an introduction to Machine Learning

Patrick van der Smagt

http://ml1.brml.org/

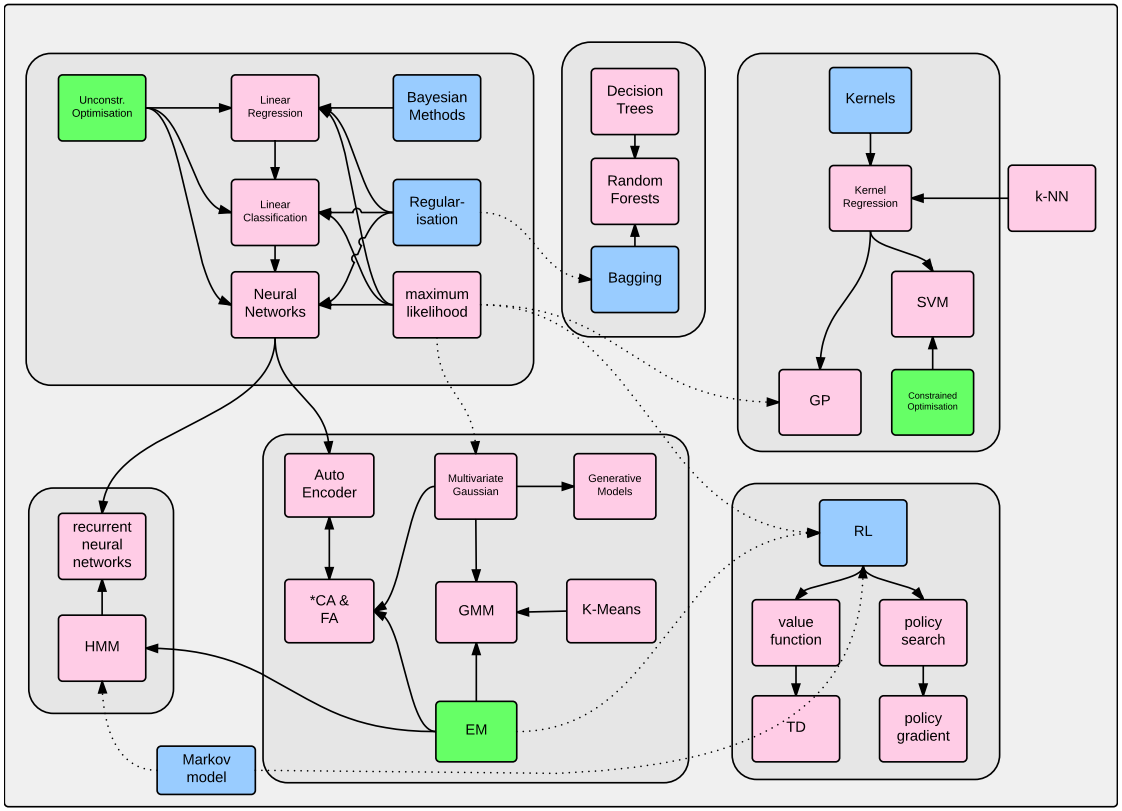
register on piazza now

or at least write down the URL and password to register later

`https://piazza.com/tum.de/fall2016/in2064`

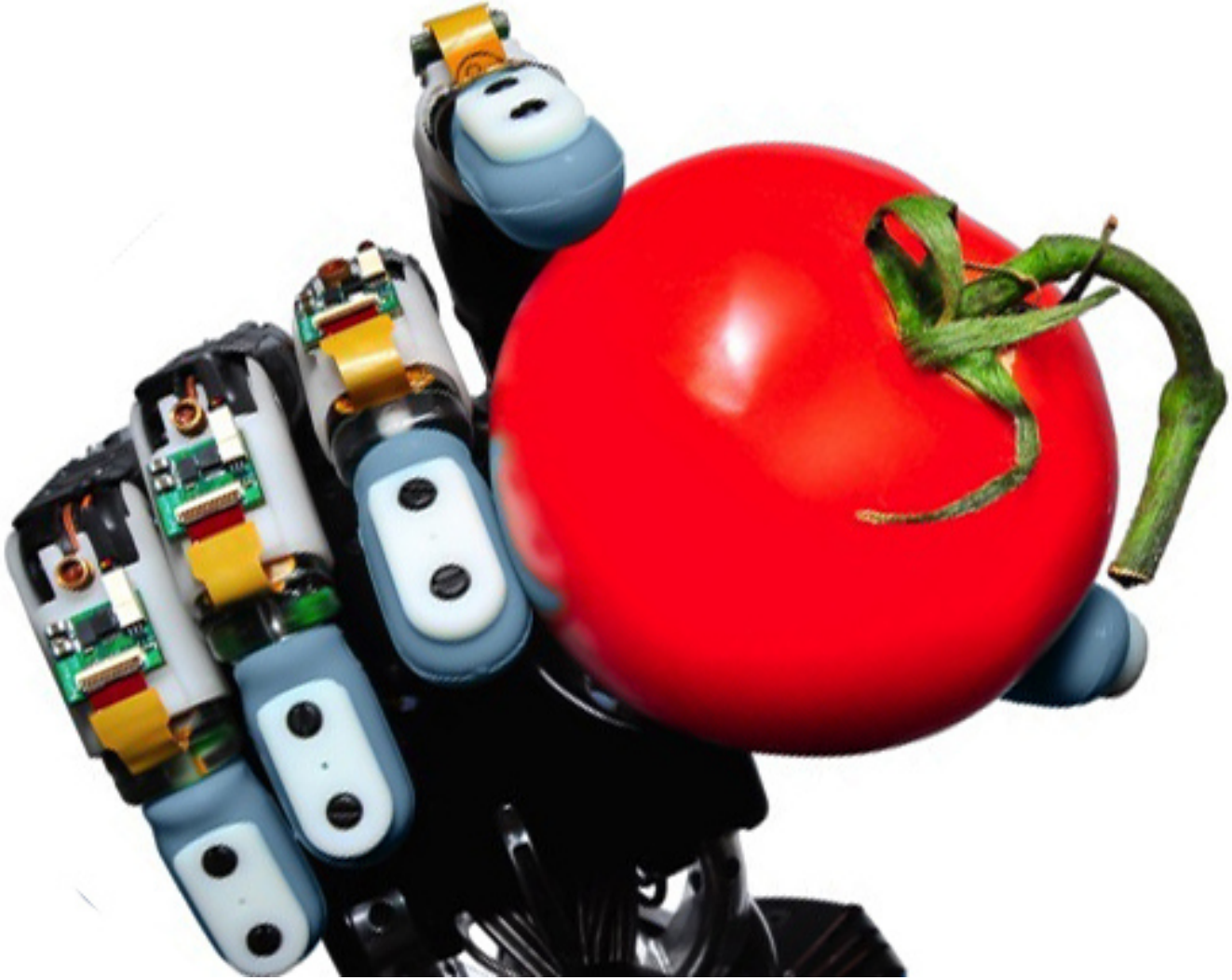
with password `p(z|x)`

what you will learn

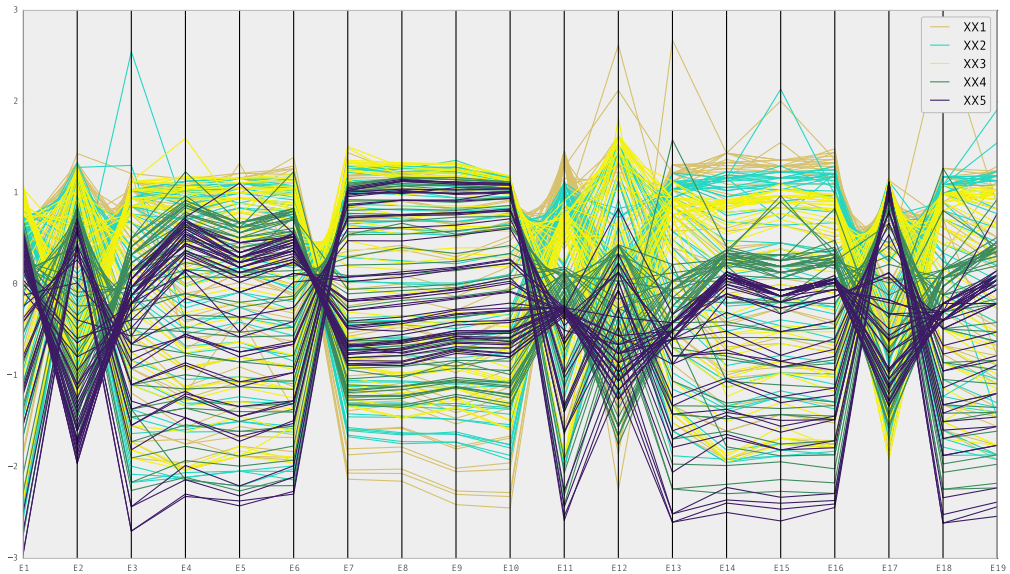


1. what?

how do we manipulate objects?



a recording of a high-density tactile sensor



many recordings of a high-density tactile sensor

. . . sensing different materials

what is machine learning?

