ETH-RAT

SDSC

SWISS DATA SCIENCE CENTER

A joint center between EPFL and ETH Zürich

Olivier Verscheure

Data Professionals Day, Zurich, May 8 2017

http://www.datascience.ch

About me

Academia

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

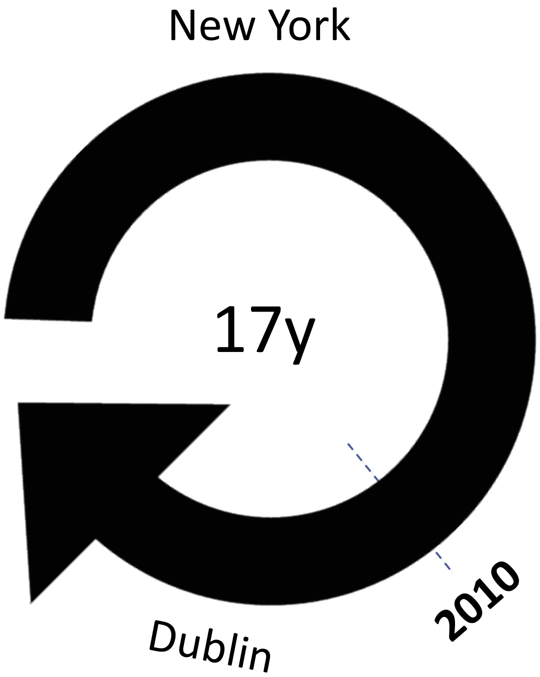
1999

2016

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

ETH zürich

IBM Research



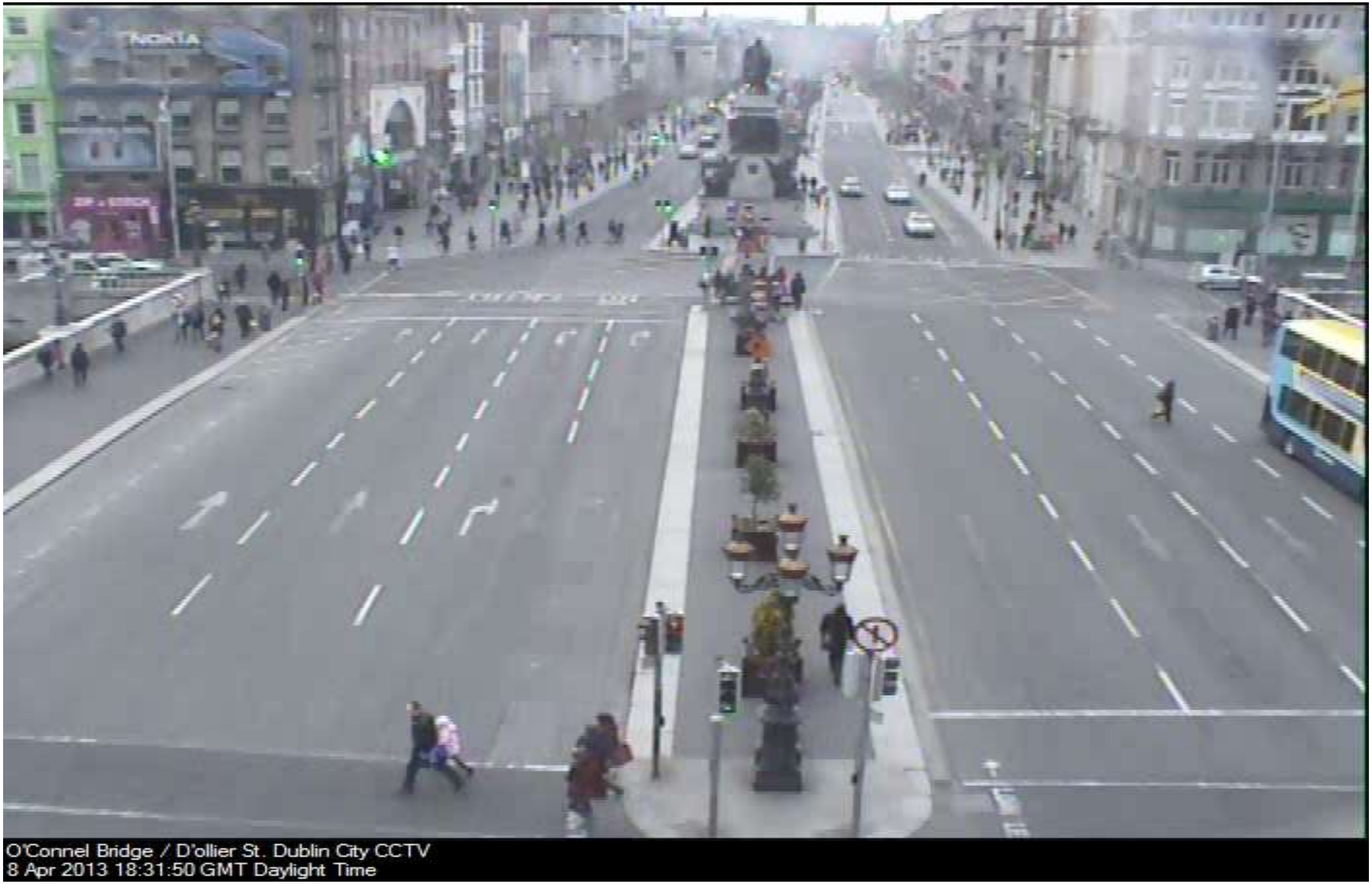
New York

17y

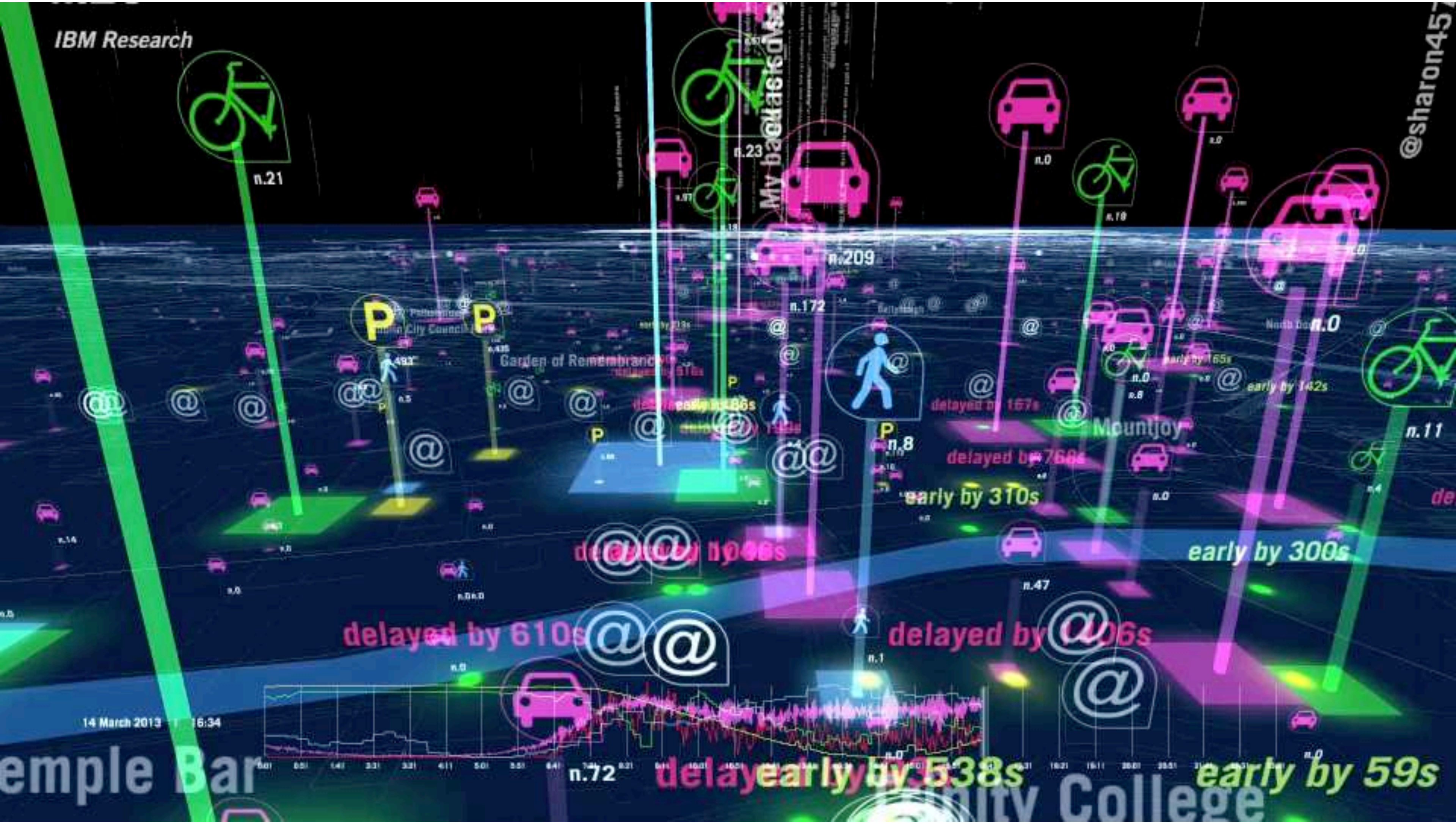
2010

Dublin

What Do You See?


NOKIA
O'Connel Bridge / D'ollier St. Dublin City CCTV
