Iterative Machine Learning Design Process for Digital Health

Borchert, Dr. Schapranow

Data Management for Digital Health

Winter 2023

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

ML Process

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

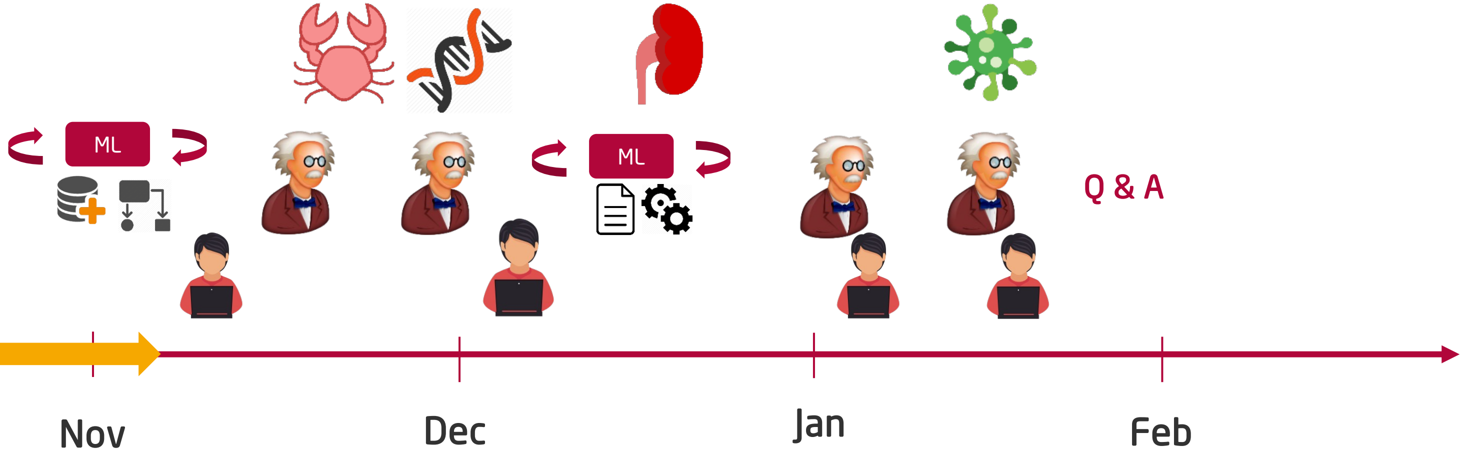
ML

Refine

Evaluate

Prediction + Probability

Lecture Schedule



- Lecture Kickoff
- Actors in Healthcare
- Digital Health Data

- Machine Learning (ML) Foundations
- Use Case Oncology
- Biology Recap

- Natural Language Processing
- Use Case Nephrology & Intensive Care
- Supervised ML & Deep Learning

- Use Case Infectious Diseases
- Unsupervised ML

Agenda

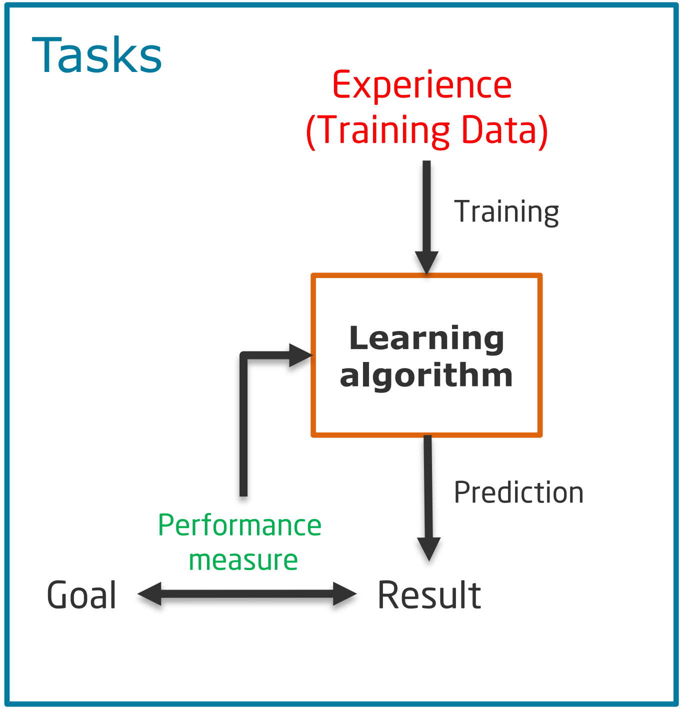
- Pitfalls of ML Systems
- ML Process Model
- Federated Learning

Using Data to Train Machines

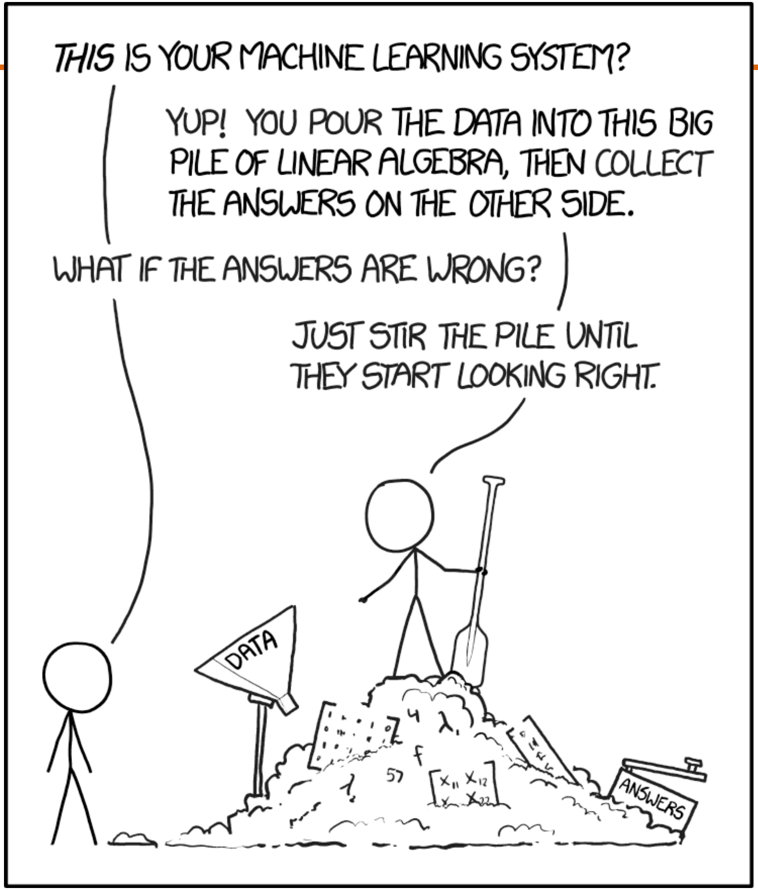
Machine Learning

"Field of study that gives computers the ability to learn without being explicitly programmed"
(Arthur Samuel, 1959)

"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."
(Tom Mitchell, 1997)



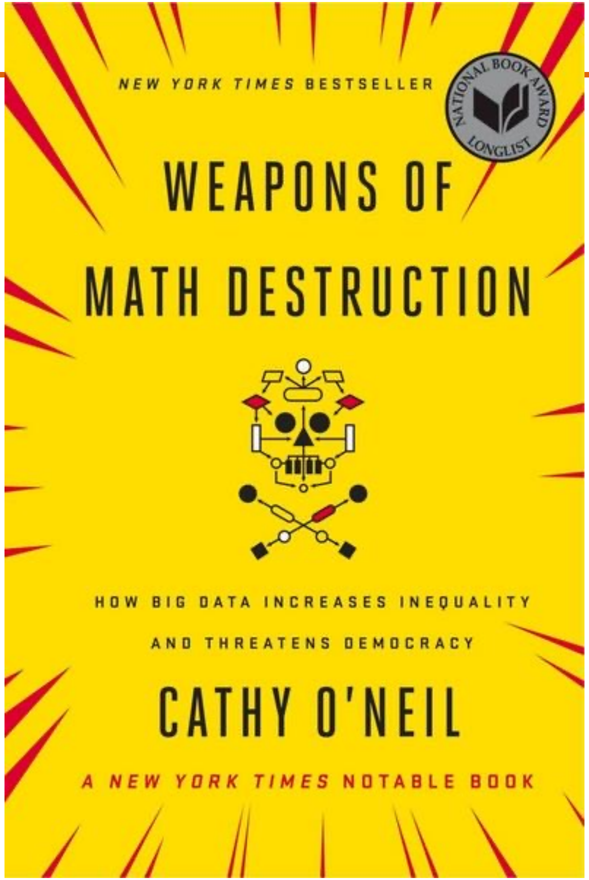
What Could Possibly Go Wrong?



https://xkcd.com/1838/

Garbage In – Garbage Out

- ML models are only as good as the data they are trained on
- When learning from data to make decisions, we may reproduce biases (racial, gender, etc.) that are inherent in the data, e.g.:
  - African-American prisoners could be predicted to be more likely to commit future crimes because they are overrepresented in prisons
  - Women could be predicted to have less a priori risk of heart attacks because they are currently more often misdiagnosed



Some Domains are Really Hard

02 Apr 2019 | 15:00 GMT

How IBM Watson Overpromised and Underdelivered on AI Health Care

After its triumph on Jeopardy!, IBM's AI seemed poised to revolutionize medicine. Doctors are still waiting

"A deeper look [...] reveals a fundamental **mismatch between the promise of machine learning and the reality of medical care** – between "real AI" and the requirements of a functional product for today's doctors."

