SDSC

Data science for Connected Health Devices

Merck - SDSC

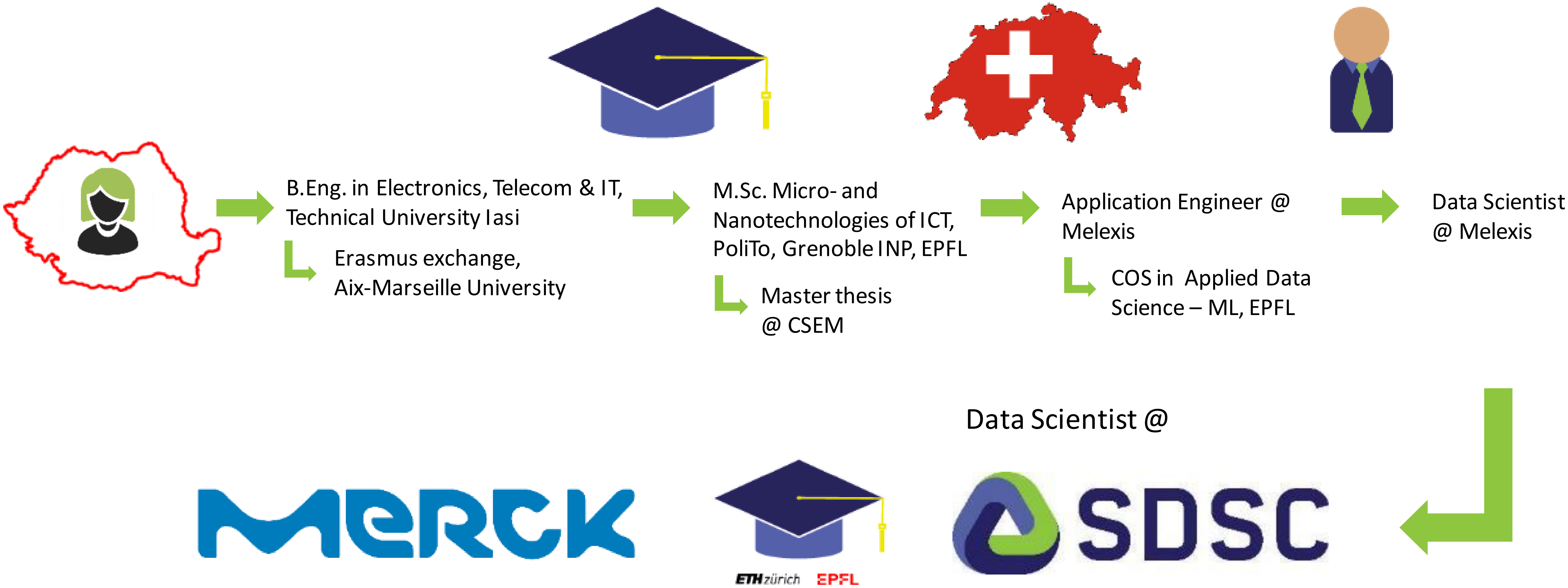
Why this presentation

“Where do you see yourself in 5 years?”

- To understand better
  - What would you do in the near future
  - What is data science in industry like
  - Why is Merck an attractive company for data science
  - What kind of problems you may need to solve

- To ask us questions

Who am I?



