Introduction to Data Science

Prof. Dr. Thomas Kopinski

Overview

- What is this course about?
- Course Goals
- How do we proceed?
- Important dates
- About myself
- Introduction to Data Science
- Introduction to Python

Data Science: Project (graded)

- pick and choose a problem
- find data for it
  - either from your company
  - or: work at a given problem setting
- project deadline: 31.03.2021
- outline deadline: 31.01.2021
- outline (2 pages):
  - describe the task/problem
  - describe the data
  - present a possible solution for the problem
- present the data science problem and solution
  - short summary: 1 cover page, 5 written pages content (2 extra pages for images)
  - a python notebook with the code and concept
  - present the result in a discussion with questions (30 mins)

Data Science: Exam (graded)

- Date: 03.03.2020
- 90 mins / 90 points, 45 to pass
- Decision (exam or project) until end of November 2020

Course goals

- fluent + efficient programming skills in Python/scikit/numpy/pandas
- understand what data science is about
- deeper understanding of data analytics
- be able to visualize and manipulate data
- understand the most important concepts and algorithms in data science
- be able to define and successfully implement a data science project

Contact

Prof. Dr. Thomas Kopinski

Office hours: appointment via mail

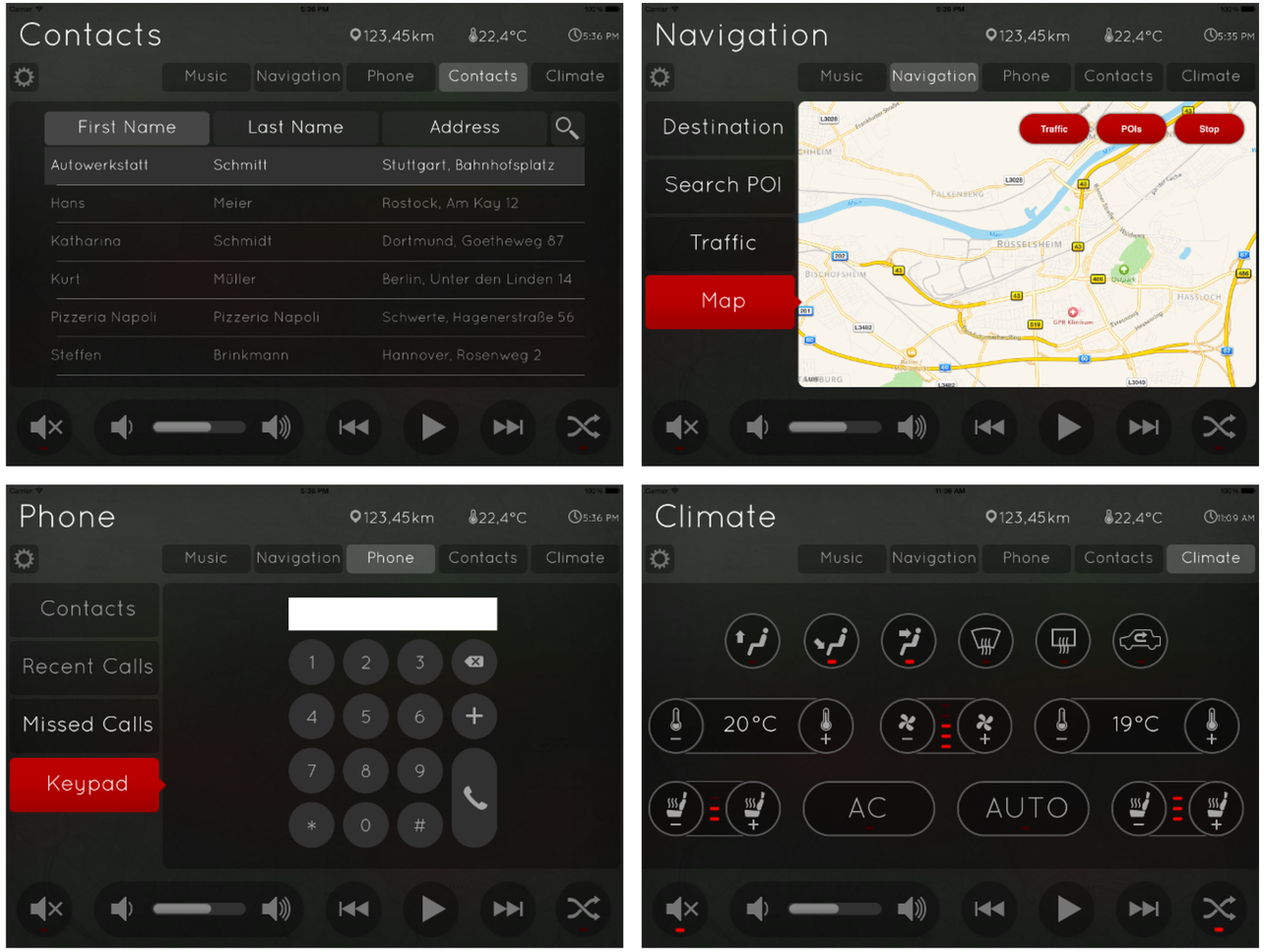
mailto: kopinski.thomas@fh-swf.de

http://www4.fh-swf.de/de/home/ueber_uns/standorte/me/doz_iw/profs_iw/kopinski/index.php