8 Apr 2013 18:31:50 GMT Daylight Time

Dublin City Data Hub



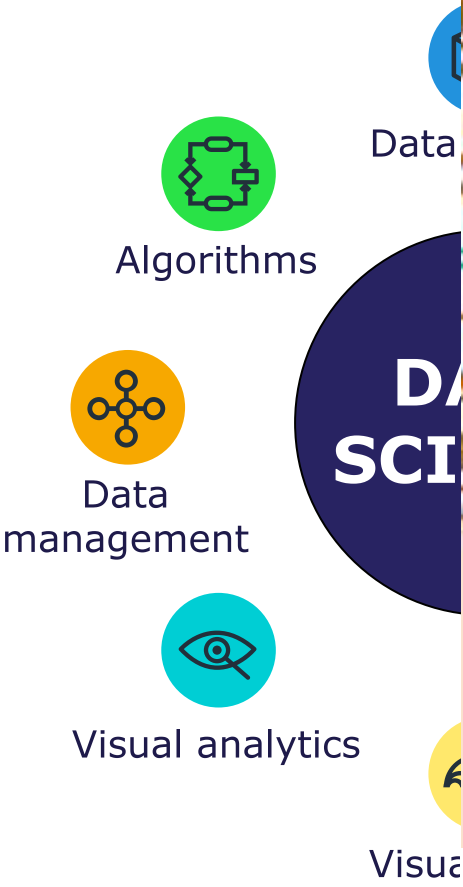
Big Data, Bad Data

WHEEP
WHEEP

"YEP... GOT MY CELLPHONE, MY PAGER, MY INTERNET LINK, MY WIRELESS FAX, AND THANKS TO THIS NIFTY SATELLITE NAVIGATING SYSTEM, I KNOW PRECISELY WHERE I AM AT ALL TIMES!"

BY LOWE FOR THE SUN-SENTINEL, FLOR

A fragmented ecosystem



What is the hyperplane that best separates two classes of points in multidimensional space?

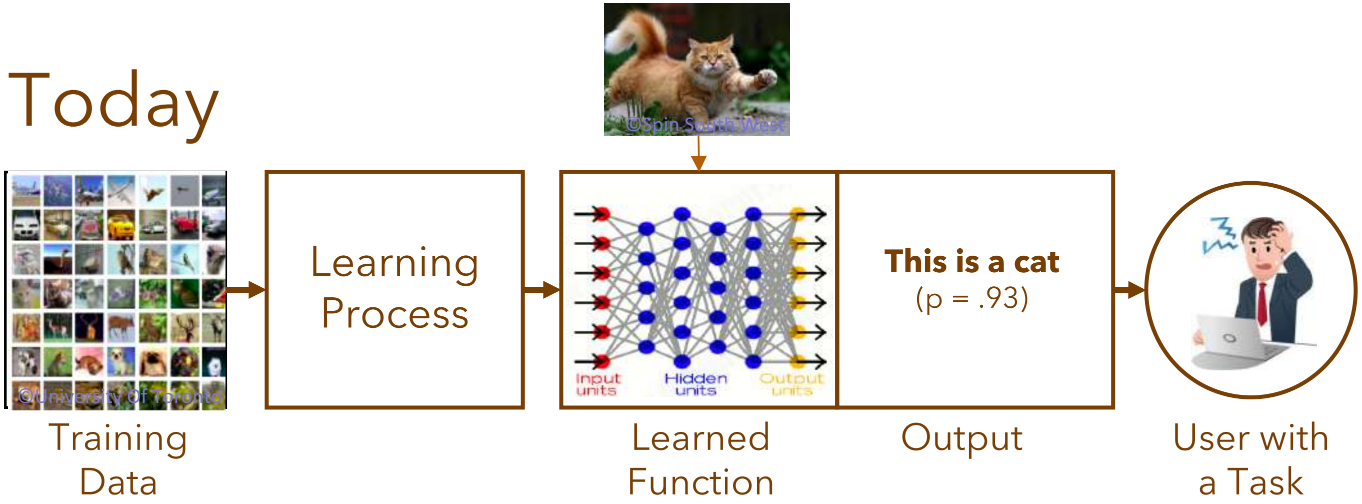
GAP

How can I best match the right drug with the right dosage to the right patient at the right time?

