Lecture 1: Introduction to Machine Learning

Ivan Provilkov

Course structure

Teachers

Iurii Efimov

Research and Development Engineer, Samsung R&D Institute & AI Center

Ivan Provilkov

Head of Machine Learning at STAI

Nikolay Karpachev

Machine Learning Developer at Yandex

Course structure

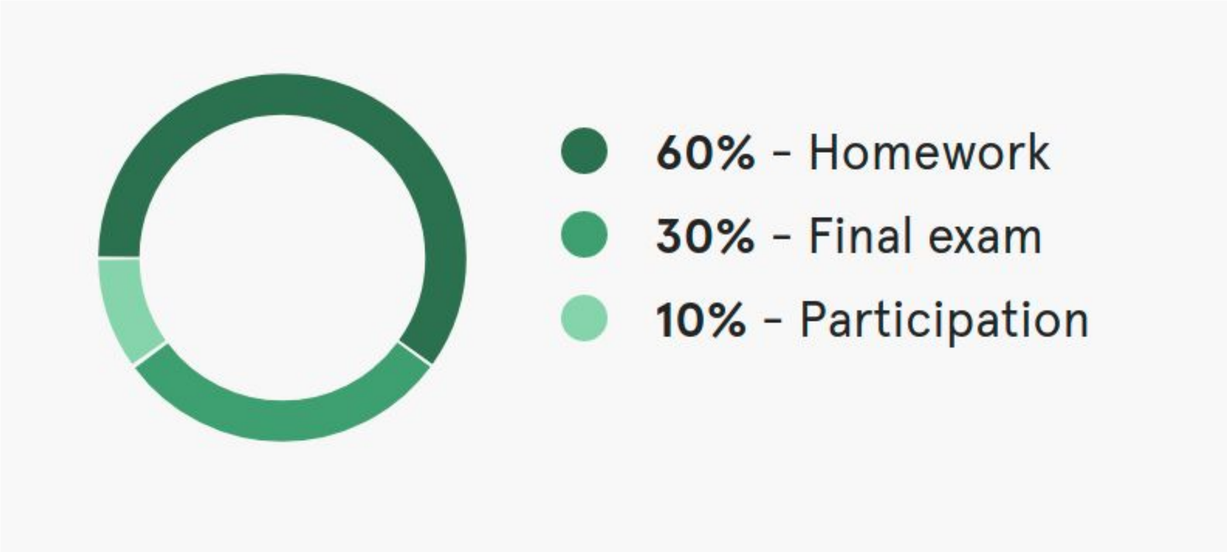
Prerequisites:

- Python
- Calculus
- Probability & Statistics
- Linear algebra

Course structure

Grading:

- Mid-term 10.02
- Exam 19.02



Course structure

Questions?

There is a lot of hype around Artificial Intelligence and Machine Learning

"Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks." — Stephen Hawking

“As a technologist, I see how AI and the fourth industrial revolution will impact every aspect of people’s lives.” — Fei-Fei Li

“Artificial intelligence would be the ultimate version of Google. The ultimate search engine that would understand everything on the web. It would understand exactly what you wanted, and it would give you the right thing. We're nowhere near doing that now. However, we can get incrementally closer to that, and that is basically what we work on.” — Larry Page

“A breakthrough in machine learning would be worth ten Microsofts” — Bill Gates

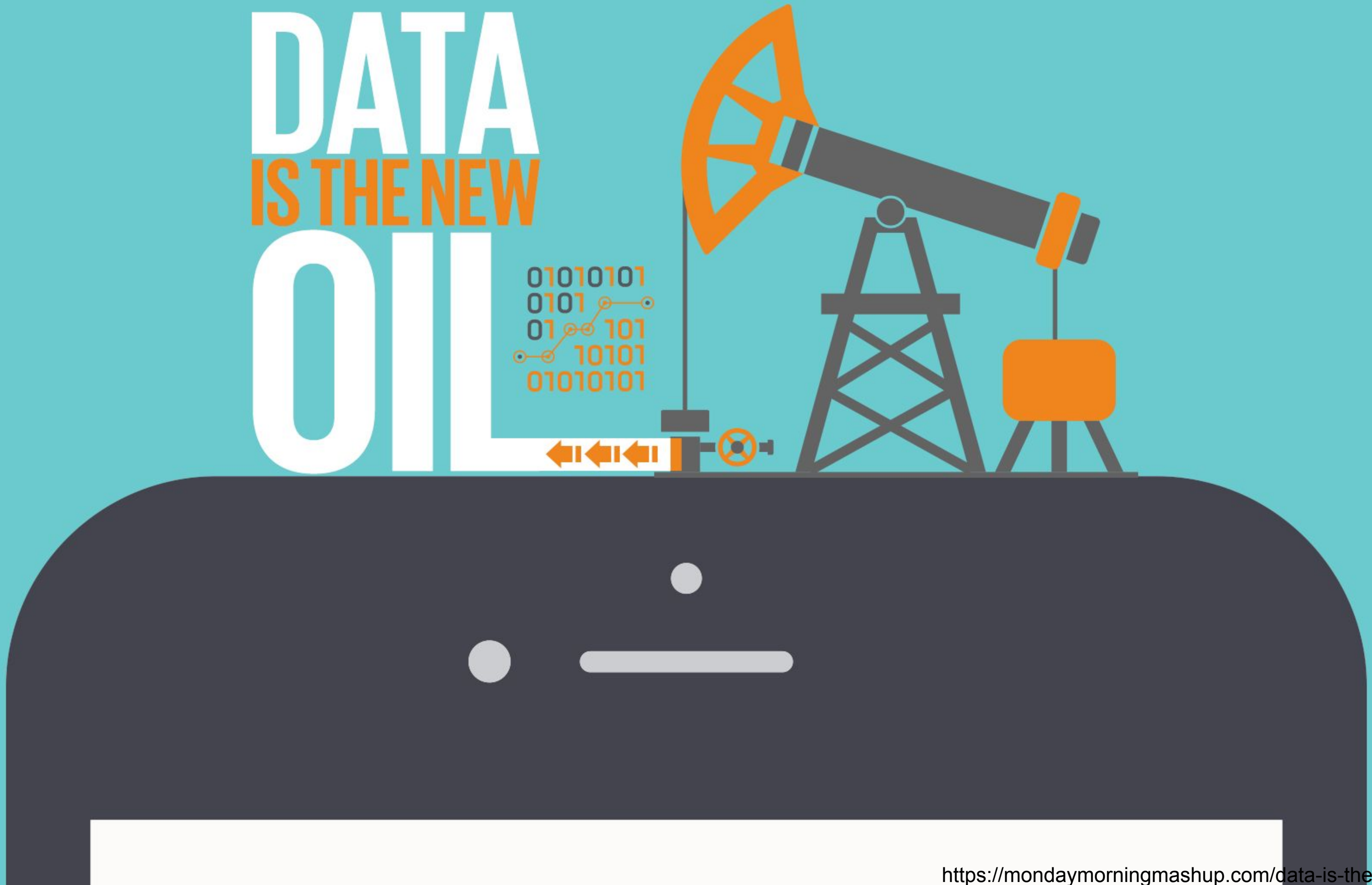
What is Machine Learning?