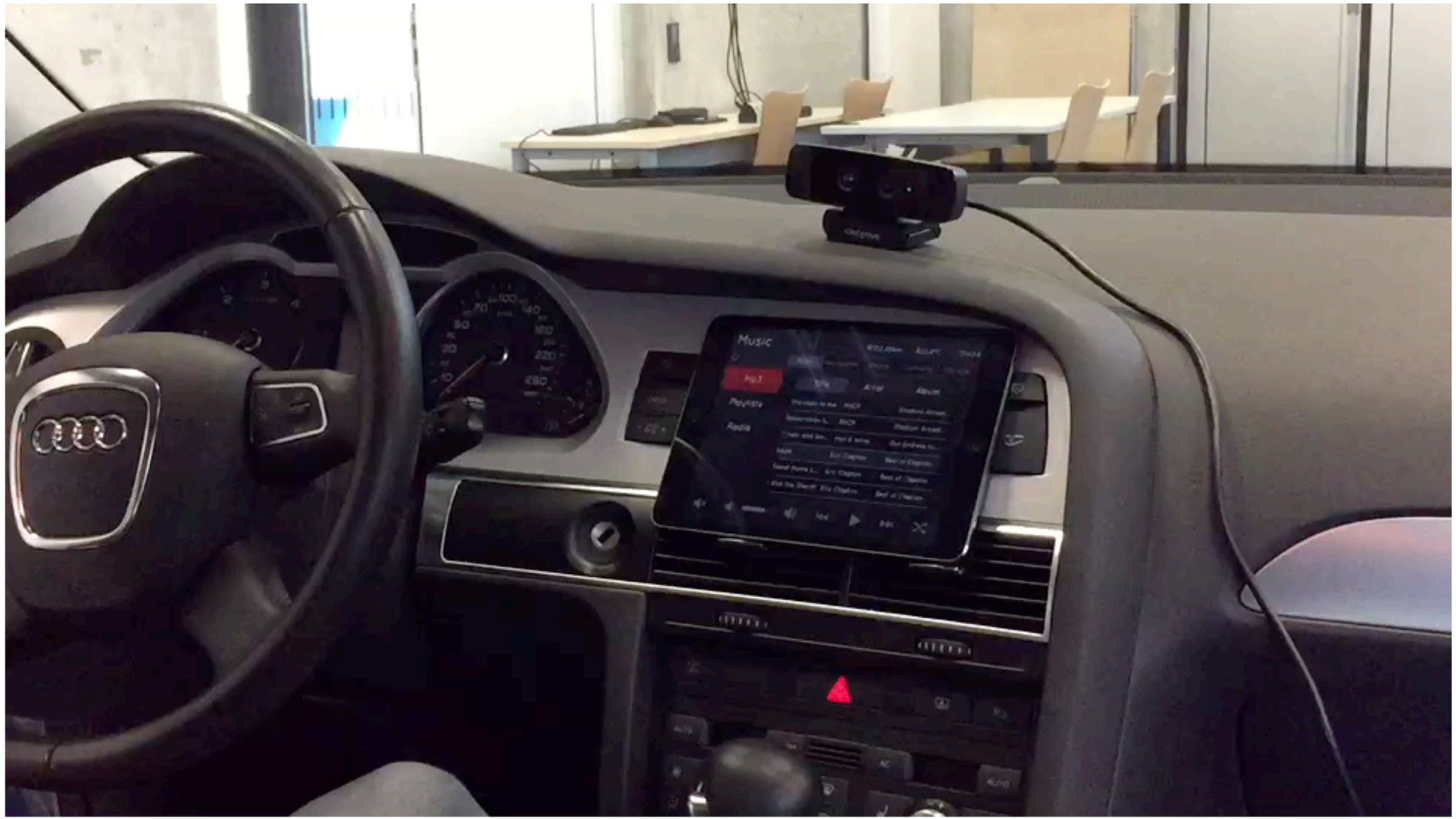
Curriculum Vitae

• Study Computer Science at TU Dortmund	|2006
• Software Developer in Hamburg	|2006-08
• Self-dependent within the mobile industry	|08 - 13
• Université Paris-Saclay: PhD Machine Learning	|13 - 16
• Université Paris-Saclay: Postdoc	|16 - 17
• March 2017: Professor for Computational Engineering and Data Science	|2017

