From Lab to Factory

Or how to turn data into value?

PyData track at PyCon Ireland

Late October 2015

peadarcoyle@googlemail.com

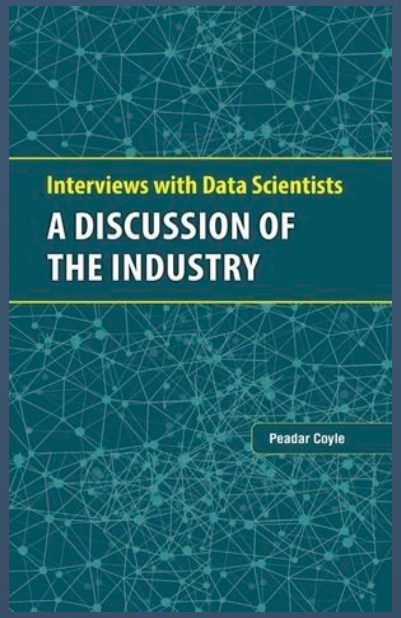
All opinions my own

Who am I?

- Type (A) data scientist - focused on analysis - c.f. Chang
- Masters in Mathematics
- Industry for nearly 3 years
- Specialized in Statistics and Machine Learning
- Passionate about turning data into products
- Occasional contributor to OSS - Pandas and PyMC3
- Speak and teach at PyData, PyCon and EuroSciPy
- @springcoil

Aims of this talk

- *"We need more success stories"* - Ian Ozsvald
- Lessons on how to deliver value quickly in a project
- Solutions to the last mile problem of delivering value



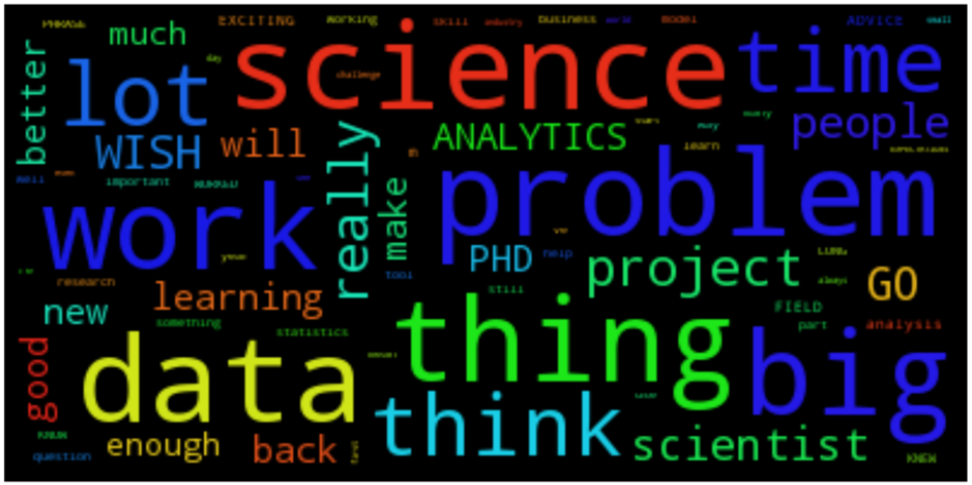
What IS a Data Scientist?

I think a data scientist is someone with enough programming ability to leverage their mathematical skills and domain specific knowledge to turn data into solutions.

