University of Zurich[UZH]

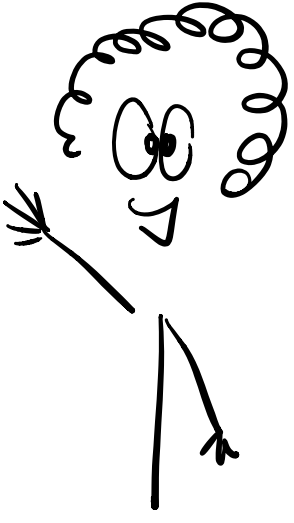
Institute of Computational Linguistics

Introduction to Machine Learning

Lesson 1

Mathias Müller, Phillip Ströbel

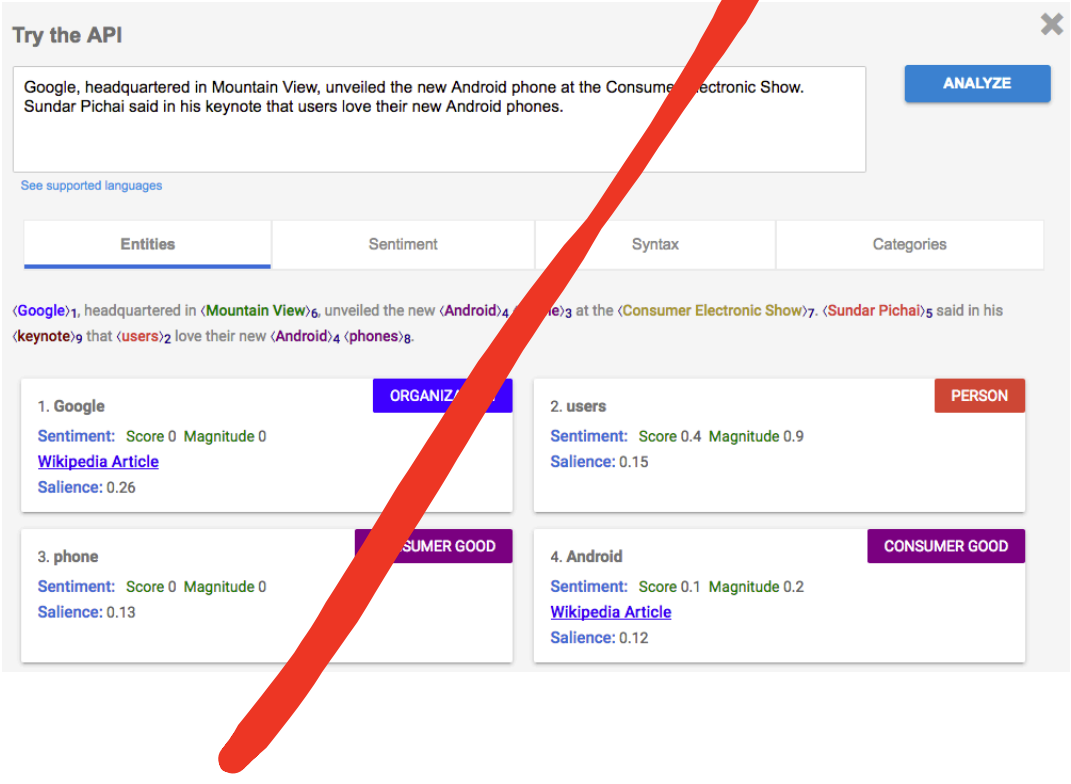
Who we are



MATHIAS

PHILLIP

Purpose of this course



University of Zurich[UZH]

Institute of Computational Linguistics

Core Concepts of Machine Learning

Topic of this lesson

- introducing core concepts and terminology of machine learning, namely: models, features, representations

What is Machine Learning?