suppose you measure some features $\mathbf{x}_i$

each feature has a target output z_i

we want to find a model $y(\mathbf{x}, \mathbf{w})$ that, for $i \in \{1, 2, \ldots, n\}$

maximises $p(z_i \mid \mathbf{x}_i, \mathbf{w}) = \mathcal{N}(z_i \mid y(\mathbf{x}_i, \mathbf{w}), \beta^{-1})$

what is supervised learning?

"learning from examples"

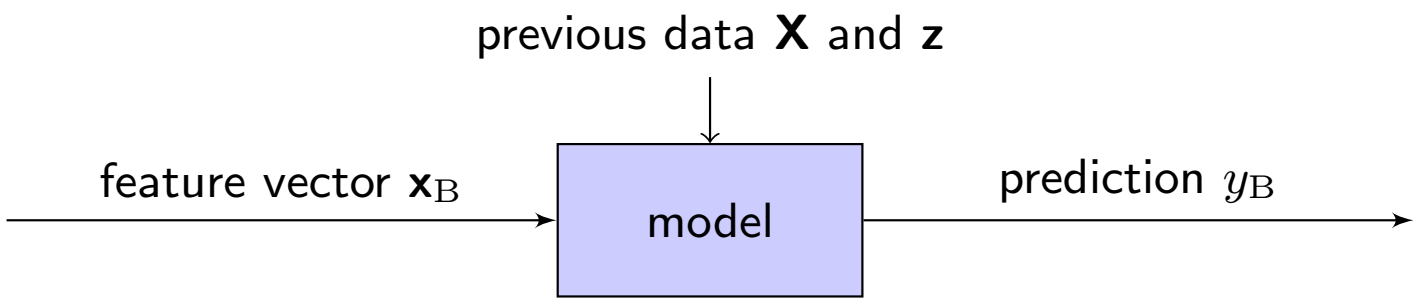
example: travel insurance

Bob applies for travel insurance. You want to insure him only if you expect his premium to be higher than his bills.

You have previous data $\mathbf{X}$ from other insurees with yearly costs $\mathbf{z}$.