SDSC

Explainable AI – What Are We Trying To Do?

Today



©Spin South West

©University Of Toronto

Training Data

Learning Process

Learned Function

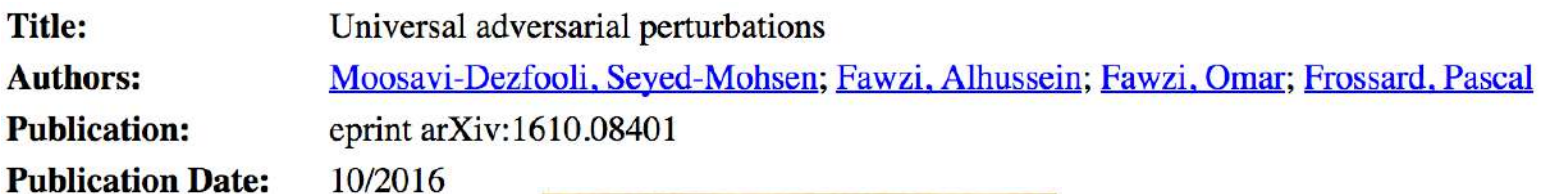
This is a cat
(p = .93)

Output

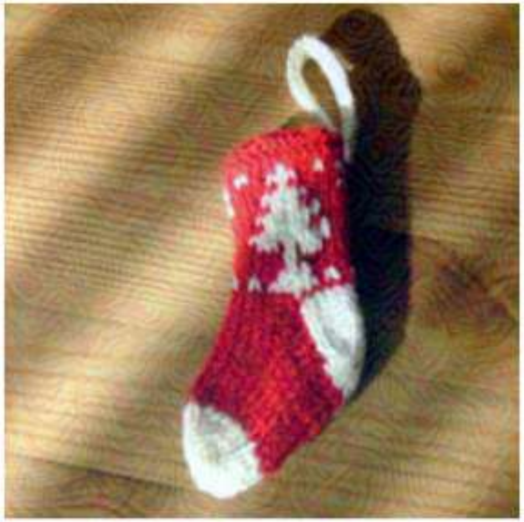
User with a Task

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Fooling deep neural net classifiers



Title:	Universal adversarial perturbations
Authors:	Moosavi-Dezfooli, Seyed-Mohsen; Fawzi, Alhussein; Fawzi, Omar; Frossard, Pascal
Publication:	eprint arXiv:1610.08401
Publication Date:	10/2016

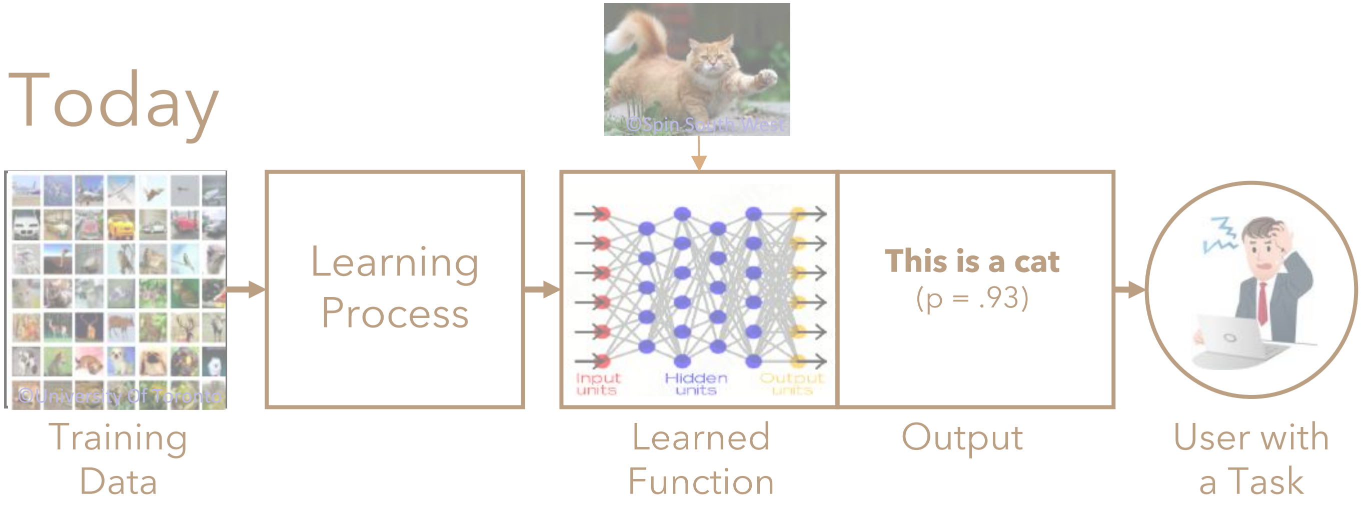


This is not a woolen sock

- It’s an Indian elephant!
- At least after adding a universal noise to the image
- Deep learning models do not mimic brain activity