- lernen aus Mustern

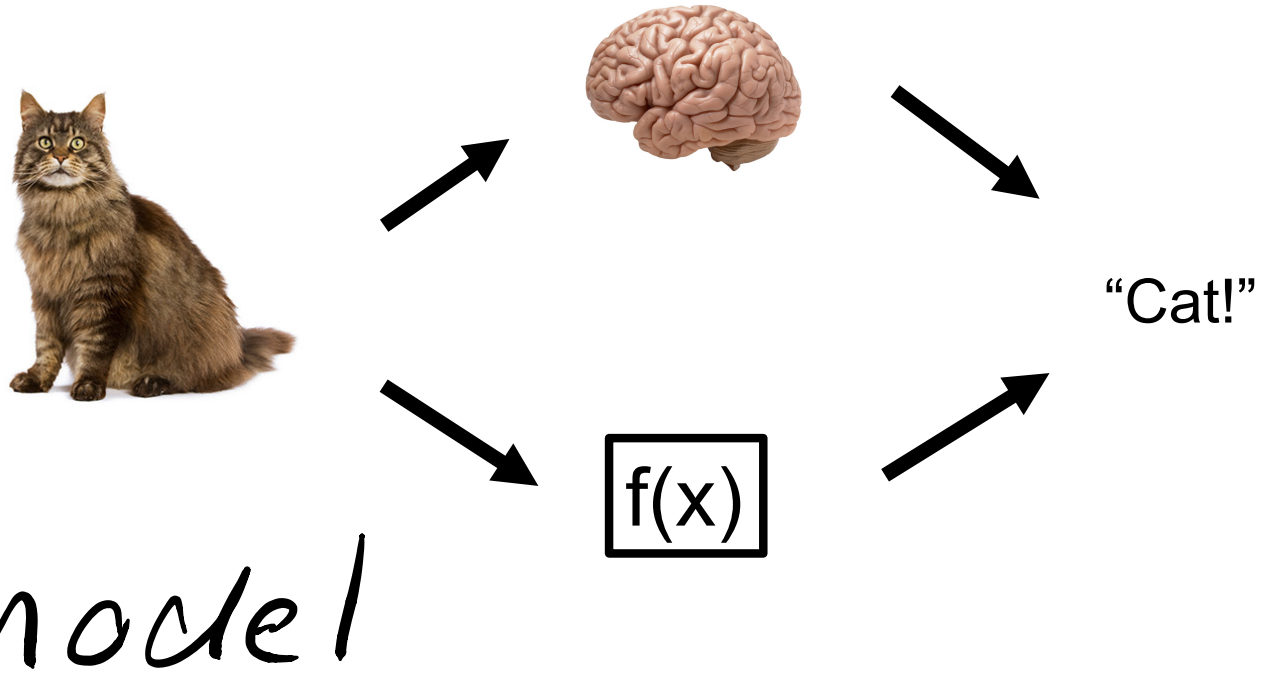
- Daten

- supervised vs unsup.

What is Machine Learning?

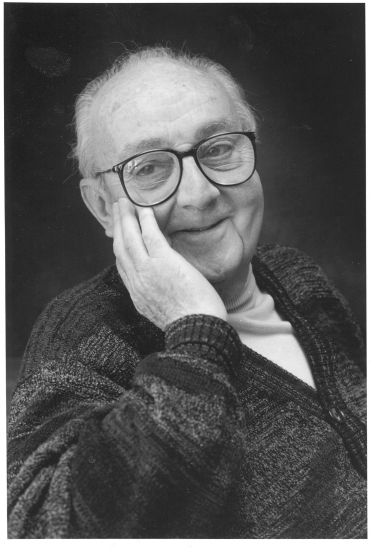
“Cat!”

What is Machine Learning?



The Real World vs. Models

"all models are wrong, but some are useful"



