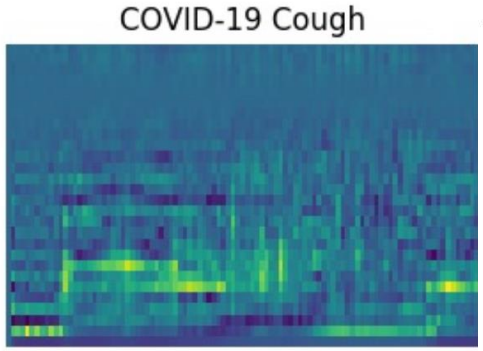
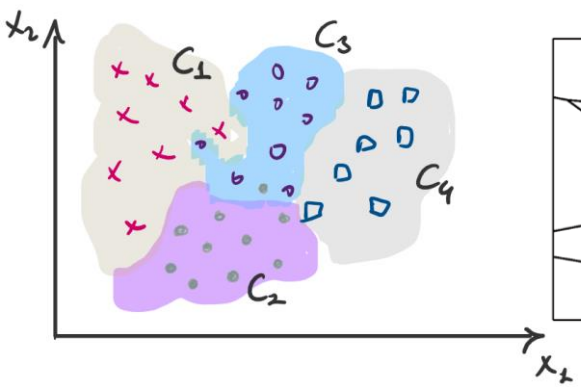
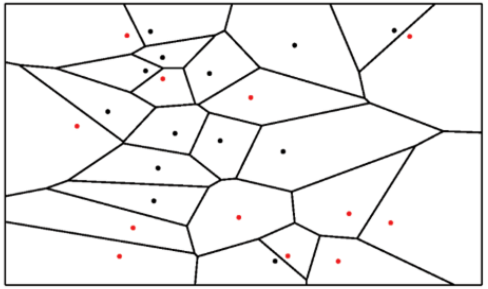
Data Driven Engineering I: Machine Learning for Dynamical Systems

Basics II: An Ode to Learning

Institute of Thermal Turbomachinery
Prof. Dr.-Ing. Hans-Jörg Bauer







Administrative Business

- Recorded lectures are online at ILIAS
- Lecture notes and active session notebooks
- Local installation guide available // Colab

One Page Summary : Intro. to DDE

* General AI ⇒ Narrow AI : One recipe cannot solve every problem
  ↳ Domain-specific expert models

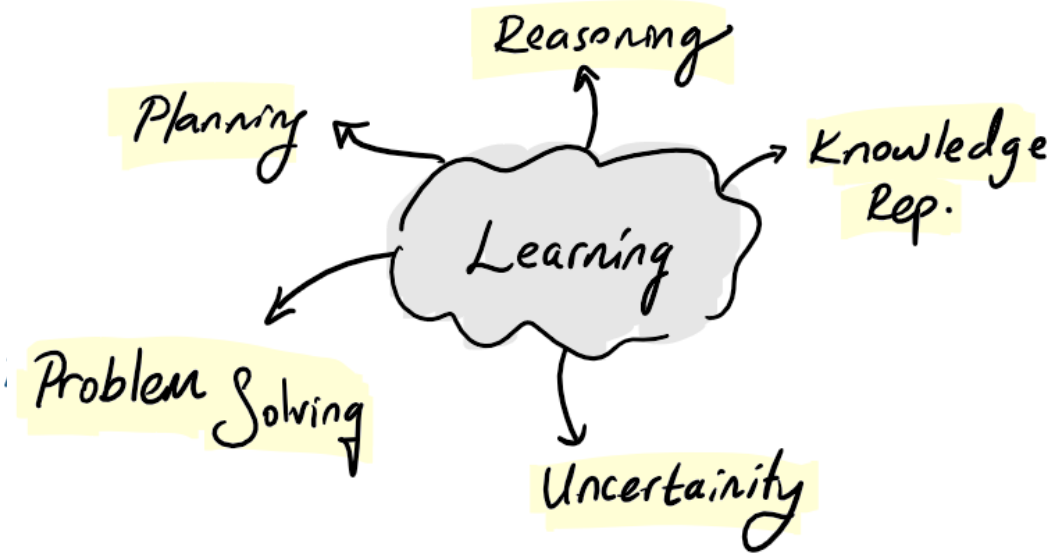
* nAI ⇨ ML := Data + Model + Hardware ⇨ Every aspect is critical
  ↳ What is new → Scale
  ↳ Accessibility

* DDE-1 ⇨ Learn to use ML Toolbox → "Cooking Class"
  Obj : end-to-end ML Project

Today's Agenda

1. What is learning?
2. How can a model learn automatically?
3. What kinds of problems can be solved?
4. What may go wrong?
5. Basic steps of a ML project

Learning landscape :



Depends on III factors,

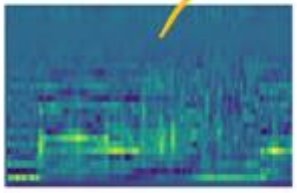
* Prior knowledge
* Data representation
* Type of feedback for learning

Learning landscape :

Eg// Covid-19 detection from sound,

* Prior knowledge := "f" should exist, $y = f(x)$ (input → x, output → y)

(i)



Data Rep.

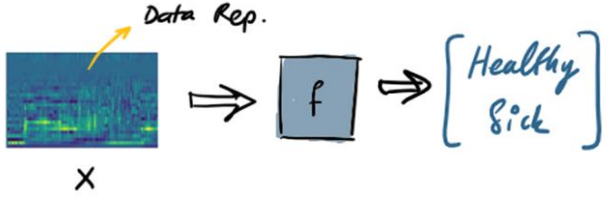
X → f → [Healthy / Sick]

? → y

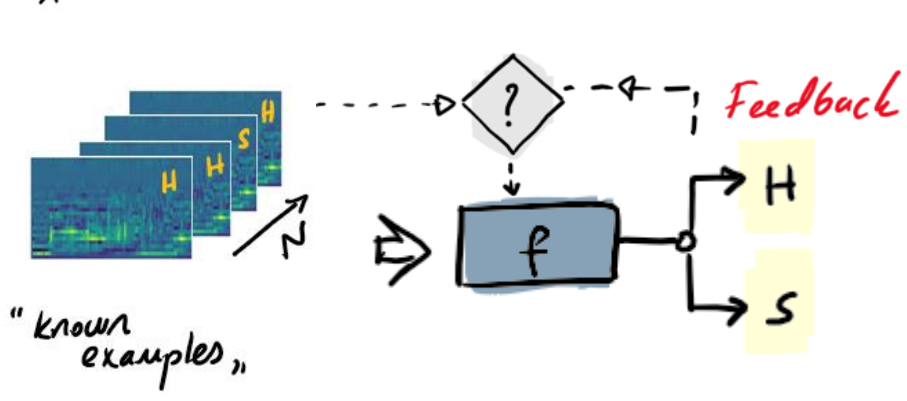
Learning landscape :

Eg// Covid-19 detection from sound,:

(i)

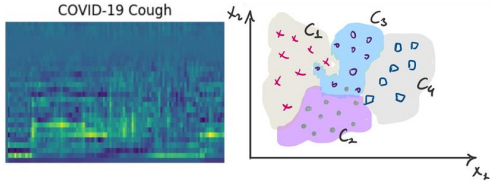


(ii)



- Prior knowledge
- Data representation
- Type of feedback for learning

Learning landscape :



Machine Learning

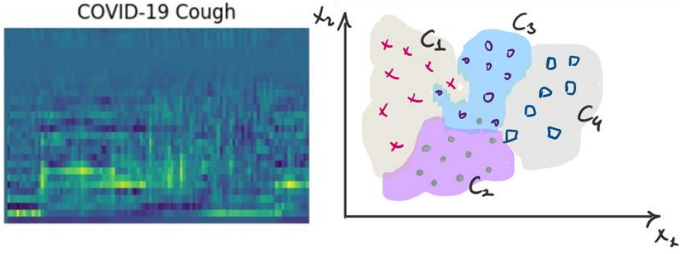
<<<<< Nature of the feedback >>>>>

- Predictive → Supervised
- Descriptive → Unsupervised
- Reward / Punishment → Reinforcement

Learning types :

- Error-based learning
- Similarity-based learning
- Information-based learning
- Probability-based learning

Learning landscape :