The solution should *ideally* be a product

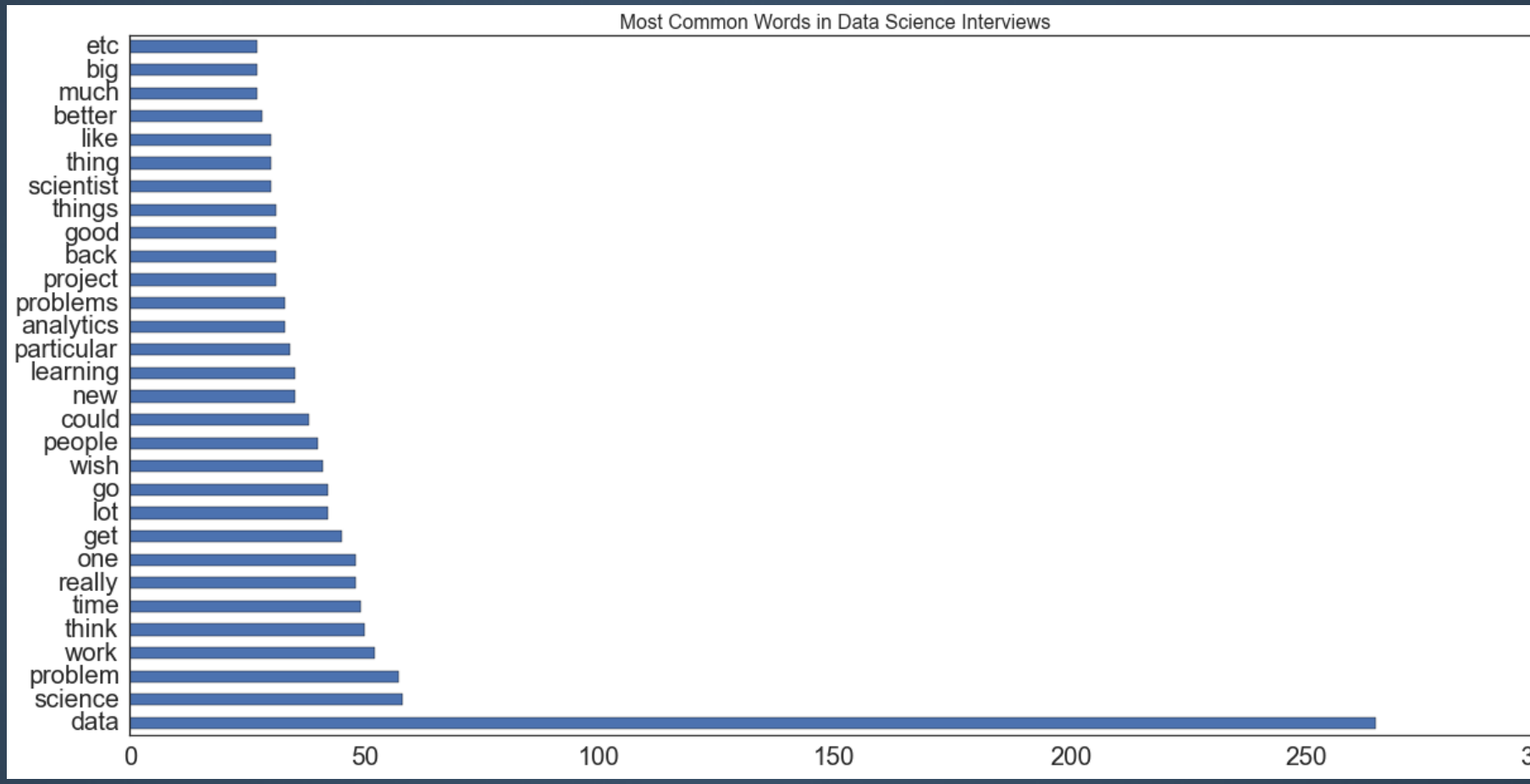
However even powerpoint can be the perfect delivery mechanism

What do Data Scientists talk about?



Based on my Dataconomy Interview series!

Some NLP on the Interviews!



HT: Sean J. Taylor and Hadley Wickham

How do I bring value as a data geek?