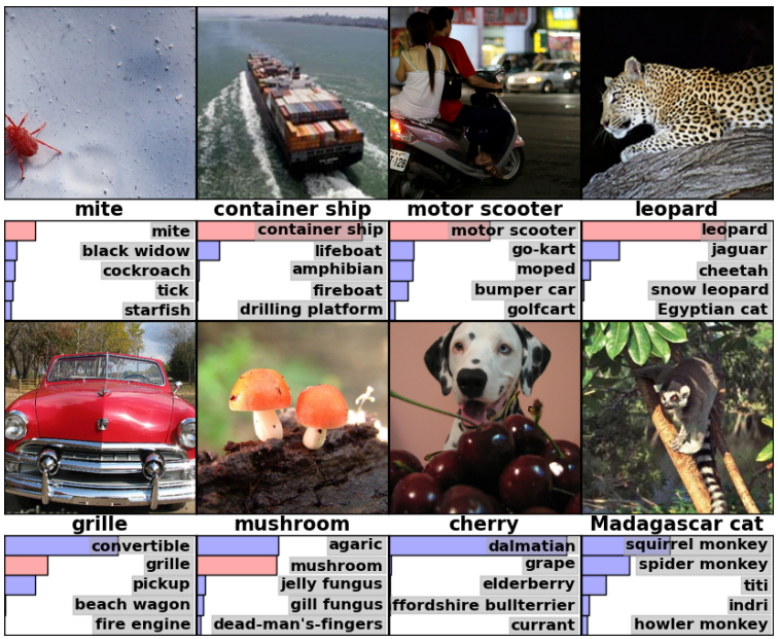
input $\mathbf{x}_i$:	sex, age, marital status, employment status, monthly income, family structure, ...
output z_i:	yearly cost

Find a model that you can give Bob's data to and that will predict how much he will cost you every year.