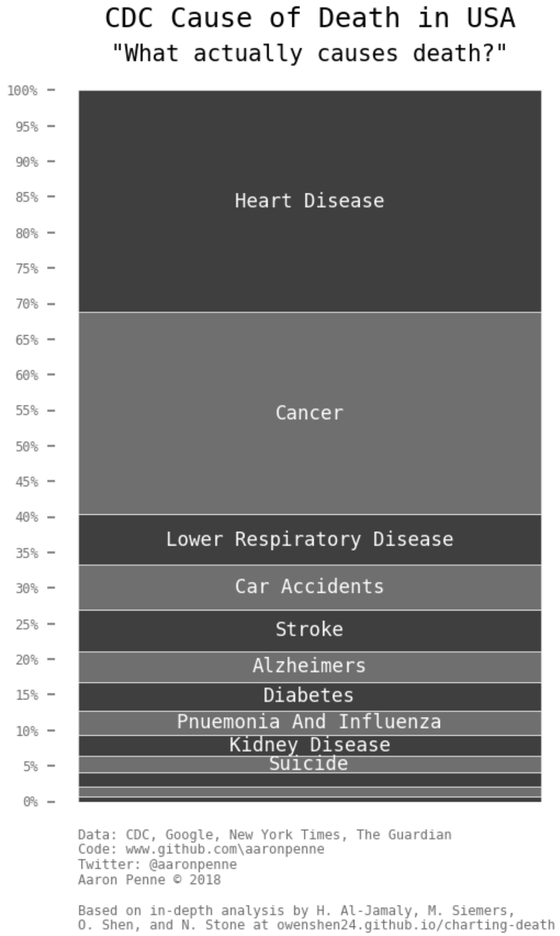
About myself: own work

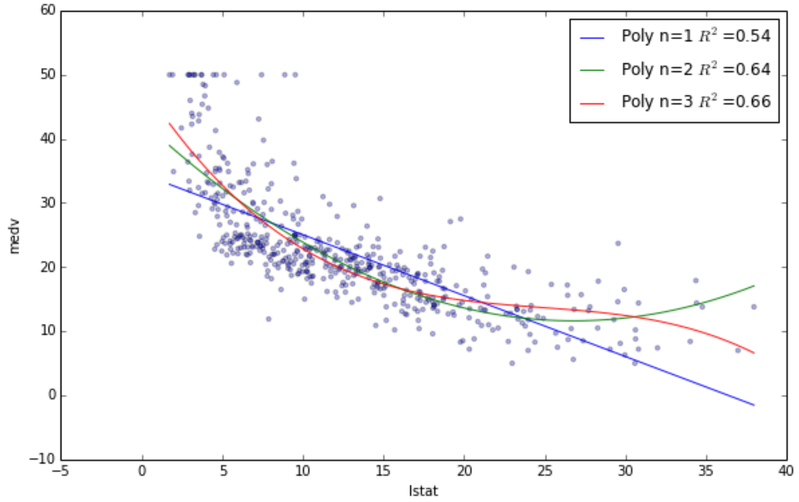


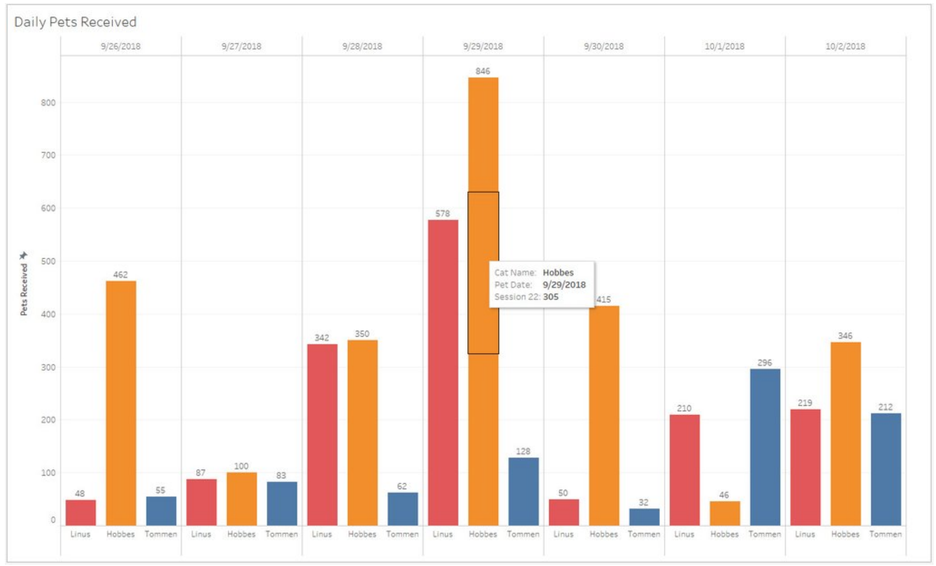
About myself: own work



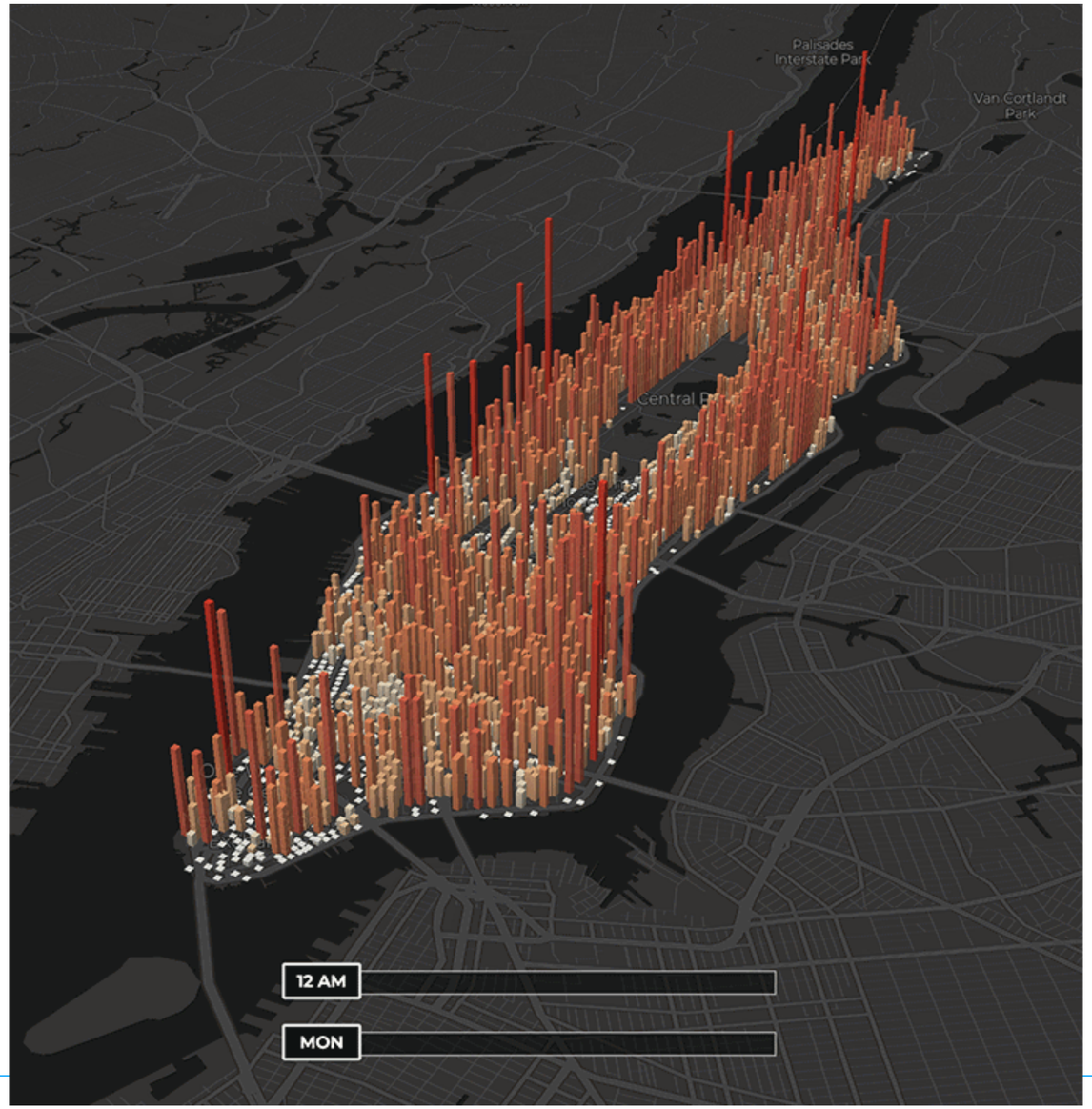
Data Science - examples



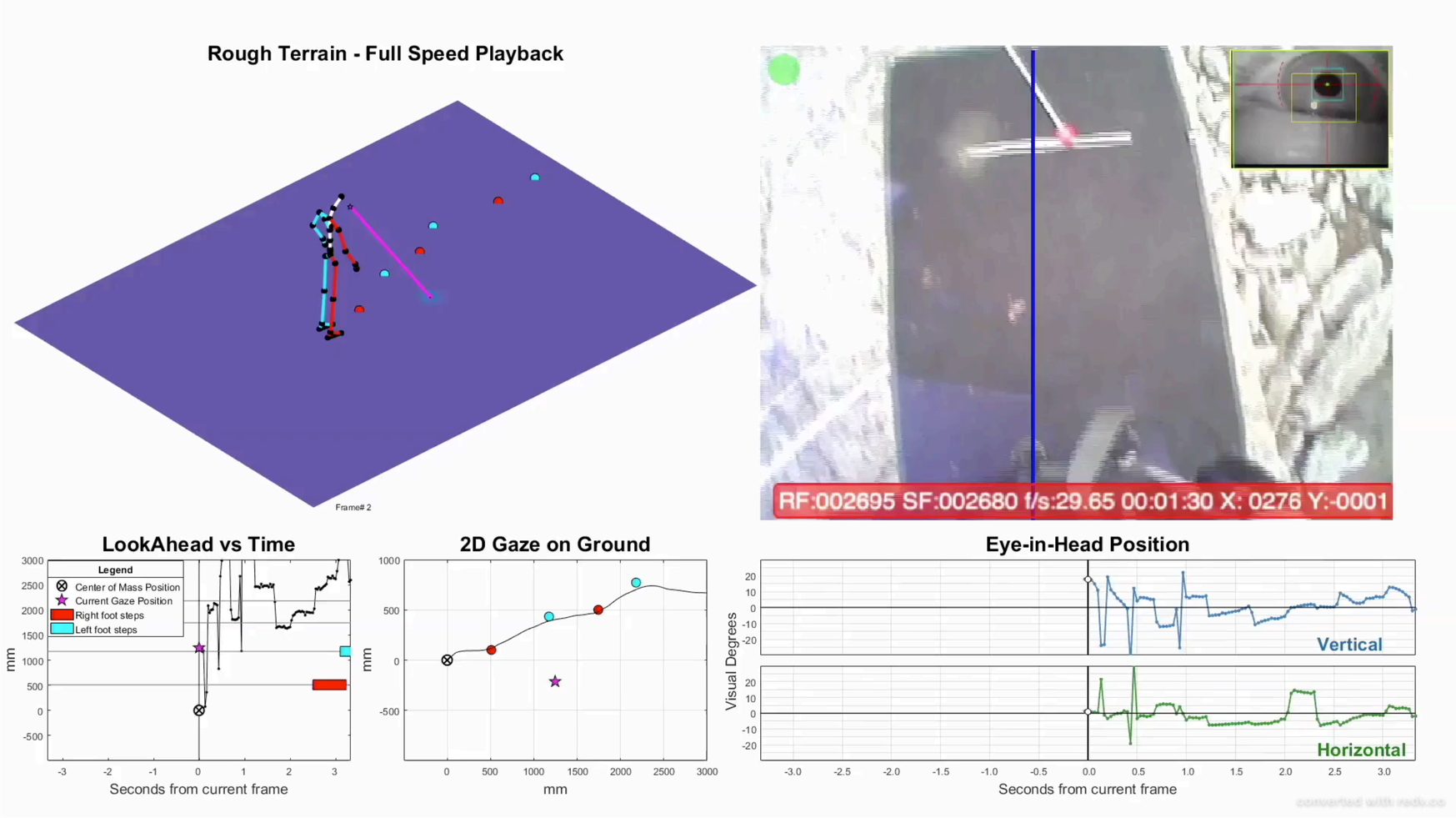




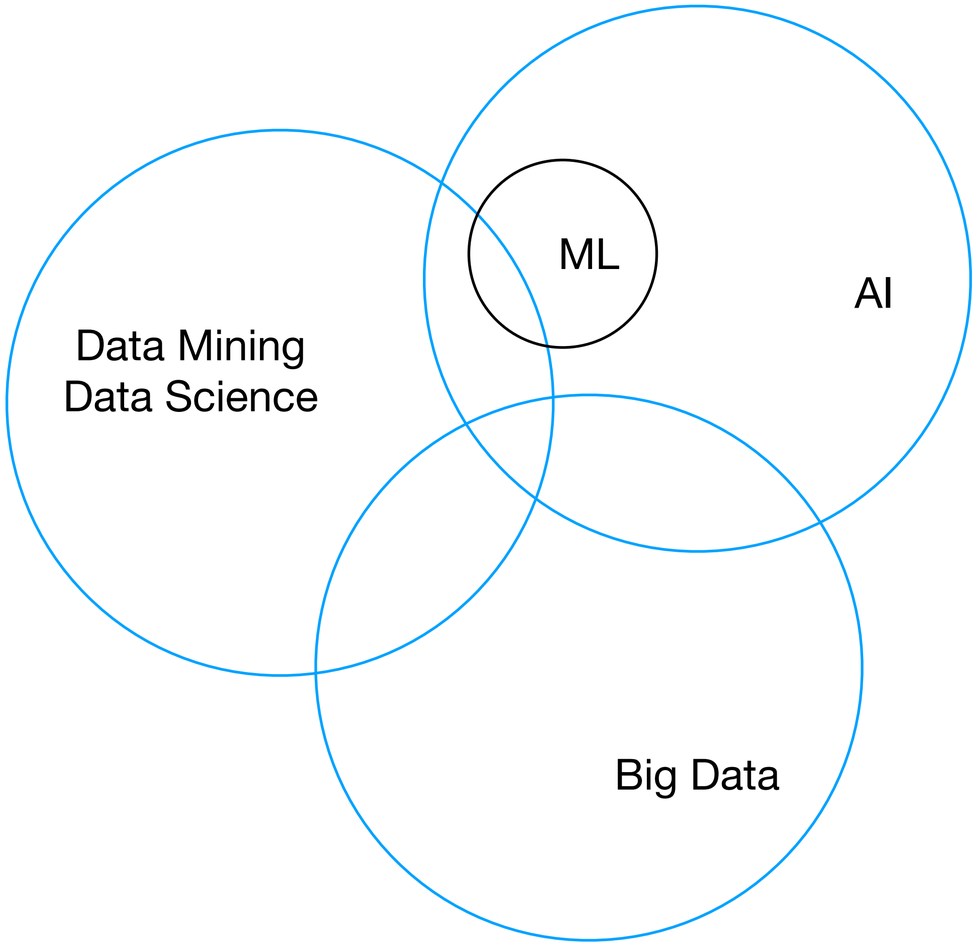
Data Science - examples



Data Science - examples



Data Science



Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician. - Josh Wills

Data Science: Skills

1. Math Skills — Statistics — Machine Learning
2. Coding —  Python
3. Databases —    

Data Science: Skills

