DATA SCIENCE IN THE REAL WORLD

CHALLENGES & OPPORTUNITIES

Olivier Verscheure, PhD
Swiss Data Science Center
EPFL & ETH Zurich
About me

Academia

1999

2016

17y

New York

Dublin
What Do You See?
Crunch time in France
Ten years on: banking after the crisis
South Korea’s unfinished revolution
Biology, but without the cells

The world’s most valuable resource

Data and the new rules of competition
Like Oil, Data Must be Refined
“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!”
Data Science, a Fragmented Ecosystem

Extract Meaning and Value out of Data

What is the hyperplane that best separates two classes of points in multidimensional space?
Machine Learning in a Nutshell

\[ f : \quad \rightarrow \quad \text{Cat} \]
Machine Learning (ML) in a Nutshell

\[
\min_{f \in \mathcal{F}} \sum_{(x, y) \in D} \ell(y, f(x))
\]
Some Key Players

- IBM
- Uber AI Labs
- Baidu
- Google
- Apple
- Facebook
- Amazon
- Microsoft
- NVIDIA
- SDSC
Recent advances in ML: Deep Learning

ImageNet Challenge

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.
Recent advances in ML: Deep Reinforcement Learning
This Success Relies On...

1. Large dataset of labelled data
2. Good quality data
3. Enough computing power
4. Clear and measurable objectives
An Unexpected Outcome

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

This is not a sock
Turning a STOP sign into a 45 mph speed limit
Structured vs Unstructured Data

Structured

Unstructured

Semi-structured
The Structured World

Health care

Predictive maintenance

Energy networks

Financial time series

Social networks
A Sobering View of Data Science

This is your machine learning system?

Yup! You pour the data into this big pile of linear algebra, then collect the answers on the other side.

What if the answers are wrong?

Just stir the pile until they start looking right.
Obstacles to a Wider Adoption in a Structured World

1. Large dataset of labelled data -> Labelling is expensive
2. Good quality data -> Data is usually missing/Increased uncertainty
3. Clear and measurable objectives -> Knowledge discovery/causality
4. Lack of interpretability/lack of trust
1. Incorporating structure knowledge in the model for data curation
2. Dealing with uncertainties
3. Promoting causality/interpretability
4. Focusing on unsupervised/semi-supervised learning
• Forecasting demand in electricity (France)

\[ y_k = \beta_{\text{Intercept}} + f_{\text{Trend}}(k) + f_{\text{LagLoad}}(y_{k-48}) + \sum_{l=1}^{6} 1(x_{k}^{\text{DayType}} = l)(\beta_{l}^{\text{DayType}} + f_{l}^{\text{TimeOfDay}}(x_{k})) + f_{\text{CloudCover}}(x_{k}) + f_{\text{Temperature/TimeOfDay}}(x_{k}) + f_{\text{LagTemperature}}(x_{k-48}) + f_{\text{LoadDecrease}}(x_{k}) + \epsilon_k. \]

Transfer functions learned from data:

© IBM Research, EDF R&D
Other Roadblocks in Data Science Ventures

© Oxford Creativity 2012

credit: oxford creativity, https://www.triz.co.uk/
Facilitate communication to foster innovation
Foster multidisciplinary collaborations

Ooh! We could make an extra smooth wheel if we use highly polished marble!

There are hundreds of possible applications!

And a collapsible wheel made from a Pteranodon's intestine!

How about a super lightweight 'pumice' version?
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?

Foster adoption of data science both in academia and industry

Data scientists

Domain experts

Data providers

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

ETH Zürich

EPFL

Swiss Data Science Center (SDSC)
Key Actor in a Complex Ecosystem

Swiss Data Science Center

- Environmental Sciences
- Personalized Health
- Manufacturing intelligence
- Predictive maintenance

- Smart sensing
- Data security & privacy
- Statistics
- Machine learning
- Operations research
- Visualization
IN GOD WE TRUST.

ALL OTHERS MUST BRING DATA.

- W. EDWARDS DEMING, STATISTICIAN, PROFESSOR, AUTHOR