UNIVERSITÄT BERN

CAS Applied Data Science - Module 1

Data Acquisition and Management

PD Dr. Sigve Haug

Bern, 2022-08-24

Module 1 Purpose and Format

Purpose

- Think about data
- Get used to the tools for working with data
- Establish the tacit skills needed for the other modules

Not very theoretical (if you already know a lot, may work with the notebook on your dataset)

Format

- Presentations intersected with discussions and hands on work
- Not inverted class room (as module 2)

Module 1 Overview

First day

- Introduction and welcome
- About data and data science
- Data Management
- Data Infrastructures

Second day

- Visualisation of data
- Web scraping and APIs

Third day

- Databases
- Project clarifications

Project/Goal

- Produce a Conceptual Design Report for a Data Science Project (deadline 2022-10-XX to be defined)

First morning

09:30 What is data and data science ?

- Data and Big Data (09:30)
- Jupyter and Colab (10:00)

0 Break

11:00 Data Management

- I/O
- Indexing, Filtering, Sorting ...

12:30 Lunch

Data (latin datum in singular = thing given)

Data

- Term first used in relation with computers in the 40-50ies
- Plural but often used in singular
- Means **any sequence of symbols**
- Needs processing and interpretation to become **information**
- A lot of information and experience become **true belief / knowledge**

- Digital (represented by 0 and 1) or analog
- Moves in serial or parallel
- Stored on magnetic (tapes, hard disks, SSD, RAM ...), optical (CD, DVD) or mechanical devices
- Most computers work on digital data and with the binary numeral system (“alphabet”)

Data

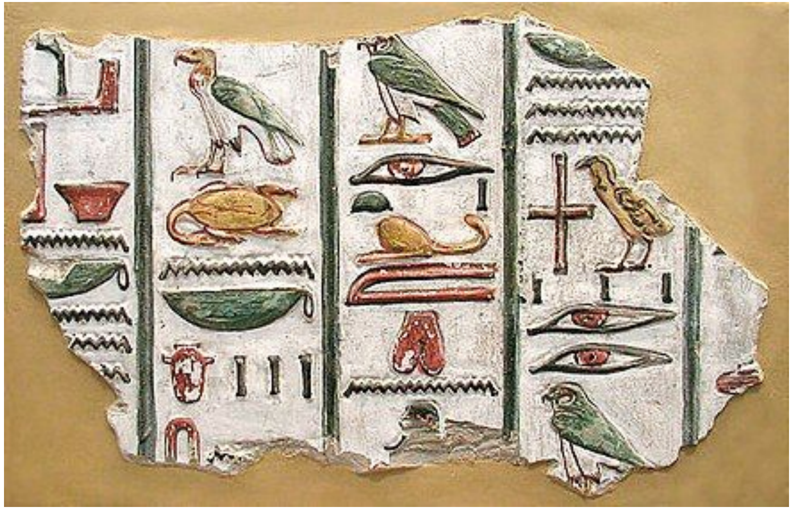
Data example

- Radius of the earth
    - 46 100 km
- In this case value with unit
- Normally only the value is stored

Metadata (data about data) example

- Unit: km
- Author: Eratosthenes of Cyrene
- Date: 240 BC
- Location: Egypt
- Method: Well, stick, sun shadow

Data examples



?

Figure shows data from Mesopotamia or Egypt (?) 3-5k years ago.

“Hello!” in binary form:

01001000 01100101 01101100 01101100 01101111 00100001

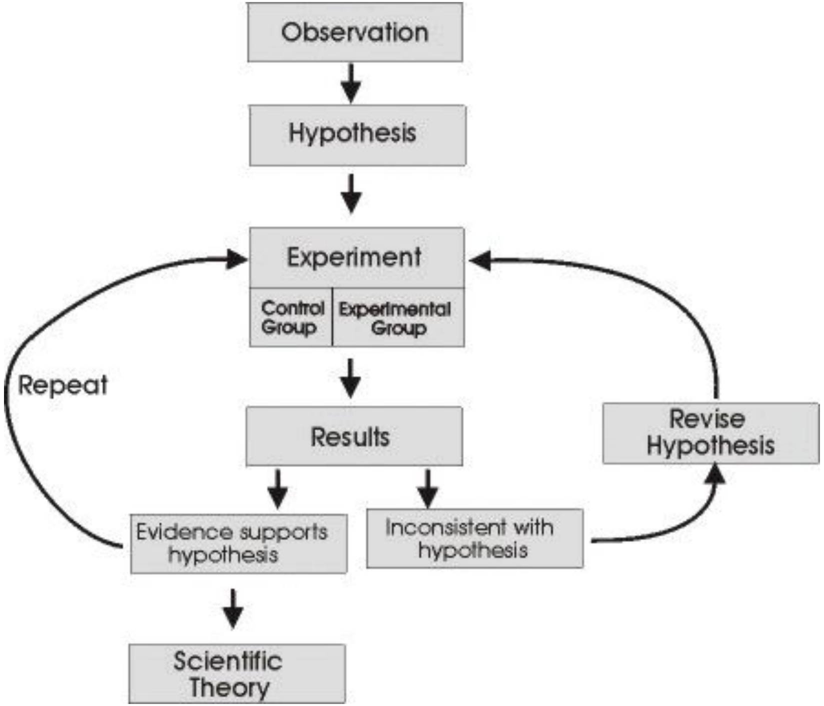
1 character has 8 bits

Data -> Information -> Knowledge -> Decision

Data and Science

Science

- The enterprise of building and organising knowledge
- Ideally based on reproducible experiments (good practise since Galileo)
- Should produce falsifiable predictions (normative definition from Karl Popper)



Data Science

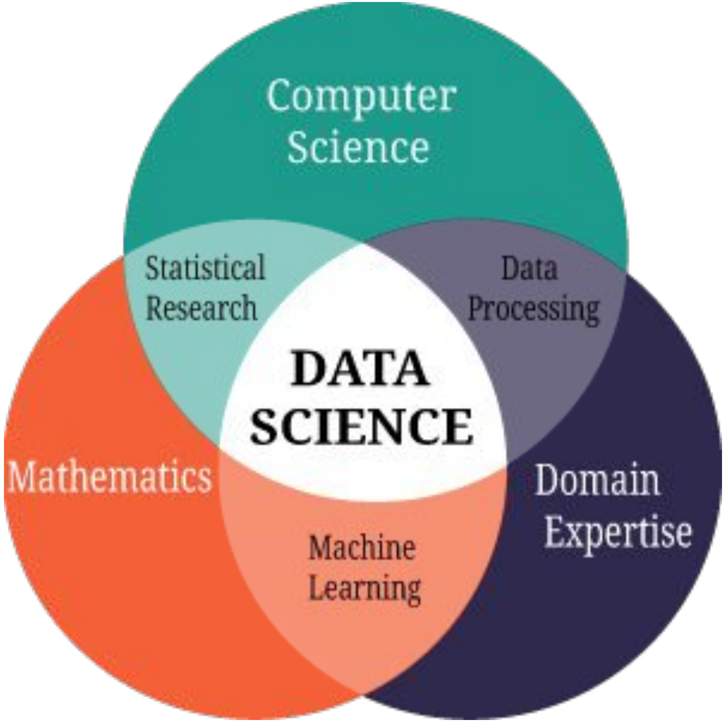
Uses

- Mathematics and Statistics
- Computer Science
- Domain expertise

on data to build information and extract knowledge (for decisions and actions)

It is the theory of making science with data.

