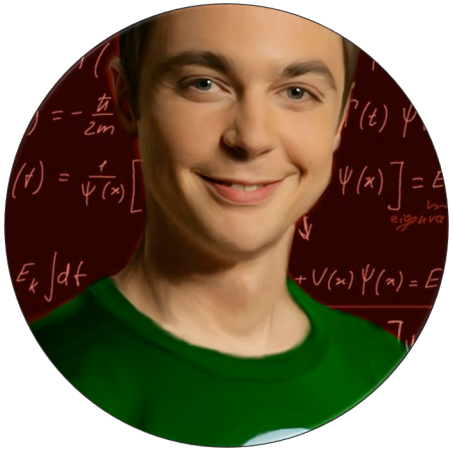
Codettes & Geekettes Code Pub Munich
Data Science Workshop

Dr. Lydia Nemec & Veronika Zellner
Data Scientist
2019-10-12

Dr. Lydia Nemec

Data Scientist @ Zeiss
Theoretical Physicist by training

but I come in a wrapping similar to Penny



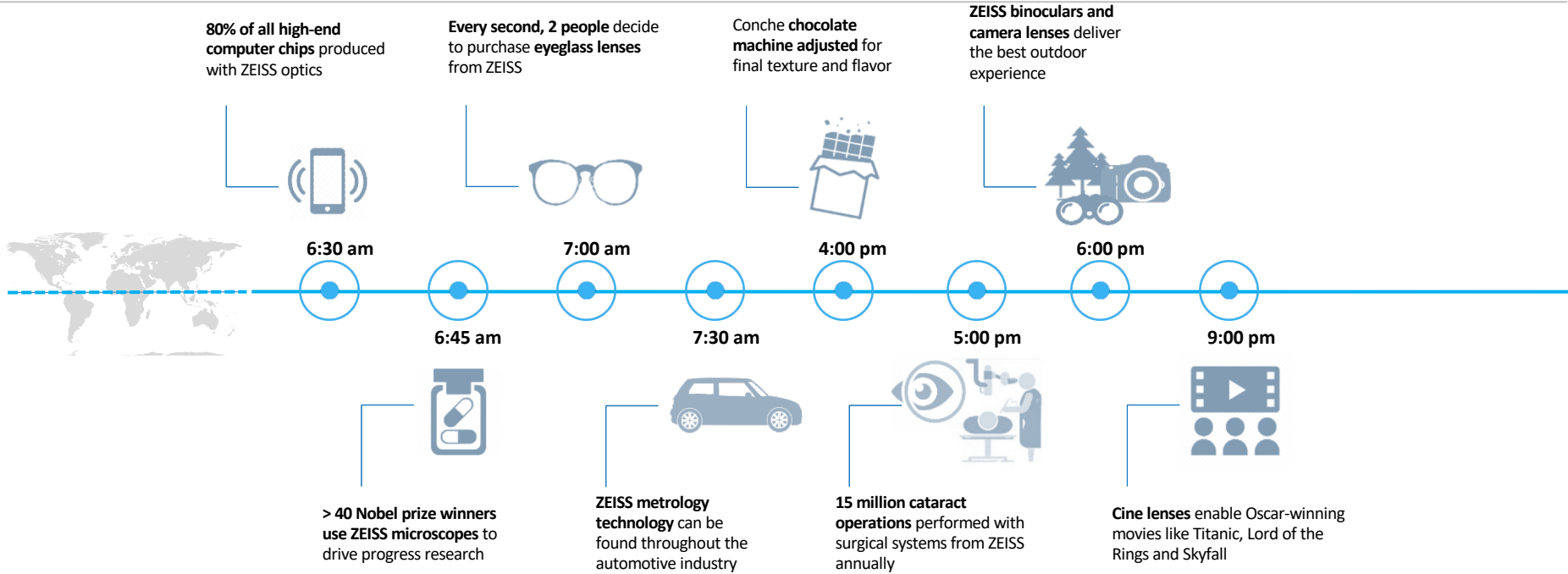
I am a bit like Dr. Sheldon Cooper,

Veronika Zellner
Mathematician
Architect for Data & AI @Microsoft

I´m am nerdy

But also girly

One day without ZEISS?
A “user journey”



~ EUR 6b in revenue

~30,000 employees

500 patent applications

11% R&D investments

25 global R&D sites

ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

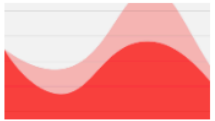
DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

WHAT TO READ NEXT



Visualizations That Really Work

T.H. Davenport and D.J. Patil; "Data Scientist: The Sexiest Job of the 21st Century" Havard Business Review (10/12)

The Skillset of a Data Scientist

Data Scientists apply numerical methods like Machine Learning to extract insights from data.

