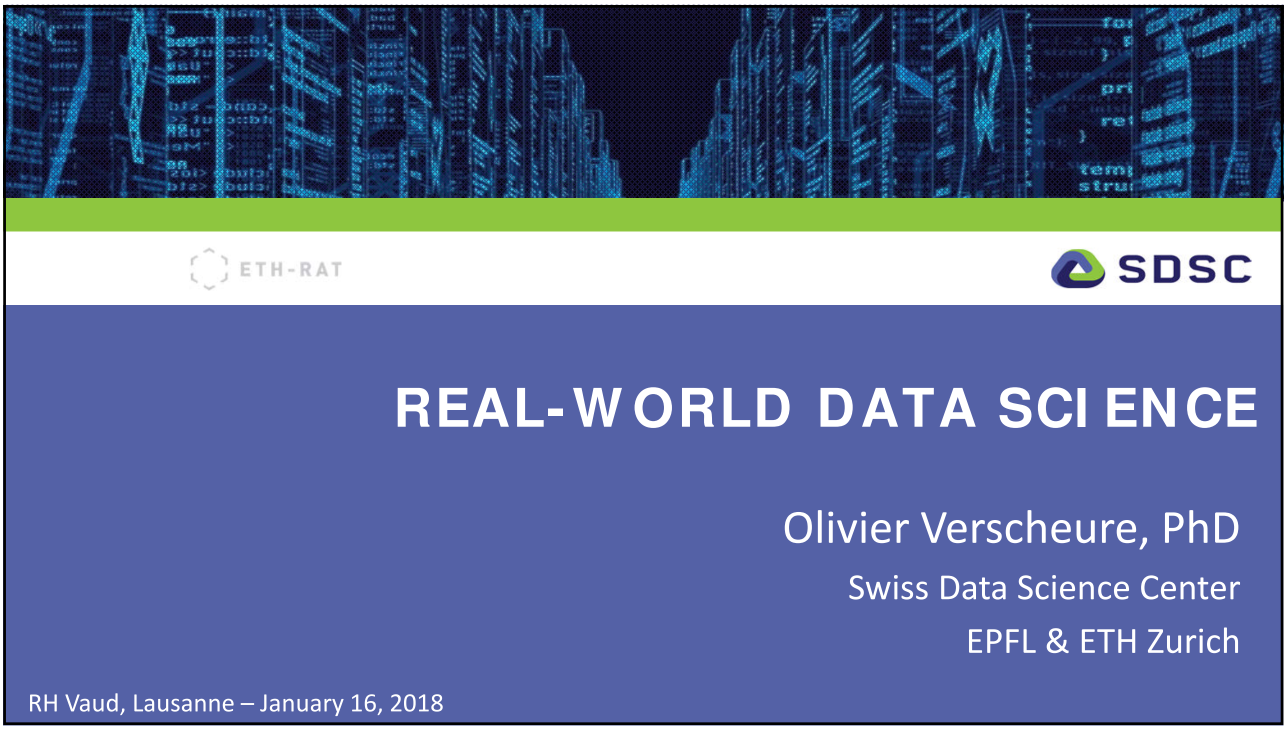
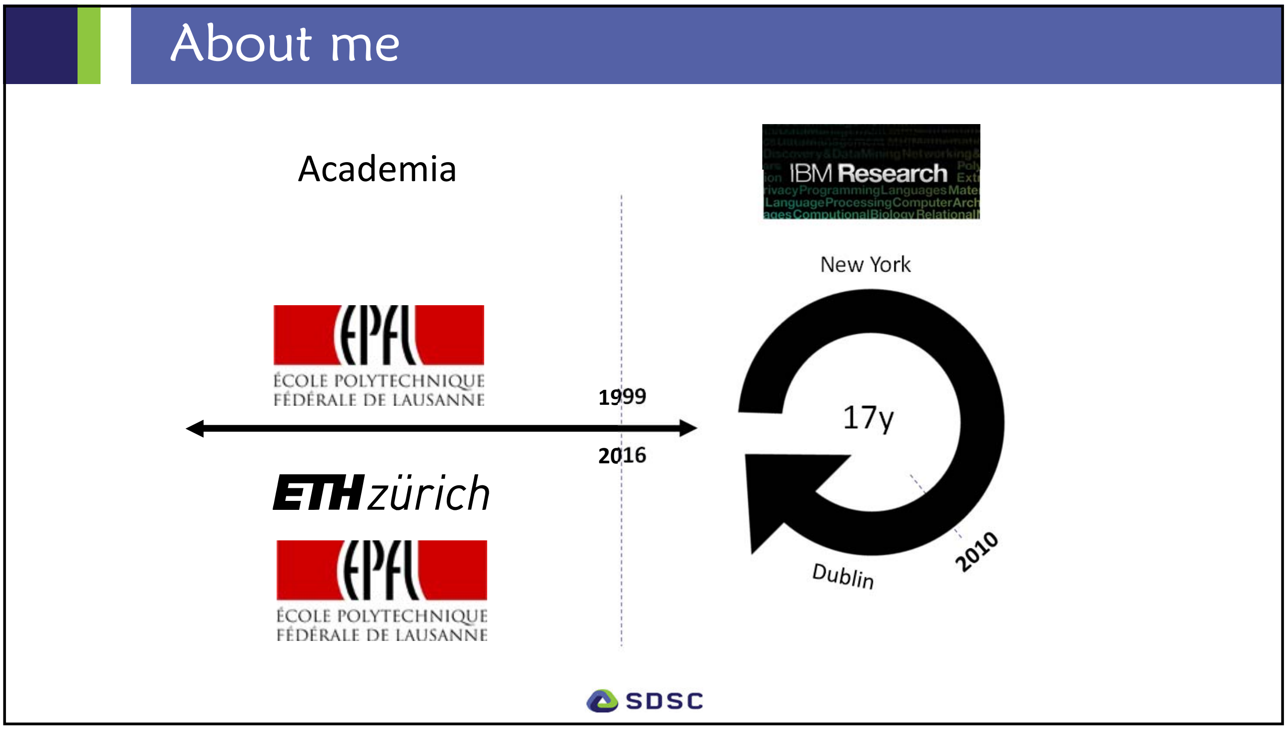


REAL-WORLD DATA SCIENCE

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

RH Vaud, Lausanne – January 16, 2018



About me

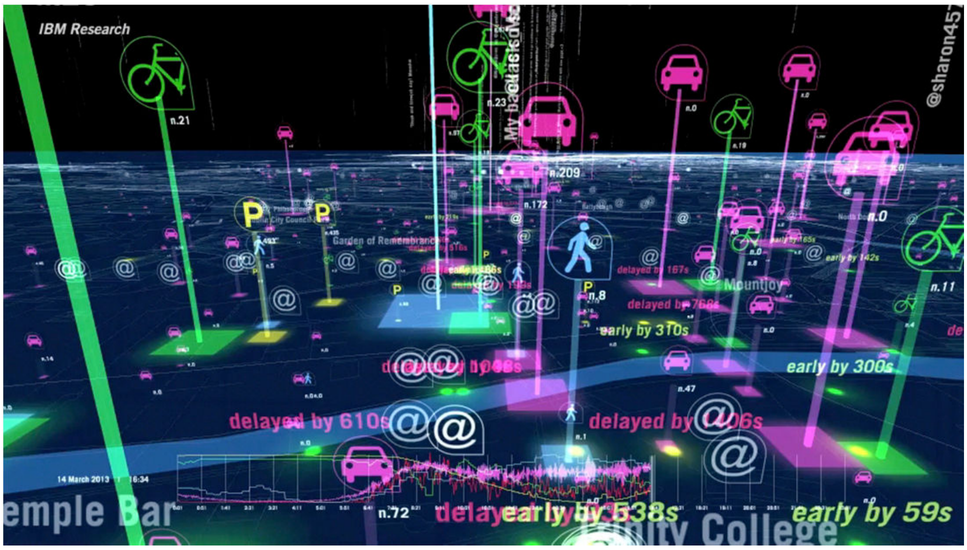
Academia

IBM Research
New York
17y
1999
2016
2010
Dublin
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
ETH zürich
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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