Machine Learning

Supervised → Predictive

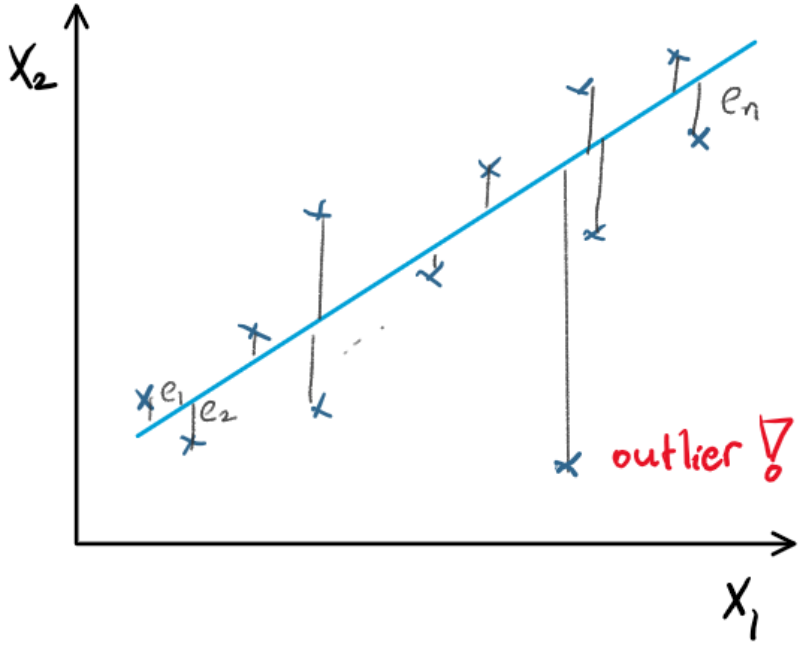
Unsupervised → Descriptive

Reinforcement → Reward / Punishment

Learning types :

- Error-based learning
- Similarity-based learning
- Information-based learning
- Probability-based learning

Error-based learning:



(i) Parameterized model ($X_2 = aX_1 + b$)

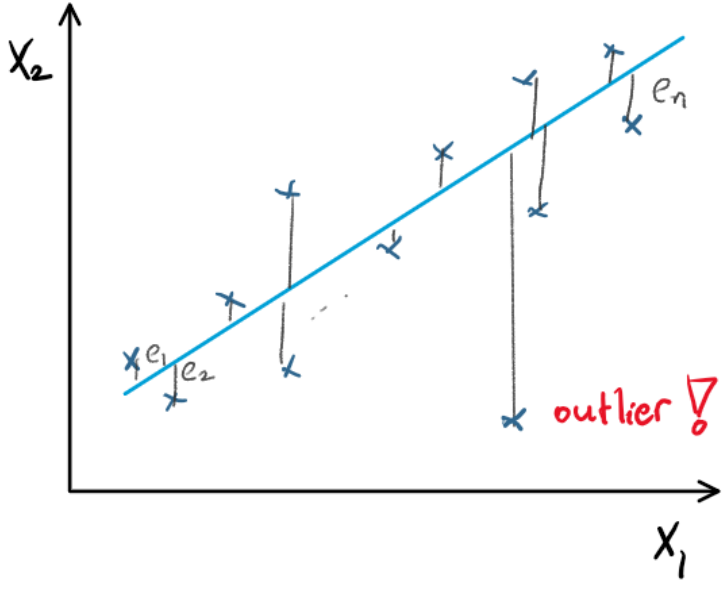
(ii) Definition of an error (MSE, MAE, R^2...)

(iii) Search for parameters
↳ minimize total error ($\sum^{n} e_i$)

How do we know correct answers?

↳ Training dataset

Error-based learning:



Common algorithm:

0) Prepare the training dataset.

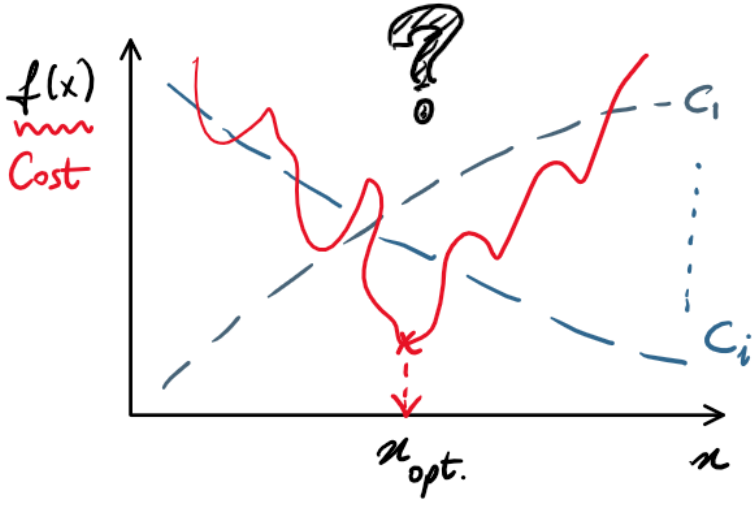
1) Initialize the trainable parameters

2) Calculate the error

3) Update the trainables in a smart way

How can we pick a smart way?

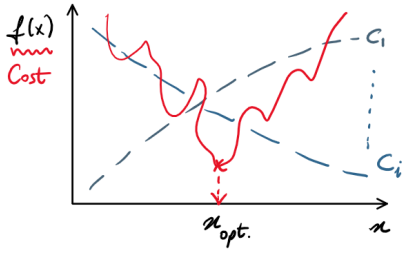
DDE ⇔ Optimization



⇨ Competing factors

⇨ Cost function

- Single / multiple variable
- Static / dynamic
- Discrete / continuous
- Contrainted / uc
- function / Data



Search Space [x]

"Calculus"

1

Newton

Random

non-guided

Las Vegas

Guided

Evolutionary algorithms

GA GP EP ...

Enumeratives

non-guided

Backtracking

Guided

Dynamic Programing

Similarity based learning :

"When see a bird that walks like a duck, swims like a duck & quacks like a duck, I call that bird duck"

- to make a Prediction := "call"
  - ↳ look what best worked in the past := "duck"
  - ↳ predict the same := "duck"

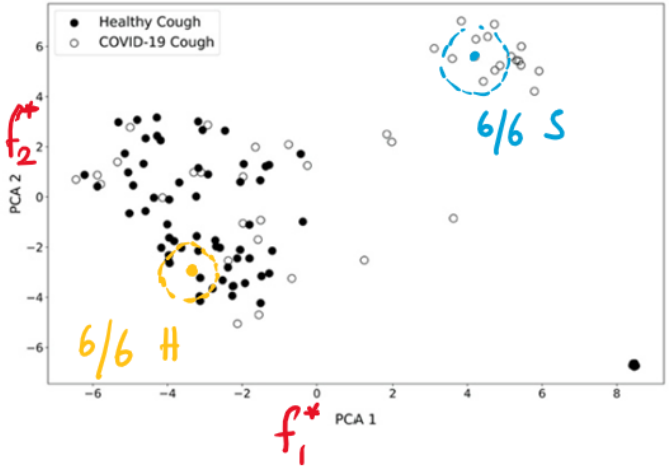
Build a system based on
- features
  - walk
  - swim
  - quack
- measure of similarity

Similarity based learning :

eg. COVID Detection

#	Age	Gender	Height	Weight	....	Voice
1	32	M	1.70	75	...	
⋮						
i	19	F	1.75	60	...	
	f_1	f_2	f_3	f_4	f_{n-1}	f_n

N → 2



* Distance based ⇒ "hyperspace" N dimensions

* Noise
* Normalization
* Feature Selection

Information-based learning:

* Source of knowledge
* Data representation

} Properties measured
↳ "Features"

? Which feature is most important for the given task ...

"informative"

! How to measure it <- - -!

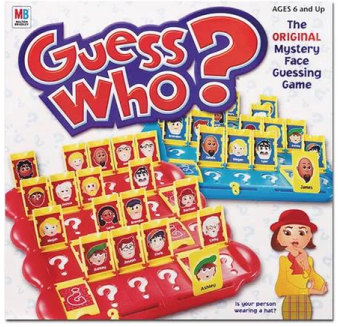
- information gain
- entropy

Information-based learning:

Eg

	Joey	Chandler	Monica	Phoebe
Man	✓	✓	x	x
Glass	✓	x	x	x
long hair	x	x	✓	x

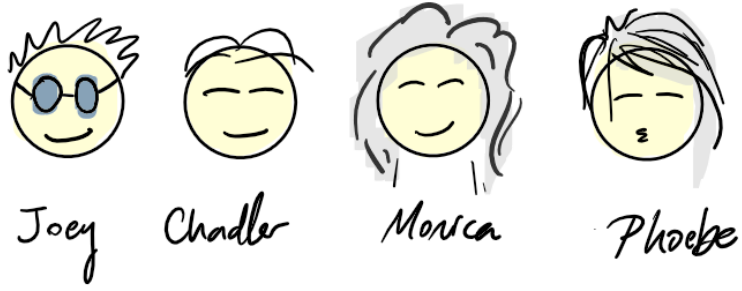
Idea: Test different features for their informativeness.



boardgamegeek.com

Information-based learning:

Eg



☑ Learn to ask right questions in the correct order.