Very general skills increasingly needed in all empirical research and business



Data Representations

For data science we need data in numeral representation

Numeral Systems

- Often data is represented by numeral systems (however also by the alphabet -> NLP)
- Writing systems for expressing numbers using e.g. digits
- Modern systems are mostly positional systems
  - $304 = 3\times10^2 + 0\times10^1 + 4\times10^0$

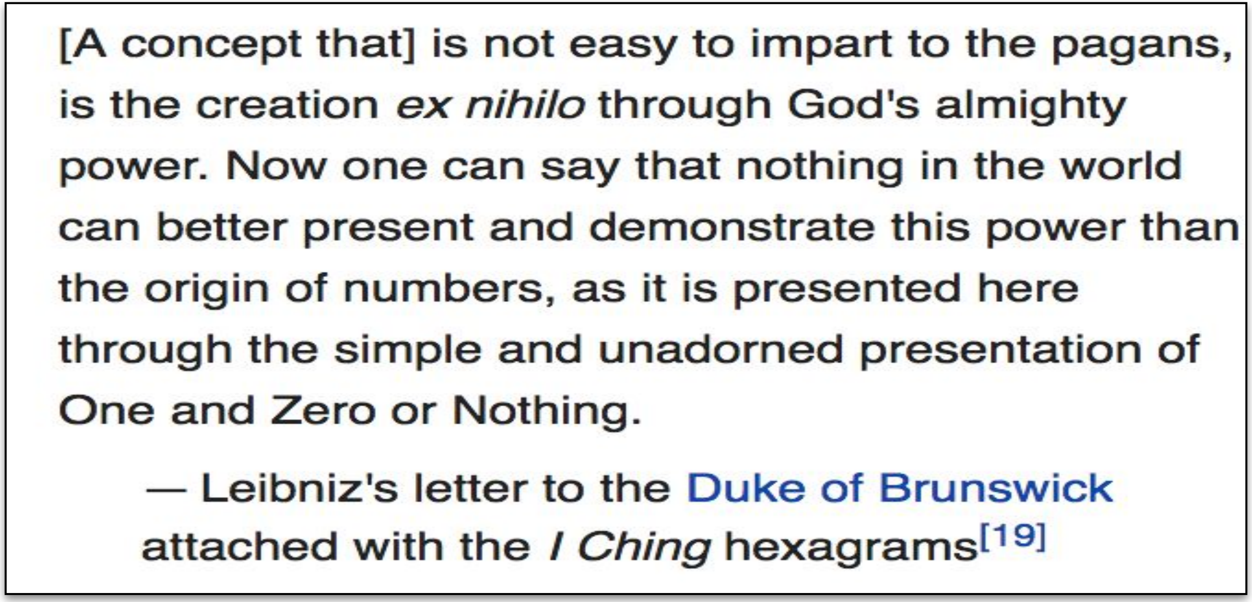
Examples Numeral Systems

- Unary (2 = //) - base 1
  - Good for human kids

simple counting

- Binary (2=10) - base 2
  - Good for computers
- Decimal (2=2) - base 10
  - Good for grown up humans
- Sexagecimal - base 60
  - (remnants in hour, minutes etc)

Most Computers understand only 0/1



> [A concept that] is not easy to impart to the pagans, is the creation *ex nihilo* through God's almighty power. Now one can say that nothing in the world can better present and demonstrate this power than the origin of numbers, as it is presented here through the simple and unadorned presentation of One and Zero or Nothing.
>
> — Leibniz's letter to the Duke of Brunswick attached with the *I Ching* hexagrams[19]

Data

Binary numbers

- Base 2, e.g. 0 and 1
- Computers work with electrical currents, either there is a current (1) or there is no current (0)
- Other numbers and characters can be represented with binary numbers
- One can do mathematics with 0 and 1

So in the end data is mostly stored as bits and processed as bits

From 0 to 3 and so on

- 0000 0000 (8 bits = 1 byte)
- 0000 0001
- 0000 0010
- 0000 0011
- ….
- 1111 1111 (2^8 = 256)

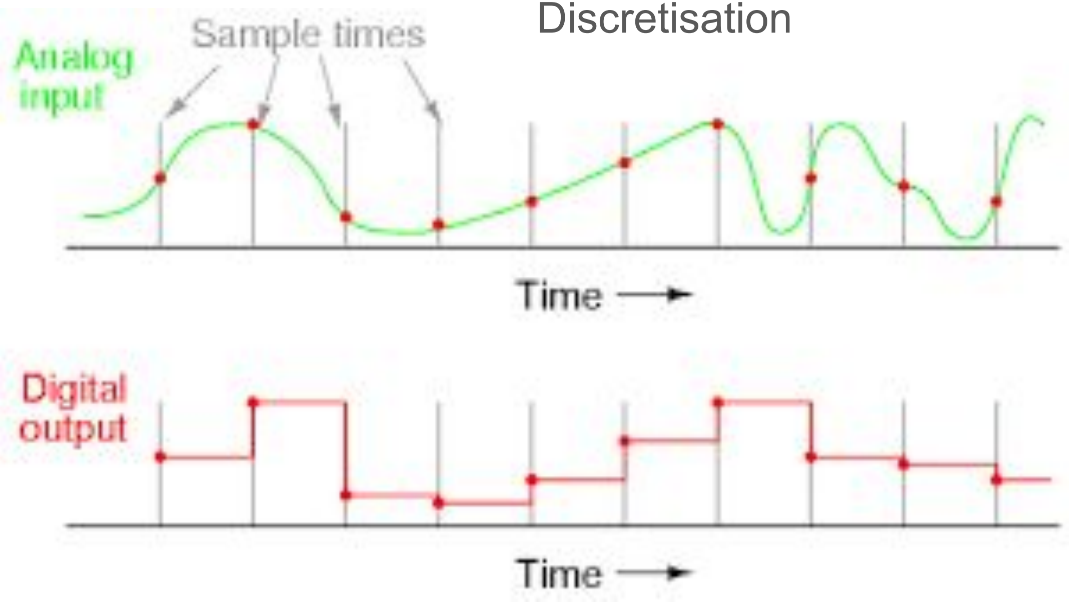
Analog to Digital Data (Digitisation) - ADC

Digitisation

- Sensors often produce analog data - often perceived as some continues wave
- Is then often converted to digital data for computation and storage as bits

Lately an inflationary usage occured - digitisation is now also a social process and every institution needs a digital strategy

Discretisation



Quantisation

Data types and structures (in programming languages)

Common types

natural numbers

- Integer (e.g. 32-bit int)
- Floating point (for real numbers), 32 or 64-bit (double) float in computer: only approximation
- Boolean (True or False)
- Character (a,b,c ...)
- String (“list of characters”)
- List or Array ([1,2,r,t,5])

Explicit and implicit declaration

- In programming languages data is loaded into variables of certain types
- In Python and R types normally don’t have to be specified.
  - counter = 2
- In C/C++, Fortran … data types must be specified
  - Int counter
  - Float strength

