DSLab 2020
The Data Science Lab

Introducing the team

Eric Bouillet
Most modules

Guillaume Obozinski
Modules 1 & 5
Weeks 4, 12-13

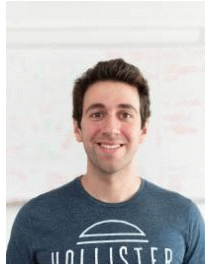
Tao Sun
Assistant

Christine Choirat
Module 1
Weeks 2-3

Sofiane Sarni
Module 4
Week 10

Olivier Verscheure
Most modules

Pamela Delgado
Module 3
Weeks 1, 7, 8 & 9



John Stephan
Teaching Assistant
EDOC-IC

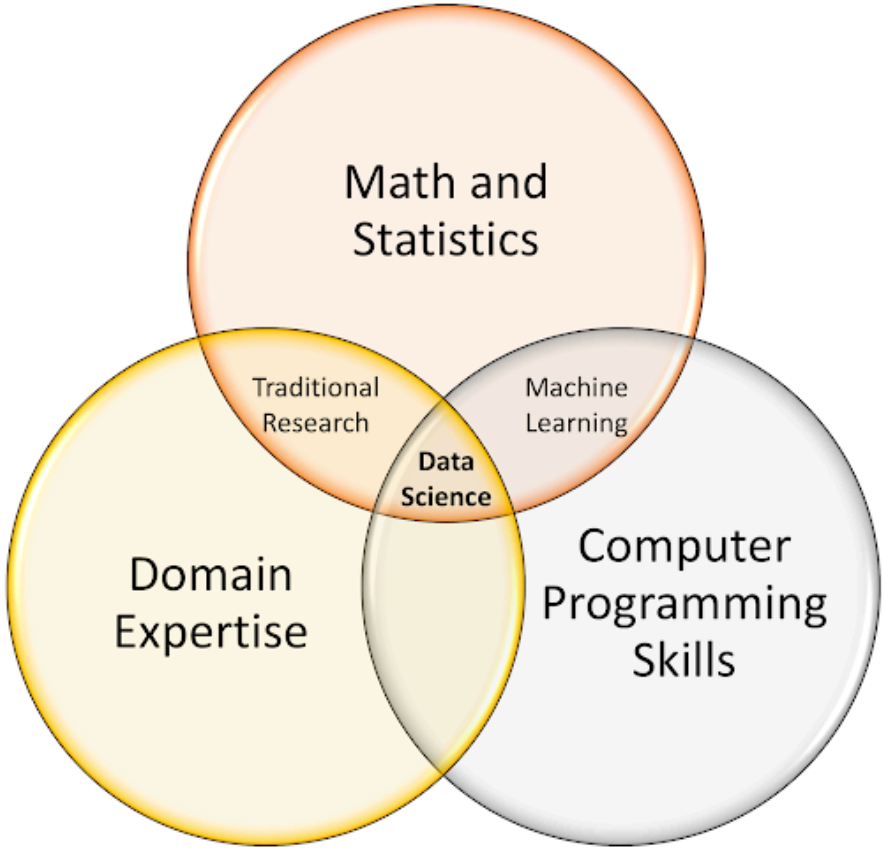
Haoqian Zhang
Teaching Assistant
EDOC-IC