Good Performance != Clinical Utility

nature machine intelligence

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature machine intelligence › analyses › article

Analysis | Open access | Published: 15 March 2021

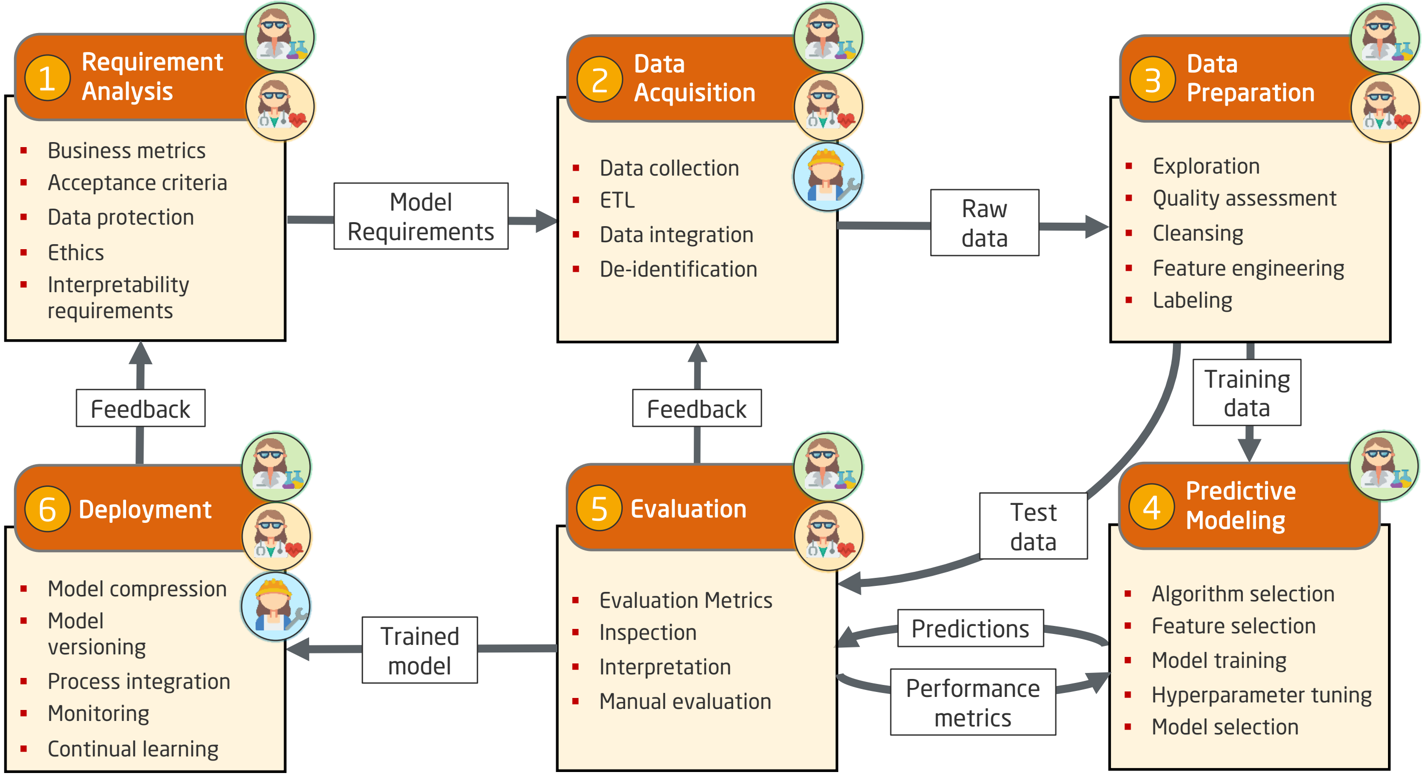
Common pitfalls and recommendations for machine learning to detect and prognosticate COVID-19 using chest radiographs and CT scans

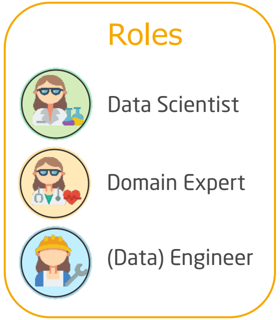
Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala & Carola-Bibiane Schönlieb

Nature Machine Intelligence **3**, 199–217 (2021) | Cite this article

in EMBASE and MEDLINE in this timeframe are considered. Our search identified 2,212 studies, of which 415 were included after initial screening and, after quality screening, 62 studies were included in this systematic review. Our review finds that none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases. This is a major weakness, given the urgency with which validated COVID-19 models are needed. To address this, we give many recommendations which, if followed, will solve these issues and lead to higher-quality model development and well-documented manuscripts.

Process Model for ML in Digital Health



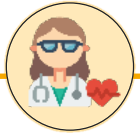


ML Process

Roles and Personas in the ML Process

Data Scientist

- Strong knowledge in statistical modelling
- Builds hypotheses and conducts experiments



Domain Expert

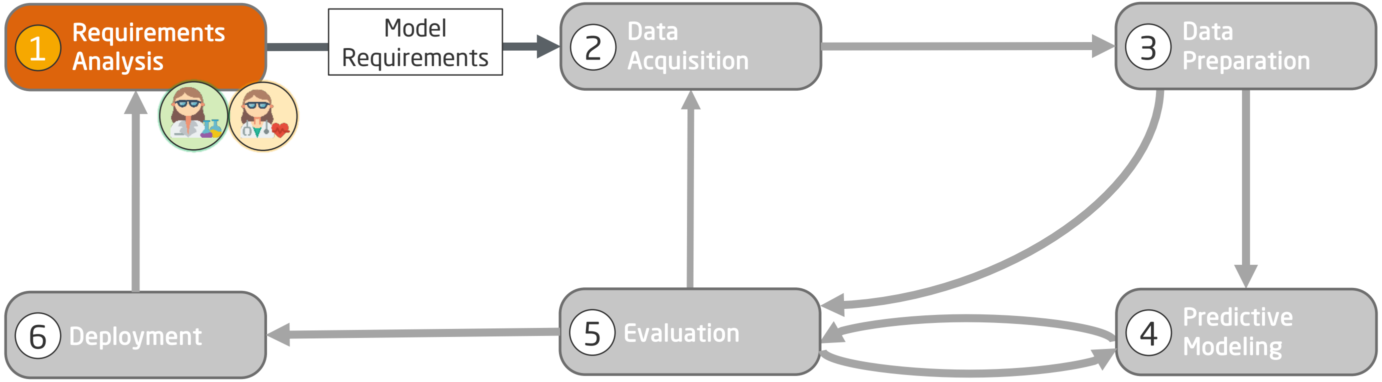
- Strong knowledge in the application domain, e.g., medicine or biology
- Often the end user of the final AI-based system
- Provides ground truth for supervised models

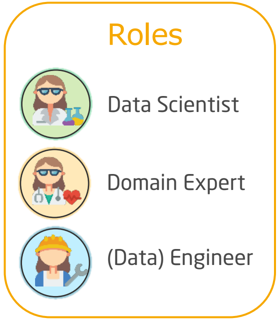
(Data) Engineer

- Strong knowledge in databases, software engineering and ETL processes
- Builds reusable infrastructure
- Optimizes system performance

Note: Individuals can have more than one role in a given project and the distinctions are sometimes blurry

Requirements Analysis



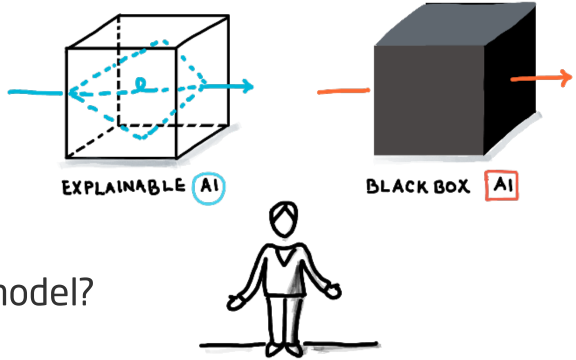


Running Example

Predict risk of coronary heart disease!