- Artificial Intelligence is a very broad term, including many aspects such as philosophy and biology
- Machine Learning is a study of computer algorithms that improve automatically through experience
- Experience = Data (in all possible forms)

Examples of Machine Learning tasks

- Image classification
- Face detection
- Churn prediction
- Personalized recommendations
- Speech recognition and generation
- Machine translation
- Question answering
- …
- Your example?

Why so much attention?



https://mondaymorningmashup.com/data-is-the-new-oil/

- Machine Learning changes paradigm of how we write software and make decisions
- “Data-driven” approach

The adjective **data-driven** means that progress in an activity is compelled by data, rather than by intuition or by personal experience

When a company employs a “data-driven” approach, it means it makes strategic decisions based on data analysis and interpretation.

https://en.wikipedia.org/wiki/Data-driven,
https://www.atinternet.com/en/glossary/data-driven/

Software 1.0

1. Decompose task into understandable pieces
2. Create logic for each task (steps)
3. Connect everything and make it work on a new data

Software 1.0

But what to do with this?

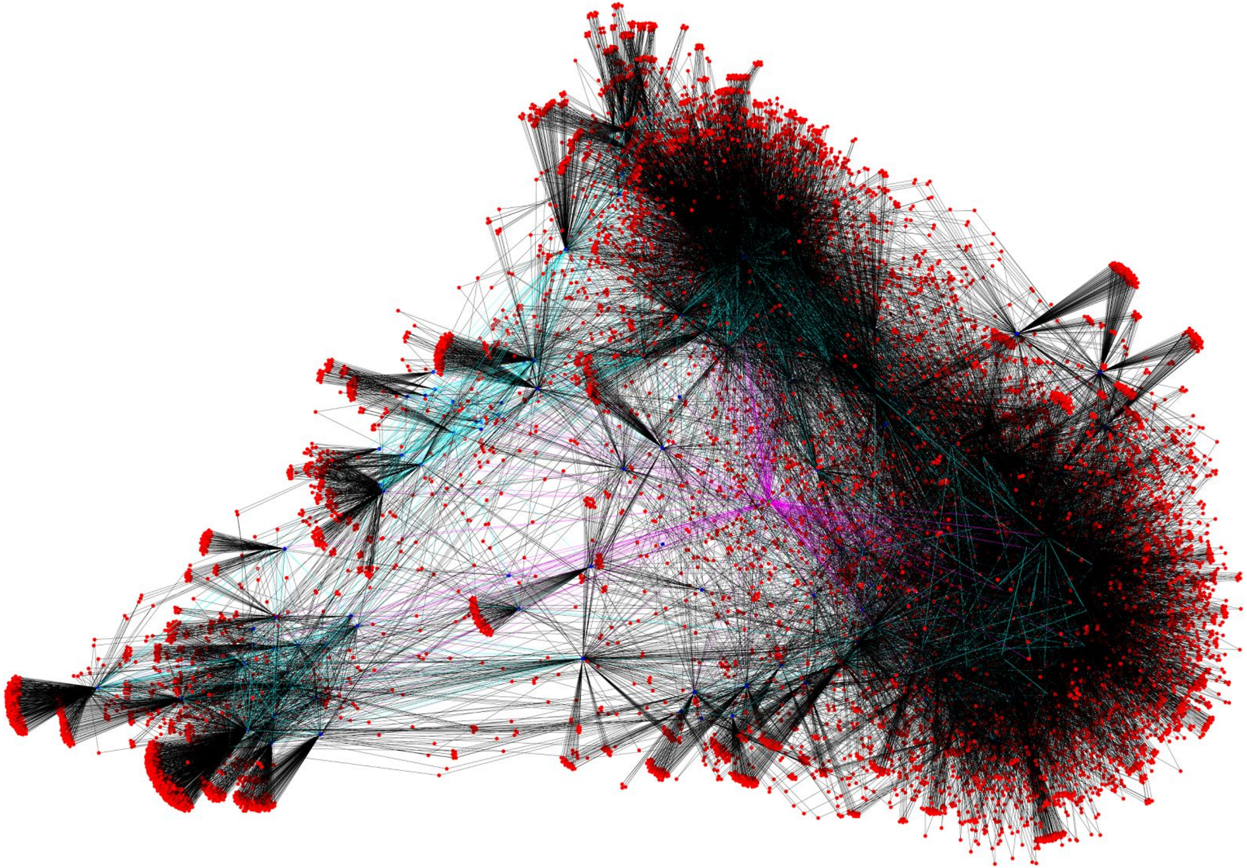
Software 1.0

Or this?