Data is the New Oil

The Economist

MAY 6TH–12TH 2017

Crunch time in France

Ten years on: banking after the crisis

South Korea's unfinished revolution

Biology, but without the cells

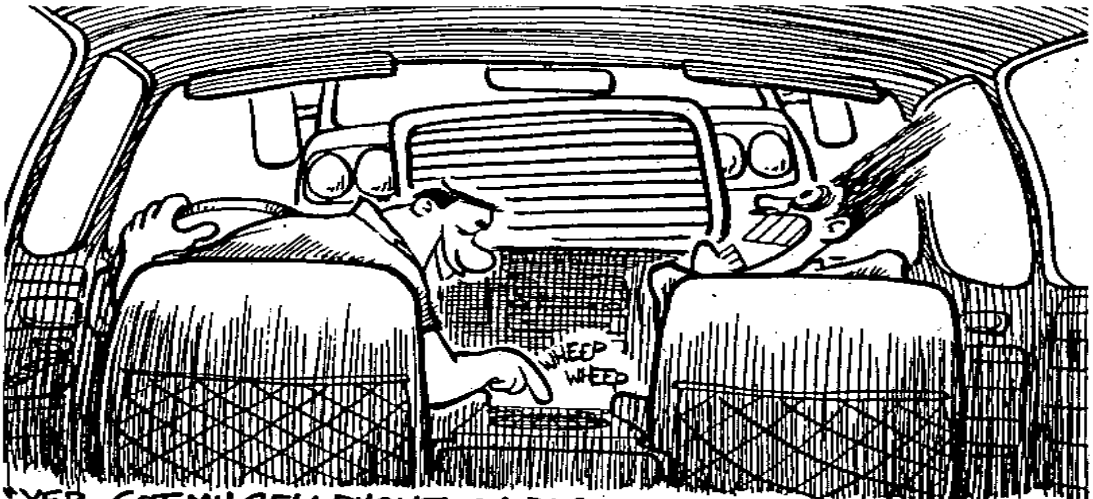
The world's most valuable resource

Data and the new rules of competition

amazon
UBER
Microsoft
Google
TESLA

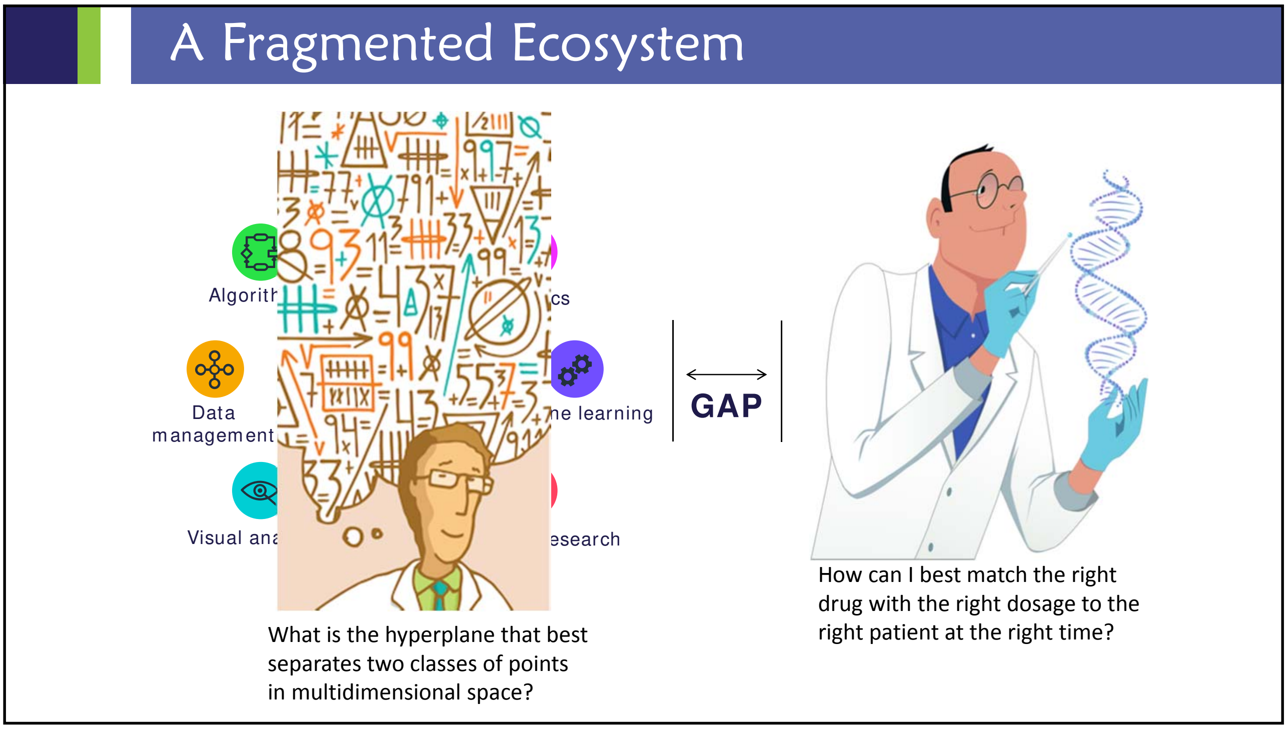
The Economist, May 2017

Big Data, Bad Data



"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!"

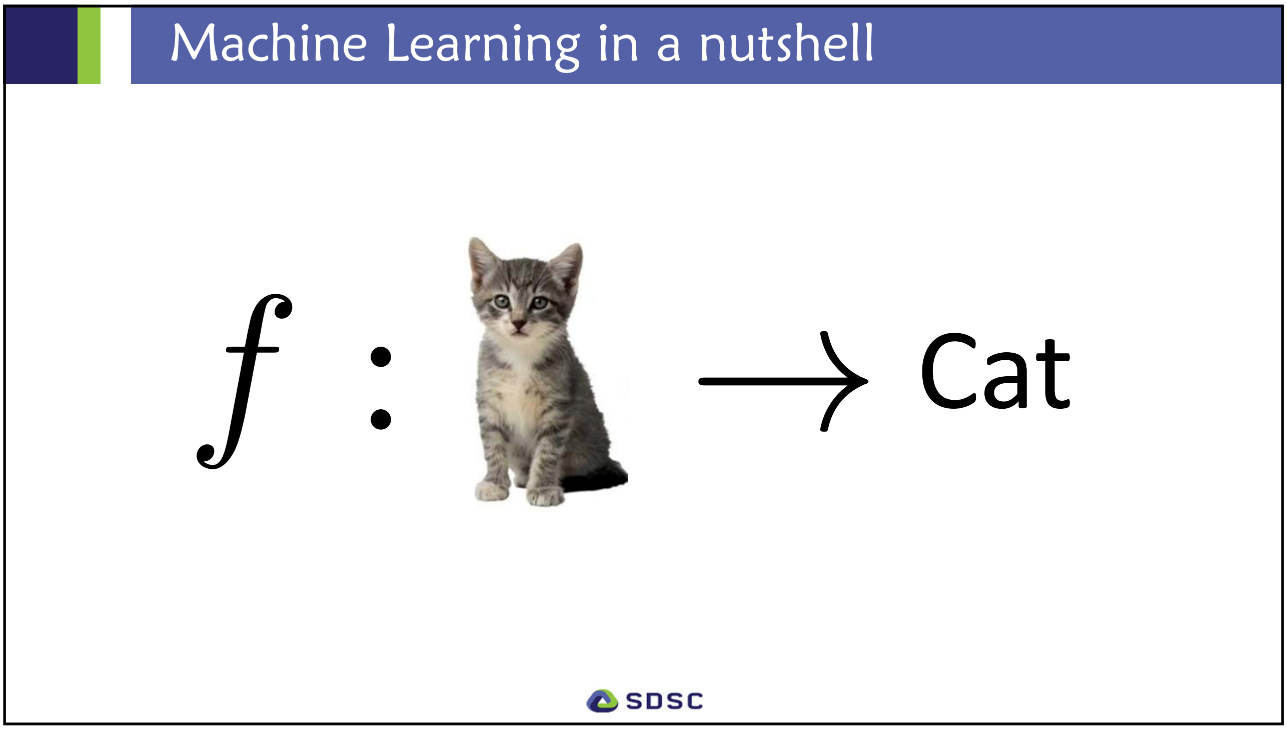
BY LOWE FOR THE SUN-SENTINEL, FLOR



A Fragmented Ecosystem

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

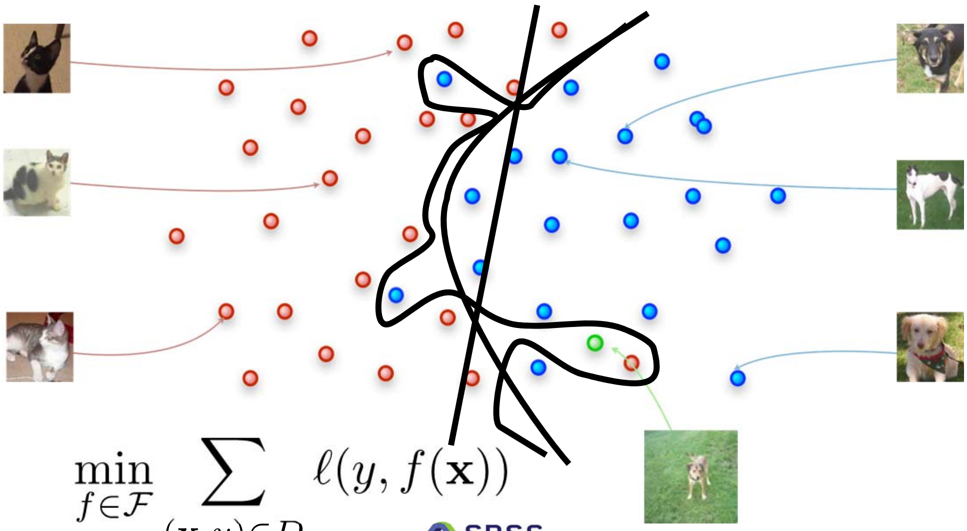
How can I best match the right drug with the right dosage to the right patient at the right time?



