Welcome to the Class of 2021 of 185A83, 23.03.2021, 17:00-20:00 CET

185.A83 Machine Learning for Health Informatics
2020S, VU, 2.0 h, 3.0 ECTS
Andreas Holzinger, Rudolf Freund
Marcus Bloice, Florian Endel, Anna Saranti

Lecture 01 – Introduction

From health informatics to ethical responsible medical AI

Contact: andreas.holzinger AT tuwien.ac.at
@aholzin #KandinskyPatterns

This is the version for printing and reading. The lecture version is didactically different.

For deeper learning effect, please always try to answer the question in the title bar first.

If you need a refresher on the basics of probability and information theory, please refer to slide deck 00.

This course follows a Research-Based Teaching Style
At the end of this lecture you will ...

- see why an integrative approach is important
- understand that machine learning can help medicine
- be fascinated by the possibilities of machine learning
- disillusioned at the limits of machine learning
- be aware of the complexity of the health domain
- realize why a human-in-the-loop is sometimes crucial
- identify the main challenges in this area
- recognize some ideas of future work
Bioinformatics = discipline, as part of biomedical informatics, at the interface between biology and information science and mathematics; processing of biological data;

Biomarker = a characteristic (e.g. body temperature (fever) as a biomarker for an infection, or proteins measured in the urine) as an indicator for normal or pathogenic biological processes, or pharmacologic responses to a therapeutic intervention;

Biomedical data = compared with general data, it is characterized by large volumes, complex structures, high dimensionality, evolving biological concepts, and insufficient data modeling practices;

Biomedical Informatics = 2011-definition: similar to medical informatics but including the optimal use of biomedical data, e.g. from genomics, proteomics, metabolomics;

Classical Medicine = is both the science and the art of healing and encompasses a variety of practices to maintain and restore health;

Genomics = branch of molecular biology which is concerned with the structure, function, mapping & evolution of genomes;

Interactive Machine Learning = defined as algorithms that can interact with both computational agents and human agents and can optimize their learning behaviour through these interactions, by bringing in a human-into-the-loop

Machine Learning = addresses the question of how to design algorithms that improve automatically through experience from big data - doing it automatically (aML) without a human-in-the-loop

Medical Informatics = 1970-definition: “... scientific field that deals with the storage, retrieval, and optimal use of medical information, data, and knowledge for problem solving and decision making”; - see the better 2011-definition by the AMIA

Metabolomics = study of chemical processes involving metabolites (e.g. enzymes). A challenge is to integrate proteomic, transcriptomic, and metabolomic information to provide a more complete understanding of living organisms;

Molecular Medicine = emphasizes cellular and molecular phenomena and interventions rather than the previous conceptual and observational focus on patients and their organs;
- **Omic data** = data from e.g. genomics, proteomics, metabolomics, etc.
- **Pervasive Computing** = similar to ubiquitous computing (Ubicomp), a post-desktop model of Human-Computer Interaction (HCI) in which information processing is integrated into every-day, miniaturized and embedded objects and activities; having some degree of “intelligence”;
- **Pervasive Health** = all unobtrusive, analytical, diagnostic, supportive etc. information functions to improve health care, e.g. remote, automated patient monitoring, diagnosis, home care, self-care, independent living, etc.;
- **Proteome** = the entire complement of proteins that is expressed by a cell, tissue, or organism;
- **Proteomics** = field of molecular biology concerned with determining the proteome;
- **P-Health Model** = Preventive, Participatory, Pre-emptive, Personalized, Predictive, Pervasive (= available to anybody, anytime, anywhere);
- **Space** = a set with some added structure;
- **Technological Performance** = machine “capabilities”, e.g. short response time, high throughput, high availability, etc.
- **Time** = a dimension in which events can be ordered along a time line from the past through the present into the future;
- **Translational Medicine** = based on interventional epidemiology; progress of Evidence-Based Medicine (EBM), integrates research from basic science for patient care and prevention;
- **Von-Neumann-Computer** = a 1945 architecture, which still is the predominant machine architecture of today (opp.: Non-Vons, incl. analogue, optical, quantum computers, cell processors, DNA and neural nets (in silico));
● (0) A few definitions – for mutual understanding
● (1) Machine Learning Health examples
● (2) A brief look at the application area health
● (3) Statistical Machine Learning
● (4) Automatic Machine Learning (aML)
● (5) Interactive Machine Learning (iML) and why we need the human-in-the-loop
● (6) Explainable AI and Methods of Explainability
● (7) Causability – Measuring the Quality of (6)
(0) A few definitions first
Health, Medicine, Informatics, Biomedical Informatics, ...?