Math, Numeric & Statistic

- ❑ Machine Learning (AI)
- ❑ Statistical modelling
- ❑ Linear Algebra & Optimization

The Scientific Mind

- ❑ Logical & independent mind
- ❑ Planning, conducting & evaluate experiments
- ❑ Excellent analytical skills
- ❑ Meticulous attention to quality and accuracy

Computer Science & Programming

- ❑ Software development
- ❑ Programming Language (e.g. python)
- ❑ Databases (SQL/ No-SQL)
- ❑ Cloud Computing

Communication, Soft Skills & Visualisation

- ❑ Collaborative, strategic, proactive, creative and innovative
- ❑ Influence without authority
- ❑ Translate data-driven insights into impactful decisions and actions
- ❑ Data Visualisation

Data Science

combines the complexity of

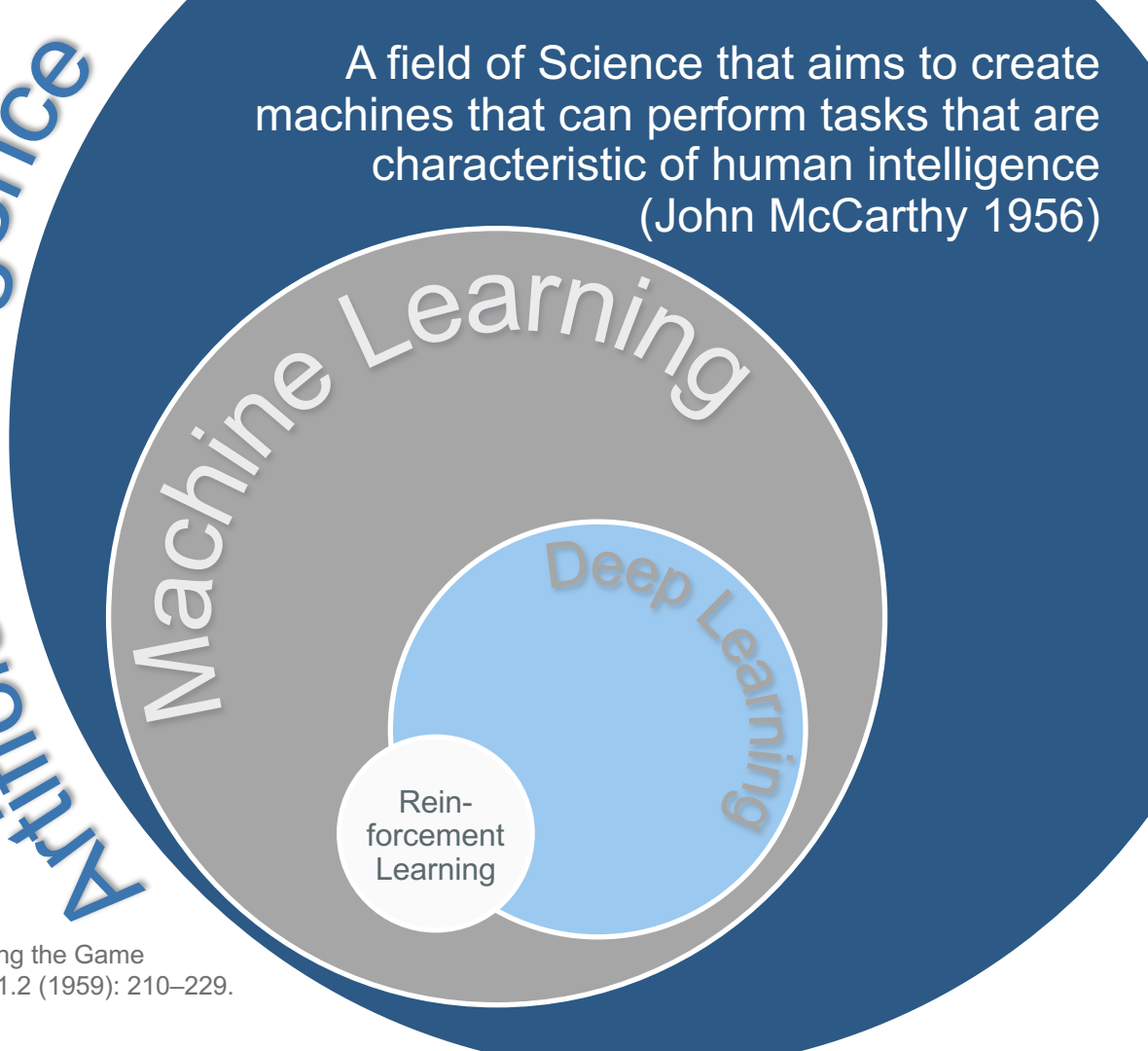
Software Development,

the challenges of applied numerical analysis

with the additional dynamic introduced by data!

