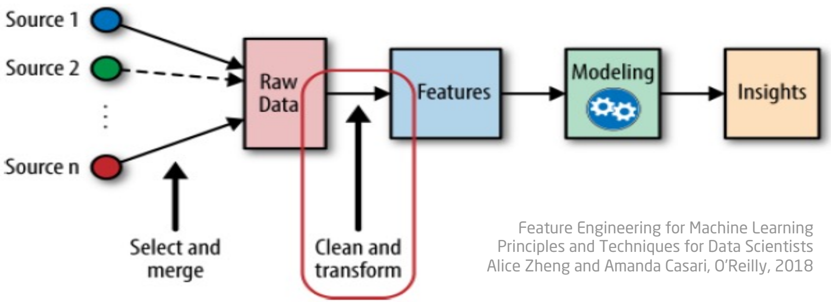
Hasso Plattner Institut

3 Data Preparation

- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

3. Data Preparation: Feature Engineering

	Feature Selection	Feature Extraction
Aim	Reduce dimension of feature space whilst representing the same information	
Approach	Select subset of features, e.g. filters, ML wrapper, or combined as embedded methods	Transform existing features into more informative features, e.g. automatic via linear PCA or non-linear autoencoder or manual extraction using subject-matter expertise
Effect	Improved model performance, reduced overfitting, faster training and inference, better interpretability, etc.	



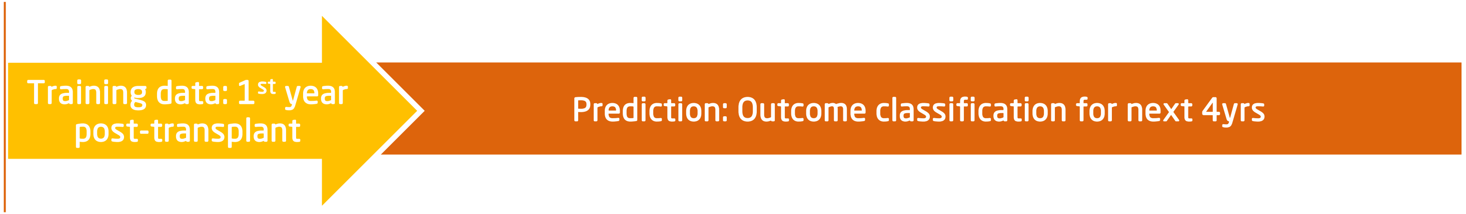
Feature Engineering for Machine Learning
Principles and Techniques for Data Scientists
Alice Zheng and Amanda Casari, O'Reilly, 2018

3 Data Preparation
- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

3. Data Preparation:
Feature Selection in NephroCAGE

- Receiver data: REC_SEX, ANONYM_DATE_BIRTH, AGE_TX, AGE_DIALYSIS,
- Donor data: DON_AGE, DON_SEX, DON_TYPE
- Organ data: CIT_HOUR, MMA, MMB, MMDR
- Lab data: CREATININ_MEAN, PROTEINURIA