(i) 4 → Glass —N→ 3 → Man —N→ 2 → Hair —N→ 1

Glass —Y→ 1

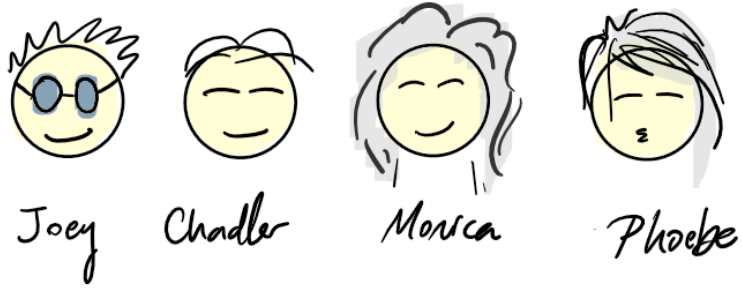
Man —Y→ 1

Hair —Y→ 1

} 2.25 Q / person

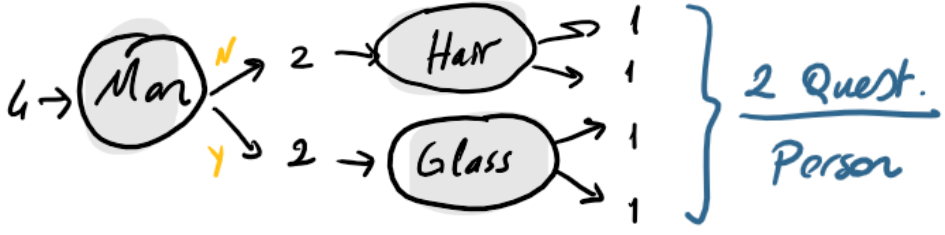
Information-based learning:

Eg



- Learn to ask right questions in the correct order.
- Method: Shannon's model of entropy

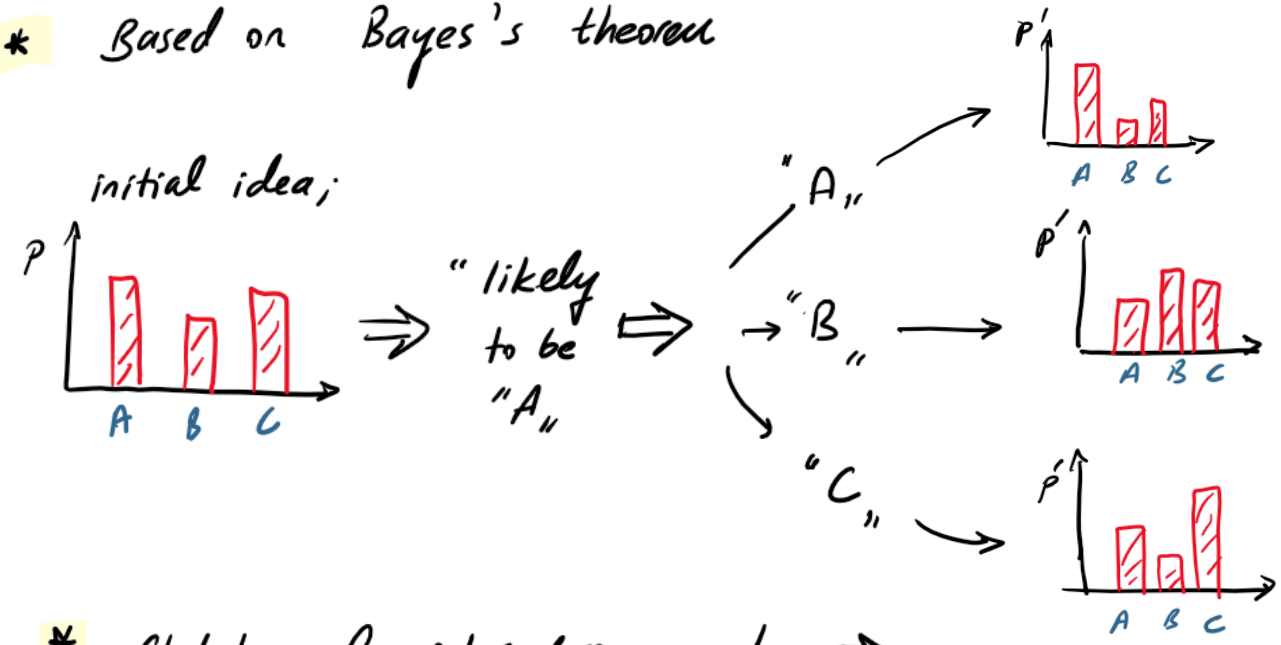
(ii)



2 Quest. / Person

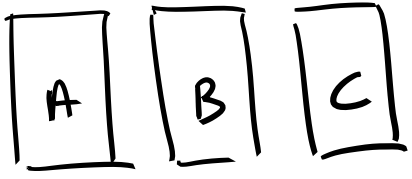
Probability based learning :

* Based on Bayes's theorem

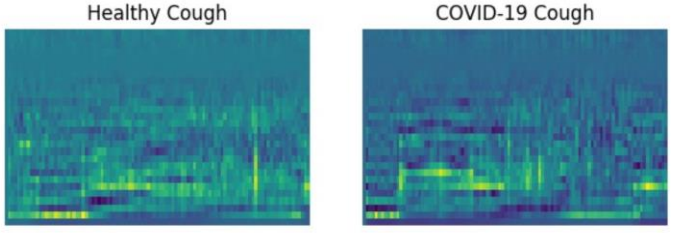


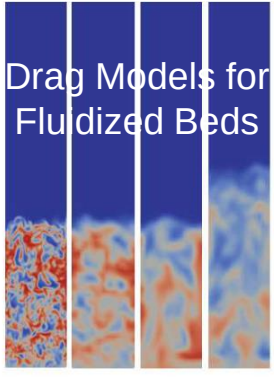
* Click here for interactive examples

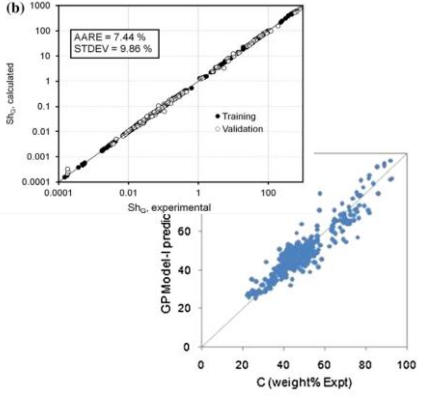
Find the Lady



Standard Learning Tasks-I



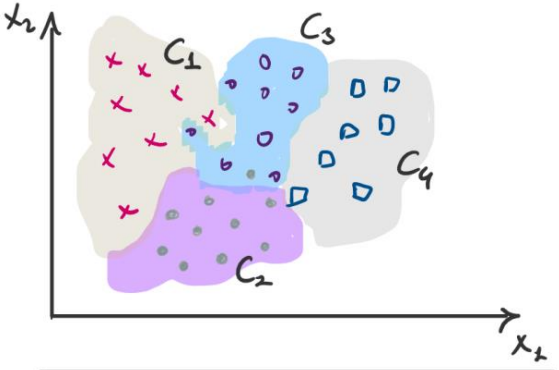


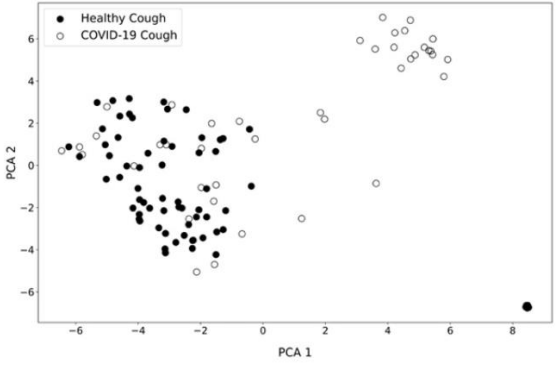


- **Classification:** problem of assigning a category to each item (**discrete**)
  - ✓ COVID-19 classifier of MIT
  - ✓ # Categories < 100
  - ✓ Unbounded classification: text classification, speech recognition