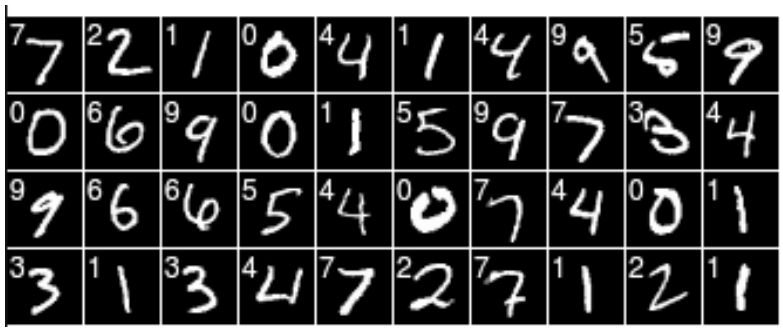


other examples of supervised learning

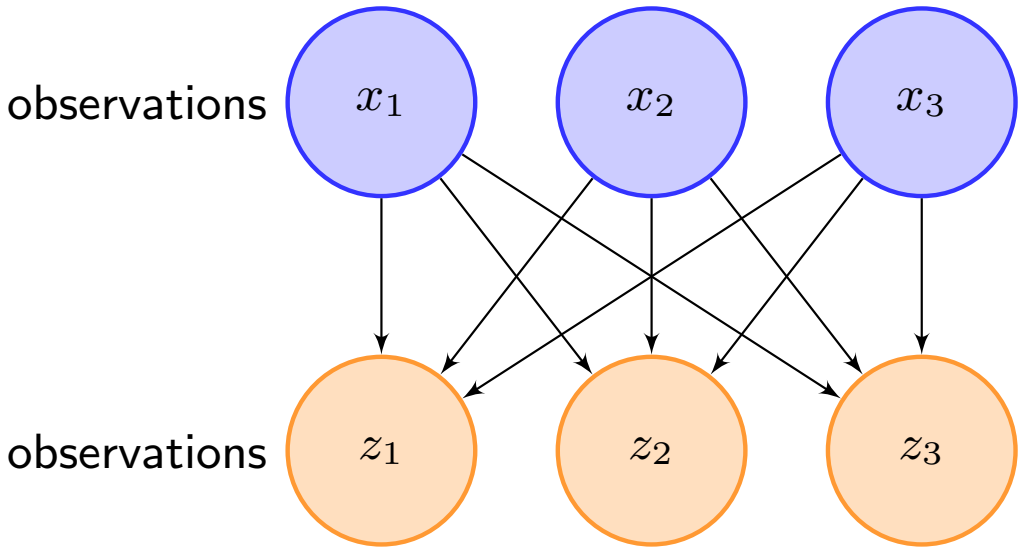
object recognition



handwritten digit recognition

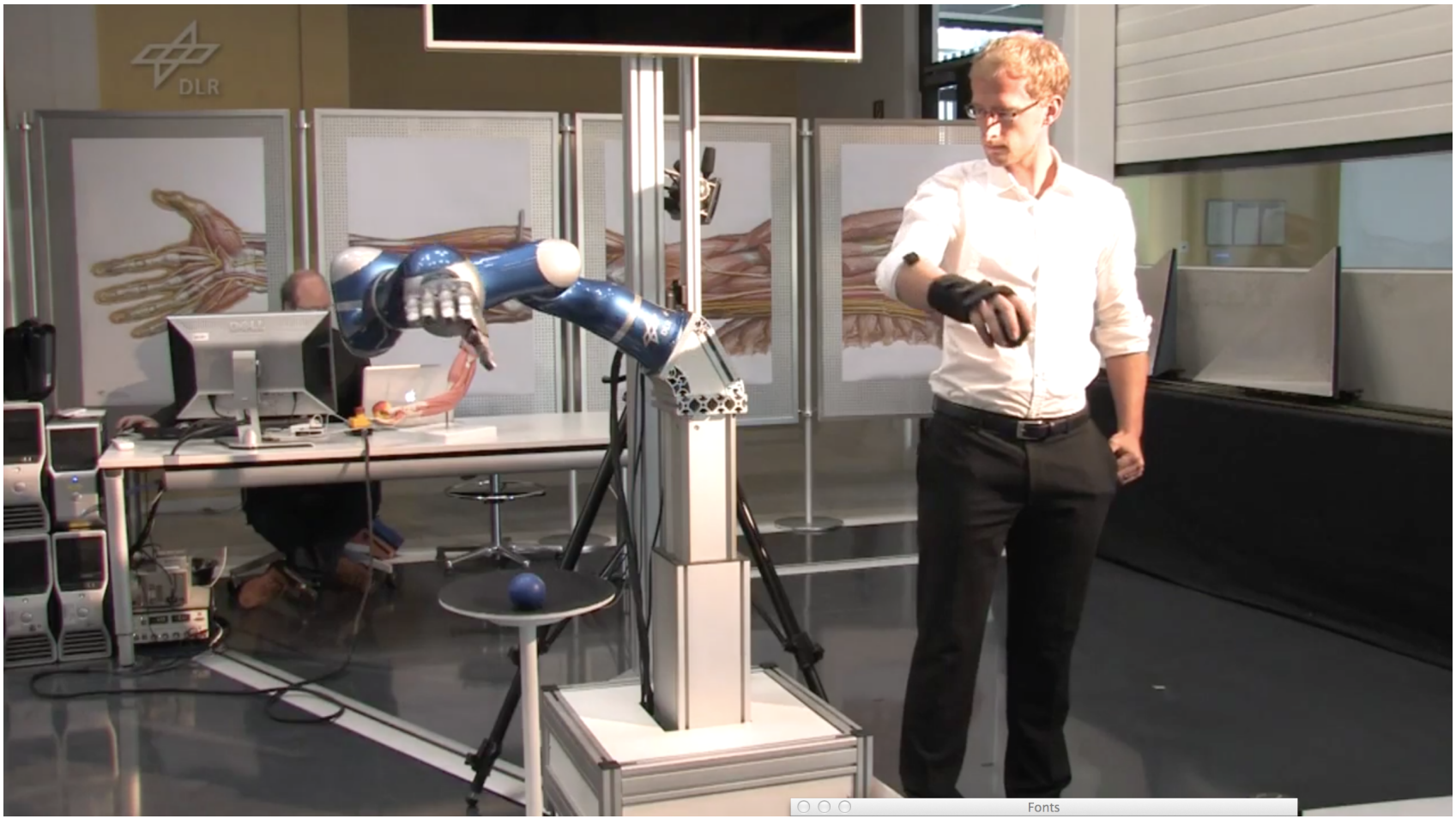


supervised learning



Both input $\mathbf{x}$ and target $\mathbf{z} = f(\mathbf{x})$ are known for some examples.
The function $f(\mathbf{x})$ is unknown, but we can sample from it.