Kayaalp Mert
Teaching Assistant
EDOC-IC

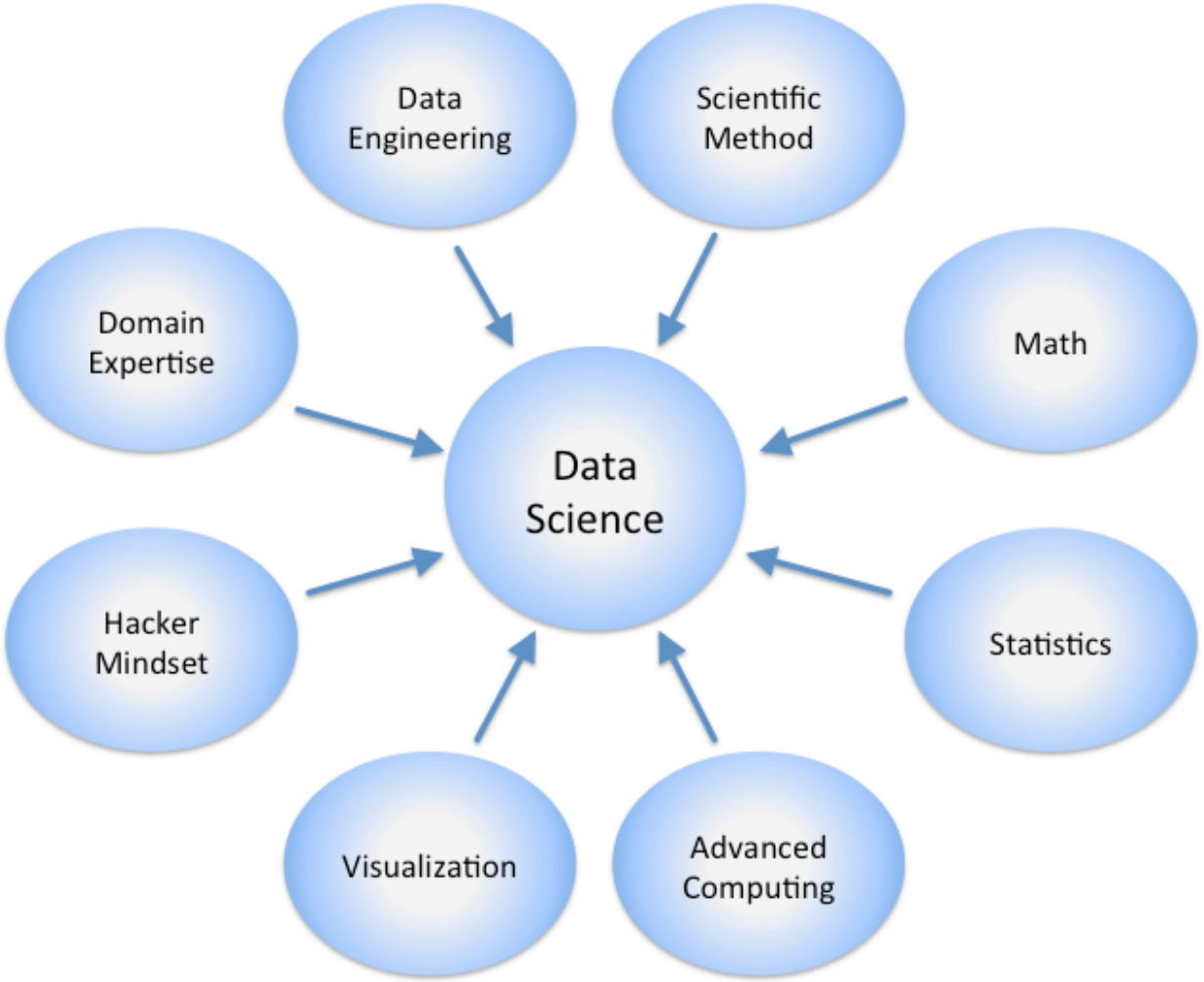
Outline

- General introduction
- An overview of our DSLab
- This week's lab
  - Crash course of Python 3.x in Jupyter Notebook

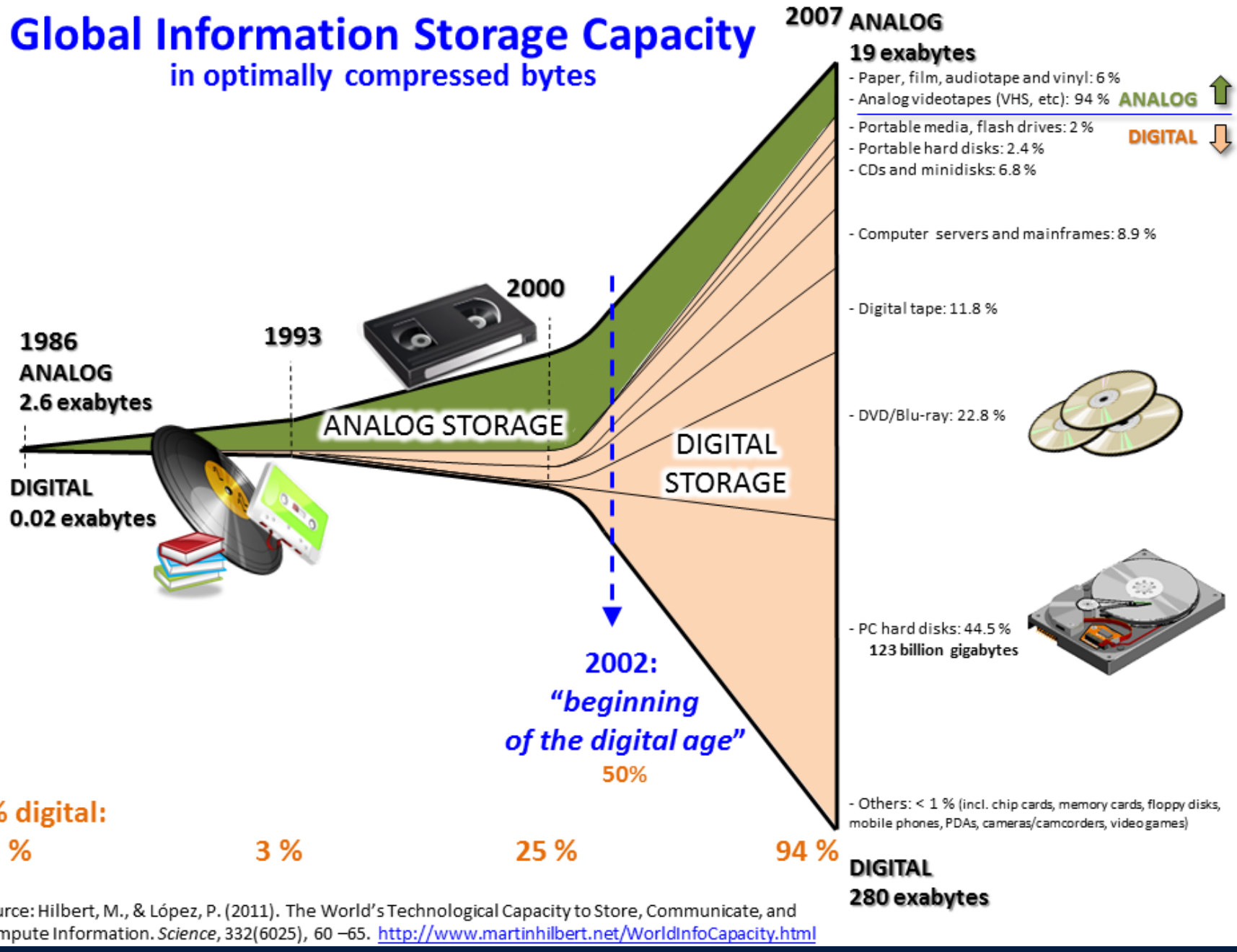
Conway's Data Science Venn diagram



What is Data Science?