- **Regression**: problem of predicting a real value for each item (**continuous**)
  - ✓ Predicting the noise of an airfoil, turbofan predictive maintenance
  - ✓ Error estimation> difference between the true and predicted values

Standard Learning Tasks-II



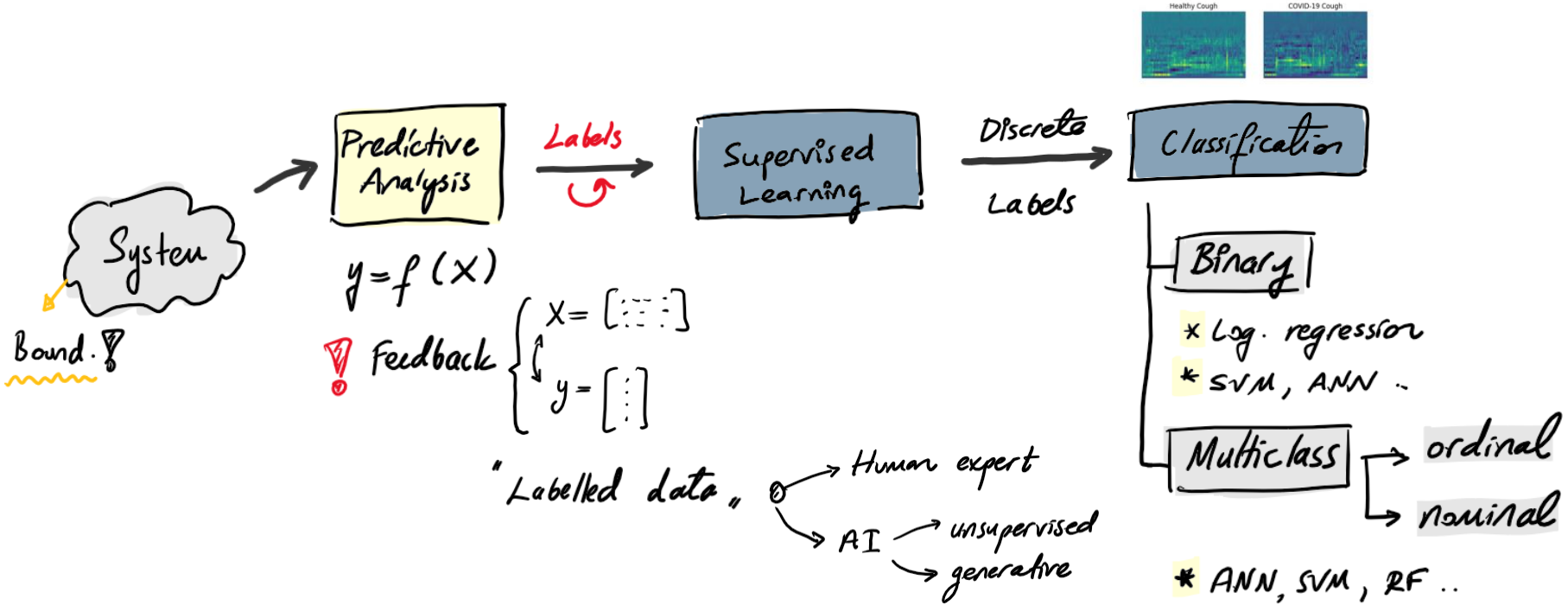


- **Clustering**: problem of partitioning a set of items into homogeneous subsets
  - ✓ Manufacturing error analysis
  - ✓ Very large data sets

- **Dimensionality reduction:** problem of transforming an initial representation of items into a lower-dimensional space
  - ✓ Manufacturing error analysis, image compression
  - ✓ Preserving the properties of the initial representation

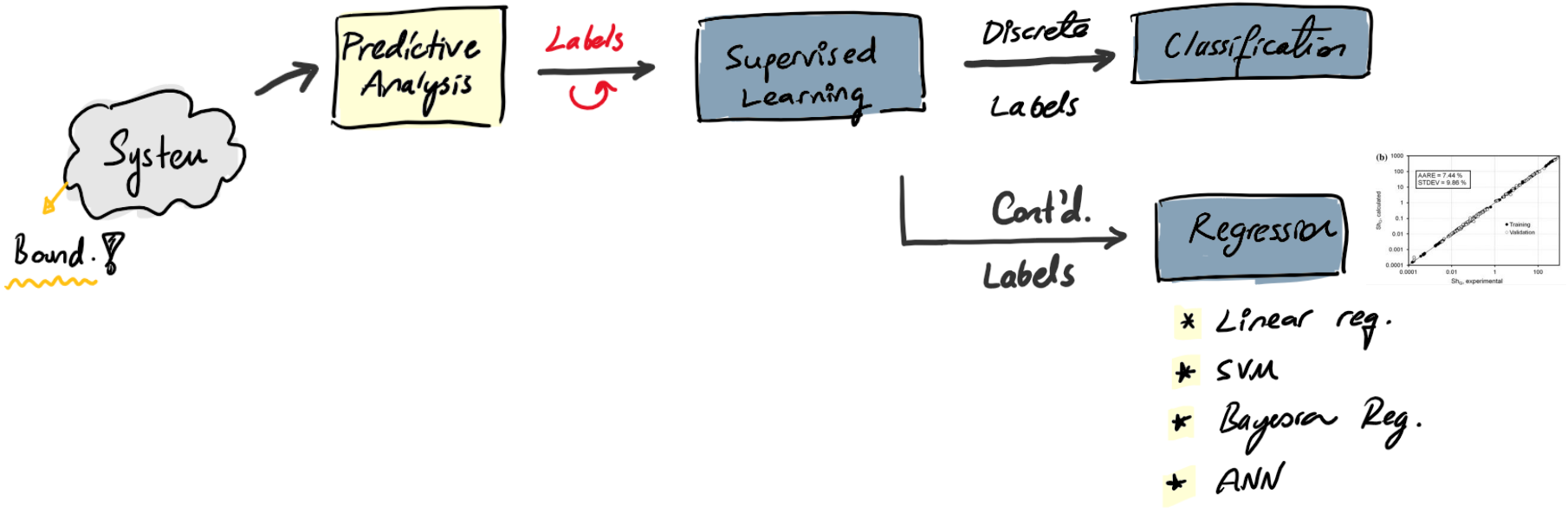
How does it work?

"Different recipes (learning strategies) for different problems"



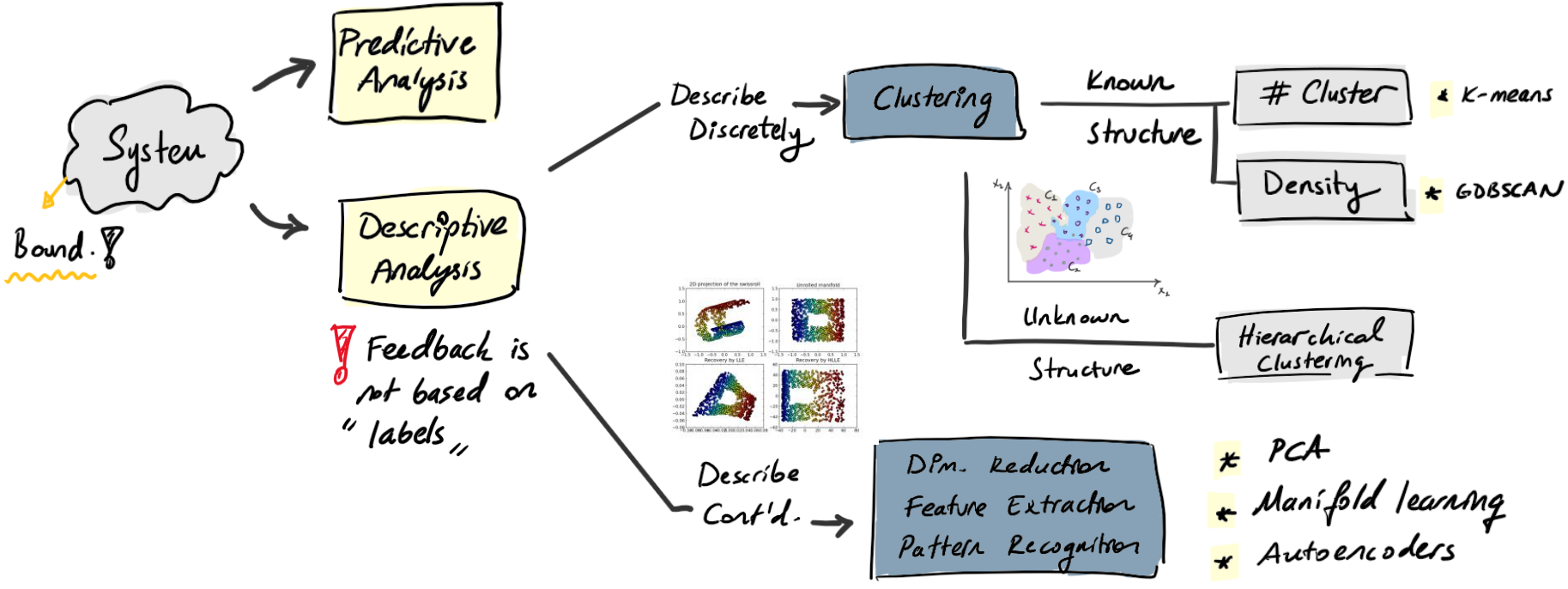
How does it work?