t_0: Transplantation

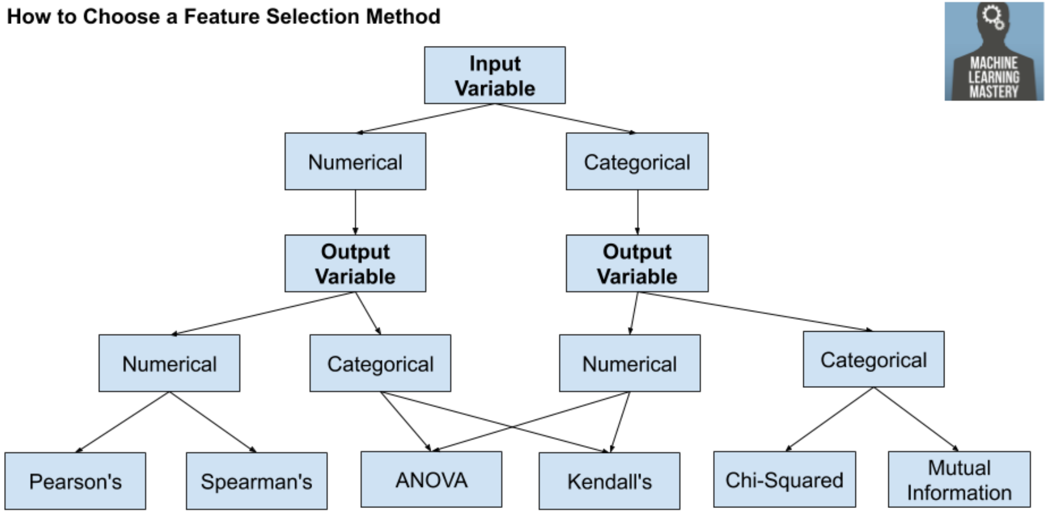


3 Data Preparation

- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

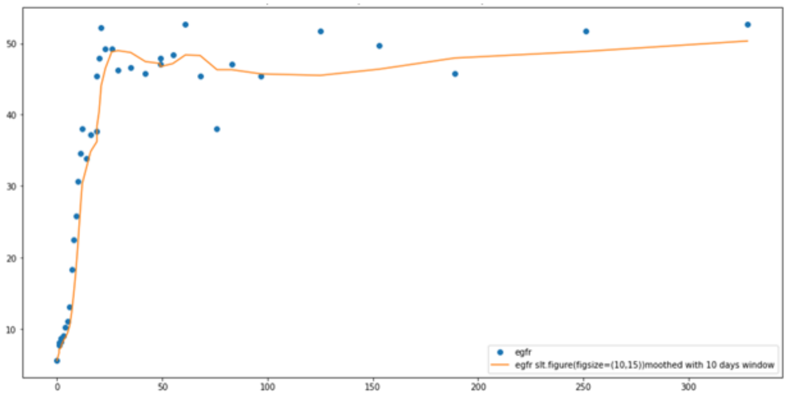
3. Data Preparation: Automatic Feature Selection

- Automatically select features that contribute most to the prediction
- Univariate feature selection: Statistical tests, helpful in Linear Models
- Recursive feature elimination: Use the model to eliminate features
- Tree-based feature selection: Elimination using feature importance, e.g. Boruta



3. Data Preparation: Example: EGFR in Nephrocage

- Incorporate domain expertise for feature selection
  - Baseline eGFR measured directly after surgery
  - Specific value per patient and transplant based on individual kidney function
- Variation in longitudinal measurement
  - Creatinine: Mean Creatinine, Variance in creatinine
  - Hospitalisation duration after surgery



3 Data Preparation

- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

NEPHRO CAGE

3. Data Preparation: Transformation of Categorical to Numerical Attributes

- Aim: Each attribute will have a value either 0 or 1
- Dummy variables encodes n categories through n dummy variables,
- Dummy variables with reference group represents n categroies through n-1 dummy variables
- Dummy variables for ordered categorical variable with reference group assumes logical ordering, e.g. S < M < L.

	X_0	X_1	X_2
Small	1	0	0
Medium	0	1	0
Large	0	0	1

	X_1	X_2
Small	0	0
Medium	1	0
Large	0	1

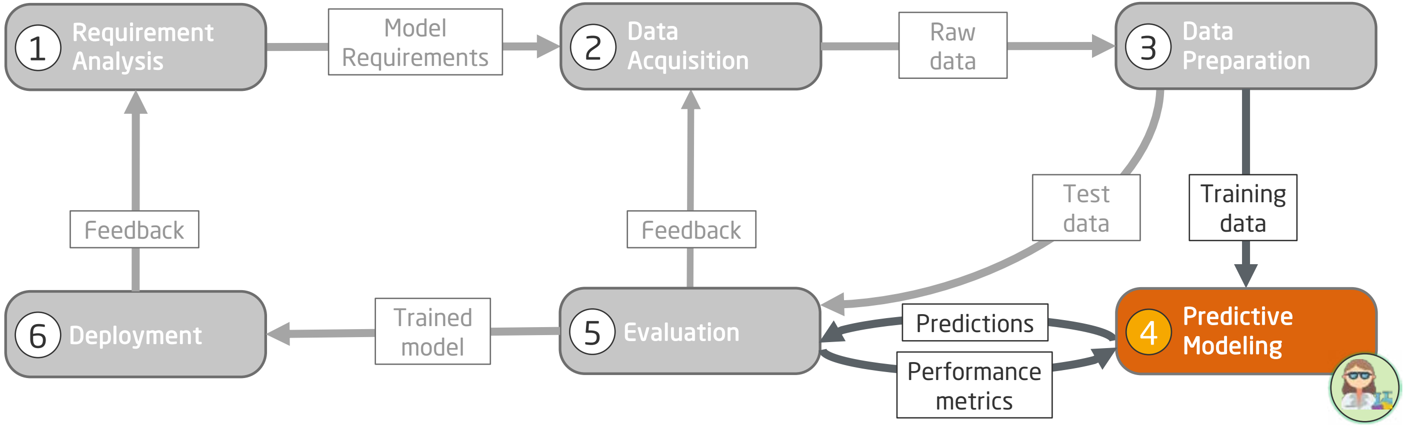
Reference Group (Small: 0 0)