Digital age



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

Big Data

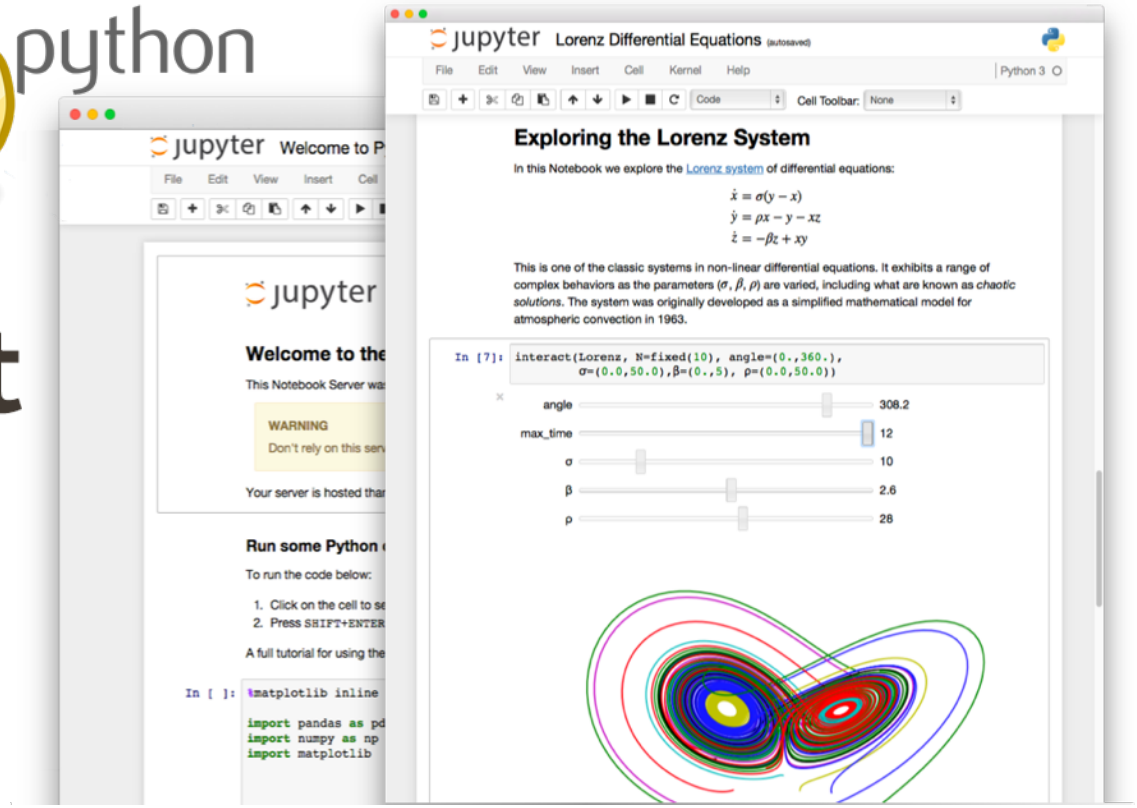
Bibliothèque Nationale de France : 14 To

- In 2013
  - Twitter 7 To/day
  - Facebook 10To/day
- In 2015, Every 60 seconds on Facebook
  - 510 comments are posted,
  - 293,000 statuses are updated,
  - 136,000 photos are uploaded…

Many banks, large stores, companies working in logistics, with sensors, with IoT, webmarketing companies, web platforms, digital factories are generating large amounts of data that are difficult to structure, model and analyze.

Instant quiz

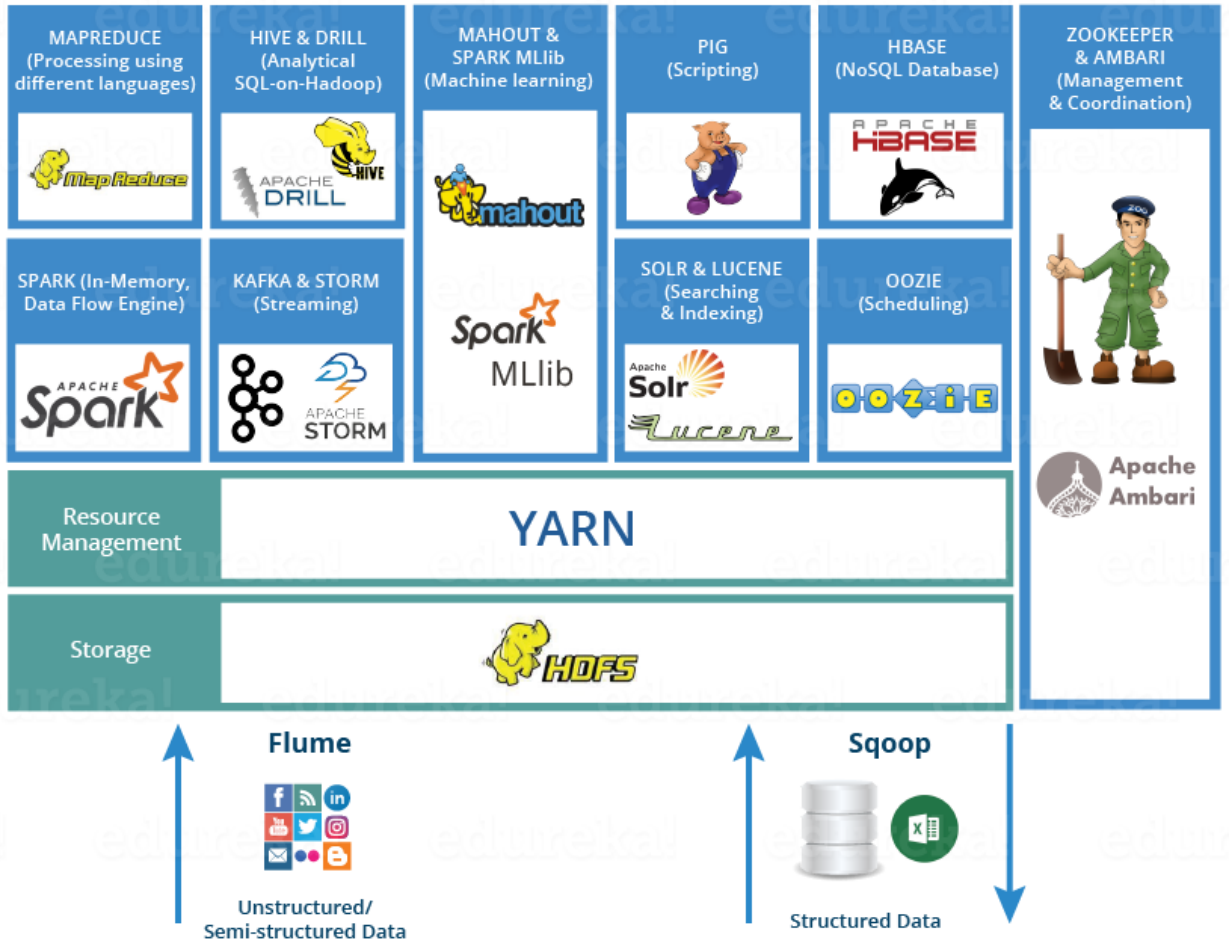
- Python & Anaconda
- Jupyter Notebooks
- Git(Lab/Hub)
- Docker containers
- Kaggle