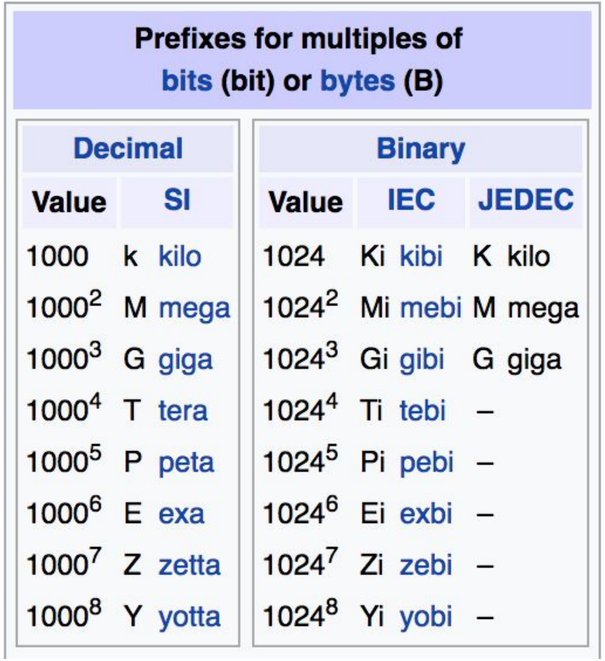
Data types and structures

Composite data types

- Arrays, matrices, **dataframes**
- In C/C++:
  - Structure example: Person (name,age,gender)
  - A structure with method (s) or functions are called classes
  - An instance of a class is often called an object
- Trees
- Graphs
- ...

In R and in the Python pandas module, **dataframe** is the essential (composite) data type, if we consider (time series as a certain type of dataframe)

Data and some vocabulary



Prefixes for multiples of
bits (bit) or bytes (B)

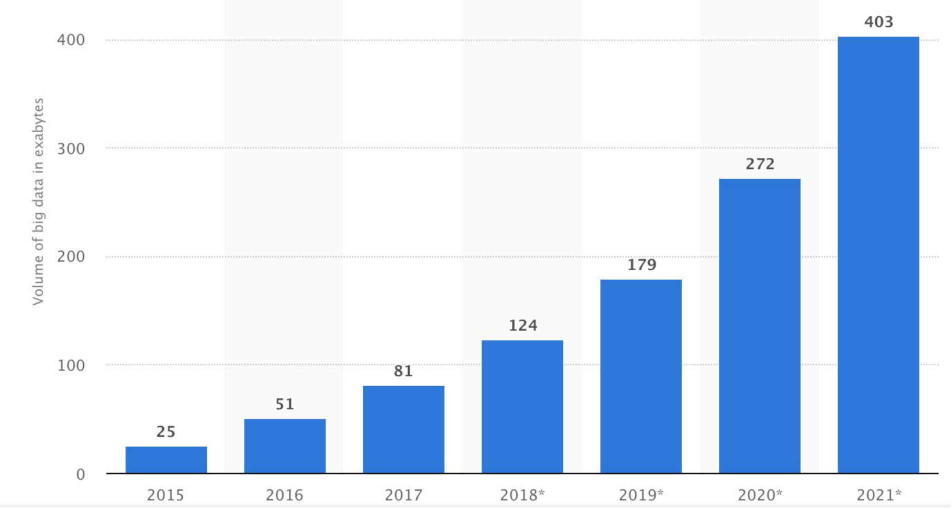
Decimal			Binary				
Value	SI		Value	IEC		JEDEC	
1000	k	kilo	1024	Ki	kibi	K	kilo
1000^2	M	mega	1024^2	Mi	mebi	M	mega
1000^3	G	giga	1024^3	Gi	gibi	G	giga
1000^4	T	tera	1024^4	Ti	tebi	–	
1000^5	P	peta	1024^5	Pi	pebi	–	
1000^6	E	exa	1024^6	Ei	exbi	–	
1000^7	Z	zetta	1024^7	Zi	zebi	–	
1000^8	Y	yotta	1024^8	Yi	yobi	–	

Unit prefixes

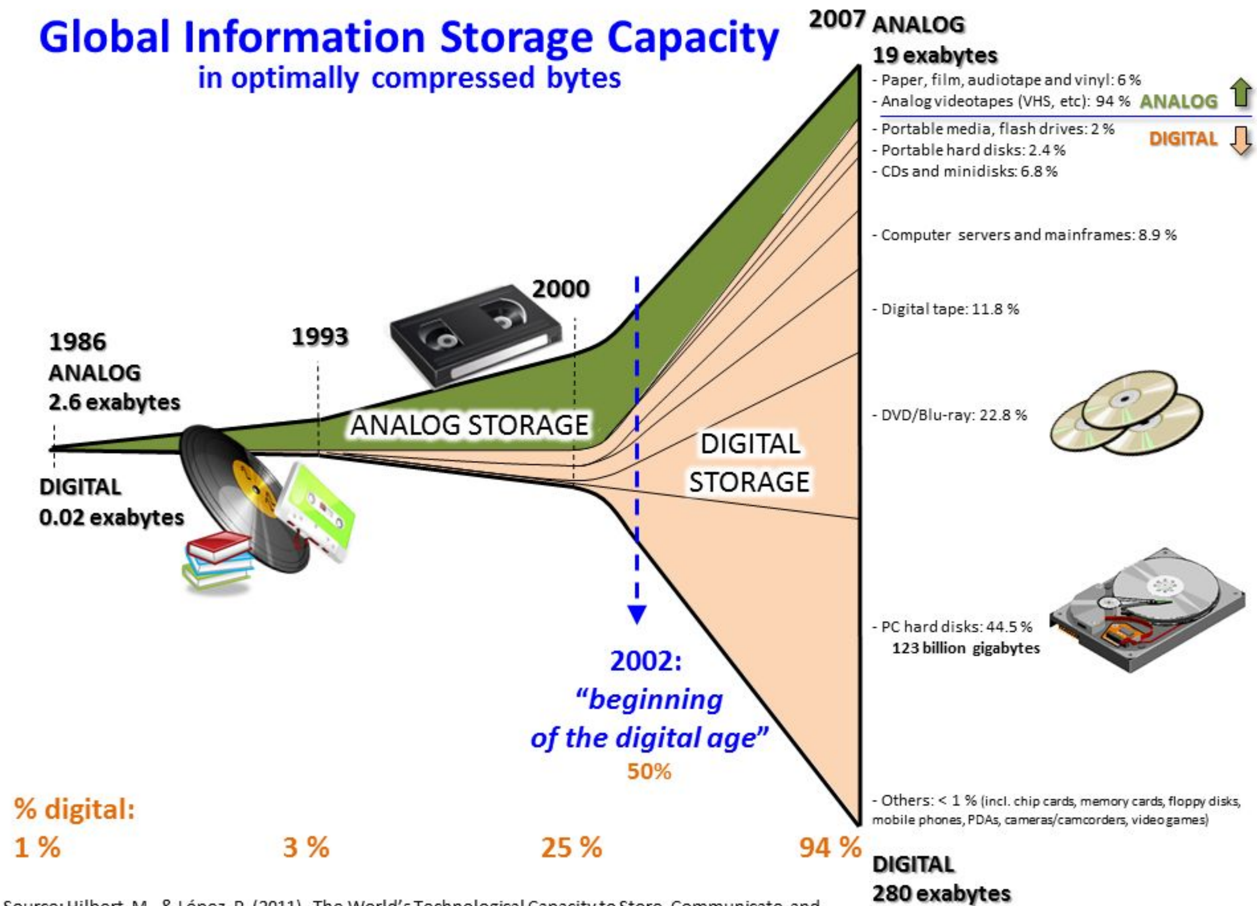
Data volumes

Sizes

- An integer number is often 4 bytes
- A character is often 1 byte
- My 2017 laptop has 8GB RAM
- In 2007 all digital data was 281 EB
- In 10 years one expects 40 EB of new genetics and genomics data
- (Microscope) Color image - 3 numbers per pixel



- Statista.com : big data in data centers



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:
excel can handle about 1 million rows
data that cant fit e.g. on your laptop

What is big data ?

