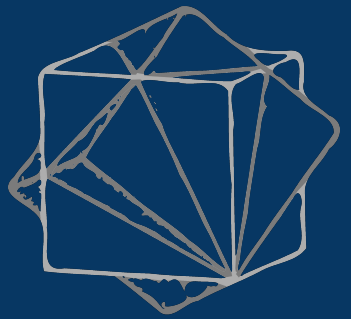
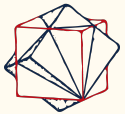


IntroML

29 Nov 2023

IntroML

Welcome



The ML ⇌Science Colaboratory

Plan for the day

Agenda

9:10	Round of introductions
Morning	**Core ML Concepts** **ML intro ⁄ learning paradigms ⁄ neural networks**
~12:30	**LUNCH BREAK**
Afternoon	**Machine Learning in Practice** **project pipeline: data, models & evaluation ⁄ ML in Science**
	Hands-on: Define your ML project **question ⁄ learning task ⁄ dataset ⁄ evaluation**
16:30	Closing – feedback survey

Practical aspects

- Pauses 2x15’ + 60’ lunch break
- Lunch options include Yellow Donkey and the Spice Kitchen food truck behind the Tübingen AI center
- Internet → MPI Lounge, share laptop if needed

?! Questions welcome at any time

Slides, links, notebook and feedback at https://mlcolab.org/public-events/introml-workshop-november-2023

Round of introductions

Tübingen is a great place to form a community of practice and mutual support around ML and science.

Let’s get to know each other! We can share

- Name & where you work
- Expectations from ML related to your research (1 sentence)

The ML ⇌ Science Colaboratory

At the Cluster ML: new perspectives in science

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Here is what ChatGPT thinks about us:

(Seth's Roast Master)

Álvaro Tejero-Cantero



Hanqi Zhou



Seth Axen



Vladimir Starostin



Here we have the modern-day Brady Bunch of tech startups:

Top left, he's the guy who looks like every generic football coach in a feel-good sports movie – you just know he's got a pep talk ready.

Top right, there's the coder who debugs software by day and moonlights as a philosopher, pondering over the existential crisis of why "Hello, World!" has to be the first program.

Bottom left, the one who's definitely the startup's 'idea guy', rocking that beard like he's about to sell you artisanal coffee or a bespoke blockchain.

And bottom right, the indie game developer who looks like he hasn't slept since that 'one last bug fix'... three days ago.

The ML ⇌ Science Colaboratory

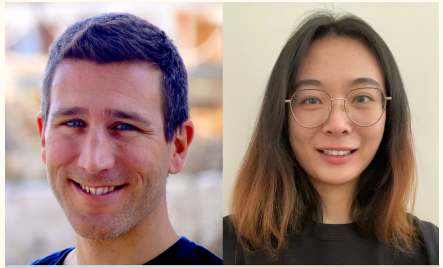
At the Cluster ML: new perspectives in science

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

«Establish machine learning across disciplines at the University of Tübingen»

via ① cooperations — ② training — ③ scientific ML software

Álvaro Tejero-Cantero



Hanqi Zhou

Seth Axen

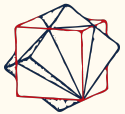


Vladimir Starostin

- **Part of the** Cluster ML in Science
- Scientists **like you**
  Astro, **atmospheric**, & **quantum physics**, environmental sciences, **neuroscience**, **genomics**, urban systems, int'l policy, **structural bio**…
- Let's work together!

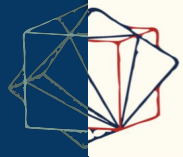
@mlcolab @mlcolab@fediscience.org

Core Concepts in Machine Learning



The ML ⇌ Science Colaboratory

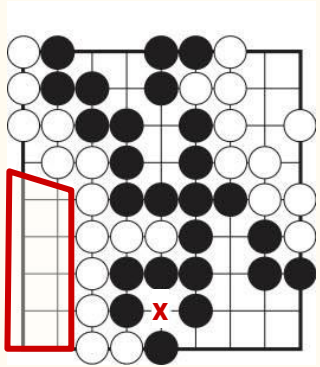
Core Concepts in Machine Learning



What is ML?

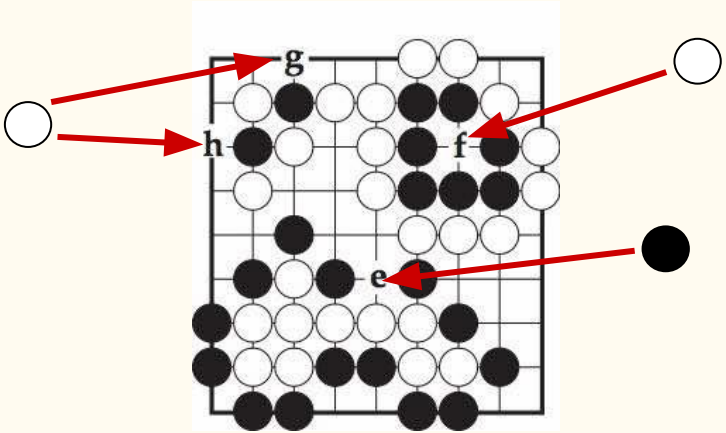
What problems does it solve?

Games and the nature of intelligence

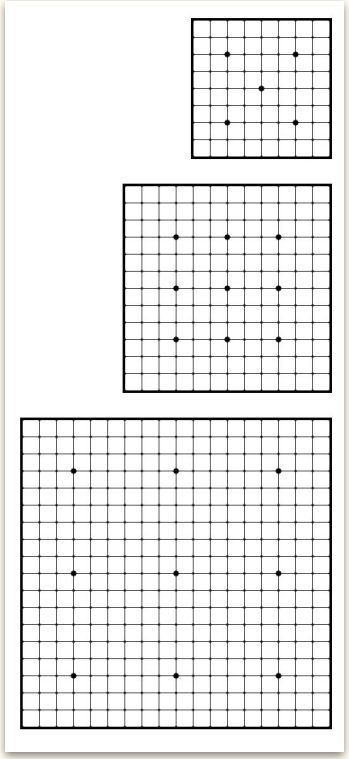


- form territories by surrounding empty areas
- capture by completely surrounding (*prisoners*) **x**

Score grows with crossing and prisoners



Real go is 19x19



British Go association

A surprising move



See also AlphaGo - The Movie

Definition of Machine Learning

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Learning

Improving with experience at some task

- Improve over **task** T,
- with respect to **performance** measure M (**metric**)
- based on **experience** D (**data**)

e.g., Learn to play Go

- T: playing Go,
- M: final score
- D: games from self-play

AlphaGo: Deep-Learning powered (ML) Monte-Carlo Tree *Search*.

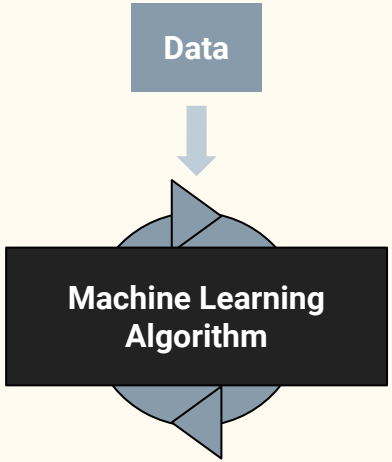
***Machine* learning**: the study of **algorithms** that allow computer programs to automatically improve **on a specific task** through **experience**

Mitchell, 1997

ML: Training and Use phases

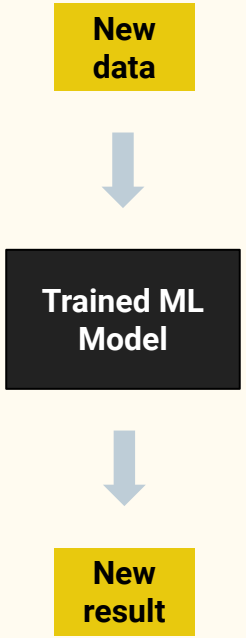
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Preparation: training



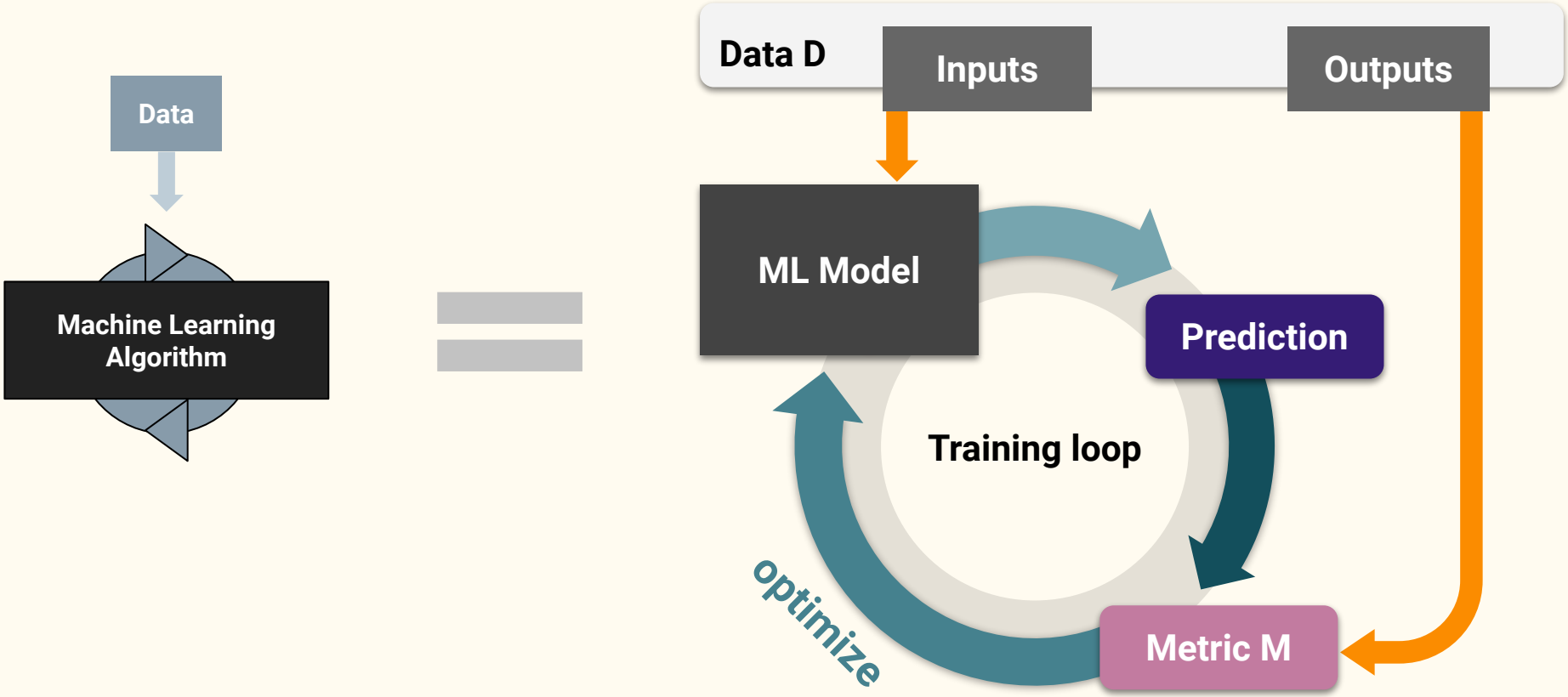


Use*

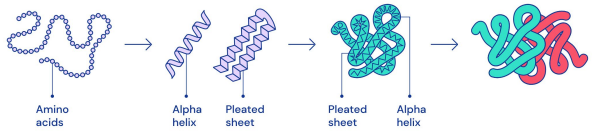


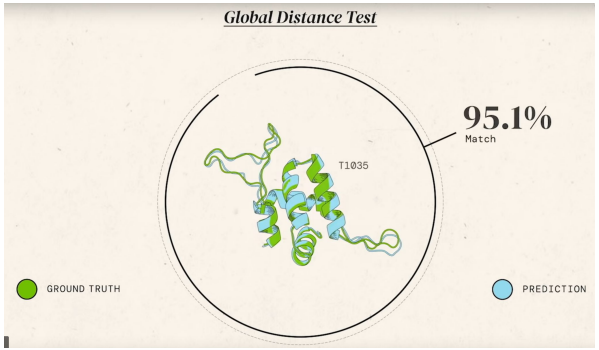
* also called prediction or “inference”