B.Eng. in Electronics, Telecom & IT, Technical University Iasi

Erasmus exchange, Aix-Marseille University

M.Sc. Micro- and Nanotechnologies of ICT, PoliTo, Grenoble INP, EPFL

Master thesis @ CSEM

Application Engineer @ Melexis

COS in Applied Data Science – ML, EPFL

Data Scientist @ Melexis

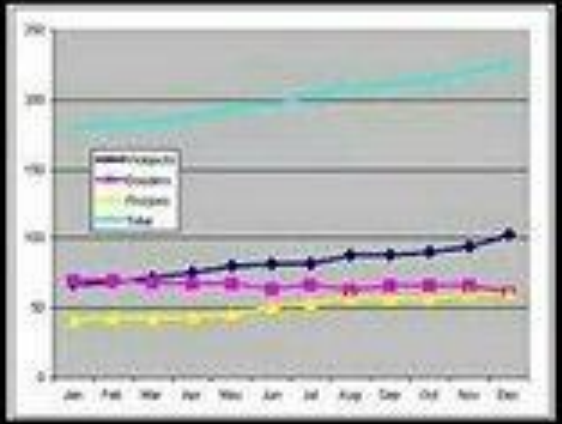
Data Scientist @

Data Scientist

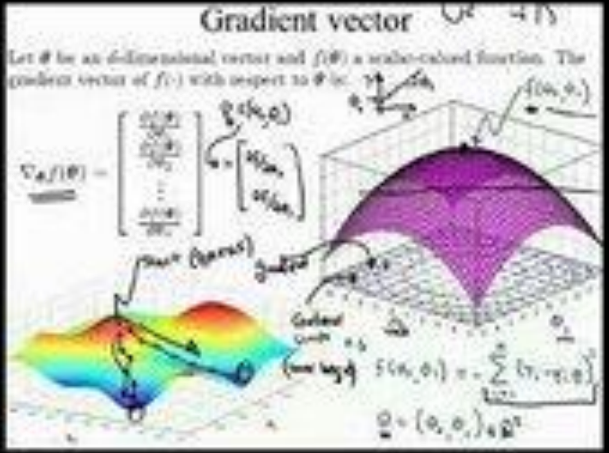
What my friends think I do

What my mom thinks I do

What society thinks I do



What my boss thinks I do

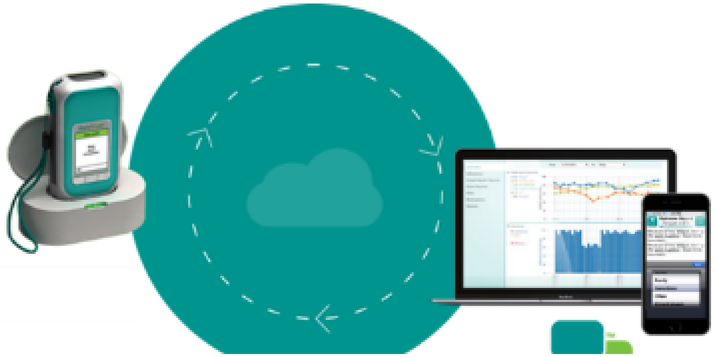


What I think I do

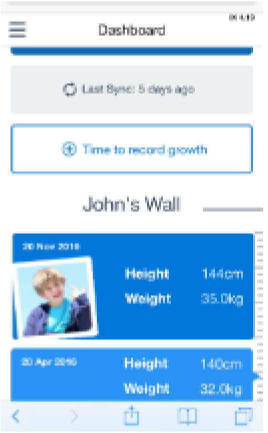
What I actually do

No, really…

Merck - Setting the Context



Connecting patients to doctors & nurses



Dashboard

Last Sync: 5 days ago

Time to record growth

John's Wall

20 Nov 2016		
	Height	144cm
	Weight	35.0kg
20 Apr 2016	Height	140cm
	Weight	32.0kg

12 years of experience

500 000 devices produced

20 000 connected patients

50 countries

Data Science in the Merck – SDSC Collaboration

The “Why”

Context → Questions

Can we cluster patients in categories ?

Can we identify improvements in the use of our services ?

What are the key parameters influencing the behaviour of a patient ?

How to visualize all this in a reproducible way?

How to extract insights from all this data?

What are the meaningful insights?

The “How” – Data Process

python™

jupyter

1

Data Gathering

- via MySQL queries into python
- > 100 tables
- Find your way and merge relevant tables

2

Cleaning

- only patients with at least 1 calculated adherence point
- only patients who started after 2014 and logged no data in 2019
- remove null-entries

3

Feature Engineering

- Patient related data
- Adherence related data
- Device settings data, aggregations, majority voting
- Creating new (relevant) features

4

Output dataset

- Thousands of patients globally
- One line per patient
  - Features
  - Target

Defining the key parameters

Adherence

Defining the key parameters

Adherence

% of intake dose vs. prescribed dose

Defining the key parameters

Duration of use

SDSC

Defining the key parameters

Duration of use

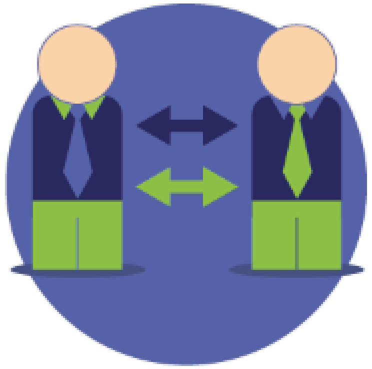
for how long is a patient following the treatment