	X_1	X_2
Small	0	0
Medium	1	0
Large	1	1

https://de.mathworks.com/help/stats/dummy-indicator-variables.html

3 Data Preparation
- Exploration
- Quality assessment
- Cleansing
- Feature engineering
- Labeling

4. Predictive Modeling



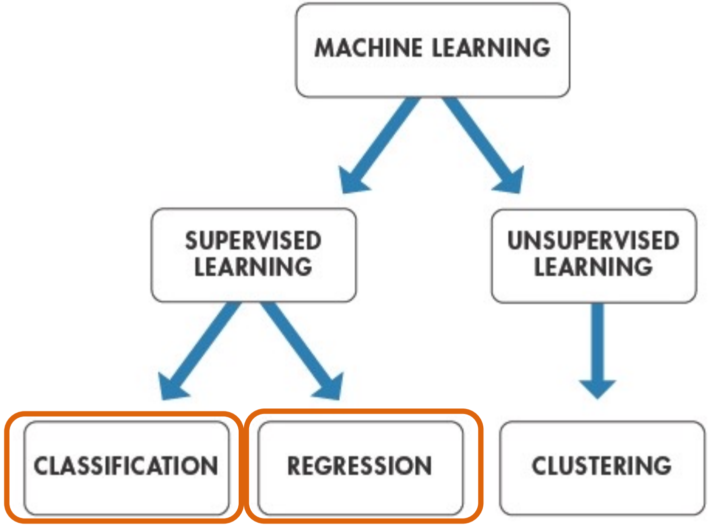
Roles

- Data Scientist
- Domain Expert
- (Data) Engineer

Icons made by Smashicons from www.flaticon.com

4. Predictive Modeling: Categories of Models

- Supervised learning
  - Labeled data is required
  - Categorical or numerical responses
  - Ex.: Decision trees, Bayesian nets, ridge regression
- Unsupervised learning
  - No data labels required
  - Performs pattern recognition
  - Ex.: Hierarchical clustering, k-means, etc.

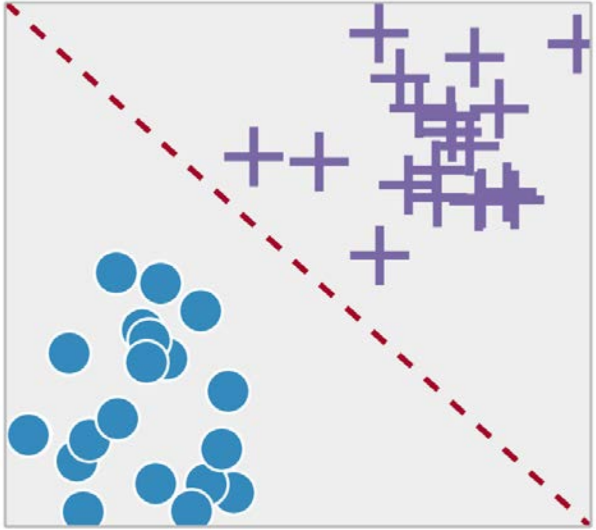


https://de.mathworks.com/help/stats/machine-learning-in-matlab.html

4. Predictive Modeling: Classification-based Models

Task: Find the best discriminants for known outcomes

- Binary class vs. multiple class classification
- Examples:
  - Logistic regression for prediction of stroke outcomes
  - Applying deep learning to diagnose cancer patients
  - Analyzing electrocardiograms to detect atrial fibrillation
  - Predict incidence of heart disease with life-style data
  - Probability for hospital re-admission



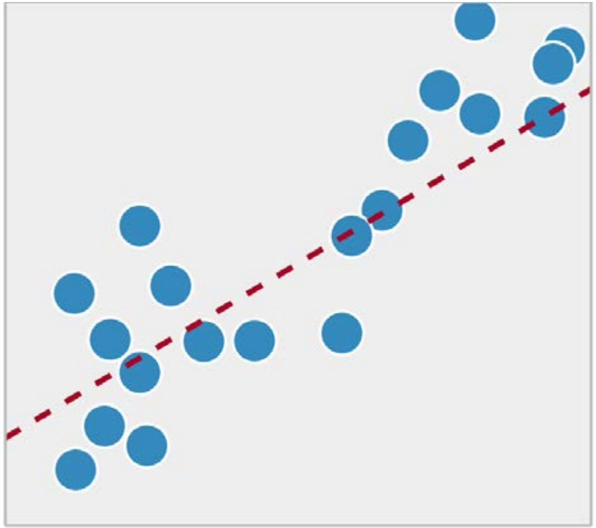
https://blog.statsbot.co/machine-learning-algorithms-183cc73197c