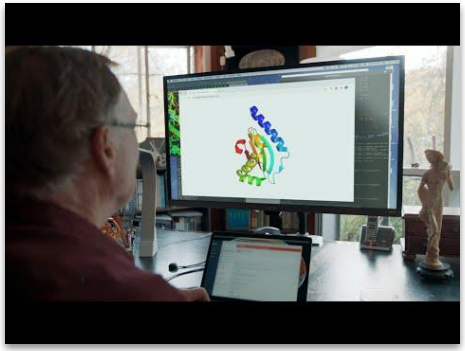
Inside the training black box



From proteins to planets, ML in science



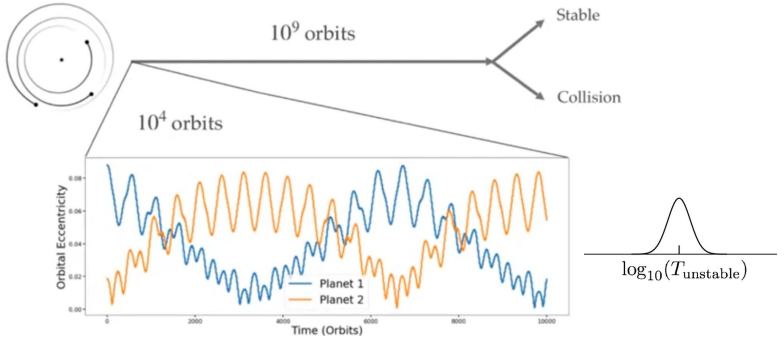




Protein structure prediction with transformers

Senior et al. 2020, Jumper et al. 2021

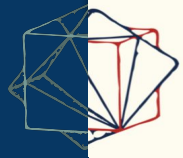




Planetary stability with Bayesian neural networks

Cranmer, Tamayo et al. 2021

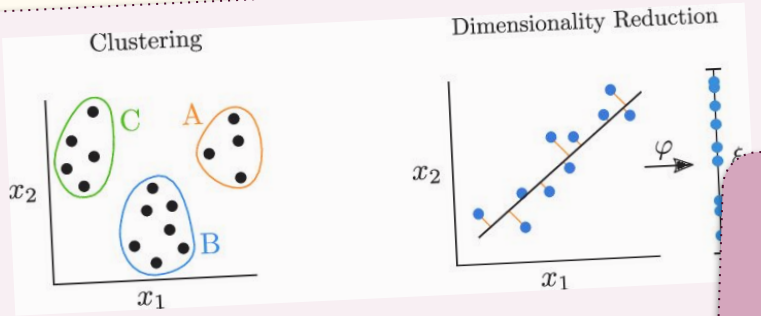
Core Concepts in Machine Learning



Learning paradigms

Main Learning Paradigms

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

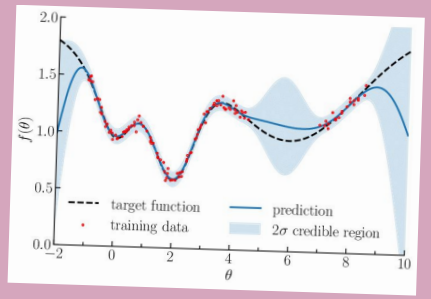


Unsupervised learning

Exploring structure

Supervised learning

Teaching to label



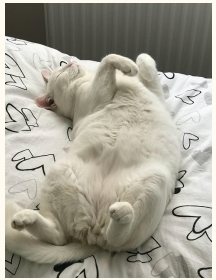
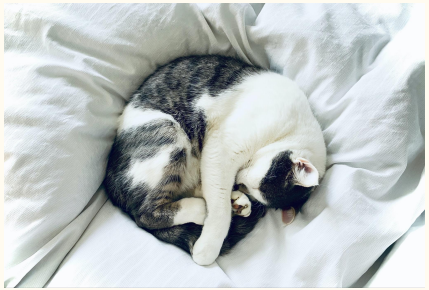
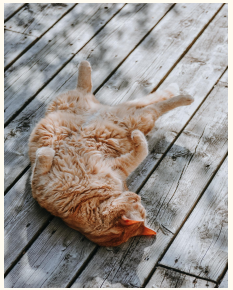
Reinforcement learning

Harnessing interaction



Figures DeepMind 2019, Leclerc 2018, jovian.ai 2021

An example dataset

Instance

ID	R01	G01	B01…
1	215	198	180 …
2	61	61	51 …
3	219	227	227 …
4	47	43	37 …

Feature

Photo by Sarah Ball on Unsplash

Photo by Gökhan Konyalı on Unsplash

Photo by Ben Wicks on Unsplash

Photo by Jacalyn Beales on Unsplash

Finding structure in the data



Photo by Sarah Ball on Unsplash

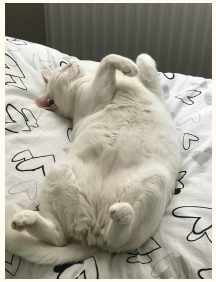


Photo by Gökhan Konyalı on Unsplash

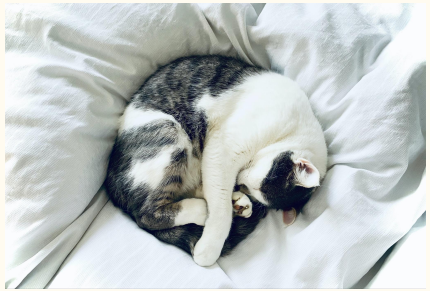


Photo by Ben Wicks on Unsplash

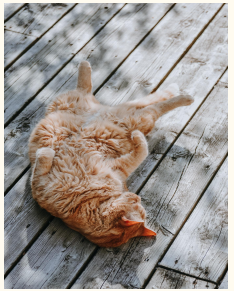
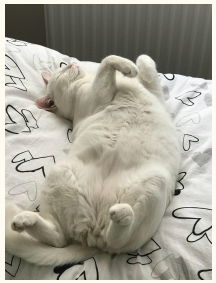


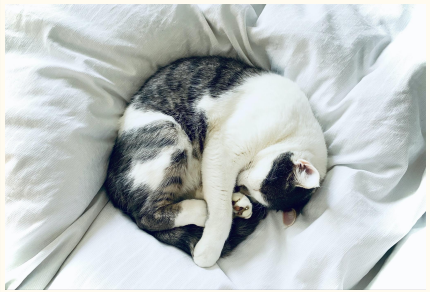
Photo by Jacalyn Beales on Unsplash

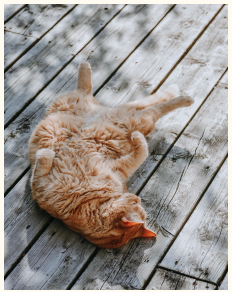
Task: group the images based on similarity

Finding structure in the data



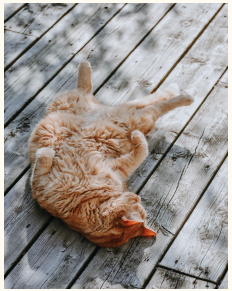
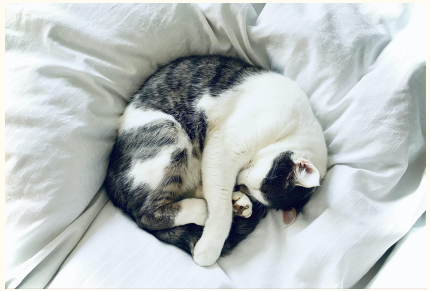
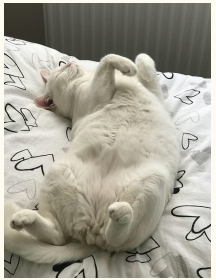






For example: 4 groups based on the breed of cat

Finding structure in the data

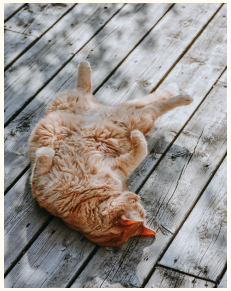





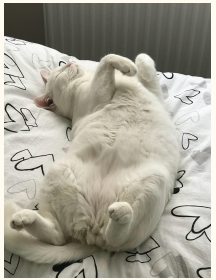
For example: 2 groups based on colors (orange or not orange)

Finding structure in the data









For example: 2 groups based on gesture

Unsupervised Learning

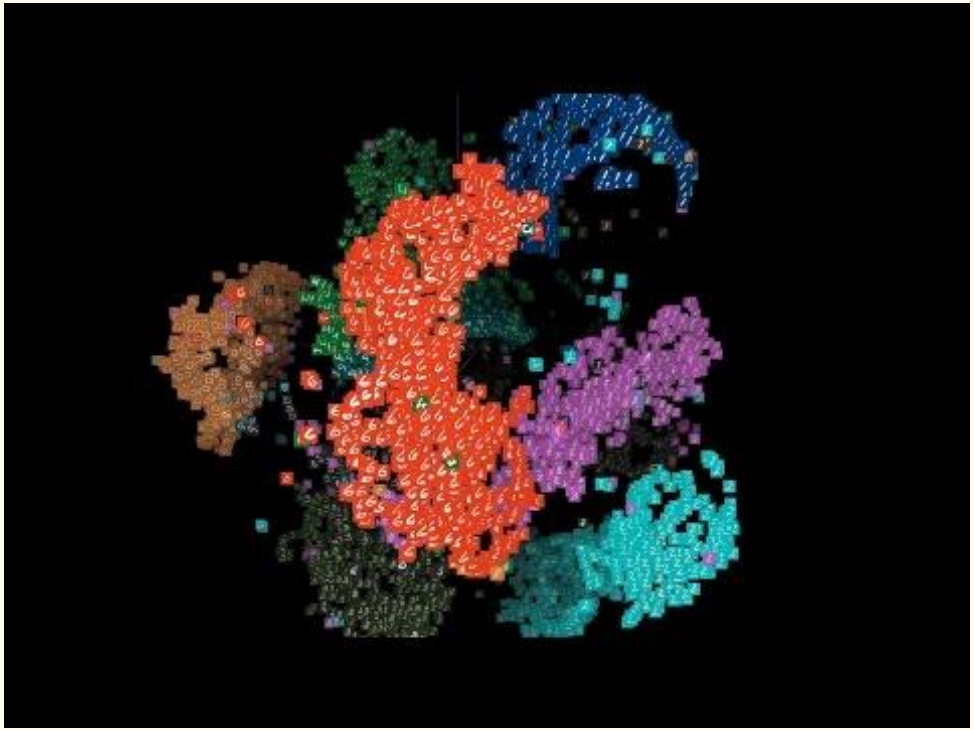
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Goals

- Structure discovery
- Dimensionality reduction
- Community detection

Requirements

- Data (unlabeled)
- Some notion of similarity of data points



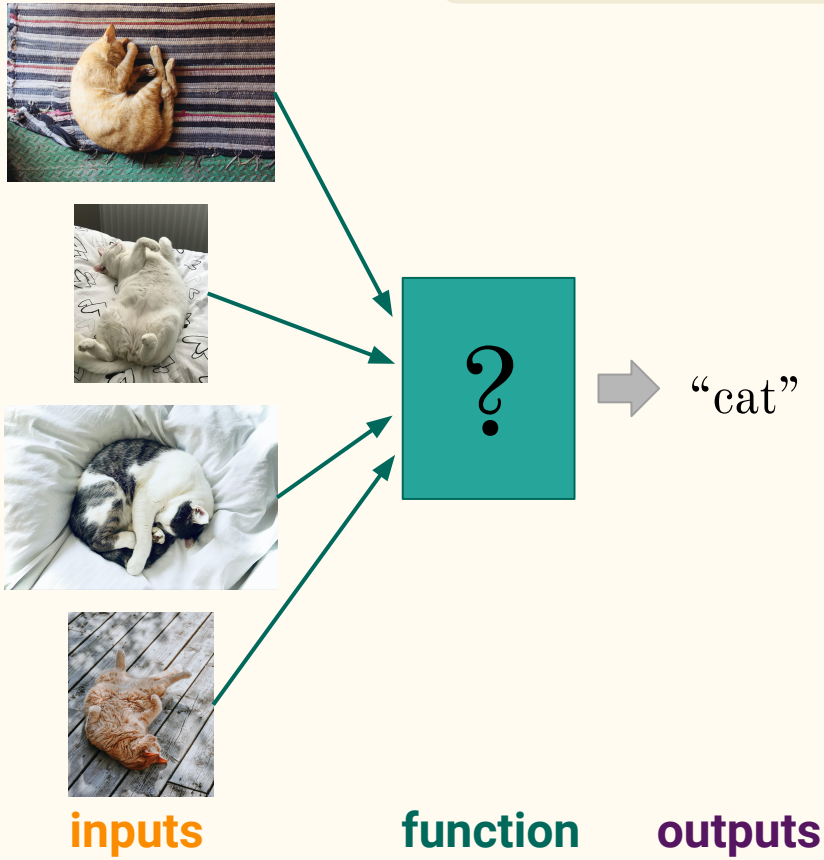
Supervised Learning

Goal
Learn a **function** to map **inputs** to **outputs**

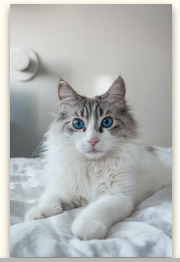
Requirement
Dataset of inputs *and outputs* (“labels”)