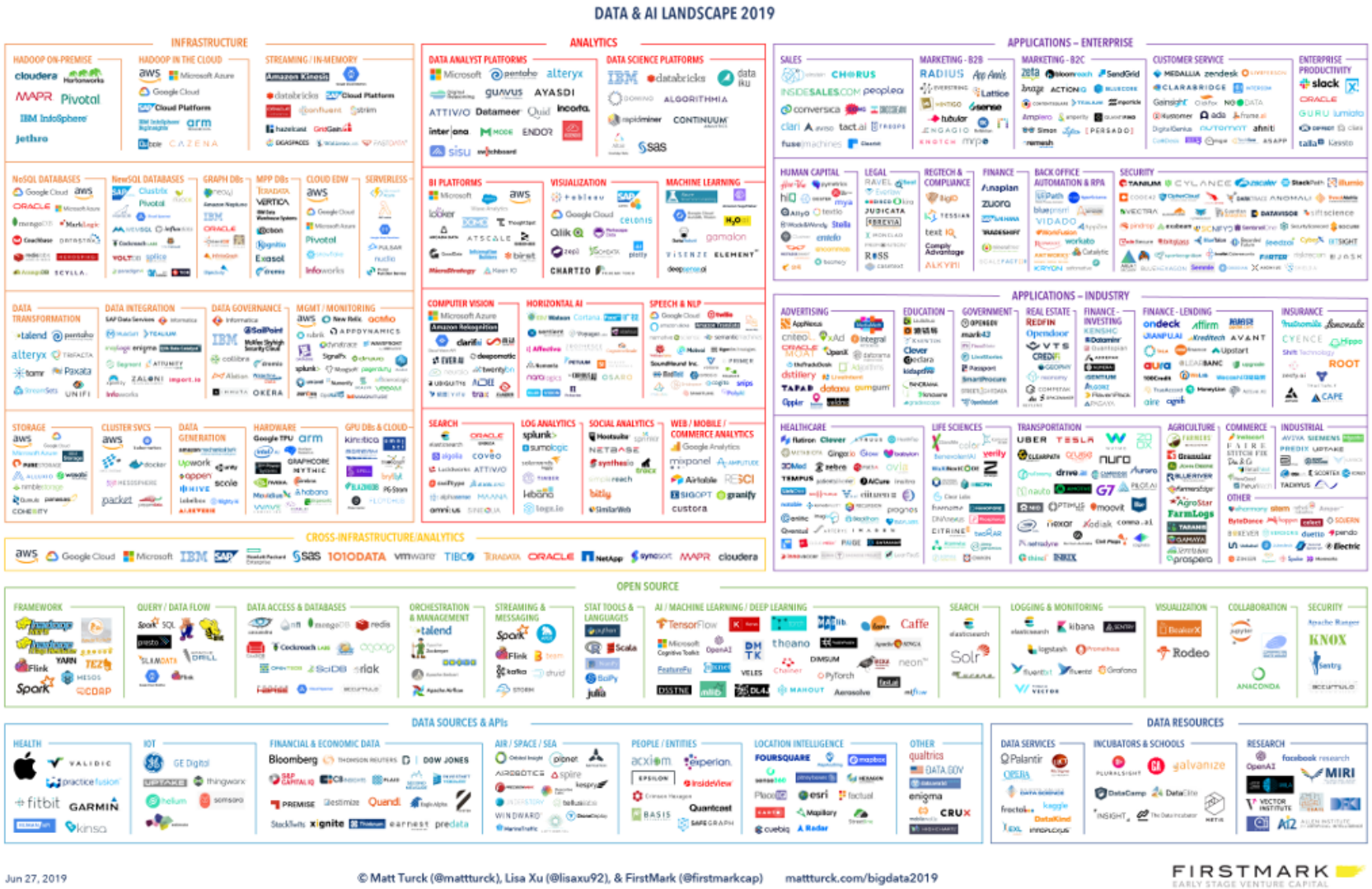
Instant quiz

- Python & Anaconda
- Jupyter Notebooks
- Git(Lab/Hub)
- Docker containers
- Kaggle

- HDFS, YARN, Hive, HBase
- Spark (Streaming)



