Data Science Toolbox:

How bad data science can destroy a good business

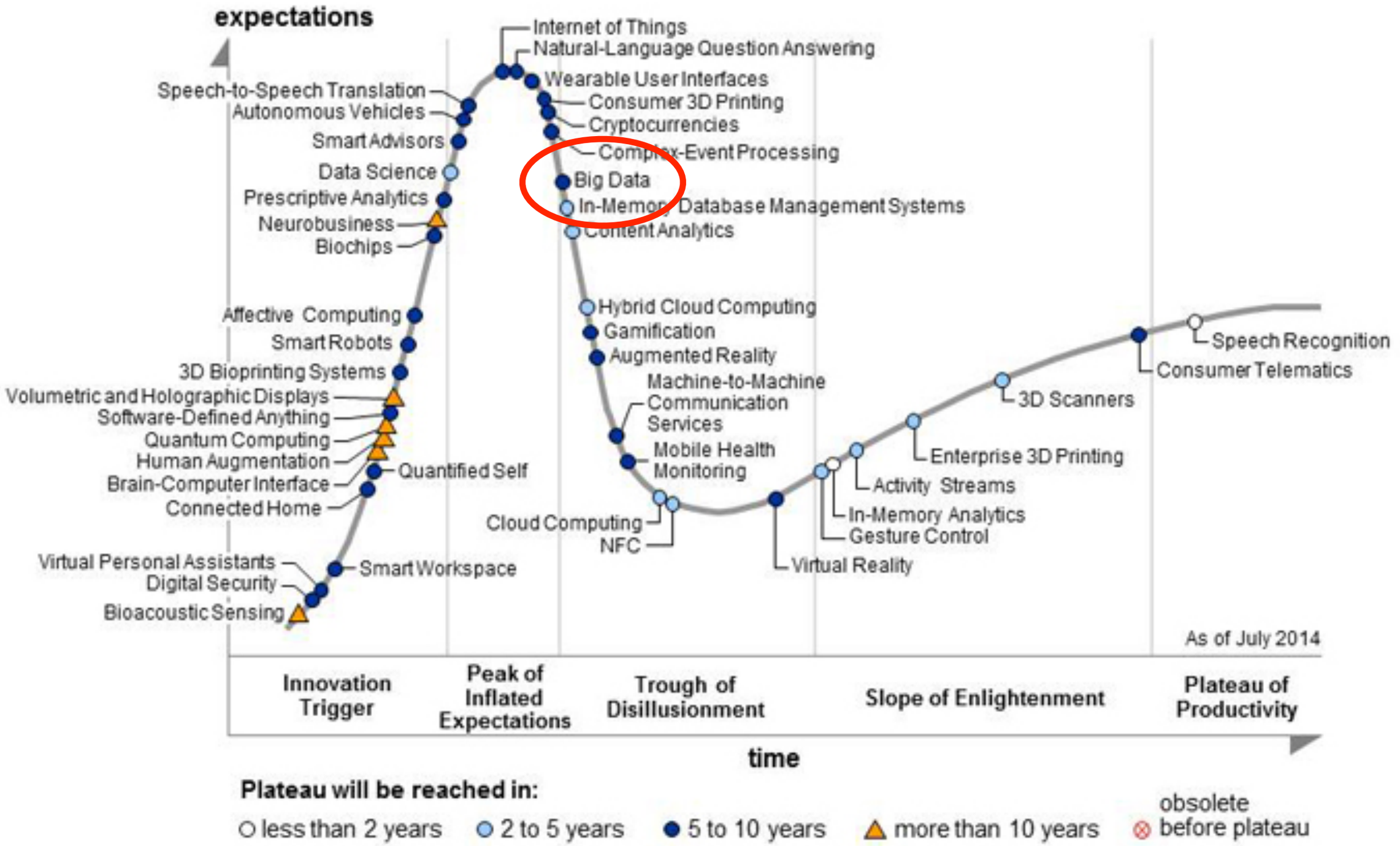
Leonid Danilchenko

About me

Center of Mathematical Finance
2015/2016

Optimum Media Direction
Media Research

Gartner's 2015 Hype Cycle



Gartner's 2016 Hype Cycle

Source: Gartner (July 2016)


ХАЛЯВА
risovach.ru

Халява:

1. Atom — Текстовый редактор 21 века. Вообще он бесплатный, но в списке присутствует.
2. AWS Education — 75-150$ на сервера амазона (!!!)
3. Crowdflower — Краудсорсинговая платформа (экономия составляет $2,500 в месяц за доступ и $50)
4. DigitalOcean — Облачный хостинг ($100 кредита на аккаунт)
5. DNSimple — DNS-сервис (бесплатно на 2 года, вместо $3 каждый месяц)
6. GitHub — Сервис Git-репозиториев (микроаккаунт $7) (!!!)
7. Microsoft Azure — Облачная технология от Microsoft (!!!)
8. Namecheap — Регистратор доменных имен (.me домен бесплатно, обычная стоимость $8.99 в год) и ssl-сертификатов (бесплатно, стоимость $9 в год)
9. Orchestrate — Поиск, геолокации, БД на основе графов и API (девелоперский аккаунт за $49 в подарок)
10. Screenhero — Общий доступ к скриншотам для работы в команде (обычная цена $9.99 в месяц)
11. SendGrid — Сервис для работы с email (обычная цена $4.95 в месяц)
12. Stripe — Веб- и мобильные платежи для разработчиков (первые $1000 без комиссии)
13. Travis CI — Сервис непрерывной интеграции (обычно в месяц $69)
14. Unreal Engine — без комментариев (что позволяет сэкономить $19 каждый месяц)

Source: https://education.github.com/pack

GitHub

Soumith Chintala
soumith

```
/\~~/\ _____________________
\0_0 /~\/ _________________
 \__~~/
```

Follow

Block or report user

Facebook AI Research
New York, USA
http://soumith.ch
Joined on Jan 7, 2012

Organizations

Overview | Repositories 121 | Stars 386 | Followers 1.6k | Following 282

Popular repositories

convnet-benchmarks
Easy benchmarking of all publicly accessible implementations of convnets
★ 1,532 Python

cvpr2015
★ 474 Jupyter Notebook

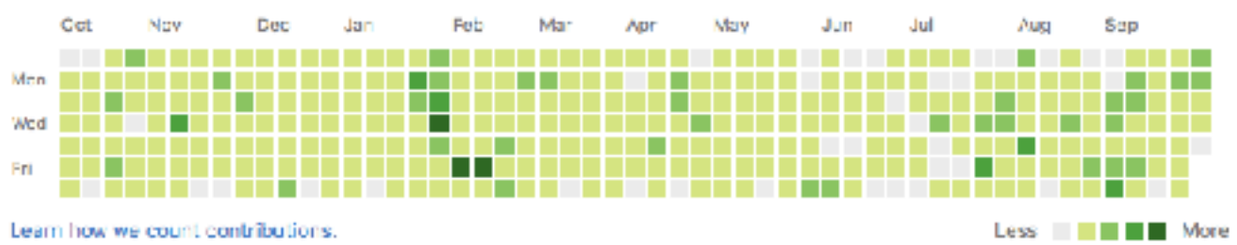
dcgan.torch
A torch implementation of http://arxiv.org/abs/1511.06434
★ 316 Lua