“Different recipes (learning strategies) for different problems”



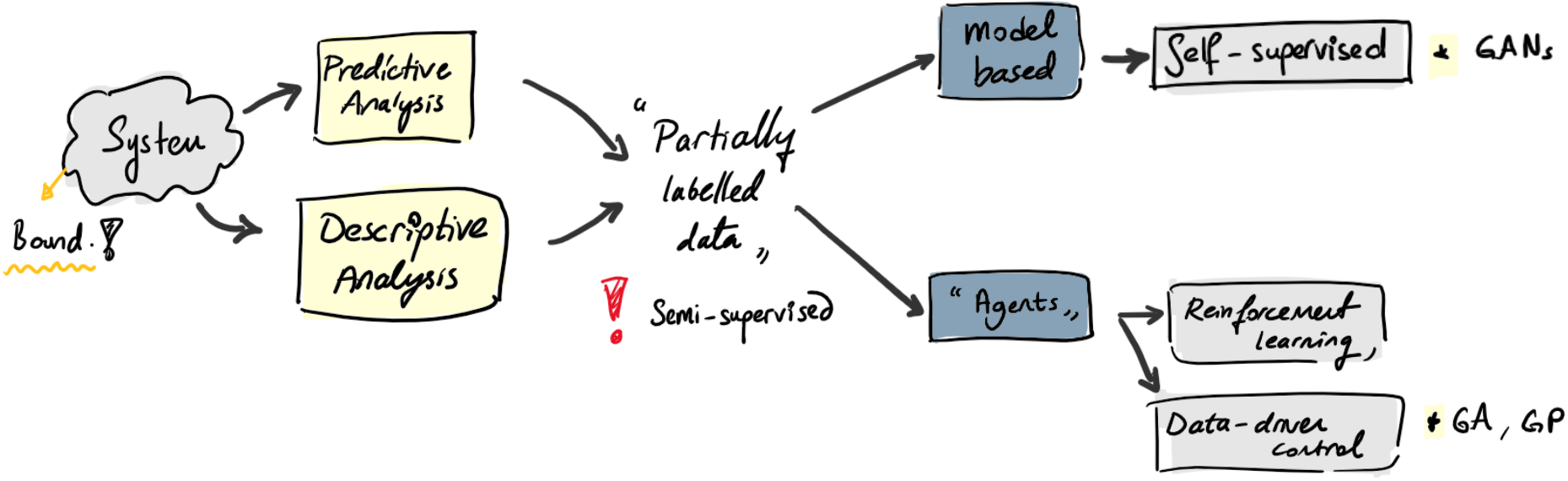
How does it work?

“Different recipes (learning strategies) for different problems”



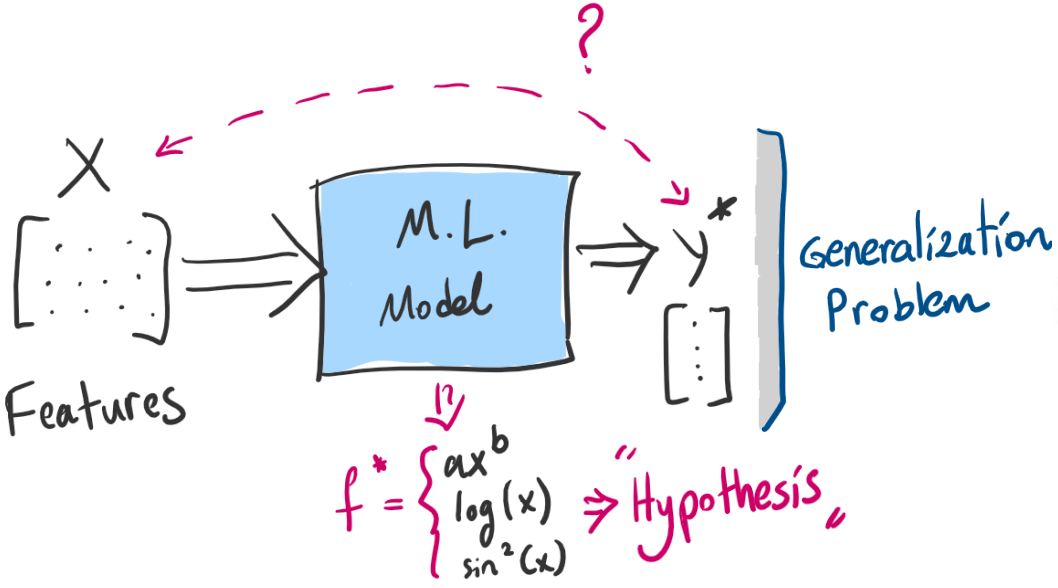
How does it work?

“Different recipes (learning strategies) for different problems”

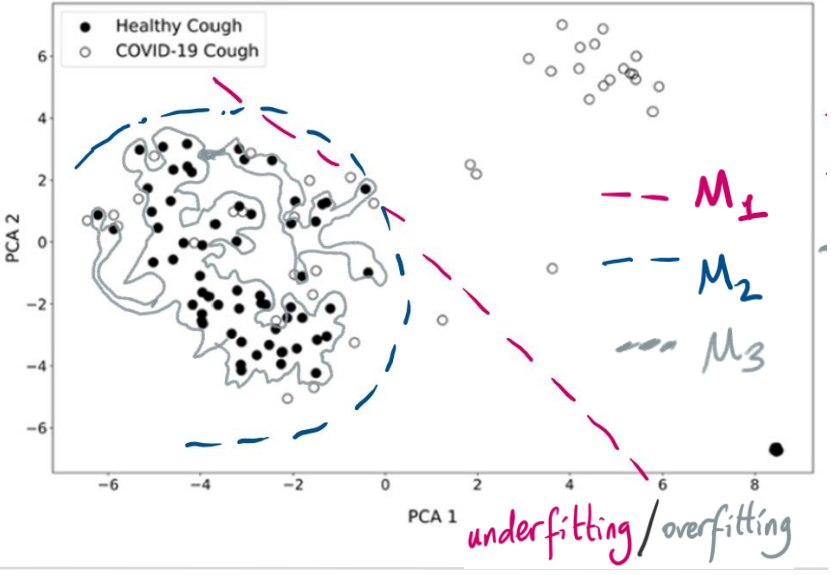


How can I select a model?

"Many recipes for the same problem"
ML is an "ill-posed problem"



How can a hypothesis be chosen?



How can I select a model?