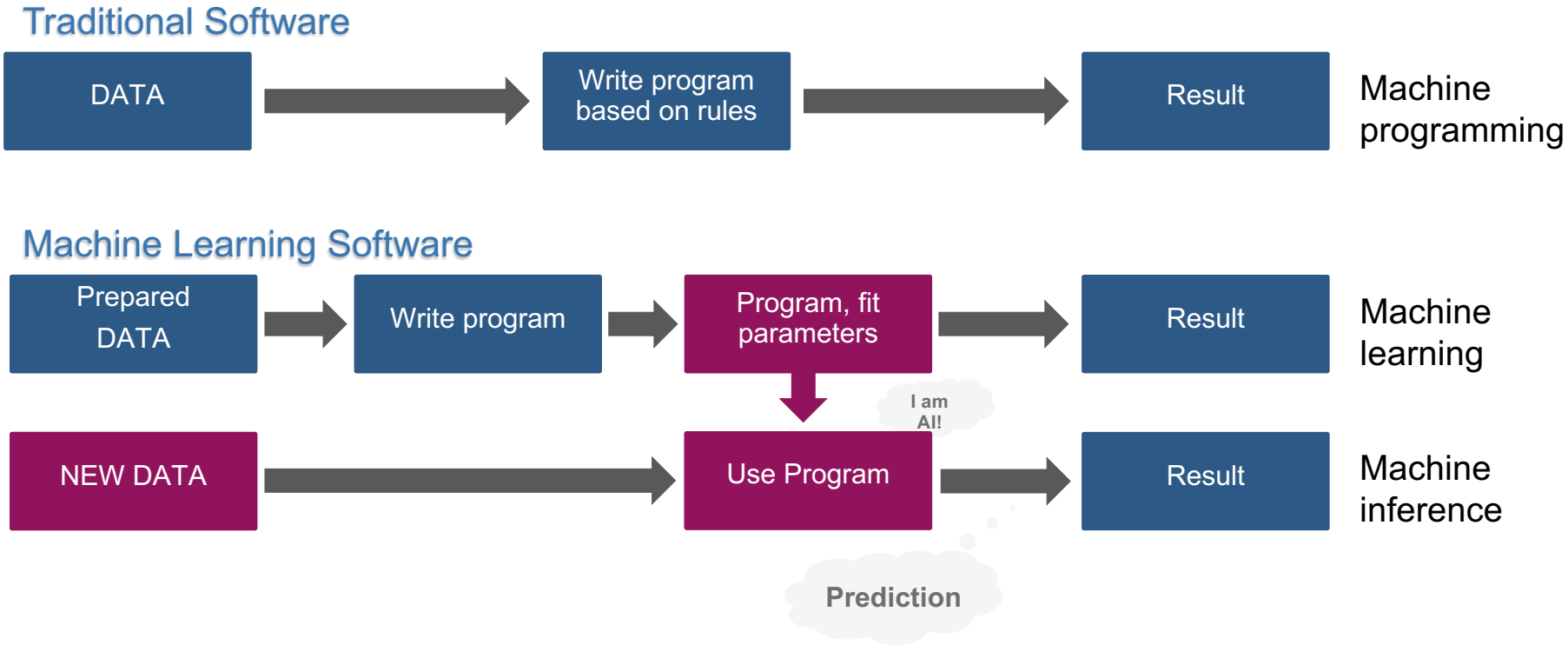
Machine Learning
refers to a set of algorithms that allow computers to learn from data without being explicitly programmed. [1]

Deep Learning
is part of a broader family of machine learning methods based on artificial neural networks. It belongs to the class of hierarchical learning algorithm.