Data Science tools landscape 2019



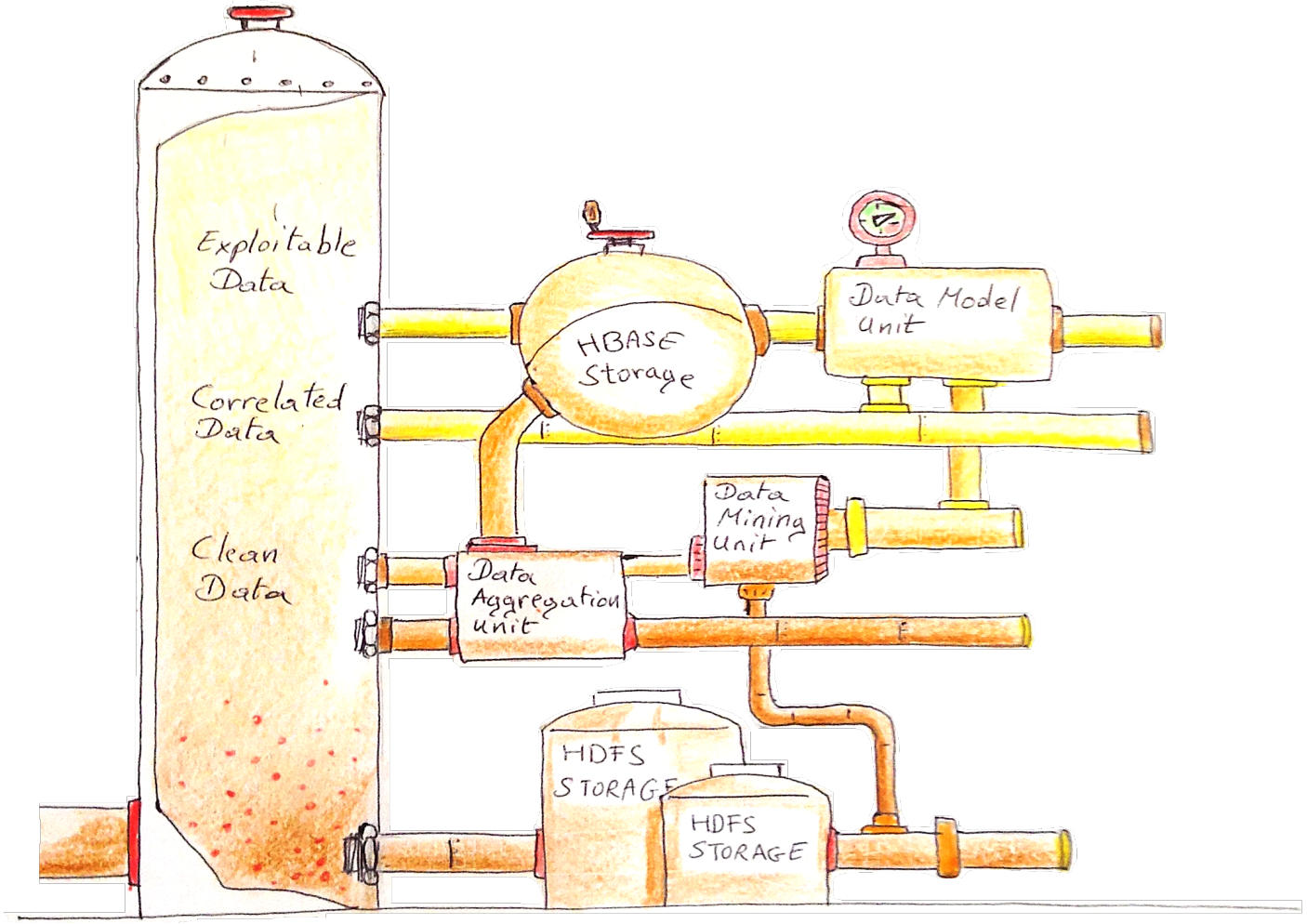
The problem and the data

- Which data for which problem formulation?
- Understanding where the data is…
- Collecting the data
- Determining the data structure (datalake, structured database)
- Finding the meta-data describing the encoding of the data
- Putting in place labelling schemes / fix existing labelling scheme

Big data and data wrangling

- GFS : Google file system
- HDFS: Hadoop Distributed File System
- MapReduce: scheme to process distributed data
- YARN: Resource manager for HDFS
- Spark: distributed cluster-computing framework
- Kafka: work with streaming data (with Spark)