Relative

- Too big for your “traditional” tools
    - Paper is not good for hundreds of rows
    - Spreadsheets are not good for more than million rows
    - Database management system may not be good for data more than the computer memory

Advanced processing

- Parallel processing over distributed systems
- Systems are orchestrated with HPC/SLURM, Hadoop, Spark … or grid technologies
- The processing is done with C++, Python, R or other programming languages with parallelism capabilities

Data Quality

Definitions

- Condition of the values of the variables
- ISO900:2015 Data quality can be defined as the degree to which a set of characteristics of data fulfills requirements.

for example: Iris is measured by milimeters - centimeter-level would be too low

Common Characteristics

- Accuracy
- Validity
- Completeness
- Availability
- ...

The best analysis, ML and inference will not help you if the data quality is poor

Data Quality

Assurance

- Profile (descriptive statistics) and cleanse data

Control

- Before and after the Assurance
  - Restrict the inputs
  - Check if the quality is according to the requirements

Data cleansing

Mostly it is needed to clean (preprocess) data with respect to

- Consistency
- Format
- NA/NAN
- Outliers
- ...

before processing/analysing it

Data is stored in files

Files

- Files have different format (standards) for different purposes
- Modern file systems on computers support sizes up to some TB
- Many small files are hard to organise
- Big files are hard to move

File formats

- Binary file (not human readable)
- Text (ascii)
- Comma separated values (csv)
- gif, pdf, tiff … (graphics)
- mp3, mp4 (sound, video ...)
- And so on

Summary and Readings

Test yourself

- What is data, science and data science?
- List common data types and structures
- What are binary numbers and why are they important?
- What is big data?
- What is data quality?

Literature

- Wikipedia on all topics
- Zacharias Voulgaris, Data Science: Mindset, Methodologies, and Misconceptions, 2017
- D. E. Holmes, Big Data - a very short introduction, 2017
- D. Faser, A. Meier, Big Data (german/swiss), 2016

Time to get started with Jupyter and Colab

So return to your notebook from which you arrived to these slides

(you may install Anaconda later)

Second half day

13:30 Infrastructures for data

- ...

14:00 Work on Notebook

14:30 Data sources and modelling

- Data flow
- Data modelling

15:00 Break

15:30 Team Work

- 16:00 Team work presentations

16:45 Summary

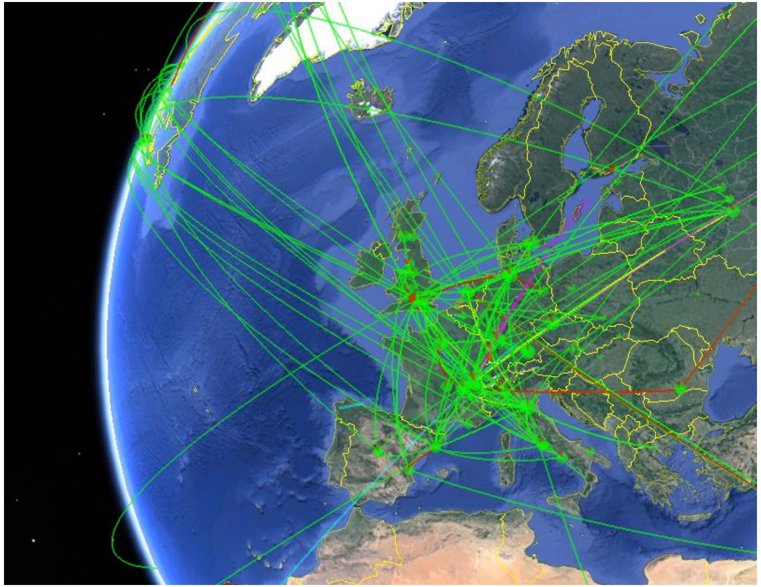
17:00 End

Infrastructures for data

Parts (of a global system)

- Sensors
- Storage
- Compute
- Network
- Interfaces for humans

The essential part is the computer



Snapshot of data movements in the WLCG data infrastructure

What is a computer ?

- (Bring a server to look at…)
- Motherboard
  - CPU, GPU
  - RAM
  - Connections / Interfaces
- Disk (hard disk, solid state disk etc)
- Network connections
- Video cards / GPU
- Power supply

An open server

Typically you need to care about CPU, RAM and disk. The rest is in the hand of the system administrators.

Infrastructures for data

From small to big - scaling

- R&D is normally done on small datasets, i.e. laptop size infrastructure
- Production may start small but can scale quickly, larger infrastructures are needed
- You may start with 10 users/customers/MB per day. One server is enough. May need to serve million times as much in some years

...

- Such scaling is not possible/feasible on premise infrastructures (buildings, electricity, procurement, expertise ..)
-
- Move application to cloud providers or other providers of larger infrastructures
- Scaling beyond the limits of one data center is traditionally called grid computing (we don’t talk much about that)

What is a cluster / data center ?

- A set of servers
  - Compute (head node, login node, computing nodes)
  - Storage
- Switch(es) with interconnect and link to internet
- Software connecting the servers via the switch

As a user you typically need to care about the login node and the connecting software, e.g. the one on UBELIX is called Slurm. (older: Pbs, Torque, GridEngines …)