- **Health** := state of physical, mental and social well-being
- **Medicine** := art, science, practice of patient diagnosis, prognosis, prevention, treatment of injury or disease.
- **Informatics** := study of information processing (Schrödinger [1]: Life is Information Processing)
- **Biomedical informatics** (BMI) := interdisciplinary field that studies and pursues the effective use of biomedical data, information, and knowledge for scientific problem solving, and decision making, motivated by efforts to improve human health [2]


Artificial Intelligence, Machine Learning, Deep Learning, ... ?

- **AI** := not clearly defined “umbrella term” of “making intelligent machines”
- **ML** := the workhorse of AI - algorithms that improve by learning from data
- **DL** := a particular family of neural networks - currently very successful


**What was the first definition of Machine Learning?**

**Some Studies in Machine Learning Using the Game of Checkers**

by A. L. Samuel


**Introduction**

The studies reported here have been on a digital computer to behave in a way which animals, would be described as involving this is not the place to dwell on the procedures, or to discourse on the philosophy of very large amount of work, now done by its demands on the intellect but does, not. We have at our command computers with with sufficient computational speed techniques, but our knowledge of the basis still rudimentary. Lacking such knowledge of methods of problem solution in minute and costly procedure. Programming could eventually eliminate the need for manual effort.

**In Memoriam**

Arthur Samuel: Pioneer in Machine Learning

Arthur Samuel (1901-1990) was a pioneer of artificial intelligence research. From 1949 through the late 1960s, he did the best work in making computers learn from their experience. His vehicle for this work was the game of checkers. Programs for playing games often fill the role in artificial intelligence research that the fruit fly Drosophila did in its time. Samuel was a modest man, and the importance of his work was widely recognized only after his retirement from IBM in 1966, in part because he didn't relish the politics that were required to have his research more vigorously followed up on. He was also realistic about the large difference between what had been accomplished in understanding intellectual mechanisms and what would be required to reach human-level intelligence.

**Field of Study that gives computers the ability to learn [from Data] without explicitly being programmed ...**

What is Standardized Medicine versus Personalized Medicine?

Preventive, Participatory, Pre-emptive, Pervasive, Predictive, Personalized?

What Kind of Healthcare Decisions Should Be Made

**Preventive**
Strategies that control risk factors of diseases will be implemented based on a mixture of individualised and population approaches.

**Participatory**
Healthcare decision making and health information will be shared by individuals and relevant practitioners.

**Pre-emptive**
Targets of intervention will be broadened beyond treatment response and remission to maintain and restore body health and functions.

How Healthcare Decisions Should Be Made

**Personalised**
Healthcare decisions will be tailor-made based on individualised modelling from genomic to system levels with reference to statistical analysis of a population.

**Predictive**
Risk of developing a disease will be constantly assessed based on the health information accumulated up-to-date.

**Pervasive**
Health services will be available to anyone, anytime and anywhere to facilitate healthcare decisions to be made whenever necessary.

How is Stratified Medicine tailoring the treatment to patient subgroups?

Interesting signals from each modality (time-based, image, structured & unstructured) are connected according to pre-defined rules. Each modality’s features lie in their own, un-aligned concept spaces.

(1) Machine learning health examples
Where do we stand with Artificial Intelligence in Medicine?

**ARTIFICIAL INTELLIGENCE IN MEDICINE**

**Where Do We Stand?**

After hearing for several decades that computers will soon be able to assist with difficult diagnoses, the practicing physician may well wonder why the revolution has not occurred. Skepticism at this point is understandable. Few, if any, programs currently have active roles as consultants to physicians. The story behind these unfulfilled expectations is instructive and, we believe, offers hope for the future.


**What Does the Future Hold?**

In 1970 an article in the *Journal* predicted that by the year 2000 computers would have an entirely new role in medicine, acting as a powerful extension of the physician’s intellect. At the halfway point, how realistic does this projection seem? It is now clear that great progress has been made in understanding how physicians solve difficult clinical problems and in implementing experimental programs that capture at least a portion of human expertise. On the other hand, it has become increasingly apparent that major intellectual and technical problems must be solved before we can produce truly reliable consulting programs. Nevertheless, assuming continued research, it still seems possible that by the year 2000 a range of programs will be available that can greatly assist the physician. It seems highly unlikely that such a goal will be achieved much before that time.
Example Dermatology: Towards Human-Level AI

Example Retinopathy: Towards Human-level AI

Jeffrey De Fauw et al. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine, 24, (9), 1342-1350

Nature medicine

What is the difference between humanoid AI and human-level AI?

Why is Machine Learning in health enormously progressing?

- Progress is driven by the explosion in the availability of big data and low-cost computation.
- Health concerns everyone ...


(1) Why can AI solve some tasks better than humans?