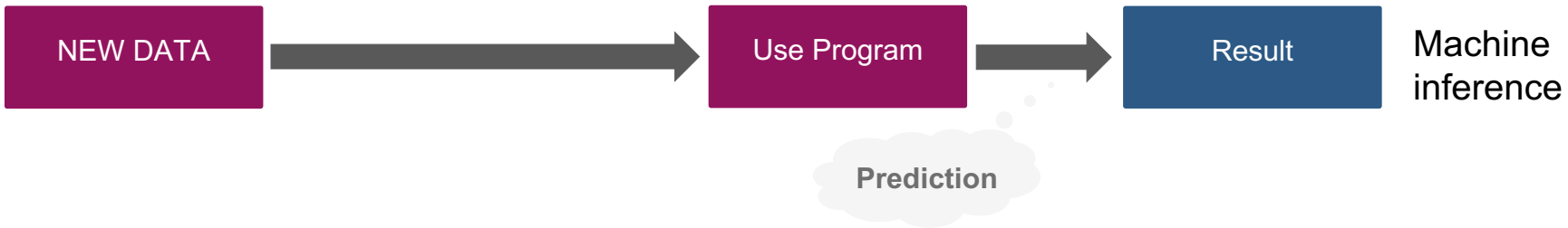
[1] Samuel, Arthur L. „Some Studies in Machine Learning Using the Game of Checkers,“ IBM Journal of Research and Development 44:1.2 (1959): 210–229.

Machine Learning: A different way of software development



Machine Learning: What do we mean with “Prediction”?

Prediction
is the process of filling in missing information!
Prediction takes information you have, often called ”data”,
and uses it to extract information you need.



Ajay Agrawal, Avi Goldfarb, Joshua Gans «Prediction Machines: The Simple Economics of Artificial Intelligence”, Harvard Business Review (04/18)

Machine Learning cost asymmetry in training and inference

Machine Learning Software

Higher computational cost during training

Prepared DATA → Write program → Program, fit parameters → Result

NEW DATA → Use Program → Result

Lower computational cost during inference

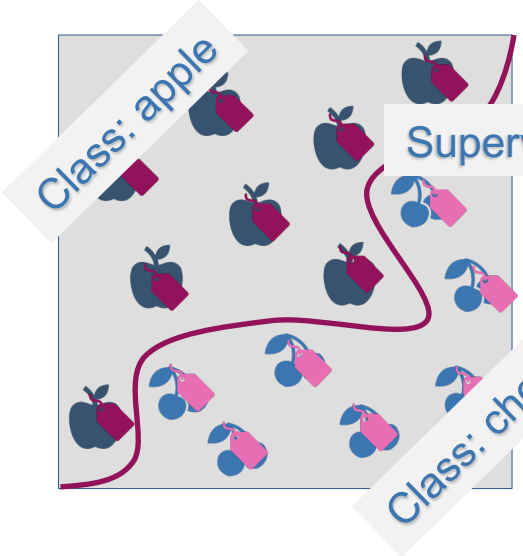
Summit Oak Ridge National Lab since November 2018 fastest supercomputer in the world

Machine Learning: Clustering, Classification & Regression

Clustering

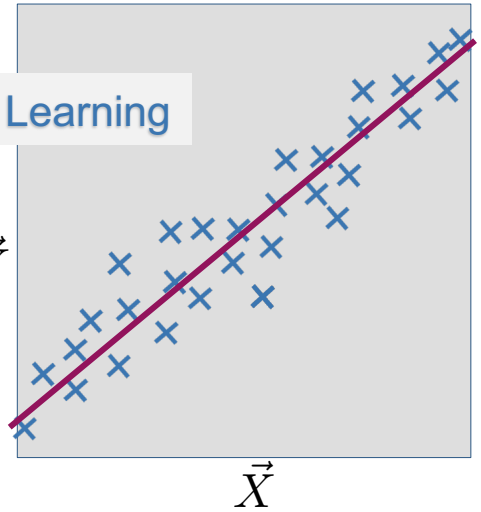
- Clustering algorithm try to find structure or pattern in uncategorized data.

Classification



- Clustering algorithm map input data to output labels.

Regression



- Regression algorithm map input data to a continuous output value.

Classification and regression model development needs labelled input data.

Machine Learning: Supervised and unsupervised learning

Unsupervised Learning

- In unsupervised learning model, only input data is given.
- Uses unlabelled data
- Number of classes is not known.
- Less accurate and trustworthy method.
- Difficult to validate the reliability of the Machine Learning model.
- Algorithm examples: Cluster algorithms (K-means, Hierarchical clustering) Signal separation (Principal & Independent component analysis) Neural Networks (autoencoders), anomaly detection (isolation forest).