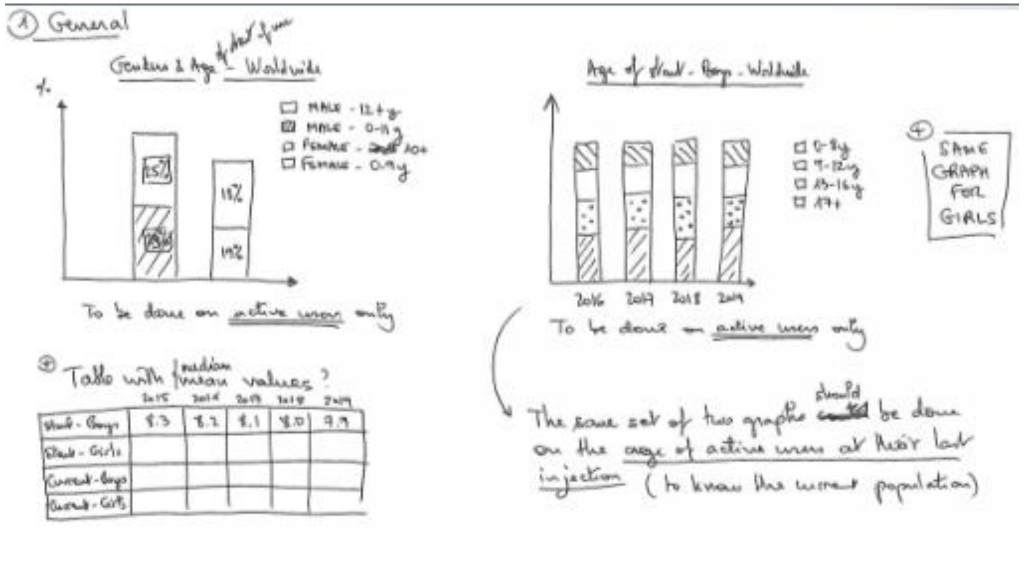
SDSC

The “What”

Regular Cross-communication

- Business to define their questions and/or demanded visualizations
- Iterative process





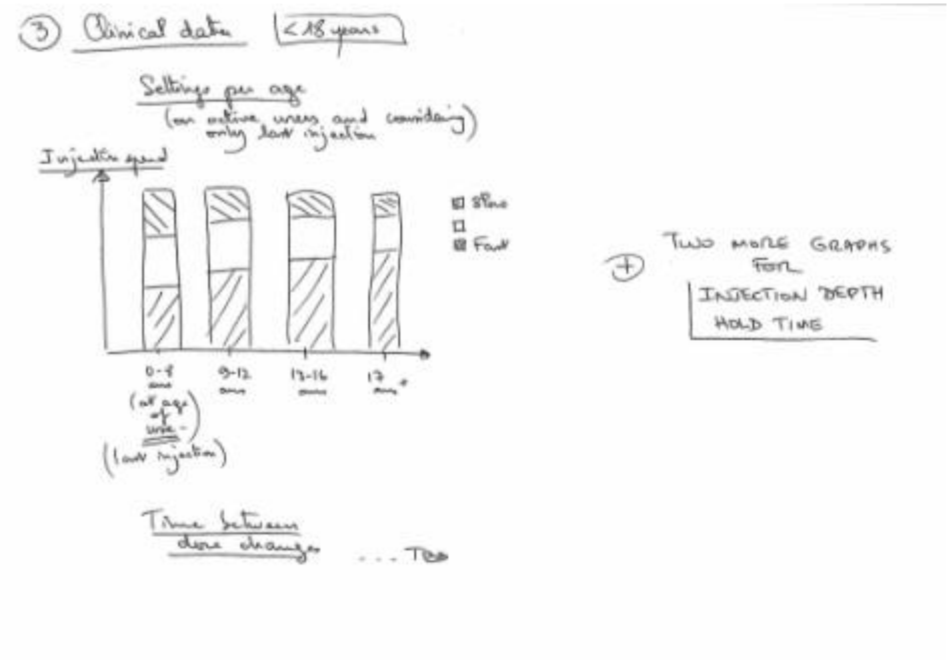
To be done on active users only

⊕ Table with median/mean values ?