5 Level up With Big Data

Velocity | Variety | Veracity | Volume

Why is Big Data processing different from other data processing

Hadoop

Grasp distributed approach to data processing and storage

Spark

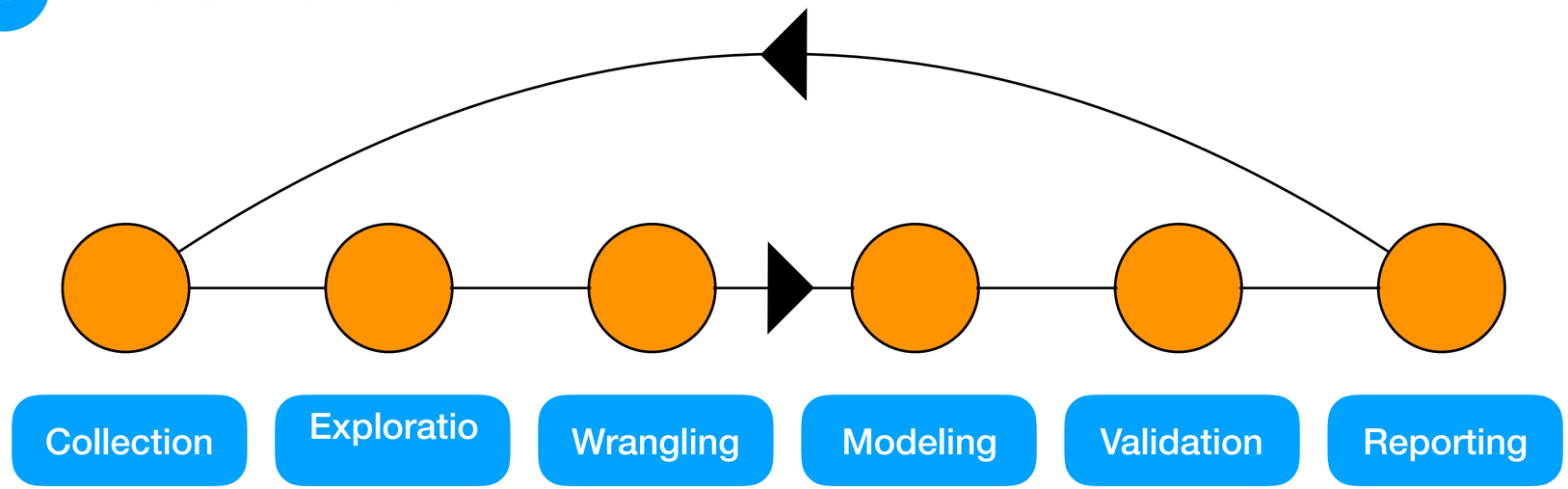
Understand the advantage of in-memory cluster computing network

Data Science: Skills

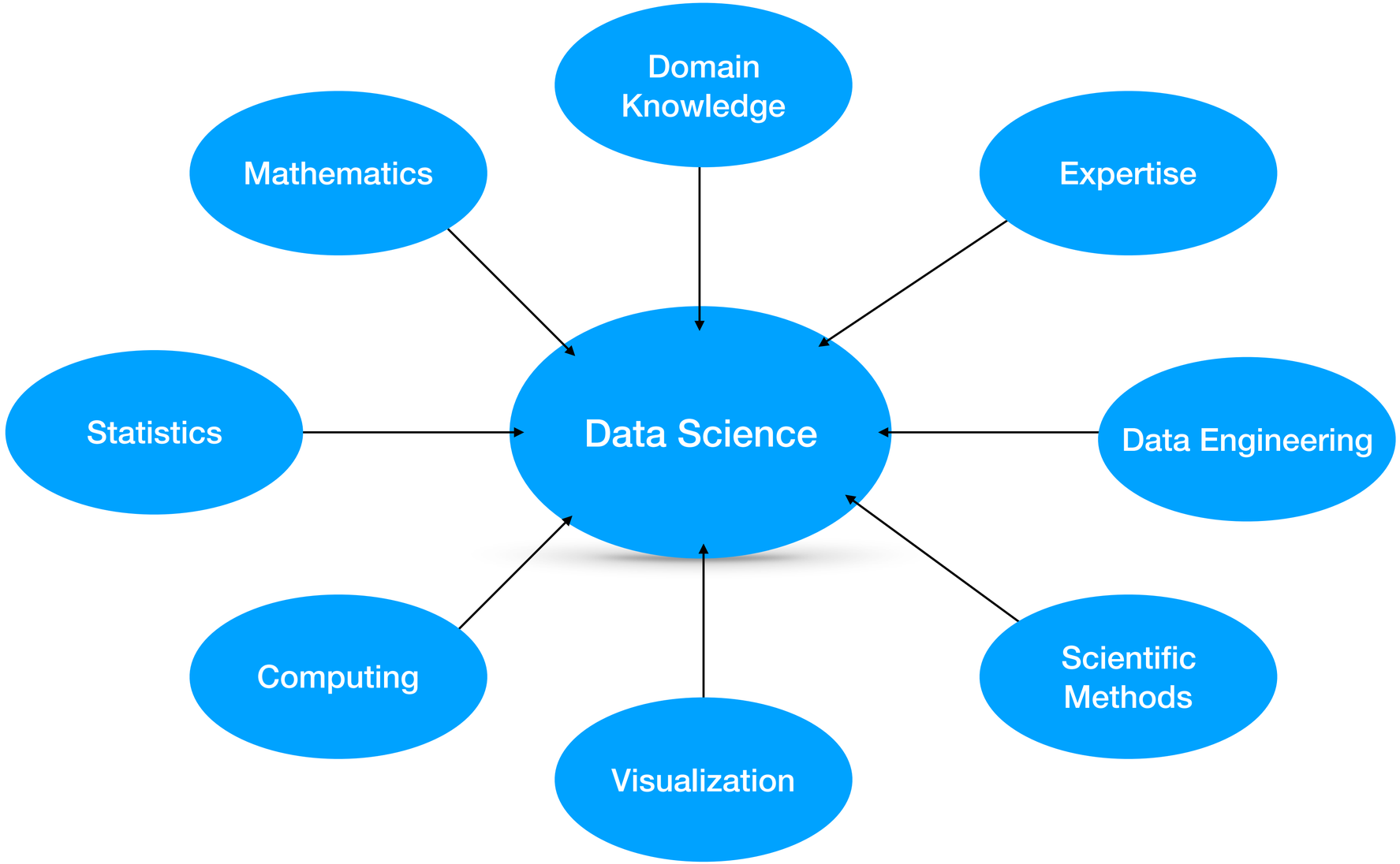
6	Grow, Connect, Learn	Get a first project, exchange with other Data Scientist, develop intuition
7	Dive in	improve skills through online projects, Bootcamps, Competitions etc.
8	Engage with community	Exchange, share your experience, follow, contribute