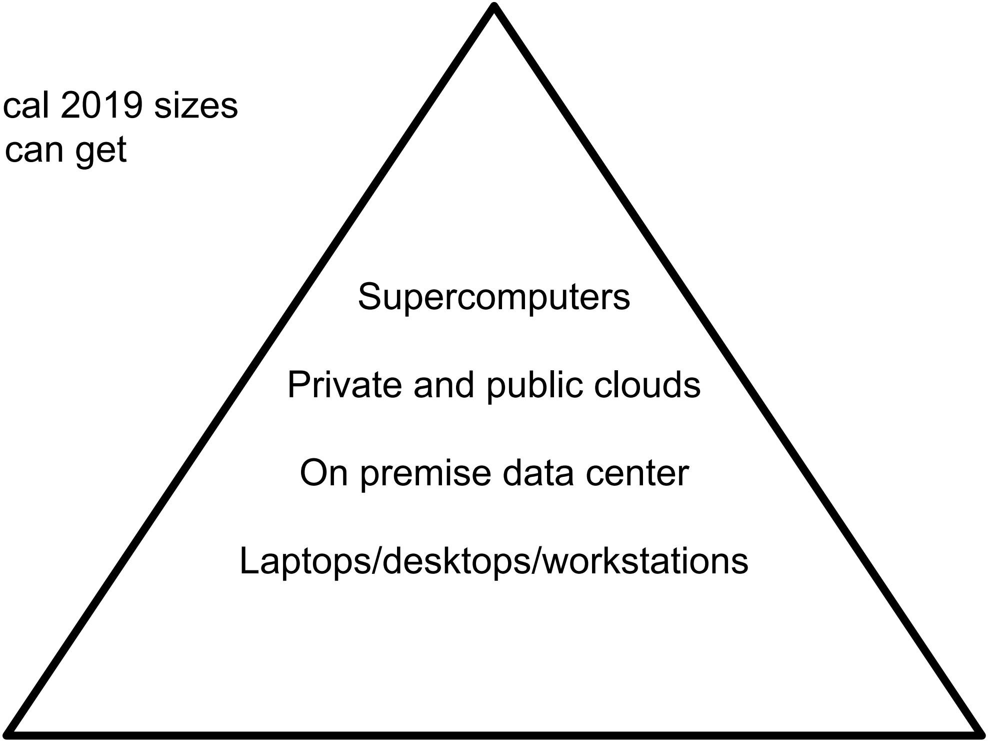
From laptops to supercomputers (academic)

Department Cluster
Central ID Cluster

CSCS

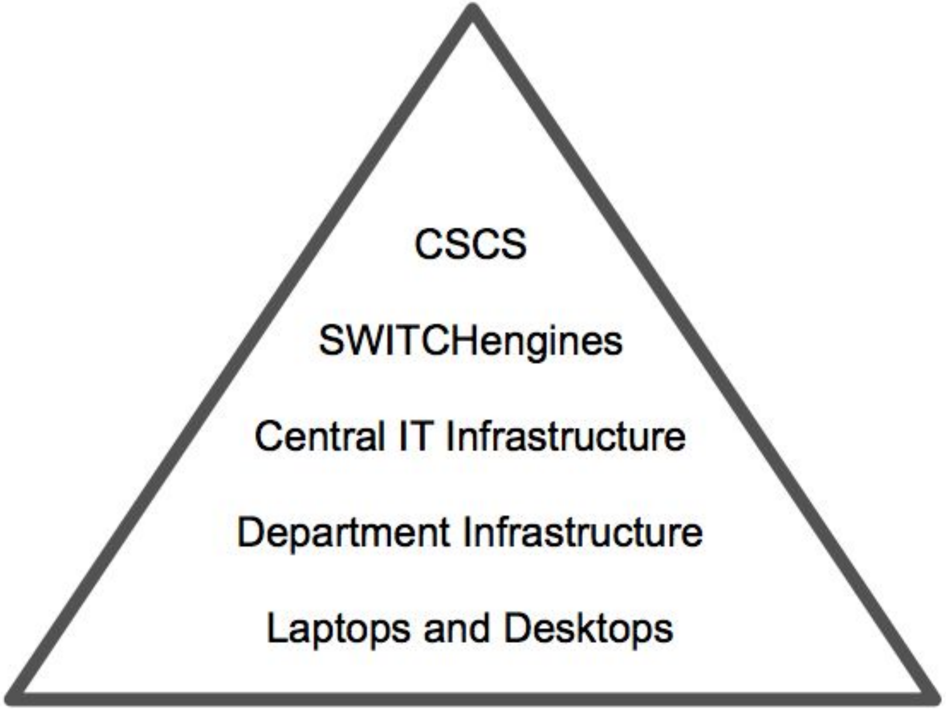
The pyramid of computing infrastructures

Typical 2019 sizes
you can get

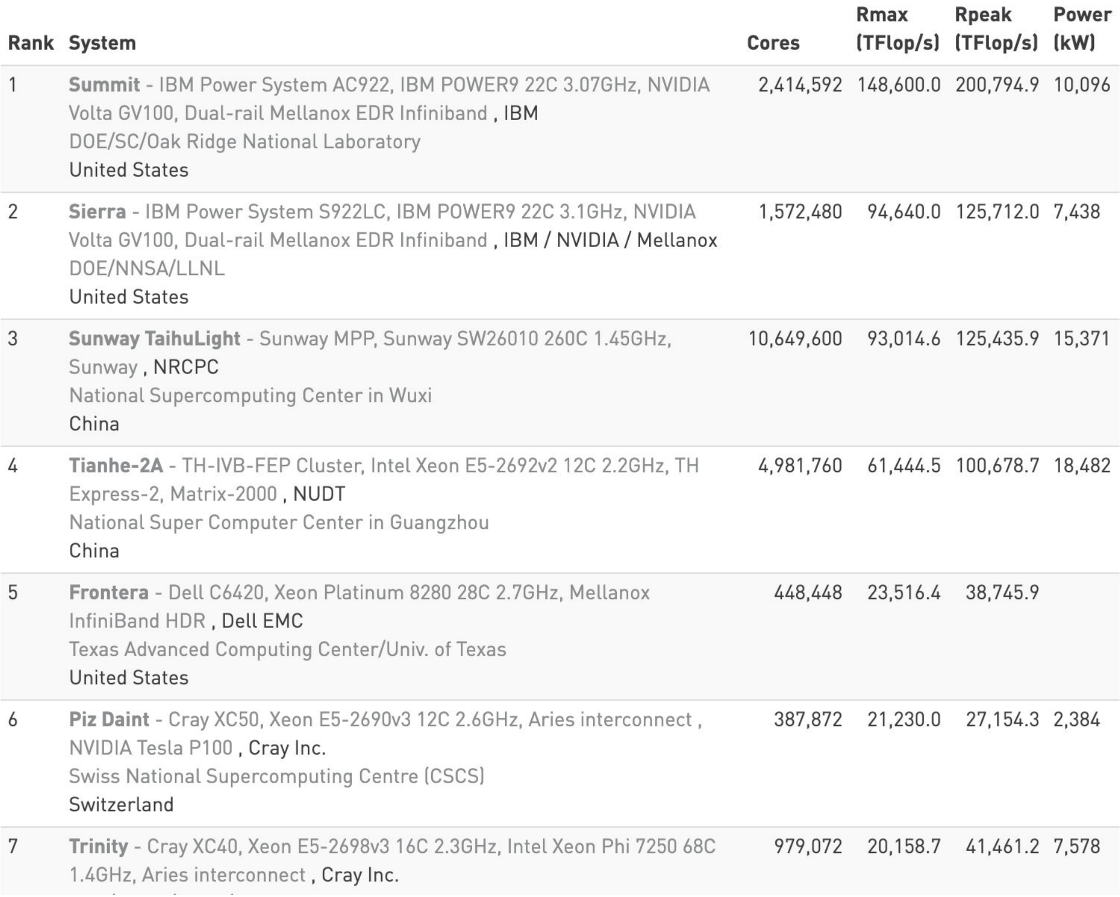


	Storage/PB	Cores
Supercomputers	10^3	10^5
Private and public clouds	10^2	10^4
On premise data center	10^1	10^3
Laptops/desktops/workstations	10^{-4}	10^1