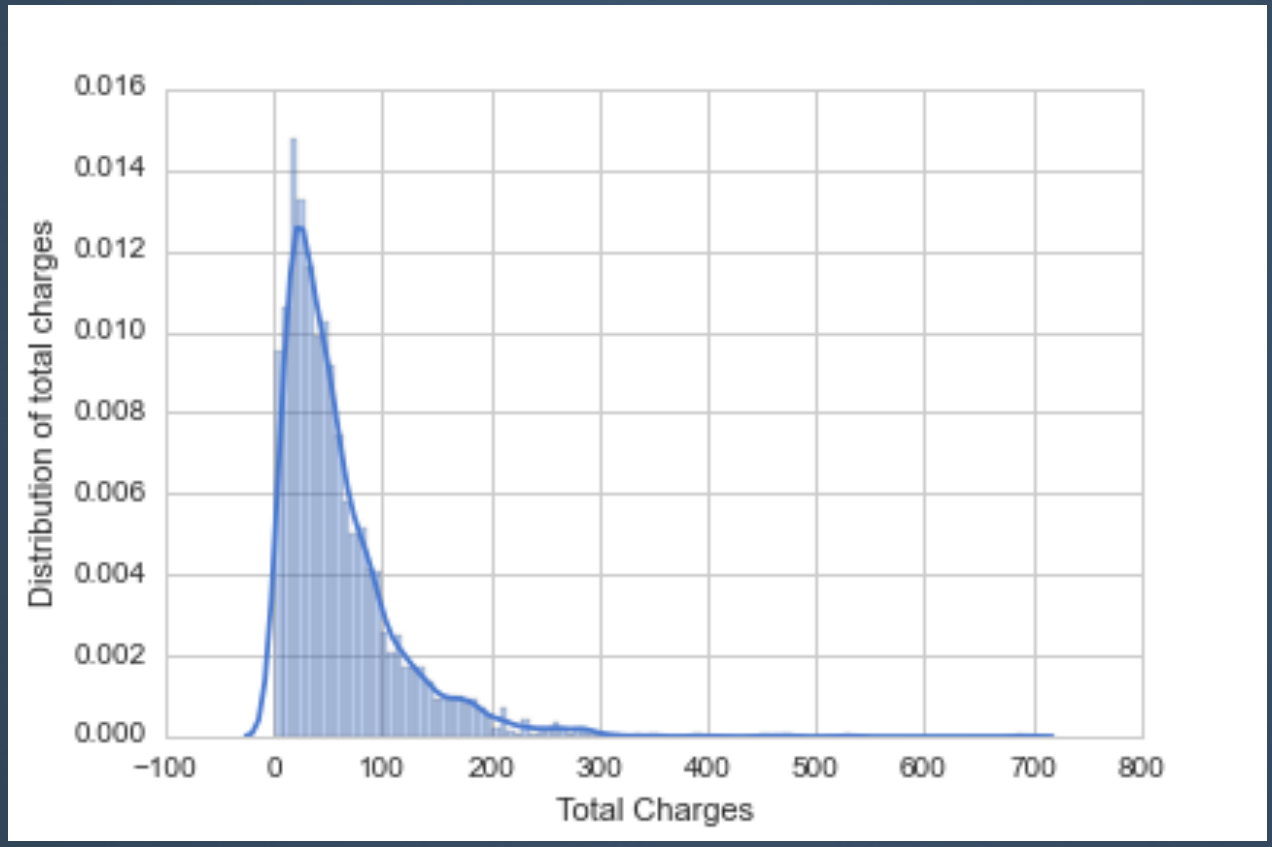
Getting models **used** is a hard problem (trust me :))

How do we turn **insight into action**?

How do we train people to trust models?

Visualise ALL THE THINGS!!

- (Relay foods dataset - HT Greg Reda)
- Consumer behaviour at a Fast Food Restaurant per year in the USA



What projects work?

