APEL Presentation – 2017-03-14

Prof. Edouard Bugnion
Vice-President and CIO (ad Interim)
Director, Data Center Systems Laboratory
Co-Scientific Director, Swiss Data Science Center
#1: The Data Science Revolution

#2: Computing is changing

#3: Software is eating the world

Prof. Edouard Bugnion
AI: From DeepBlue to Watson and AlphaGo
Why do neural networks work?

- Anything humans can do in 0.1 sec, the right big 10-layer network can do too

Source: Jeff Dean, Google. "Large-scale Deep Learning"
Why this works

● Massive data sets
  ● Facebook uploads 600M pictures a day (not counting Instagram)
  ● Each picture goes through 2 DNN within 2 seconds [LeCun]

● Massive computational power
  ● Model parallelism: Google’s largest model runs on 144 machines (1 replica) [Dean]
  ● Data parallelism -- replicate the model, train, merge (n replicas)
SWISS DATA SCIENCE CENTER
A joint center between EPFL and ETH Zürich
A fragmented ecosystem

There is an inherent gap between data scientists, domain experts and data providers
How is my data protected?

How private is it?

How exactly is it used?

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

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

Swiss Data Science Center (SDSC)

Fostered by ETH Zurich, EPFL, academic funding, and industrial collaboration

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

Academic funding & Industrial collaboration

Scientific breakthroughs

Domain experts

Data scientists

Data providers

Capacity to federate
Where does SDSC fit?

- Environmental Sciences
- Personalized Health
- Manufacturing intelligence
- Digital Humanities

Domain expertise:
- Applied Research
- Basic Research in Data science

Swiss Data Science Center

- Data management
- Data security & privacy
- Statistics
- Machine learning
- Operations research
- Visualization
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.
• 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
  • Explainable / interpretable machine learning
  • Resilient / robust data science
Open Insights Factory

- 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
Open Insights Factory – Software Stack

- Domain specific notebook and analytics
- Environmental Science
- Transport
- Personalized Health
- Social science

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

- Common API

- Data Processing Applications

- Open Big Data Platform Stack

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

- Users, data scientists
- SDSC and partners
- IaaS
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)

- **Infrastructure / IaaS / PaaS**
  - SWITCH / SWITCHEngine as P-2 SCALE UP use case
Automated Knowledge Graph

Find
Access
Interoperate
Reuse

Distributed Graph in Federated Environment

subscribe(Data4)
publish(Data5)

Insights Factory

Application Repository

SDSC
Open (Data) 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
Immediate next steps

The center is fully operational as of January 2017

**Center set-up**
- Hiring R&D staff
- Defining governance
- Planning budget
- Identifying projects in selected domains to bootstrap the design of the platform

**Kick-off event**
- One-day workshop in Bern
- Joint opening ceremony given by the Presidents of EPFL and ETH Zürich

**Annual call for projects**
- Evaluation criteria:
  - scientific merit
  - impact of results
  - novelty of approach
- Center covers:
  - use of SDSC platform
  - compute / storage resources
  - Help from SDSC personnel

In progress

6 February 2017

March 2017

SDSC
#1: The Data Science Revolution
#2: Computing is changing
#3: Software is eating the world

Prof. Edouard Bugnion
Cloudy clouds

I hired a consultant to help us evolve our products to cloud computing.


It's as if you're a technologist and a philosopher all in one!

Blah blah platform.

www.dilbert.com; 2011/1/07; (licensed)
Very large datacenters

Every day, AWS adds enough new server capacity to support all of Amazon’s global infrastructure when it was a $7B annual revenue enterprise

Source: James Hamilton, 2015
In financial terms – 4Q2016

- **AWS 4Q**
- **Sales** = 3,5B (+47%)
- **Income** = 926M$ (26% of sales)
And the winner is (according to the Source: Dell ‘Oro 2Q15)
Rethinking scalability

Source: Adrian Cockcroft, NetflixOSS, 2013
#1: The Data Science Revolution
#2: Computing is changing
#3: Software is eating the world

Prof. Edouard Bugnion
Software is eating the world

THE LARGEST COMPANIES BY MARKET CAP
The oil barons have been replaced by the whiz kids of Silicon Valley

<table>
<thead>
<tr>
<th>Year</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>Exxon</td>
<td>Microsoft</td>
<td>EXON</td>
<td>Citi</td>
<td>Walmart</td>
</tr>
<tr>
<td></td>
<td>$406B</td>
<td>$365B</td>
<td>$272B</td>
<td>$261B</td>
<td>$260B</td>
</tr>
<tr>
<td>2006</td>
<td>Exxon</td>
<td>Exxon</td>
<td>TOTAL</td>
<td>Microsoft</td>
<td>Citi</td>
</tr>
<tr>
<td></td>
<td>$446B</td>
<td>$383B</td>
<td>$327B</td>
<td>$293B</td>
<td>$273B</td>
</tr>
<tr>
<td>2011</td>
<td>Exxon</td>
<td>Apple</td>
<td>PalcoCela</td>
<td>Shell</td>
<td>CBBC</td>
</tr>
<tr>
<td></td>
<td>$406B</td>
<td>$376B</td>
<td>$277B</td>
<td>$237B</td>
<td>$228B</td>
</tr>
<tr>
<td>2016</td>
<td>Apple</td>
<td>Alphabet</td>
<td>Microsoft</td>
<td>Amazon</td>
<td>Facebook</td>
</tr>
<tr>
<td></td>
<td>$582B</td>
<td>$556B</td>
<td>$452B</td>
<td>$364B</td>
<td>$359B</td>
</tr>
</tbody>
</table>
Fictional Case study: Garbage Collection
Old-style garbage collection

Negotiate with municipality

Buy trucks

Optimize the route
Introduce a small amount of technology

- Add sensor to bin
- Upload data
- Optimize route
- Renegociate
- Sell trucks
Garbage Collection – full stack edition

- Give out sensors
- Analyze
- Optimize route

Global

Local
GARBAGE IS SEXY: It's A $1 Trillion Market That Finnish Startup Enevo Wants To Disrupt
Technology & Disruption

Enabling Technology

3G sensors

Disrupted industry

Garbage Collection?

smartphone
“Before Uber and Lyft were started, there were multiple startups that tried to build software that would make the taxi and limo industry more efficient. Then they went out and knocked on the door of taxi companies and pitched them on their software”

… and the rest is history

Source: http://a16z.com/2015/01/22/the-full-stack-startup/
More examples covering all industries

- Uber
- Airbnb
- Nest
- Amazon
- Netflix
- Wealthfront
“Software is eating the world.”

“Every business is in the software business.”

“It’s an application economy.”
Summary

Efficiency of cloud computing
- Deliver compute, storage, networking **at scale** as a commodity
- The industrial revolution of ICT

Business is IT!
- Software is eating the world
- Combination of connected devices and analytics

Digital Sovereignty and Privacy
- Requires an updated, innovative, pragmatic, legal definition
- Independence for the 21st century

The Data Science Revolution
- Deliver human-level insights (rather than improve productivity)
- Essential skill for the 21st century – cannot be viewed as a commodity