Cleophas, T. J., & Zwinderman, A. H. (2013). Machine Learning in Medicine (Vol. 53). Springer Berlin Heidelberg. http://doi.org/10.1017/CBO9781107415324.004

4. Predictive Modeling: Regression-based Models

Task: Fit the best curve to predict a continuous variable

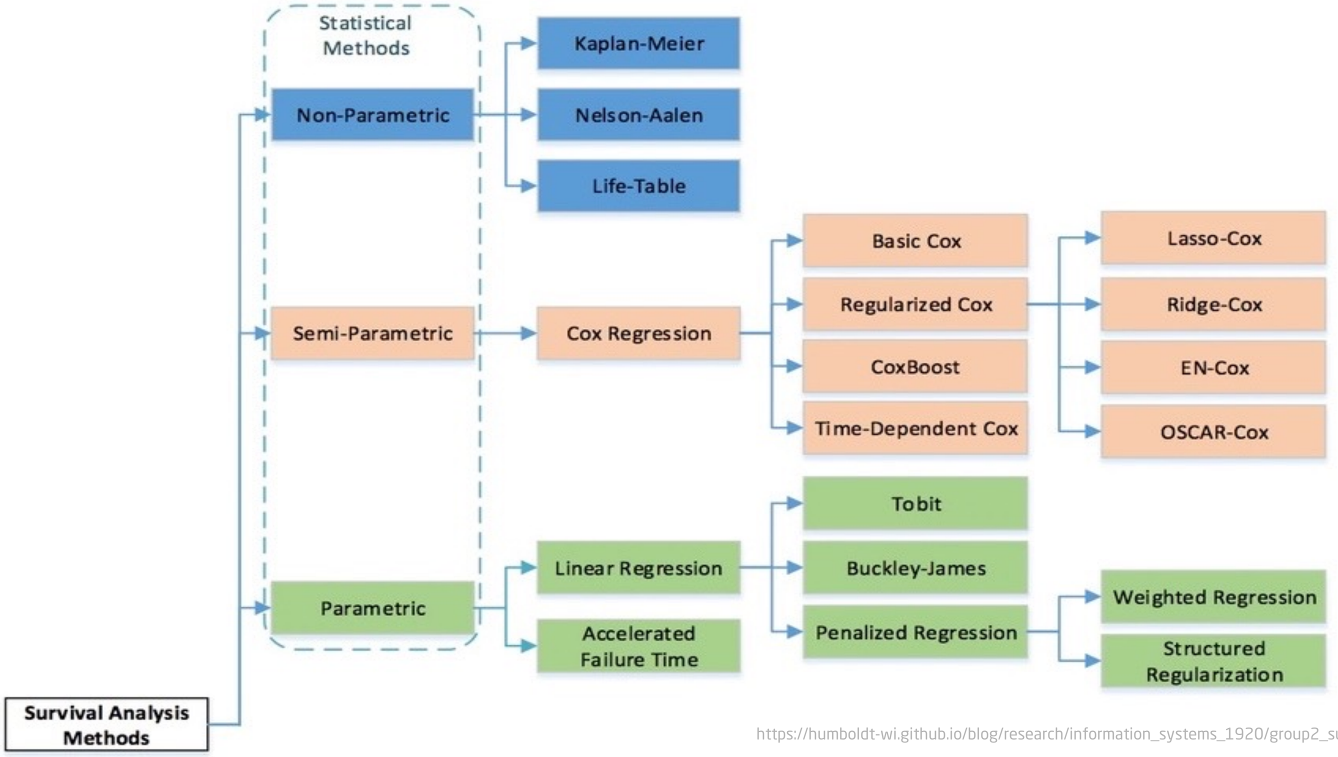
- Examples:
  - Predicting cancer survival time using a Cox model
  - Forecasting reduction of viral load after treatment using Support Vector Regression (SVR)
  - Predicting Length of Stay (LoS) of ICU patients using local polynomial regression Optimal drug dosage
  - Survival analysis / survival curve
  - Time-to-event prediction, e.g. cancer mortality
- Bear in mind: <u>Correlation does not imply causation</u>