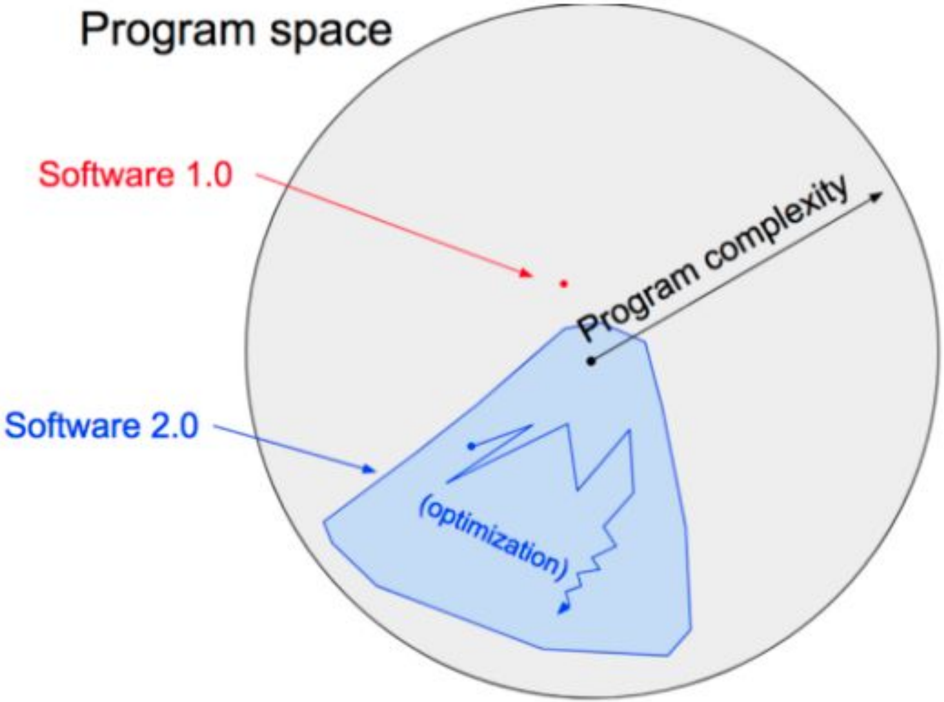
连续几年的*人类发展报告*表明，生活在全球多数国家的大部分人口其人类发展水平正在稳步提高。科技、教育的进步和收入的增加为人们过上更长寿、更健康、更安心的生活提供了可靠的保障。[1]总的来说，全球化对人类发展产生了重要的积极作用，尤其是在许多南方国家。但在当今世界，不安全感仍普遍存在，无论是在生计、人身安全、环境还是全球政治方面。[2]在人类发展的一些

Software 1.0

Or this?



Software 2.0

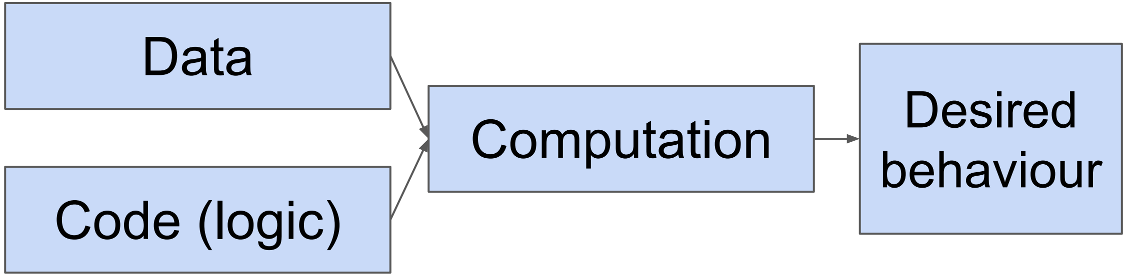


https://medium.com/@karpathy/software-2-0-a64152b37c35

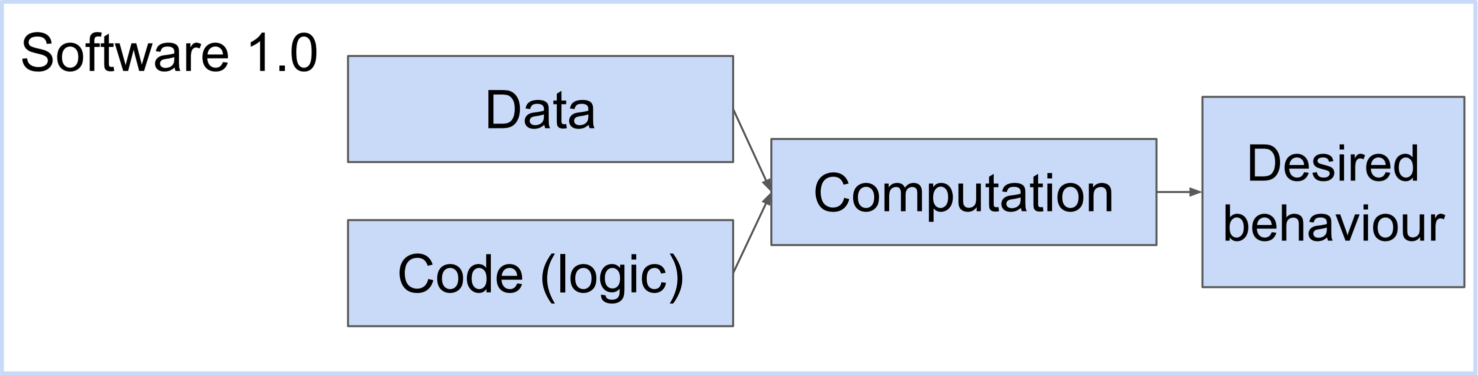
Software 2.0

- Find data that describes desired behaviour (Example: Image->Class)
- Specify architecture that is capable to learn this behaviour
- Optimize this architecture
- Use the trained architecture in Software 1.0