Modes

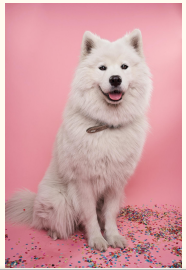
- Classification (category outputs)
- Regression (continuous outputs)



Supervised Learning: Classification



Cat



Dog



Cat



Dog

Category Label

ID	R01	G01	B01…	Type
1	184	187	187 …	Cat
2	218	157	164 …	Dog
3	61	61	51 …	Cat
4	236	236	236 …	Dog

Photo by Esteban Chinchilla on Unsplash Photo by Peri Stojnic on Unsplash Photo by Gökhan Konyalı on Unsplash Photo by FLOUFFY on Unsplash

Supervised Learning: Regression



47



0



2



-33

Continuous Label

ID	R01	G01	B01…	Head pose (degree)
1	116	98	81…	47
2	157	146	164 …	0
3	23	22	21…	2
4	137	130	129 …	-33

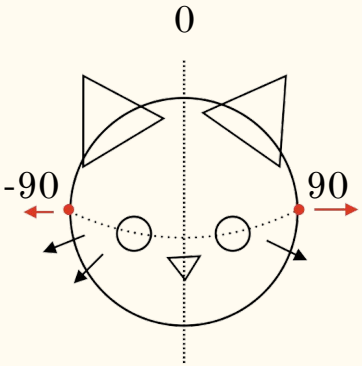


Photo by: cat-breeds-classifier

Reinforcement Learning

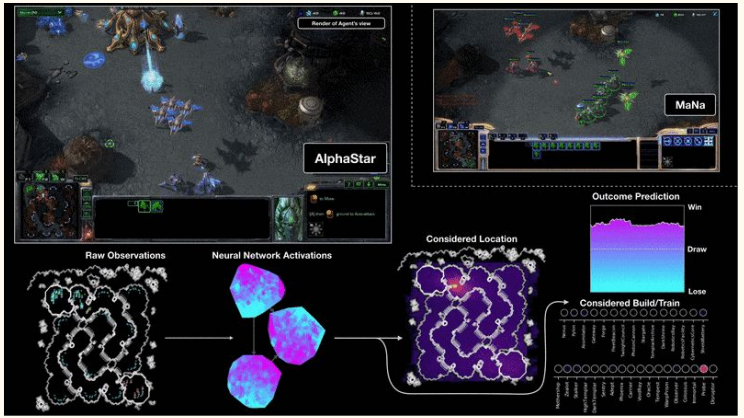
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Goal

Train an agent to interact successfully with some environment

Requirements

- Agent interacting with an environment (real or simulated)
- Reward mechanism for actions taken



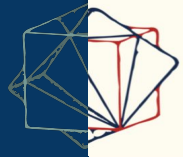
Reinforcement Learning



Reinforcement Learning



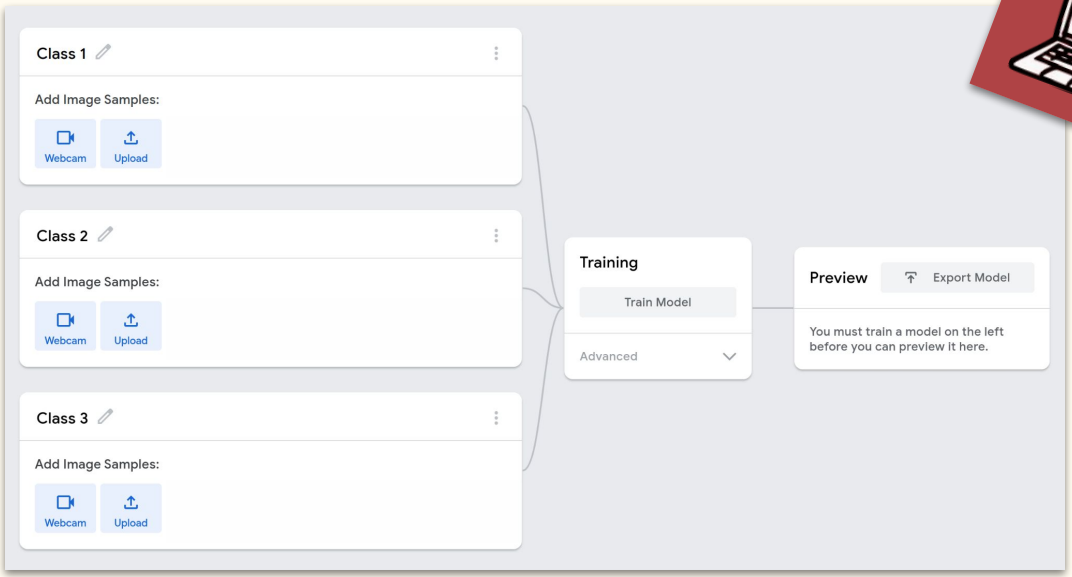
Core Concepts in Machine Learning



Supervised learning

One step at a time

Let's train a supervised classifier





Navigate to:
https://teachablemachine.withgoogle.com/train/image

We've trained a supervised classifier



- Collect data
  image frames + labels
- Train*
  neural network
  * fine-tune
- Infer or predict
  class of new image

ML brings the senses to computers

HOW? Supervised Learning Tutorial

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



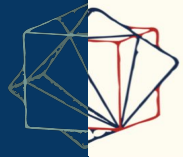
Supervised learning: One step at a time

This notebook was created for the ML ⇌ Science Colaboratory's workshop Introduction to Machine Learning.

We here focus on supervised learning. As we progress from a simple linear regression to a first example of a neural network, we introduce the basic concepts needed to understand machine learning.

Interactive notebook downloadable at https://mlcolab.org/resources/introml-november-2023-workshop-materials

Core Concepts in Machine Learning



Neural Networks

Layer upon layer upon layer

Building a neural network - terminology

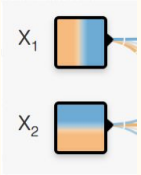
playground.tensorflow.org

DATA

Inputs: coordinates

X_1 horizontal

X_2 vertical



optional: *(hand-crafted)* **features**

X_1^2, X_2^2, X_1X_2, sin X_1, sin X_2 ← augment input

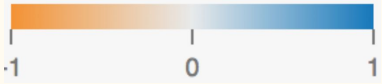
Outputs: Binary classes **orange (-1)** or **blue (+1)** → **TASK: CLASSIFICATION**

MODEL

Hidden layers = number of nestings in the NN; *here: fully connected* } **architecture**

Neurons = units in each layer

Model output: f(X_1, X_2, *features*(X_1, X_2))



Neural network playground – Training



blobs

What is the simplest architecture you can find?



xor

What minimal architecture can you find using only X_1 and X_2?

Can you improve using features?





concentric

Does the default setting solve the problem?

Can you simplify the network using features?



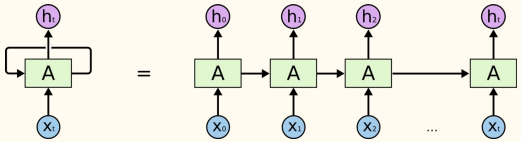
spiral

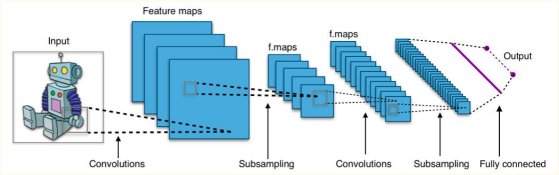
Use 1 layer and features → “feature engineering”

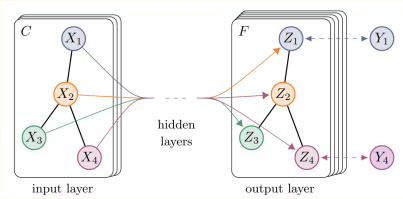
Construct a deep network, but only use X_1 and X_2

Task-specific architectures

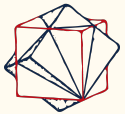
Data	**Time series, sequences (1D)**	**Images (2D)**	**Networks**
e.g.	Text, speech, temperature, …	Microscopy, astronomical shots, density maps, ...	Social networks, molecular graphs, knowledge graphs, …
NN types	RNNs *(recurrent) NNs* Transformers	CNNs – *(convolutional) NNs*	GNNs – *graph NNs*





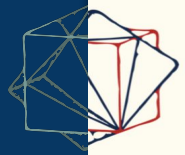


Machine Learning in Practice



The ML ⇌ Science Colaboratory

Machine Learning in Practice



ML development cycle

From data to scientific results

Zooming out: the ML dev cycle

Let's talk about how to…

1. … define the task
2. … obtain data, and prepare it
3. … implement the model
4. … evaluate, interpret and report output

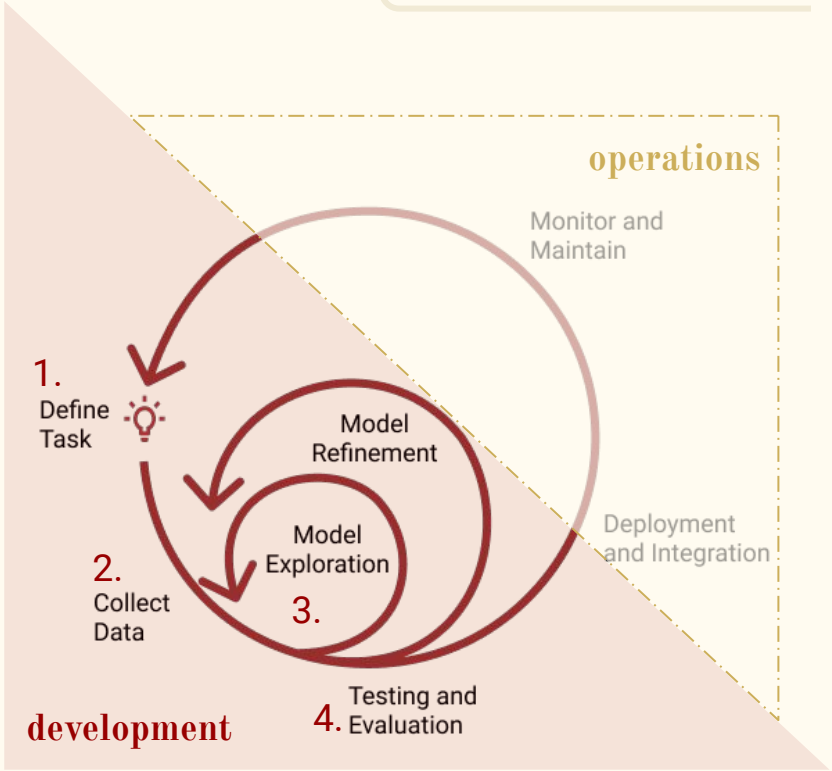


Figure based on Jordan 2018: organizing ML projects

1. Task: Feasibility & Impact of ML Project

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Do you know your data acquisition?

- All relevant conditions covered?
- Cost of additional (labeled) data?

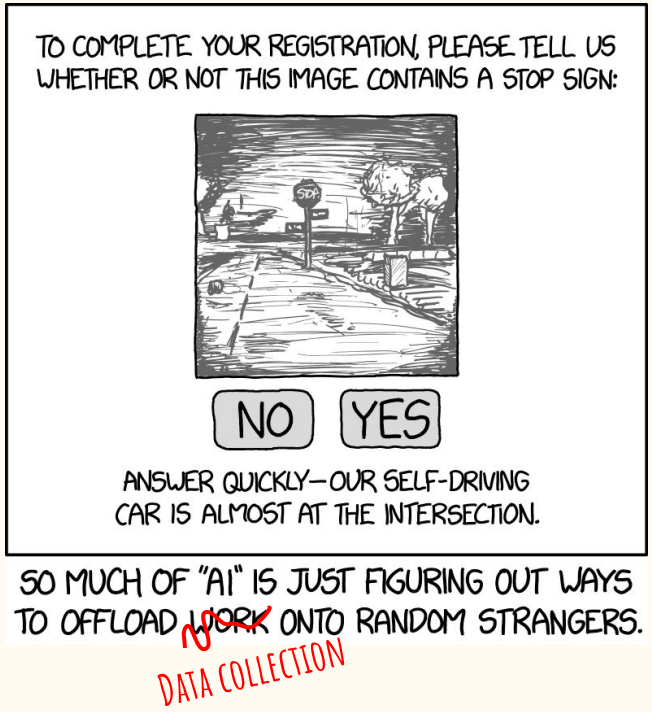
Model requirements, alternatives

- Are there non-ML models known to work well?
- Are there pre-trained models?
- Do you need interpretable results?