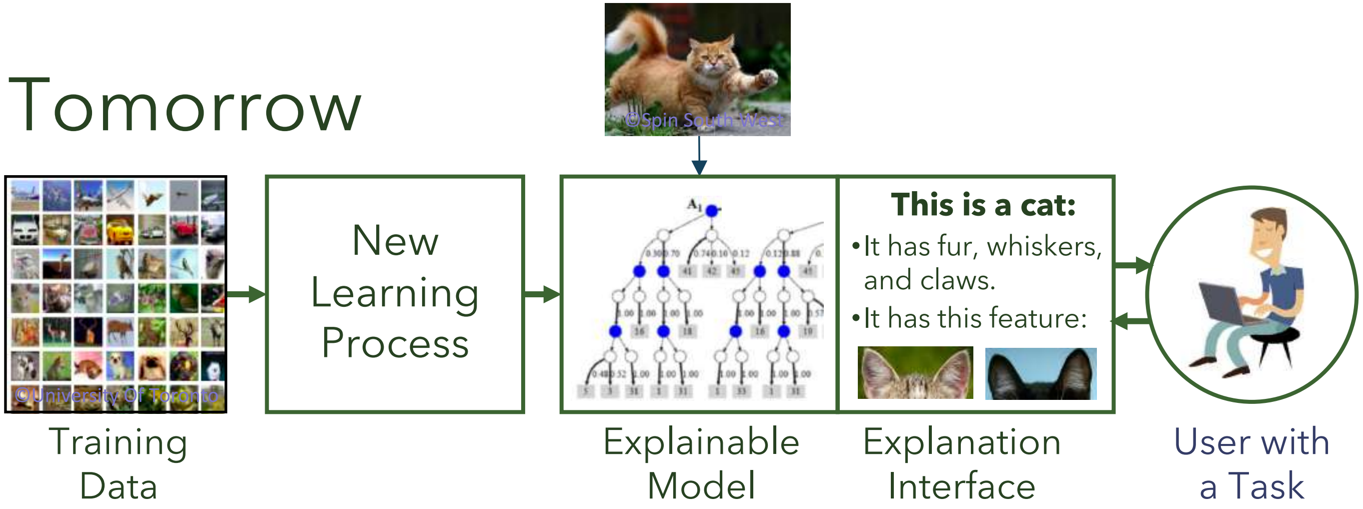
Explainable AI – What Are We Trying To Do?

Today



- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Tomorrow



This is a cat:

- It has fur, whiskers, and claws.
- It has this feature:

- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

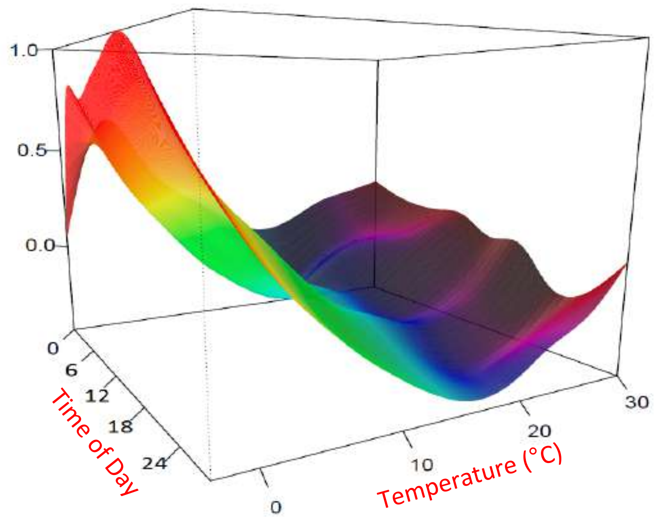
Anecdotal digression

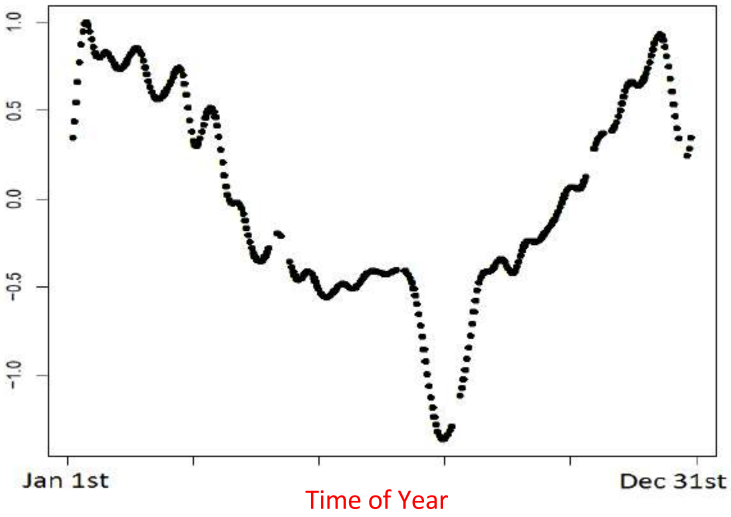
- Forecasting demand in electricity (France)

$$
\begin{aligned}
y_k \;=\; & \beta^{\text{Intercept}} + \overset{\text{Trend}}{f^{\text{Trend}}(k)} + \overset{\text{Lag load}}{f^{\text{LagLoad}}(y_{k-48})} + \overset{\text{Day-type specific daily pattern}}{\sum_{l=1}^{6} \mathbf{1}(x_k^{\text{DayType}} = l)(\beta_l^{\text{DayType}} + f_l^{\text{TimeOfDay}}(x_k))} \\
+ \; & f^{\text{CloudCover}}(x_k) + f^{\text{Temperature/TimeOfDay}}(x_k) + f^{\text{LagTemperature}}(x_{k-48}) \\
+ \; & f^{\text{TimeOfYear}}(x_k) + x_k^{\text{LoadDecrease}} f^{\text{LoadDecrease}}(x_k) + \epsilon_k .
\end{aligned}
$$

Lag temperature (accounting for thermal inertia)

Transfer functions learned from data:



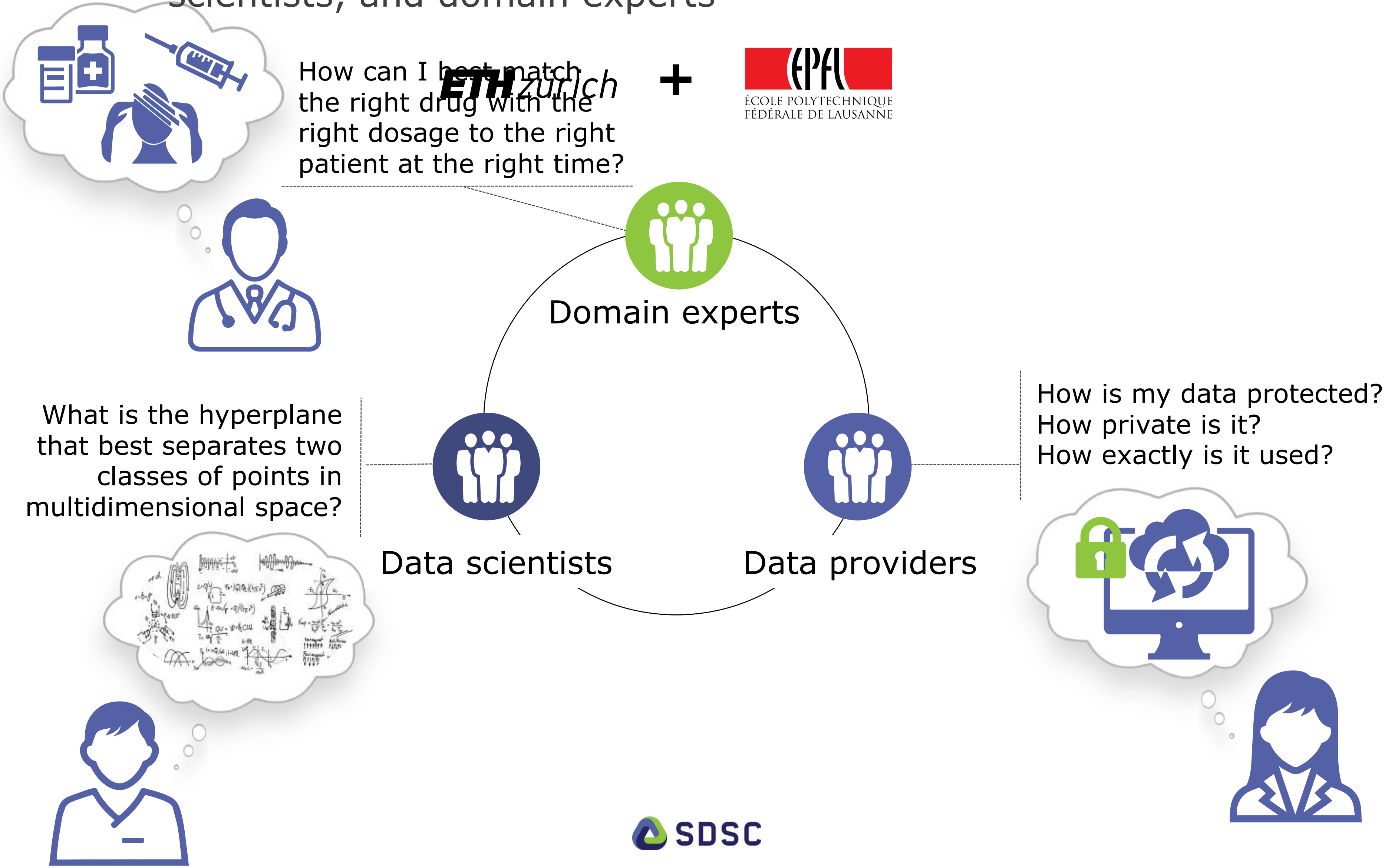


Swiss Data Science Center (SDSC)

Foster adoption of data science both in academia and industry

Multi-disciplinary team of 40 full-time computer and data scientists, and domain experts

ETH zürich + ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE



How can I best match the right drug with the right dosage to the right patient at the right time?

Domain experts

What is the hyperplane that best separates two classes of points in multidimensional space?

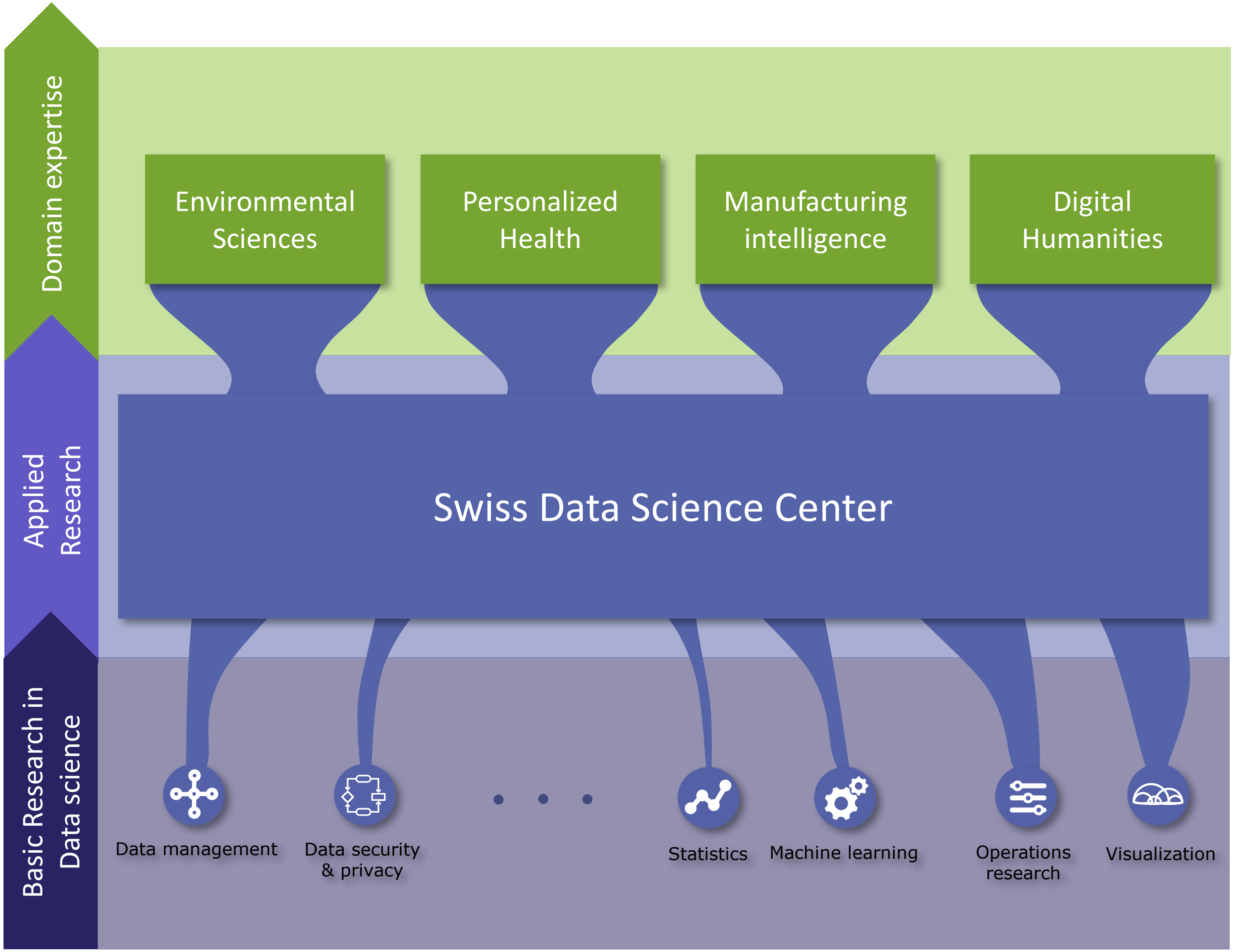
Data scientists

How is my data protected?
How private is it?
How exactly is it used?