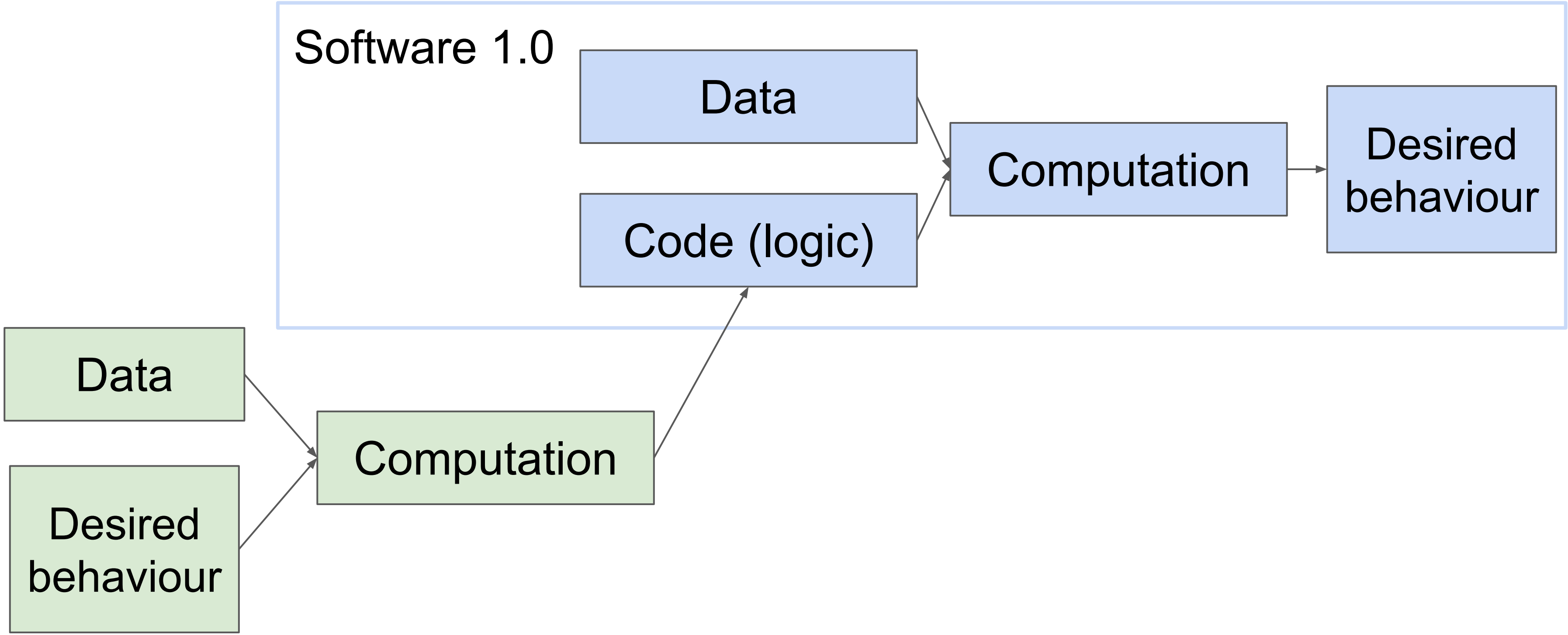
Software 1.0 vs Software 2.0



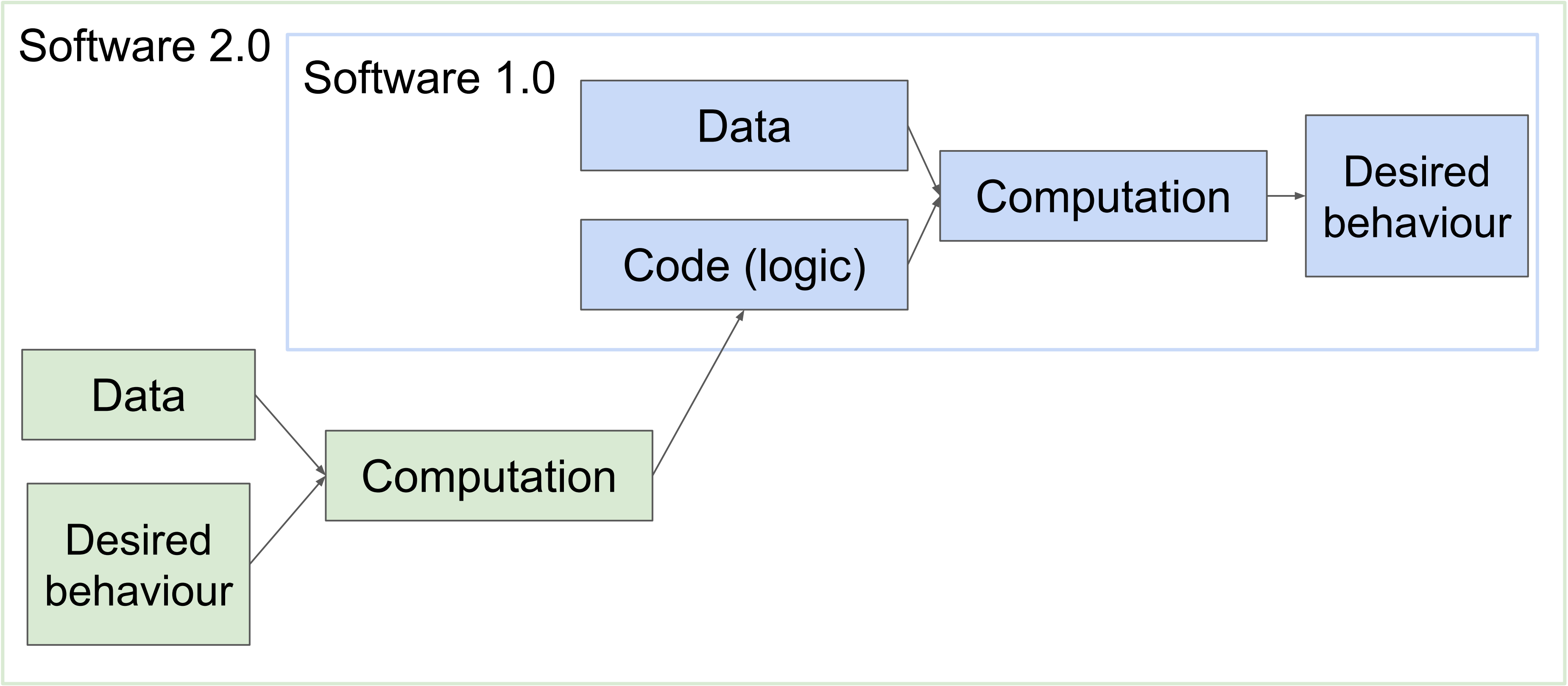
Software 1.0 vs Software 2.0



Software 1.0 vs Software 2.0

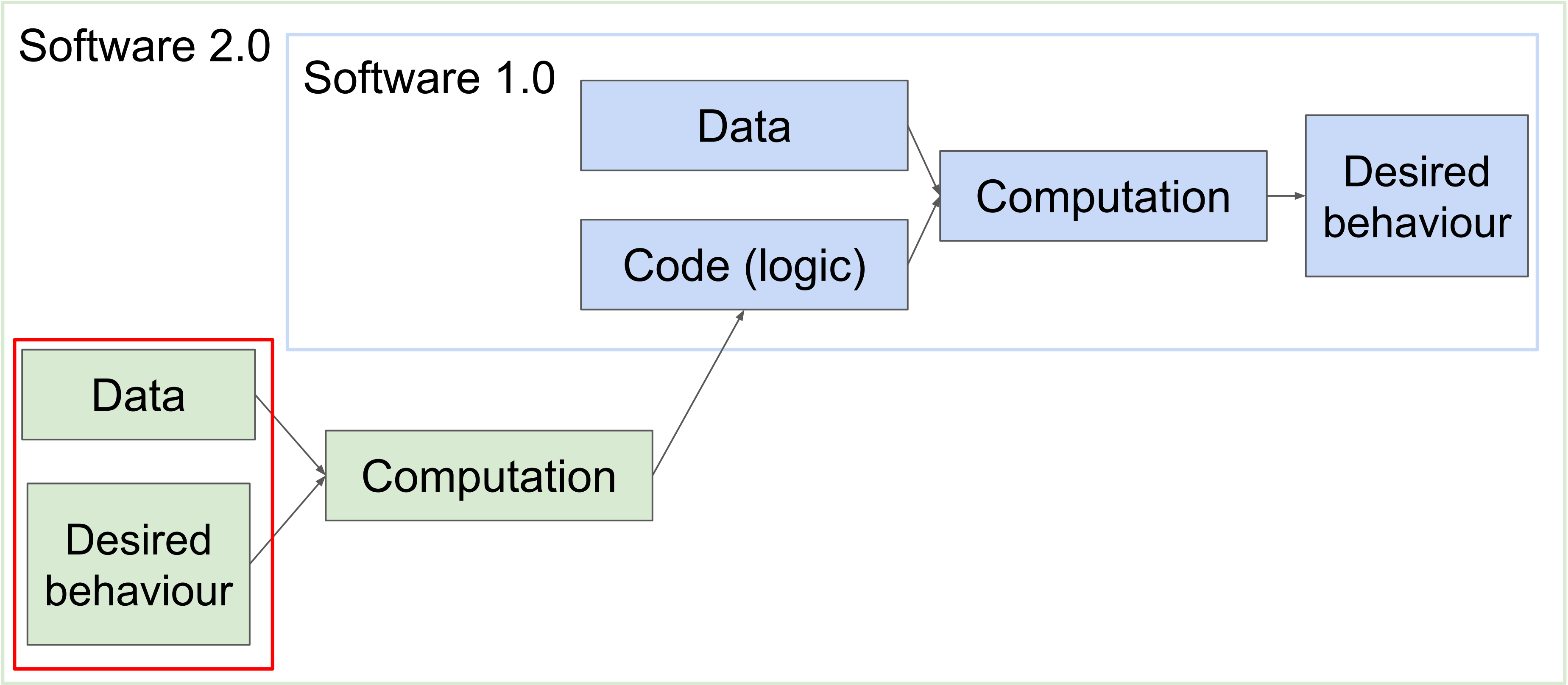


Software 1.0 vs Software 2.0



Autopilot Example

Software 1.0 vs Software 2.0



Autopilot Example

- Collect data: everything that is needed to drive a car