Machine Learning in a nutshell

SDSC

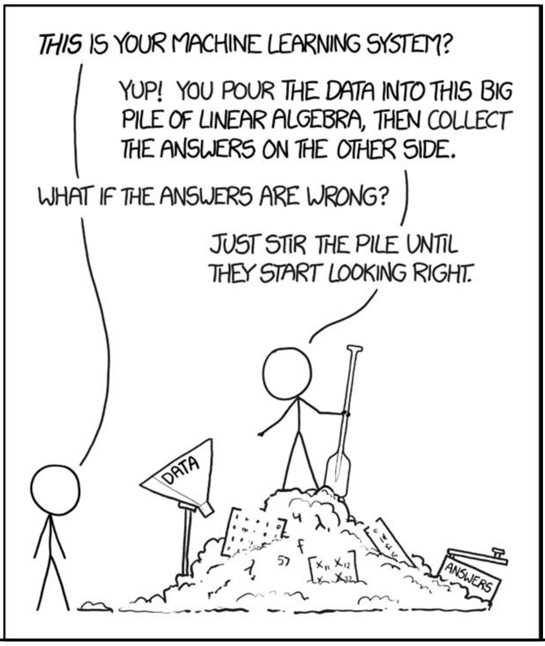
Machine Learning in a nutshell



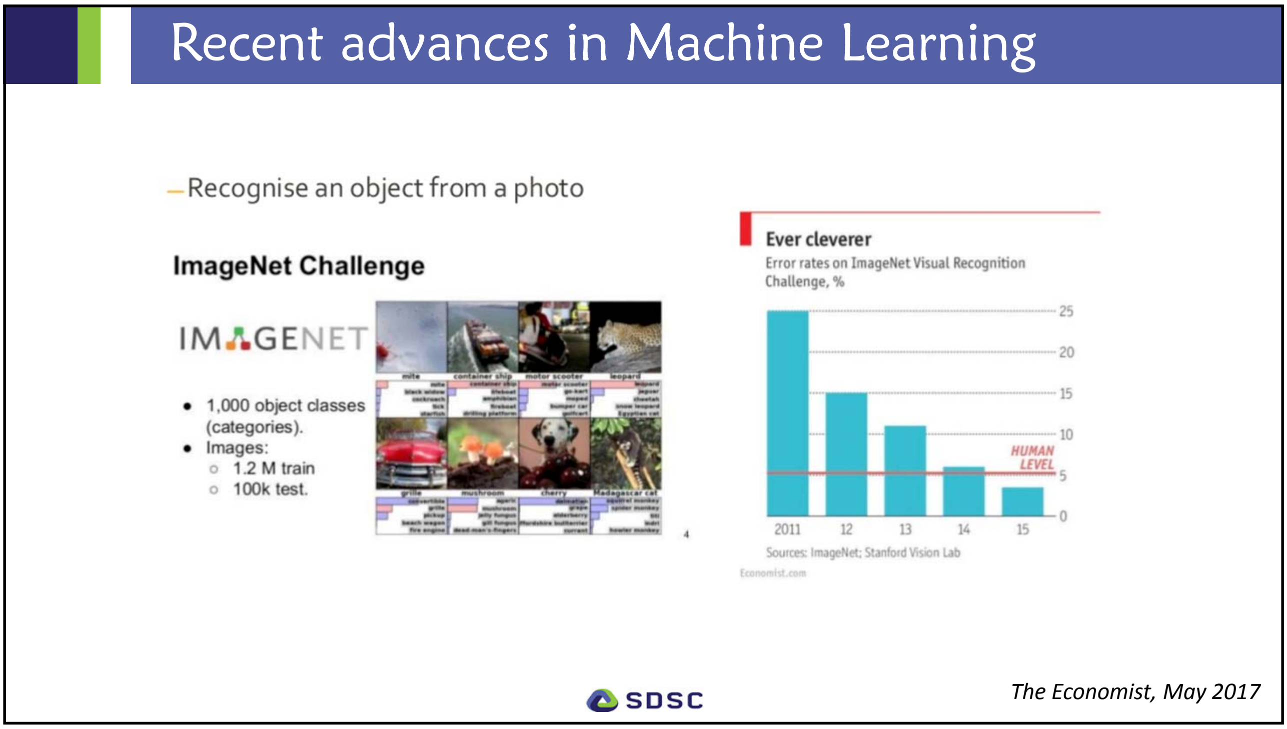
$$\min_{f\in\mathcal{F}} \sum_{(\mathbf{x},y)\in D} \ell(y, f(\mathbf{x}))$$

SDSC

A Sobering View on Machine Learning


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.
DATA
ANSWERS

© XKDC



Recent advances in Machine Learning

–Recognise an object from a photo

ImageNet Challenge

IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.

mite
container ship
motor scooter
leopard
grille
mushroom
cherry
Madagascar cat
4

Ever cleverer

Error rates on ImageNet Visual Recognition Challenge, %

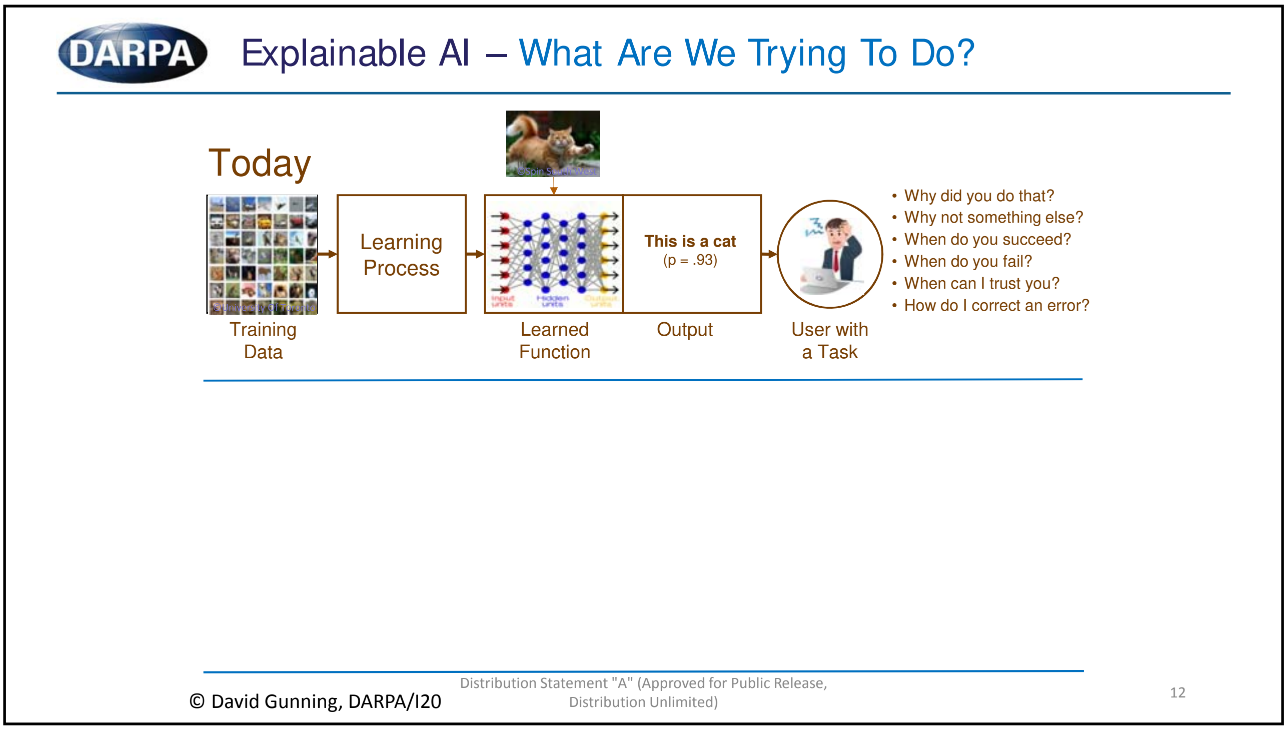
25
20
15
10
5
0
HUMAN LEVEL
2011
12
13
14
15

Sources: ImageNet; Stanford Vision Lab

Economist.com

SDSC

The Economist, May 2017



DARPA Explainable AI – What Are We Trying To Do?

Today

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?