Requirements Analysis

- Define business metrics / outcomes, e.g.:
  - Mortality
  - False detection rate
  - Cost for hospital stays
  - ➔ Conflicting goals must be balanced
- Define the baseline, e.g.:
  - existing (clinical) algorithm
  - human performance
- Define acceptance criteria for an ML solution
  - What performance is considered "good enough"?
  - Interpretability: will your users trust a black box model?



http://kiskipby.org/group-meeting/
https://medium.com/ritual-design/culture-meets-artificial-intelligence-a2ad6dc82bb7

1 Requirements Analysis

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

Requirements Analysis
How to get data and labels?

- Ethics: you might need IRB approval for your ML project!

- Data protection: to work with personal health data, you need consent (or data donation) or use de-identified data

- Obtaining data and ground-truth labels comes at a cost
  - Ask yourself: do you really need ML to solve the problem?

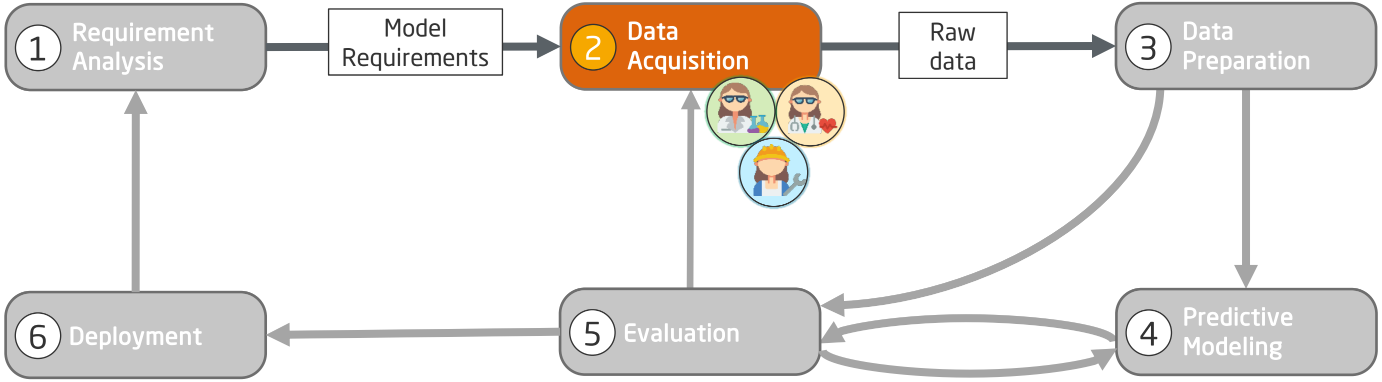
$

Hasso Plattner Institut

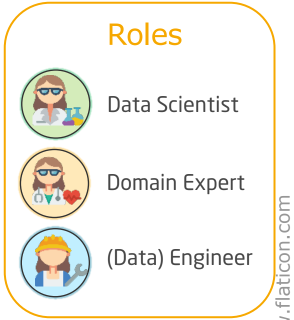
1 Requirements Analysis

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

Data Acquisition



Hasso Plattner Institut



Icons made by Smashicons from www.flaticon.com

Data Acquisition
Data collection & ETL

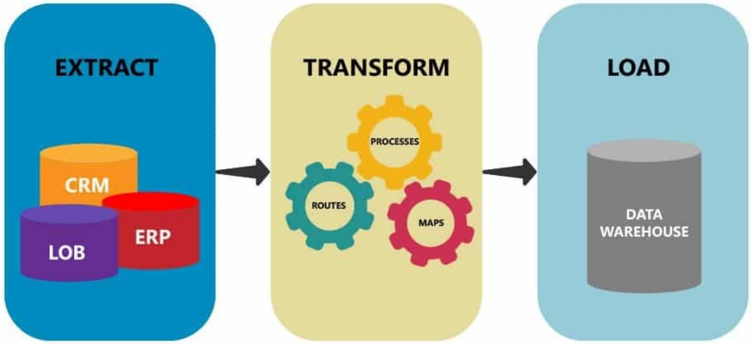
Hasso Plattner Institut

- Data in Digital Health is often scattered across multiple (legacy) systems
- Obtaining additional data:
  - Existing datasets (internal or open source) can be helpful
  - External knowledge bases are available for many Digital Health applications
- Extract, Transform, Load (ETL) processes have to be implemented to join different data sources

ID	var1	var2	var3
588	2	d	1
654	1	y	1
527	1	o	0
955	2	c	0
954	1	t	0

ID	var1	var2	var3
588	290	Apples	Breakfast
654	81	Bananas	Snack
527	63	Apples	Snack
955	6	Pears	Snack
954	146	Pears	Breakfast

ID	var1	var2	var3	var4	var5	var6
588	2	d	1	225	Apples	Breakfast
654	1	y	1	56	Bananas	Snack
527	1	o	0	245	Apples	Snack
955	2	c	0	46	Pears	Snack
954	1	t	0	121	Pears	Breakfast



2 Data Acquisition

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

https://www.displayr.com/what-is-data-merging/
https://www.astera.com/de/Typ/Blog/etl-prozess-und-schritte/

Example
Collecting Data From Different Data Sources

Patient ID	Age	Systolic Blood Pressure
43	15	120
71	15	118
293	16	101
433	16	108
280	16	114
...	...	...
408	61	180
19	62	158
395	64	180
204	64	128
226	15330	143

Dataset from GP

+

Patient ID	Coronary Heart Disease
293	False
80	False
433	False
410	False
280	False
...	...
395	False
408	True
316	False
459	False
399	True

Dataset from Health Insurance Company

→

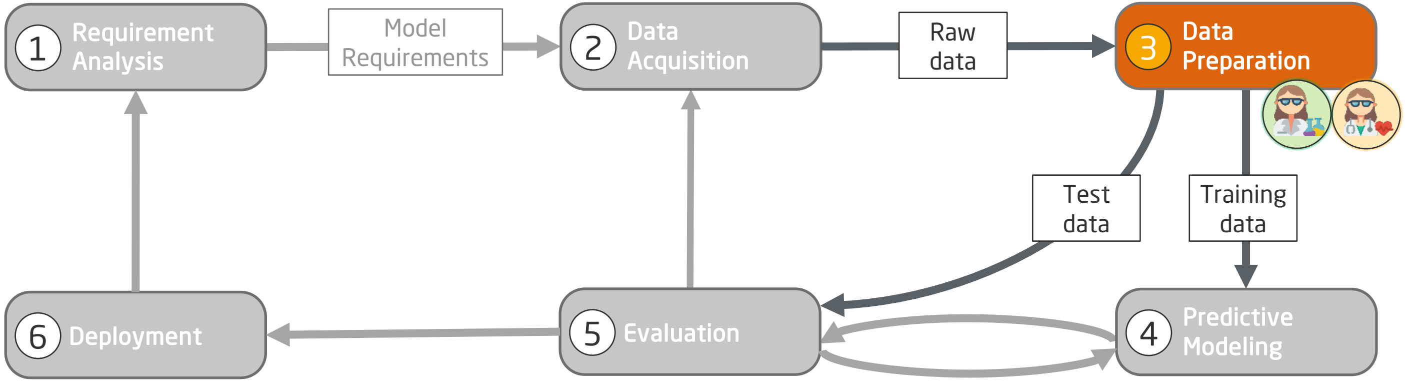
Patient ID	Age	Systolic Blood Pressure	Coronary Heart Disease
293	16	101	False
80	39	108	False
433	16	108	False
410	40	112	False
280	16	114	False
...	...	...	...
395	64	180	False
408	61	180	True
316	50	190	False
459	58	214	False
399	48	218	True

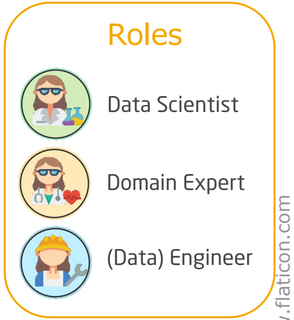
Merged dataset

2 Data Acquisition

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

Data Preparation





Data Preparation
Labeling

- (Ground truth) labels required for supervised ML
- Sometimes, we can just define one attribute as the label
- But: some labels can be very expensive to obtain, e.g., if they require manual annotation
- Labels can take different shape depending on the setting:
  - Category (classification)
  - Scalar value (regression)
  - Vectors / matrices (structured prediction)

	Patient ID	Age	Systolic Blood Pressure	Coronary Heart Disease
291	293	16	101	False
79	80	39	108	False
431	433	16	108	False
408	410	40	112	False
278	280	16	114	False
...	...	...	...	...
393	395	64	180	False
406	408	61	180	True
314	316	50	190	False
457	459	58	214	False
397	399	48	218	True

„Coronary Heart Disease“ (0/1) = Binary label (classification problem)

→ Two features remain: Age & Systolic Blood Pressure

Why is Patient ID not a good feature?