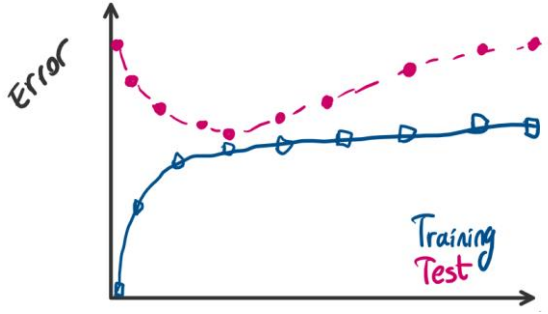
"Many recipes for the same data"

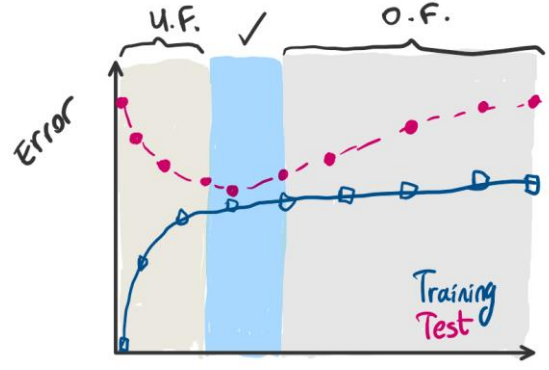
How can we pick the right model?

We can calculate the error !

↳ Prediction error ⇒ "overfitting

↳ Objective: Minimize the generalization error

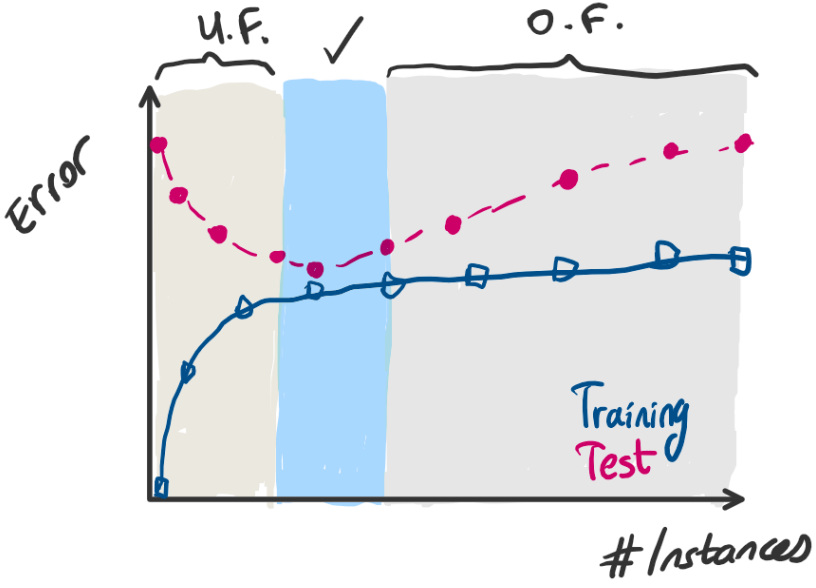




How can I select a model?

"Many recipes for the same data"

How can we pick the right model?

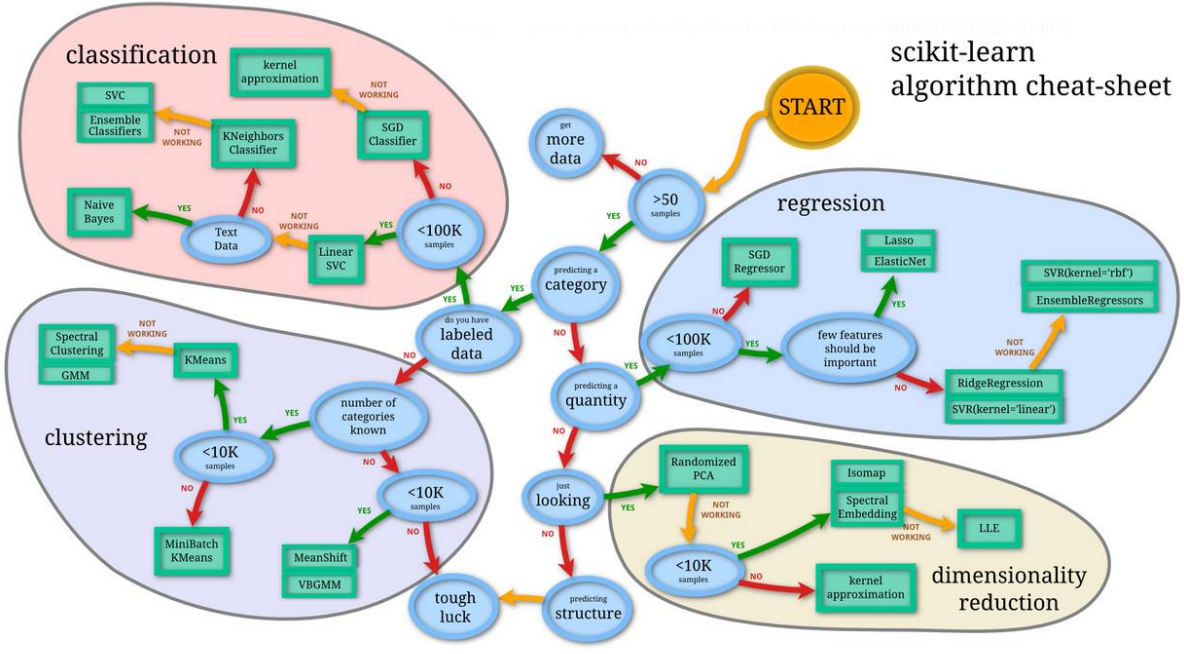


* Problem: We can not use test set to pick the right degree of complexity.

$x_0 = f(R_1)$
$x_1 = f(R_2)$
⋮
} Cross Validation

How does it work?

"Different recipes for different problems"



No free lunch theorem

Wolpert, 1996

There is no universally best model

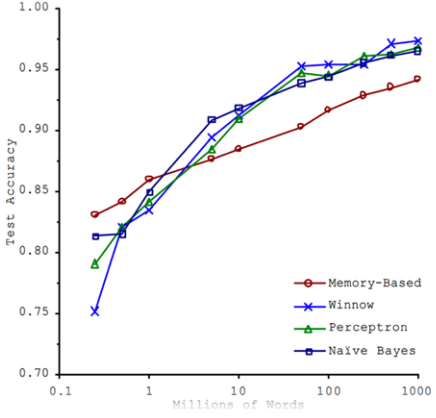
Assumptions that works well in one domain may work poorly in another.

How can I select a model?

"Insufficient **Data** for the Training"

- **High accuracy:** high volumes of data required
  - ✓ very simple problems ~ 1000s examples
  - ✓ Complex problems ~ millions of examples
- **Large datasets** >> computational burden
  - SoA Deep Learning training: Energy eq. of the electricity consumption of a city for a few days
- **Why bother with larger datasets?**

"The Unreasonable Effectiveness of Data"



"Scaling to Very Very Large Corpora for Natural Language Disambiguation"

Michele Banko and Eric Brill
Microsoft Research
2001

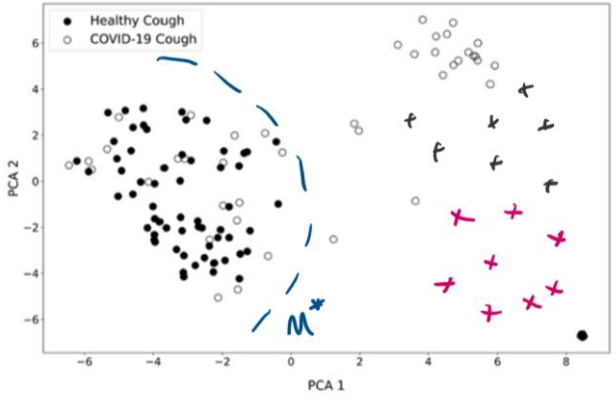
- ✓ Size of the data mattered far more than the choice of ML approach
- ✓ Differences became very small as the data grew large

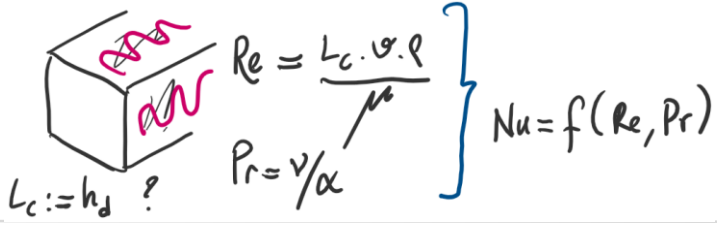
How can I select a model?