© David Gunning, DARPA/I20

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

12

Fooling deep neural net classifiers

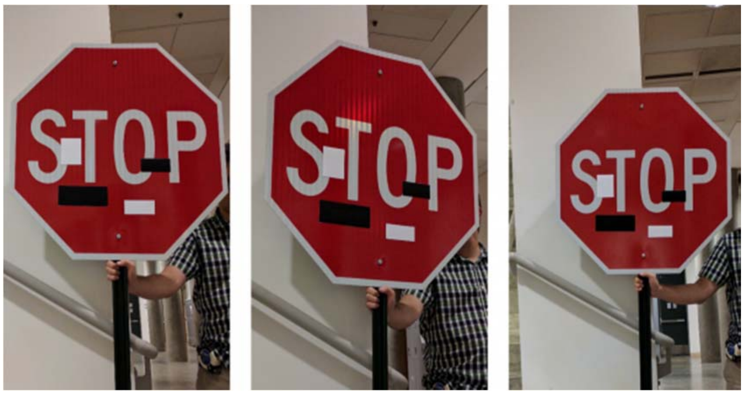
Title: Universal adversarial perturbations
Authors: Moosavi-Dezfooli, Seyed-Mohsen; Fawzi, Alhussein; Fawzi, Omar; Frossard, Pascal
Publication: eprint arXiv:1610.08401
Publication Date: 10/2016

This is not a sock

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

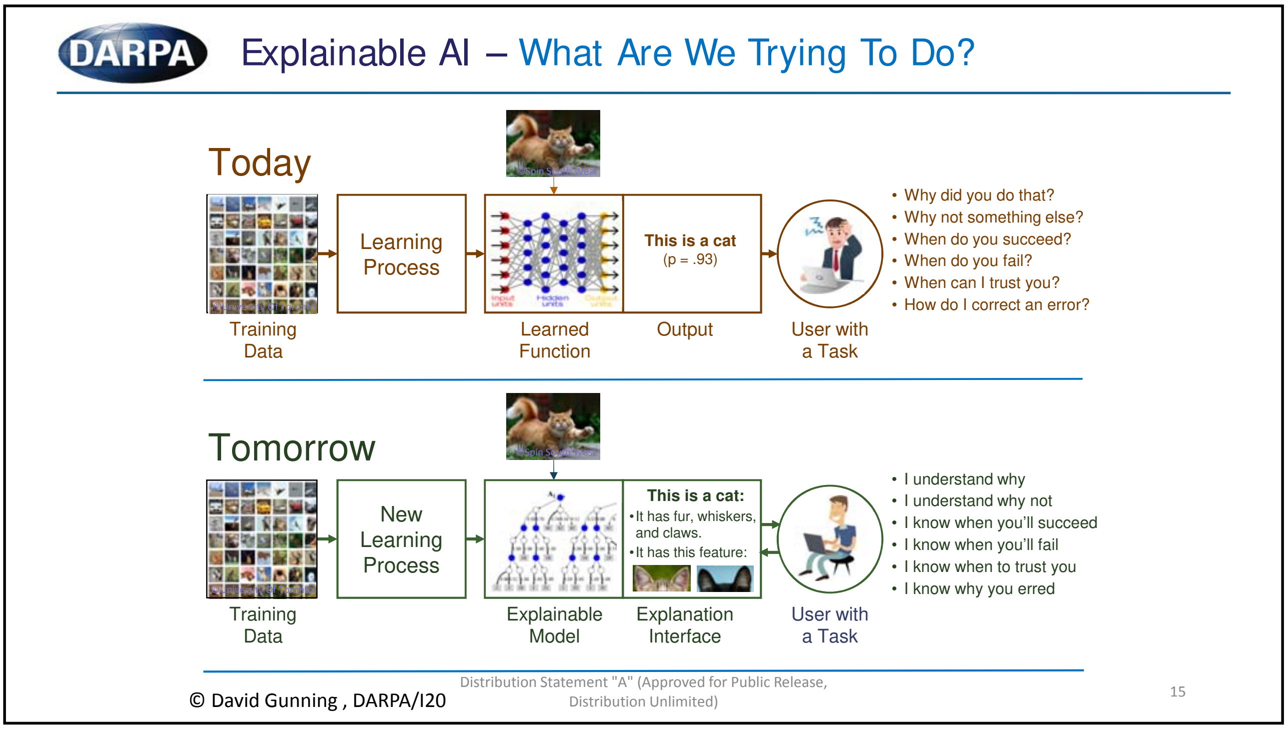
Fooling deep neural net classifiers

- Autonomous driving!



- Turning a **STOP** sign into a **45 mph speed limit**

SDSC



Explainable AI – What Are We Trying To Do?

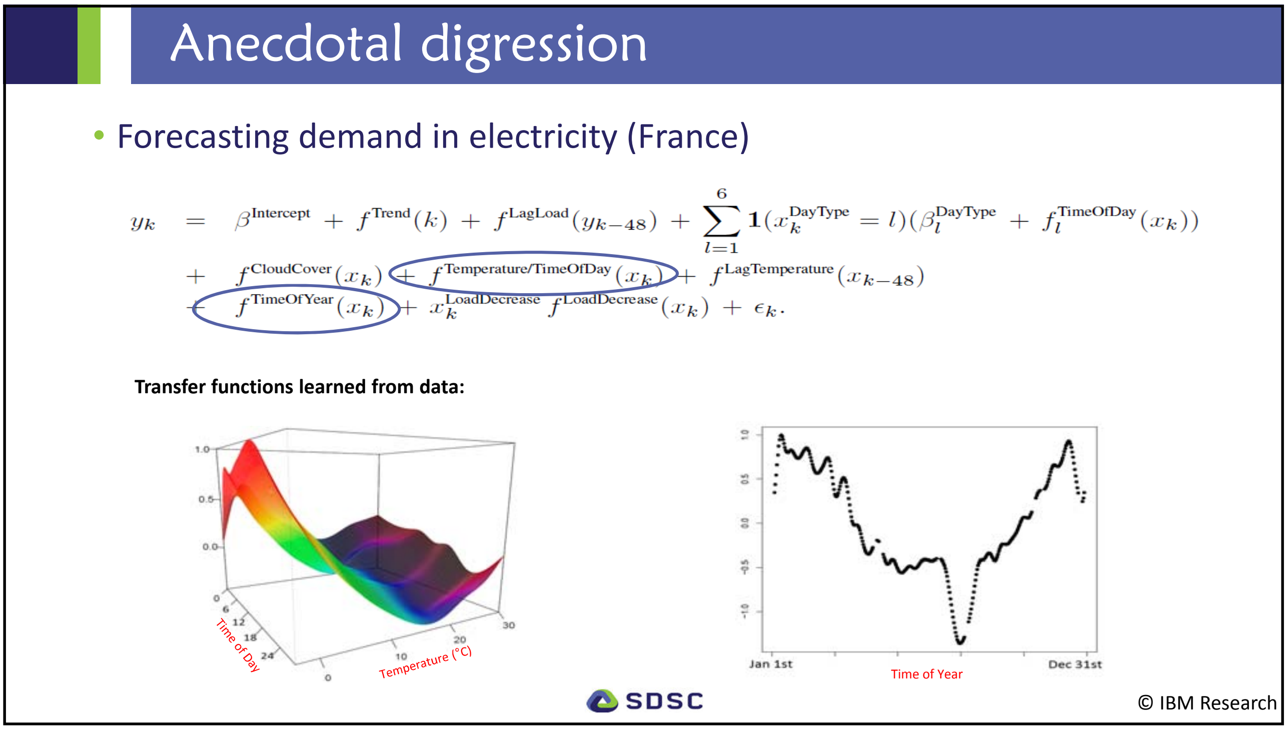
- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

- 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

© David Gunning , DARPA/I20

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

15



Anecdotal digression

- Forecasting demand in electricity (France)

$$
\begin{aligned}
y_k \quad = \quad & \beta^{\text{Intercept}} + f^{\text{Trend}}(k) + f^{\text{LagLoad}}(y_{k-48}) + \sum_{l=1}^{6} \mathbf{1}(x_k^{\text{DayType}} = l)(\beta_l^{\text{DayType}} + f_l^{\text{TimeOfDay}}(x_k)) \\
& + f^{\text{CloudCover}}(x_k) \boxed{+ f^{\text{Temperature/TimeOfDay}}(x_k)} + f^{\text{LagTemperature}}(x_{k-48}) \\
& \boxed{+ f^{\text{TimeOfYear}}(x_k)} + x_k^{\text{LoadDecrease}} f^{\text{LoadDecrease}}(x_k) + \epsilon_k .
\end{aligned}
$$

Transfer functions learned from data:

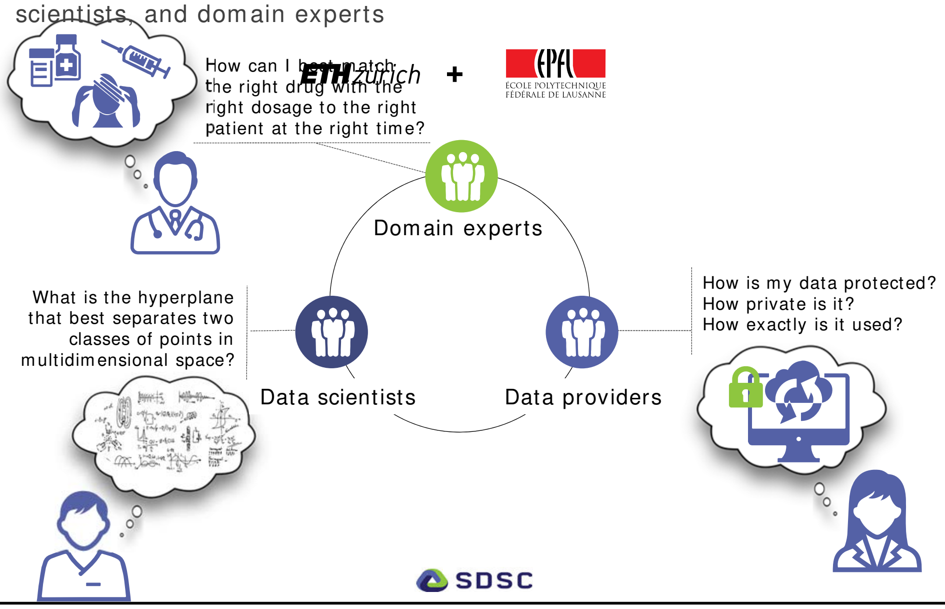
© IBM Research

Swiss Data Science Center (SDSC)

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

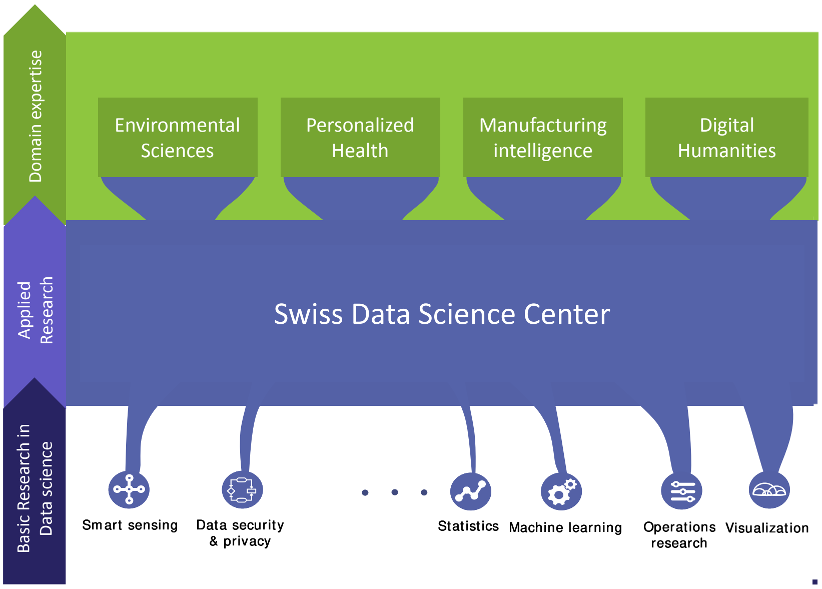
Foster adoption of data science in both academia and industry

ETH zürich + EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

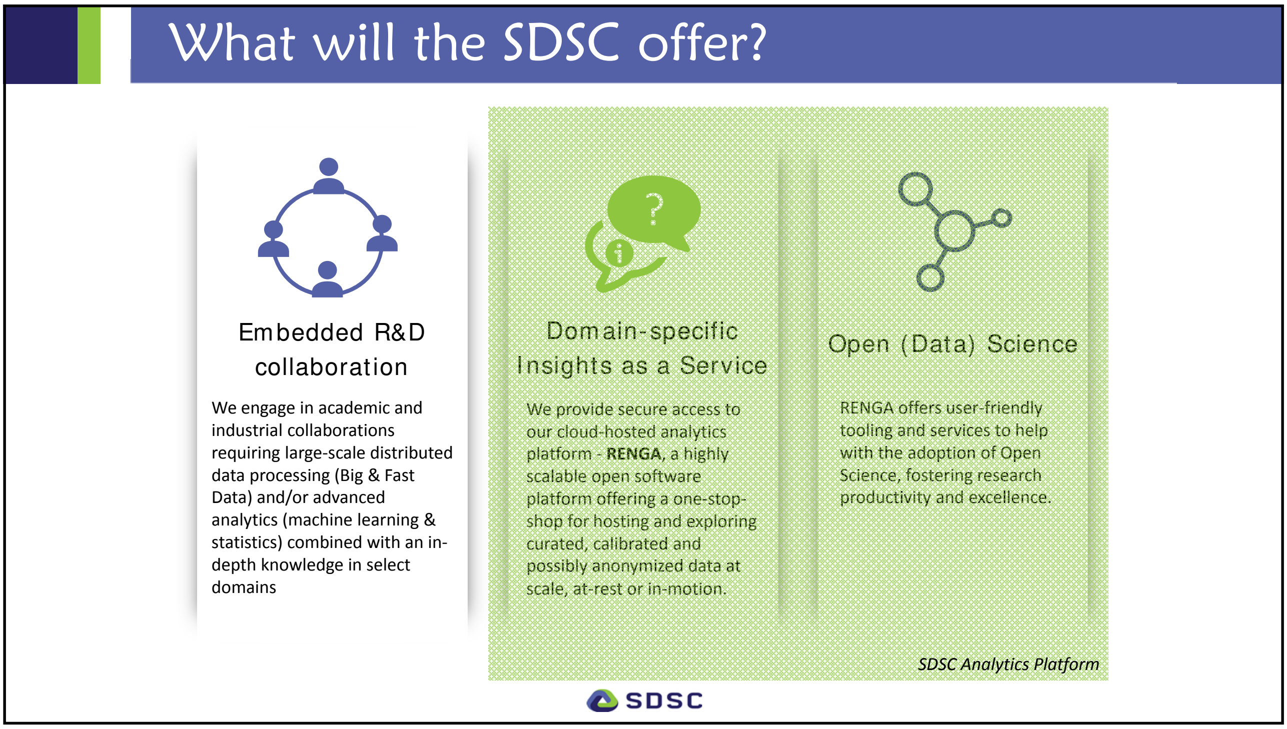

How can I best match the right drug with the right dosage to the right patient at the right time?
Domain experts
What is the hyperplane that best separates two classes of points in multidimensional space?
Data scientists
Data providers
How is my data protected?
How private is it?
How exactly is it used?

SDSC

Where does SDSC fit?



SDSC



What will the SDSC offer?

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

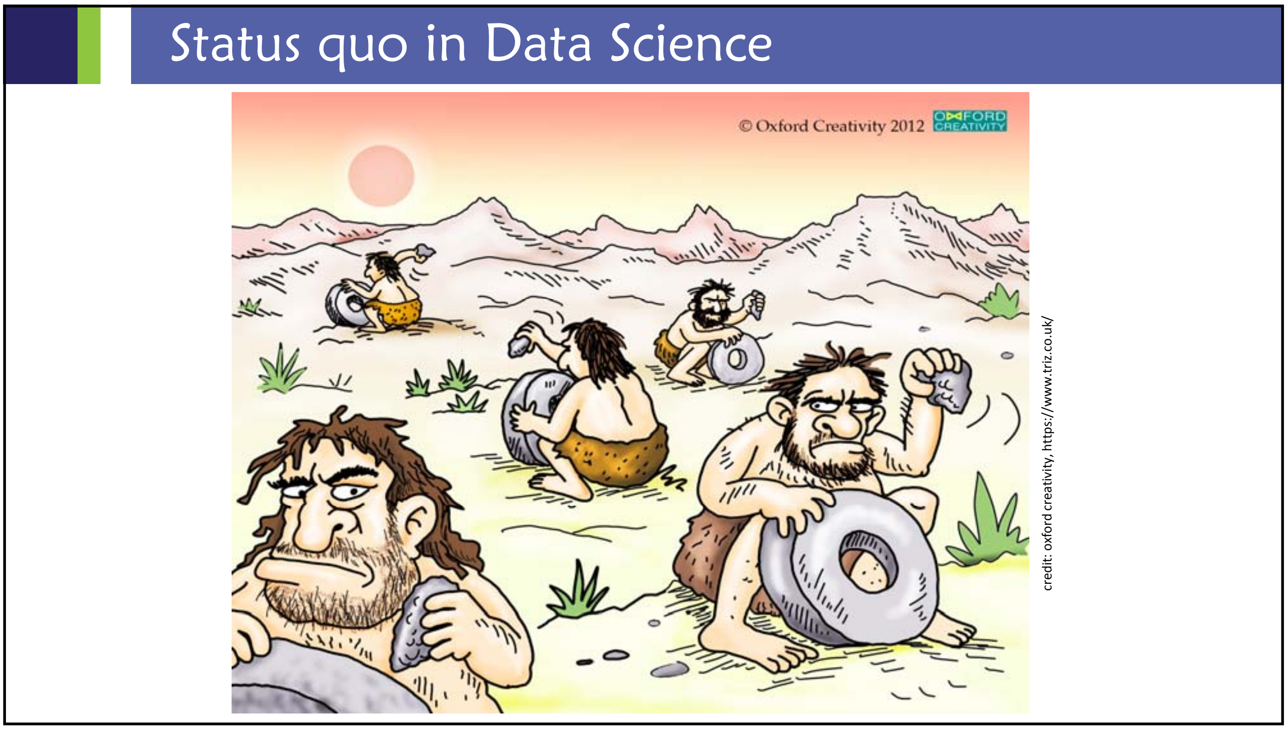
We provide secure access to our cloud-hosted analytics platform - **RENGA**, a highly scalable open software platform offering a one-stop-shop for hosting and exploring curated, calibrated and possibly anonymized data at scale, at-rest or in-motion.