Supervised Learning

- In a supervised learning model, input and output variables are given.
- Uses labelled data
- Learning a link between the input and the outputs.
- Possible to develop highly accurate and trustworthy method.
- Difficult to get a sufficient amount of high quality labelled data.
- Algorithm examples: Support vector machine, (deep) neural network, Linear and logistics regression, random forest, and Classification trees (forests)

Machine Learning: A fitting problem

Linear Regression; the simplest example of a Machine Learning Algorithm

$$\vec{Y} \approx f(\vec{X}; \{p_1, p_2, ..., p_N\})$$

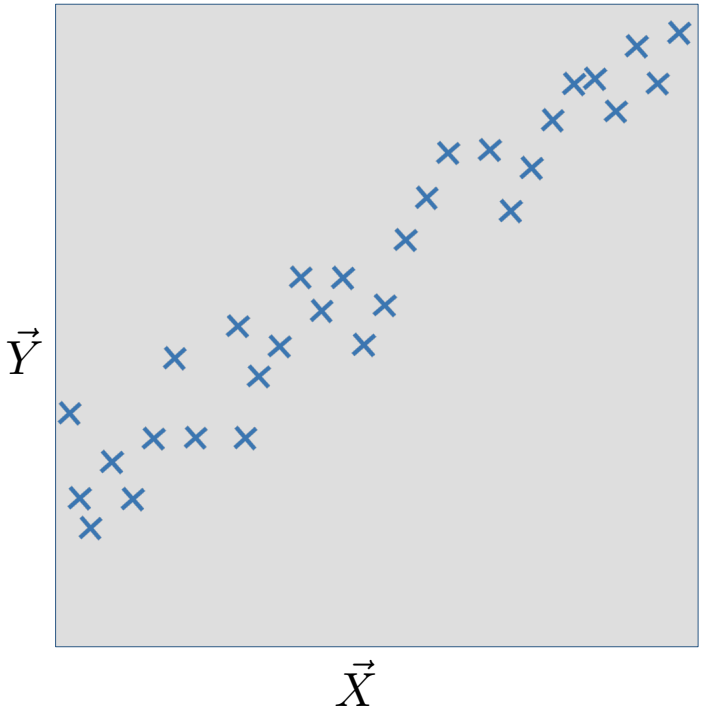
- The function f describes the relationship between $\vec{X}$ and $\vec{Y}$.
- In general, f is not known.
- $\vec{X}$ and $\vec{Y}$ can be multidimensional vectors

Example:

x duration of a taxi drive in seconds
y price of the Taxi drive

Example multidimensional Input:

$\vec{X}$ (x_1) duration and (x_2) distance of a taxi drive
y price of the Taxi drive



Machine Learning: A fitting problem

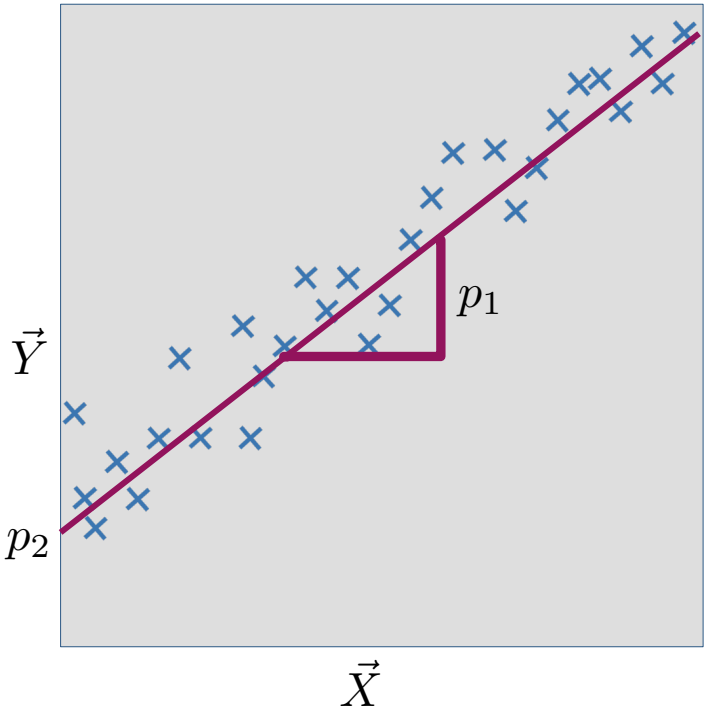
Linear Regression; the simplest example of a Machine Learning Algorithm

$$\vec{Y} \approx f(\vec{X}; \{p_1, p_2, ..., p_N\})$$

- The simplest form of f approximating the relationship between $\vec{X}$ and $\vec{Y}$ is a straight line.
- $\{p_1, p_2, ..., p_N\}$ are the parameter determining the Function f

$$Y = f(X; p_1, p_2)$$
$$N = 2$$
$$Y = p_1 X + p_2$$

- p_1 slope and p_2 Y-intersection of the linear equation
- The equation is fully defined by p_1, p_2