"Insufficient **Quality** for the Training"

- **Representative data**
  - training data ~ new cases
  - Sample is too small >> "sampling noise"
  - Sampling method flawed >> "sampling bias"

- **Data quality**
  - errors, outliers, and noise (sensor, model)
  - spend time cleaning up (outliers)
  - missing features (ignore / guess / omit)
  - Dimen. Reduction
  - "Feature engineering": select, extract, combine



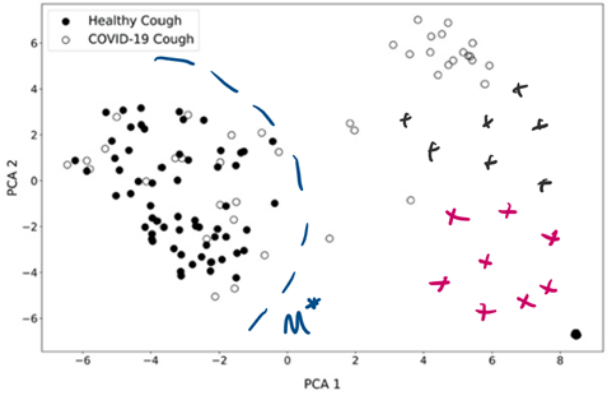


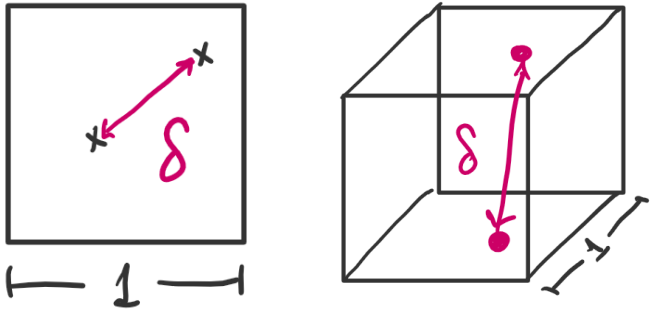
How can I select a model?

"The **curse** of **dimensionality**"

- Typical ML feature space ~ millions
  >> higher dimensional space
  >> How can I "**draw**" separation "**curve**"

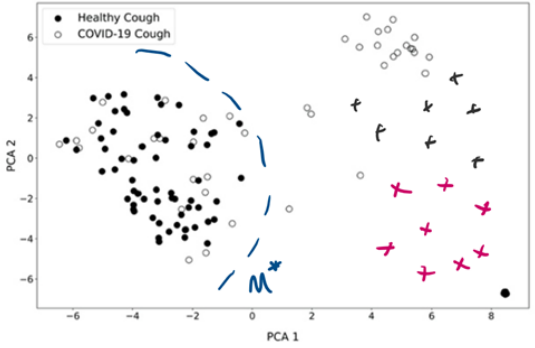
- ✓ Pick two points randomly in a unit square:
  >> the distance ~ 0.52
- ✓ Pick two random points in a unit 3D cube:
  >> the distance ~ 0.66
- ✓ Pick two random points in a unit 1M-D hypercube:
  >> the distance ~ 408





How can I select a model?

"The **curse** of **dimensionality**"



- Typical ML feature space ~ millions
  - >> higher dimensional space
  - >> How can I "**draw**" separation "**curve**"

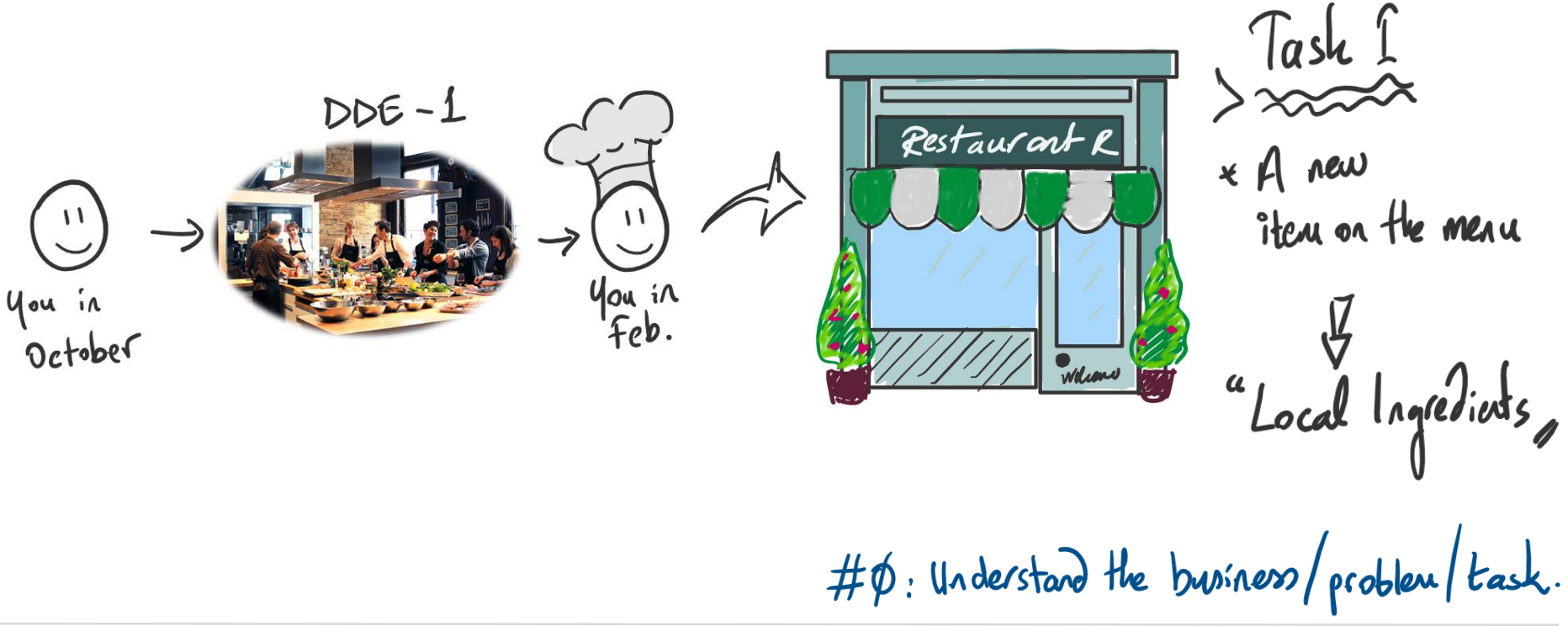
Why this is a problem?

- New instance will be far away from any training example
- Higher the dimension ~ greater the risk of overfitting
  - >> Solution 1: increase the size of the training set
  - >> Solution 2: dimensionality reduction

The Style

How does it work?

"PFD of a ML project"



How does it work?

"PFD of a ML project"

#1: Understand the data: The Sources available & the type.

How does it work?

"PFD of a ML project"

Remove the spoiled food

The form: pealed; chopped
boiled; frozen

size, shape, texture

"Recipe"

Data: Type, Scale
Structure

ML

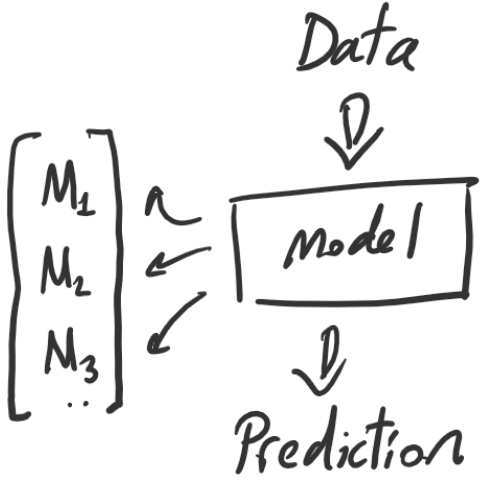
Predictions

#2 Data Preparation / Exploratory Data Analysis

How does it work?

"PFD of a ML project"

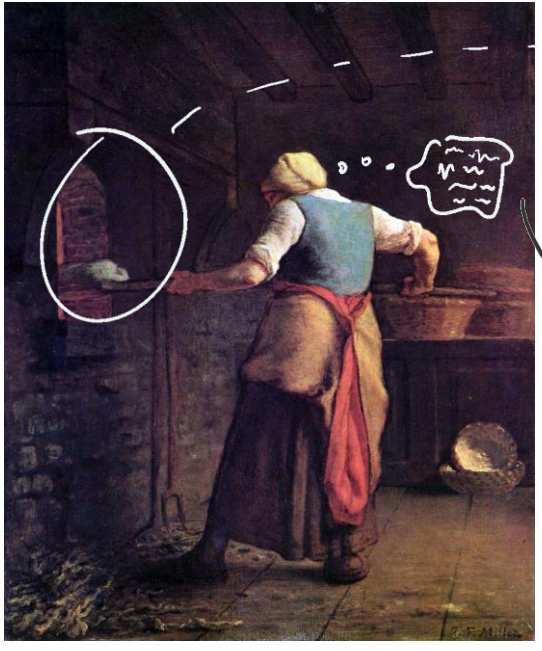
⇒ Chosing the right cooking method



#3 Modeling: Try different ML approaches

How does it work?