3 Data Preparation

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

Data Preparation
Exploration & Quality Assessment

- Data preparation can take up most of the time of ML project
- Explore dataset and **look** at single examples
  - Get descriptive statistics (e.g., number of samples, class imbalance, mean, variance, etc.)
  - Visualize dataset (can also be done using unsupervised ML methods for high-dimensional data)
  - Assess data quality (missing data, input errors, outliers, labeling errors, etc.)

	Patient ID	Age	Systolic Blood Pressure	Coronary Heart Disease
291	293	16	101	False
79	80	39	108	False
431	433	16	108	False
408	410	40	112	False
278	280	16	114	False
...	...	...	...	...
393	395	64	180	False
406	408	61	180	True
314	316	50	190	False
457	459	58	214	False
397	399	48	218	True

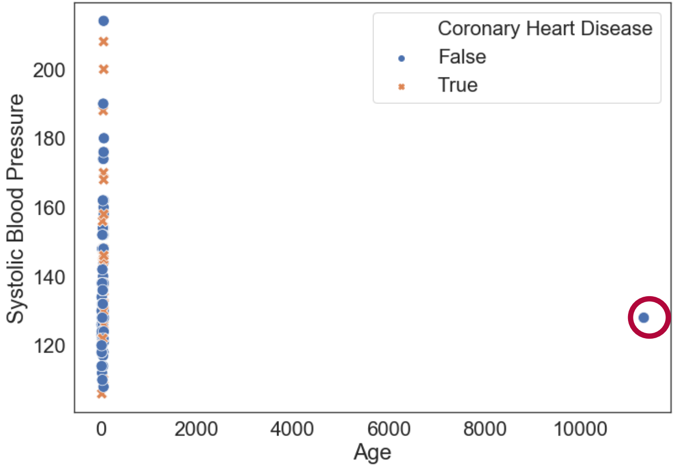
	Age	Systolic Blood Pressure
count	100.00	100.00
mean	157.54	140.65
std	1,127.12	22.27
min	16.00	106.00
25%	32.00	124.00
50%	49.50	136.00
75%	58.25	154.00
max	11,315.00	214.00

Hasso Plattner Institut

3 Data Preparation

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

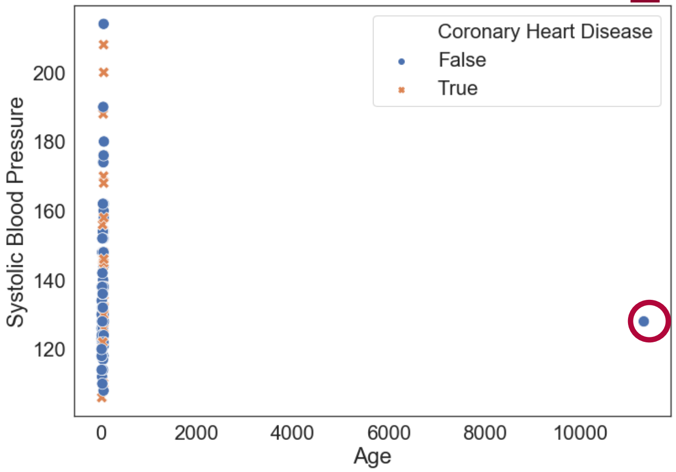
Data Preparation
Cleansing

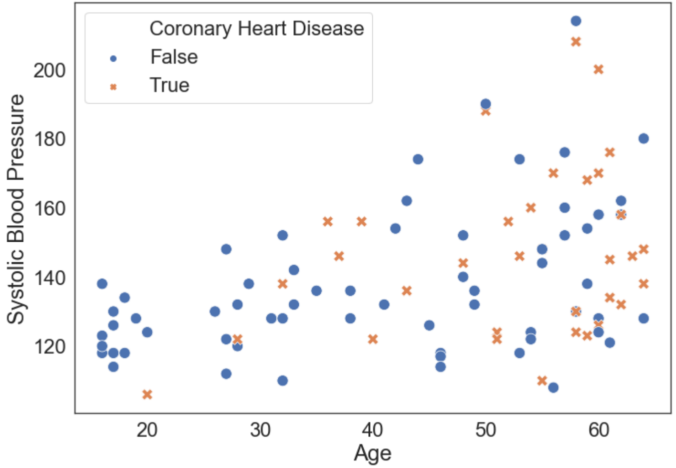


	Patient ID	Age	Systolic Blood Pressure
103	104	11315	128

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

Data Preparation
Cleansing





Data entry error?

$11315 = 31 \cdot 365$

3 Data Preparation

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

Data Preparation
Feature Engineering

- What are the relevant features that describe the data points?

→ Requires domain knowledge, takes a lot of time and used to be one of the defining success factors for ML projects (less important with Deep Learning)

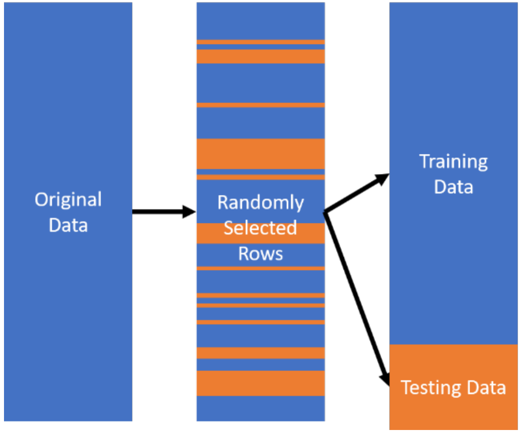
- Transformations are applied to the data to get it into a form to use it for modelling. Common operations:
  - Normalization, Scaling, e.g.: $\log(Age)$
  - Non-linear interactions, e.g.: Age^2 or $Age \cdot Syst.BP$
- Come up with reasonable estimates for missing values (Imputation) or remove them

	Patient ID	Age	Systolic Blood Pressure
291	293	16	101
79	80	39	108
431	433	16	108
408	410	40	112
278	280	16	114
...	...	...	...
393	395	64	180
406	408	61	180
314	316	50	190
457	459	58	214
397	399	48	218

3 Data Preparation

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

Training & Test Sets



- Part of the data needs be set aside as a proxy for "unseen data"

  ➔ Never look at the labels in your test set while training or you will overestimate performance