Autopilot Example

- Collect data: everything that is needed to drive a car

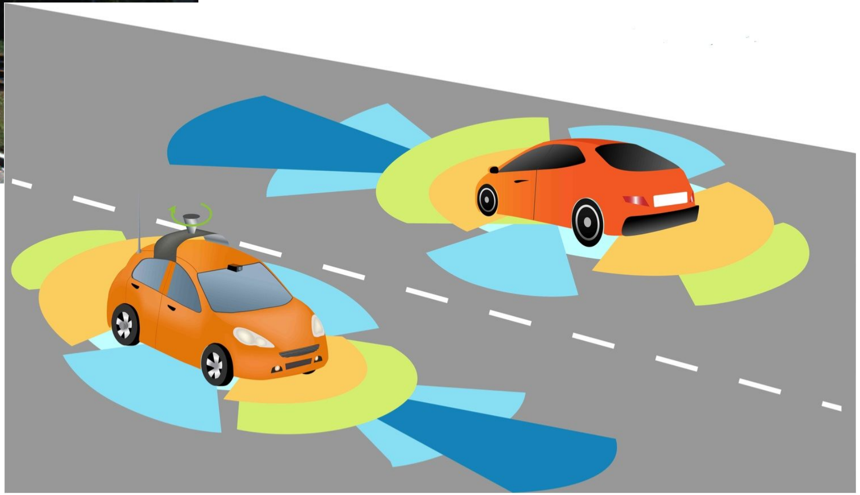
Images

Autopilot Example

- Collect data: everything that is needed to drive a car

Images

Sensor data



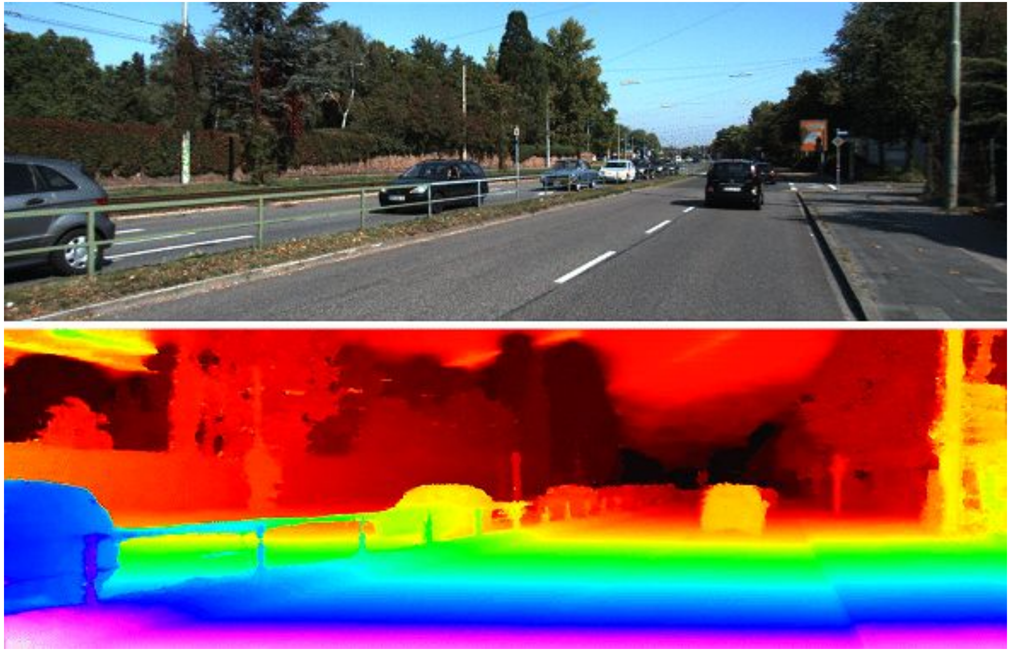
Autopilot Example

- Desired behaviour is driving, but for that we need to know a lot of additional information