(2) How did AI get these results in the first place?

(3) What happens if I change, replace, disturb, remove, ... input data?
What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?

Robustness & Explainability

What do we need to make AI even more successful?
What do we need to reach robust intelligence?

- 1) learning from few **data**
- 2) extracting **knowledge**
- 3) **generalize**
- 4) fight the curse of **dimensionality**
- 5) disentangle the **independent** explanatory factors of data, i.e.
- 6) **causal understanding** of the data in the **context** of an application domain
(2) Application Area

Health

Image Source: LKH Feldbach, Steiermark
Why is health a complex application domain?

Our central hypothesis:
Information may bridge this gap

What are the main problems in “building a bridge”?

Heterogeneity
Dimensionality
Complexity
Uncertainty

Why is probabilistic information so important in medicine?

- 400 BC Hippocrates (460-370 BC):
  - A medical record should reflect the course of a disease ...
  - ... and should indicate the **probable cause** of a disease

- **1890** William Osler (1849-1919):
  - Medicine is a science of uncertainty and an art of probabilistic decision making

- Today
  - Prediction models are based on data features, the patient health status is modelled as high-dimensional feature vectors ...
(3) Probable Information and Statistical Machine Learning
Probability \( p(x) \) is the formal study of laws of chance and managing uncertainty; allows to measure (many) events

- **Frequentist** view: coin toss
- **Bayesian** view: probability as a measure of belief (this is what made machine learning successful)
- \( p(x) = 1 \) means that all events occur for certain
- Information is a measure for the reduction of uncertainty
- If something is 100% certain its uncertainty = 0
- Uncertainty is max. if all choices are equally probable (I.I.D = independent and identically distributed)
- Uncertainty (as information) sums up for independent sources: \( \sum_x p(x = X) = 1 \)

*) Bayesian vs. Frequentist - please watch the excellent video of Kristin Lennox (2016): [https://www.youtube.com/watch?v=eDMGDhyDxuY](https://www.youtube.com/watch?v=eDMGDhyDxuY)
Who laid the foundations for modern statistical machine learning?

"Il est remarquable qu'une science qui a commencé avec l'ère la prise en compte des jeux de hasard ... aurait dû devenir l'objet le plus important de la connaissance humaine."

Pierre Simon de Laplace, 1812

\[ p(\theta | \mathcal{D}) = \frac{p(\mathcal{D} | \theta) \cdot p(\theta)}{p(\mathcal{D})} \]

Probability Theory is nothing, but common sense reduced to calculation ...


Bayes’ Rule is a corollary of the Sum Rule and Product Rule:

\[
p(x_i | y_j) = \frac{p(y_j | x_i) p(x_i)}{\sum p(x_i, y_j) p(x_i)}
\]

Why is the work of Laplace so important for us?

What is the simplest mathematical operation for us?

\[ p(x) = \sum_x (p(x, y)) \]  

(1)

How do we call repeated adding?

\[ p(x, y) = p(y|x) \times p(y) \]  

(2)

Laplace (1773) showed that we can write:

\[ p(x, y) \times p(y) = p(y|x) \times p(x) \]  

(3)

Now we introduce a third, more complicated operation:

\[ \frac{p(x, y) \times p(y)}{p(y)} = \frac{p(y|x) \times p(x)}{p(y)} \]  

(4)

We can reduce this fraction by \( p(y) \) and we receive what is called Bayes rule:

\[ p(x, y) = \frac{p(y|x) \times p(x)}{p(y)} \quad p(h|d) = \frac{p(d|h)p(h)}{p(d)} \]  

(5)
What are the fundamentals of the work of Bayes-Price-Laplace?

- 1763: Richard Price publishes posthum the work of Thomas Bayes
- 1781: Pierre-Simon Laplace: Probability theory is nothing, but common sense reduced to calculation...
- 1812: Théorie Analytique des Probabilités, now known as Bayes' Theorem, should be correctly named as Bayes-Price-Laplace T.

\[ p(h|d) \propto p(d|h) \times p(h) \]

- **Hypothesis** \( h \in \mathcal{H} \) (uncertain quantities (Annahmen))
- **Data** \( d \in \mathcal{D} \) ... measured quantities (Entitäten)
- **Prior probability** \( p(h) \) ... probability that \( h \) is true
- **Likelihood** \( p(d|h) \) ... “how probable is the prior”
- **Posterior Probability** \( p(h|d) \) ... probability of \( h \) given \( d \)

\[ p(h|d) = \frac{p(d|h)p(h)}{p(d)} \]
Who are our giants in the foundations of statistical machine learning?