Machine Learning: A fitting problem

Linear Regression; the simplest example of a Machine Learning Algorithm

$$\vec{Y} \approx f(\vec{X}; \{p_1, p_2, ..., p_N\})$$

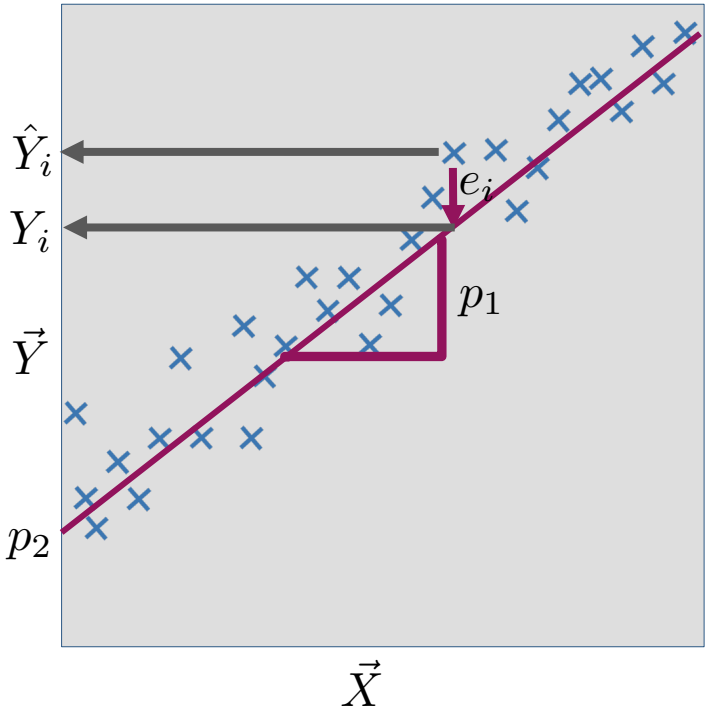
- Error e_i between the function value Y_i and the data point $\hat{Y}_i$

$$e_i = |Y_i - \hat{Y}_i|^2$$

- The machine Learning part is finding p_1, p_2
- p_1, p_2 is determined by minimising the sum of e_i

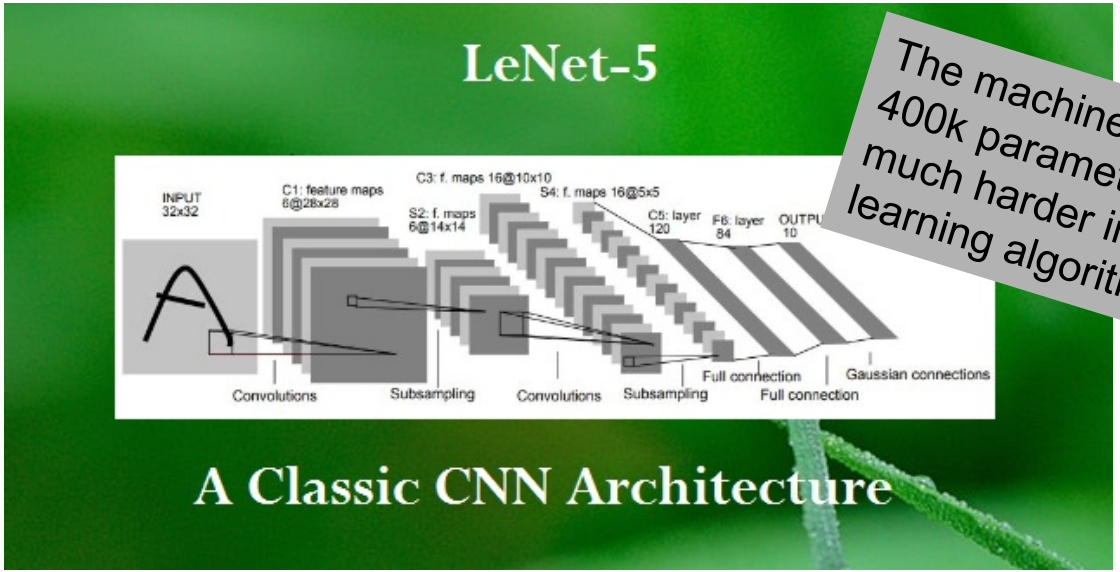
$$MSE = \frac{1}{N} \sum_{i=1}^{N} |Y_i - \hat{Y}_i|^2$$

- Using gradient descent: updating p_1, p_2 to reduce the cost function (MSE).