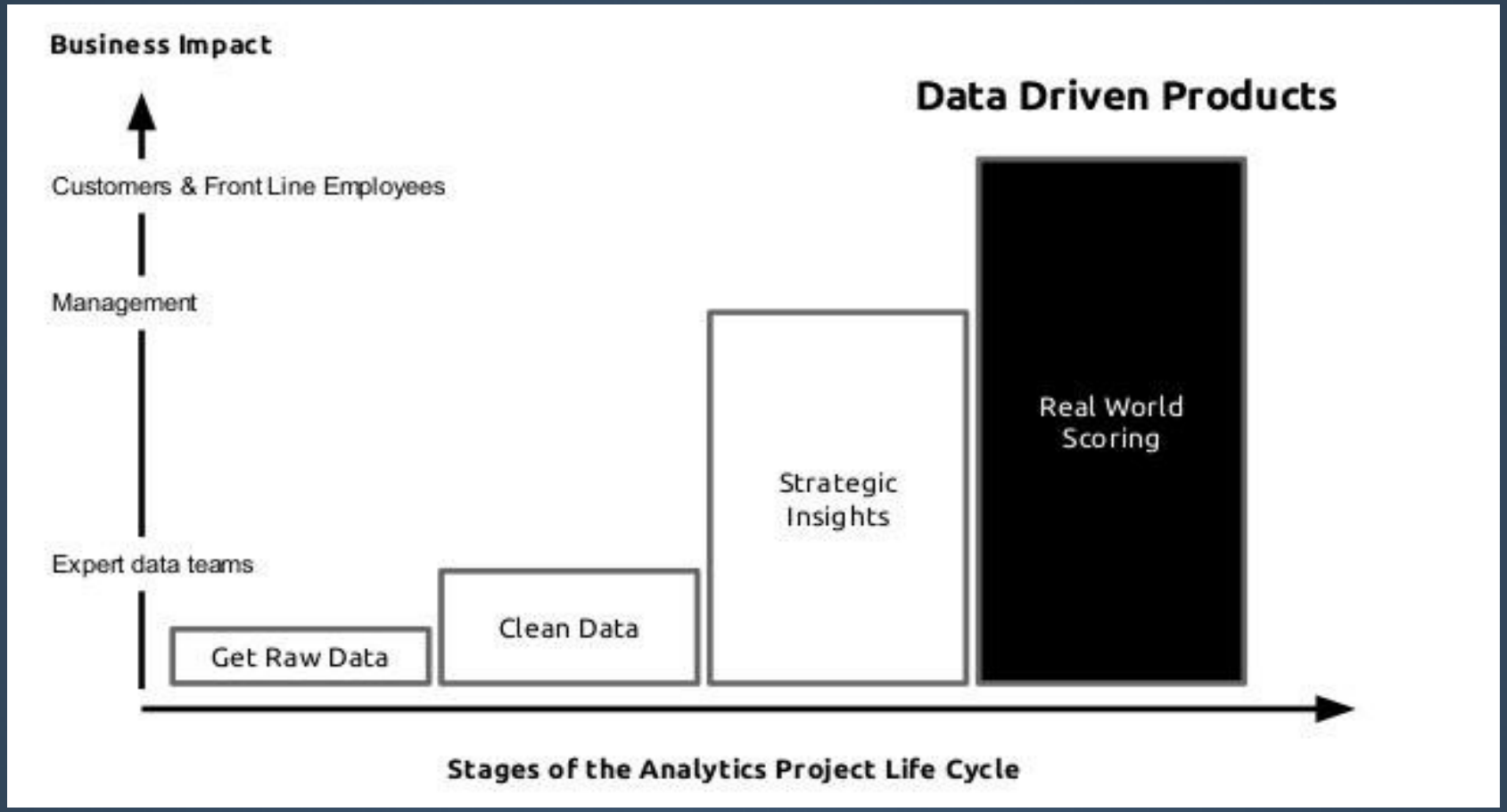
- Explaining existing data (visualization!)
- Automate repetitive/ slow processes
- Augment data to make new data (Search engines, ML models)
- Predict the future (do something more accurately than gut feel)
- Simulate using statistics :) (Rugby models, A/B testing)

Data Science projects are risky!

Many stakeholders think that data science is just an engineering problem, but research is **high risk and high reward**

Derisking the project - how? Send me examples :)

https://github.com/ianozsvald/data_science_delivered



(HT: The Yhat people - www.yhathq.com)

What are the blockers?

- Domain knowledge and understanding - can't be faked :)
- Difficult to extract information and produce good visualizations without engineering and business.
- Example - it took me months to be able to do good correlation analysis of Energy markets

"You need data first" - Peadar Coyle

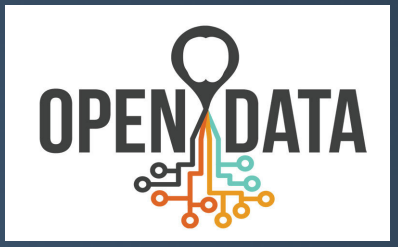
- Copying and pasting PDF/PNG data
- Messy csv files and ERP output
- Scale?
- Getting data in some areas is hard!!
- Months for extraction!!!
- Some tools for web data extraction
- Messy APIs without documentation :(