Data providers

SDSC

Where does SDSC fit?



Domain expertise

Applied Research

Basic Research in Data science

Environmental Sciences

Personalized Health

Manufacturing intelligence

Digital Humanities

Swiss Data Science Center

Data management

Data security & privacy

. . .

Statistics

Machine learning

Operations research

Visualization

SDSC

What will the SDSC offer?

Excellence in academic research backed by strong industrial experience

Embedded R&D collaboration

We engage in academic and industrial collaborations requiring large-scale distributed data processing (Big & Fast Data) and/or advanced analytics (machine learning & statistics) combined with an in-depth knowledge in select domains

Domain-specific Insights as a Service

We provide secure access to our cloud-hosted analytics platform - the Open Insights Factory, a highly scalable open software platform offering a one-stop-shop for hosting and exploring curated, calibrated and possibly anonymized data at scale, at-rest or in-motion.

Open (Data) Science

The Insights Factory offers user-friendly tooling and services to help with the adoption of Open Science, fostering research productivity and excellence.

Answering Researchers Challenges

- A data lake, not a data swamp!
  - Where can I upload my data, and make it available?
  - What other data is available? And where is it?
  - How was this data created? Who created it?
  - How trustful is it? Can I build my research on it?

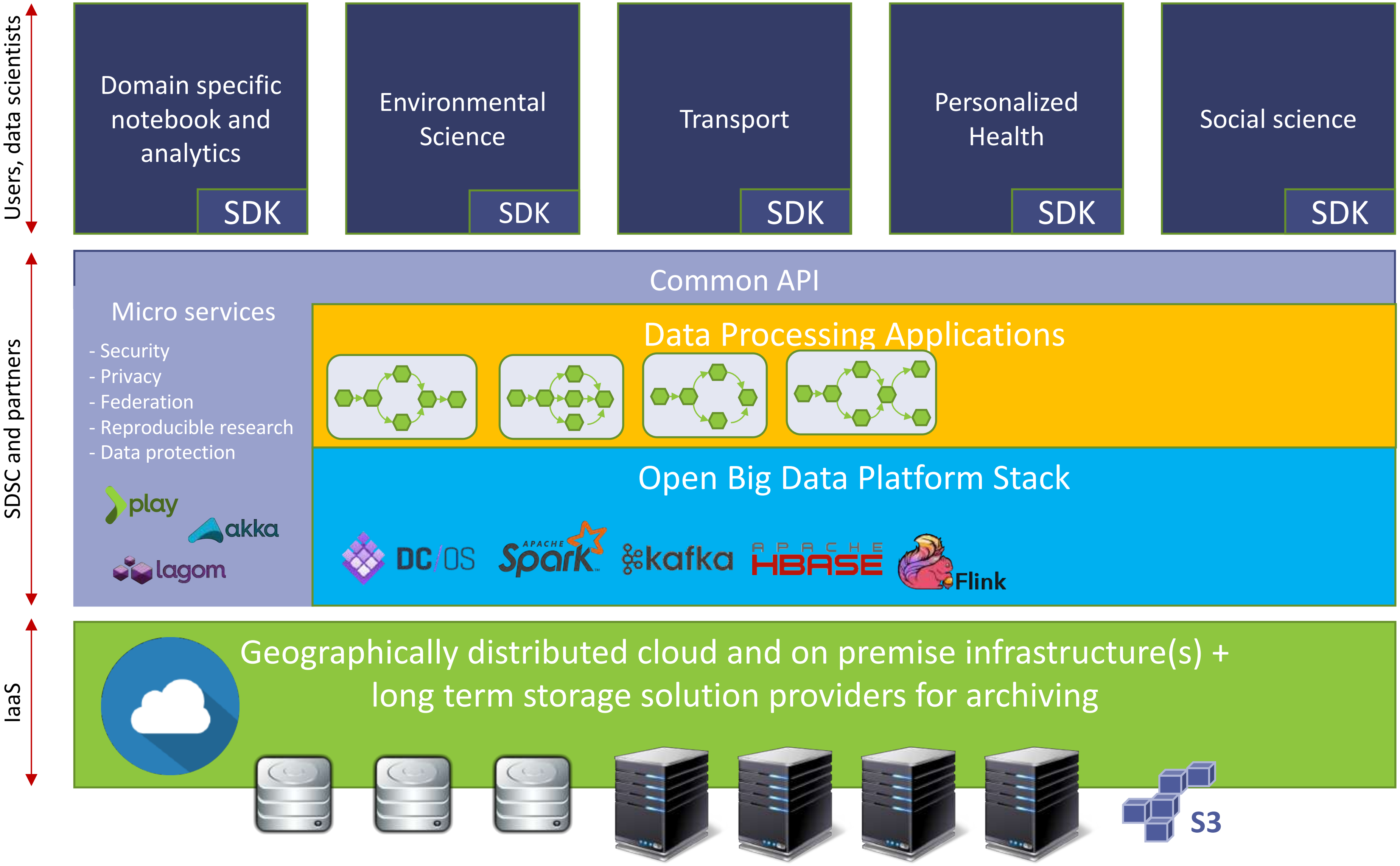
… impedes **collaboration** between scientists, and **reusability** and **reproducibility** of research

- Data science made simple & trustable
  - Combining human expertise and machine intelligence
  - Making learning methods robust against uncertainties
  - Designing methods for interpretable machine learning

Hosted Analytics Platform

- Highly-scalable open software platform offering domain-specific insights as a service, featuring:
  - Data protection and digital rights management
  - Secure computing across (semi-)autonomous entities
  - Reusable research data and reproducible science
  - Agile data science via interactive IDE for rapid R&D
  - Domain-specific analytics SDK and frameworks