	2015	2016	2017	2018	2019
Start - Boys	8.3	8.2	8.1	8.0	7.9
Start - Girls					
Current - Boys					
Current - Girls					

To be done on active users only

The same set of two graphs ~~could~~ should be done on the age of active users at their last injection (to know the current population)


③ Clinical data
< 18 years
Settings per age
(on active users and considering only last injection)
Injection speed
▨ Slow
□
▨ Fast
0-8
9-12
13-16
17+
(at age of use - (last injection)

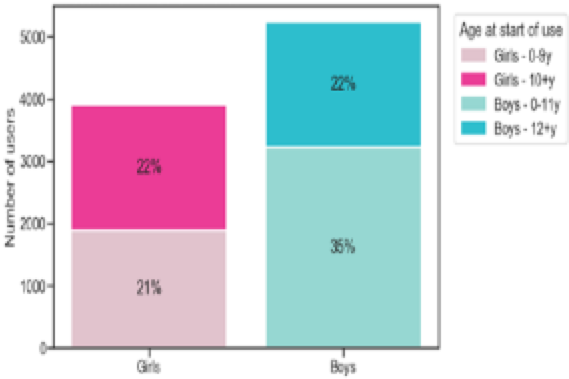
⊕ TWO MORE GRAPHS FOR INJECTION DEPTH HOLD TIME

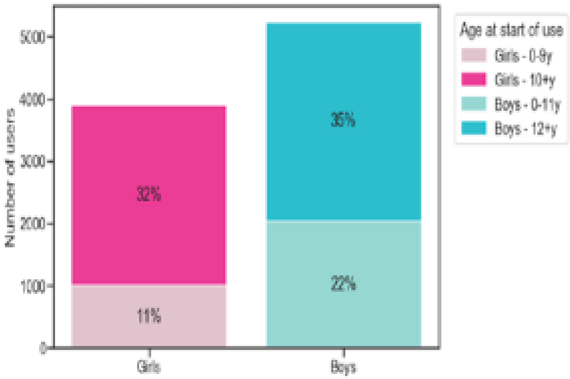
Time between dose changes ... TBD

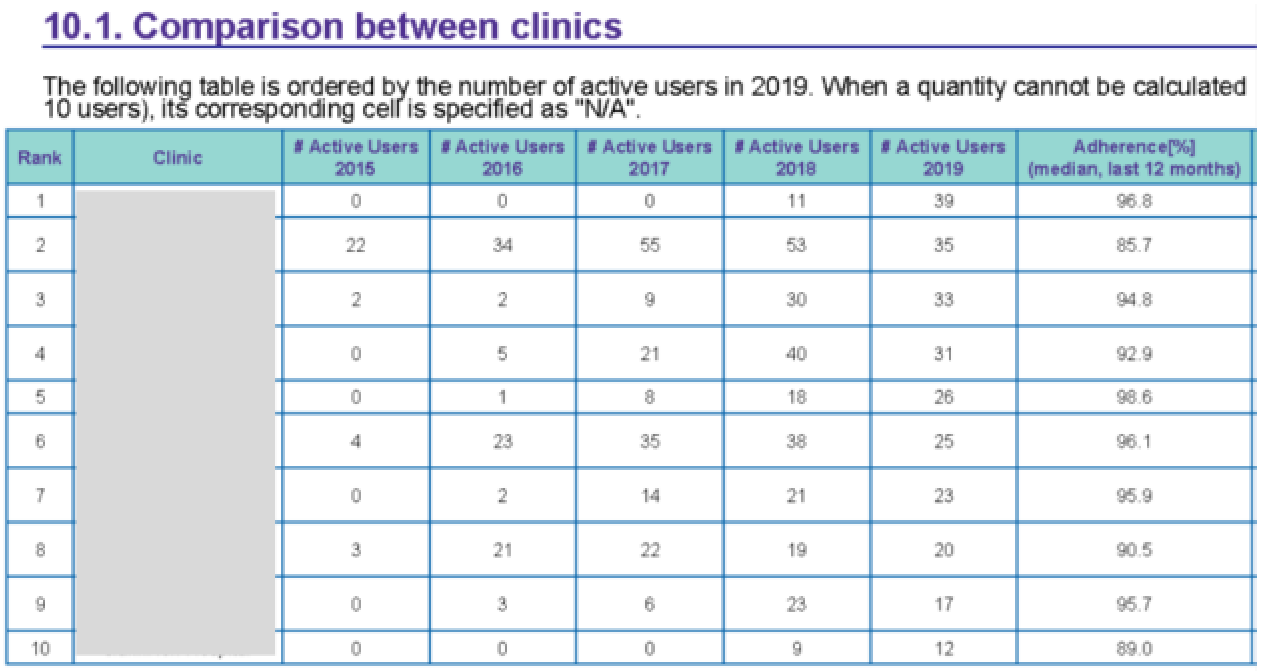
Reproducible reports

- Automatic generation of reports
  - Worldwide
  - Per country

```
generate_graphs.ipynb
generate_reports.ipynb
```