Machine Learning: A fitting problem

Example of a Machine Learning Algorithm: Deep neural network for computer vision

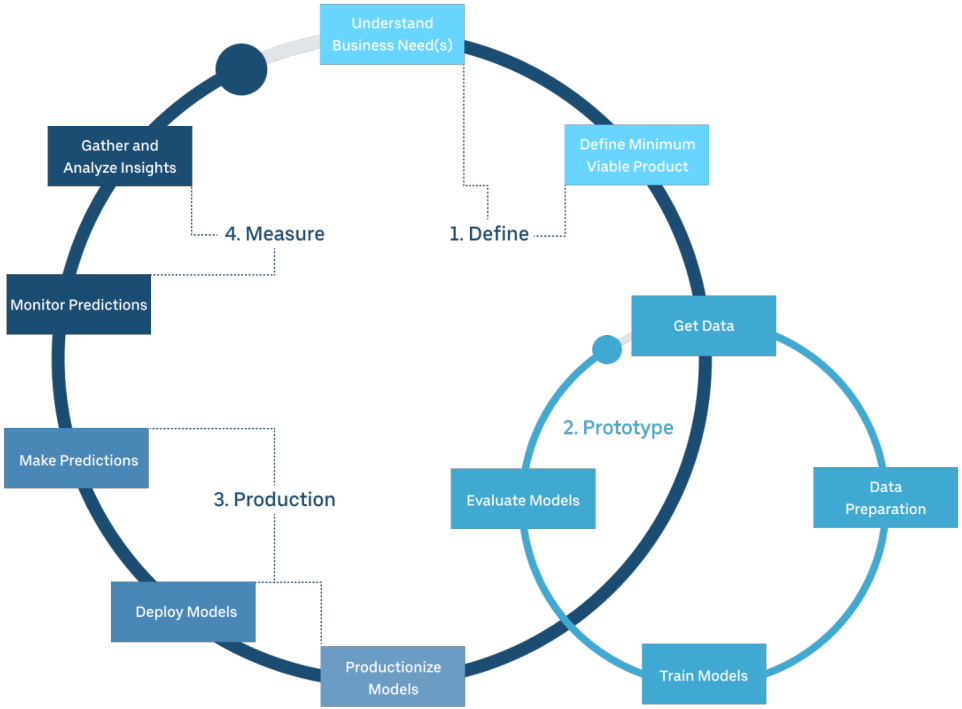


The machine Learning part of fitting the 400k parameter $\{p_1, p_2, \ldots, p_{388272}\}$ is much harder in advanced machine learning algorithm like deep neural nets.

- LeNet-5 is a typical example of a deep neural network for computer vision for a 32 x 32 pixel image the Net contains **~400k parameters** to optimise.

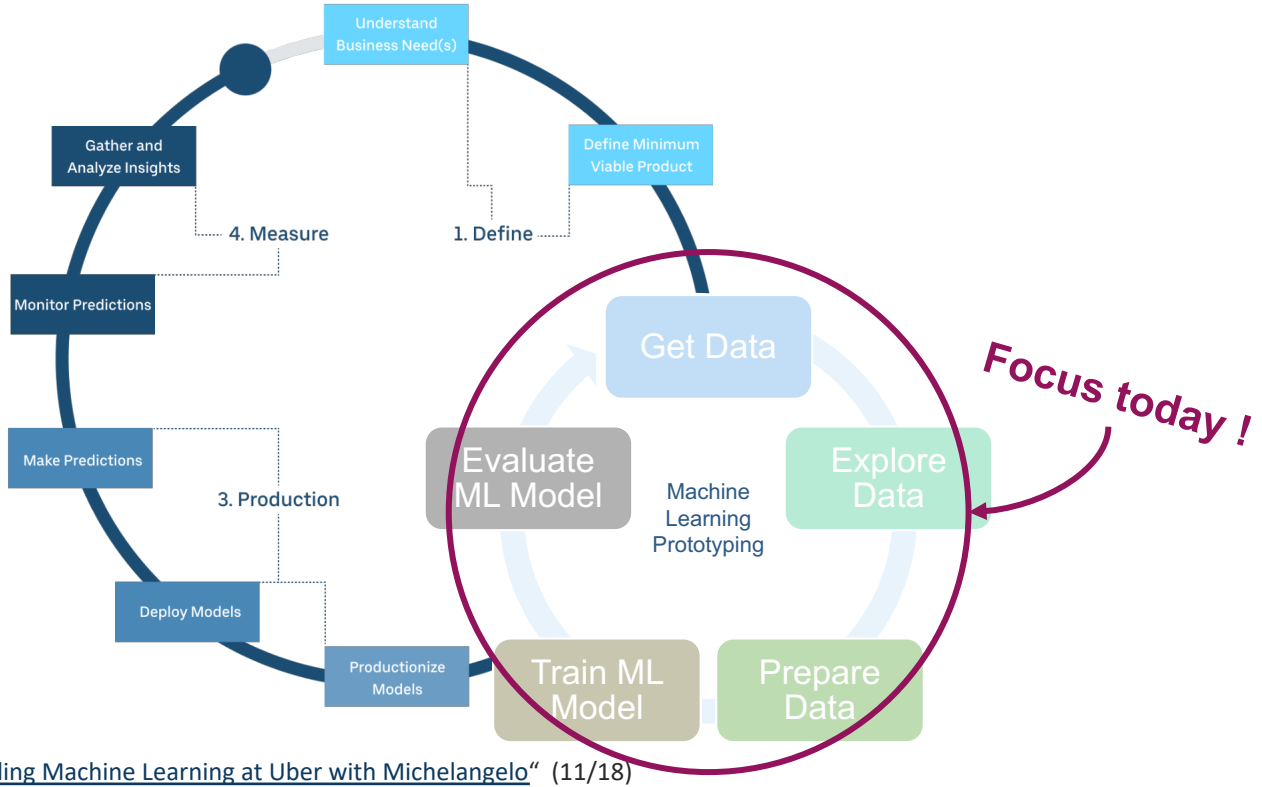
https://engmrk.com/lenet-5-a-classic-cnn-architecture/ http://yann.lecun.com/exdb/lenet/