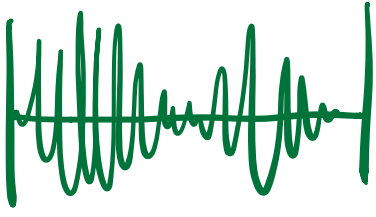
George Box

Box

The Real World vs. Models

Representations

- Objects must be represented as lists of numbers

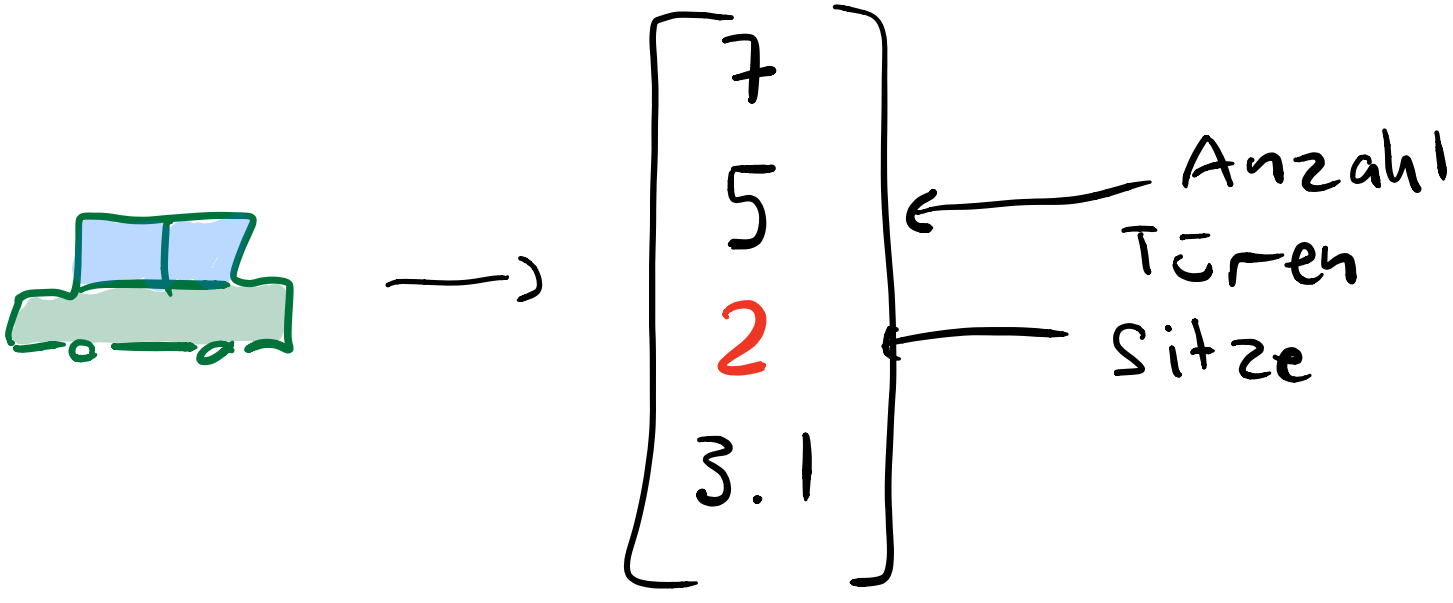


"Bern – gern."

$$\begin{bmatrix} 17 \\ 3 \\ 2 \end{bmatrix} \quad \begin{bmatrix} 17 \\ 8 \end{bmatrix} \quad \begin{bmatrix} 2 \\ \cdot \\ 8 \end{bmatrix}$$

Features

- A feature is a single number



Test your understanding

$$\begin{bmatrix} 3 \\ 7 \\ 5.1 \\ 8 \end{bmatrix}$$

- How many features are there?
- What do the features mean?
- Is this a machine learning model?

Summary

- Machine learning means modelling parts of our world, and learning behaviour
- a model is an imperfect abstraction of reality
- real-world objects are described by features which must be numbers

University of Zurich[UZH]

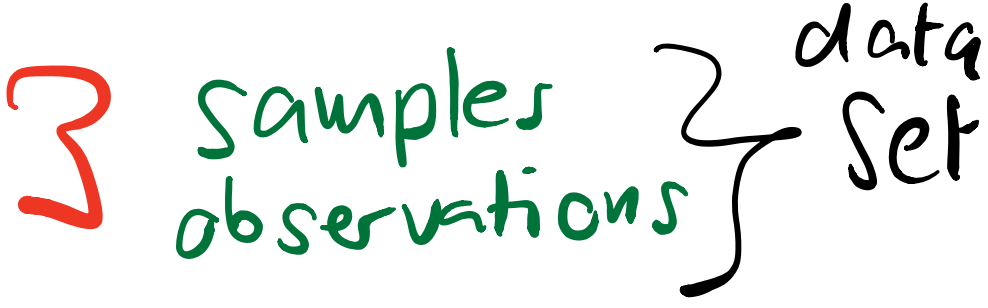
Institute of Computational Linguistics

Data, Phases and Splits

Topic of this lesson

- how to talk about data
- phases of machine learning
- data set splits

Data: Samples



Name of Beverage	Sugar (g)	Vitamin C (mg)
Coke	10	Zero ☺
Orange juice	9	120
Tap water	0	0 (traces!)