cudnn.torch
Torch-7 FFI bindings for NVIDIA CuDNN
★ 231 Lua

imagenet-multiGPU.torch
an imagenet example in torch.
★ 212 Lua

torch-android
Torch-7 for Android
★ 148 CMake

3,050 contributions in the last year



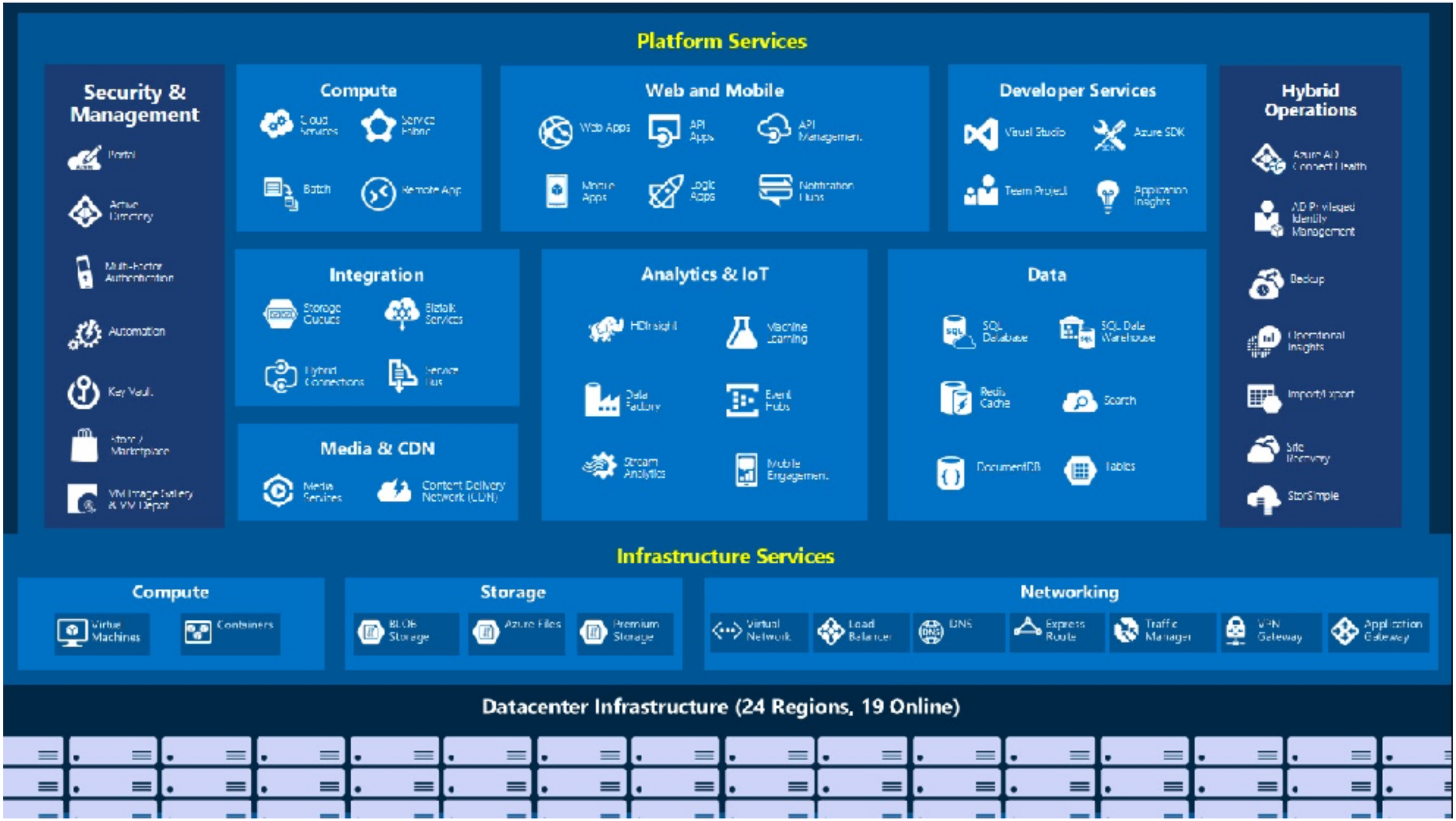
Халява:

1. MS Office 365 — пакет офиса на 3 года
2. Microsoft Azure — облачная платформа. всего очень много
3. Windows Client
4. Visual Studio Premium
5. Visual Studio Ultimate
6. Visio
7. Project
8. OneNote
9. BizTalk Server
10. SharePoint Serve

И многое другое

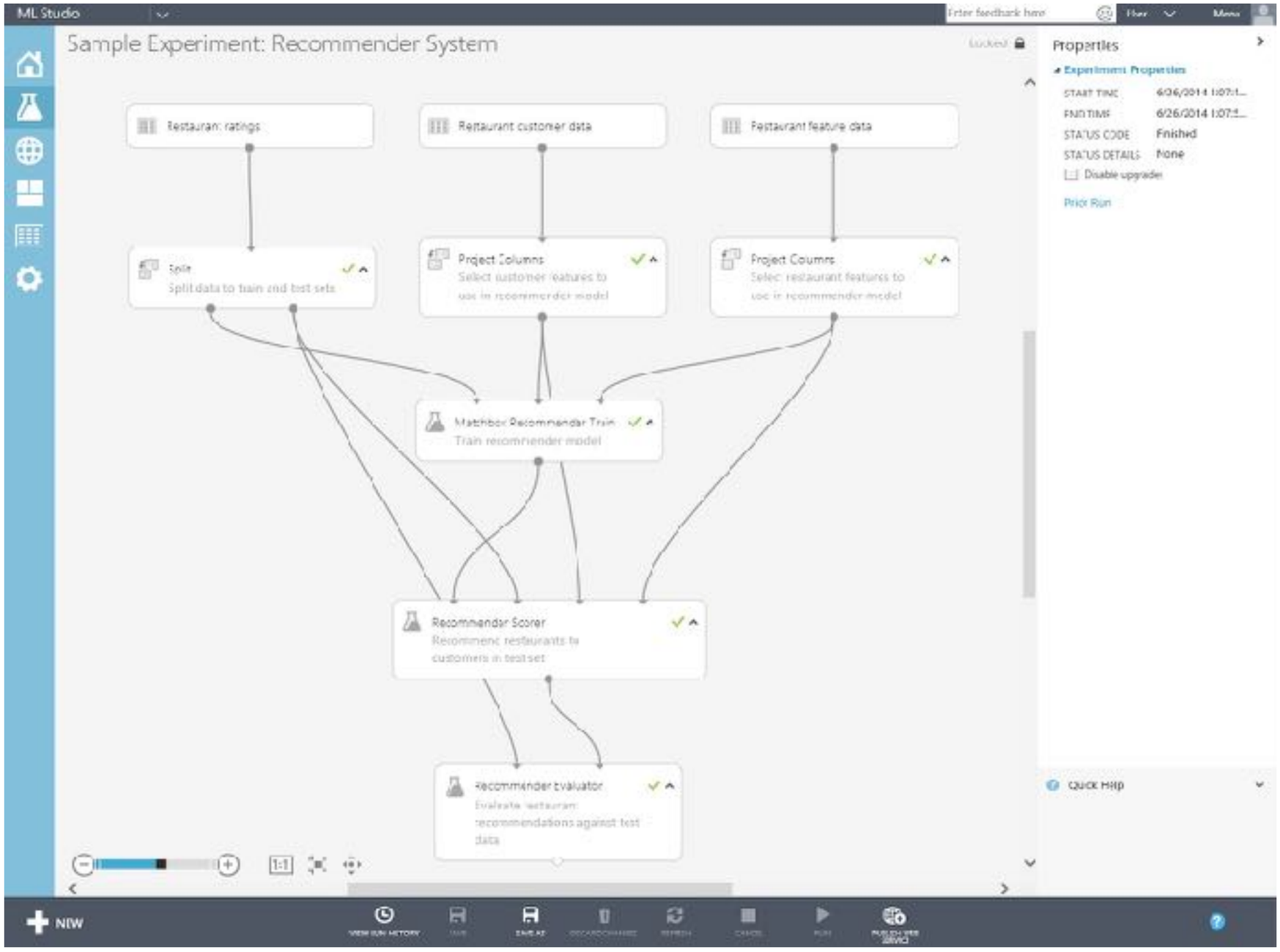
Source: https://catalog.imagine.microsoft.com/en-us