How to measure performance?

- Can it be measured? How?
- Can humans do it? How well?
- Are there baselines? (even if non-ML)
- What are the consequence of errors?

2. Data: How to get It



xkcd.com/1897

Strategies for data preparation (YOU!)

- **Complement** with public ancillary data
- **Reduce** via pretrained models
- **Augment** with synthetic or modified data ("Data augmentation")

Data sources

- Your own *sensor* data. Cheap!
- Your own *survey* data
- Public data sources

Data preparation: labeling

- Make it efficient:
  good user interfaces, active learning
- Consider 3rd party services: gamification

2. Data: Common Modalities

- Time series
- Images
- Relational (graph) data
- Tabular
- Video
- Georeferenced measurements
- Spectra
- A mix!

Data modalities provide knowledge that can be exploited, e.g.



The future depends on the past



Close points are usually similar



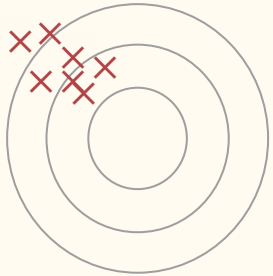
The curvature of the Earth matters

Icons from Icons8

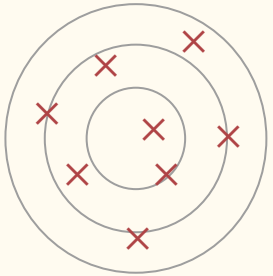
2. Data: Pitfalls

Data might not represent the sampling distribution because of

- Systematic error (bias)
- Noise (i.e. “random” error)
- Sampling bias
- Independence across samples
- Missing data
- Multi-scale phenomena
- Censored data



systematic error



random error

2. Data: Processing

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

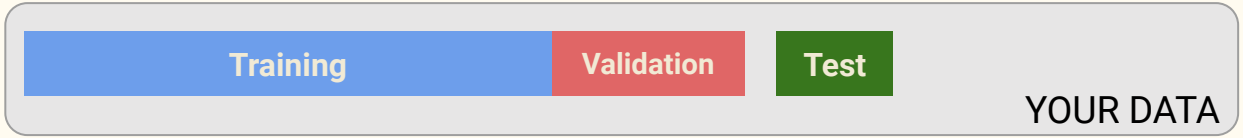
- Validation – verify **data quality**
- Aggregation
- Cleaning
- Transforming
- Reduction
- Summarization, Visualization



Take home:

- Data processing assumes some model!
- Make data processing programmatic for reproducibility!
- **Never** overwrite **raw** data with **processed** data!

3. Model: Training, Validating and Testing



- Training data ▮▮▯ to tune **parameters:** minimize loss function

- Validation data* ▯▯▮ to set ***hyper*parameters:** evaluate metrics
  *or *development set*

TRAINING

EVALUATE

- Test data ▯▯▯ ▮🔒 to estimate **generalization performance:** compare metrics <u>after</u> training with **baselines** (random or specific) and toplines (human performance). *Held out* of training.

3. Model: Implementation & Embodiments



- **“Classical ML”**
  - \+ stable, documented, interpretable
  - ○ CPU arrays (numpy, R)

  `R, sklearn, mlj`

- **Probabilistic programming**
  - \+ uncertainty quantification
  - \+ interpretable
  - – CPU mostly, expensive
  - – expert knowledge

  `PyMC, Pyro, Stan, turing`

- **Deep learning**
  - \+ expressive, versatile
  - – hard to interpret
  - ○ Training vs inference
    - ■ Resources & devices
  - ○ GPU

  `PyTorch, TensorFlow, jax, flux`

- **Pretrained**
  - ○ API serving (query only) or
  - ○ download-and-finetune (transfer learning)

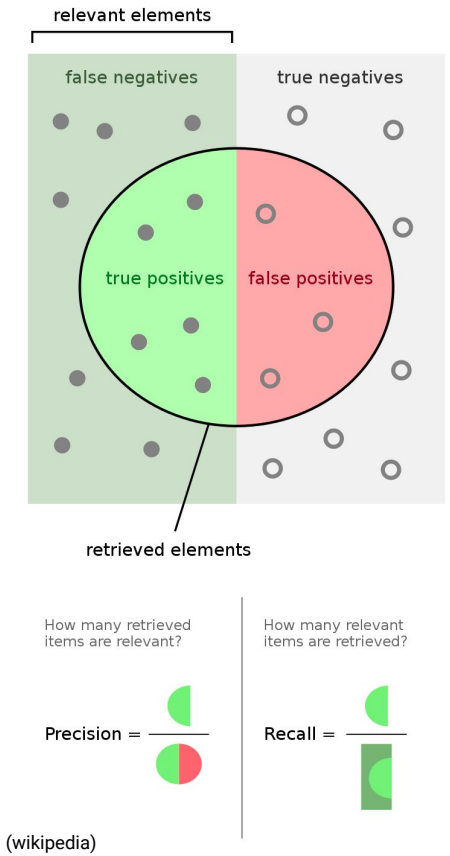
  `Huggingface, GPT, ...`

4. Evaluation & Performance Metrics

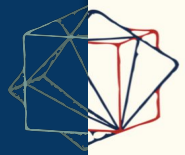
- Use held-out test data 🗄🗄🗄 🔒 to report & compare models
  - *Benchmarks* are combinations of task + dataset used to drive development in ML

- You get what you optimize for…
  … but not all you care about we can optimize for
  → need **metrics**, not just loss

- Metrics for:
  - **Regression**: MSE (cf. poly regression), …
  - **Classification**: precision, recall, F-score, …



(wikipedia)

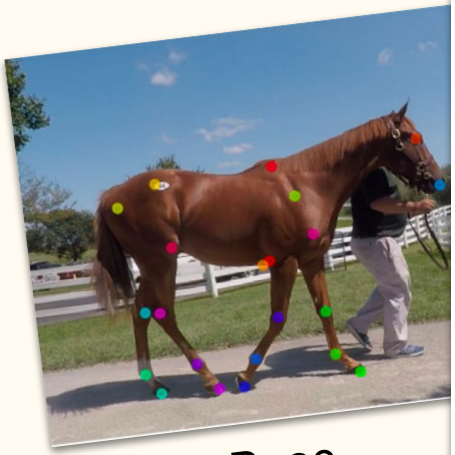


Machine Learning in Science

Learning from language, images, simulations...

Computer Vision (CV)

Data: images or videos



Pose estimation

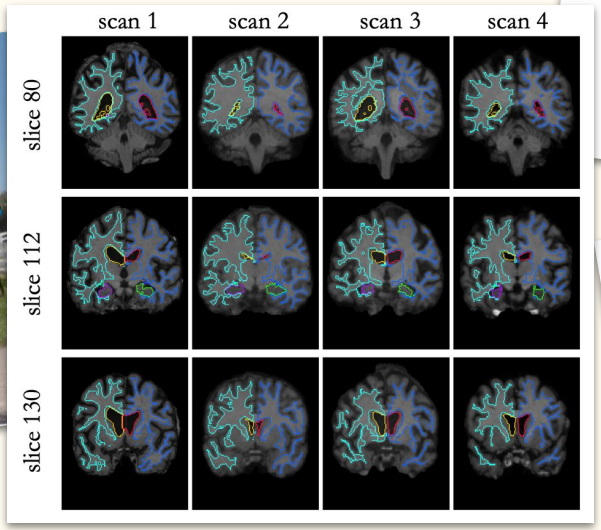


Image segmentation



Classification

Balakrishnan, Zhao et al. 2019,
Nath, Mathis et al.2018, Resler, Yeshurun et al. 2021

DeepLabcut: pose estimation without body markers

Hypothesis we can detect body parts of multiple animal species from unstaged video without body markers

Dataset video frames with *labeled* positions of body parts

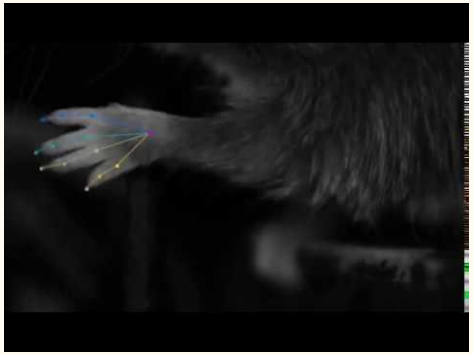
Task frame-based pose estimation

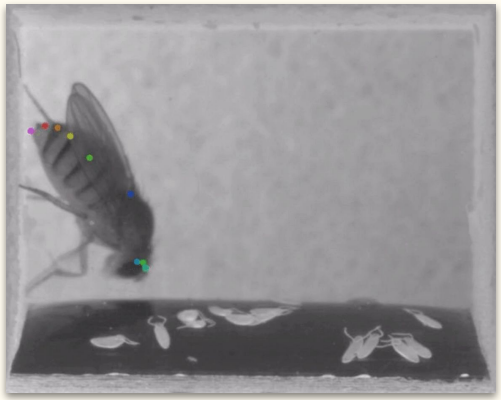
Evaluation *MSE* between predicted and labeled positions

Model *pre-trained* ResNet with fully convolutional upsampling

Supervised learning (requires labeling)

Transfer learning





Fruit fly

Nath, Mathis et al.2018

Anomaly Detection for Medical Images

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

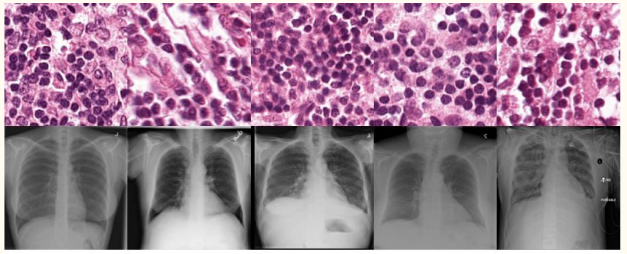
Hypothesis we can use data from healthy patients to detect rare anomalies with ML

Dataset *unlabeled* images (medical scans, etc.)

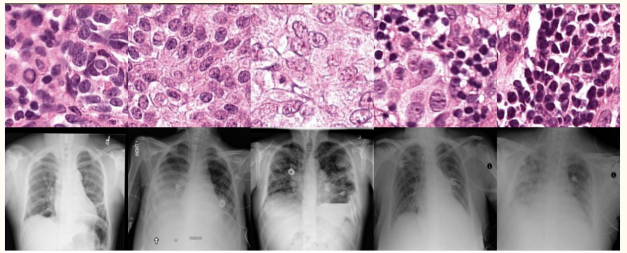
Task anomaly detection: “classification” in normal/ abnormal (=samples outside training data distribution)

Evaluation classification performance on normal and abnormal data

Model autoencoder



Normal



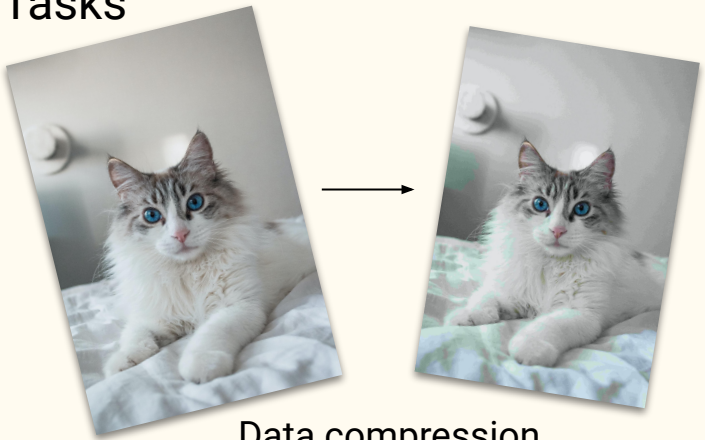
Abnormal

N. Shvetsova et al., 2021

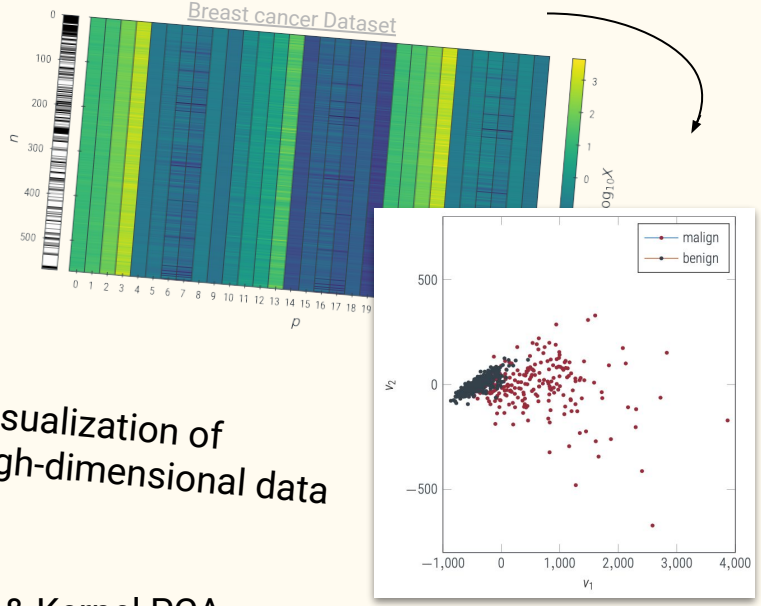
Unsupervised learning (no labeling)