Further reading: *That Sugar Book* (2015). Damon Gameau. Pan Macmillan Books, Sidney.

Collecting Data

Bias

My employees like me, the percentage is even higher for people with a temporary contract! – Anonymous CEO

We have reached out to our customers about mobile network coverage. Almost 9 out of 10 Americans contacted on their phones are satisfied with coverage.– Telecommunications Corporation

response bias

Name of Beverage	Sugar (g)	Vitamin C (mg)
Coke	10	Zero ☺
Orange juice	9	120
Tap water	0	0 (traces!)

Data: Labels

Label

Target

Response

goldstandard

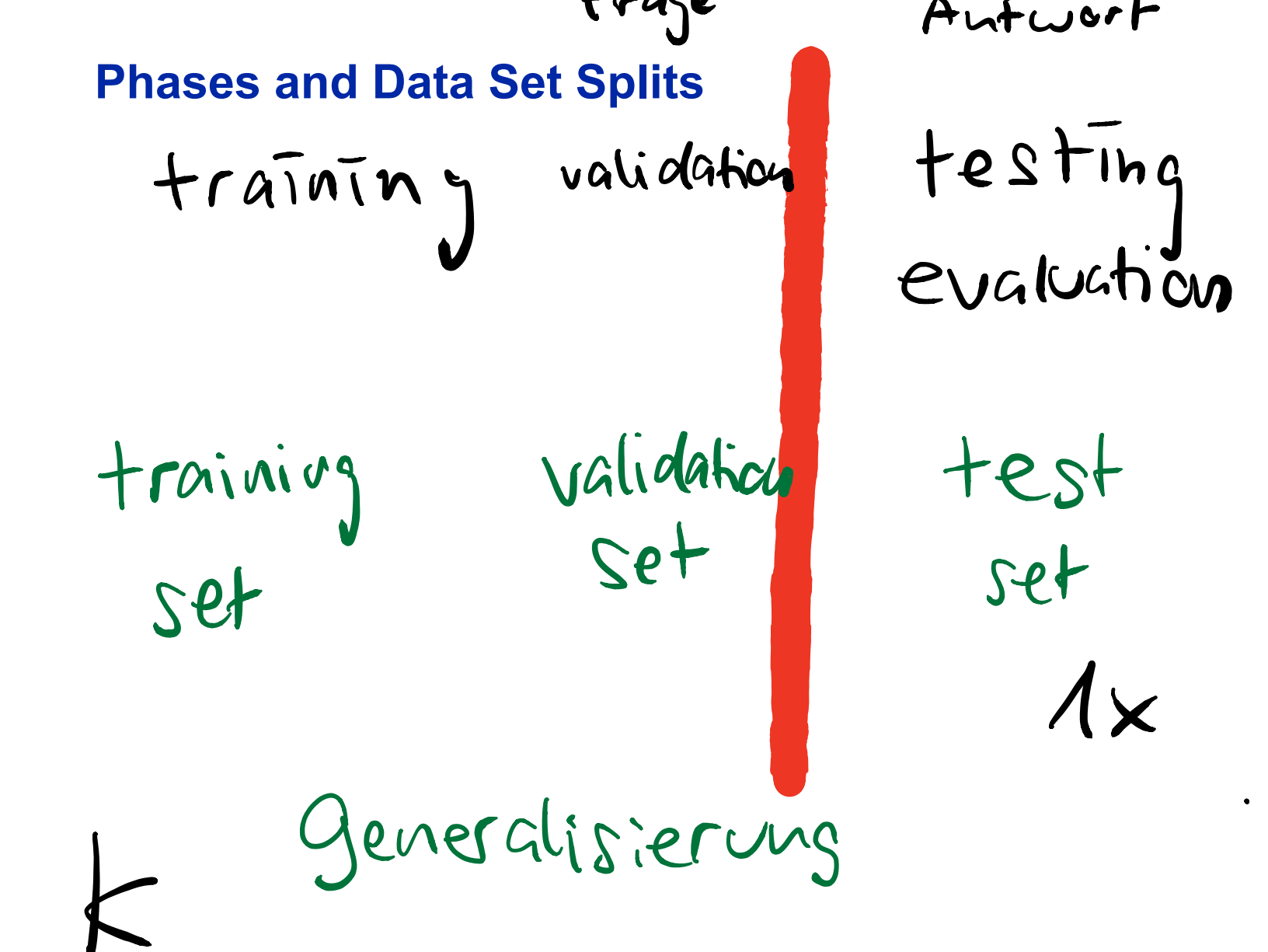
Label

& 10

Name of Beverage	Sugar (g)	Vitamin C (mg)	good/bad?
Coke	10	Zero ☺	bad
Orange juice	9	120	bad
Tap water	0	0 (traces!)	good