10.1. Comparison between clinics

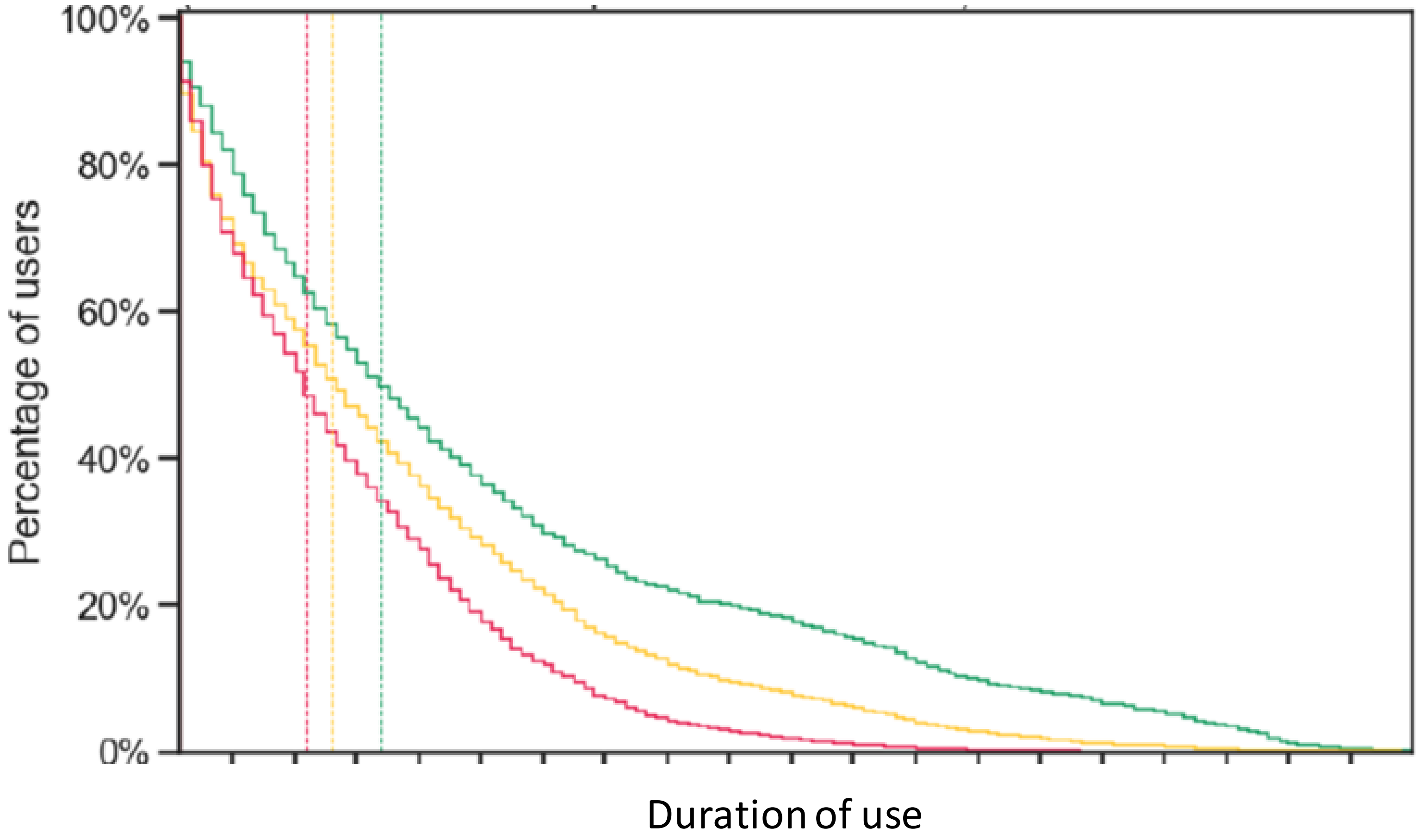
The following table is ordered by the number of active users in 2019. When a quantity cannot be calculated 10 users), its corresponding cell is specified as "N/A".