Dimensionality reduction

- Data
  - High-dimensional data (e.g. k space data)
- Tasks



Data compression



Visualization of high-dimensional data

- Models
  - Unsupervised models
    - Principal component analysis (PCA) & Kernel PCA
    - Autoencoders
    - t-SNE, UMAP (for visualization)
- NOTE: Often for pre-processing of the data

Unsupervised learning (no labeling)

Cell-type prediction for single-cell transcriptomics

Hypothesis we can identify similar cells via gene expression profiles

Dataset gene expression matrices (read counts per million)

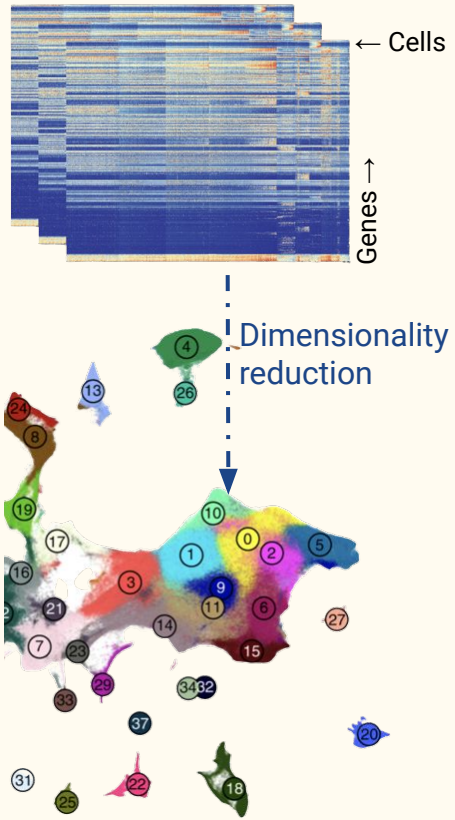
Task *dimensionality reduction*

Evaluation % neighbor cells preserved in lower-dim space

Model *t-SNE* with PCA initialisation

Unsupervised learning (no labeling)

Luecken,Theis et al.2019



Kobak, Berens et al.2019

Natural Language Processing: tasks



huggingface.co: tasks

Natural Language Processing: self-supervised learning

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Next token prediction task - no labeling required!

Unlabeled text: *The sun sets over the horizon, casting a golden glow on the sea as a gentle breeze whispers through the palm trees…*

Input (previous text)	Output (next word)
The sun sets over the horizon, casting a golden	glow
The sun sets over the horizon, casting a golden glow	on
The sun sets over the horizon, casting a golden glow on	the
The sun sets over the horizon, casting a golden glow on the	sea …
…	

self-supervised learning

ChatGPT compresses (a chunk of) an internet, learning language structure and facts (and some reasoning, too)

NLP in genomics: DNABERT

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Hypothesis we can treat DNA sequences as sentences and use next-token prediction (or masks) to train large language models on unlabeled genomics data

Data Unlabeled DNA datasets

Task next token / masked token prediction (foundation model for multiple tasks in genomics)

Model bidirectional encoder representations from transformers (BERT)

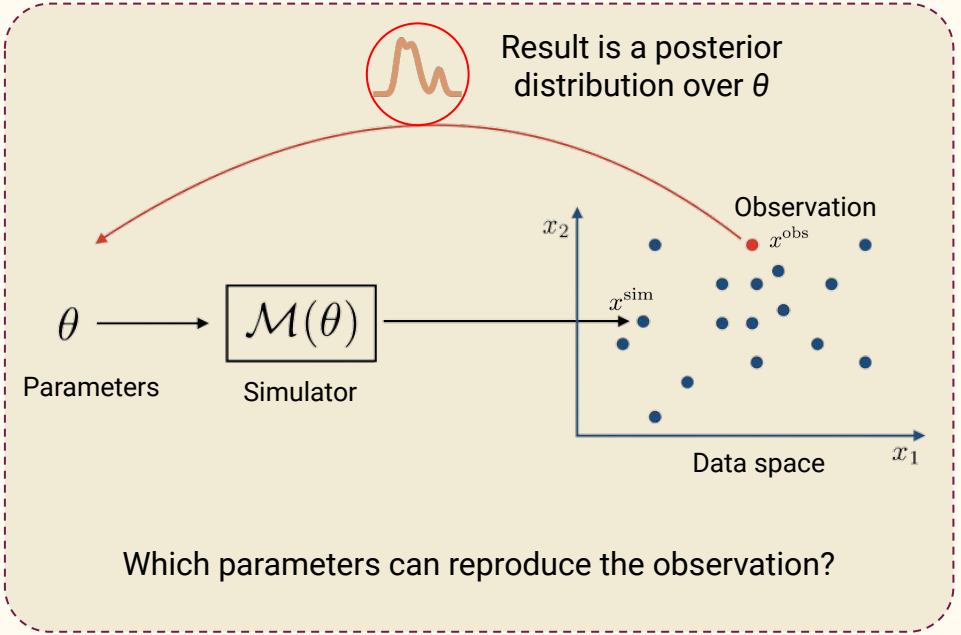
self-supervised learning

Yanrong Ji et. al., 2021

Simulation-based inference (SBI)

Tasks

- Inverse problems
- Especially beneficial for costly or intractable likelihoods
- Particularly suited for black-box simulators $x^{sim} = \mathbf{M}(\theta)$.
- How can we learn a distribution over the parameters θ such that $\mathbf{M}(\theta) \approx x^{obs}$



Learning from simulated data

Note: we held a 3-day SBI workshop and can repeat it in the future

SBI: Estimating parameters of colliding black holes from gravitational waves

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Hypothesis simulations of gravitational waves can be used to estimate parameters of colliding black holes / neutron stars

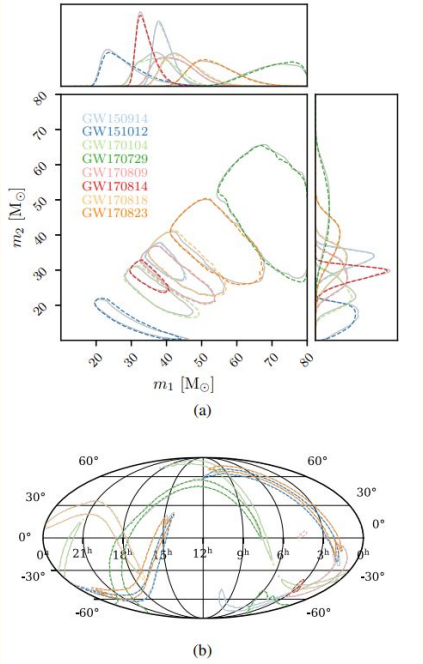
Dataset simulated gravitational waves x^{sim} = **M**(θ) from black hole parameters θ (mass, rotation, …)

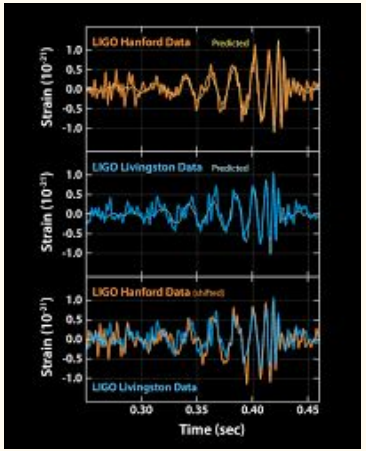
Task solve the inverse problem, i.e. get "$\theta=\theta(x^{\text{obs}})$", but *Bayesian*: $p(\theta|x^{\text{obs}})$

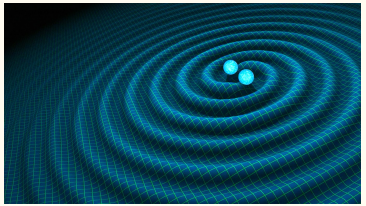
Model normalizing flows (invertible NNs) as conditional density estimators

Learning from simulated data

M. Dax et. al., 2021







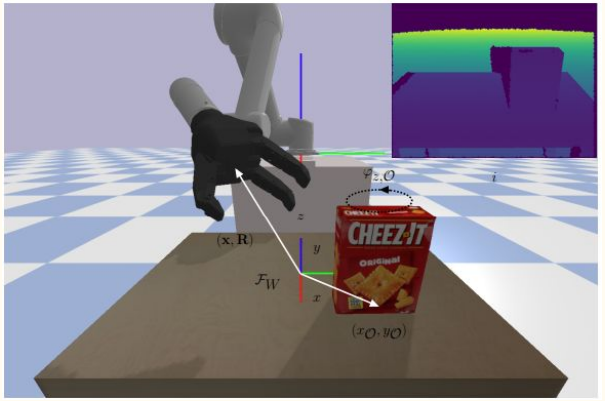
SBI: Multi-Fingered Robotic Grasping

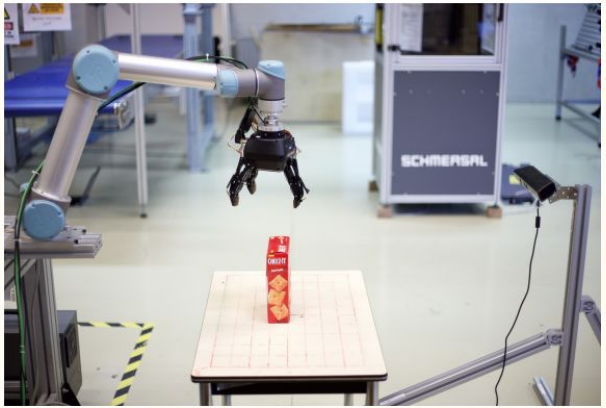
Hypothesis correct hand pose for robotic grasping can be learned by simulating various (random) scenarios and simulate result

Dataset Simulated scenarios (hand poses θ, scene depth image x^{sim} with object)

Task Bayesian solution for the inverse problem "$\theta=\theta(x^{obs})$"

Model normalizing flows (invertible NNs) as conditional density estimators





Learning from simulated data

Marlier, N. et al., 2021

Limitations of Today's ML for Science

- Inductive paradigm, limited by data
  - Bad at extrapolation
  - Building in inductive biases still a craft
- Compute intensive
- Correlational, mostly not causal
  - But can help you fit mechanistic models that are causal
- Interpretability (mostly post-hoc)
- Uncertainty quantification

Should I use ML for my problem?

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

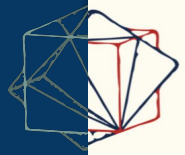
Use ML when the problem…

- … has a simple objective,
- is too complex for explicit rules,
- is constantly changing,
- is perceptive, or
- is observable, but unstudied

Don't use if:

- there exist performant classical models
- every decision must be explainable
- errors are high-consequence,
- getting dense data is infeasible or costly…

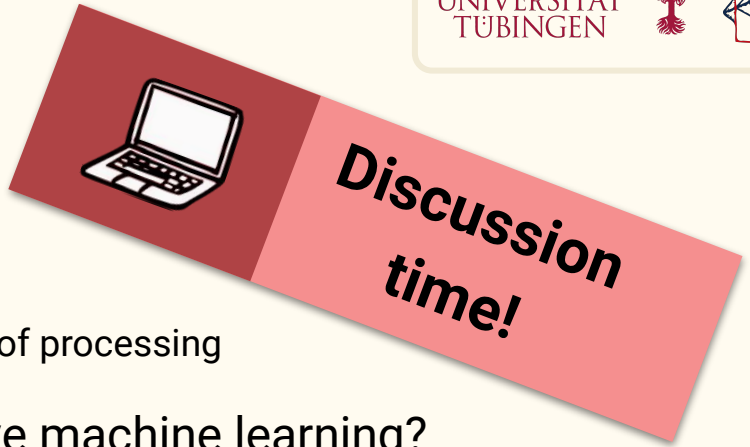
Machine Learning in Science



Bring Your Science Problem

Task, data, metrics...