The CH pyramid of academic computing infrastructures



	Storage/PB	Cores
CSCS	10^3	10^5
SWITCHengines	10^1	10^4
Central IT Infrastructure	10^1	10^4
Department Infrastructure	10^1	10^3
Laptops and Desktops	10^{-4}	10^1



Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
2	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
3	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
4	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 , NUDT National Super Computer Center in Guangzhou China	4,981,760	61,444.5	100,678.7	18,482
5	**Frontera** - Dell C6420, Xeon Platinum 8280 28C 2.7GHz, Mellanox InfiniBand HDR , Dell EMC Texas Advanced Computing Center/Univ. of Texas United States	448,448	23,516.4	38,745.9	
6	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , Cray Inc. Swiss National Supercomputing Centre (CSCS) Switzerland	387,872	21,230.0	27,154.3	2,384
7	**Trinity** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , Cray Inc.	979,072	20,158.7	41,461.2	7,578

www.top500.org

Amazon/google/etc data centers are an order of magnitude larger than CSCS

How to use HPC/Supercomputers

For example the UBELIX HPC Cluster at UNIBE

Short demonstration - skip it

Interested people may sign up to an HPC course here: https://www.scits.unibe.ch/training/training_and_workshops/trainings_and_workshops/

Infrastructures for data

Vertical systems

- Laptops/workstations/confined clusters scale vertically
- You can add more RAM and more storage to a certain limit
- Cons: Expensive as non-mass produced components needed, limited to the sortiment of the provider

Horizontal systems

- Just add more commodity hardware (stays cheap)
- Do it in a “cloud” or another data center provider (no hardware procurement or maintenance)
- Cons: More knowledge, expertise and software technology needed

Infrastructures for data

Vertical systems

- Optimal performance / alignment between storage, CPU and network (limited scaling)
- Often a turn-key ready solution designed and tested for a certain scale

Horizontal systems

- Shared parallel file system (spfs) or Hadoop like solutions
- Spfs may require high bandwidth network (typically on HPC/supercomputers, not in the cloud)
- Spfs is data to compute
- “Hadoop” is compute to data

Infrastructures for data

Cloud

- Pay as you go with credit card
- Little support for small users and special needs
- Economy of scale
- Scales well

On premise/HPC/supercomp

- Up front investment
- Limited amount of use cases -> better support
- Scaling beyond planning is difficult

Infrastructures for data - price considerations

Cloud

- Monthly billing on your credit card
- Prices are public (but not always very transparent)
- If the usage of a machine/system is below 50%, for sure something to consider, only pay for usage
- Infrastructure within minutes, minimal bureaucracy (only money)

On premise

- Mostly upfront investment
- Infrastructure used at 80-90% may be cheaper, but doesn’t scale
- If you don’t see electricity, sys admin costs, etc (subsidized by company/organisation) something to consider if usage is around 30%

30-50% of a datacenter cost is electricity for the computers and cooling

Infrastructures for data

Lifetime considerations

- Typical replacement of equipment every four years
- Runs longer but failure risk increases
- Out of warranty
- Electricity consumption per calculation and storage goes down with new equipment

...

- Procurement, installation and commissioning is expensive and time consuming
- Cloud solutions hide all this

Infrastructures for data - main providers

Private cloud providers

- Amazon AWS
- Microsoft Azure
- IBM
- Google
- Oracle
- Alibaba
- ...

CH academic sector

- Your department infrastructure
- Central IT infrastructure (often for free)
- SWITCHengines (cloud provider)
- Swiss Super Computing Center (CSCS) in Lugano

Infrastructures for data - price considerations

Examples

- Amazon, google, Azure
- SWITCHengines (Swiss academic)
- Buy your own computer (center)

Data acquisition (from primary data sources)

The physics

- There are 4 fundamental interactions in nature
  - Gravitation
  - **Electromagnetic**
  - Weak nuclear force
  - Strong nuclear force

Examples

- Keyboard (macroscopic pressure)
- Microphone (sound)
- Camera (light, em wave)
- Temperature ()
- Eye, ear, skin

All about releasing electrical charges

Data acquisition (from primary data sources)

Process

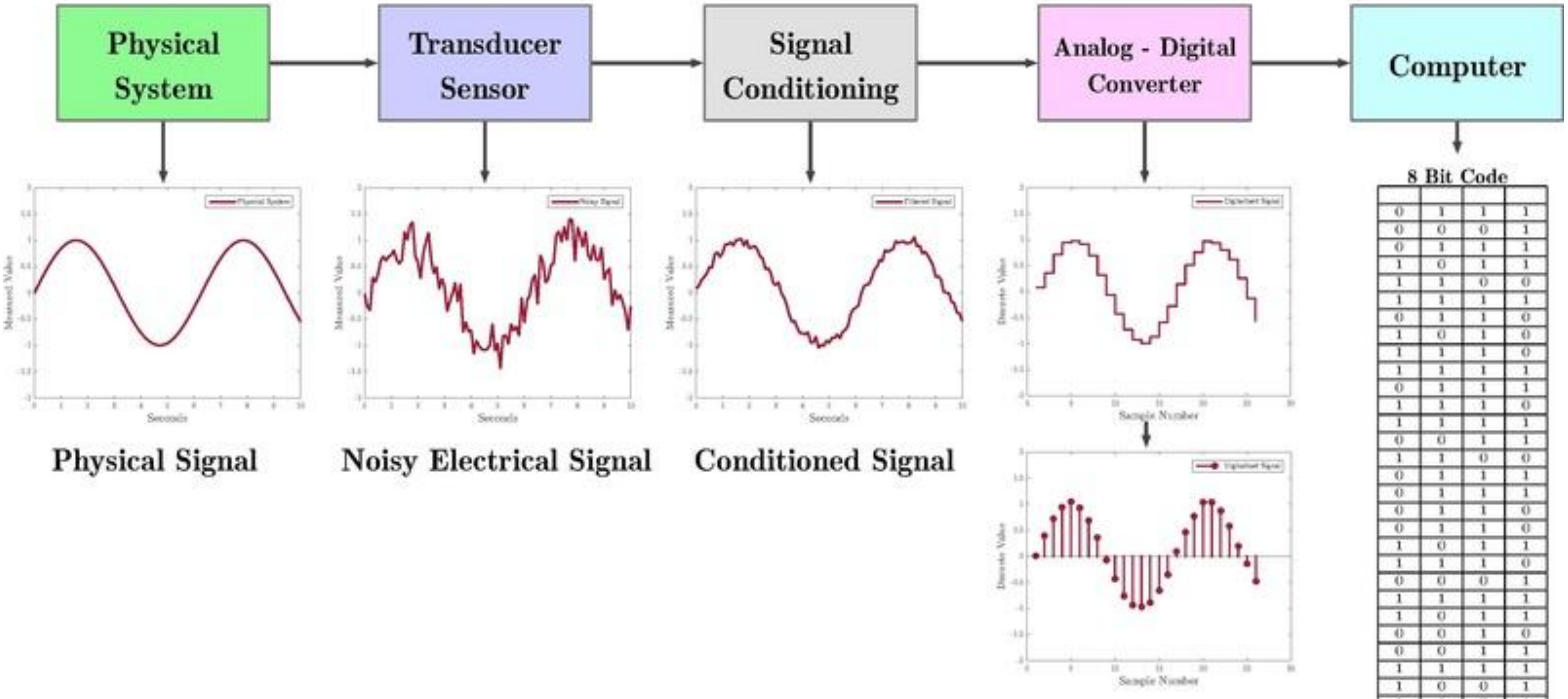
- Physical signal is detected by **sensors** converted into electrical signal
- Electrical signal can be conditioned
- Electrical signal is digitised
- Maybe amplified, transported
- Stored to a physical device