Data Science: Skills

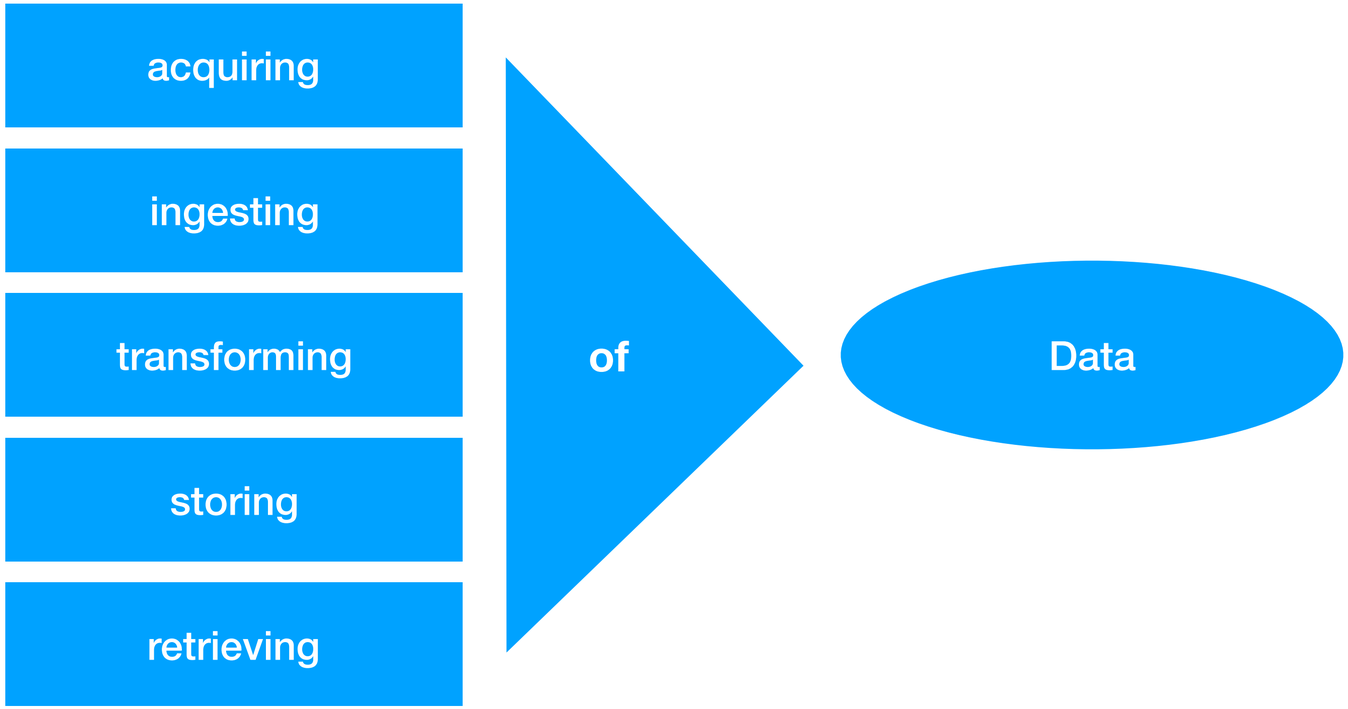
4 Data Science workflow



What is Data Science?

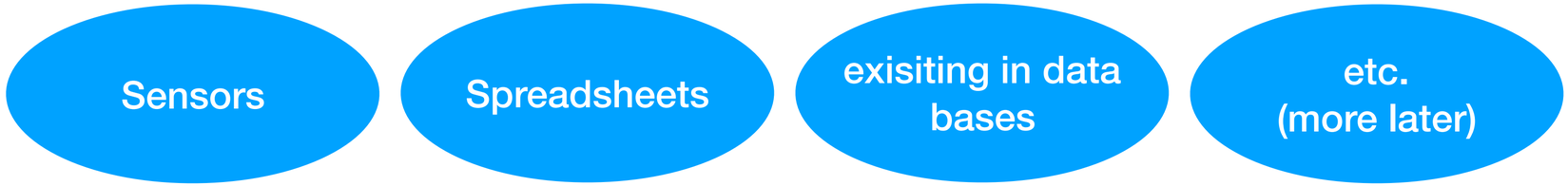


Data Engineering



Data Acquisition

- where is the data coming from?
- how does the data look like?
- how do I get access to the data?



Data Ingestion and Transformation

- getting the data into the system
  - how much data is coming?
  - what is the velocity of the data flow?
  - is there enough space?
  - does the incoming data need to be filtered in any way?
- Data often comes as CSV-files

```
20180101,070014,40,112,2,100.45
20180101,070015,40, 111,1,112.45
```

What could this be?

Year	Month	Day	Hour	Minute	Second	MWay	Cam	Lane	Speed
2018	1	1	7	0	14	40	112	2	100.45
2018	1	1	7	0	15	40	111	1	112.45

- Knowing the Metadata (Data about data) is important!

Data Storage

- data can be stored in various forms
- save all data coming from the camera in a single file per day
- leads to large files
- fast to read, little functionality
- how to compute the number of cars going above 100 km/h?
- store data in a data base
- longer reading times - more functionality
- we extract the relevant information directly to solve a task/question!

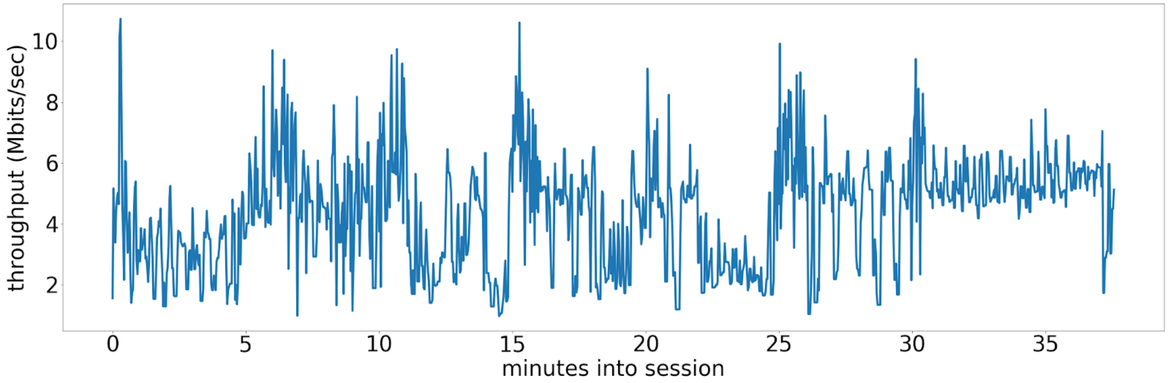
Netflix: Improve Streaming Quality

- Goal: Give the user a best streaming experience
- 117M members worldwide; > 50% outside of U.S.
- growing audience with varying viewing behavior and viewing capabilities
  —> ‚one-size-fits-all‘ solution does not work
- mobile devices work different than smart TVs
- cellular networks are volatile / unstable
- networks in some markets can be congested
- different device groups have different capabilities

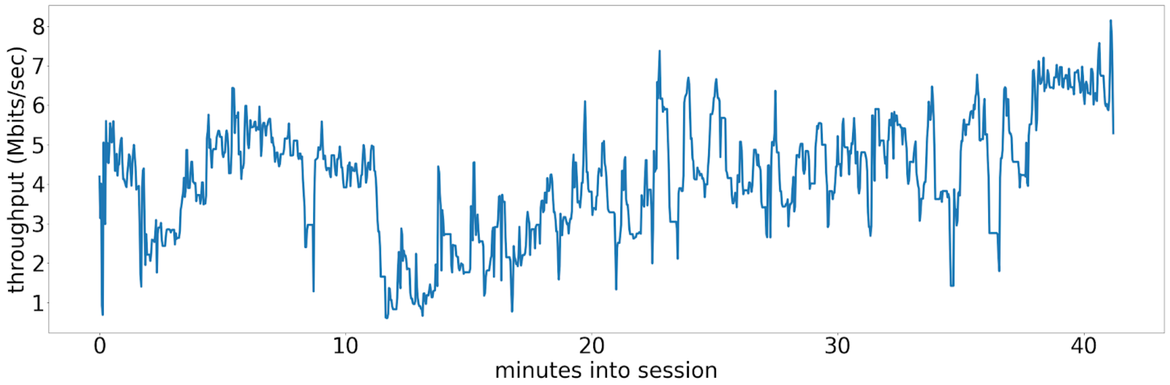
Predicting network quality

- average bandwidth / round trip time are well-known indicators
- unknown factors: stability + predictability