Newton, Leibniz, ... developed calculus – mathematical language for describing and dealing with rates of change

Bayes, Laplace, ... developed probability theory - the mathematical language for describing and dealing with uncertainty

Gauss generalized those ideas
Bayes’ Rule is a corollary of the Sum Rule and Product Rule:

\[ p(x_i | y_j) = \frac{p(y_j | x_i)p(x_i)}{\sum p(x_i, y_j)p(x_i)} \]


How does inference work in practice?

\( d \) ... data

\( h \) ... hypotheses

\[ \mathcal{H} \ldots \{H_1, H_2, ..., H_n\} \quad \forall \ h, d \ ... \]

\[
p(h|d) = \frac{p(d|h) \cdot p(h)}{\sum_{h \in H} p(d|h') \cdot p(h')}\]

Prior Probability

Likelihood

Posterior Probability

Problem in \( \mathbb{R}^n \rightarrow \) complex

Feature parameter \( \theta \)
What is the Probability Density Function vs. Probability Distribution?

The Probability Density Function (PDF) is a function that describes the relative likelihood for a continuous random variable to take on a given value. It is defined as:

\[ p(x) \]

The Probability Distribution is the probability of a random variable taking on a certain value. It is defined as the integral of the PDF over a range:

\[ P(x) = \int_{-\infty}^{x} p(x') dx' \]

\[ p(x) \geq 0 \quad \int_{-\infty}^{\infty} p(x) \, dx = 1 \]

For a continuous random variable, the probability of it being in a specific range is given by the integral of the PDF over that range:

\[ P(x \in (a, b)) = \int_{a}^{b} p(x) \, dx \]


https://brilliant.org/wiki/multivariate-normal-distribution
How can we use this in medicine to make predictions?

\[ \mathcal{D} = x_{1:n} = \{x_1, x_2, \ldots, x_n\} \]

\[ p(\mathcal{D}|\theta) \]

\[
p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) \ast p(\theta)}{p(\mathcal{D})}
\]

posterior = \frac{\text{likelihood} \ast \text{prior}}{\text{evidence}}

The inverse probability allows to learn from data, infer unknowns, and make predictions.
How do we reason under uncertainty for decision making in medicine?

- Take patient information, e.g., observations, symptoms, test results, -omics data, etc. etc.
- Reach conclusions, and **predict** into the future, e.g. how likely will the patient be ...
- Prior = belief before making a particular observation
- Posterior = belief after making the observation and is the prior for the next observation – intrinsically incremental

\[
p(x_i | y_j) = \frac{p(y_j | x_i)p(x_i)}{\sum p(x_i, y_j)p(x_i)}
\]
Expectation and Expected Utility Theory
Why does uncertainty matter?

\[
\mathbb{E}[f] = \int p(x) f(x) \, dx
\]

\[
\mathbb{E}[f] \approx \frac{1}{N} \sum_{n=1}^{N} f(x_n)
\]

Image by Katharina Holzinger
What is the result of the Expected Utility Theory $E(U|d)$?

For a single decision variable an agent can select $D = d$ for any $d \in \text{dom}(D)$. 

The expected utility of decision $D = d$ is

$$E(U \mid d) = \sum_{x_1, \ldots, x_n} P(x_1, \ldots, x_n \mid d) U(x_1, \ldots, x_n, d)$$

An optimal single decision is the decision $D = d_{max}$ whose expected utility is maximal:

$$d_{max} = \arg \max_{d \in \text{dom}(D)} E(U \mid d)$$

Example: Learning to complement humans

**Discriminative approaches**

**Fixed discriminative**: Predictive model trained in isolation, followed by training of policy for querying

**Joint discriminative**: Predictive model and policy trained jointly

**Decision-theoretic approaches**

**Fixed VOI**: Predictive model trained in isolation, followed-up with principled VOI analysis

**Joint VOI**: Predictive model and policy trained jointly, with principled VOI analysis.

\[ E_{(x,y,h)} \sim P \left[ q_\theta(x) \left( u(y, m_\theta(x, h)) - c \right) \\ + (1 - q_\theta(x)) \left( u(y, m_\theta(x)) \right) \right] \]

\[ u_{q(h|x)} = \max_{\hat{y} \in \mathcal{Y}} \left( \sum_{y \in \mathcal{Y}} p_{\alpha}(y|x) u(\hat{y}, y) \right) \]

\[ u_q = E_{h \sim p_\beta(h|x)} \left[ \max_{\hat{y} \in \mathcal{Y}} \left( \sum_{y \in \mathcal{Y}} p_{\gamma}(y|x, h) u(\hat{y}, y) \right) \right] - c \]