Augmenting data and using API's

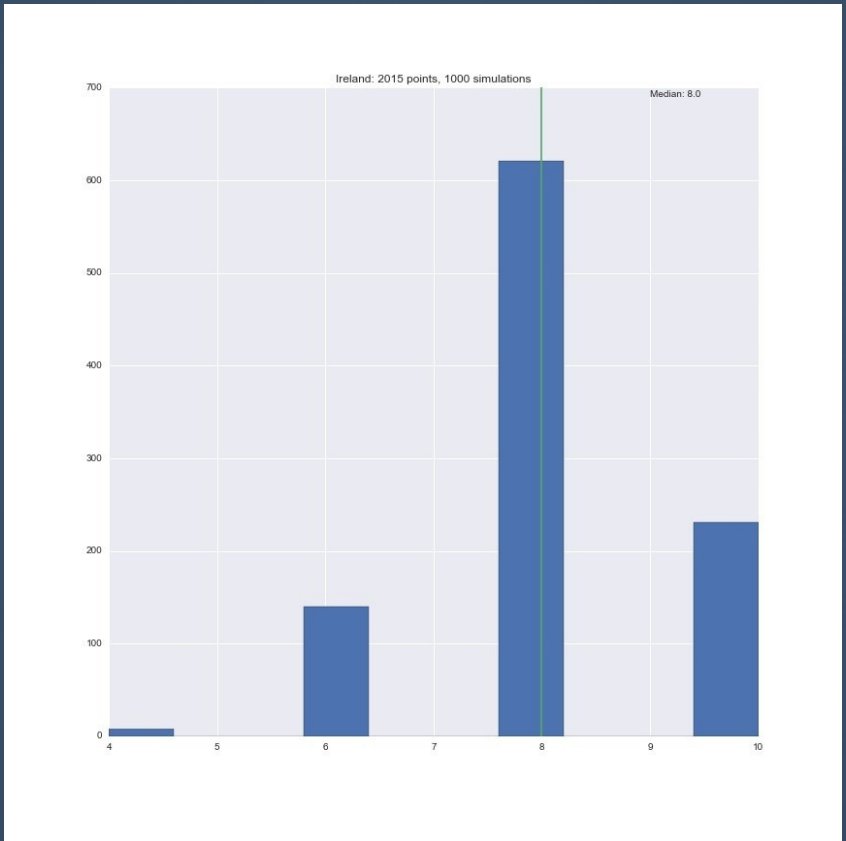
Sentiment analysis

Improving risk models with data from other sources like Quandl

Air Traffic data blend - many many API's.



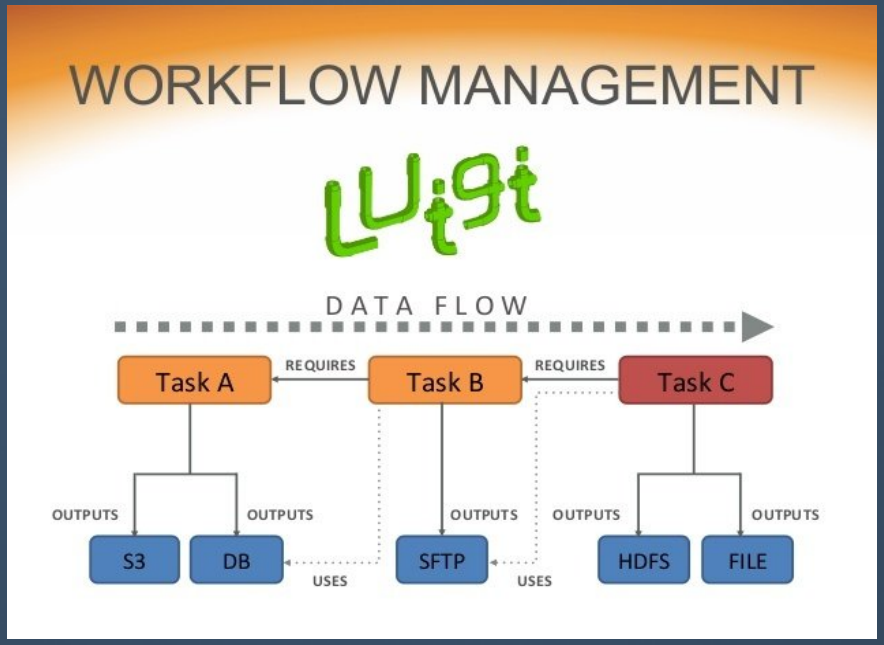
Simulate: Six Nations with MCMC (PyMC3)