Data Science Projects: A Highly Iterative Process



[1] Jeremy Hermann and Mike Del Balso; „Scaling Machine Learning at Uber with Michelangelo“ (11/18)

Data Science Projects: A Highly Iterative Process



[1] Jeremy Hermann and Mike Del Balso; „Scaling Machine Learning at Uber with Michelangelo“ (11/18)

Data Science: The surrounding Infrastructure is Vast and Complex

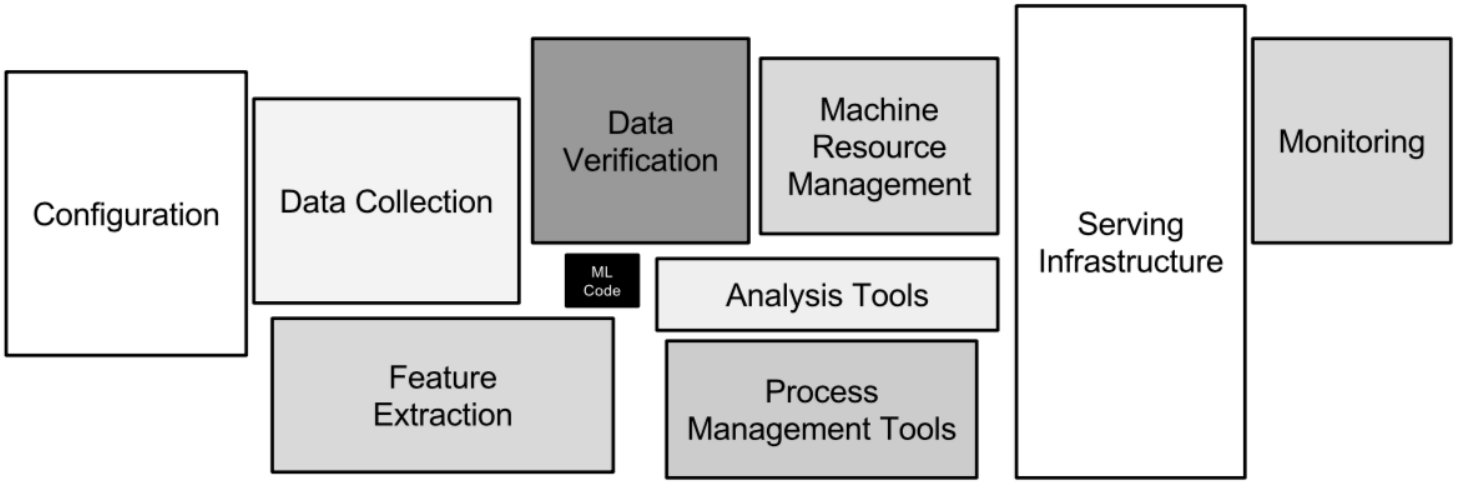


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

D. Sculley, G. Holt, D. Golovin et al.; "Hidden Technical Debt in Machine Learning Systems" Google, Inc., ISBN: 978-1510825024, pp. 2503 — 2511 NIPS 2015

The Use Case today
New York City Taxi …..

Thank you for your attention

Dr. Lydia Nemec
Data Scientist @ Zeiss

https://www.linkedin.com/in/lydianemec/

@LydiaNemec

Veronika Zellner
Architect for Data & AI @Microsoft

https://www.linkedin.com/in/veronika-zellner/


ZEISS