Rank	Clinic	# Active Users 2015	# Active Users 2016	# Active Users 2017	# Active Users 2018	# Active Users 2019	Adherence[%] (median, last 12 months)
1	[redacted]	0	0	0	11	39	96.8
2	[redacted]	22	34	55	53	35	85.7
3	[redacted]	2	2	9	30	33	94.8
4	[redacted]	0	5	21	40	31	92.9
5	[redacted]	0	1	8	18	26	98.6
6	[redacted]	4	23	35	38	25	96.1
7	[redacted]	0	2	14	21	23	95.9
8	[redacted]	3	21	22	19	20	90.5
9	[redacted]	0	3	6	23	17	95.7
10	[redacted]	0	0	0	9	12	89.0

SDSC

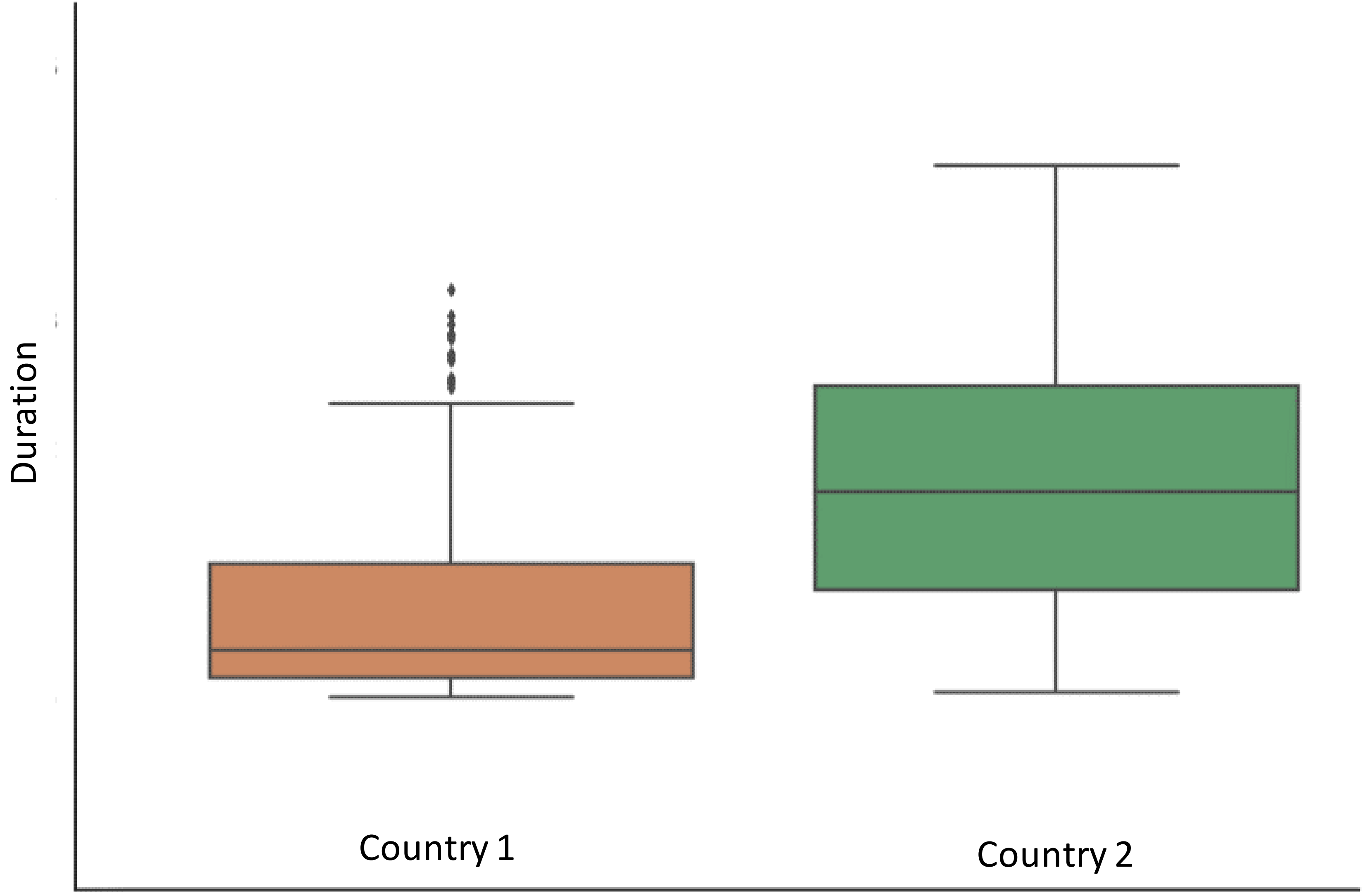
Duration of use of a device, per age group

- What are meaningful insights?