**Algorithm 1** Joint VOI training

1: **for** \( T \) iterations **do**
2: \( \text{Sample a minibatch } B \subseteq [n] \)
3: **for** \( i \in B \) **do**
4: \( \text{for } \hat{y} \in \mathcal{Y} \) **do**
5: \( u_{q(h)} = \sum_{y \in \mathcal{Y}} p_{\alpha}(y|x_i) u(\hat{y}, y) \)
6: **end for**
7: \( u_{q} = \sum_{\hat{y} \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} \exp(u_{q(h)}(\hat{y}, y)) \)
8: \( \text{for } \hat{y} \in \mathcal{Y} \) **do**
9: \( u_q(\hat{y}, h) = \sum_{y \in \mathcal{Y}} p_{\gamma}(y|x_i, h) u(\hat{y}, y) \)
10: **end for**
11: \( u_q = \sum_{h \in \mathcal{Y}} p_{\beta}(h|x) \sum_{\hat{y} \in \mathcal{Y}} \exp(u_{q(h)}(\hat{y}, h)) \)
12: \( q = \frac{\exp(u_{q(h)})}{\exp(u_{q})} \)
13: \( \ell_{i}^\text{combined} = \ell(q p_{\gamma}(h|x_i, h)) \)
14: \( + (1 - q) p_{\alpha}(h|x_i) + q c \)
15: **end for**
16: **Backpropagate** \( \frac{1}{|B|} \sum_{i \in B} \ell_{i}^\text{combined} \)
17: **Every r iterations: update calibrators**
18: **end for**
What was the origin of probabilistic decision making?

Image source: Staffordshire University Computing Futures Museum http://www.fcet.staffs.ac.uk/jdw1/sucfm/malvern.htm

How does this work in medical decision making?

Two doctors, with equally good training, looking at the same CT scan, will have the same information ... but they may have a **different bias/criteria**!
What does a correct rejection mean?

Remember: Two doctors, with equally good training, looking at the same CT scan data, will have the same information … but they may gain different knowledge due to bias/criteria.

Positive = identified and negative = rejected
True positive = correctly identified (hit)
False positive = incorrectly identified, false alarm, type I error
True negative = correctly rejected (correct rejection)
False negative = incorrectly rejected, miss, type II error


And please look up the Wikipedia page:
Why do we need specificity and sensitivity?

Andrew P. Bradley 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30, (7), 1145-1159, doi:http://dx.doi.org/10.1016/S0031-3203(96)00142-2
What is the advantage of the concordant pAUC?

Why do we need to take care of ROC for understanding & interpretation?

- Optimal performance is critical for decision-making.
- Common performance measures may be too general or too specific.
- AUC too general because including unrealistic decision thresholds.
- Accuracy, sensitivity or the F1 score are measures at a single threshold that reflect an individual single probability or predicted risk, rather than a range of individuals or risk.
- Deep ROC examines groups of probabilities or predicted risks for more insightful analysis.
- That can improve model selection in some cases and provide interpretation and assurance for patients in each risk group.


[https://github.com/Big-Life-Lab/deepROC](https://github.com/Big-Life-Lab/deepROC)
What is the principle of Bayesian diagnostic in medicine?

How does Bayesian inference work?

Your MD has bad news and good news for you.

Bad news first: You are tested positive for a serious disease, and the test is 99% accurate if you are infected (T)

Good news: It is a rare disease, striking 1 in 10,000 (D)

How worried would you now be?

$posterior \ p(x) = \frac{\text{likelihood} \ast \text{prior} \ p(x)}{\text{evidence}}$

$p(h|d) = \frac{p(d|h)p(h)}{p(d)}$

$p(T = 1|D = 1) = p(d|h) = 0,99 \ \text{and}$

$p(D = 1) = p(h) = 0,0001$

$p(D = 1 \mid T = 1) = \frac{(0,99) \ast (0,0001)}{(1-0,99) \ast (1-0,0001)+0,99 \ast 0,0001} = 0,0098$
What is Inference?

Inference

Parameters

Program

Output

Parameters

Program

Observations

\[ p(x|y) \]

\[ p(y|x)p(x) \]

\[ y \]

CS     Probabilistic Programming     Statistics

What is the Connection to Cognitive Science/Decision Making?

Can we scale into the high dimensions?

What is a Gaussian process?

How can we reduce uncertainty?

\[
\mathbb{E}[f] = \int p(x) f(x) \, dx
\]

\[
\mathbb{E}[f] \approx \frac{1}{N} \sum_{n=1}^{N} f(x_n)
\]

Why do we take the human-out-of-the-loop?