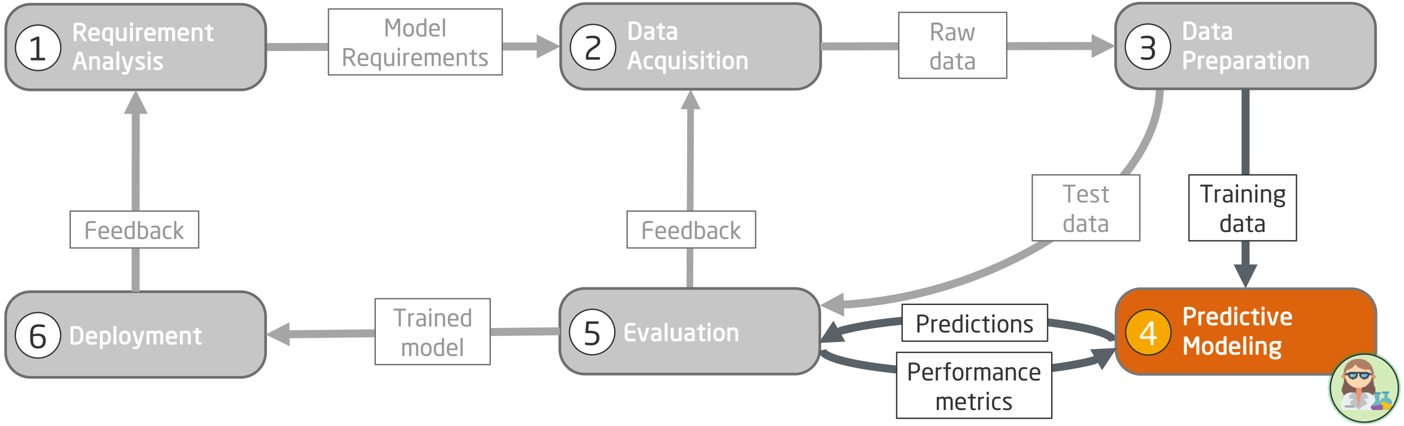
  - Common approach to get a test set: pick x% of training data randomly (e.g., 90% training, 10% test)
  - Use stratified sampling to get same distribution in training and test set
- Beware of data leakage (using information that is not available in production), e.g. :
  - use information from future to predict the past (e.g., using treatment to predict disease)
  - have data from same patient in training and test set

https://conlanscientific.com/posts/category/blog/post/avoiding-data-leakage-machine-learning/
https://www.jerseywater.je/water-services/get-connected/leakage/

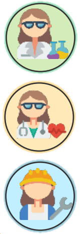
4 Predictive Modeling

- Algorithm selection
- Feature selection
- Model training
- Hyperparameter tuning
- Model selection

Predictive Modeling



Roles



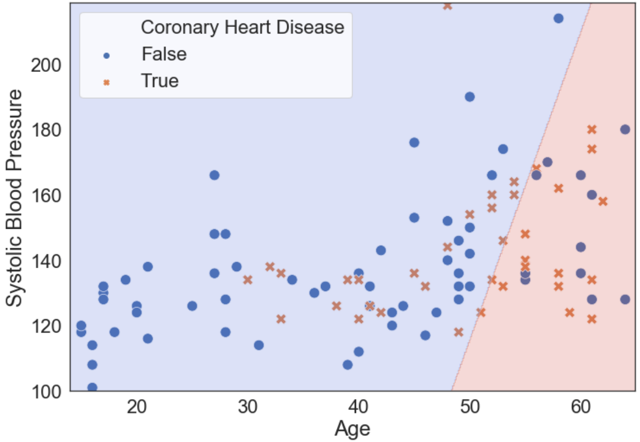
Data Scientist
Domain Expert
(Data) Engineer

Predictive Modelling
Simple Binary Classification Problem

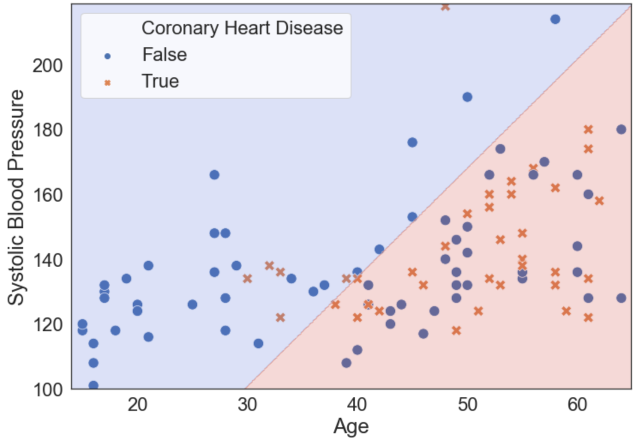
Age	Systolic Blood Pressure	Coronary Heart Disease
17	118	0
46	117	0
53	146	1
62	158	1
20	106	1
...	...	...
20	124	0
48	144	1
42	154	0
51	124	1
58	214	0

Recall Goal in ML:

Find function (model of the data) $f: X \rightarrow Y$



Decision boundary for CHD classification



Another decision boundary for CHD classification

4 Predictive Modeling

- Algorithm selection
- Feature selection
- Model training
- Hyperparameter tuning
- Model selection

Predictive Modelling

- Highly experimental and iterative process
- Can require lots of computational resources depending on:
  - size of dataset
  - complexity of model
  - hyperparameter search space
- Choosing a relevant subset of features (Feature selection) can improve speed, predictive performance and interpretability
- Details strongly depend on problem setting (supervised / unsupervised) and learning algorithm ➔ will be covered in other lectures

4 Predictive Modeling

- Algorithm selection
- Feature selection
- Model training
- Hyperparameter tuning
- Model selection

Simple Binary Classification: Logistic Regression

	Age (x_1)	Systolic Blood Pressure (x_2)	Coronary Heart Disease (y)
$x^{(1)}$	17	118	0
$x^{(2)}$	46	117	0
$x^{(3)}$	53	146	1
$x^{(4)}$	62	158	1
$x^{(5)}$	20	106	1
.	...	...	...
.	20	124	0
.	48	144	1
.	42	154	0
	51	124	1
$x^{(n)}$	58	214	0

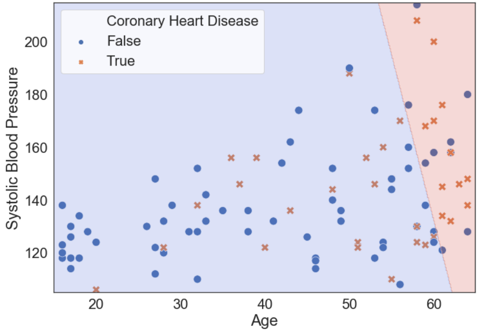
Model $P\left(y^{(i)} = 1 \,\middle|\, \boldsymbol{x}_i\right) = \sigma(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b)$

$= \sigma(w_1 \; x_1^{(i)} + w_2 \; x_2^{(i)} + b)$

Features $\boldsymbol{x}_i = (x_i^{(1)}, x_i^{(2)})$

Weights $\boldsymbol{w} = (w_1 \,, w_2 \,)$, bias b

Logistic function $\sigma(z) = \frac{1}{1+e^{-z}}$



→ Result of running logistic regression:

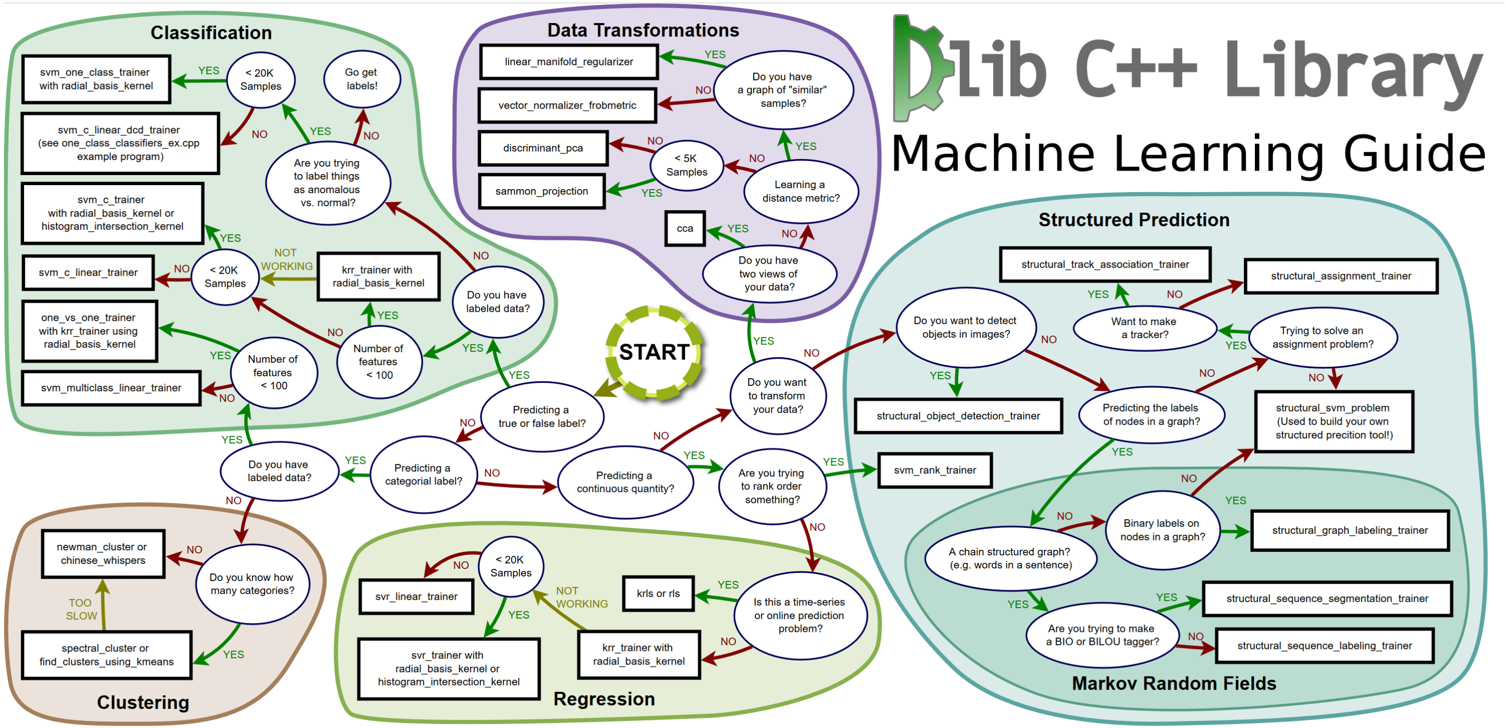
$w_1 \;= 0.057 \quad w_2 \;= 0.0044 \quad b = -3.857$

