LESSONS LEARNED IN DATA-DRIVEN SCIENCE

www.datascience.ch
@SDSCdatascience
The Swiss Data Science Center

Accelerate the adoption of data science and AI in Switzerland

- An initiative from the ETH Domain, started in 2017
- Offices in Zurich and Lausanne
- Academic and industry collaborations

Academic projects
Industry collaborations
RENKU platform
How is my data protected?
How exactly is it used? By whom?

What is the best architecture for my deep Convolutional Neural Network (CNN)?

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

Closing the gaps in the data science journey

Data scientists

Domain scientists

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

BY LOWE FOR THE SUN-SENTINEL, FLO
From raw data to unbiased information
Antarctic Circumnavigation Expedition

Expedition boat with 22 teams from South Africa to Australia to Chile in 90 days

**Context:** Foster collaboration between teams of scientists, breaking data silos

**Initial problem:** Model relationships between ocean / wave parameters and aerosols
The Data-driven Science journey

Data + Algorithms $\Rightarrow$ Knowledge $\Rightarrow$ Benefits

- Big Data / Data lake
- Machine Learning / “Dumb” A.I.
$f : \text{Cat} \rightarrow \text{Cat}$
Explainable AI – What Are We Trying To Do?

Today

- Training Data
- Learning Process
- Learned Function
- 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?

Tomorrow

- Training Data
- New Learning Process
- Explainable Model
- Explanation Interface
- User with a Task

• 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
Data-Driven Acoustical Design

Joint research collaboration with the Architecture group at ETH Zurich

**Problem:** Modeling sound propagation and diffusion in everyday rooms

**Initial results:** Estimation of impulsive response from different walls

From the wall, we predict the signal that would be received on each microphone.
Deep Learning for Observational Cosmology

Joint research collaboration with the Cosmology Research Group at ETH Zurich

**Problem:** Observational cosmology relies on computationally expensive simulations

**Results:** Using a generative adversarial network (GAN), we can generate new approximate simulations for a fraction of the computational resources.
From AI to Data Science

DATA SCIENCE SKILLSET

- **Hacking Skills**: Necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.

- **Math and Statistics Knowledge**: Allows a data scientist to choose appropriate methods and tools in order to extract insight from data.

- **Substantive Expertise**: In a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.

- **Traditional Research**: Lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.

- **Machine Learning**: Stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.

- **Danger Zone**: Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

Data science, due to its interdisciplinary nature, requires an intersection of abilities: **hacking skills**, **math and statistics knowledge**, and **substantive expertise** in a field of science.
Sharing data and knowledge, or lack thereof

credit: oxfordcreativity, www.triz.co.uk
1. How did I compute this result?

2. How does new data change this result?

3. How did you compute your result?
   Can I use your data to reproduce it?
   With your code?
   On your infrastructure?

4. Has anyone ever used an <XYZ-algorithm> on this data? How?

5. Who is using my data? and my algorithm?
   Why are they not citing me?!
Reproducibility
Reusability
Collaboration