Machine Learning (HT: Ian Ozsvald)

Models are a small part of a problem

- Only 1% of your time will be spent modelling
- Stakeholder engagement, managing people and projects
- Data pipelines and your infrastructure matters - Eoin Brazil Talk
- How is your model used? How do you get adoption?



Lessons learned from Lab to Factory

1. The 'magic quickly' problem is a big problem in any data science project - our understanding of time frames and risk is unrealistic :)

2. Lack of a shared language between software engineers and data scientists - but investing in the right tooling by using open standards allows success.

3. To help data scientists and analysts succeed your business needs to be prepared to **invest in tooling**

4. Often you're working with other teams who use different languages - so micro services can be a good idea

How to deploy a model?

- Palladium (Otto Group)
- Azure
- Flask Microservice
- Docker

Invest in tooling

- For your analysts and data scientists to succeed you need to invest in infrastructure to empower them.
- Think carefully how you want your company to spend its **innovation tokens** and take advantage of the excellent tools available like ScienceOps and AWS.
- I think there is great scope for entrepreneurs to take advantage of this **arbitrage opportunity** and build good tooling to empower data scientists by building platforms.
- Data scientists need better tools :) For all parts of the process :)

Data Product Development

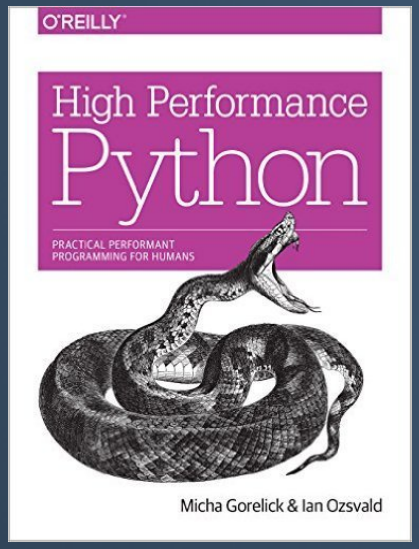
- Software Engineers aren't data scientists and shouldn't be expected to write models in code.
-
- **A high value use of models is having them in production**

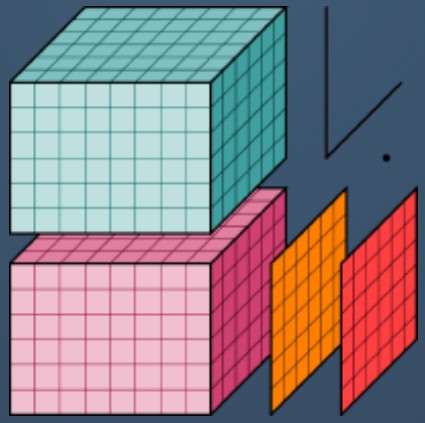
- Getting information from stakeholders is really valuable in improving models.

(I gave a talk on Data Science models in Production using Yhat tech)

Use small data where possible!!