$P(y = 1 \mid \boldsymbol{x}^{(1)}) \;= \sigma(0.057 \cdot 17 + 0.0044 \cdot 118 \;- 3.857) = \mathbf{0.08}$

$P(y = 1 \mid \boldsymbol{x}^{(4)}) \;= \sigma(0.057 \cdot 62 + 0.0044 \cdot 158 \;- 3.857) = \mathbf{0.56}$

Predictive Modelling
Selecting the appropriate model / ML algorithm

Hasso Plattner Institut

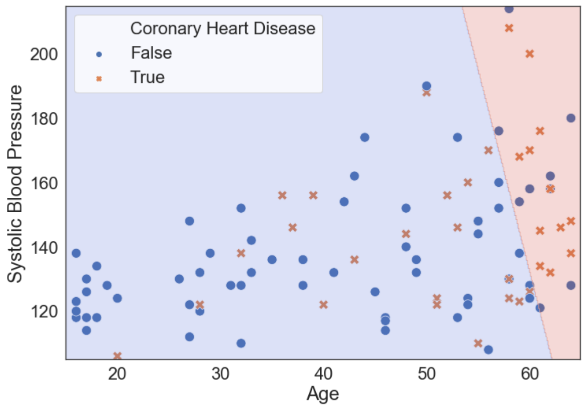


4 Predictive Modeling

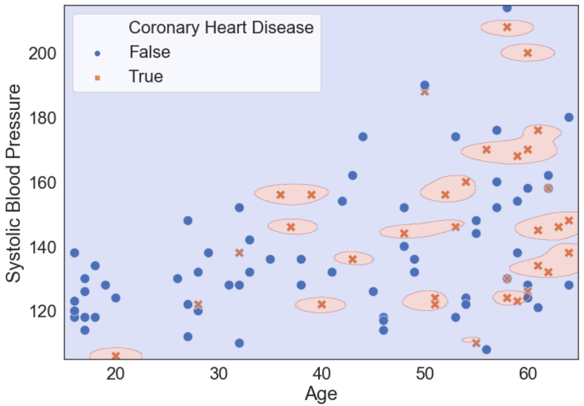
- Algorithm selection
- Feature selection
- Model training
- Hyperparameter tuning
- Model selection

Predictive Modelling
Model Selection

Linear model might be too simple → just use a more complex model?



A



B

Poll: Which model would you prefer?

Predictive Modelling
Model Selection

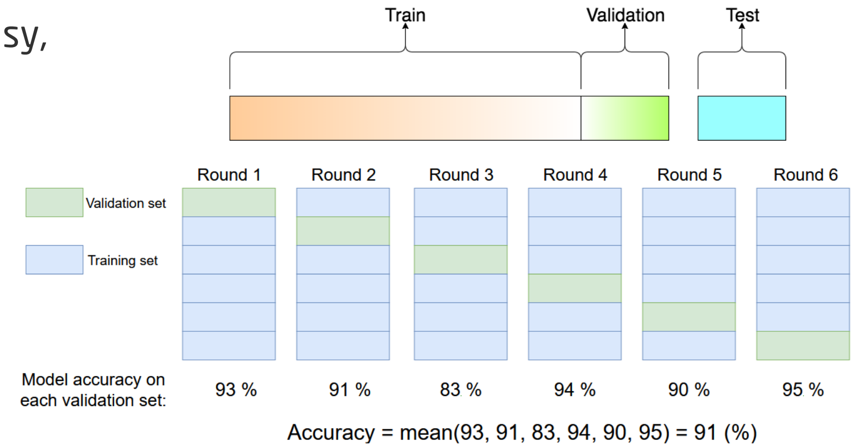
- Getting good results on training data is relatively easy, especially on small training datasets
- Recall goal in ML: make predictions for unseen data
- Model selection (picking one f^* out of your class of functions) should not be done on the training set

4 Predictive Modeling

- Algorithm selection
- Feature selection
- Model training
- Hyperparameter tuning
- Model selection

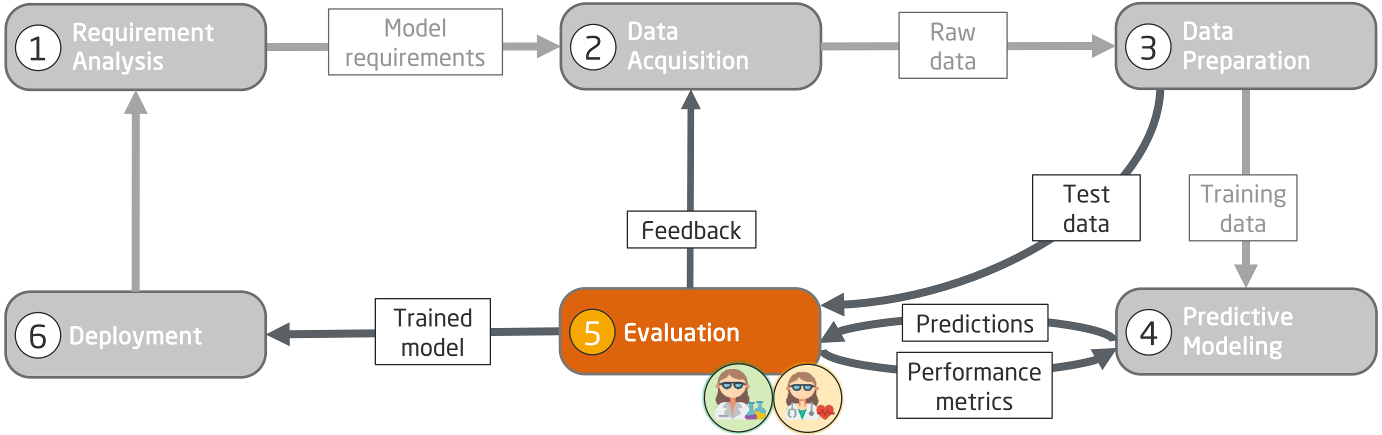
Predictive Modelling
Model Selection

- Getting good results on training data is relatively easy, especially on small training datasets
- Recall goal in ML: make predictions for unseen data
- Model selection (picking one f^* out of your class of functions) should not be done on the training set
- Common approach: split dataset
  - Training set used for model training
  - Validation set used for model selection
  - (Hold-out) Test set used to report performance estimate
- If you have little data: n-fold cross-validation (care must be taken not to overestimate performance)



https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7
https://www.machinelearningtutorial.net/2017/04/01/training-set-vs-test-set-vs-validation-set-whats-the-deal/

Evaluation



Roles
- Data Scientist
- Domain Expert
- (Data) Engineer

Icons made by Smashicons from www.flaticon.com

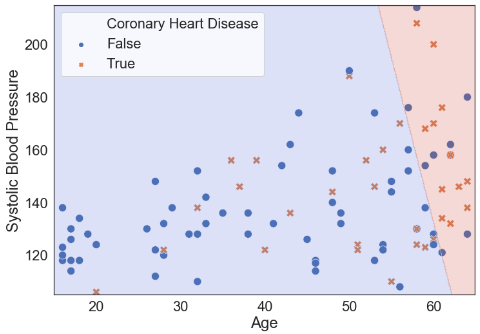
Evaluation

- Different evaluation metrics for different ML settings:
  - Classification: Accuracy, Precision, Recall, …
  - Regression: Root Mean Squared Error, Max-Error, …
  - Structured Prediction: domain specific
- Choosing right metric is not trivial and should be aligned with “business” goals
- Statistical significance tests should be performed to determine if model is really better or just by chance