 Algorithm 1 Bayesian optimization

1: for \( n = 1, \ldots \) do
2: select \( x_{n+1} \) by optimizing acquisition function \( \alpha \)
3: query objective \( y_{n+1} \)
4: at \( x_{n+1} \), \( y_{n+1} \)
5: end for

**Taking the human out of the loop**: A review of Bayesian optimization. 
(4) aML
What does level 5 autonomy mean?

What does a recommender system do?

Where can autonomous machine learning help?

**Cyber-Physical Systems (CPS):**
*Tight integration of networked computation with physical systems*

What is a best practice example for the success of Deep Learning?

What do we need: more data or better data?

How much training data does deep learning need?

\[ x^* = \arg \min_x f(x; W, H), \quad \text{subject to } ||x||_2 = 1. \]


When does deep learning fail?


- Sometimes we do not have “big data”, where aML-algorithms benefit.
- Sometimes we have
  - Small amount of data sets
  - Rare Events – no training samples
- NP-hard problems, e.g.
  - Subspace Clustering,
  - k-Anonymization,
  - Protein-Folding, ...

[https://human-centered.ai/project/iml](https://human-centered.ai/project/iml)
What is now our big problem with deep learning approaches?

- High dimensionality (curse of dim., many factors contribute)
- Complexity of medical problems (medical world is non-linear, non-stationary, non-IID *)
- Need of large top-quality data sets
- Sensitive to small disturbances (noise, bias, one-pixel attacks, ...)
- Little prior data (no mechanistic models of the data)
  - *) = Def.: a sequence or collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent
- However, most of all ...
Why are such approaches difficult in terms of the EU GDPR?
05 iML
What is interactive machine learning?

- iML := algorithms which interact with agents*) and can optimize their learning behaviour through this interaction
  
*) where the agents can be human

Humans can understand abstract concepts!

Image Source: 10 Ways Technology is Changing Healthcare http://newhealthypost.com Posted online on April 22, 2018
Image is used according UrhG §42 lit. f Abs 1 as “Belegfunktion” for discussion with students
Why are humans Bayesian reasoners?

This was presented on December, 8, 2002 as Nobel Prize Lecture by Daniel Kahneman from Princeton University, and has later been published as:


What is Deductive vs Inductive vs Abductive Reasoning?

- **Deductive Reasoning** = Hypothesis > Observations > Logical Conclusions
  - DANGER: Hypothesis must be correct! DR defines whether the truth of a conclusion can be determined for that rule, based on the truth of premises: A=B, B=C, conclusion: A=C

- **Inductive reasoning** = makes broad generalizations from specific observations
  - DANGER: allows a conclusion to be false if the premises are true
  - generate hypotheses and use DR for answering specific questions

- **Abductive reasoning** = inference = to get the best explanation from an incomplete set of preconditions.
  - Given a true conclusion and a rule, it attempts to select some possible premises that, if true also, may support the conclusion …
  - Example: "When it rains, the grass gets wet. The grass is wet. Therefore, it might have rained." This kind of reasoning can be used to develop a hypothesis, which in turn can be tested by additional reasoning or data.

Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg & Andreas Holzinger 2018

What is ground truth?

- := information provided by direct observation (empirical evidence) in contrast to information provided by inference
- Empirical evidence = information acquired by observation or by experimentation in order to verify the truth (fit to reality) or falsify (non-fit to reality).
- Empirical inference = drawing conclusions from empirical data (observations, measurements)
- Causal inference = drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect.
  - Causal inference is an example of causal reasoning.

Even Children can make inferences from little, noisy, incomplete data ...

What can we do with rare examples in typical classification tasks?

Feature $y$
e.g. ear-geometry

Feature $x$
e.g. snout-length
What is the difference between Correlation and Causation?

\[ x \times 0.007 \times \frac{\text{sign}(\nabla_x J(\theta, x, y))}{\text{sign}(\nabla_x J(\theta, x, y))} = x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)) \]

Panda: 57.7% confidence
Nematode: 8.2% confidence
Gibbon: 99.3% confidence


What is adversarial machine learning?

Adversarial Examples that Fool both Computer Vision and Time-Limited Humans

Gamaleldin F. Elsayed*  
Google Brain  
gamaleldin.elsayed@gmail.com

Shreya Shankar  
Stanford University

Brian Cheung  
UC Berkeley

Nicolas Papernot  
Pennsylvania State University

Alex Kurakin  
Google Brain

Ian Goodfellow  
Google Brain

Jascha Sohl-Dickstein  
Google Brain  
jaschasd@gmail.com

Abstract

Machine learning models are vulnerable to adversarial examples: small changes to images can cause computer vision models to make mistakes such as identifying a school bus as an ostrich. However, it is still an open question whether humans are prone to similar mistakes. Here, we address this question by leveraging recent techniques that transfer adversarial examples from computer vision models with known parameters and architecture to other models with unknown parameters and architecture, and by matching the initial processing of the human visual system. We find that adversarial examples that strongly transfer across computer vision models influence the classifications made by time-limited human observers.