https://blog.statsbot.co/machine-learning-algorithms-183cc73197c

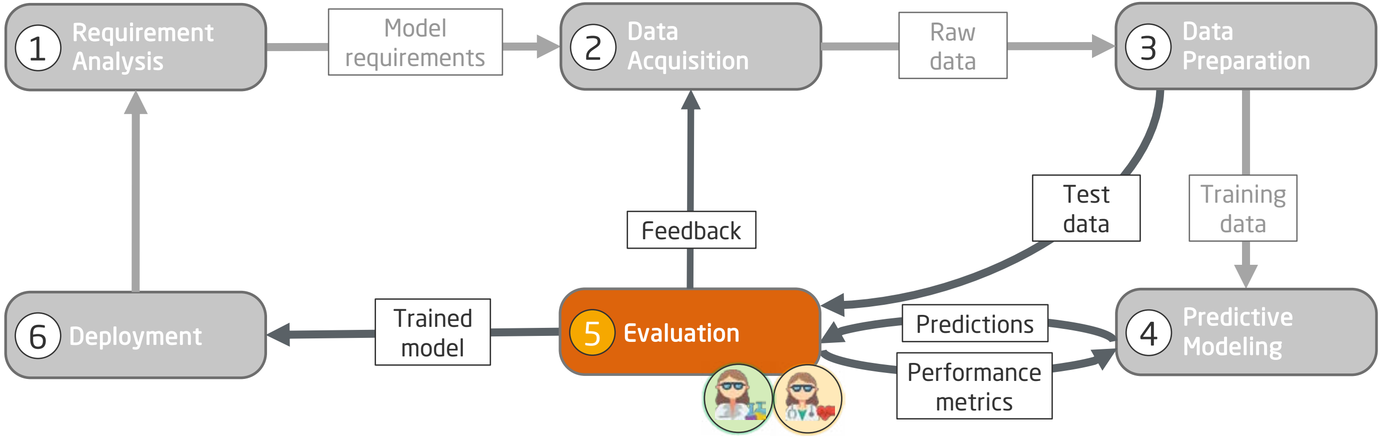
Cleophas, T. J., & Zwinderman, A. H. (2013). Machine Learning in Medicine (Vol. 53). Springer Berlin Heidelberg. http://doi.org/10.1017/CBO9781107415324.004

4. Predictive Modeling: Censored Data (cont'd)



https://humboldt-wi.github.io/blog/research/information_systems_1920/group2_survivalanalysis/

5. Evaluation



Roles

- Data Scientist
- Domain Expert
- (Data) Engineer

Icons made by Smashicons from www.flaticon.com

5. Evaluation
Measures of Performance

Measure
Sensitivity and specificity
Discrimination (ROC/AUC)
Predictive values: positive, negative
Likelihood ratio: positive, negative
Accuracy: Youden index, Brier score
Number needed to treat or screen
Calibration: Calibration plot, Hosmer-Lemeshow test
R^2 statistical significance: p-value (e.g. likelihood ratio test)
Magnitude of association, e.g., β coefficients, odds ratio
Model quality: Akeike IC/ Bayes IC
Net reclassification index and integrated discrimination improvement
Net benefit
Cost-effectiveness

Lee, Y.-H., Bang, H., & Kim, D. J. (2016). How to Establish Clinical Prediction Models. Endocrinology and Metabolism, 31(1), 38. http://doi.org/10.3803/EnM.2016.31.1.38

5. Evaluation
F1 score vs. MCC

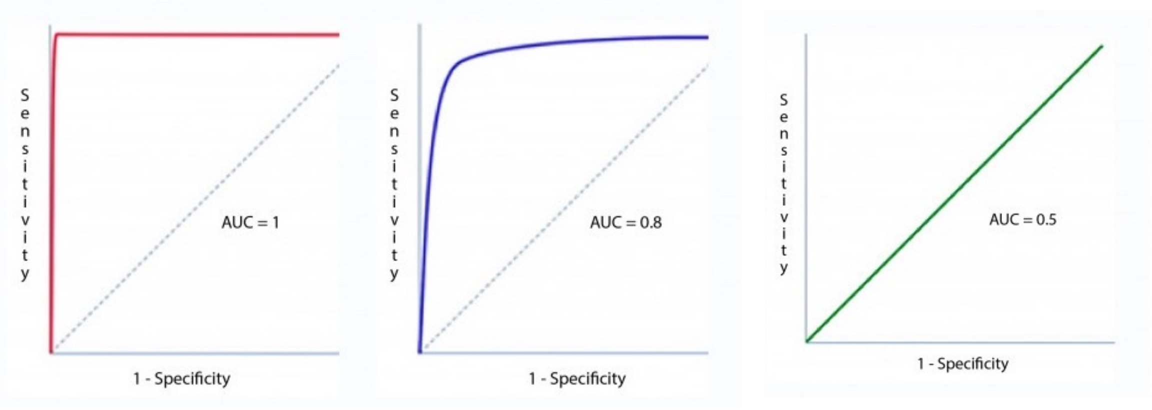
- F1 score combines precision and recall
- Disadvantages of using F1 score:
  - It is not normalized
  - It is not symmetric (when swapping positive and negative classes)
- Matthew's Correlation Coefficient (MCC) is normalized and symmetric