Open (Data) Science

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

SDSC Analytics Platform

SDSC



Status quo in Data Science

© Oxford Creativity 2012

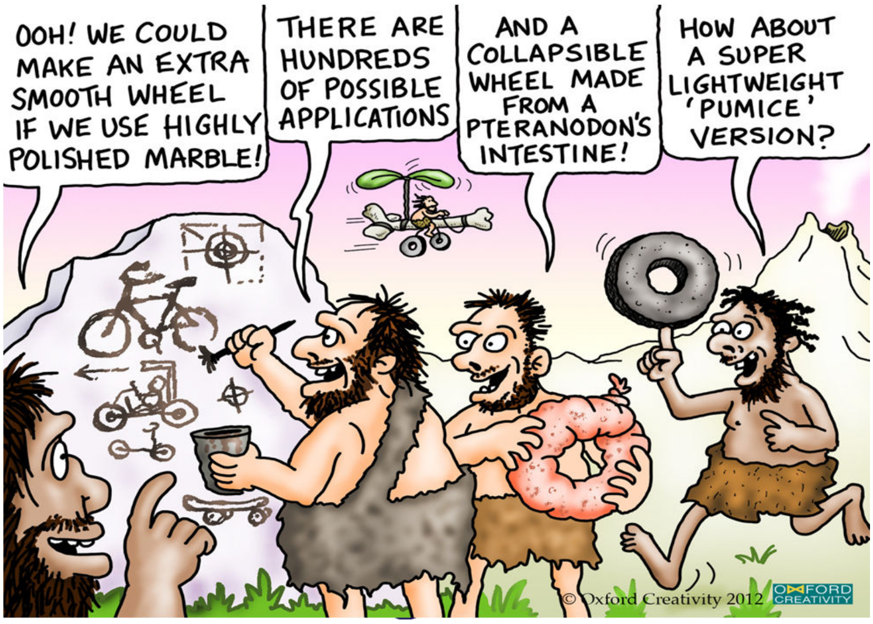
credit: oxford creativity, https://www.triz.co.uk/

Facilitate communication to foster innovation

© Oxford Creativity 2007
OXFORD CREATIVITY
LOOK WHAT I'VE INVENTED!
WHAT USE IS IT?
GRIND

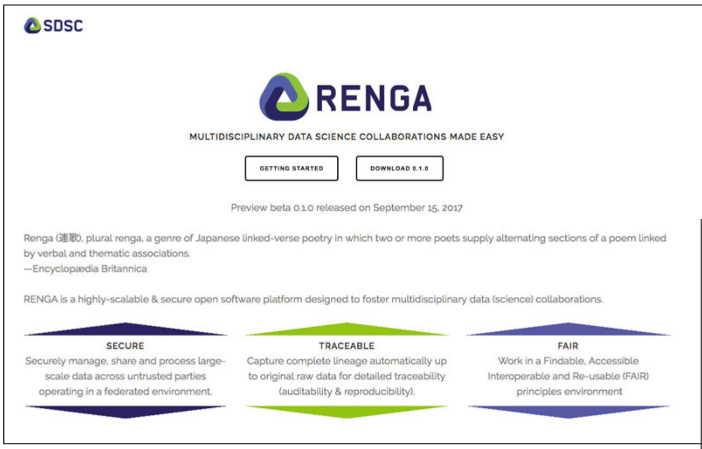
credit: oxford creativity, https://www.triz.co.uk/

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?
© Oxford Creativity 2012
OXFORD CREATIVITY

credit: oxford creativity, https://www.triz.co.uk/

Available as Open Source (Apache v2)



SDSC

RENGA

MULTIDISCIPLINARY DATA SCIENCE COLLABORATIONS MADE EASY

Preview beta 0.1.0 released on September 15, 2017

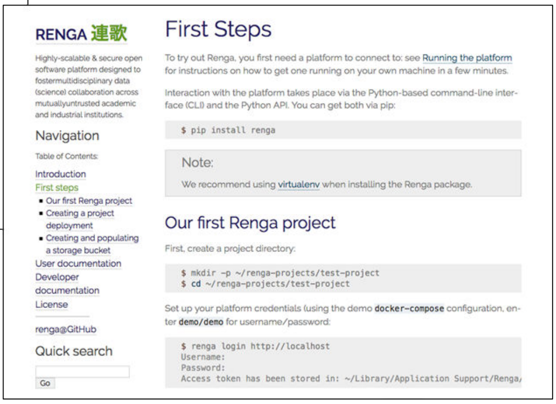
Renga (連歌), plural renga, a genre of Japanese linked-verse poetry in which two or more poets supply alternating sections of a poem linked by verbal and thematic associations.
—Encyclopædia Britannica

RENGA is a highly-scalable & secure open software platform designed to foster multidisciplinary data (science) collaborations.

SECURE
Securely manage, share and process large-scale data across untrusted parties operating in a federated environment.

TRACEABLE
Capture complete lineage automatically up to original raw data for detailed traceability (auditability & reproducibility).

FAIR
Work in a Findable, Accessible Interoperable and Re-usable (FAIR) principles environment



RENGA 連歌

Highly-scalable & secure open software platform designed to fostermultidisciplinary data (science) collaboration across mutuallyuntrusted academic and industrial institutions.

Navigation

Table of Contents:

- Introduction
- First steps
  - Our first Renga project
  - Creating a project deployment
  - Creating and populating a storage bucket
- User documentation
- Developer documentation
- License

renga@GitHub

Quick search

First Steps

To try out Renga, you first need a platform to connect to: see Running the platform for instructions on how to get one running on your own machine in a few minutes.

Interaction with the platform takes place via the Python-based command-line interface (CLI) and the Python API. You can get both via pip:

```
$ pip install renga
```

Note:

We recommend using virtualenv when installing the Renga package.

Our first Renga project

First, create a project directory:

```
$ mkdir -p ~/renga-projects/test-project
$ cd ~/renga-projects/test-project
```

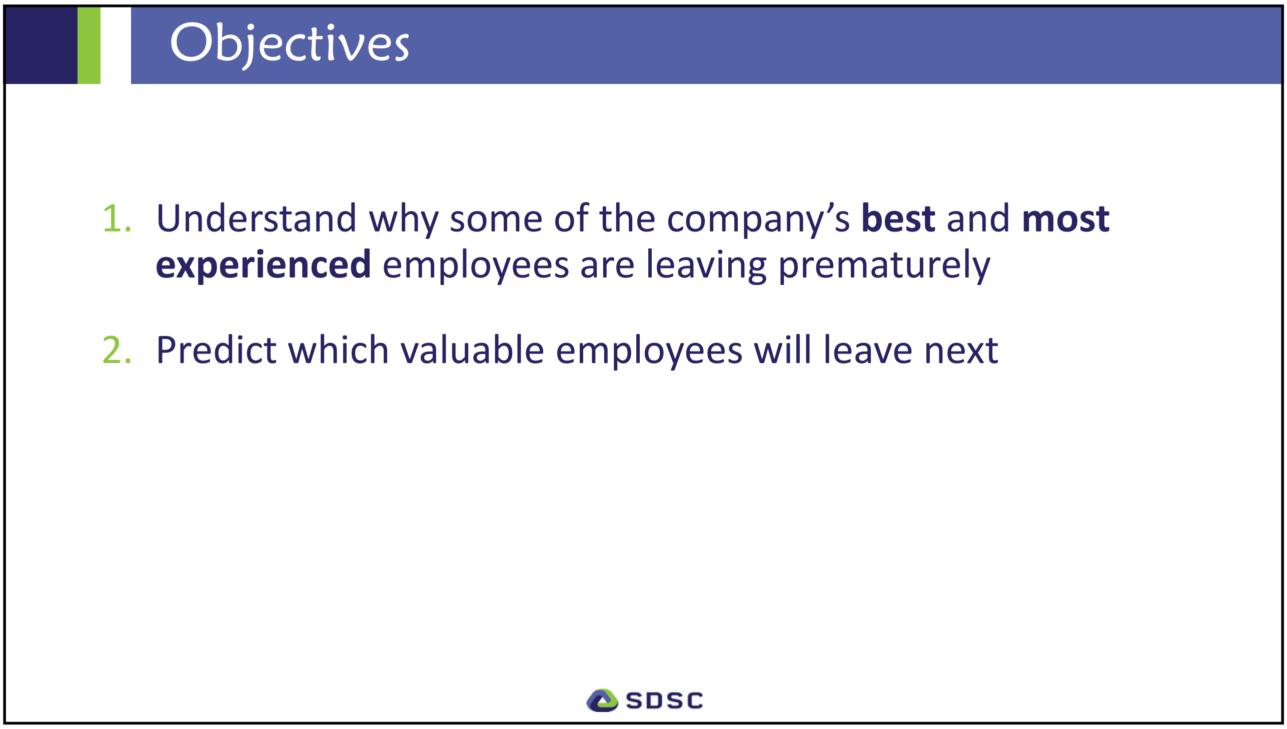
Set up your platform credentials (using the demo `docker-compose` configuration, enter `demo/demo` for username/password:

```
$ renga login http://localhost
Username:
Password:
Access token has been stored in: ~/Library/Application Support/Renga/
```