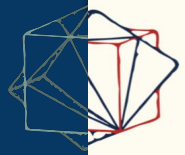
Checklist for group discussion



1. What is the scientific question?
2. **Data:** Describe your dataset:
   - data modalities, labels (if there are), size, state of processing
3. **Task:** Which part of the work might involve machine learning?
   - Identify the learning paradigm that applies to your problem (supervised, unsupervised, ...)
4. **Performance:** How could you measure ‘success’ in your problem?
5. Are there relevant baselines/toplines for performance?

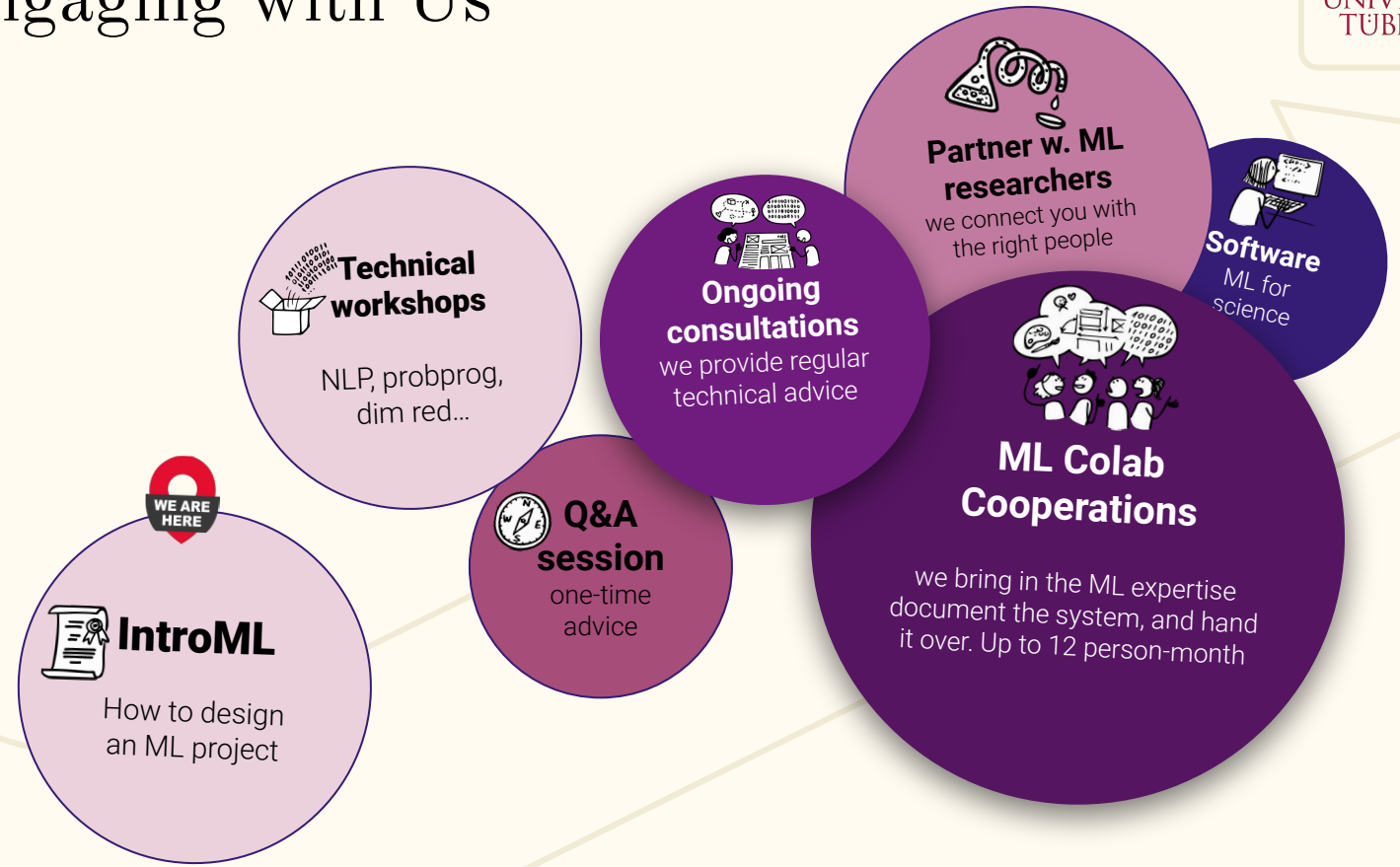
The ML ⇌ Science Colaboratory

Connecting ML with Science



Engaging with Us

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



@mlcolab
https://mlcolab.org
mlcolab@inf.uni-tue.de

The Methods Center – *Methodenzentrum*

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

An institute for quantitative analysis **in the social sciences**.

Expertise

- multivariate & multilevel methods,
- (nonlinear) latent variable models,
- item-response theory,
- structural equation models,
- longitudinal studies, and
- causal mediator models.

Services

- consulting
- cooperations
  - publications
  - joint applications.

✉ office@mz.uni-tuebingen.de

In closing...

Thank you very much for attending our IntroML workshop!



Seth



Álvaro



Vladimir



Hanqi

One last thing,

… we thrive on critical, constructive feedback

please visit https://mlcolab.org/resources/introml-november-2023-workshop-materials and give us some!

Additional materials

Slides we can build into the program, with cover slides explaining what, where and why.

Creative? (give this as part of reading recommendations)



Slides about ML vs. software and what ML is appropriate for

—

We had them at the beginning, but they may work better after seeing a few concrete examples

Great Potential for ML in Research

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

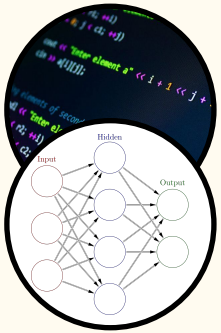
+ More data in sciences (sensors) and humanities (text understanding)

- **ML is hard to use for domain researchers**
  - **Resources —** abundant data and powerful compute
  - **Planning —** coupled model-data development
  - **Skills —** exploding ML literature, software engineering, training lore
  - **Limitations for research —** biases, interpretation, uncertainty, and causality

- **ML-to-domain *application gap*** Do research with *existing* ML algorithms?
  - Incentives misalignment in standard collaborations

Machine learning as ‘software 2.0’

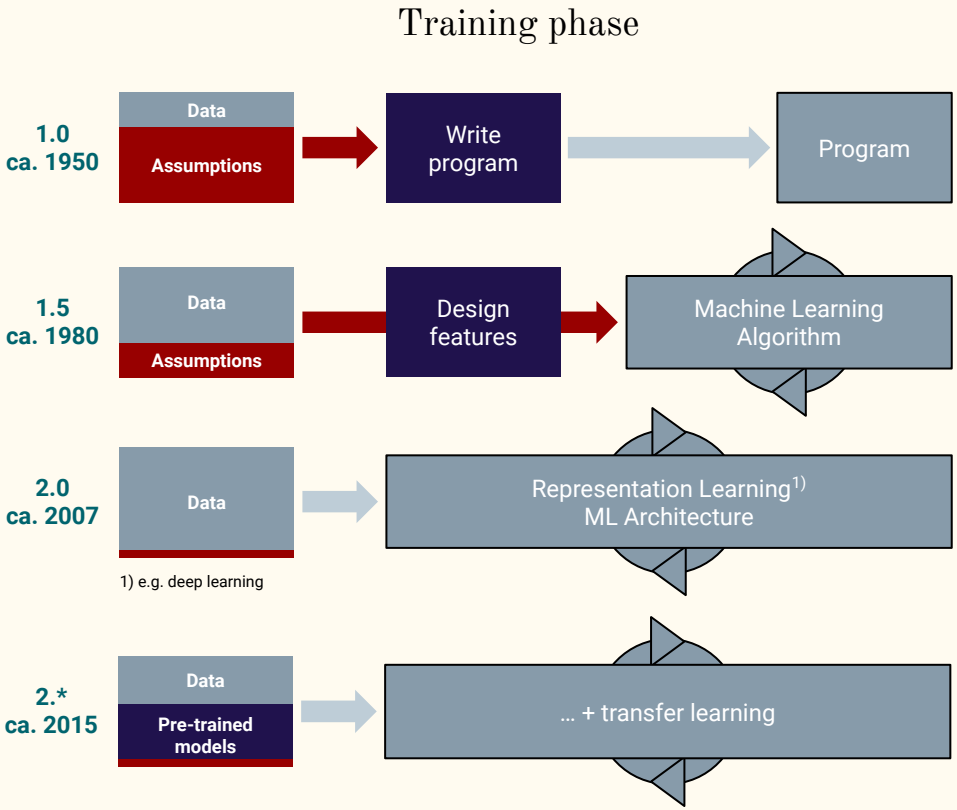
- ML produces computer algorithms, called “ML models”
  - Like traditional computer programming based on **rules**, the result is computer code
  - But ML models are substantially based on **data**, not pre-specified rules
  - In fact, ML models *use* traditional code containing general **rules for learning specific rules**
- Stages of an ML model and its internal parameters (IPs)
  - **Untrained** - it has not seen any data; IPs are typically random
  - **Training** - changing IPs progressively reflect presented data/experience
  - **Operational** - IPs fixed by training to values that satisfy a desired performance level



- Who interacts with ML when?
  - **ML researchers** design learning algorithms, train and test them, often on standardized tasks
  - **Users** train ML algorithms with their data for their *use case*
  - **Everybody** (almost) supplies data for 3rd-party ML, or uses ML trained by others

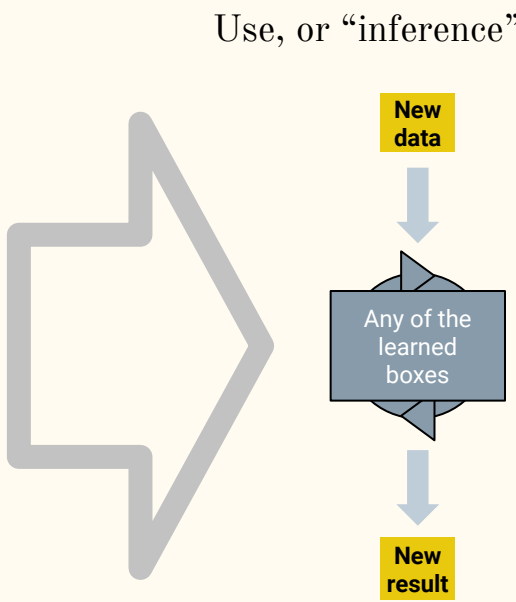
Karpathy 2017: Software 2.0, Jordan 2018: organizing ML projects

Towards Software 2.*





Towards "Software 2.0" (Karpathy, 2017)



Why ML now?

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN





ML coming of age ca 2010, a coalition of factors

- Data: commoditization of sensors and distributed data collection
- Hardware: parallelized linear algebra, originally for gaming (“GPUs”)
- Software:
  - Usable automatic differentiation engines (e.g. Theano 2007)
  - Languages for composing models easily
- Models
  - Multiple layers and training tricks (grad clipping, dropout, batchnorm, …) to stabilize training
  - Compositionality of neural networks
  - Transfer learning

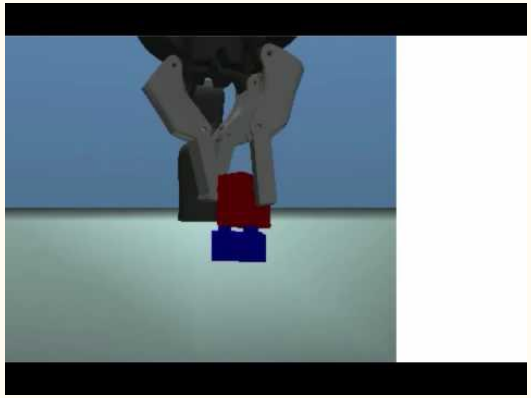
Half-finished slides

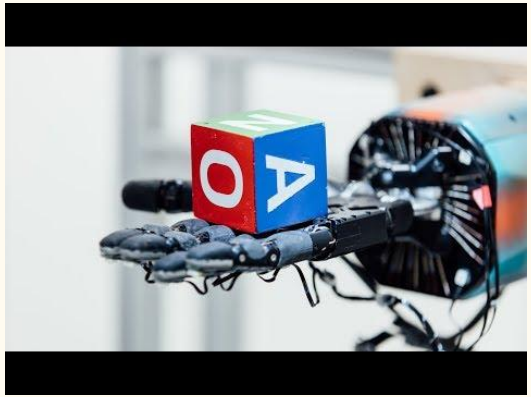
1. Data: Processing

- Cleaning
  - Outlier detection/removal
  - Integrating data from different sources
- Transforming
  - Standardizing (shifting and scaling)
  - Smoothing/denoising
  - Imputation of missing data
  - Augmentation
- Reduction
  - Downscaling
  - Feature selection
- Take-homes
  - Data processing assumes some model!
  - Always make data processing programmatic for reproducibility!
  - Never overwrite raw data with processed data!