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

- Similarly interpretable as Pearson's correlation coefficient [-1,1], i.e.
  - 1 = perfect prediction
  - 0 = random prediction
  - -1 = negative prediction

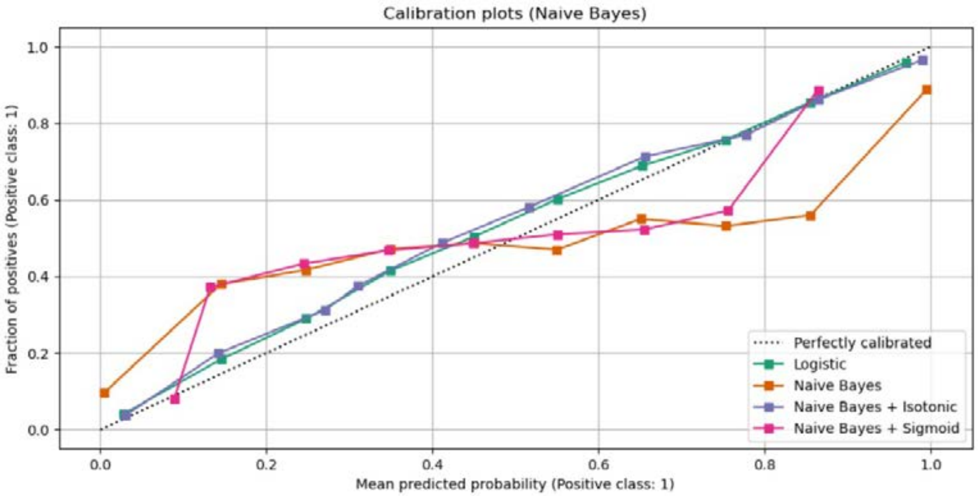
5. Evaluation
Recap: Receiver Operating Characteristic (ROC) Curve

- Allows comparison between classifiers (popular in CPMs)
- But: Not suitable for imbalanced classes (common for CPMs)
- F1 score and Matthew's Correlation Coefficient (MCC) are better suited



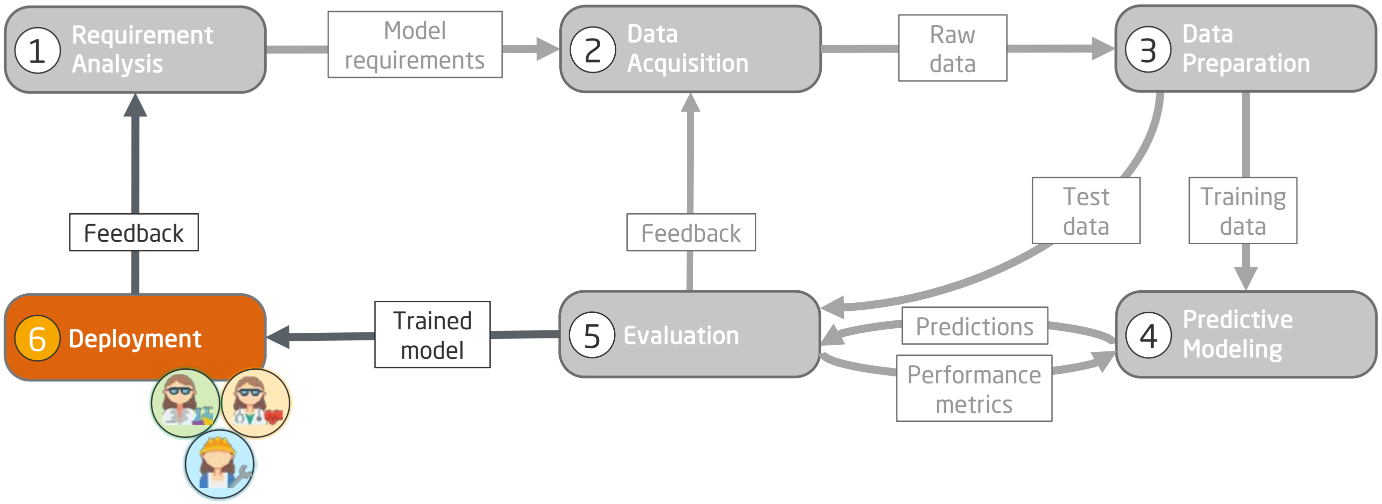
5. Evaluation
Calibration Plot

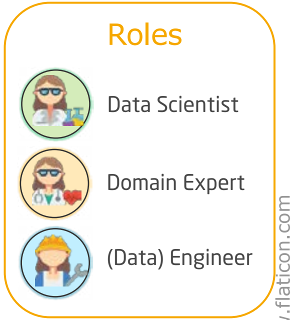
- X-axis: Mean predicted value
- Y-axis: Fraction of positive predictions
- Ideal calibrated model would be a straight line



https://scikit-learn.org/stable/auto_examples/calibration/plot_calibration_curve.html