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

Evaluation
Model Inspection

	x_1	x_2	y
	Age	Systolic Blood Pressure	Coronary Heart Disease
$x^{(1)}$	17	118	0
$x^{(2)}$	46	117	0
$x^{(3)}$	53	146	1
$x^{(4)}$	62	158	1
$x^{(5)}$	20	106	1
.	...	...	...
.	20	124	0
.	48	144	1
.	42	154	0
	51	124	1
$x^{(n)}$	58	214	0



→ Result of running logistic regression:

$$w_1 = 0.057 \quad w_2 = 0.0044 \quad b = -3.857$$

Interpretation:

Risk of CHD increases with age and also with SBP, but more strongly with age

→ Just inspecting the weights usually not so easy for more complex models

5 Evaluation

- Evaluation Metrics
- Inspection
- Interpretation
- Manual evaluation

Evaluation
Manual Evaluation & Interpretation

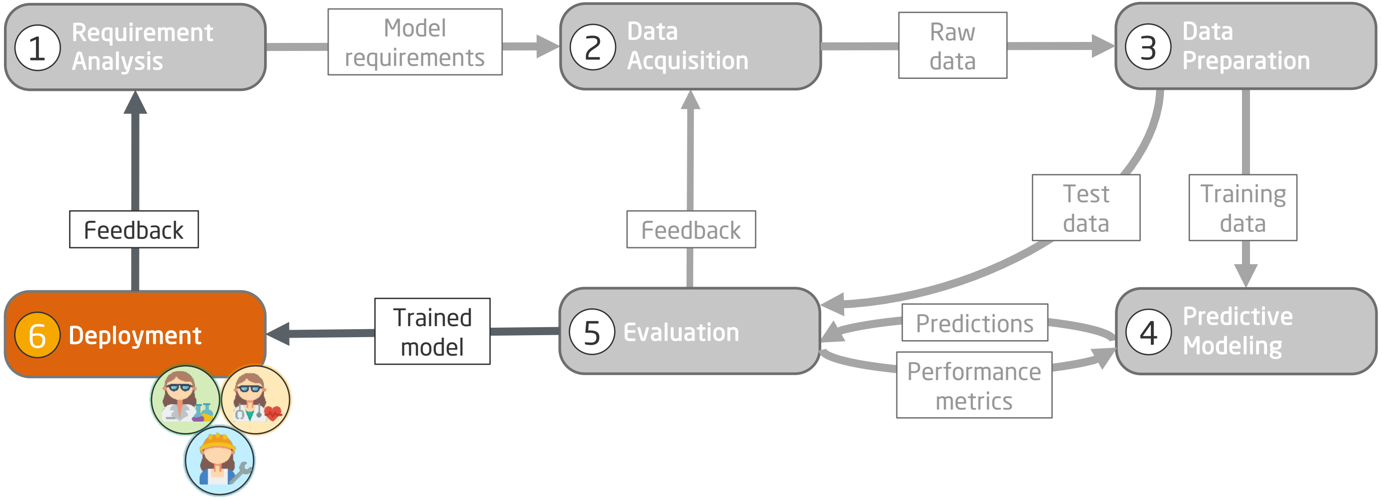
(a) Original image (b) Grad-CAM for biased model (c) Grad-CAM for unbiased model

5 Evaluation

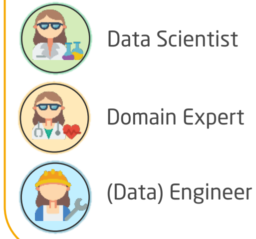
- Evaluation Metrics
- Inspection
- Interpretation
- Manual evaluation

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

Deployment



Roles



Deployment

- Integration into existing applications & processes
- Simplification / compression of models to run on target hardware might cost some accuracy
- Target hardware includes:
  - Web servers
  - Smart watches / Mobile devices
  - Imaging device
  - Hospital PCs

  ➔ Usually not as much computational power required compared to training (unless inference needs to happen in real time)
- Continuous monitoring of model performance on real-world data is vital
- Continual learning to update models regularly as new data becomes available

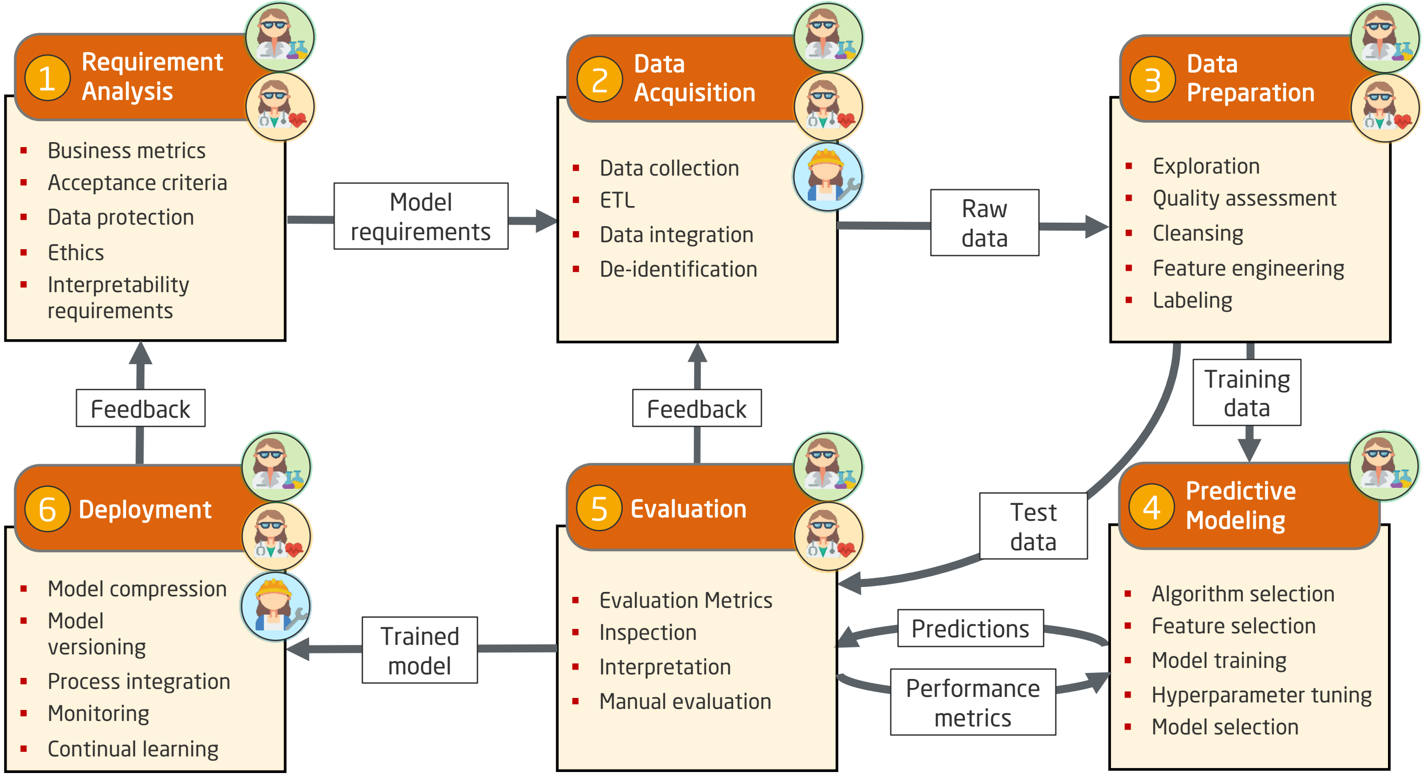
https://www.bbc.com/news/health-45275071

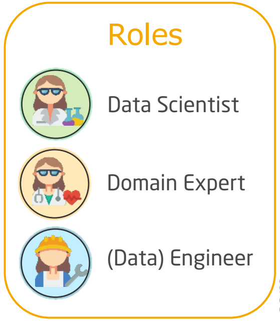
https://venturebeat.com/2018/11/28/apple-watch-series-4-ecg-app-release/