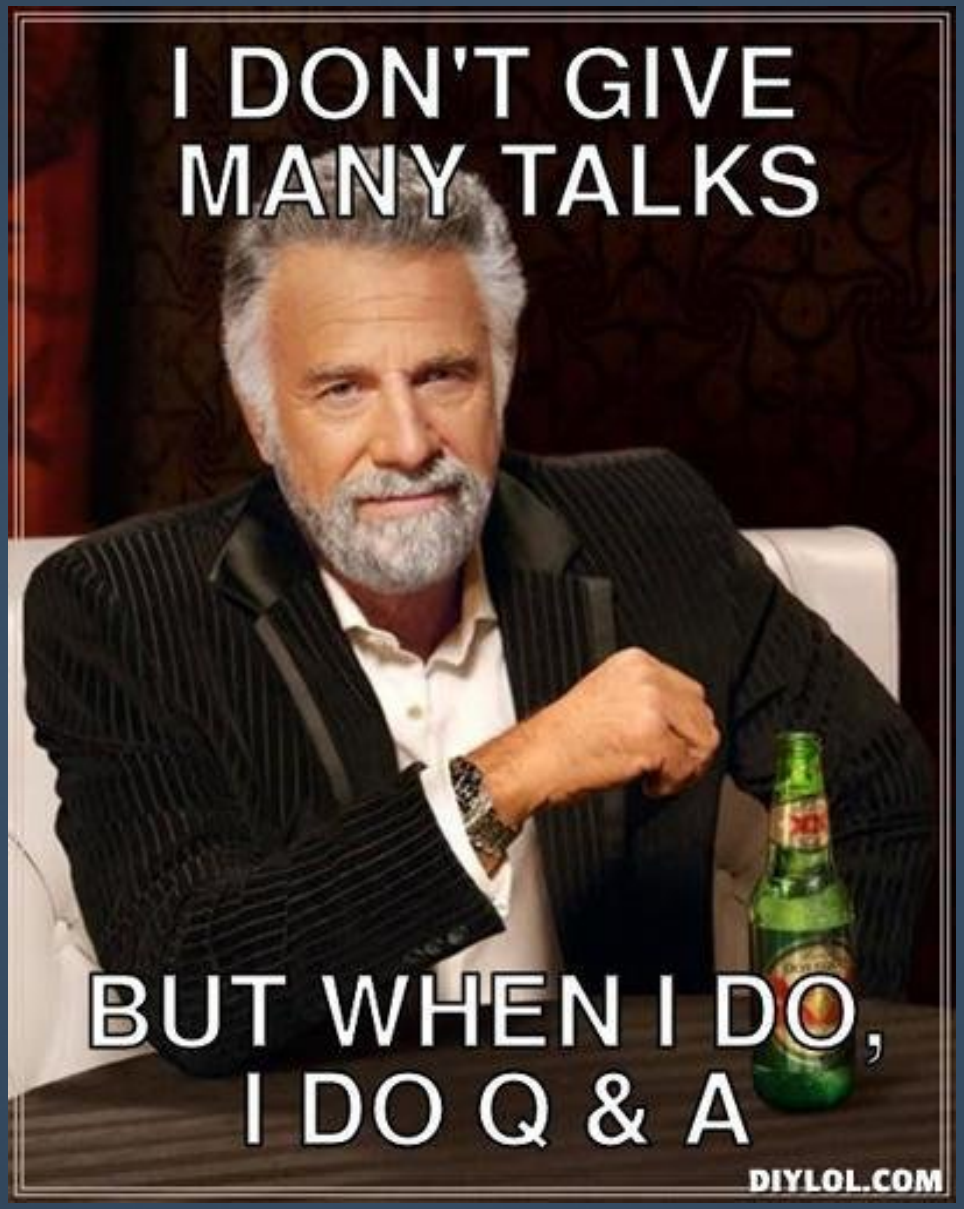
- Small problems with clean data are more important - (Ian Ozsvald)
- Amazon machine with many Xeons and 244GB of RAM is less than 3 euros per hour. - (Ian Ozsvald)
- Blaze, Xray, Dask, Ibis, etc etc - PyData Bikeshed
- "The mean size of a cluster will remain 1" - Matt Rocklin





Closing remarks

- Dirty data stops projects
- There are some good projects like Icy, Luigi, etc for transforming data and improving data extraction
- These tools are still not perfect, and they only cover a small amount of problems
- Stakeholder management is a challenge too
- Come speak in Luxembourg
- *It isn't what you know it is who you know...*
- On Dataconomy I did a series of interviews with Data Scientists
- Send me your dirty data and data deployment stories :)
- My website


I DON'T GIVE
MANY TALKS
BUT WHEN I DO,
I DO Q & A
DIYLOL.COM

What is the Data Science process?