often we have data with ground truth

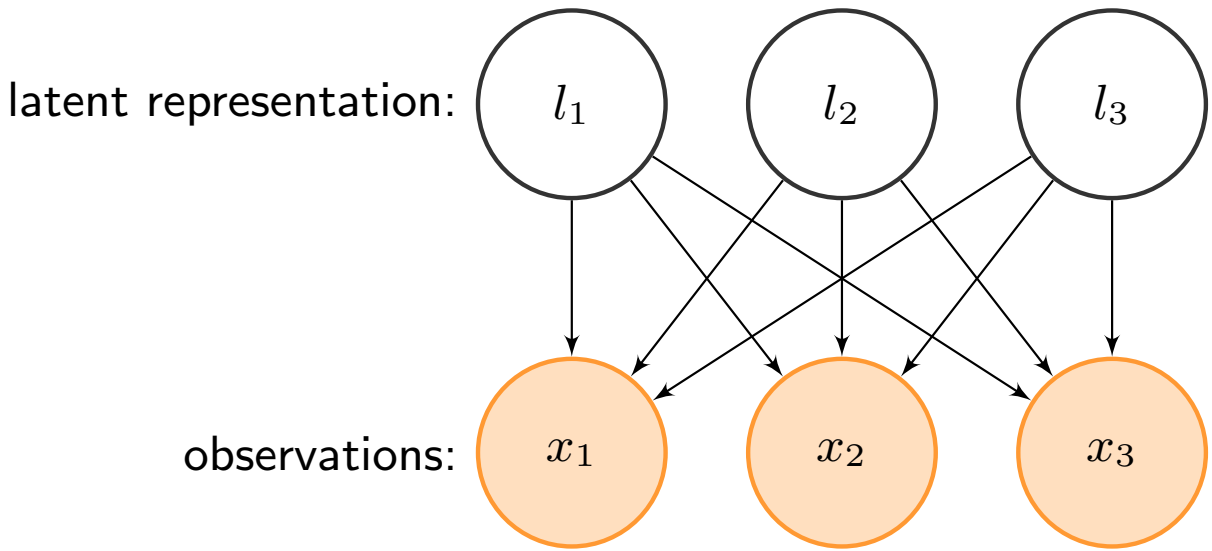


sometimes we have data without ground truth

Unsupervised Learning

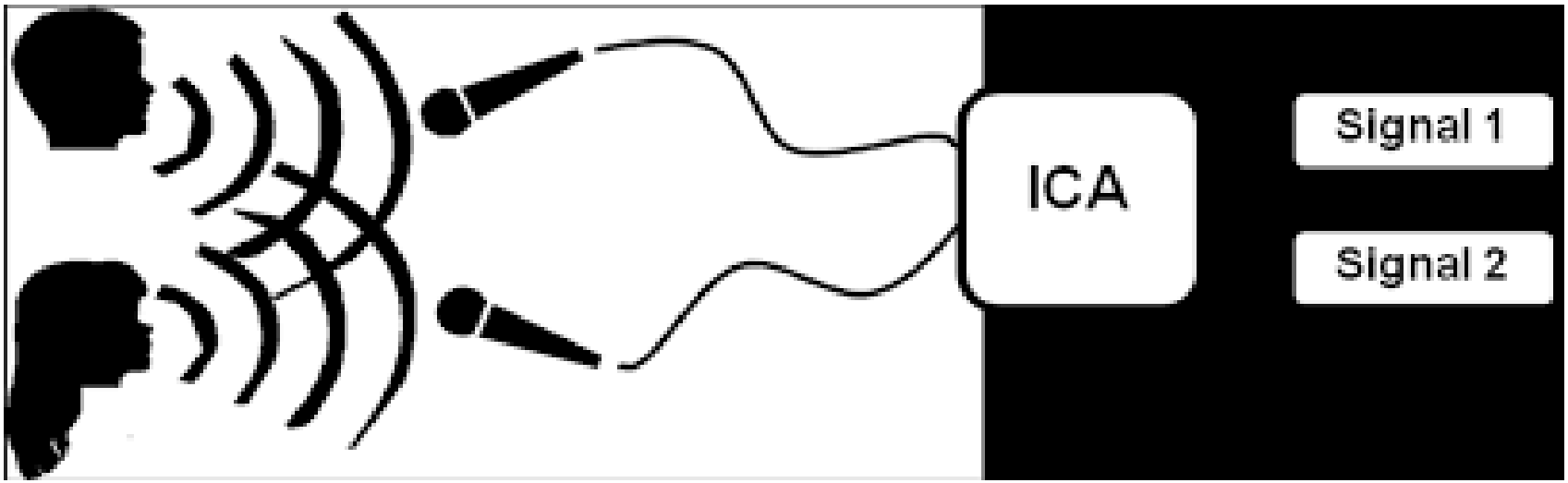
In unsupervised learning we assume that our observations **X** are driven by an unknown process.

We have no ground truth for the latent representation.