Hasso Plattner Institut

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

Process Model for ML in Digital Health

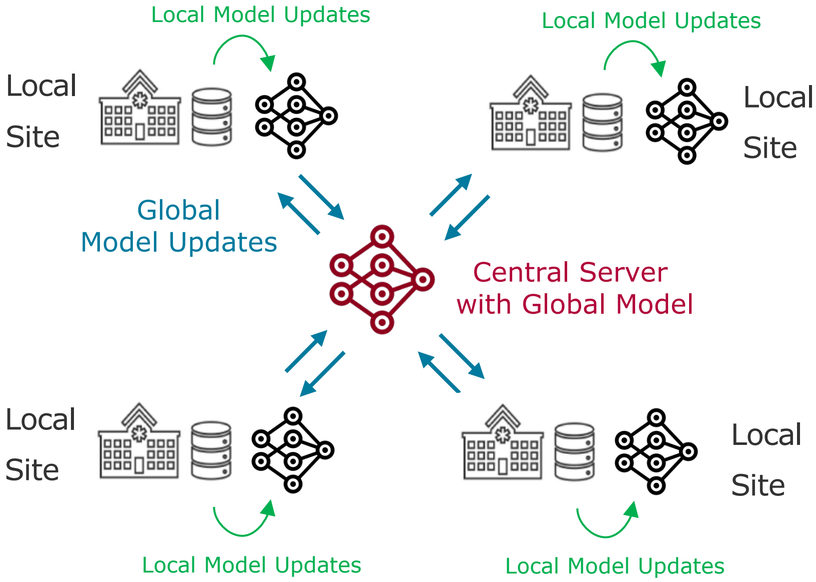




Federated Learning

- Paradigm for distributed training of ML models across sites
- No data sharing, just parameter updates

- In the centralized version, a central server orchestrates the learning process:
  - Selection of nodes for next iteration
  - Result aggregation
- Disadvantage of centralized FL:
  needs a trustworthy 3rd party, which is the single point of failure

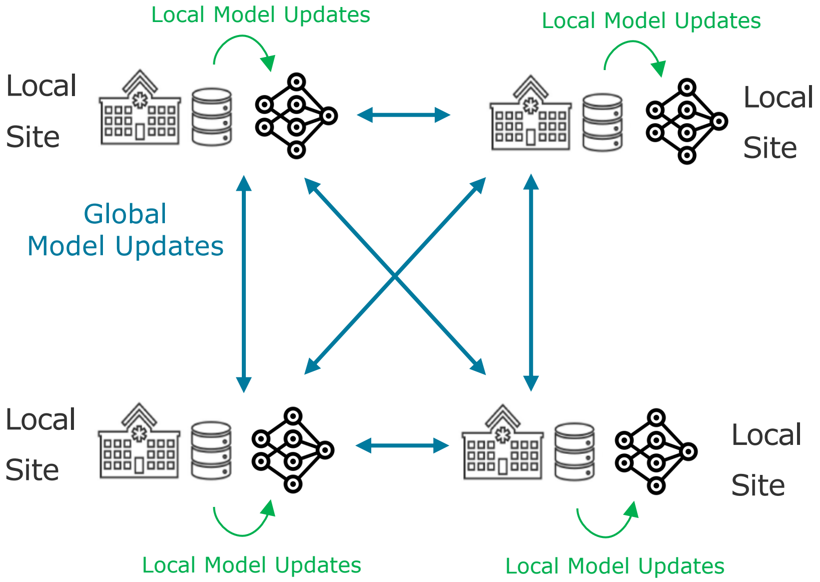


Pfitzner, B., Steckhan, N., & Arnrich, B. (2021). Federated learning in a medical context: a systematic literature review. *ACM Transactions on Internet Technology (TOIT)*, *21*(2), 1-31.

Decentralized Federated Learning

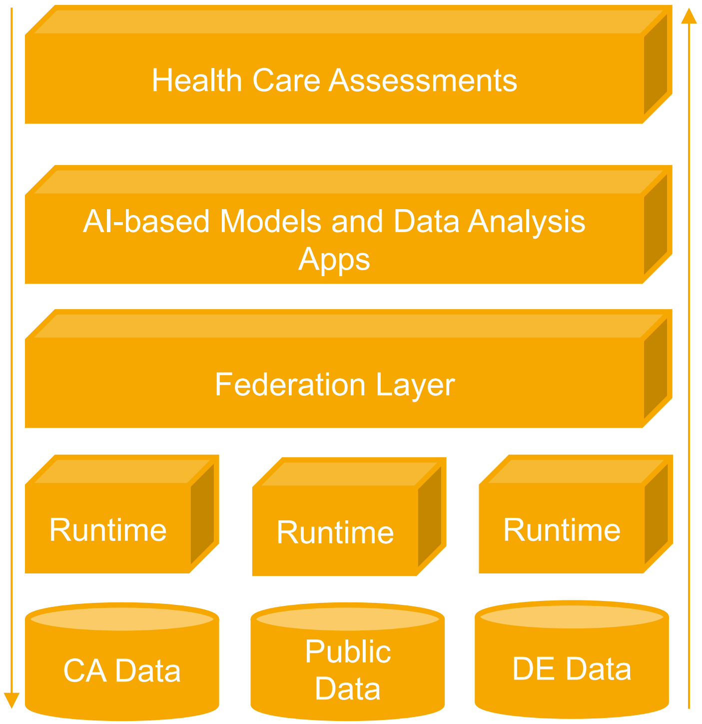
- Nodes coordinate themselves to obtain global model
- Model updates are propagated in a peer-to-peer manner

- Disadvantages:
  communication overhead, performance depends on topology of the network



NephroCAGE Federated Learning Software Architecture

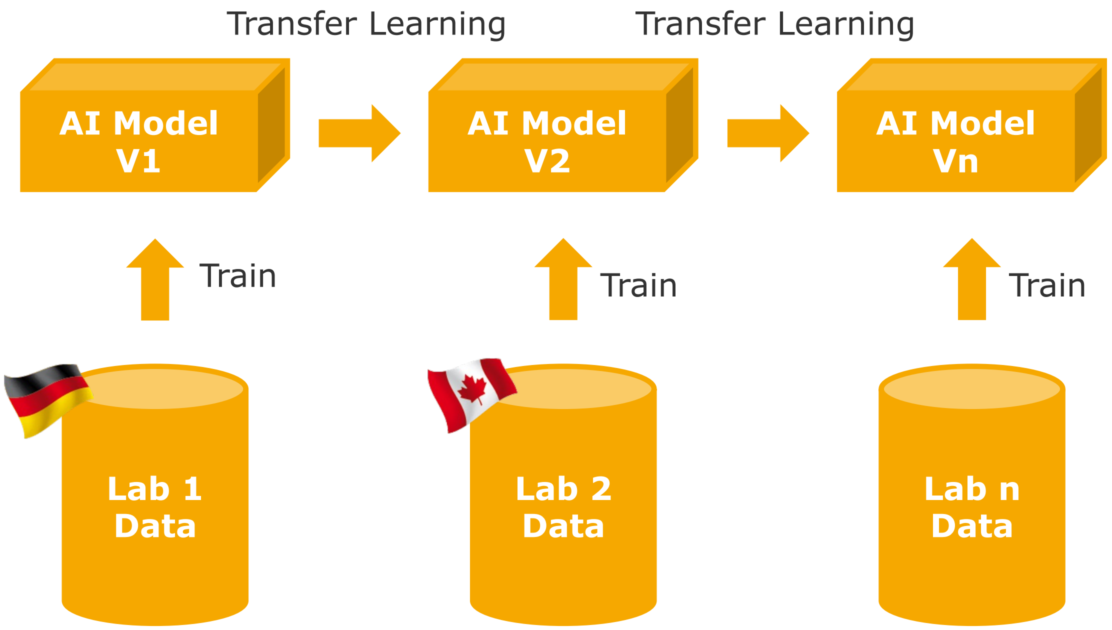
- Assess real-world transplant data from German and Canadian medical centers
- Access to 10yrs+ transplant data
- Healthcare data remains protected
- AI algorithms travel to data
- Federated learning enables data analysis whilst keeping data protected



NephroCAGE Federated Learning Software Architecture (cont'd)

Hasso Plattner Institut

NEPHRO CAGE



What to take home?

- "Garbage In – Garbage Out"
- Activities according to the ML process model
- Federated Learning as an emerging paradigm for digital health

New Jupyter Notebook!

(relevant for Exercise 1)

https://github.com/hpi-dhc/dm4dh-2023/blob/main/2_ML_Process.ipynb