Royal Netherlands
Meteorological Institute
Ministry of Infrastructure and Water Management

FROM GATHERING TO ANALYSIS: DATA PIPELINE EXAMPLES AT KNMI

Andrea Pagani

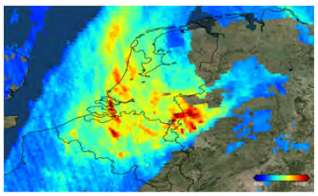
10th October 2019

EO3S – NSO, The Hague

1

Context: Big Data Age

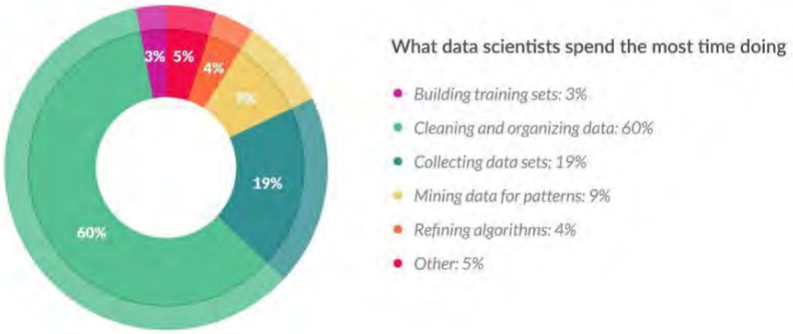
- More data
- Higher resolution
- More sources
- Unconventional
- **New opportunities**



2

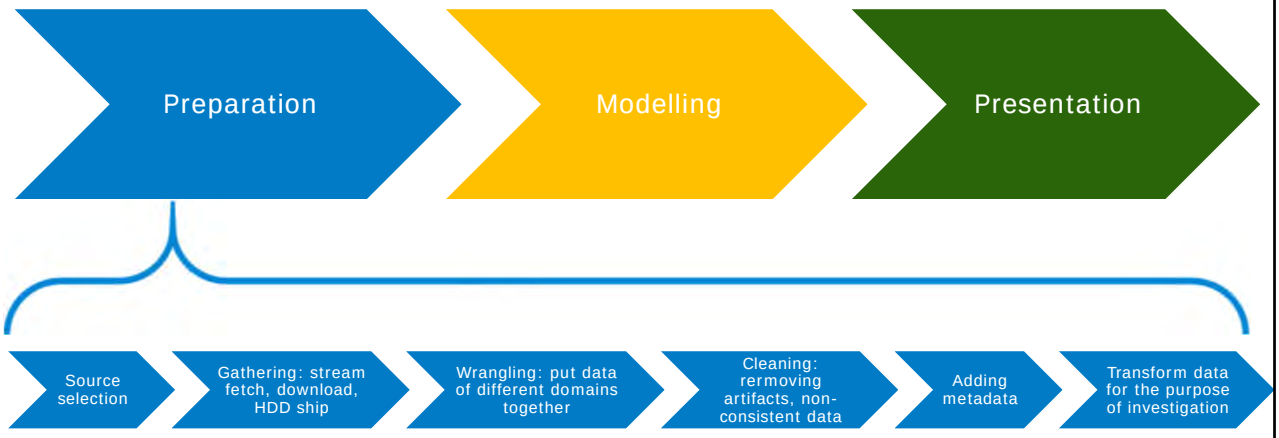
Data Science work: the real story

Data preparation *accounts for about 80% of the work of data scientists*



Source: Forbes -- Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says

What is preparation?



My opinion

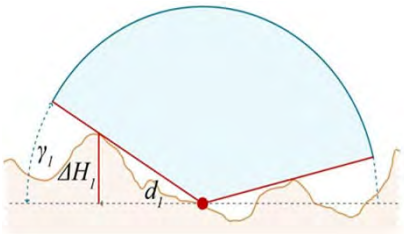
- Laborious and non-exciting work
- BUT extremely important to have it done WELL
- It takes already some decision steps that might influence the analysis afterwards
  - E.g., resolution of the spatial transformation
  - E.g., time approximation for wrangling phenomena
  - E.g., good understanding of cleansing policies
- Might trigger further questions to the experts of the data/phenomenon

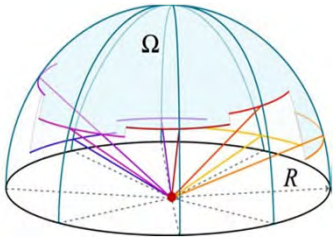
Data pipeline example: **Sky View Factor from digital elevation model (AHN2)**

Goal: create high value dataset - the sky view factor at high resolution (1m grid) for the entire Netherlands

What's SVF:

"The sky view factor (SVF) denotes the ratio between radiation received by a planar surface and that from the entire hemispheric radiating environment and is calculated as the fraction of sky visible from the ground up..." – Source: wikipedia





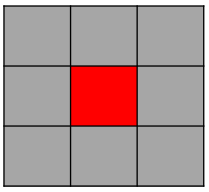
Data pipeline example: Sky View Factor from AHN2

Data sources info:

- 1.5TB point cloud data of height of objects for the Netherlands (AHN2) in raw LAZ format
- Data gathering via Dutch e-Science Center
- 40000+ LAZ files (i.e., tiles) with geo metadata in the filename
- Tile re-arranging for corner cases
- Interpretation of NAs (water bodies)

Steps:

- Gridding to 1m resolution (very memory intensive)
- Keep tracking of tiles geo locations and store in the grid file format
- Logic to merge/subset tiles done ad hoc for tile boundaries
- Compute SVF



7

Data pipeline example: Sky View Factor from AHN2

Computation part:

- Computation to be performed in R (nice library for the purpose: horizon)
- High memory requirement to process multiple tiles
- Test on high parallel machine (24CPUs/128GB ram)
- Distributed on Amazon AWS EC2 (80CPUs)

Opportunity offered: AWS Cloud Credits for Research

8

Data pipeline example: **from traffic pictures to fog detection**

Goal: train a deep learning model able to detect fog in pictures

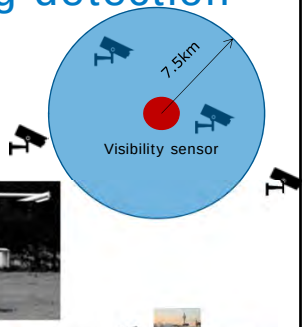
Source: stream of real-time pictures from traffic cameras

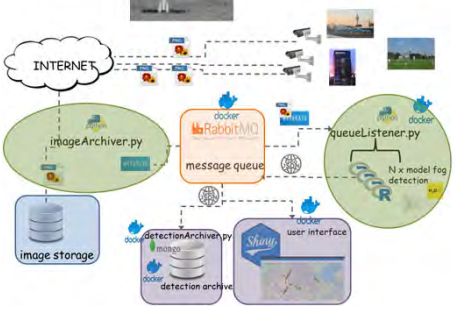
Preparation steps:

- Fetch pictures
- Addition of metadata: time, location, filepath, ID
- Store metadata in DB for easier access
- Setup a message queue to streamline interaction between components
- Label the pictures with nearby KNMI visibility observations
- Resize images to use less computation power

Data pipeline example: from pictures to fog detection

- Labeling:
  - Filter cameras in range of KNMI visibility sensor
  - Fetch KNMI 10-min observation database
  - Merge visibility value with picture based on location and timestamp
- Once deep learning model is trained the pipeline goes on





Data pipeline example: **urban air quality**

- Goal: create real time (hourly) maps of urban air quality
- Based on open data as much as possible
- Several ingredients needed:
  - Observations (official + low cost sensors)
  - Dispersion model
  - Proxy data for emission

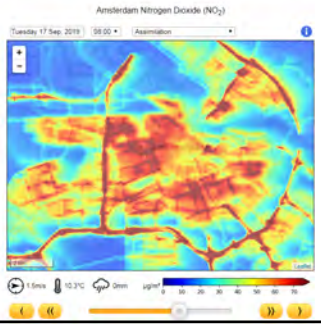
	Variable	Source
Observations	Hourly in-situ measurements NO_2	Official + crowd-sourced stations
	Background concentrations NO_2 and ozone	CAMS
Meteo	Surface meteorology	Meteo stations
	Upper air meteorology	Radiosonde
Emission proxies	Road network and road classification	Open Street Map
	Traffic flow	Dutch traffic datawarehouse (NDW)
	Population density	National statistics center (CBS)

Data pipeline example: urban air quality

- Each source needs special treatment for access, clean
- Data are combined spatially by re-projection on a local grid

Raw data gathering → Pre-processing to get relevant values → Data cleansing (e.g., interpolation for missing data, outliers removal) → Store in database → Wrangling → Modelling → Output in web and FTP endpoint

- Currently working in real time for Amsterdam, Madrid and Barcelona



Conclusion

- Data preparation is taking a considerable part of data science work
- It is a non-exciting but very important step
- It is often underestimated:
  - For time and effort required
  - For the importance to have a good working pipeline
- Every project has different preparation requirements
- Flexible computation environment needed
- 3 different examples with challenges and solutions presented

- Bonus: For geo-spatial data access tools such as data cubes using WCS and WPS services help