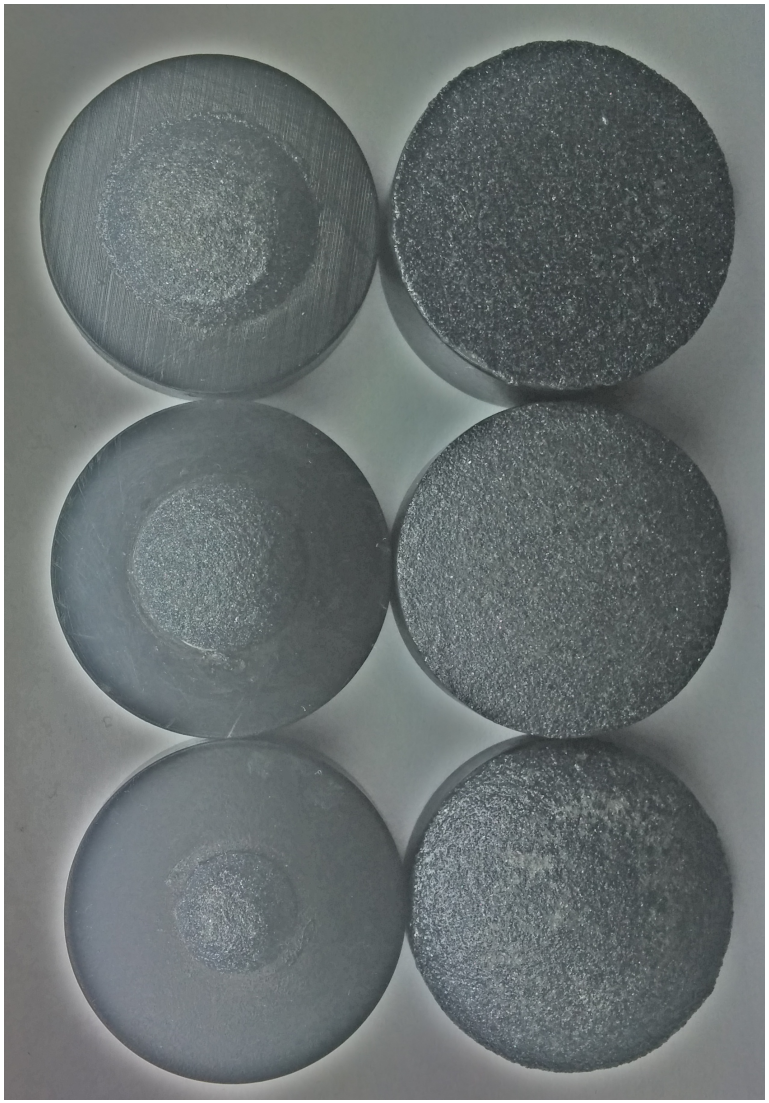


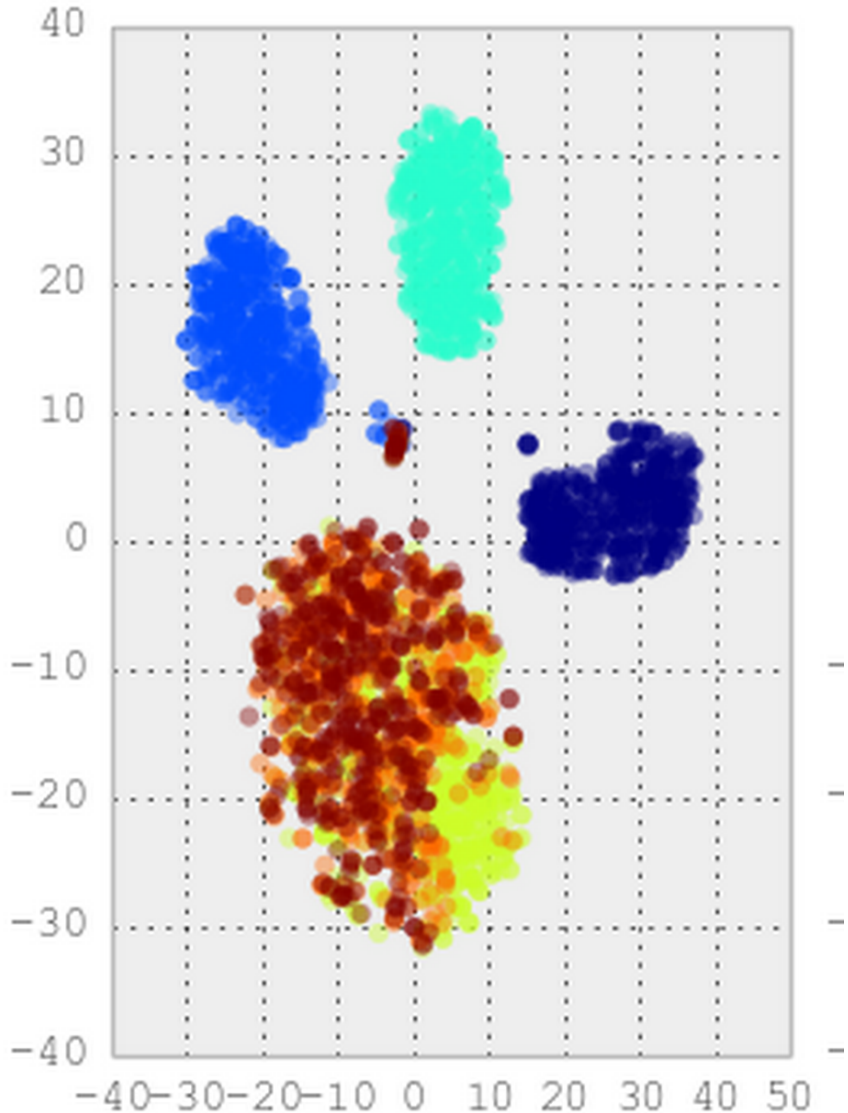
examples of unsupervised learning

source separation



clustering of tactile curvature measurements





is that all?

what if. . . there is a delayed reward?

1. what is the optimal market price of a car (GM use[ds] this) to maximise revenue?
2. where should the elevator park to minimise waiting time?
3. helicopter control
4. . . .

Reinforcement learning

A system in state s_i has to select an optimal action a_i. After performing action a_i the environment gives bac a reward r_i.

In many cases, r is only known after a whole sequence of actions (did my helicopter crash?)

The goal of reinforcement learning is to find a policy which maximises the future reward

$$E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots] \quad \text{with } 0 \leq \gamma \leq 1$$

machine learning vs. statistics

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = \$1,000,000	large grant= \$50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

2. how?

1st generation machine intelligence: expert systems (1960–1980s)

> *An expert system is a computer system that emulates, or acts in all respects, with the decision-making capabilities of a human expert.*
>
> Edward Feigenbaum, Stanford University

- major components: knowledge base, inference engine, and user interface
- knowledge is encoded as IF . . . THEN rules

Problem:

- combinatorial explosion of rules
- rules are made by hand from human experts
- suffer from the challenge of uncertainty

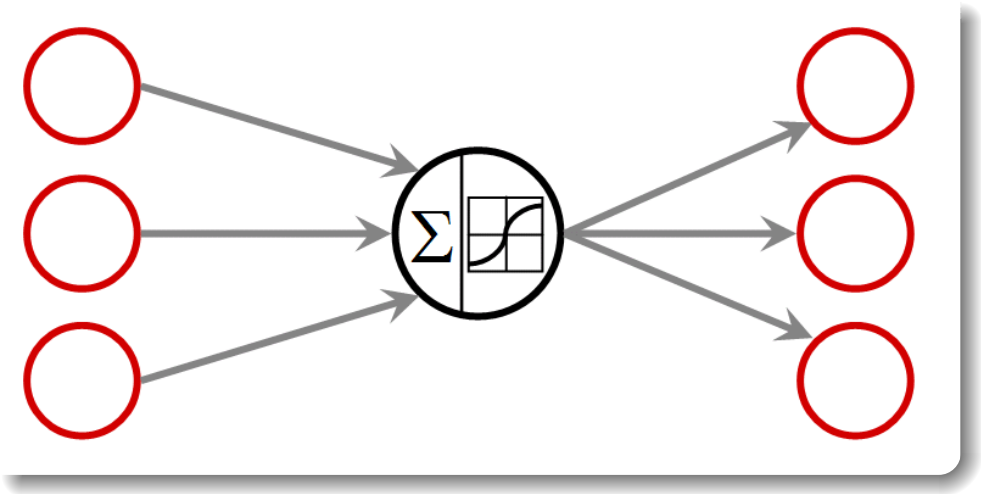
—it does not generalise.

2nd generation machine intelligence: neural networks (1980s–1990s)

- learn from data only
- can (theoretically) represent any kind of data

Simulated Neural Networks

a simulated neural network contains. . .



neurons sum up inputs and put them through a transfer function.

connections are represented by real number weights, which modify the signals.

networks are created such that a structured system arises (often layered).

1949: Hebb learning rule

basic biological synapse learning rule, formulated by Hebb in 1949:

$$\Delta w_{ij} = \alpha v_i v_j$$

in words, "cells that fire together, wire together."

1957: Rosenblatt's perceptron

in Rosenblatt's perceptron[1], the class of the output is one when

$$y = \phi(b + \mathbf{x}^T \mathbf{w})$$

with

$$\phi(x) = \begin{cases} +1 & \text{if } x > 0 \\ -1 & \text{otherwise} \end{cases}$$

perceptron learning rule:

1. choose a random $\mathbf{w}_0$
2. select an input vector $(\mathbf{x}_i, c_i)$ with $c_i \in \{-1, 1\}$
3. if $y \neq c_i$ then $\mathbf{w}_{i+1} = \mathbf{w}_i + c_i \mathbf{x}_i$
4. go back to 2.

This is modelled after the Hebb rule, but here no weights are changed when the classification is correct. Convergence can be proven, but there are better methods.