http://get-renga.io

SIMPLE USE CASE

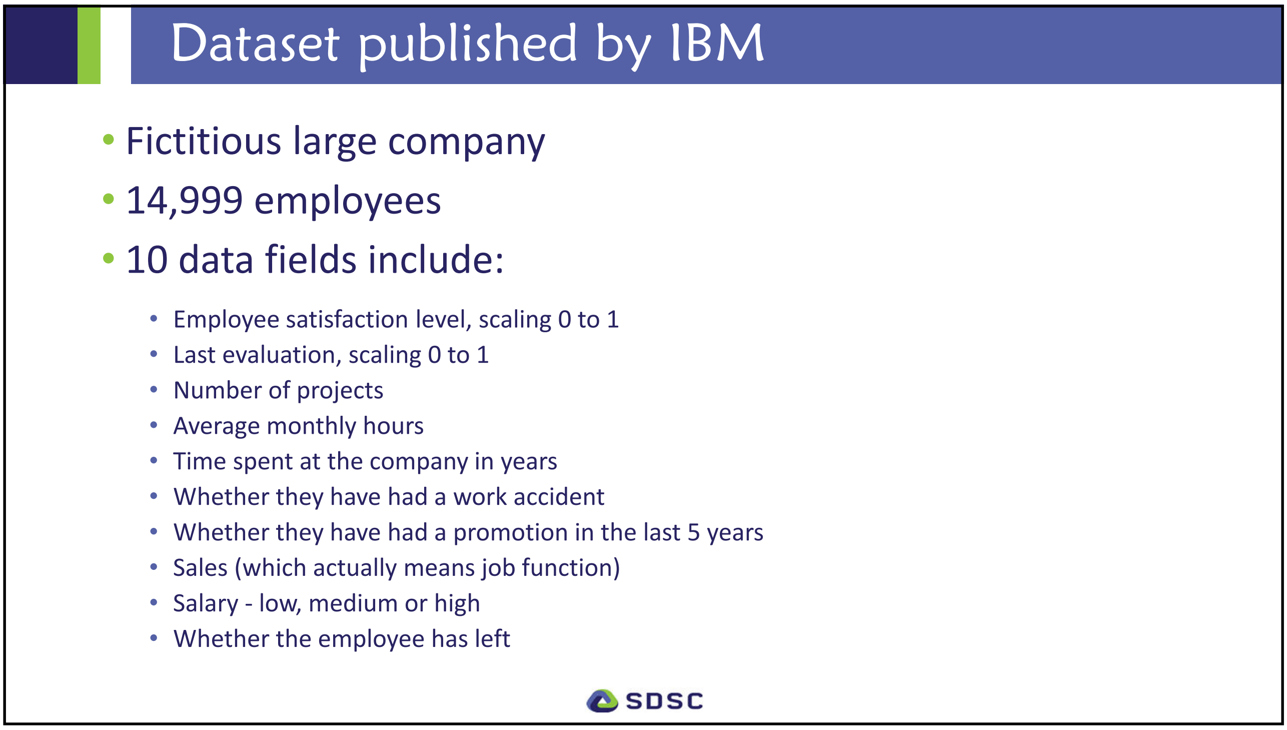
PREDICTING WHEN PEOPLE QUIT THEIR JOBS



Objectives

1. Understand why some of the company’s **best** and **most experienced** employees are leaving prematurely
2. Predict which valuable employees will leave next

SDSC

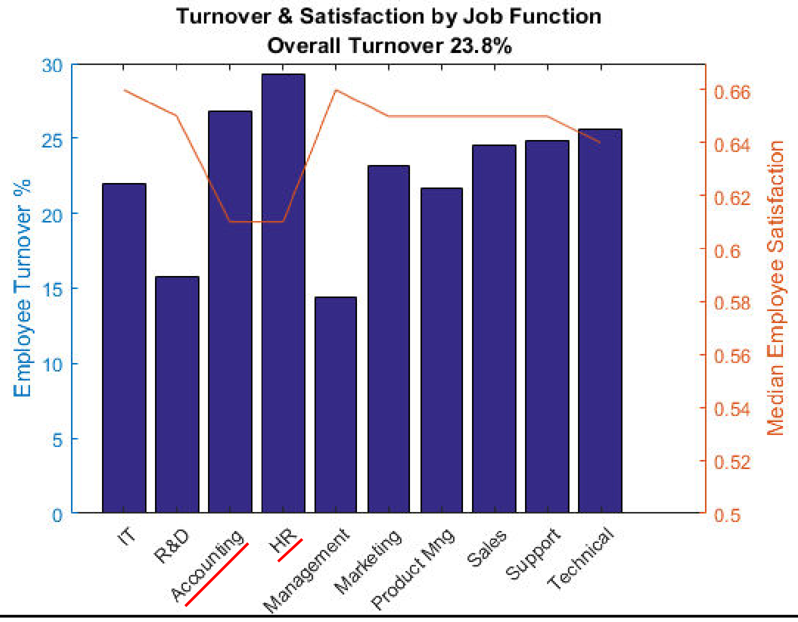


Dataset published by IBM

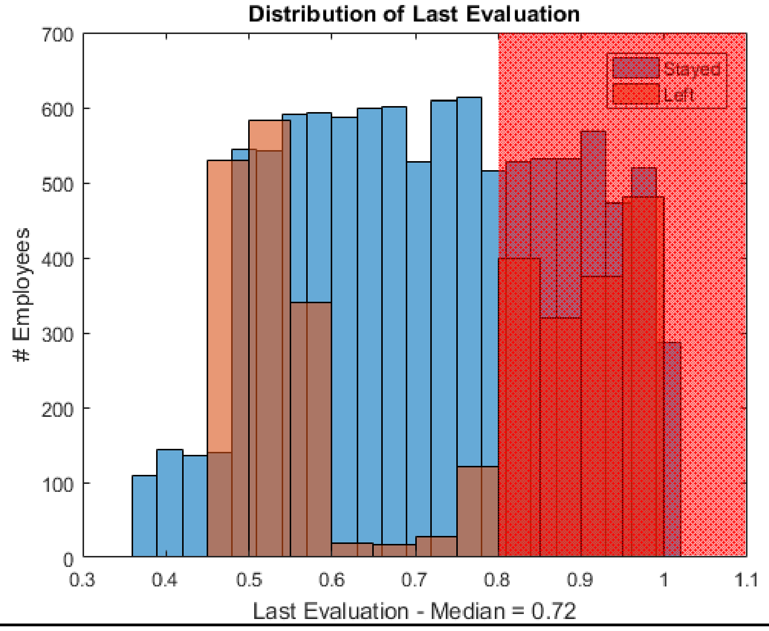
- Fictitious large company
- 14,999 employees
- 10 data fields include:
    - Employee satisfaction level, scaling 0 to 1
    - Last evaluation, scaling 0 to 1
    - Number of projects
    - Average monthly hours
    - Time spent at the company in years
    - Whether they have had a work accident
    - Whether they have had a promotion in the last 5 years
    - Sales (which actually means job function)
    - Salary - low, medium or high
    - Whether the employee has left

SDSC

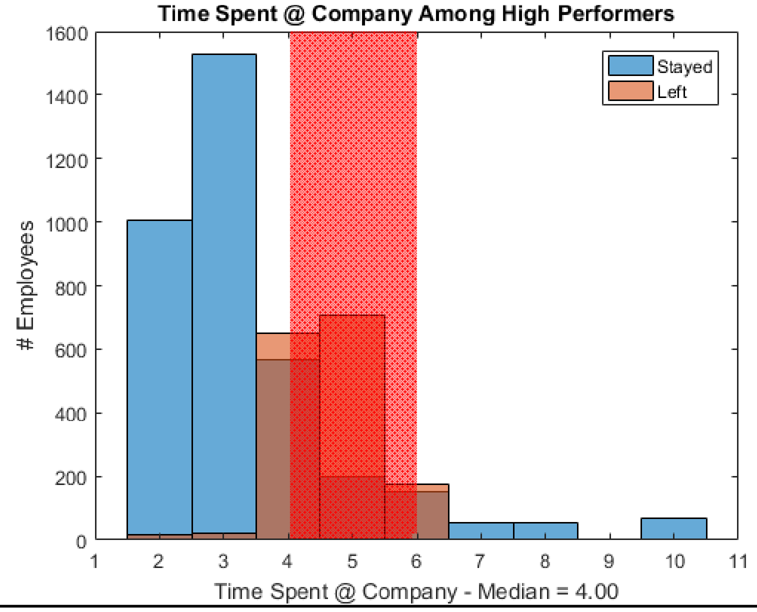
How Bad Is Turnover at This Company?