Like oil, data must be refined

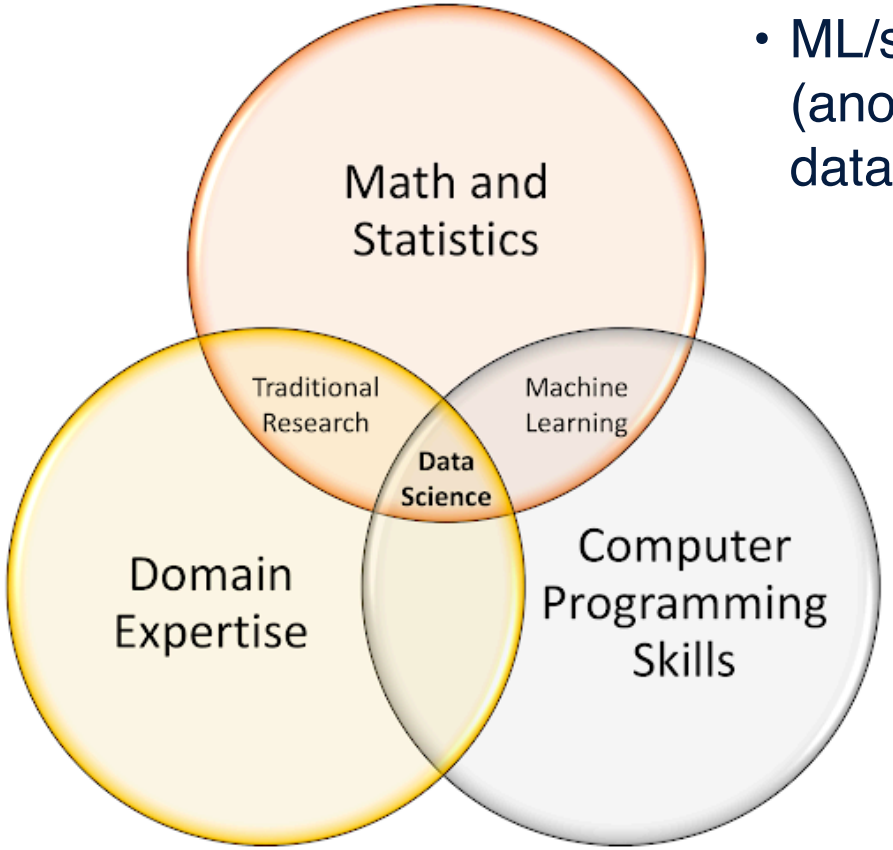


Preparing the data

- 🛢️🛢️ Merge databases
  - 🧰 Record linkage
- 👾 **Errors**
  - Inconsistencies
    - 🧰 Detect / fix / remove
  - Duplicate entries
    - 🧰 Deduplication
  - 👾 **Outliers**
    - 🧰 Anomaly Detection
    - 🧰 Robust Machine Learning
- 👾 **Missing data**
  - 🧰 Imputation techniques
- 👾 **Non-stationarity**
  - Seasonal effects
  - Drifts
    - Set horizon to retrain
  - Sudden changes:
    - 🧰 Change point detection

What you will learn in this lab

- Hear about a number of concrete data science projects on which the Industry Team at SDSC works on with industry partners



- ML/stats for real world data (anomalies, outliers, missing data, etc)

- Hadoop, Spark, Kafka
- Work with large scale data
- **Batch** or **streaming data**

Mission of the Swiss Data Science Center:

Accelerating the adoption of Data Science and Machine Learning techniques within academic disciplines of the ETH Domain, the Swiss academic community at large, and the industrial sector in Switzerland.

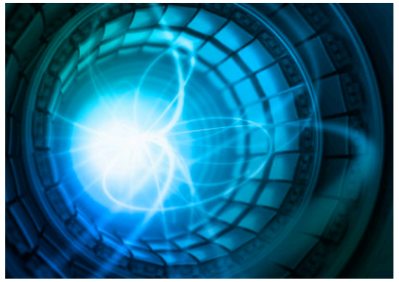
Academic team: 16, Industry team: 12, Renku/engineering team: 15

SDSC website: https://datascience.ch

Master Students projects: https://www.epfl.ch/research/domains/sdsc/

A few of our academic projects

Beyond Weather – Extending the PrEdiCTability of the Atmosphere over Europe



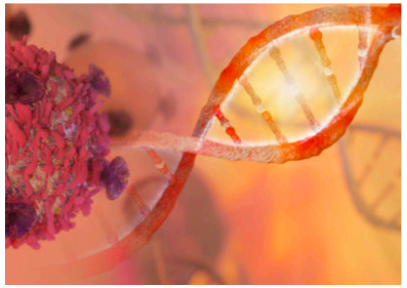
PACMAN – Particle Accelerators and Machine Learning

DEAPSnow – Improving snow avalanche forecasting by data-driven automated predictions

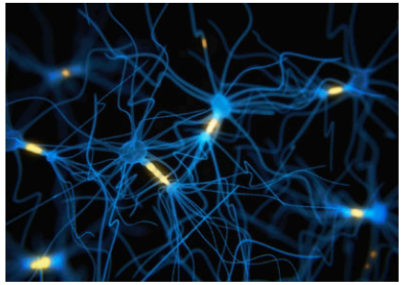
TAPEDA – Towards Automated Post-Earthquake Damage Assessment

MSEI – Molecular structure elucidation by integrating different data mining strategies

EconMultiplex – Multiplex Networks in International Trade



Transposcriptome-based identifier for precision medicine – TE4med



Extracting Neural Activity Signals from Large-scale Calcium Imaging Data – Neuro-Choice

SDSC

EPFL

Introducing the lecturing team

Eric Bouillet
Most modules

Guillaume Obozinski
Modules 1 & 5
Weeks 4, 12-13

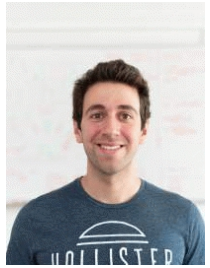
Tao Sun
Assistant

Christine Choirat
Module 1
Weeks 2-3

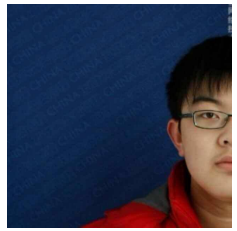
Sofiane Sarni
Module 4
Week 10

Olivier Verscheure
Most modules

Pamela Delgado
Module 3
Weeks 1, 7, 8 & 9



John Stephan
Teaching Assistant
EDOC-IC



Haoqian Zhang
Teaching Assistant
EDOC-IC

Kayaalp Mert
Teaching Assistant
EDOC-IC

An overview of our lab

4+1 Modules in 14 weeks

Final project (**3 weeks**)

1. Crash-course in Python for data scientists (**2 weeks**)
2. Distributed computing with a Hadoop Distribution (**3 weeks**)
3. Distributed machine learning with Apache Spark (**3 weeks**)
4. Real-time data acquisition and processing (**2 weeks**)

- Data science as a journey!
- Very hands-on and practical
- 3+ instructors for every lab

Course webpage: http://epfl-dsplab2020.github.io

The labs using Renku

RENKU 連句

A software platform for reproducible and collaborative data science.