[1] Bishop, pp. 192

1960: Widrow and Hoff's delta rule

they introduced the *adaline* (ADAptive LInear NEuron; later: ADAptive LINear Element), very similar to the perceptron, but:

- the nonlinear transfer function ϕ is the identity function:

$$y(\mathbf{x}) = \sum_i w_i x_i$$

- learning is done with the delta rule

delta rule: (aka Widrow-Hoff rule)

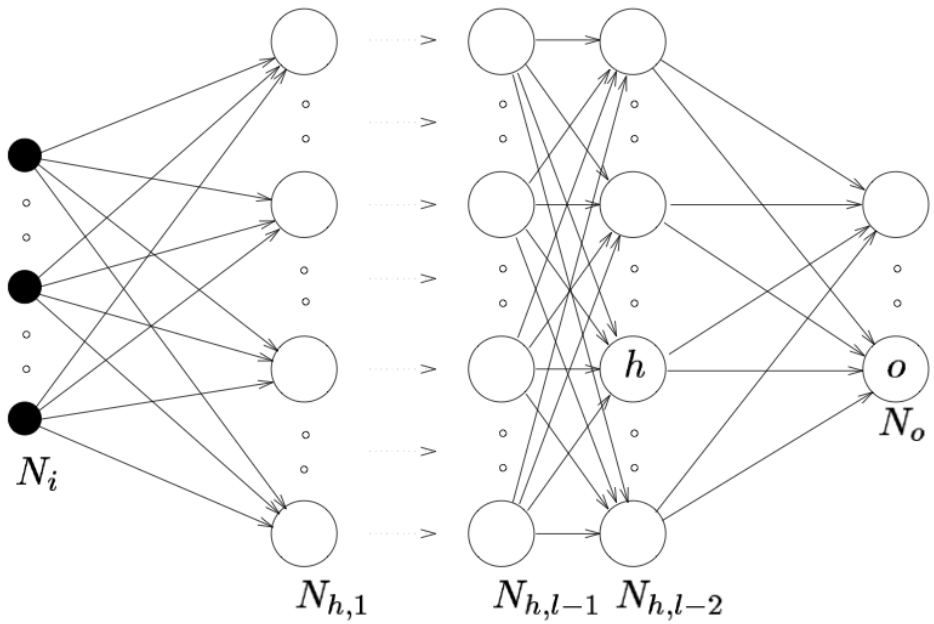
1. choose a random $\mathbf{w}_0$
2. select a date sample $(\mathbf{x}_i, z_i = F(\mathbf{x}_i))$
3. $\mathbf{w}_{i+1} = \mathbf{w}_i + \alpha[y_i - z_i]\mathbf{x}_i$
4. go back to 2.

we know α as the learning rate

the 1969 crisis: Minsky & Papert

in 1969, Minsky and Papert showed that the adaline could only separate linearly. As an example, the XOR problem was given: it cannot be solved with linear classification.

they also showed a solution to the problem: multi-layered perceptrons can solve that. However, a general rule to optimally find the weights **w** was not discovered until 1974 (Paul Werbos) or 1985 (LeCun) and 1986 (Rumelhart *et al.*): back propagation.

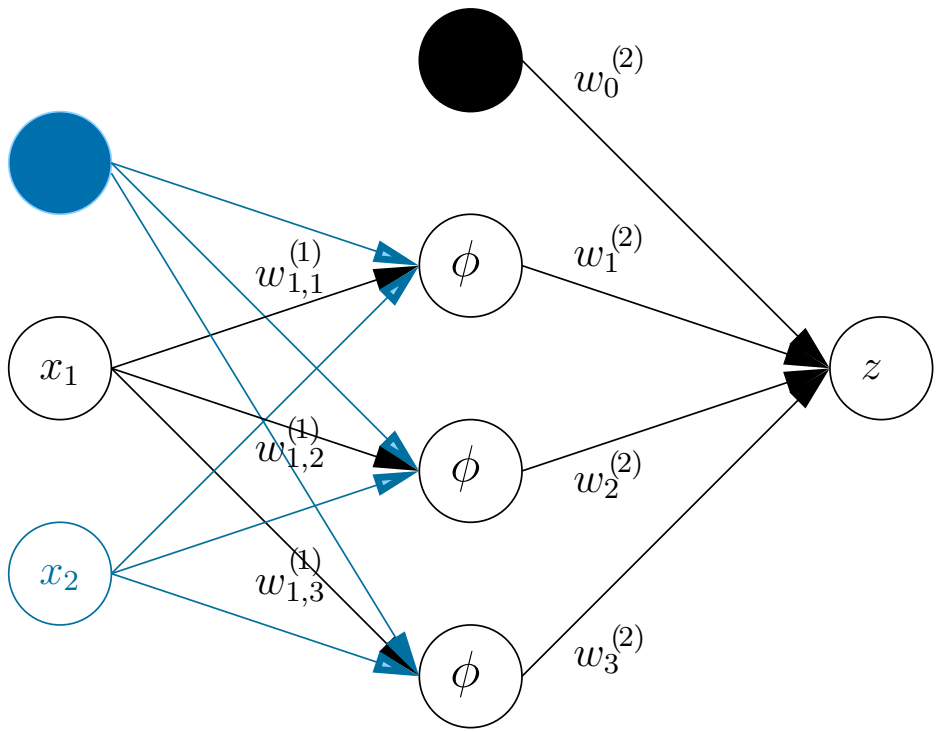


the multi-layered perceptron

we can compactly write the above network with *one hidden layer* as

$$y(\mathbf{w}, \mathbf{x}) = \mathbf{w}^{(2)} \cdot \phi(\mathbf{w}^{(1)}\mathbf{x})$$