Data acquisition (DAQ) system

Example Mobile Device

- Bike with GPS and internet
    - Time and position
    - Camera and sound
    - Temperature, vibrations …
    - …
- Pushes data to an internet server

Digital Data Acquisition System



Physical Signal

Noisy Electrical Signal

Conditioned Signal

Digitalized Signal

8 Bit Code

0	1	1	1
0	0	0	1
0	1	1	1
1	0	1	1
1	1	0	0
1	1	1	1
0	1	1	0
1	0	1	0
1	1	1	0
1	1	1	1
0	1	1	1
1	1	1	0
1	1	1	1
0	0	1	1
1	1	0	0
0	1	1	1
0	1	1	1
0	1	1	0
0	1	1	0
1	0	1	1
1	1	1	0
0	0	0	1
1	1	1	1
1	0	1	1
0	0	1	0
0	0	1	1
1	1	1	1
1	0	0	1
1	1	0	1

Data acquisition (from secondary data sources)

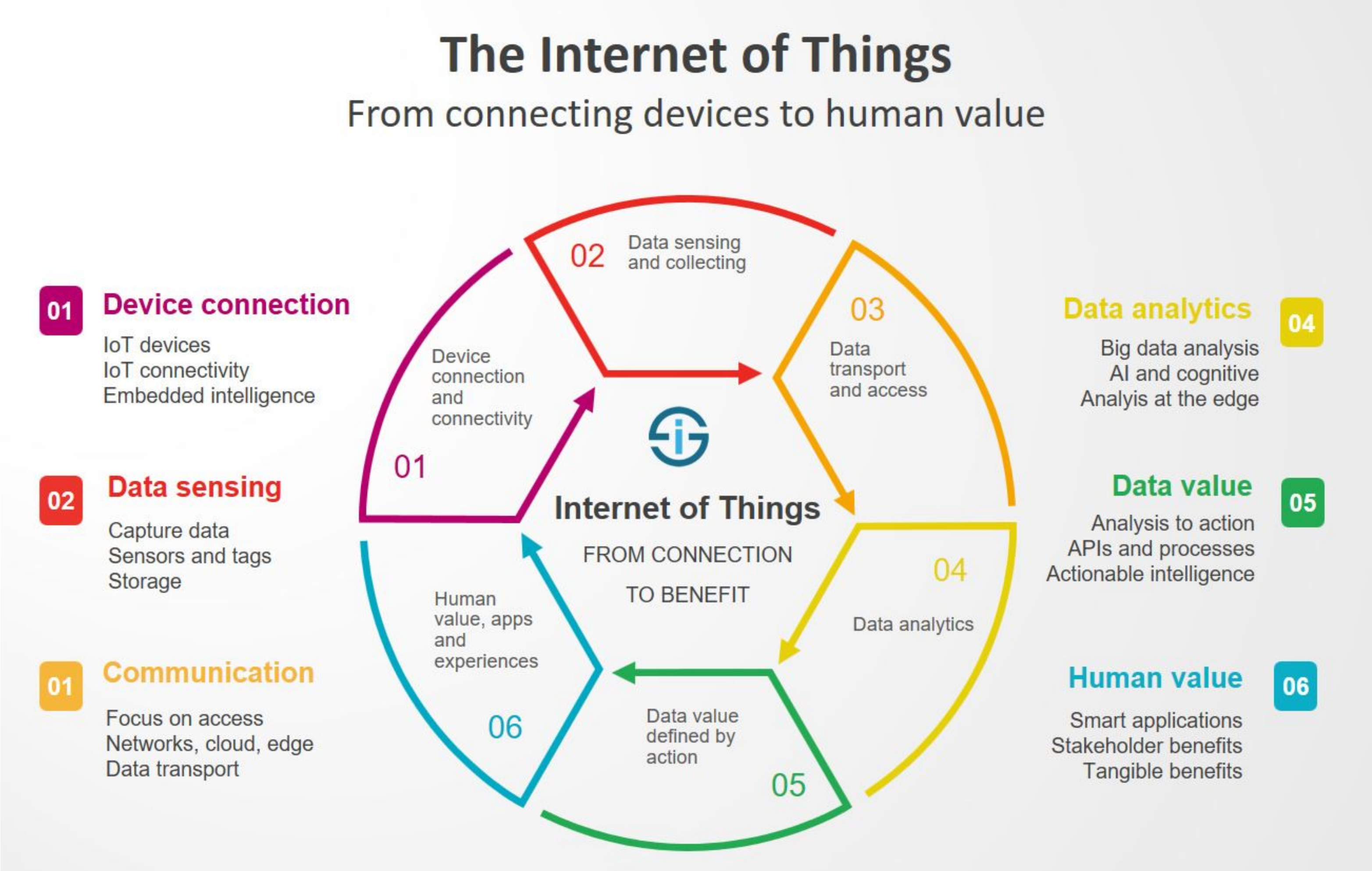
Secondary data sources

- Data storage (wood, stone, paper, tapes, hard disks, solid state disks, human memory, computer memory)
- Online data storage (accessible with wired or wireless internet)

www increases the access to secondary (and primary) data sources dramatically. We will look at data collection from www on day 3

IoT

- Internet of Things
- Increases the access to primary data
- Devices have sensors connected to internet
- Cars, houses, drones, fridges, animals, humans ...)



The Internet of Things

From connecting devices to human value

01 **Device connection**

IoT devices
IoT connectivity
Embedded intelligence

02 **Data sensing**

Capture data
Sensors and tags
Storage

01 **Communication**

Focus on access
Networks, cloud, edge
Data transport

02
Data sensing and collecting
03
Data transport and access
04
Data analytics
05
Data value defined by action
06
Human value, apps and experiences
01
Device connection and connectivity
Internet of Things
FROM CONNECTION
TO BENEFIT

Data analytics **04**

Big data analysis
AI and cognitive
Analyis at the edge

Data value **05**

Analysis to action
APIs and processes
Actionable intelligence

Human value **06**

Smart applications
Stakeholder benefits
Tangible benefits

Much competitive business and research must leverage the new situation - www, IoT, big data, ML, AI, computing power

New skills are required from the data scientists

Continue to work on the notebook till :

15: 30 Break

A nice data science page: www.kaggle.com

Planning: Conceptional Design Report

- Objectives --> Goals, what am I going to achieve (e.g. with PubliBike map street surface condition)

- Data Description --> size (how much data), variables (location, weather, etc.)