6. Deployment

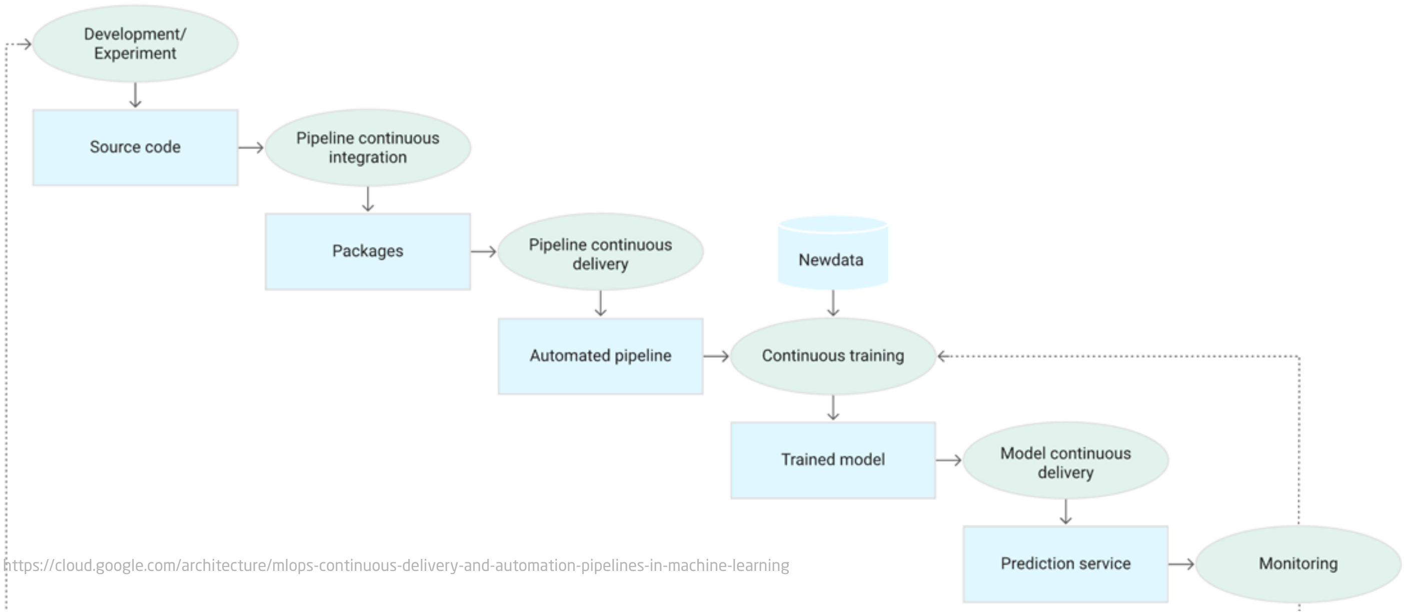




Icons made by Smashicons from www.flaticon.com

6. Deployment (cont'd)

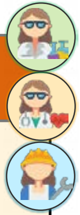
- Packaging for CPM models comparable to applications
- Meta data description required, e.g. input data definition, training data, etc.
- Continuous Integration and Continuous Delivery (CI/CD) of ML systems are achieved



https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

6 Deployment

- Model compression
- Model versioning
- Process integration
- Monitoring
- Continual learning



6. Deployment: Monitoring and Continual Learning

- CPM needs to be monitored for stability metrics in data, model performance metric, and software development operations metrics.
- Data Shift Metric: Helps identify various shifts in data distribution between the training data and production data.
- Continual Learning mitigates data shift
  - Incremental learning: Learn frequently without losing old model
  - Model retraining: Retrain on new data
  - Online learning: Continuously improve model through new real-world data

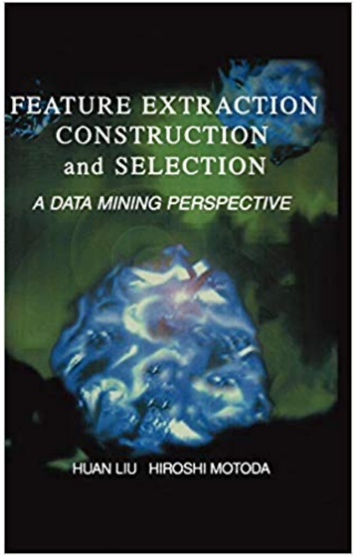
6 Deployment

- Model compression
- Model versioning
- Process integration
- Monitoring
- Continual learning