"PFD of a ML project"

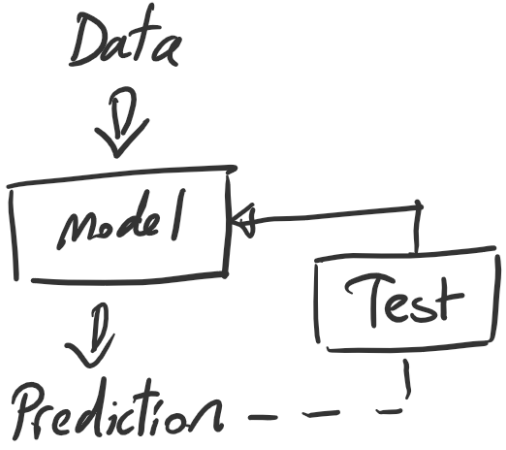


Cook the dish by using the recipe

Compare its taste to the dishes tasted before.

How much creative you can be?

- Pure randomness
- "Regularization"



#4 Training the ML model

How does it work?

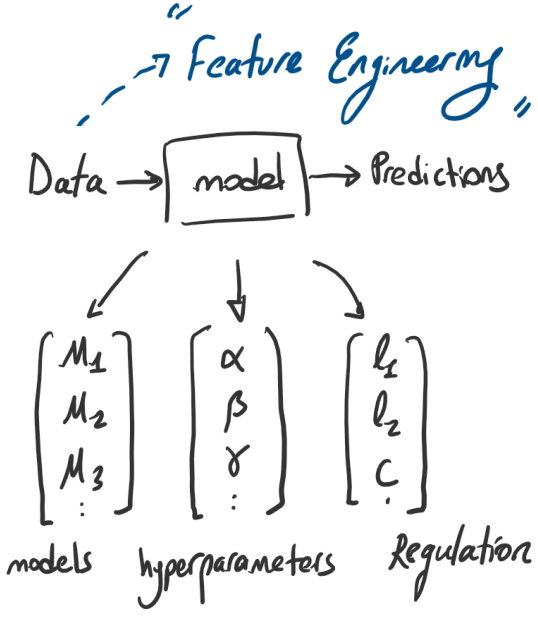
"PFD of a ML project"

⇒ how satisfied are you with the taste?

↳ Adjust ingredients

↳ Add more complexity "

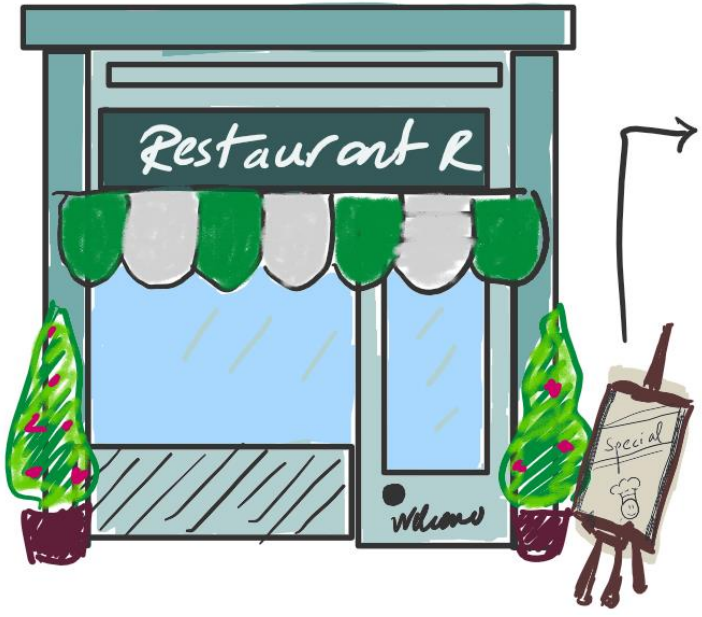
↳ Change the cooking method



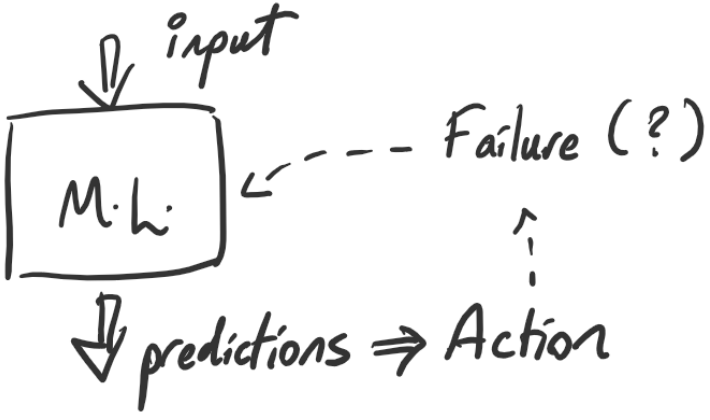
#5 Evaluation: capable of making accurate predictions

How does it work?

"PFD of a ML project"



Dish in on the menu ?



#6 Deployment: Model is ready for usage

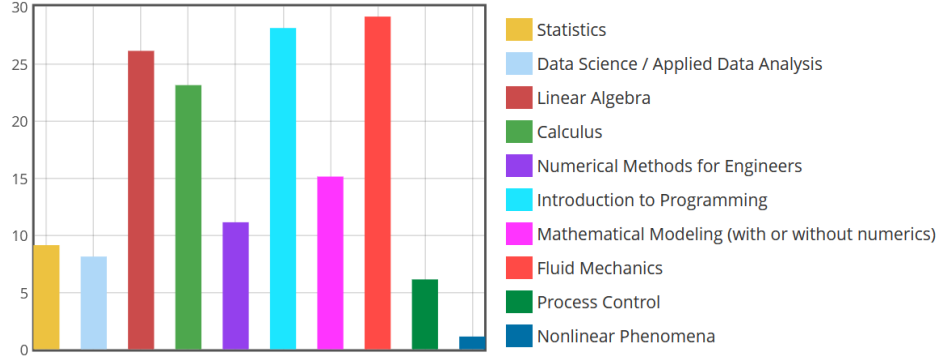
Colab: An Introduction

❑ Introduction to Python with Colab

Objective:

✓Introduce Colab environment

✓Introduce some basics



Today's Agenda

- ✓ Theory of Learning and Learning Types
- ✓ Machine Learning: Overview, Means and Goals
- ✓ Problem Solving
- ✓ Planning: How a ML project is organized

- ❑ Local installation guide >> ILIAS
- ❑ Next Week: Classification Methods in ML **with active session!**

Additional Notes

Some important keywords to know…

- **Examples:** Items or instances of data used for learning or evaluation
- **Features:** The set of attributes, often represented as a vector, associated with an example.
- **Labels:** Values or categories assigned to examples. In classification problems, examples are assigned specific categories, (e.g. healthy // sick)
- **Hyper-parameters:** Free parameters that are not determined by the learning algorithm, but rather specified as inputs to the learning algorithm.
- **Training sample:** Examples used to train a learning algorithm.

Some important keywords to know…

- **Validation sample:** Examples used to tune the parameters of a learning algorithm when working with labelled data.
- **Test sample:** Examples used to evaluate the performance of a learning algorithm. The test sample is separate from the training and validation data and is not made available in the learning stage.
- **Loss function:** A function that measures the difference, or loss, between a predicted label and a true label.
- **Hypothesis set:** A set of functions that maps the features (feature vectors) to the set of labels (assumed relationship possibilities between the features and labels).

Wiki is being updated...

cihan.ates › Data Driven Engineering › Wiki › **Toolbox**

Last edited by **cihan.ates** 18 hours ago

Page history

Toolbox

- Python for ML applications
- ML libraries and tools
- Datasets
- Scikit learn
- TensorFlow
- Useful tools
- Others places to look for answers

Clone repository

Home

- Recommended resources
- Toolbox
- Miscellaneous topics

DDE 1: ML for Dynamical systems

- Ode to Learning
- Outlier Detection

DDE 2: Advanced topics

- Evolutionary learning