- Data Quality --> e.g. right resolution (e.g. meters)

- Methods --> Tools to do it (e.g. budget, Sensors on the bikes, Software)

- Dataflow --> data collection (e.g. bike sensors) > database / server somewhere > to dataframe (e.g. via python) > manipulations > visualization

- Data Model --> conceptual (what will I do), logical (what columns will be there in my dataframe / multiple dataframes), physical (model of the hardware necessary)

- Initial Numbers (Pilot-Data)

Data Flow Analysis
Data Modelling
(Brief)

Analyse the dataflow

Why

- Know your data, needed for good inference / data science
- Data flow may have impact on the data quality
- Identify risks and incompatibilities in the DAQ and computing system
- Make appropriate data model and structure

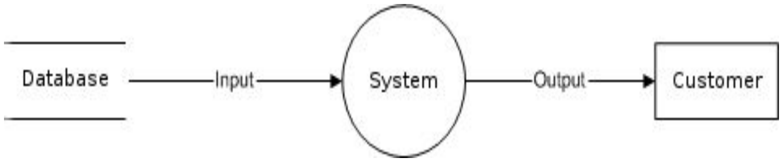
Data flow

- From input source via
  - Network
  - Storage
  - Preprocessing
  - Analysis
  - Output storage
- Presentation (publication, talk)
- Lobbying / Convincing
- Decision

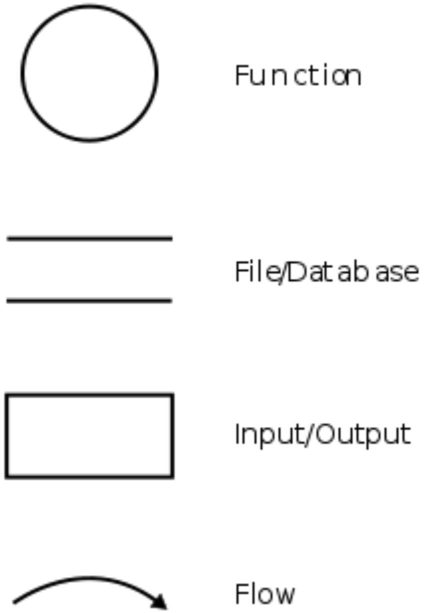
Data flow diagrams

How

- Some diagram standardisation suggestions exist
- However, important is that the diagram is serving its purpose



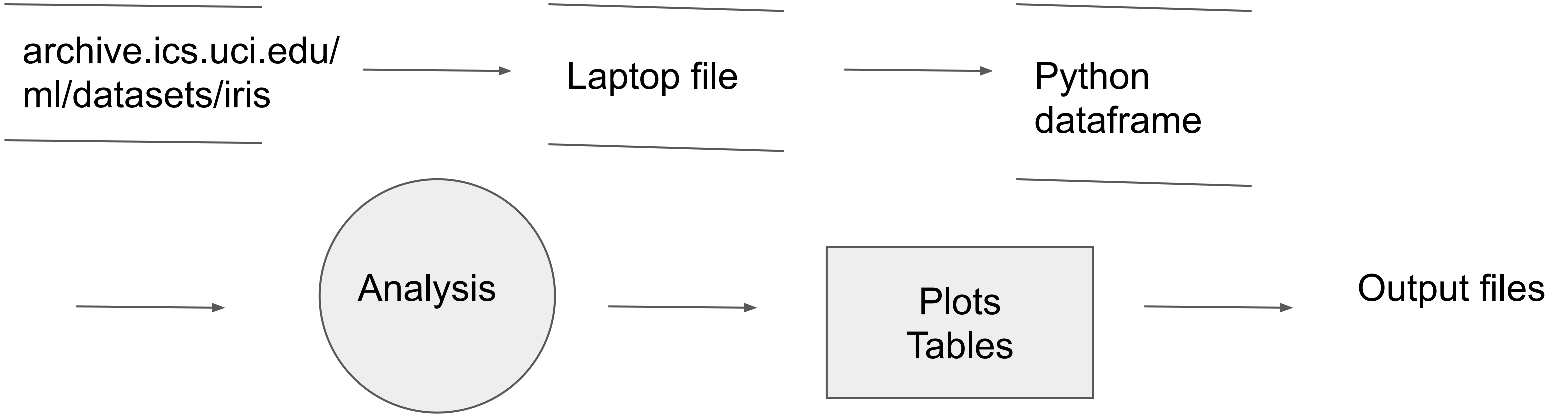
Symbol suggestions



Data flow diagrams

Our Iris data set case

-



Analyse the data flow - physical data model

Concerns

- Data set size at each step
- Storage types, backup
- Velocity (time between steps)
- Memory considerations
- Lifetime
- Formats
- Analysis tools

...

- Analysis time
- CPU capacity
- Bandwidth capacity
- ...

All gives requirements to your data infrastructure

Analyse the data flow - logical data model

Concerns

- Data set organisations (multiple files)
- Metadata organisation
- File organisation
- Data organisation within files
- Data organisation within databases (tables, columns ...) see day 3

...

- Data structures in analysis tool
    - Pandas dataframes
    - R dataframes
    - Some classes
    - etc

Analyse the data flow - conceptual data model

High level

- Terms/entities and relationships between them
- Not the actual data model design (logical)
- Not the physical data model

Order

- Normally
    - Conceptual
    - Logical
    - Physical
- However, all levels may restrict each other

Team work - 7 Teams (30 Minutes)

You can do this in a notebook, on paper in the room or on some slides

- Formulate a data science challenge or take this one:
  - Mapping Bern road surface conditions with Publibikes
- Sketch/tabulate the conceptual data model
- Sketch the data flow diagram
- Sketch/tabulate the logical data model
- Sketch/tabulate the physical data model

Presentations at 16:00

Team work - 7 Teams (30 Minutes)

- Group 1
- Group 2
- Group

Summary (16:45 - 17:00)

...   .

- https://jakevdp.github.io/PythonDataScienceHandbook/

CAS Applied Data Science - Module 1

Data Acquisition and Management

End of day 1 - tomorrow data visualisation and databases

Bern, 2020-08-19

Data science platform(s)

Anaconda

- Free open source
- 6M users in 2018
- Comes with GUI, Python, R RStudio, Jupyter …
- Linux, MacOS, Windows

…

- https://en.wikipedia.org/wiki/Anaconda_(Python_distribution)
-

Install Anaconda, launch Jupyter

Steps

- Go to Download on the Anaconda webpage
- Download and install Anaconda with Python 3 for your OS
- Install R from the command line: conda install -c r r-irkernel

...

- Install more packages ...
- Start the Graphical User Interface (GUI) : anaconda-navigator
- Launch Jupyter Lab and open the first notebook on Ilias

Create your GitHub CAS repo

If you don’t have one, create your GitHub account

- If you don’t have one, create your GitHub account
- Create a CAS repository
- Upload your CAS material there
- You will learn more about GitHub and git in Module 4

Datasets to work with

Choose your own or take

- Bring your own from work or research (on laptop)
- Search internet
- If your set is too big … tell me

a public dataset

- https://archive.ics.uci.edu/ml/datasets.html
- https://www.kaggle.com/datasets

You may (ideally) work with one dataset through the whole CAS or use severals according to what is to be done.