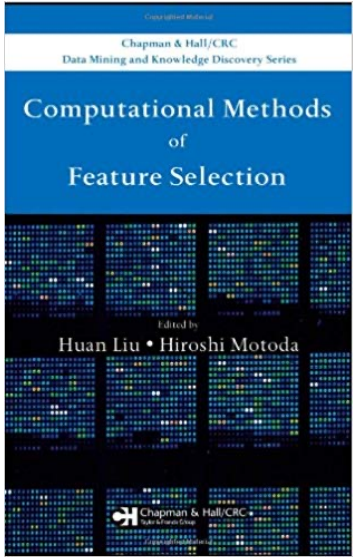
What to Take Home?

- Many steps could be automated with a software program but some steps need domain experts.
- Data acquisition: Consider what data is available, what data is missing and what data can be removed.
- Data preparation: Organize your data by formatting, cleaning and sampling from it.
- Data transformation: Identify relevant features for CPM development
- Model evaluation: performance metrics need to be defined prior development, e.g. use of F1 score vs. MCC on imbalanced data
- Deployment and monitoring of CPMS is crucial for clinical use

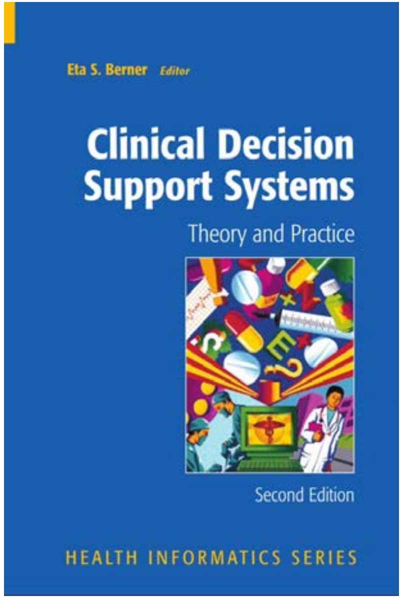
To Know More

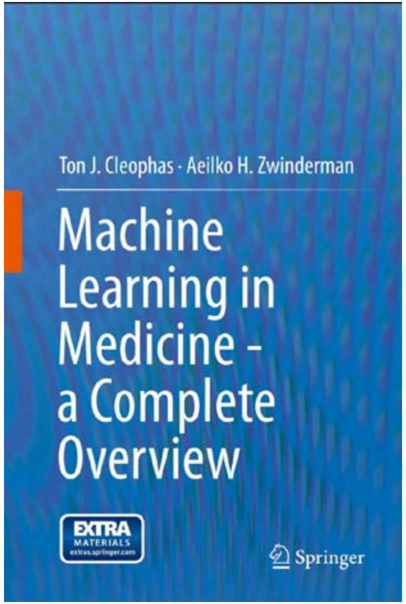

FEATURE EXTRACTION
CONSTRUCTION
and SELECTION
A DATA MINING PERSPECTIVE
HUAN LIU HIROSHI MOTODA

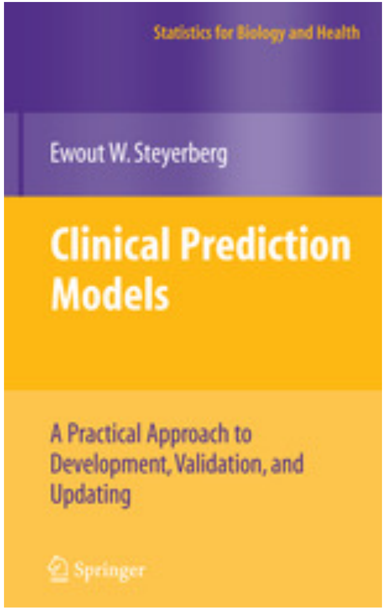

STUDIES IN FUZZINESS AND SOFT COMPUTING
Studies in Fuzziness and Soft Computing
Isabelle Guyon · Steve Gunn
Massoud Nikravesh · Lotfi A. Zadeh
Editors
Feature Extraction
Foundations and Applications
Springer
Extra Materials


Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series
Computational Methods
of
Feature Selection
Edited by
Huan Liu • Hiroshi Motoda
Chapman & Hall/CRC

To Know More


Eta S. Berner Editor
Clinical Decision Support Systems
Theory and Practice
Second Edition
HEALTH INFORMATICS SERIES


Ton J. Cleophas · Aeilko H. Zwinderman
Machine Learning in Medicine - a Complete Overview
EXTRA MATERIALS
Springer


Statistics for Biology and Health
Ewout W. Steyerberg
Clinical Prediction Models
A Practical Approach to Development, Validation, and Updating
Springer