The Software Stack



Users, data scientists

Domain specific notebook and analytics — SDK

Environmental Science — SDK

Transport — SDK

Personalized Health — SDK

Social science — SDK

SDSC and partners

Common API

Micro services

- Security
- Privacy
- Federation
- Reproducible research
- Data protection

Data Processing Applications

Open Big Data Platform Stack

IaaS

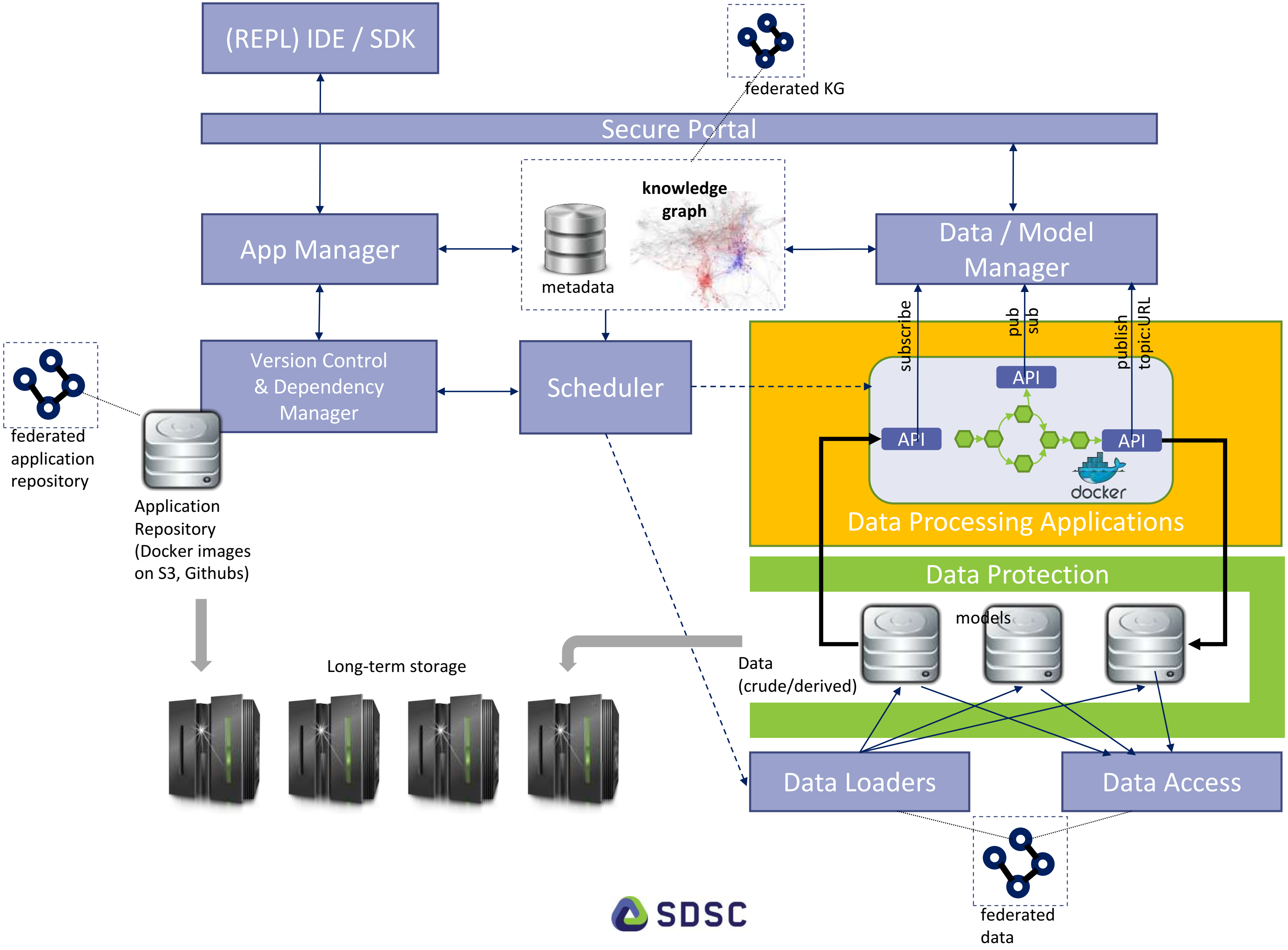
Geographically distributed cloud and on premise infrastructure(s) + long term storage solution providers for archiving

Use Cases in Environmental Science

- Addressing several data science challenges
  - From data ingest to insights discovery
  - Dealing with complex data
    - Network of physical sensors
    - Mix of streaming & historical data
  - Physics-informed machine learning
  - Reusability of research data
  - Reproducibility of science

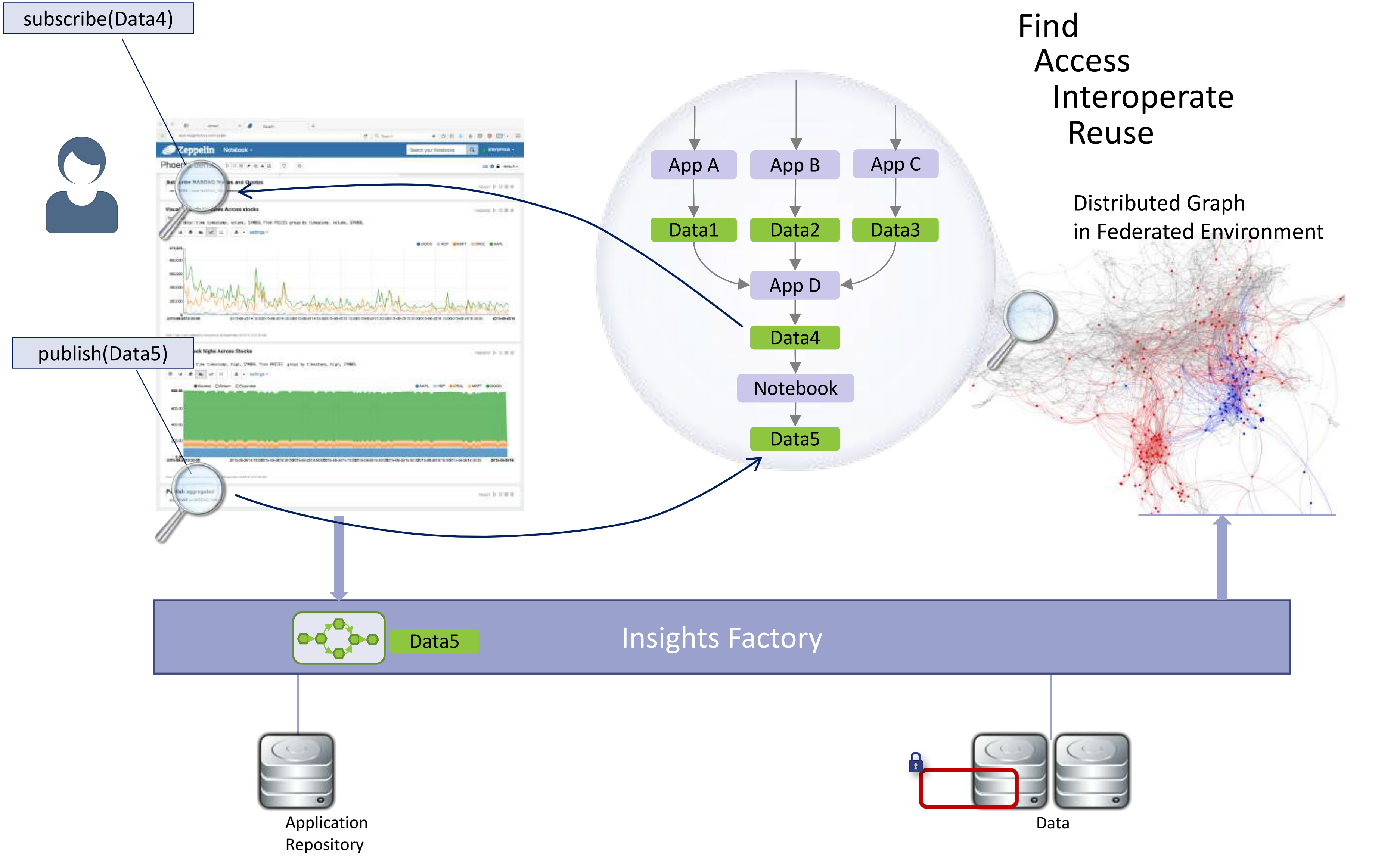
- Demonstrators
  - CarboSense with Empa & Swisscom (Nano-Tera Gateway)
  - Grassland Science with Nina Buchmann (ETH Zurich)
  - ecoHydrology with Tom Battin (EPFL)