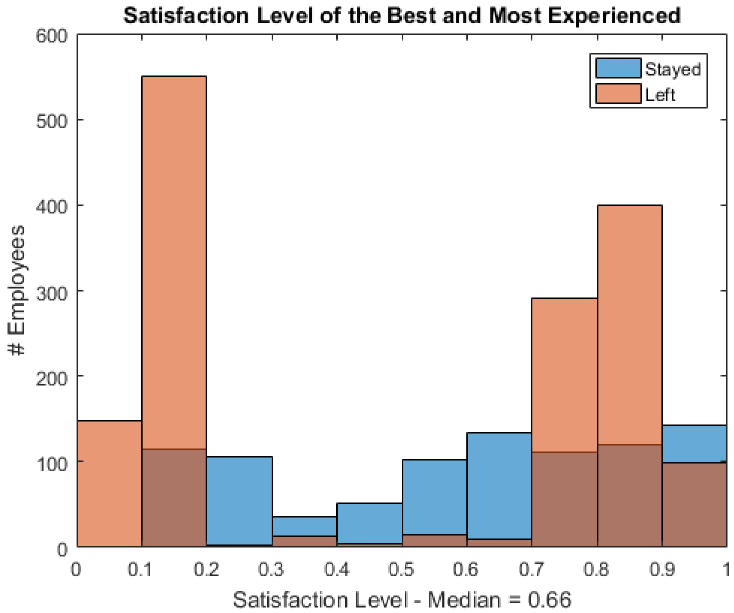
Defining Who Are the "Best"



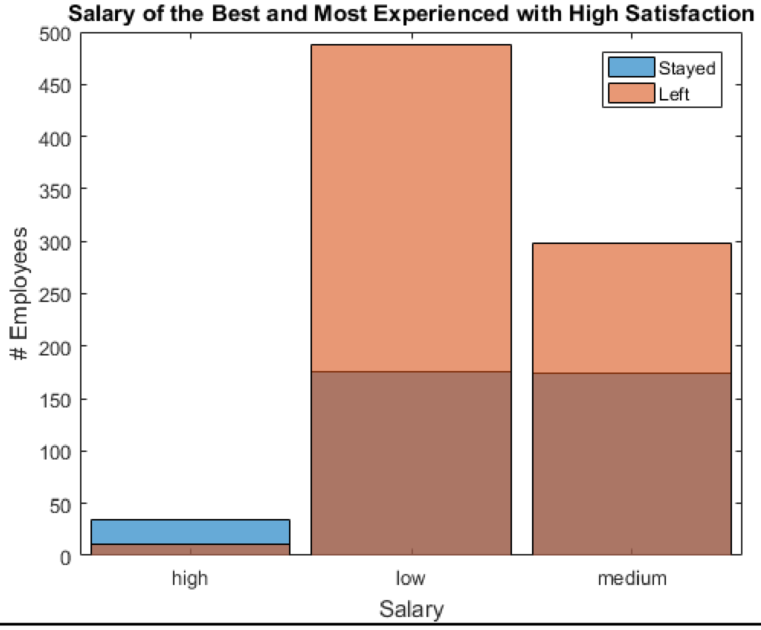
Defining Who Are the ”Most Experienced"



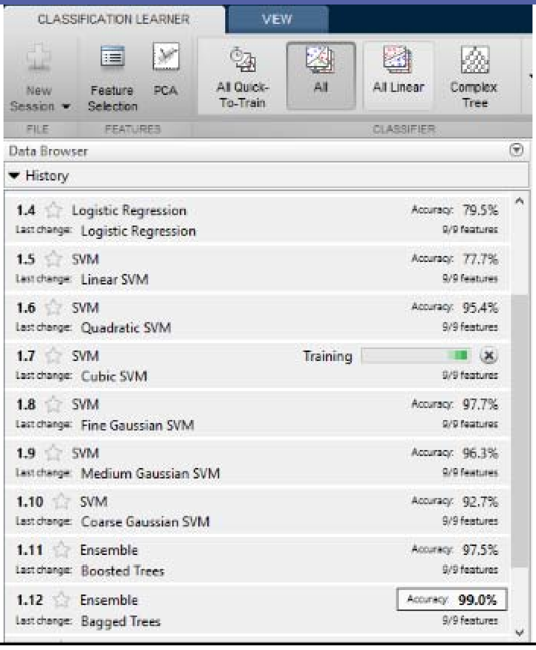
Job Satisfaction Among High Risk Group



Was It For Money?



Building a Predictive Model



1.4 ☆ Logistic Regression Accuracy: 79.5%
Last change: Logistic Regression 9/9 features

1.5 ☆ SVM Accuracy: 77.7%
Last change: Linear SVM 9/9 features

1.6 ☆ SVM Accuracy: 95.4%
Last change: Quadratic SVM 9/9 features

1.7 ☆ SVM Training
Last change: Cubic SVM 9/9 features

1.8 ☆ SVM Accuracy: 97.7%
Last change: Fine Gaussian SVM 9/9 features

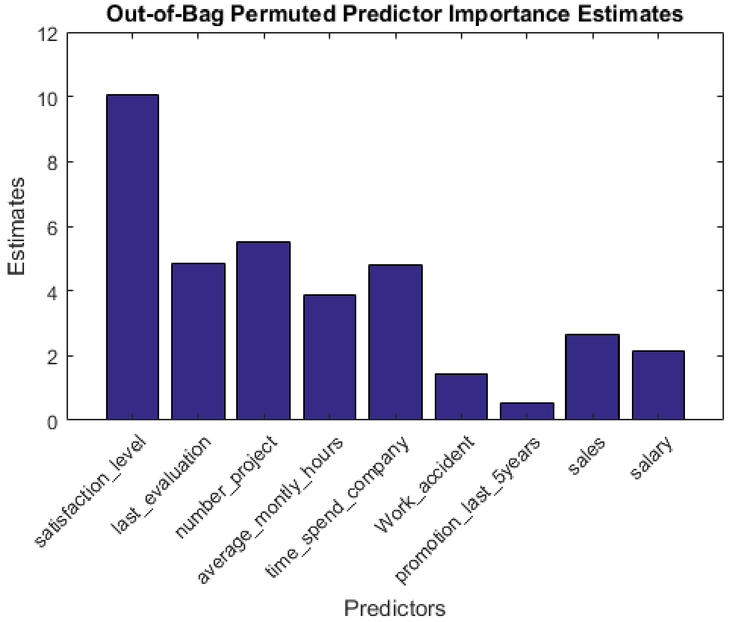
1.9 ☆ SVM Accuracy: 96.3%
Last change: Medium Gaussian SVM 9/9 features

1.10 ☆ SVM Accuracy: 92.7%
Last change: Coarse Gaussian SVM 9/9 features

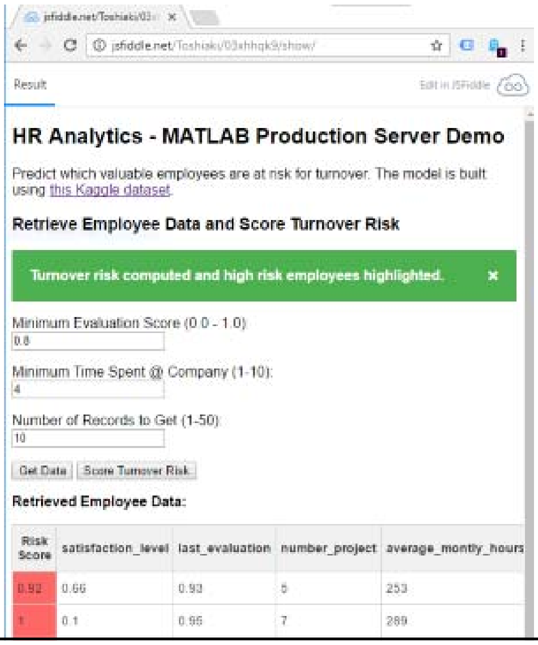
1.11 ☆ Ensemble Accuracy: 97.5%
Last change: Boosted Trees 9/9 features

1.12 ☆ Ensemble Accuracy: 99.0%
Last change: Bagged Trees 9/9 features

Explaining the Model



Operationalizing Action



HR Analytics - MATLAB Production Server Demo

Predict which valuable employees are at risk for turnover. The model is built using this Kaggle dataset.

Retrieve Employee Data and Score Turnover Risk

Turnover risk computed and high risk employees highlighted.

Minimum Evaluation Score (0.0 - 1.0)
0.8

Minimum Time Spent @ Company (1-10):
4

Number of Records to Get (1-50):
10

Retrieved Employee Data:

Risk Score	satisfaction_level	last_evaluation	number_project	average_montly_hours
0.92	0.66	0.93	5	253
1	0.1	0.95	7	289

THANK YOU!

http://www.datascience.ch

Twitter: @SDSCdatascience