Obtain

Scrub

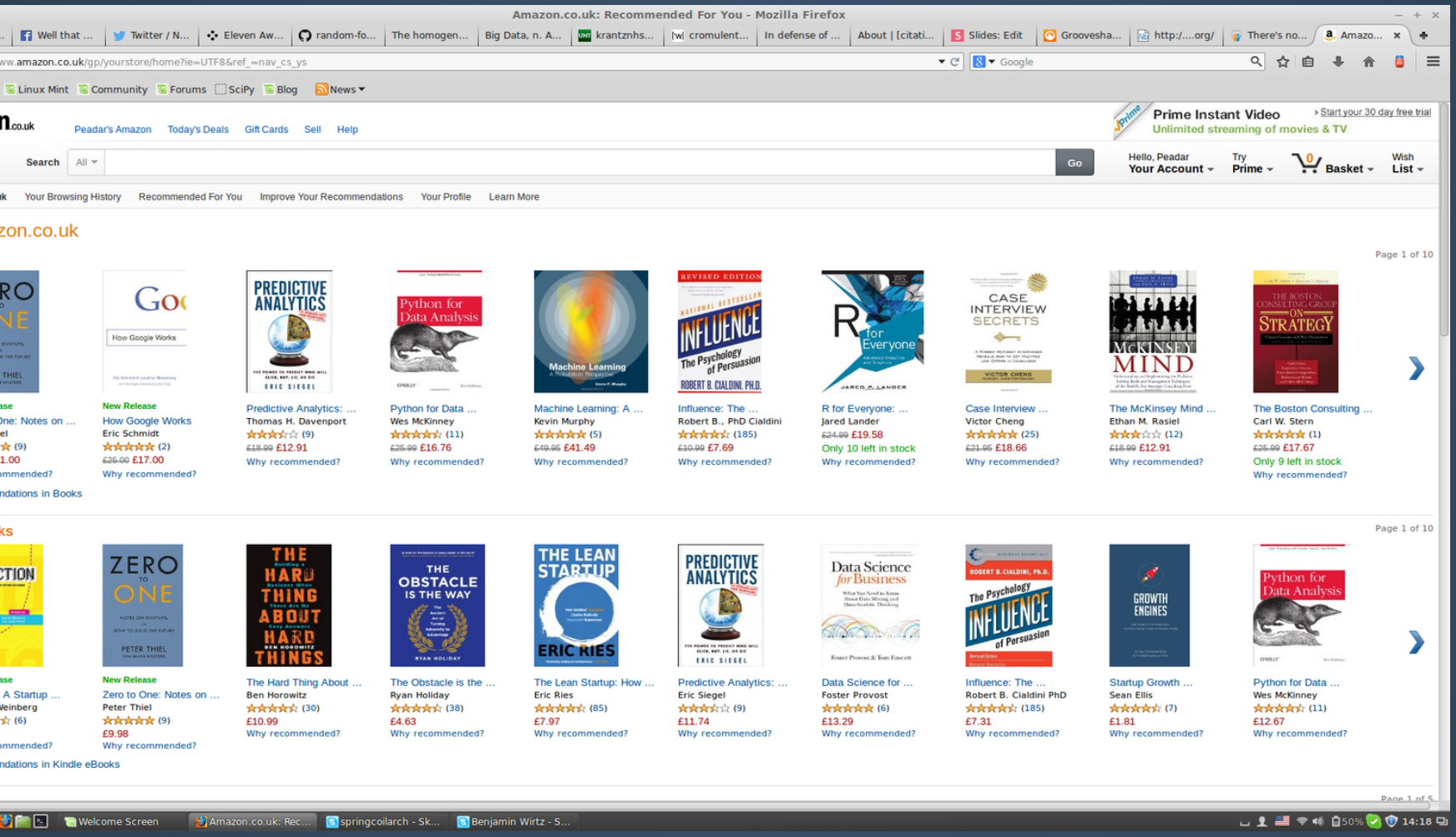
Explore

Model

Interpret

Communicate (or Deploy)

A famous 'data product' - Recommendation engines




ONE DOES NOT SIMPLY
BUILD A MODEL IN ONE DAY