network throughput:



noisy
+
fluctuating



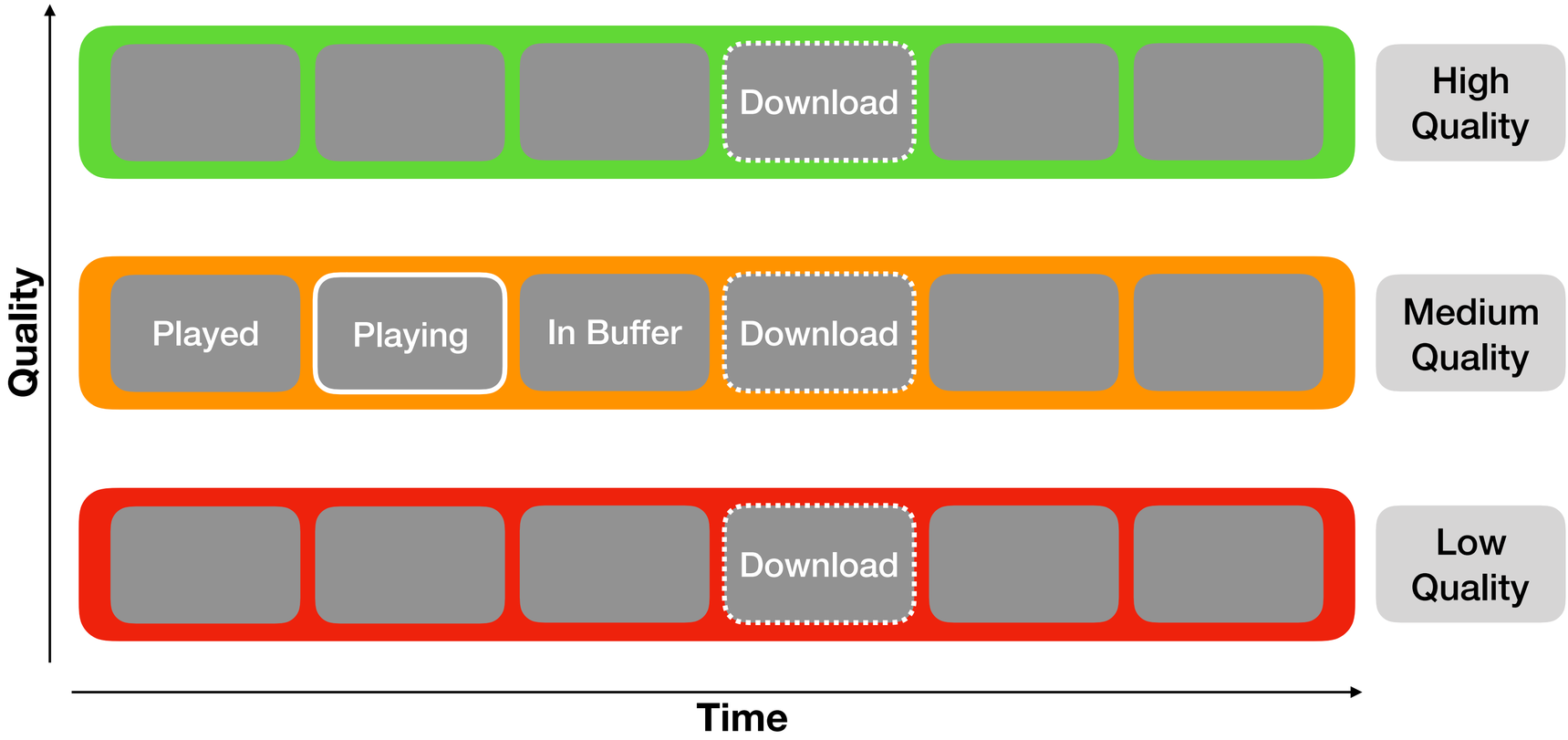
are predictions possible given the last minutes ?

[1] https://medium.com/netflix-techblog/using-machine-learning-to-improve-streaming-quality-at-netflix-9651263ef09f

Predicting network quality

- are predictions possible given the last minutes ?
- can we extract information about the device given some historical data?
- what kind of data can be provided for the server to adapt optimally?
- other alternative: predict the *distribution* of throughput
- maybe combine temporal data/information with other contextual information?

Netflix: Improve Streaming Quality



- aggressive download - high chance of rebuffer
- download data / video upfront at wait time cost
- feedback signals come delayed and sparsely
- ‚credit-assignment problem‘

Predictive Caching

- develop statistical model: predict what the user will play in order to cache parts on the device before it is played
- simple example: predict the user will watch the next episode of a series
- combine:

viewing history

user interaction

context information

Device anomaly detection

- thousands of different device types: Smart TVs, tablets, smarphones, laptops, streaming sticks etc.
- firmware updates sometimes cause problems in user experience
- detecting a problem in a device is challenging —> which criteria are suitable?

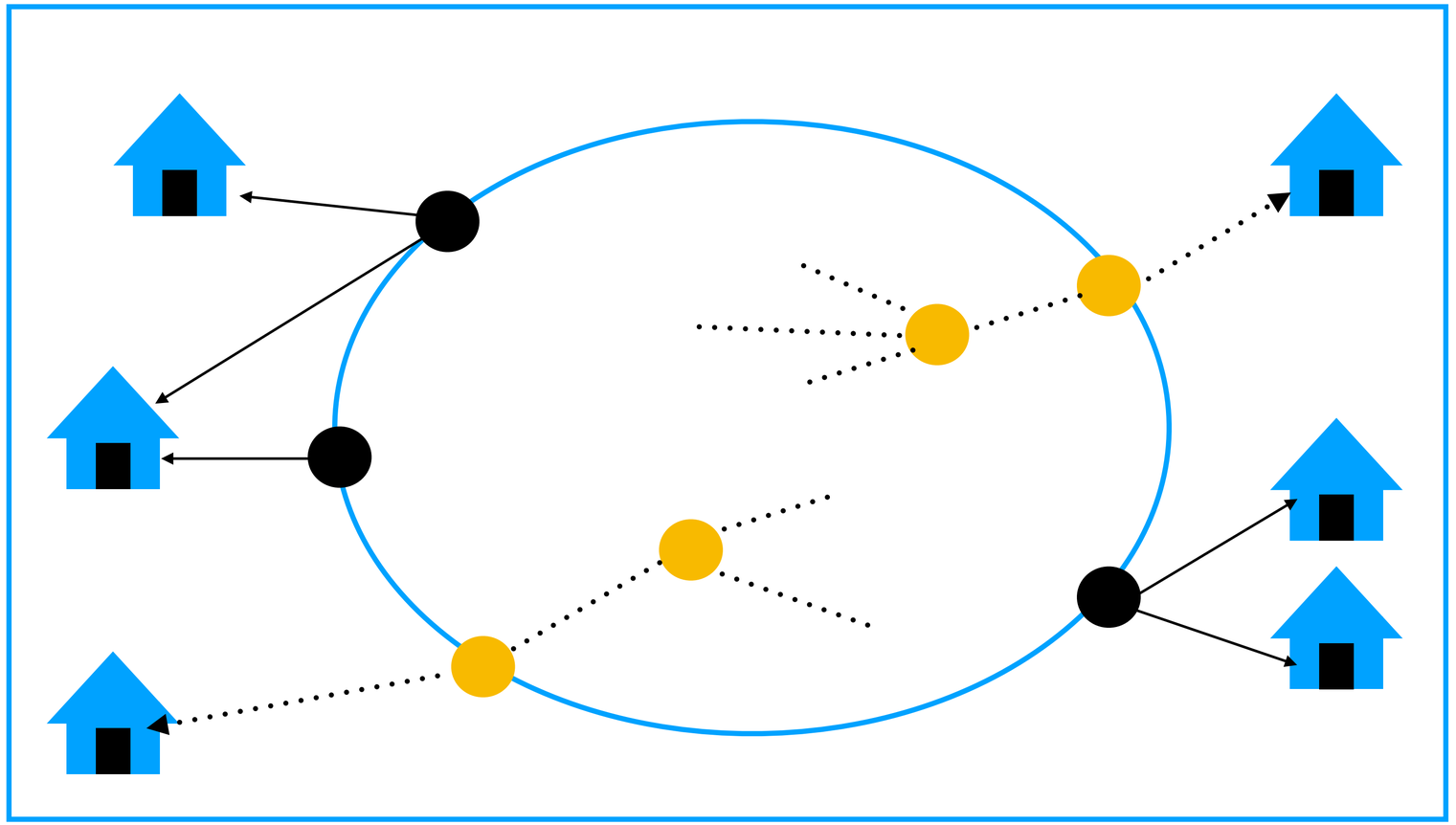
	Device not OK	**Device OK**
Test: Problem	true negative	false negative
Test: No Problem	false positive	true positive