Autopilot Example

- Desired behaviour is driving, but for that we need to know a lot of additional information

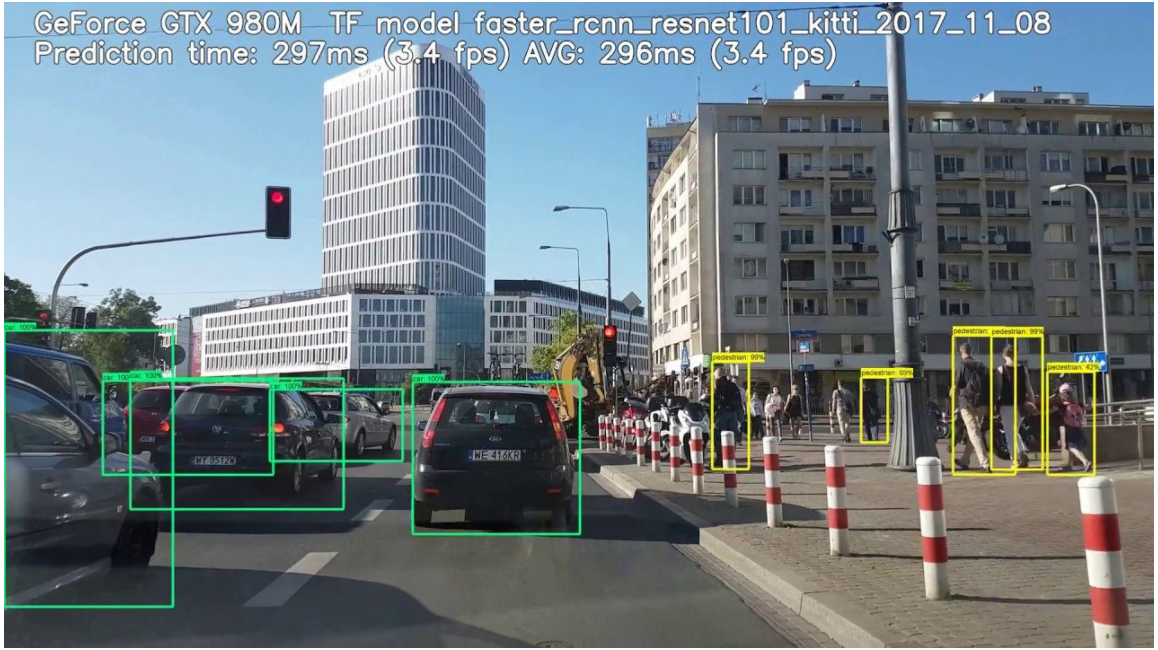
Depth



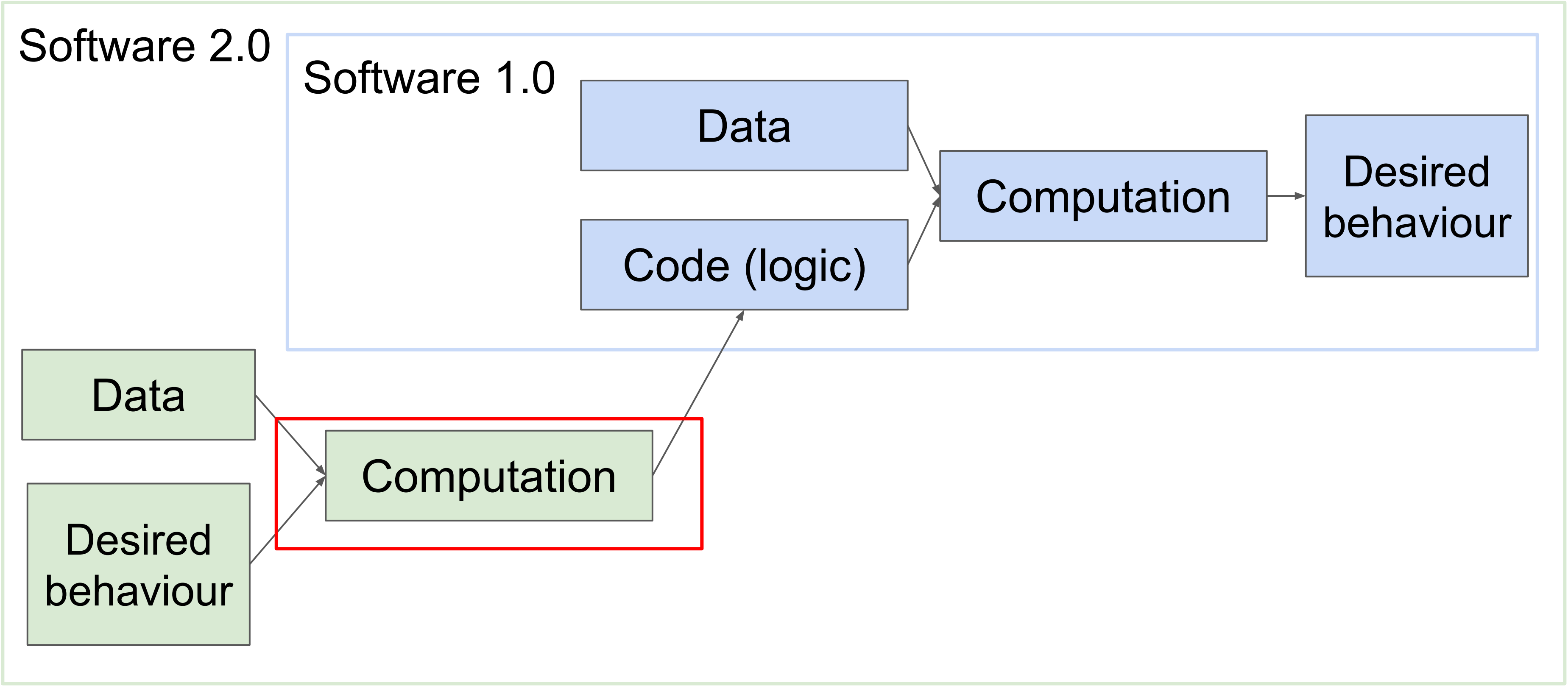
Autopilot Example

- Desired behaviour is driving, but for that we need to know a lot of additional information