Microsoft Azure Infrastructure

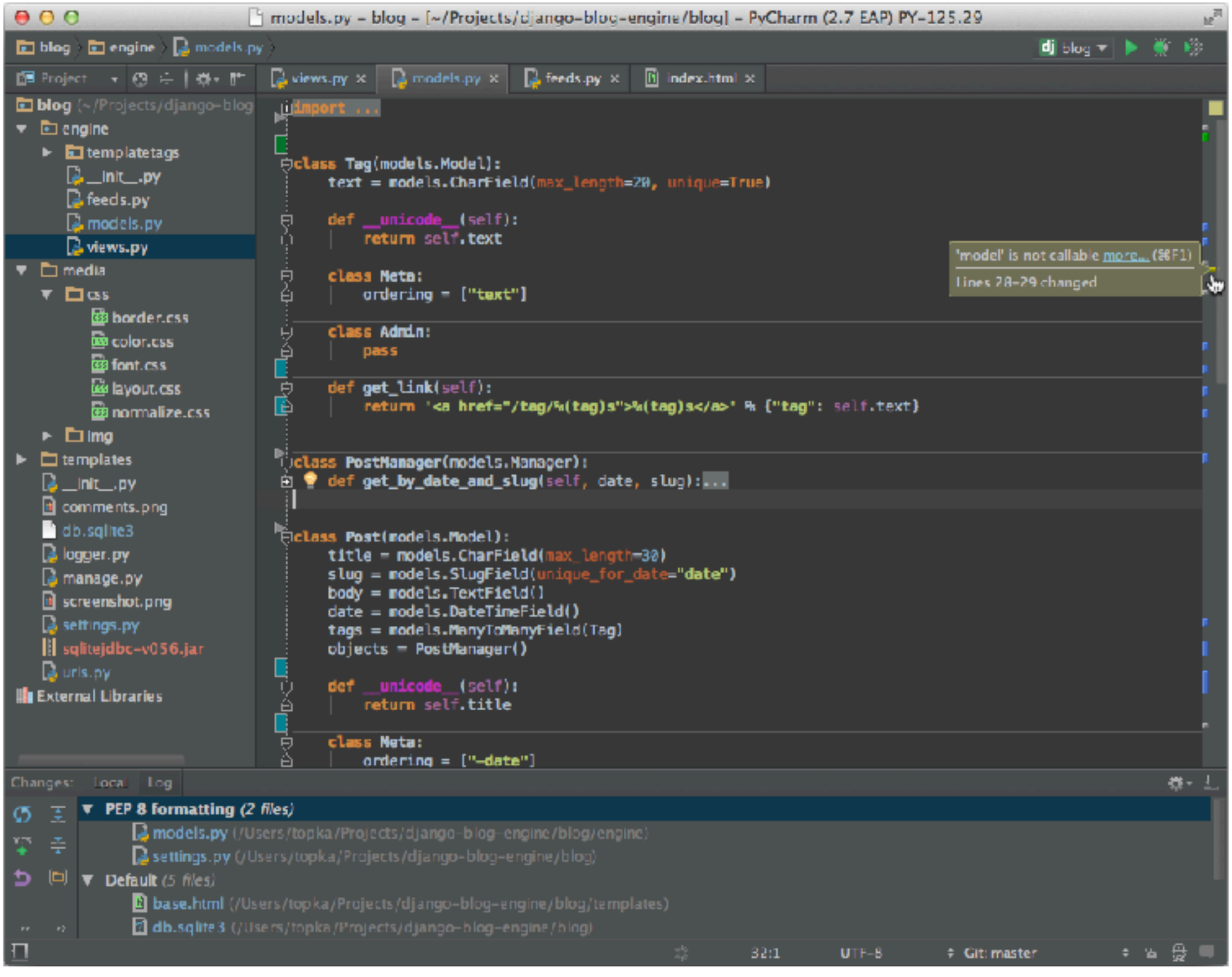


Microsoft Azure

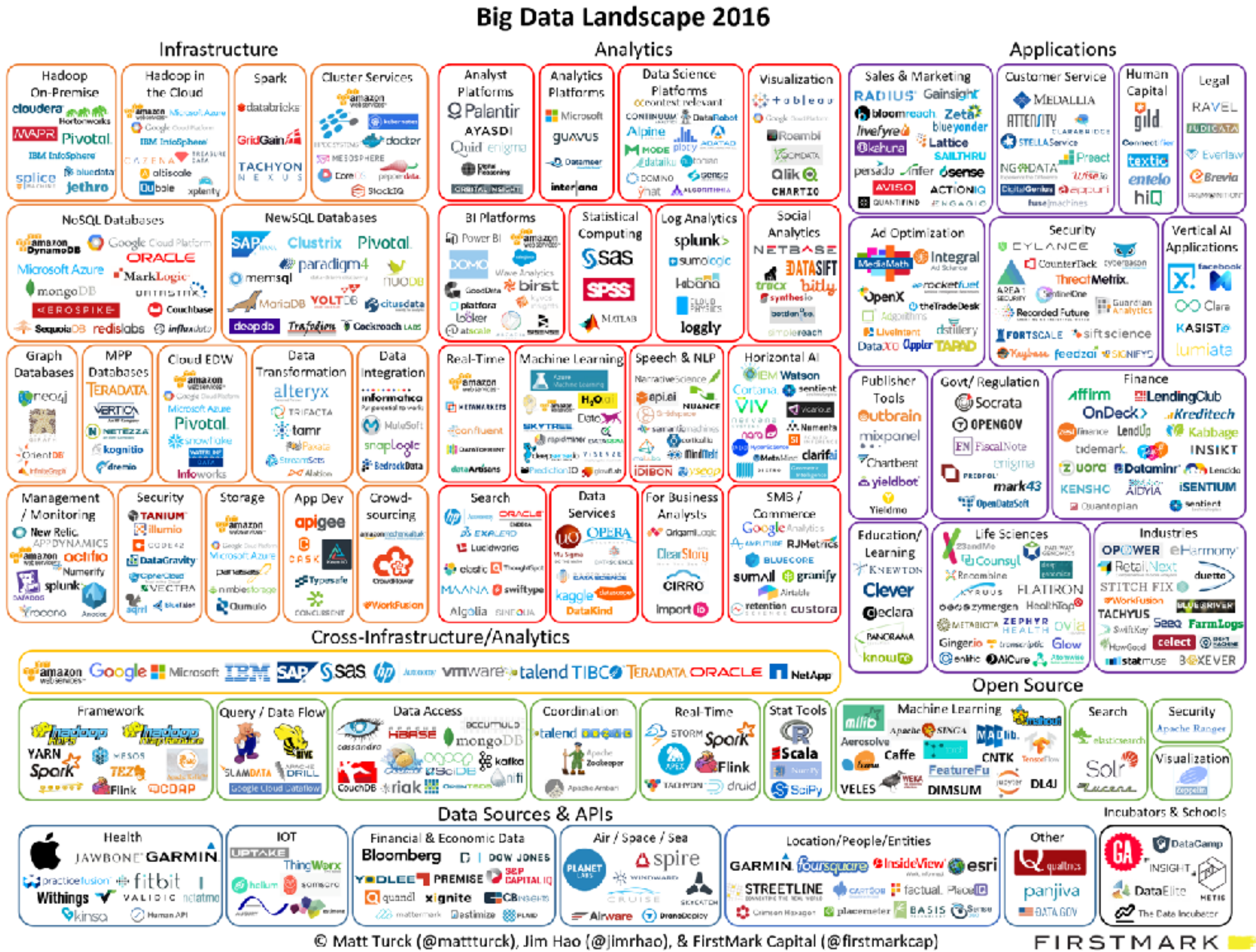
Microsoft Azure ML Studio



PyCharm



Source: https://www.jetbrains.com/student/


Big Data Landscape 2016
Infrastructure
Analytics
Applications
Hadoop On-Premise
Hadoop in the Cloud
Spark
Cluster Services
Analyst Platforms
Analytics Platforms
Data Science Platforms
Visualization
Sales & Marketing
Customer Service
Human Capital
Legal
NoSQL Databases
NewSQL Databases
BI Platforms
Statistical Computing
Log Analytics
Social Analytics
Ad Optimization
Security
Vertical AI Applications
Graph Databases
MPP Databases
Cloud EDW
Data Transformation
Data Integration
Real-Time
Machine Learning
Speech & NLP
Horizontal AI
Publisher Tools
Govt/ Regulation
Finance