- Feature creation
- Processing
- Reverse cumulative distribution
- Churn analysis & prediction

Duration of use of a device, per country

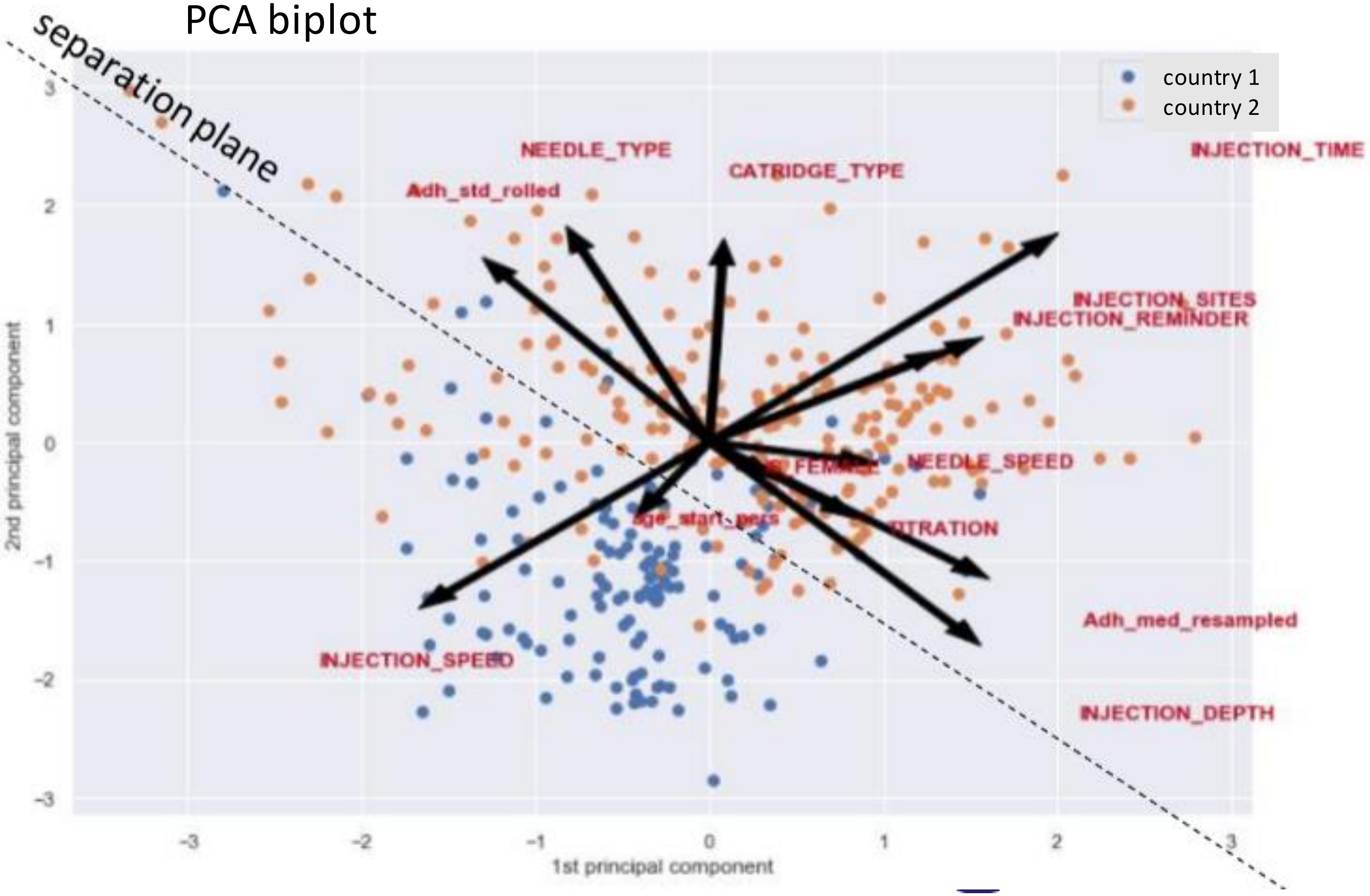


SDSC

Interpretability

- What are the key parameters influencing the outcome?