Features

Phases and Data Set Splits

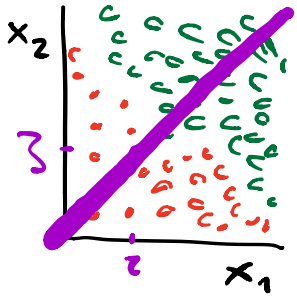


Why Validation?

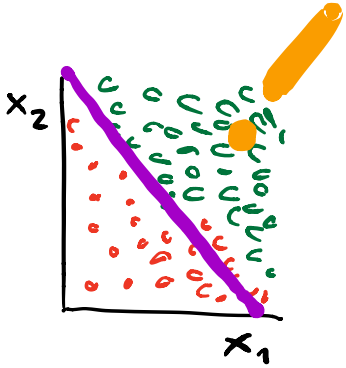
overfitting

fitting = training

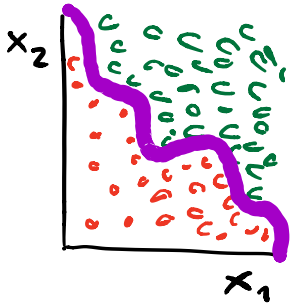
$\vec{w} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$



underfitting



fitting



overfitting

Evaluation

Generalization

Metrics

k = 3, 4, 13

Algorithm A 17, 18, 15

Alg. B 26

Classification Accuracy

Summary

- Machine learning typically requires data
- data collection can introduce biases
- data is typically split into:

train valid test

- evaluation tests generalization

University of Zurich[UZH]