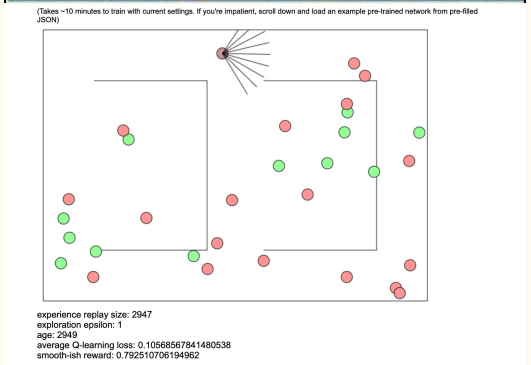


Reinforcement Learning: additional examples









https://cs.stanford.edu/people/karpathy/convnetjs/demo/rldemo.html

We’re not sure these slides are necessary at all

—

Sometimes, it is not that easy to classify....



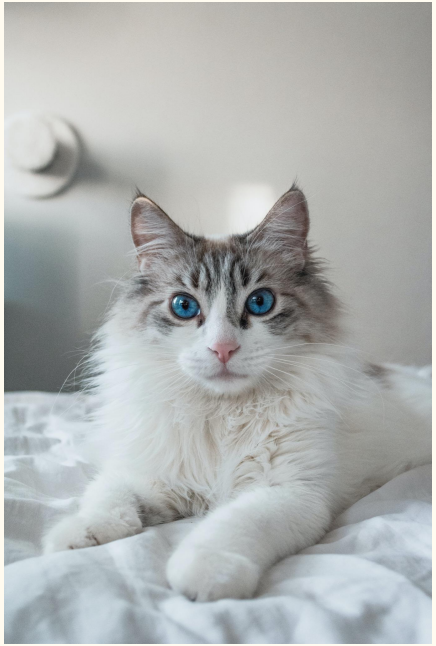




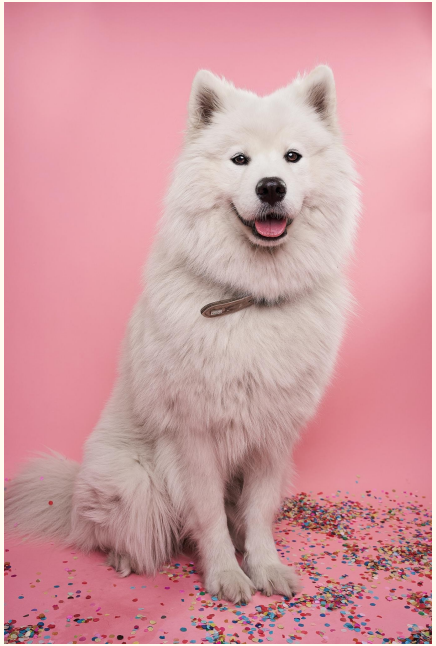
Animal or
food?

Source: thecoastnet.nz

Sometimes, it is not that easy to classify....



Cat



Dog



?

Cat ? or Dog ?

Photo by Esteban Chinchilla on Unsplash Photo by Peri Stojnic on Unsplash Marc Brockschmidt MSR Cambridge

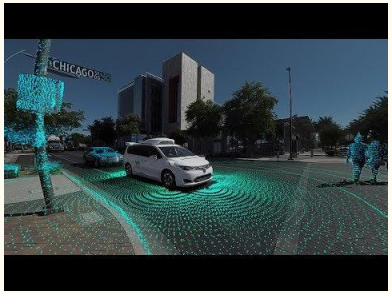
More application examples

Those which are not scientific we should take out; scientific ones can be developed to replace current examples if appropriate

CV: A Deep Convolutional Encoder-Decoder for Image Segmentation

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- Task
  - Pixel-wise semantic segmentation
- Data
  - CamVid road scenes dataset (for training)
  - SUN RGB-D dataset (for test scenes)



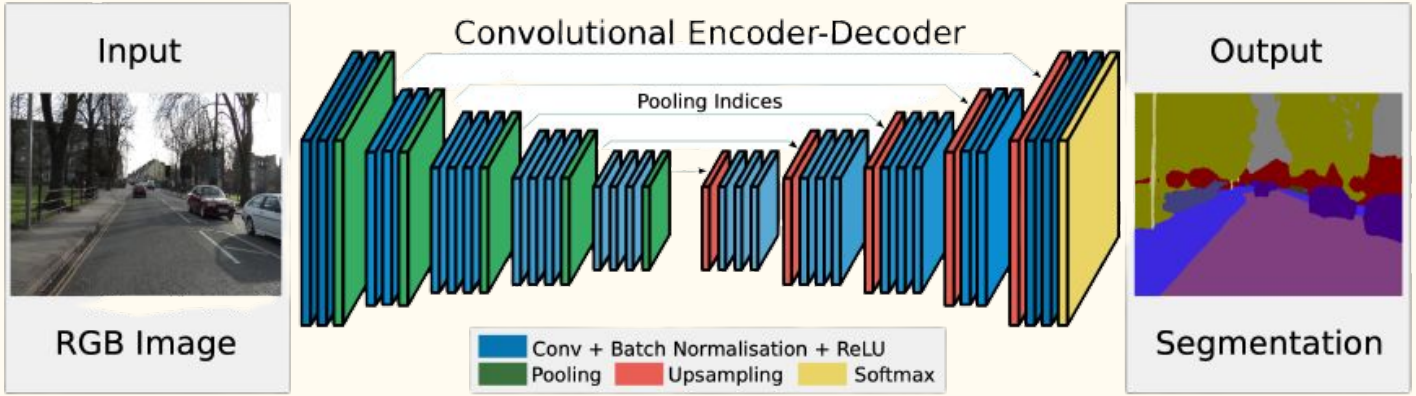
Input

SegNet



CV: A Deep Convolutional Encoder-Decoder for Image Segmentation

- Model
  - Encoder Network
  - Decoder Network
  - Pixelwise classification layer



CV: Object recognition in archaeology(skip)

- Task & evaluation
  - Classify objects from images
- Dataset
  - 12,364 photographs of 6770 artefacts, southern Levant, Lower Palaeolithic
  - Training set (81%), Testing set (19%)
- Problem: visual similarity ?
  - Similar artefacts belong to different classes
  - Visually distinct artefacts belong to same classes



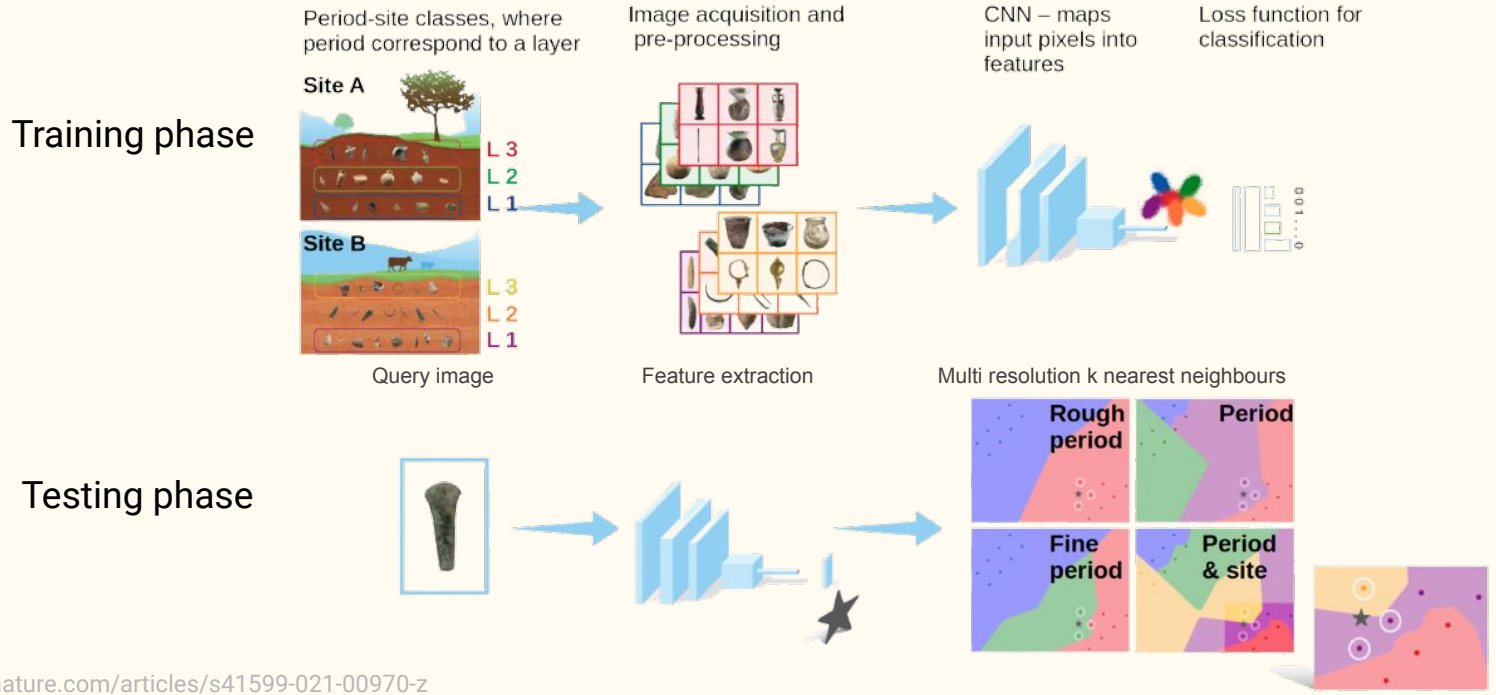
Different class

Same class

Reference: https://www.nature.com/articles/s41599-021-00970-z

CV: Object recognition in archaeology(skip)

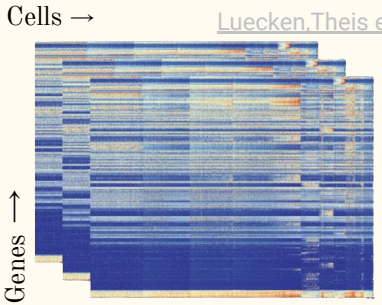
- Model
  - Convolutional Neural Network (CNN) for **classification**



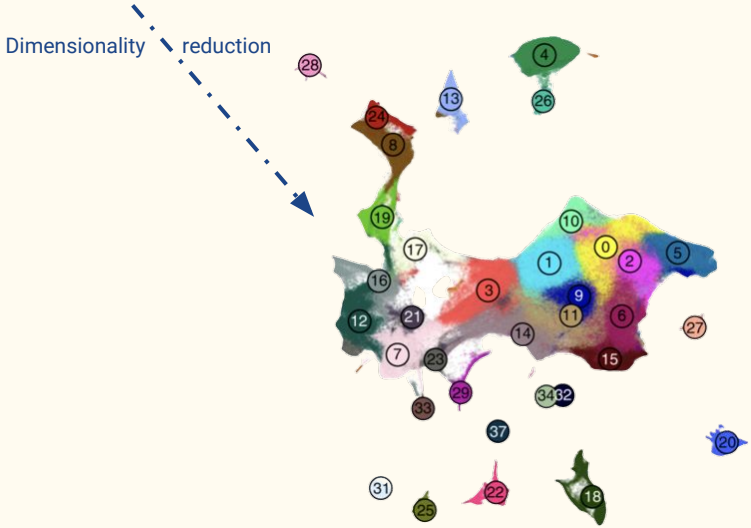
Reference: https://www.nature.com/articles/s41599-021-00970-z

DimRed: Data visualization for single-cell transcriptomics

- Task
  - Visualizing cells according to gene expression
  - Clustering
- Data
  - Gene expression matrix
    ~ thousands of genes
    ~ millions of cells
  - Expression of many genes is correlated
- Model
  - t-SNE with PCA initialisation