Building the Knowledge Graph



SDSC

Data Science Governance

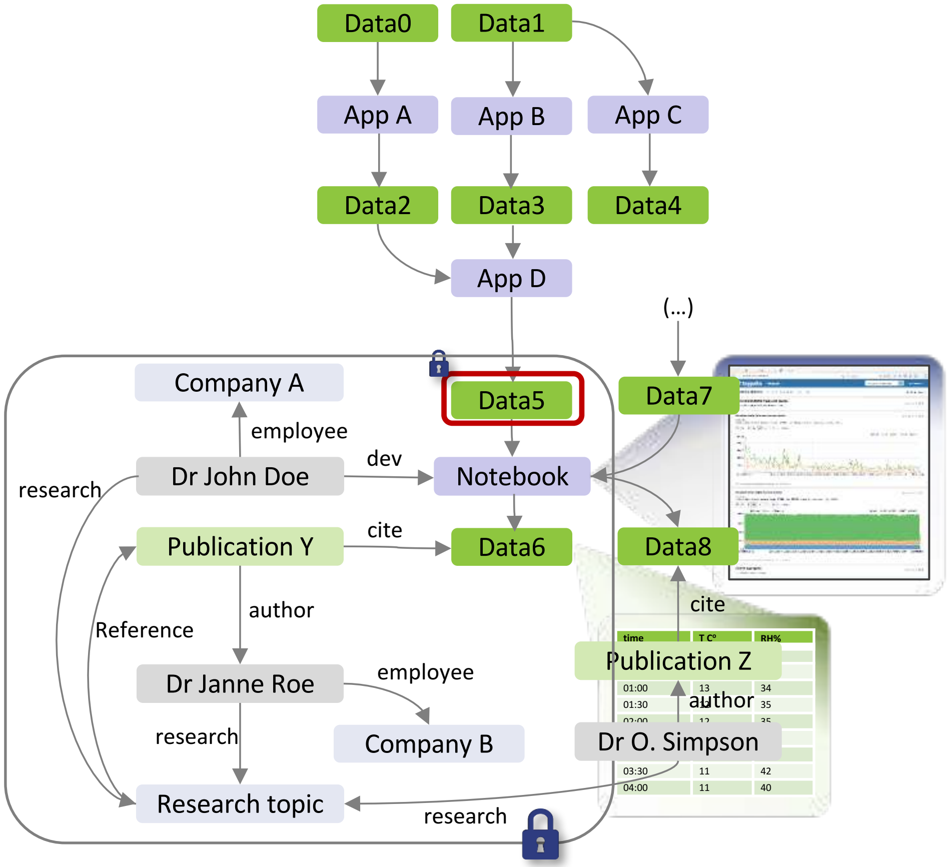


Find
Access
Interoperate
Reuse

DEMO

Automated Open Science

- Reproducible Research
  - See the (versioned) algorithms
  - See the data
  - Replay a workflow
  - Compare workflows, validate robustness
- Reusability
  - Reuse data on new workflows
  - Clone and modify workflows
- Knowledge Graph
  - Data popularity, H-index
  - Who is using the data?
  - For what?
- IP Protection
  - Decide who sees the data,
  - The algorithms,
  - The data I use,
  - And how I use it



Platform Development Milestones

- **2017.06** – Beta (limited functionality, no guarantee for forward compatibility)
  - Data manager with meta-data, data lineage for data provenance graph
  - Versioned application repository with dependency manager
  - Semantic search on data and algorithms
  - Semantic pub/sup pluggable analytics
  - Automated orchestration of pluggable analytics and data movement
- **2018.01** – Early release for internal use
  - Federated platform
  - Notebook IDE (in collaboration with *Data Fellas*)
  - SDK extension(s) for selected domain(s)
  - Reproducible research
  - Social network services (data H-Index, who works with whom and on what data, …)
- **2018.06** – Open source release
  - (Free) Public license: without 3rd party/partners technology, community support only
  - (Pay-for) Enterprise license: extended features
- **Post-open source release**
  - Roll out of new domain-specific SDK extensions
  - Contribute additional data science algorithms
  - Continuous support to maintain advantage of state of the art and evolving technology

Current Status & Next Steps

The center is fully operational as of January 2017

Center set-up

In progress

- Hiring R&D staff
- Developing hosted platform
- **Collaborating across the Swiss academic community**
  - Personalized health
  - Environmental science
- **Engaging with industry**
  - Preventive maintenance

Call for Academic Research Proposals

March 2017

Motivations

- Foster and accelerate the adoption of data science across the ETH Domain
- Promote Open Science

Research themes

- Data science meets domain science
- Data science methods for the real-world

SDSC Industry Day

October 2017

Objectives

- Showcase R&D activities of the center
- Offer a platform for industry to engage with SDSC

Current SDSC Staff

THANK YOU!

http://www.datascience.ch

Twitter: @SDSCdatascience