where $\mathbf{w} = (\mathbf{w}^{(1)}, \mathbf{w}^{(2)})$



the neural network saga

1960's:	linear perceptron (Rosenblatt et al)
1969:	the perceptron cannot do XOR (Minksy & Papert)
1970's–80's:	nonlinear neural networks with back-propagation (Linnainmaa; Dreyfus; Werbos; Rumelhart)
early 1990s:	one hidden layer suffices to represent any (Borel-measurable) function
mid 1990's:	neural networks can't do everything / don't generalise / too slow
mid 1990's:	Enter SVM
1995–2000:	SVMs are too expensive / slow because of having too many Support Vectors

the neural network saga

- 2000–: probabilistic models for machine learning
- 2006: deep neural networks, trained with RBM + back-propagation
- 2009: deep NNs can also be trained by good computing power (GPUs) and having many data
- 2012: dropout prevents overfitting
- 2012–: cNNs beat many vision benchmarks
- 2012–: recurrent neural networks increasingly stable
- 2013–: probabilistic neural networks (variance propagation; variational autoencoder; ...)
- 2014–: end-to-end learning applications in robotics
- 2015–: key players variational inference; deep reinforcement learning

3rd generation: modern machine learning

Bayesian inference!

- integrate statistical domain knowledge with learning

and that what's this course is about.

list of topics

1. history, intro, basic concepts
2. probability theory
3. trees, k-Nearest Neighbour, entropy
4. parameter inference
5. multivariate Gaussian
6. linear regression and kernels
7. linear classification
8. neural networks
9. neural networks and optimisation
10. Gaussian Processes (GP)
11. unsupervised learning
12. variational inference and Expectation Maximisation (EM)

You'll see different teachers (from my lab: Wiebke, Nutan, Max, Max, Grady) and the tutor, Georg Groh. Don't expect me every week.

recommended reading

Our official reading recommendation:

- Christopher M. Bishop, *Pattern Recognition and Machine Learning*. Springer, Berlin, New York, 2006.

but we also like:

- Kevin Murphy, *Machine Learning: A probabilistic perspective*. MIT Press, 2012.

Other recommended books:

- David Barber, *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012 (free, online version available).
- David J. C. MacKay, *Information theory, inference, and learning algorithms*. Cambridge University Press, 2008 (free, online version available).

organisational stuff

class hours:

Mon, 11:59: homework due

Tue, 10:15–13:00, MW 0001: flipped classroom lecture

Wed, 16:15–19:00, Interims HS 2: tutorial

Thu, 14:00–16:00, IN 02.13.10: small homework tutorial

Thu: new slides, video, and homework on piazza

On Wednesday tutorial lectures, additional exercises are solved. Homework solutions are provided online; in addition, a small tutorial on Thursdays discusses homework, too. The latter is intended for students who desire additional homework feedback.

re homework

Exercises are handed out on Piazza on Thursdays of week n.
The exercises have two parts:

- homework part 1: short questions related to the lecture video, which must be answered by 11:59am, Monday week $n + 1$ by filling them in at `http://ml1.brml.org`
- homework part 2: due 11:59am by Monday week $n + 2$: homework which should be uploaded as PDF as typeset document at `http://ml1.brml.org`

Some of the exercises may be discussed during tutor meetings.

You **must** hand in your homework yourself, but can solve the tasks in groups of two (, . . .). If you do, please indicate this on the submitted PDF.

Grading Schema

The course has a closed-book written final exam at the end of the semester (duration 120 minutes), giving grade X. You are allowed to bring one A4-sheet of handwritten notes.

IF

- at least 66% of the homework part 1 has been graded as "sufficient attempt", AND
- at least 66% of the homework part 2 has been graded as "sufficient attempt", AND
- you passed the final exam

the final grade will be $\min(X, X - 0.3)$.

Homework is voluntary and can only improve the grade of the final exam. Doing homework prepares you optimally for the exam.

We *strongly* suggest you do the homework.

background information

The *official* webpage of the class is at http://piazza.com/ and join the TUM class IN2064 there.
There is a secret code to get in. . . that code is `p(z|x)`.

All announcements and handouts will be distributed via the piazza site. So if you do not join, you may miss some essential information.

flipped classroom

Class topics are available as video; links will be posted to piazza. (Nearly) each topic consists of a video, tutorial, a class, and homework (in that order). The video gives all information, and the class is there to answer your questions. Please study as follows:

- on Friday, prior to the class, download and study the video and the related stuff in Bishop; download the slides and study them; download the homework.
- by Monday 11:59am, submit the homework of last week's class, and the simple homework of this week's class. Also, post questions related to that week's lecture on piazza, so that your teacher can prepare.
- on Wednesday, go to the tutorial to play with the material.

the symbols that we will use

$\mathcal{V}$	a calligraphic symbol typically denotes a set of random variables
$p(x, y)$	joint probability of x and y
$p(x\|y)$	probability of x conditioned on y
x	input
y	actual/computed output
z	desired/target output (note: Bishop uses t)
$\mathcal{D}$	data set $\{(x_i, z_i)\}$
n	data index
N	number of data set training points
$\mathcal{N}$	Gaussian
$\langle\cdot\rangle$	mean
$\sigma(x)$	the logistic sigmoid $1/(1+\exp(-x))$
$x_{a:b}$	$x_a, x_{a+1}, \ldots, x_b$

linear algebra and calculus

During the rest of the week, make sure that you are familiar enough with linear algebra and calculus. Best watch our videos:

- linear algebra `https://youtu.be/a8Zf9o09R-I`
- calculus `https://youtu.be/zDrVZRLdACY`

We also like this book `http://www.janmagnus.nl/misc/mdc2007-3rdedition` on linear algebra and calculus.

— *there is no homework related to these topics* —

next week: Probability Theory

- tomorrow: NO tutorium
- Thursday: material uploaded for Probability Theory. **watch the PT videos before the lecture!**
- Monday: simple homework due, related to the PT videos