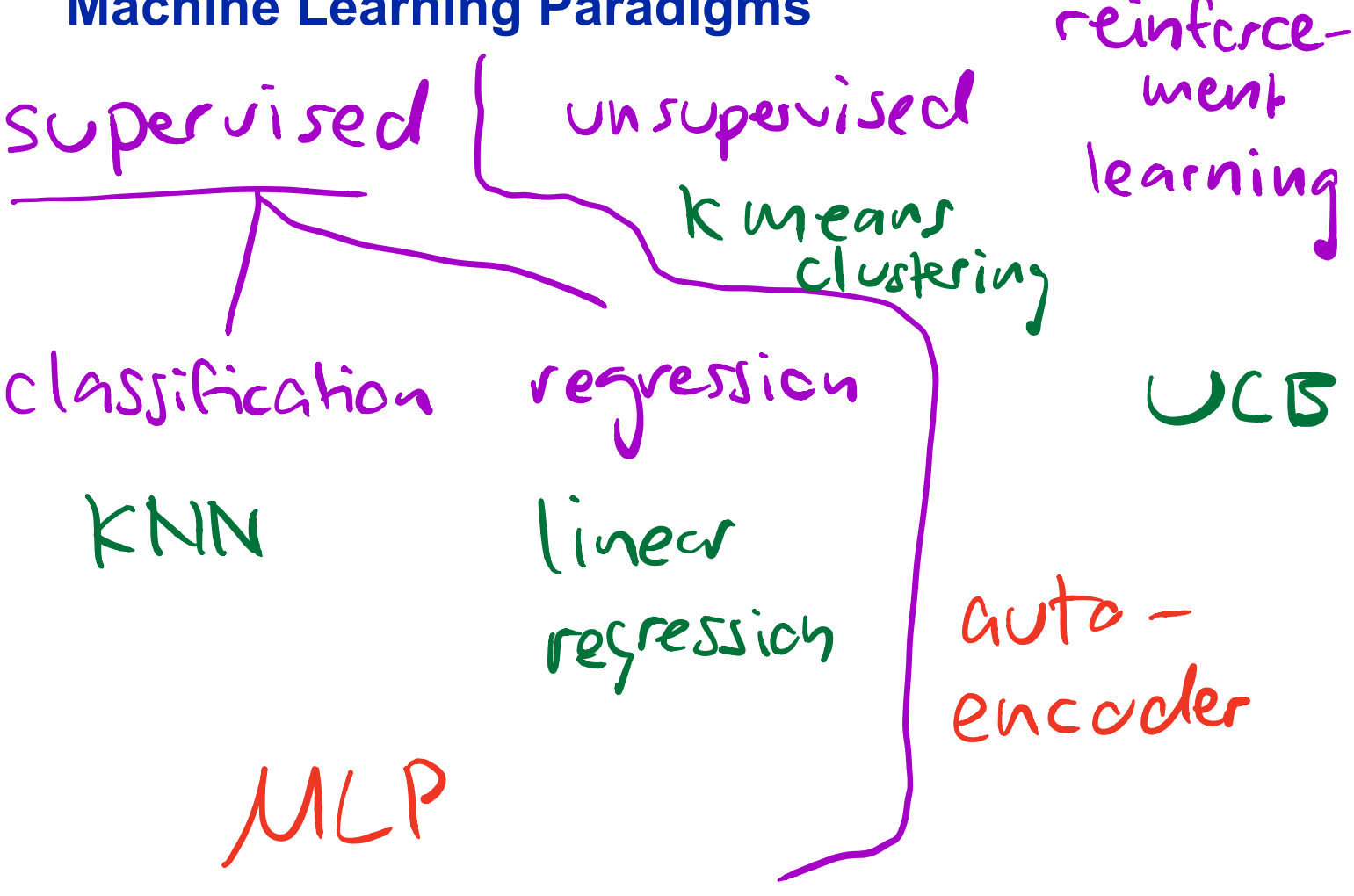
Institute of Computational Linguistics

Machine Learning Paradigms

Topic of this lesson

- major paradigms of machine learning
- supervised vs. unsupervised algorithms
- classification vs. regression algorithms

Machine Learning Paradigms



Supervised vs. Reinforcement Learning

e4 e5 → …

Bc4 Bc5 → …

Nc3 Nf6 → …

d3 Nc6 → ✓

Supervised vs. unsupervised: Demo

Unsupervised learning

clustering

Pleiaden

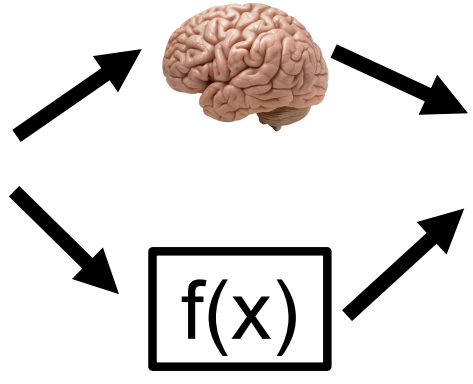
dimensionality reduction

$$\begin{bmatrix} 17 \\ 4 \\ 1 \\ 2 \end{bmatrix} \longrightarrow \begin{bmatrix} 3 \\ 7 \end{bmatrix}$$

Classification vs. regression

Classification ("group membership")

Name of Beverage	Sugar	Vitamin C
Coke	10	Zero ☺
Orange juice	9	120
Tap water	0	0 (traces!)

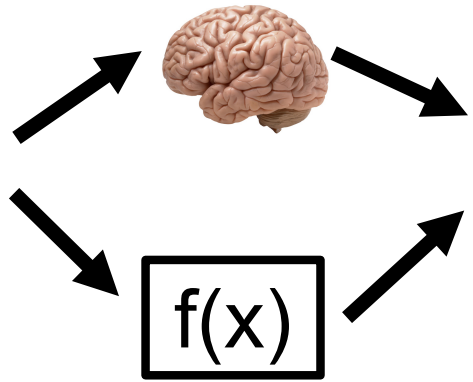


good/bad?
bad
bad
good

2

Regression (value prediction)

Name of Beverage	Sugar	Vitamin C
Coke	10	Zero ☺
Orange juice	9	120
Tap water	0	0 (traces!)



body temperature
40
39
37

Where does Deep Learning fit in?

Summary

- major paradigms are:

supervised unsup RL

- main difference between supervised and unsupervised methods:

training data incl. labels

- main difference between classification and regression:

fixed list of classes

classification regression

KNN