Correlation ≠ Causality
Why we need the Human-in-the-loop
Where is the benefit of the human-in-the-loop?

iML = human inspection – bring in human “intuition” – abstract concept learning and context understanding!

Hans Reichenbach (1891-1953): Common Cause Principle
Links causality with probability:
- If A and B are statistically dependent, there is a C influencing both
- Whereas:
  - A, B, C ... events
  - p ... probability density

Remember: Correlation is NOT Causality

How can we make AI more robust?

Generalization error

Generalization error plus human experience

iML = human inspection – bring in human conceptual knowledge

(Sometimes – not always!) humans are able ...

- to understand the context
- to make inferences from little, noisy, incomplete data sets
- to learn relevant representations
- to find shared underlying explanatory factors,
- with a causal reasoning
  \[ P(Y|X) \quad Y \rightarrow X \text{ (predict cause from effect) or} \]
  \[ P(Y|X) \quad X \rightarrow Y \text{ (predict effect from cause)} \]

(6) explainable AI and Methods of Explainability
What are the expectations for explainable AI?

- **Trust** – interpretability as prerequisite for trust (as propagated by Ribeiro et al (2016)); how is trust defined? Confidence?

- **Causality** – inferring causal relationships from pure observational data has been extensively studied (Pearl, 2009), however it relies strongly on prior knowledge.

- **Transferability** – humans have a much higher capacity to generalize, and can transfer learned skills to completely new situations; compare this with e.g. susceptibility of CNNs to adversarial data (please remember that we rarely have iid data in real world).

- **Informativeness** – for example, a diagnosis model might provide intuition to a human decision-maker by pointing to similar cases in support of a diagnostic decision.

- **Fairness and Ethical decision making** – interpretations for the purpose of assessing whether decisions produced by algorithms conform to ethical standards, avoiding bias and misconceptions..

What are interpretable models vs. interpreting models?

- **Interpretable Models**, = ante-hoc - the "glass-box" model itself is ante-hoc interpretable, e.g. Regression, Naïve Bayes, Decision Trees, Graphs, ...

- **Interpreting Black-Box Models**, = post-hoc - the model is not interpretable and needs a post-hoc interpretability method $M$

Why is explainable AI only a first step in human-centered AI?

What are typical examples of post-hoc methods of explainable AI?

- 1) Gradients
- 2) Sensitivity Analysis
- 3) Simple Taylor expansions
- 5) Excitation Backpropagation
- 6) Optimization (LIME, BETA, Smooth Grad, ...)
  BETA transparent approximation, ...)
- 7) Deconvolution (Occlusion-based, meaningful perturbations, ...)
- 8) Qualitative Testing with Concept Activation Vectors TCAV
LRP Layer-Wise Relevance Propagation

\[ f(x) \approx \sum_{d=1}^{V} R_d \]

\[ R_i = \left\| \frac{\partial}{\partial x_i} f(x) \right\| \]

For forward propagation:

\[ a_j^{(l+1)} = \sigma \left( \sum_i a_i^{(l)} w_{ij} + b_j^{(l+1)} \right) \]

For layer-wise relevance propagation:

\[ R_i^{(l)} = \sum_j \sum_{k} z_{ij} R_j^{(l+1)} \]


(7) Causability measures the quality of explanations obtained from (6).
Example: How do human pathologists make diagnoses?
What is ground truth? Where is the ground truth?

- := information provided by direct observation (empirical evidence) in contrast to information provided by inference
  - Empirical evidence = information acquired by observation or by experimentation in order to verify the truth (fit to reality) or falsify (non-fit to reality).
  - Empirical inference = drawing conclusions from empirical data (observations, measurements)
  - Causal inference = drawing conclusions about a causal connection based on the conditions of the occurrence of an effect
  - Causal machine learning is key to ethical AI in health to model explainability for bias avoidance and algorithmic fairness for decision making

KANDINSKYPatterns – where does the name come from?


Note: Image is in the public domain and is used according to UrhG §42 lit. f Abs 1 as “Belegfunktion” for discussion with students.

Wassily Kandinsky (1866 – 1944)
Hubel & Wiesel (1962): Our world is compositional!


Source: https://www.intechopen.com/books/visual-cortex-current-status-and-perspectives/models-of-information-processing-in-the-visual-cortex
Definition 1: A Kandinsky Figure is ...