Object detection



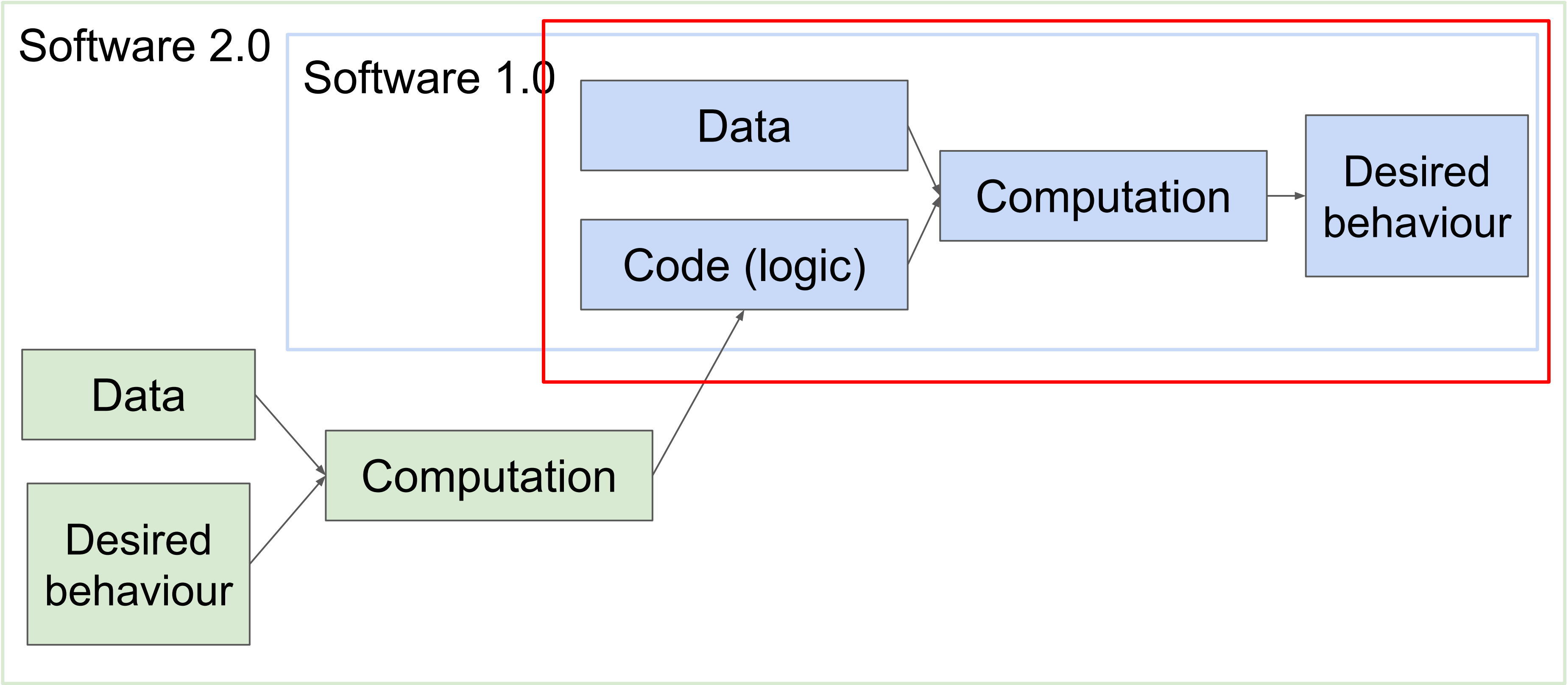
Software 1.0 vs Software 2.0



Autopilot Example

- Optimize model to provide desired behaviour
  Depths, classes, object detection and so on
- It is possible to learn how to drive directly, but it is too complex for modern systems, that’s why we decompose this task

Software 1.0 vs Software 2.0



Autopilot Example

- Software 1.0 part, putting all the logic together, create handcrafted rules to make it work

General Ideas

- Data is in the core of Machine Learning, always make it as good and large as you can
- Task decomposition is beneficial. If your machine cannot learn the end goal, try to decompose task into simple parts
- Machine Learning is not a magic, you should understand your task and introduce prior knowledge into architectures and data

Terminology

Dataset - data and labels

Dataset - data and labels

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

Observation - one element from dataset

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

In many cases the observations are supposed to be ***i.i.d.***

- ***independent***
- ***identically distributed***

Feature - a property of an observation

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

ML thesaurus

These all are features

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

ML thesaurus

These all are features

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

These all are features

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

ML thesaurus

Target represents the information we are interested in.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

Target can be either a ***number*** *(real, integer, etc.) – for* ***regression*** problem

Target represents the information we are interested in.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Target (passed)
John	22	5	4	Brown	English	5	TRUE
Aahna	17	4	5	Brown	Hindi	4	TRUE
Emily	25	5	5	Blue	Chinese	5	TRUE
Michael	27	3	4	Green	French	5	TRUE
Some student	23	3	3	NA	Esperanto	2	FALSE

Or a ***label*** – for ***classification*** problem

Assume that we have some model

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)
John	22	5	4	Brown	English	5
Aahna	17	4	5	Brown	Hindi	4
Emily	25	5	5	Blue	Chinese	5
Michael	27	3	4	Green	French	5
Some student	23	3	3	NA	Esperanto	2

ML thesaurus

The ***prediction*** contains values we predicted using some ***model***.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Predicted (mark)
John	22	5	4	Brown	English	5	4.5
Aahna	17	4	5	Brown	Hindi	4	4.5
Emily	25	5	5	Blue	Chinese	5	5
Michael	27	3	4	Green	French	5	3.5
Some student	23	3	3	NA	Esperanto	2	3

One could notice that prediction just averages of Statistics and Python marks. So our ***model*** can be represented as follows:

$$\widehat{\mathrm{mark}}_{ML} = \frac{1}{2}\mathrm{mark}_{Statistics} + \frac{1}{2}\mathrm{mark}_{Python}$$