PCA biplot

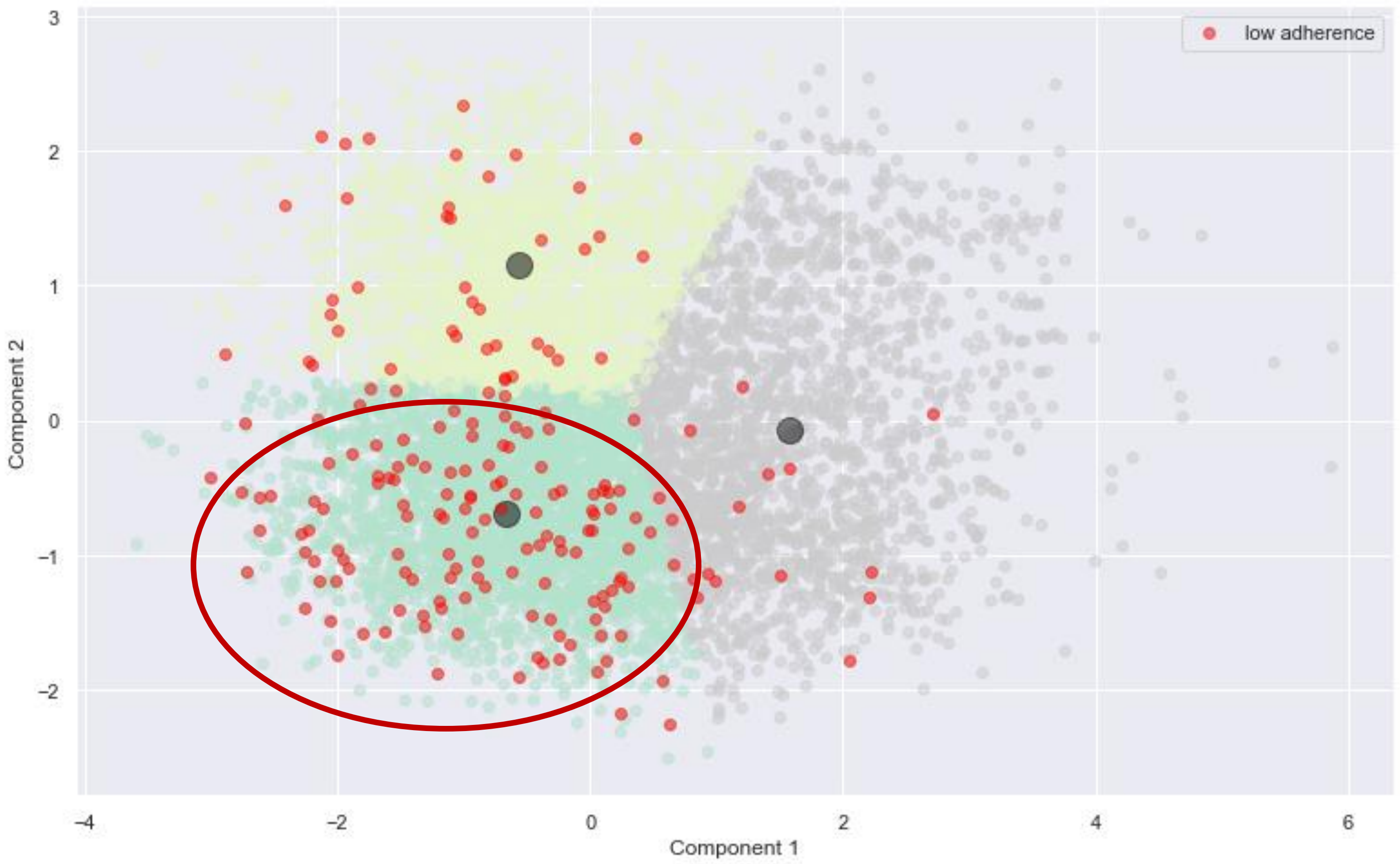


Main **discriminators:**

- **Injection speed**
- **Injection time**
- **Injection sites**

Clustering for patient segmentation

- Can we cluster patients in different categories?



70% of patients with low adherence are in this cluster

SDSC

- K-Means clustering
- Dimensionality reduction

Conclusions

Pharma is a very rigorous industry, and **interpretability is key**

“Traditional” machine-learning algorithms are therefore chosen more frequently than Deep Learning

The feeling of seeing your work reaching so many different people is just great!

Thank you!

Swiss Data Science (@SDSCdatascience)

Questions?