Eric Bouillet

Rok Roskar

Olivier Verscheure

Christine Choirat

- Renku is a form of Japanese collective poetry
- Renku= a platform entirely developed at SDSC (12 senior software and systems engineers)
- Goal: reproducible collaborative research in data science
- It is version-control solution for your whole data science environment (code, data, execution pipeline)
- Environment independent thanks to Dockers
- Useful to teach in hands on computer science
- It supports open science, traceability and reproducibility of science

- https://datascience.ch/renku/

Assessment

- 60% continuous assessment during the semester
  - One project per module
  - Groups of 4 students
  - Projects graded within 2 weeks

- 40% final project
  - Final project in the classroom
  - Groups of 4 students

Hardware / Software Resources

- Please bring your own laptop!
- Renku platform
  - IC Cluster of 4 servers
- Hadoop Cluster
  - Hortonworks Data Platform
  - IC Cluster of 12 servers

Logistics

- Lab on Wednesday's 13:00 – 16:00 in INF 01
- Lab's github: https://epfl-dslab2020.github.io/
- Slack epfl-dslab2020.slack.com
- Office hours will be announced during homeworks

Module 1: Crash-course in Python for data scientists

- Week #1
  - Jupyter Notebooks
  - Python 3.x
  - NumPy, Pandas, Matplotlib, Scikit-Learn
- Week #2
  - Reproducible data science
  - Git, Docker, Renku

Module 2: Distributed computing with Hadoop

- Week #3
  - Introduction to big data, best practices and guidelines
  - Loading & querying data with Hadoop
  - HDFS, Hive
- Week #4
  - Data wrangling with Hadoop
  - Assessed project 1
- Week #5
  - Introduction to distributed computing and the Spark runtime architecture
  - Python on Spark
  - Use basic RDD manipulations

Module 3: Distributed processing with Apache Spark

- Week #6
  - Spark data frames
  - Assessed project 2

    Scaling up to Hadoop cluster with Hive and Spark
- Weeks #7
  - Advanced python for Spark, Spark optimization
  - Spark pipelines, Spark MLlib, classifiers
- Week #8
  - Assessed project 3

    Machine mining with Spark

Module 4: Real-time data acquisition and processing

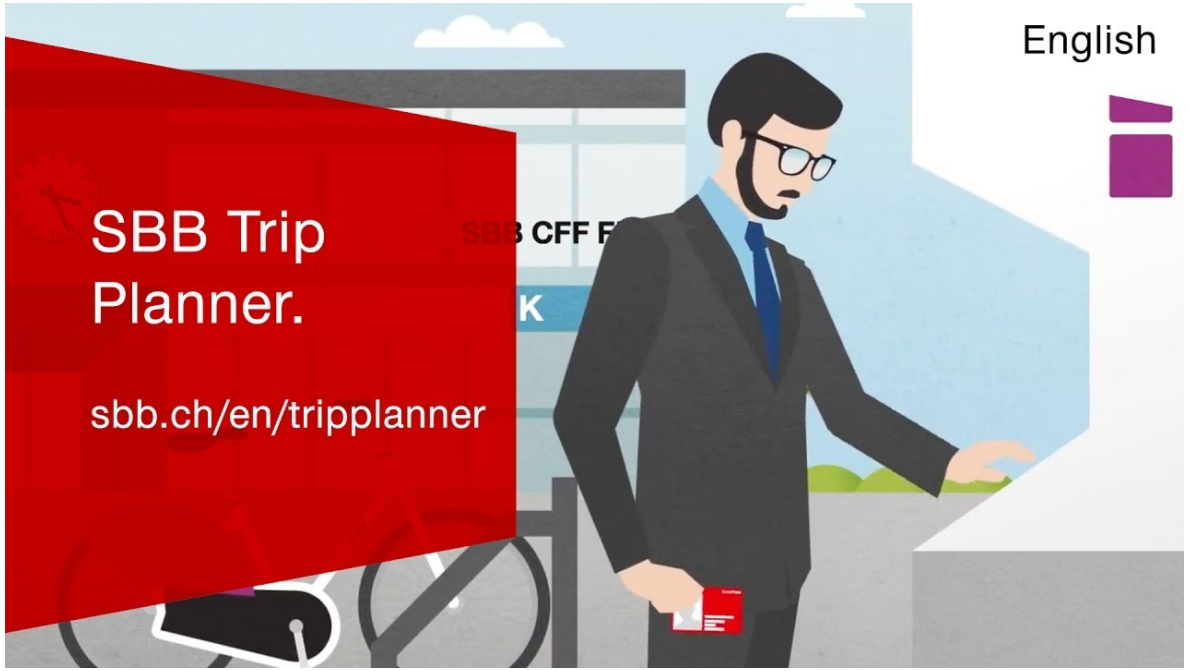
- Week #9
  - Introduction to data stream processing
  - Overview of MQTT as sensor protocol for IoT
  - Apache Kafka for stream processing

- Week #10
  - Advanced data stream processing concepts on Spark with Kafka
  - Assessed project

    Process streaming data from real-time train geolocation data

Module 5: Final assignment

Robust Journey Planning



Module 5: Final assignment

- Week #11 - #17
  - Teams of 4
  - 6min video presentation + 10 mins Q&A

Today’s assumption of a deterministic world

Meeting Zurich HB @ 10:30… from St-Sulpice

NFB 701 Direction Morges, gare 07:20 — 09:56	2 h 36 min	1	1. 2.
NFB 701 Direction Morges, gare 07:40 — 10:28	2 h 48 min	2	1. 2.
NFB 701 Direction Lausanne, Bourdonnette 07:47 — 10:28	2 h 41 min	3	1. 2.
NFB 701 Direction Morges, gare 08:20 — 10:56	2 h 36 min	1	1. 2.