Management / Monitoring
Security
Storage
App Dev
Crowd-sourcing
Search
Data Services
For Business Analysts
SMB / Commerce
Education/ Learning
Life Sciences
Industries
Cross-Infrastructure/Analytics
Open Source
Framework
Query / Data Flow
Data Access
Coordination
Real-Time
Stat Tools
Machine Learning
Search
Security
Visualization
Data Sources & APIs
Incubators & Schools
Health
IOT
Financial & Economic Data
Air / Space / Sea
Location/People/Entities
Other
© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)
FIRSTMARK

Data Science Toolkit:

Visualization:
ggplot2, matplotlib, D3, GraphViz

Modeling:
Python, R, Scala, C++, Java

Reporting:
PowerBI, Shiny, Tableau, Zeppelin

Deep Learning:
Theano, CUDA, ThensorFlow

Hadoop:
Hadoop, Cloudera, Amazon EMR, Microsoft Azure

Sharing:
Git(GitHub/GitLab/etc), SVN

SQL for Hadoop:
Hive, Spark, Pig

One Love:
Excel, PowerPoint

and more..


НЕКОТОРЫЕ ЛЮДИ ДУМАЮТ, ЧТО УЧЕНЫЕ ГОВОРЯТ
ЭВРИКА!
КОГДА ЭКСПЕРИМЕНТИРУЮТ...
НО, СКОРЕЕ ВСЕГО, ОНИ СКАЖУТ...
ЧУШЬ!
ВОТ БЛИН!
Б*Я!!!
ЧЕРТ!
ГРЕБАНАЯ ЖЕЛЕЗЯКА!
Я НЕНАВИЖУ НАУКУ
twistedtoodles.com

[Some] Data Science Principles:

- approximately **80% of the costs** for data-related project get spent on data preparation - data is always dirty. Deal with it

- most valuable skill:
  - learn to use programmable tools that prepare data
  - learn to generate compelling data visualizations
  - learn to estimate the confidence for reported results
  - learn automate work, making analysis repeatable

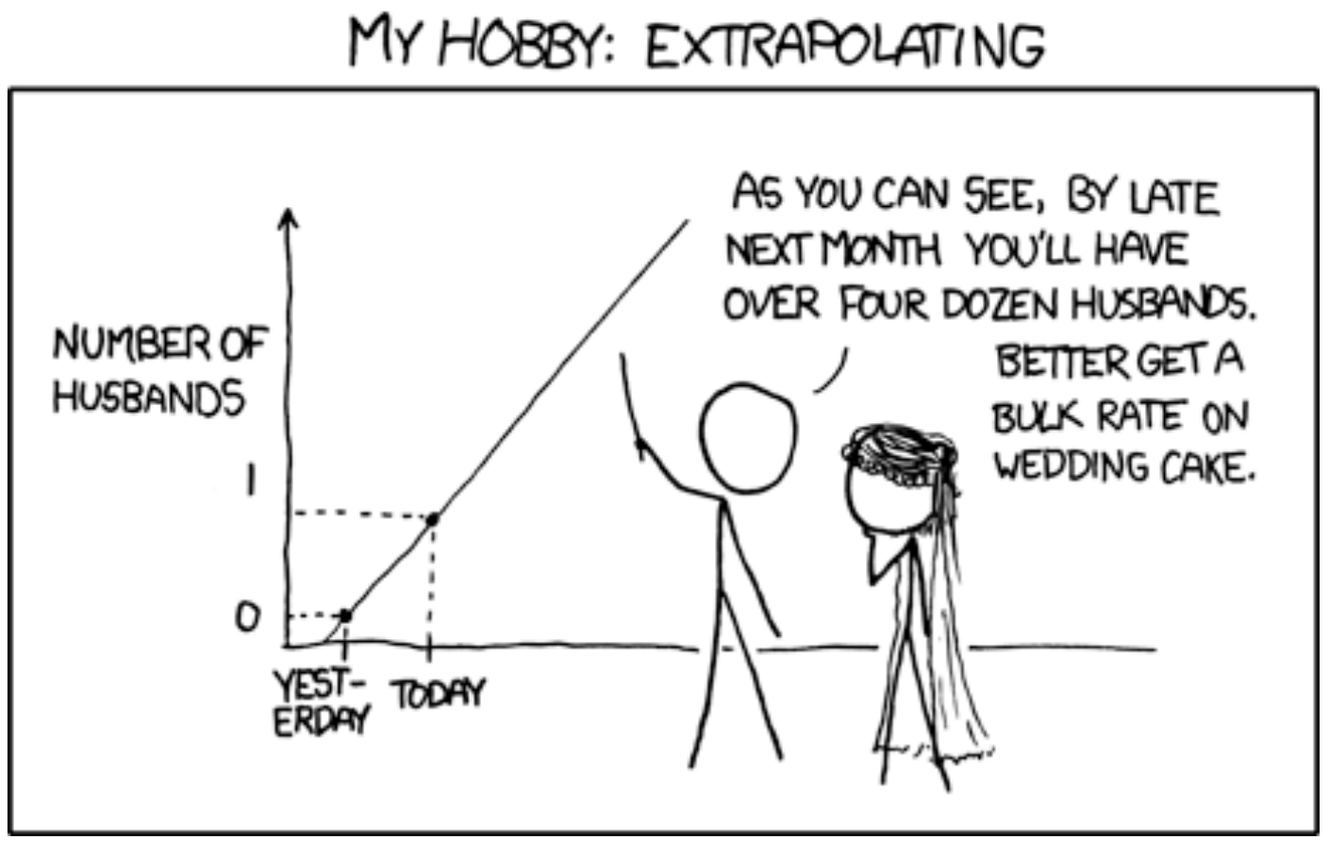
The rest of the skills - modeling, algorithms, etc. - those are secondary