- a liberal trigger may cause many false positives
- a strict trigger may miss many problems
- statistical models can help finding the root of the problem

The basis

- sufficient amount of data (117M users)
- use Machine Learning algorithms to help improve user experience
- high-dimensionality of data —> need to select variables for problems
- **rich structure** of data: collective network usage, human preferences, device hardware capabilites

Edge Caching

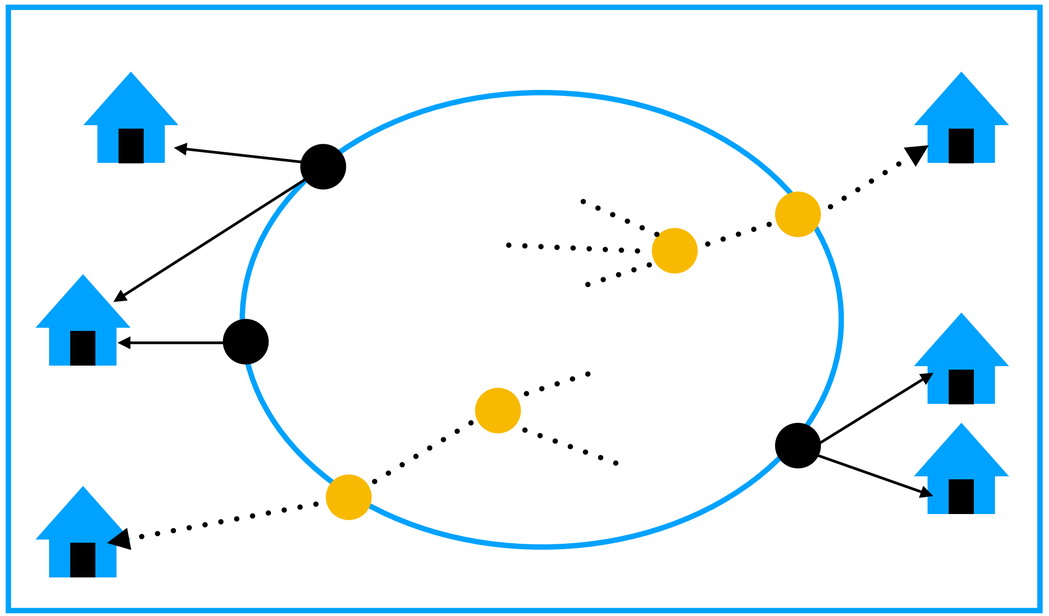


High QoE, efficient delivery: Content is present near a Netflix member

QoE impact, congestion problem: Content is not present at Edge

Edge Caching

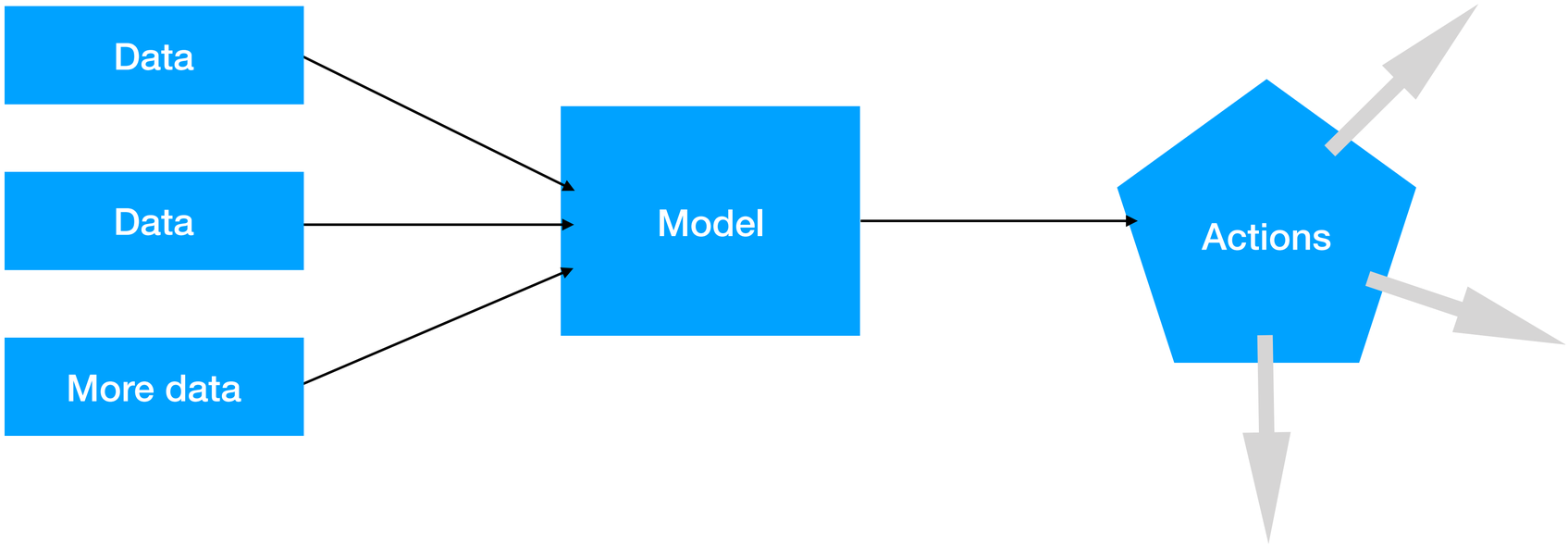
- Goal: Use data on popularity demand to prioritize content updates
  - a) for caching the most popular files
  - b) minimize the number of file replacements (traffic reduction)
- granularity on regional demand level: e.g. for popularity of shows
- granularity on user demand level: e.g. for text files, metadata, encodings etc.
- technical approach: time series forecasting, constrained optimization, network modeling,



Predictive Analytics

Predictive Analytics

- use of historical data to predict future events
- build a model which incorporates past trends
- utilize this model on present data to predict near-future outcome
- optimally: suggest next steps or actions to be taken