16 minutes to catch train in Morges

NFB 701 Direction Morges, gare
07:30 — 07:44 | 14 min | 0 | 1. 2.

07:30 **St-Sulpice VD, Chantres**

NFB 701 1030 Direction Morges, gare

07:44 **Morges, gare**

3 ' Parcours à pied

08:00 Morges, Voie 3

IR 90 **1711** Direction Brig 1. 2.

08:12 Lausanne, Voie 3

Changement

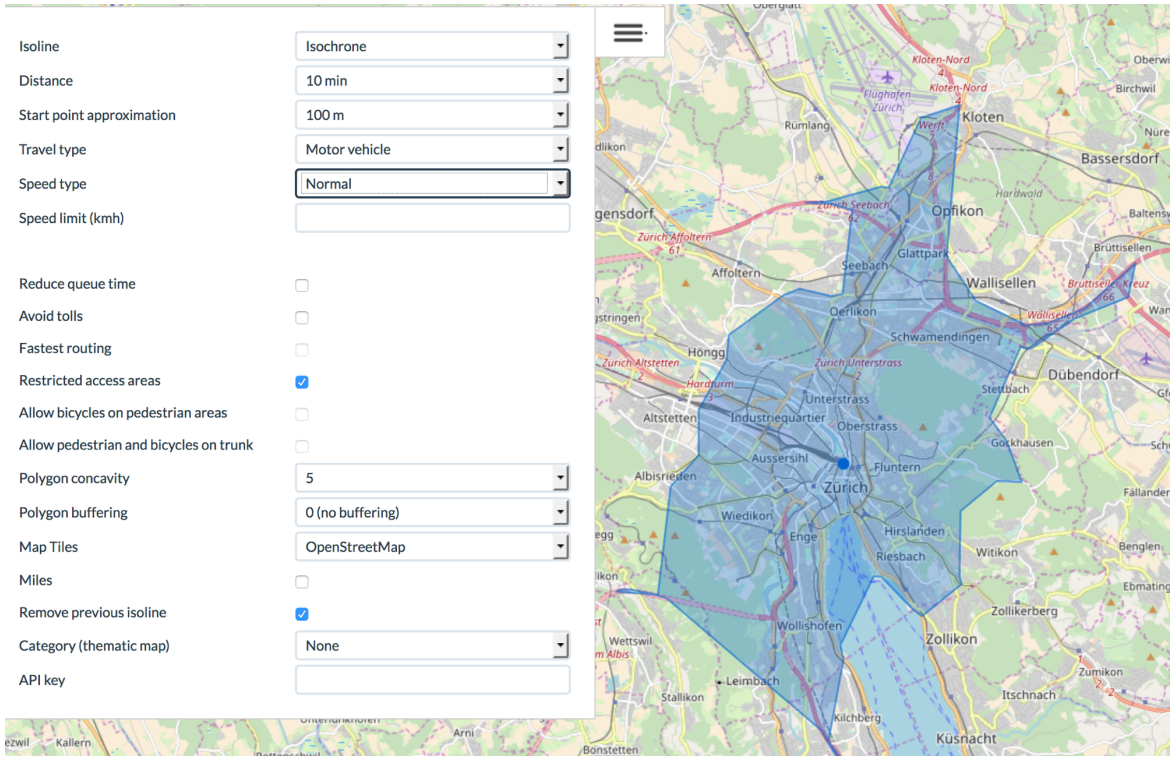
08:20 Lausanne, Voie 1

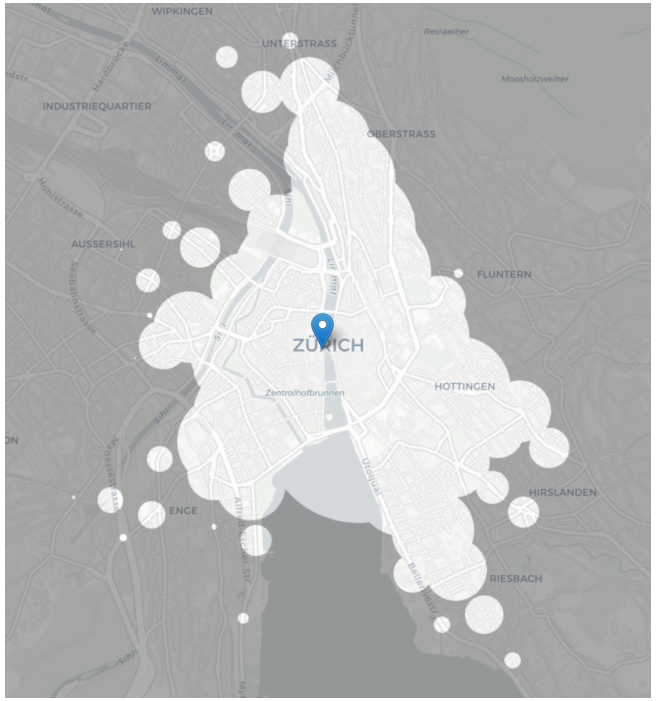
IC 1 **711** Direction St. Gallen 1. 2. FA BZ RZ R

10:28 **Zürich HB, Voie 33**

Overall objective

- **Display isochronous map**
  - Start your journey e.g. at Zurich HB
  - **How far can you go within M minutes Q% of the time?**





Rest of today's module

Sylvia Quarteroni
Presentation of the
Industry team

Pamela Delgado
Jupyter notebooks with Renku
Python starter and scientific toolkits