Some points:

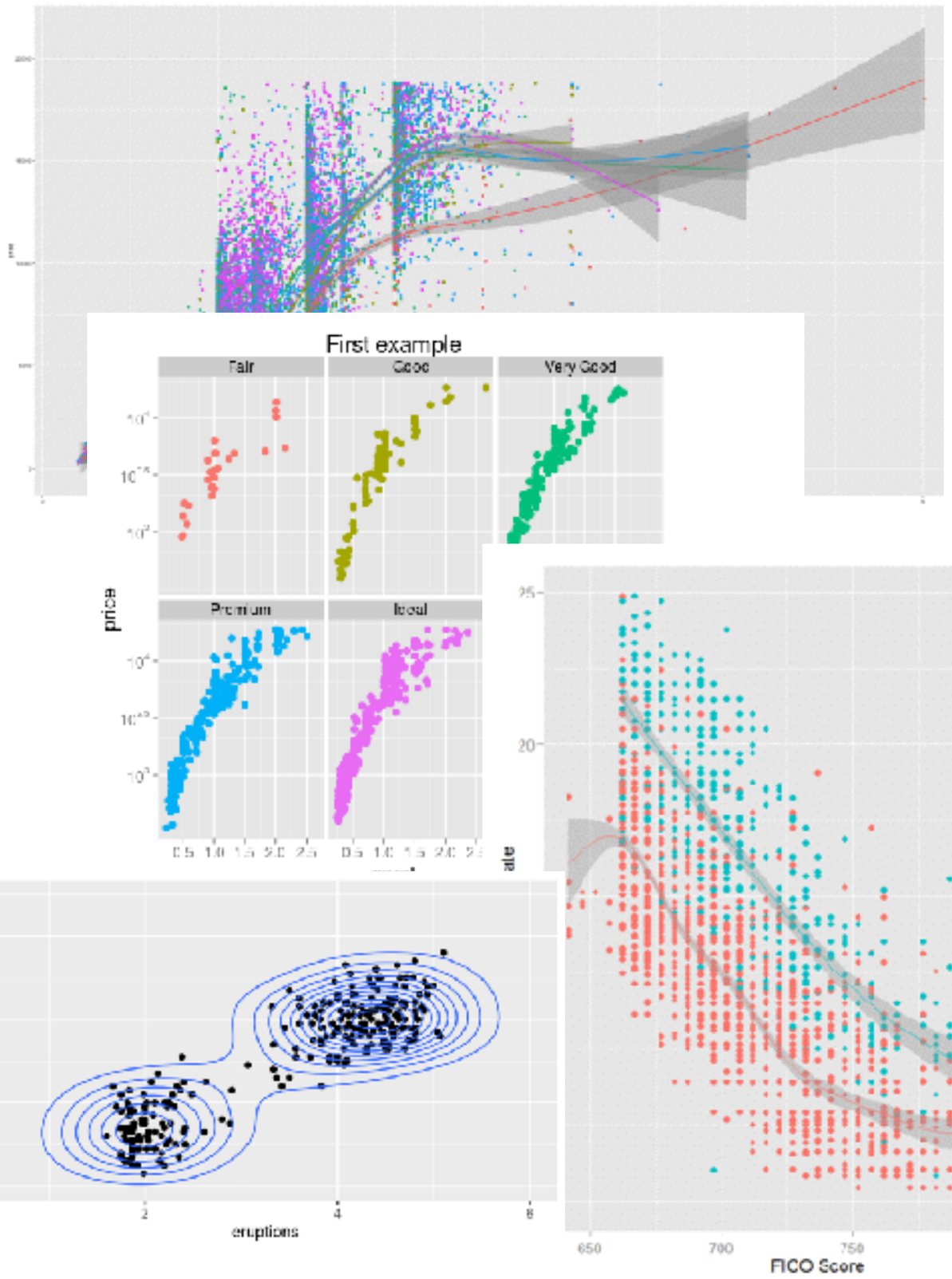
- The phrase "This data cannot be correct!" may be an early warning about the organization itself