Big Data

Big Data

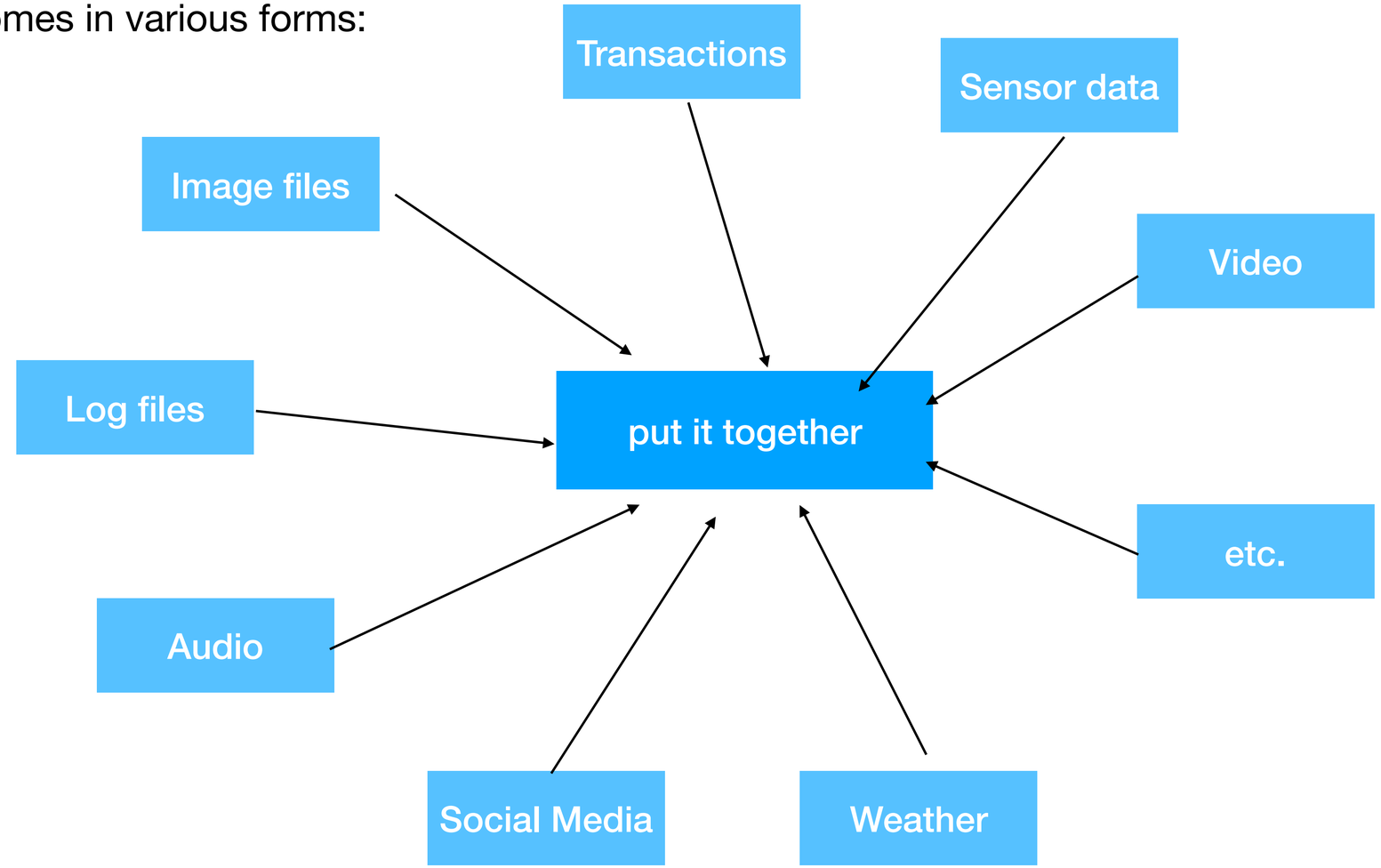
- Data from: Sensors, instruments, connected Systems
- Within a company the following kind of data may be generated e.g.:
  - transaction data
  - sales
  - customer data
  - marketing data
  - data from social media
- Idea: make data-driven decisions based on insights gained from this data

Gaining Advantage

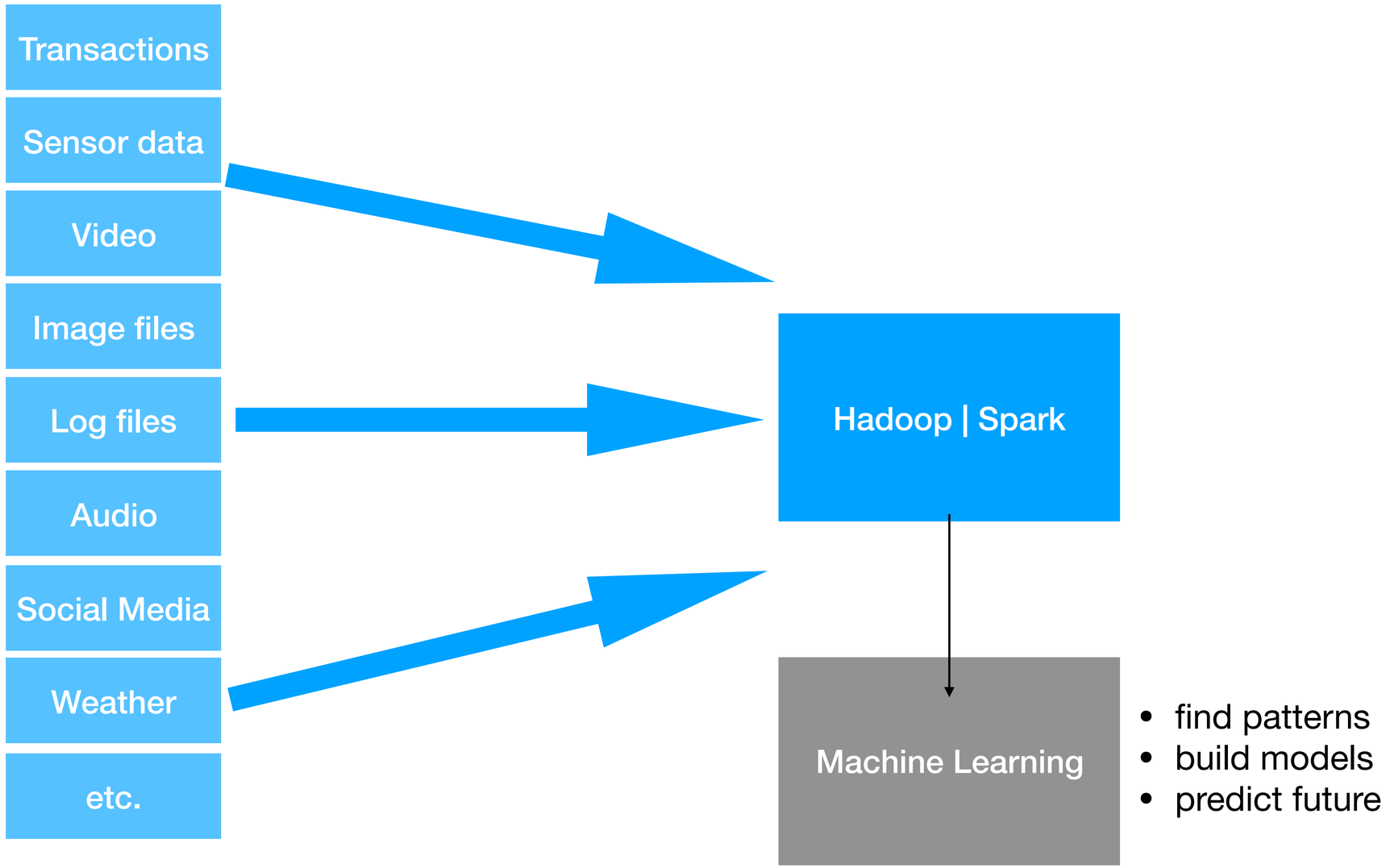
- Innovating and improving may be cumbersome on hardware alone
- it takes a lot of time, money, expertise etc. to develop a new product
- Adding a tool for predictive analysis hence may improve current processes
- Example: Equipment maintenance
  - predict when parts need to be replaced in a factory
  - when does equipment need cleaning, maintaining, renewing?
- forecast failure, energy consumption
- sensor measures heat above a threshold—> replace part causing the heat
- forecast for energy demand in a grid
- forecast weather: rain in Hamburg on Day n —> 60% chance for rain in Berlin on Day n + 1

Data

- Data comes in various forms:



Data


Transactions
Sensor data
Video
Image files
Log files
Audio
Social Media
Weather
etc.
Hadoop | Spark
Machine Learning
• find patterns
• build models
• predict future

Examples for Predictive Analytics

Automotive

- collect data from various sensors in the car
- build better Advanced Driver Assistance Systems (ADAS)
- improve maps from road data, better break systems etc.

Aerospace

- collect data from various sensors in the plane
- collect data from customer behavior
- reduce flight costs, fuel consumption, plane maintenance

Energy sector

- collect data from power consumption in the power grid
- forecast energy consumption (historic + seasonal events)
- optimize power grid functionality

Financial sector

- collect data from transactions, customer behavior
- detect fraud, misuse
- predict finance trends (make money :)

Industrial Automation

- collect data from various machine sensors
- predict machine failure, manufacturing errors
- improve maintenance, reduce machine downtime

Predictive Analytics

- fuse information from analytics, statistics, Machine Learning (ML)
- create a predictive model
- forecast future events
- Question:
  - What can be done to reduce waste?
  - When do we need to put this machine into maintenance?
  - What kind of diseases are trending this winter?

- common ground: some kind of (business) goal