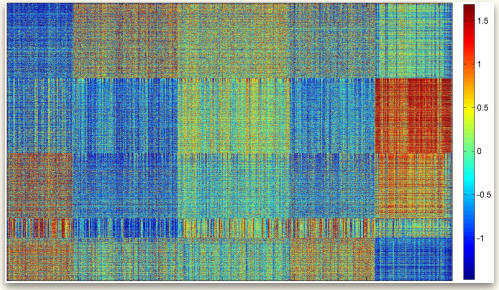


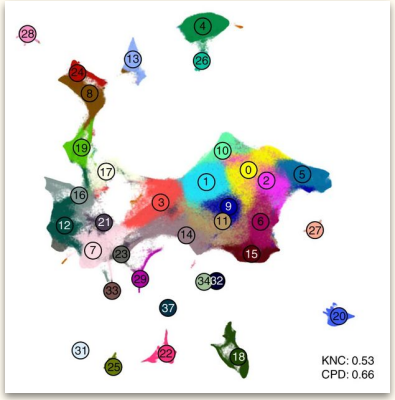
Luecken,Theis et al.2019



Kobak, Berens et al.2019

Backup images

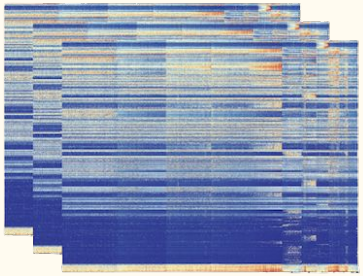




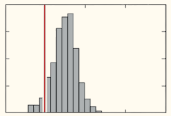
DimRed for single-cell transcriptomics

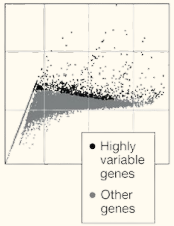


Count matrices



Quality control
Data correction





Feature selection
Normalization





Pre-Processing

Visualization

Clustering

Luecken,Theis et al.2019

Kobak, Berens et al.2019

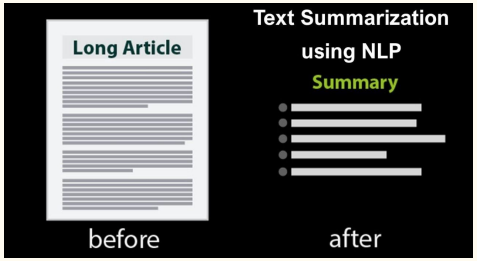
Natural Language Processing (NLP)

- Document summarization
- Image captioning
- Topic classification
- Sentiment analysis
- Machine translation
- Question answering
- Speech-to-text; text-to-speech



a bus that has pulled into the side of the street
a bus is parked at the side of the road
a white bus is parked near a curb with people walking by

Image captioning

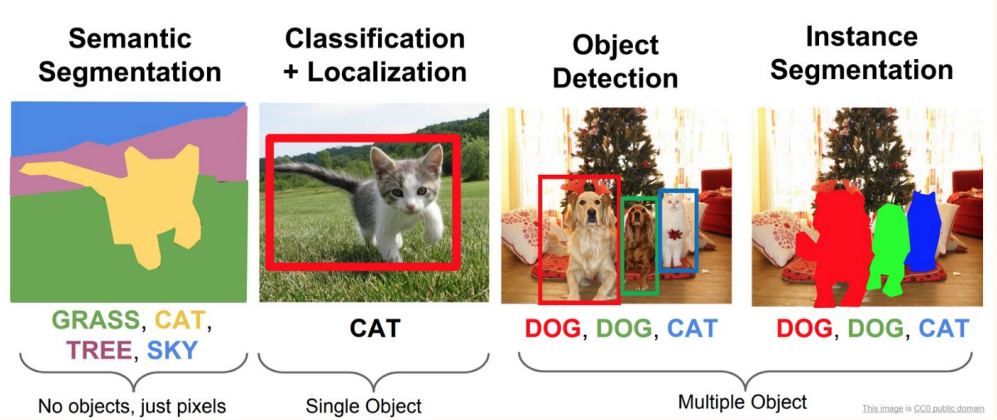


TODO: find example figures

- Data
- Tasks
  - Document summarization
    - Semantic segmentation
    - Instance segmentation
  - Object recognition (classification)
  - Object detection
  - Optical character recognition (OCR)

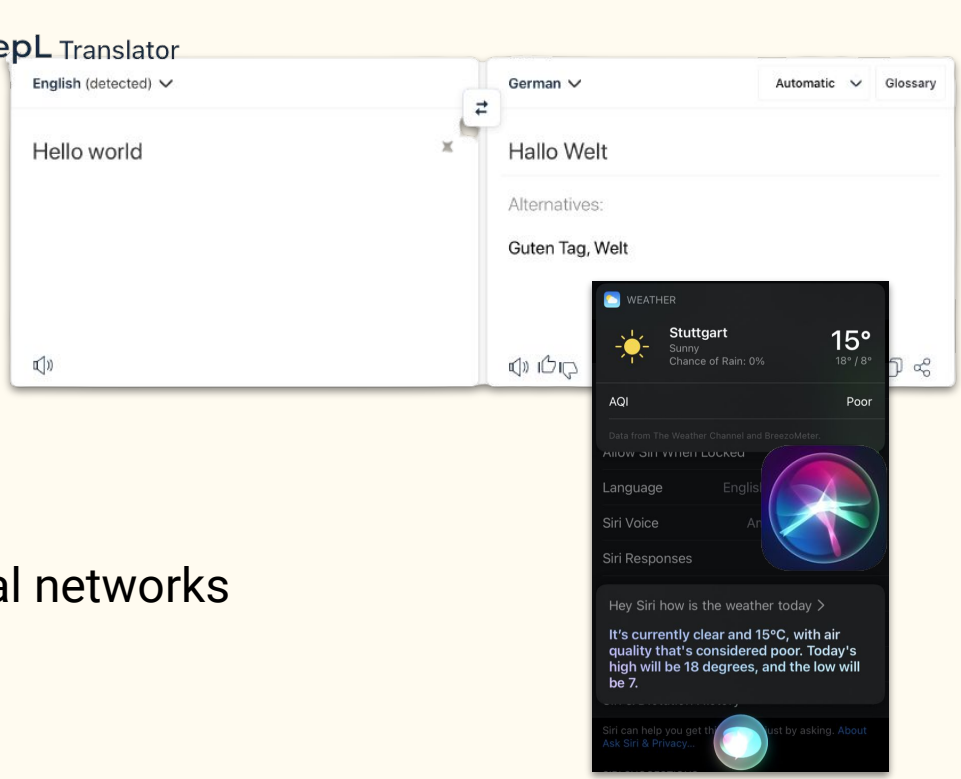
Computer Vision (CV)

- Data = images or videos
- Tasks
  - Image segmentation
    - Semantic segmentation
    - Instance segmentation
  - Object recognition (classification)
  - Object detection
  - Optical character recognition (OCR)
- Model: often pre-trained convolutional neural networks



Natural Language Processing (NLP)

- Data: Text or audio
- What is NLP used for?
  - Text summarization
  - Language translation
  - Topic classification
  - Question answering
  - Speech-to-text; text-to-speech
  - . . . .
- Model: pre-trained recurrent neural networks

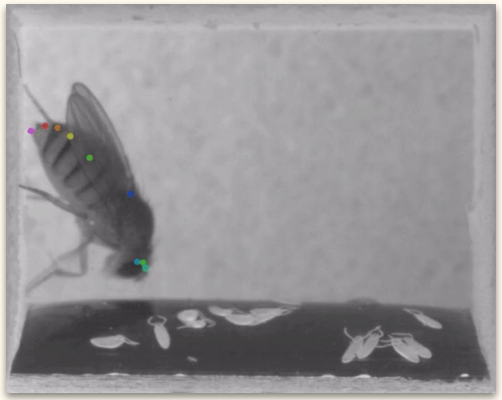


NLP: Biomedical Text Mining

- Problem
  - 3000 new articles are published every day
  - Word distributions of general and biomedical corpora are different
- Goal
  - Easy access to the information locked within biomedical literature
- Data
  - NER datasets
  - Species 800 datasets
- Model
  - Pre-trained BioBERT model
- Highlight
  - Achieve state-of -art performance on various biomedical task with small model modifications
  - Outperform on
    - biomedical named entity recognition
    - Biomedical relation extraction
    - Biomedical question answering
  - Help to understand complex biomedical texts

CV: pose estimation without body markers

- Problem
  - Noninvasive behavioral tracking is important, but extracting the pose of animals without using markers in dynamically changing background is challenging.
- Learning Goal
  - Detect user defined body parts in varying background
- Input: Videos
  - Output: Extracted pose files
  - Toolbox: DeepLabCut
- Highlight
  - Use for small-scales operations in laboratory
  - Human-level labeling accuracy
  - Capture behaviors with minimal occlusion
  - Easy to use without requiring programming skill
  - Not only for laboratory, also for sports



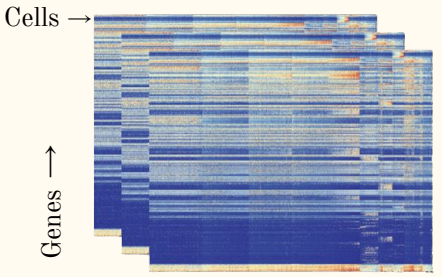
Fruit fly

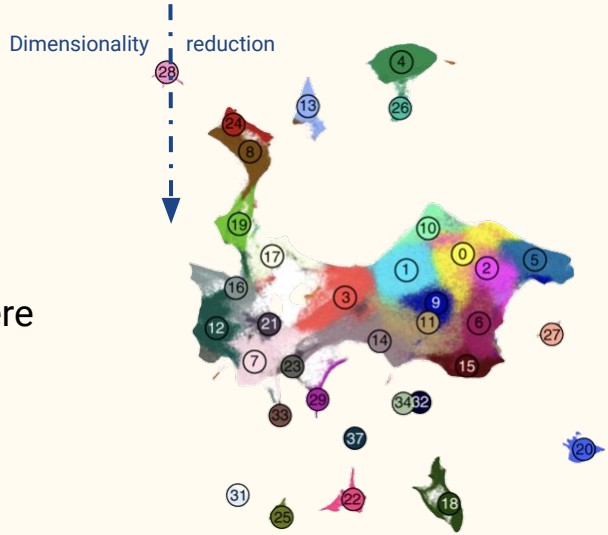
Nath, Mathis et al.2018

DimRed: Data visualization for single-cell transcriptomics

- Problem
  - Data is messy (thousands of genes and millions of cells)
- Goal
  - Visualizing cells according to gene expression
  - Clustering
- Input data
  - Gene expression matrix(matrices of molecular counts)
- Output
  - Visualize and the dataset in 2 dimensions and clustering
- Model
  - t-SNE with PCA initialisation
- Highlight
  - Place all the point in 2D plane and let them interact as if they were physics particles
  - The result can preserve the global geometry of the data
  - The final visualization is better than the state-of-art
  - Can position new data on existed model

Luecken,Theis et al.2019





Kobak, Berens et al.2019

Meta slides

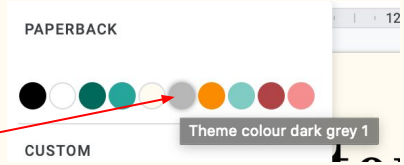
Those which are not scientific we should take out; scientific ones can be developed to replace current examples if appropriate

Instructions for slide developers

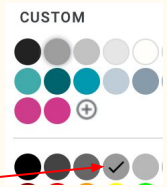
- Refer to our Notion content sketch for the structure
- Make background white as soon as figures would look bad, but try first with drop shadow.
- Participants e-mail template, signup selection
- Example citation style (copy & modify) → Assael, Sommerschield et al. 2022

Format

- Normal Content, explanations in page
  - Roboto(style) 18(font size), black
- Title in page
  - Old standard TT (style) 27(font size) color(black)
- Subtitle in page
  - Old standard TT(style) 16 (font size)color(black)
- Caption of figure
  - Old standard TT(style) 8(font size)color(dark grey 1)
  - Exception: speaker name with black color
- Section title
  - Old standard TT(style) 33(font size) color(background blue)
- Subsection title
  - Old standard TT(style) 42(font size) color(#fffbf0)
- Subsubsection title
  - Old standard TT(style) 24(font size) color(dark grey 1)
- Bulleted list(or something like that)
  - Big black dot: Roboto(style) 18(font size), black
  - Small hollow dot: Roboto(style) 14(font size), black
- Reference Link
  - Roboto(style) 9(fontsize), dark grey 2







This should replace the slide with a screen capture

—

But only if people can find the fonts

We've trained a supervised classifier



- Collect data
  image frames + labels
- Train*
  neural network
  * fine-tune
- Infer or predict
  class of new image

ML brings the senses to computers