The ***prediction*** contains values we predicted using some ***model***.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Predicted (mark)
John	22	5	4	Brown	English	5	4.5
Aahna	17	4	5	Brown	Hindi	4	4.5
Emily	25	5	5	Blue	Chinese	5	5
Michael	27	3	4	Green	French	5	3.5
Some student	23	3	3	NA	Esperanto	2	3

Different models can provide different predictions:

$$\hat{\mathrm{mark}}_{ML} = \frac{1}{2}\mathrm{mark}_{Statistics} + \frac{1}{2}\mathrm{mark}_{Python}$$

The ***prediction*** contains values we predicted using some ***model***.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Predicted (mark)
John	22	5	4	Brown	English	5	1
Aahna	17	4	5	Brown	Hindi	4	5
Emily	25	5	5	Blue	Chinese	5	2
Michael	27	3	4	Green	French	5	4
Some student	23	3	3	NA	Esperanto	2	3

Different models can provide different predictions:

$$\hat{\text{mark}}_{ML} = \text{random}(\text{integer from } [1; 5])$$

The ***prediction*** contains values we predicted using some ***model***.

Name	Age	Statistics (mark)	Python (mark)	Eye color	Native language	Target (mark)	Predicted (mark)
John	22	5	4	Brown	English	5	1
Aahna	17	4	5	Brown	Hindi	4	5
Emily	25	5	5	Blue	Chinese	5	2
Michael	27	3	4	Green	French	5	4
Some student	23	3	3	NA	Esperanto	2	3

Different models can provide different predictions.

*Usually some **hypothesis** lies beneath the model choice.*

Loss function measures the error rate of our model.

Square deviation	Target (mark)	Predicted (mark)
16	5	1
1	4	5
9	5	2
1	5	4
1	2	3

- ***Mean Squared Error*** (where **y** is vector of targets):

$$MSE(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N}\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 = \frac{1}{N}\sum_i (y_i - \hat{y}_i)^2$$

Maximum Likelihood Estimation

Likelihood

Denote dataset generated by distribution with parameter θ

Likelihood function:

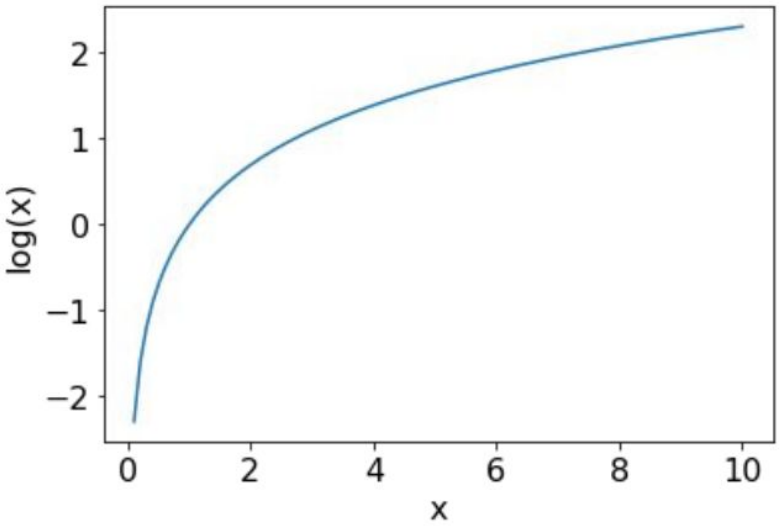
$$L(\theta|X,Y) = P(X,Y|\theta)$$

$$L(\theta|X,Y) \longrightarrow \max_{\theta}$$

samples should be i.i.d.

$$L(\theta|X,Y) = P(X,Y|\theta) = \prod_i P(x_i, y_i|\theta)$$

Maximum Likelihood Estimation



Likelihood

Denote dataset generated by distribution with parameter θ

Likelihood function:

$$L(\theta|X,Y) = P(X,Y|\theta)$$

$$L(\theta|X,Y) \longrightarrow \max_{\theta}$$

samples should be i.i.d.

$$L(\theta|X,Y) = P(X,Y|\theta) = \prod_i P(x_i, y_i|\theta)$$

equivalent to

$$\log L(\theta|X,Y) = \sum_i \log P(x_i, y_i|\theta) \longrightarrow \max_{\theta}$$

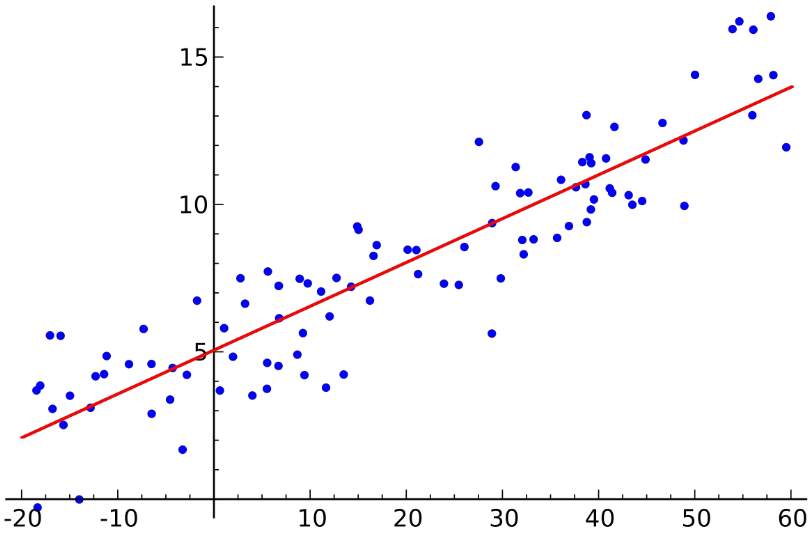
Machine Learning problems overview

Supervised learning problem statement

Let's denote:

- Training set $\mathcal{L} = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$, where
  - $(\mathbf{x} \in \mathbb{R}^p,\ y \in \mathbb{R}\ )$ for regression
  - $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \{+1, -1\}$ for binary classification
- Model $f(\mathbf{x})$ predicts some value for every object
- Loss function $Q(\mathbf{x}, y, f)$ that should be minimized

- Regression problem



Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$

- Regression problem
- Classification problem