- ... a square image containing 1 to \( n \) geometric objects.
- Each object is characterized by its shape, color, size and position within this square.
- Objects do not overlap and are not cropped at the border.
- All objects must be easily recognizable and clearly distinguishable by a human observer.
Related Work (1): Bongard Problems

Bongard, M. Mikhail, 1967. The problem of recognition (in Russian), Moscow, Nauka (1970 in English)

Douglas R. Hofstadter (1979)
Related Work (2): CLEVR

https://cs.stanford.edu/people/jcjohs/clevr/

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C. Lawrence Zitnick & Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017 Hawaii. IEEE.
Related Work (3): CLEVRER

Related Work (4): Bongard-LOGO


(a) free-from shape problem

(b) basic shape problem

(c) abstract shape problem
Related Work (5): CURI

for-all \( x \in S \) (color?(x) = “blue”) and (all (size?(S) = size?(x)))

All objects are blue and have the same size

for-all \( x \in S \) (all (color?(x) = color?(S)))

All objects in the scene have the same color

exists \( x \in S \) (color?(x) = “blue”) and all (shape?(S_{\{-x\}}) = “square”)

There exists a blue object in the scene and the rest of the objects are squares

\( G \):
Context Free Grammar

Variables
\( x \triangleq \text{Object in scene} \)
\( S \triangleq \text{All objects} \)
\( S_{\{-x\}} \triangleq S/\{x\} \)

Quantifiers
for-all exists

Functions
color? location? shape? size? material? all

Operators
and Greater(>) or Lesser(<) not =

Definition 2 A statement $s(k)$

- about a Kandinsky Figure $k$ is ...
- either a mathematical function $s(k) \rightarrow B$; with $B (0,1)$
- or a natural language statement which is true or false

- The evaluation of a natural language statement is always done in a specific context.
- we follow well known concepts from human perception and linguistic theory.
- If $s(k)$ is given as an algorithm, it is essential that the function is a pure function, which is a computational analogue of a mathematical function.

“Intelligence Test for Machines”

Difficult Concepts

Part of the pattern

Not part of the pattern

S8  Basic Pattern 8
Title: Mickey Mouse
Every figure contains a pattern which is made out of a big yellow circle and two smaller blue ones and looks like a Mickey Mouse.
A) True (the cells are smaller and closer together – it is a tumor ...)

B) False

C) Counterfactual (What if the cells are slightly bigger ?)
#KANDINSKYPatterns our Swiss-Knife for the study of explainable-AI

**ABSTRACT**

KANDINSKYPatterns (inspired after the famous artist Wassily Kandinsky) are mathematically describable, simple, self-contained, hence controllable test data sets for the development, validation and training of explainability in artificial intelligence (AI) and machine learning (ML). Whilst our KANDINSKY Patterns have these computationally manageable properties, they are at the same time easily distinguishable from human observers. Consequently, patterns can be described by both humans and algorithms.

We define a KANDINSKY Pattern as a set of KANDINSKY Figures, where for each figure an "infallible authority" (ground truth) defines that this figure belongs to the KANDINSKY Pattern. With this simple principle we build training and validation data sets for automatic interpretability and context learning.

**KANDINSKY PATTERNS AT TEDX**

**KANDINSKY ARTIFICIAL INTELLIGENCE EXPLANATION CHALLENGE**

Here we challenge the international machine learning community to generate machine explanations.

**KANDINSKY HUMAN INTELLIGENCE EXPLANATION CHALLENGE**

Here we challenge any human individual to take part in this experiment and to generate human explanations.

**HCAI GITHUB REPOSITORY**

How can we measure the quality of explanations?

Measuring the quality of Explanations: The Systems Causability Scale

Causability is neither a typo nor a synonym for Causality
Causa-bil-ity ... in reference to ... Usa-bil-ity.
While xAI is about implementing transparency and traceability, Causability is about the measurement of the quality of explanations.

**Explainability** := technically highlights decision relevant parts of machine representations and machine models i.e., parts which contributed to model accuracy in training, or to a specific prediction.
- Explainability does not refer to a human model!

**Causability** := the measurable extent to which an explanation of a statement to a user achieves a specified level of causal understanding with effectiveness, efficiency, satisfaction in a specified context of use.
- Causability does refer to a human model!

How can we measure the quality of explanations?

Conclusio
Constructing a multi-modal interaction & correspondence graph (ICG)

Interesting signals from each modality (time-based, image, structured & unstructured) are connected according to pre-defined rules. Each modality’s features lie in their own, un-aligned concept spaces.

Explainability needs a framework to ensure common understanding and adaptive Question/Answering Interfaces

Thank you very much!
Appendix
“The most interesting facts are those which can be used several times, those which have a chance of recurring ... which, then, are the facts that have a chance of recurring? In the first place, simple facts.”

Henri Poincare (1854-1912), Sciences et Methods (1908), (1913)