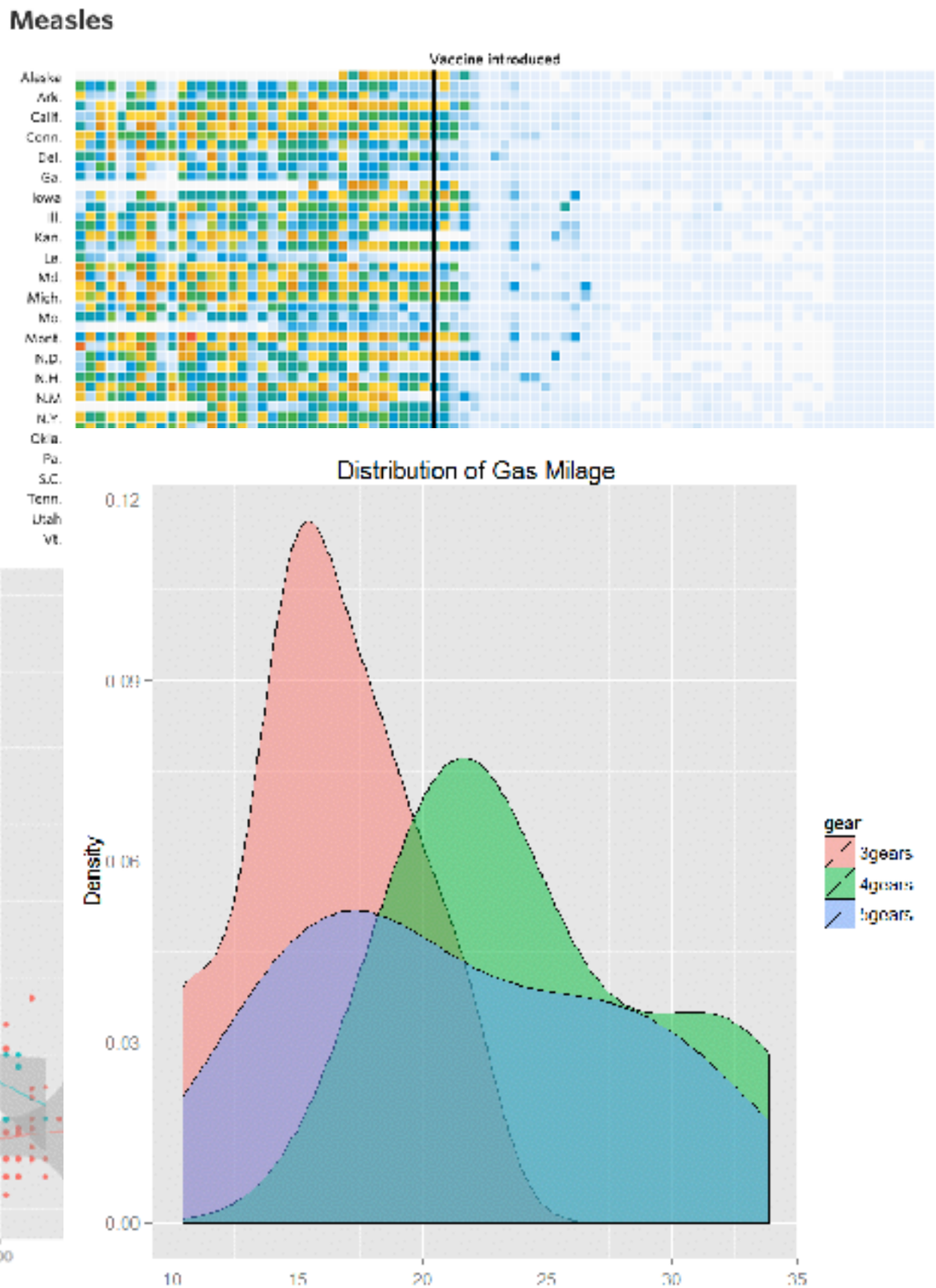
- Much depends on how the people whom you work alongside tend to arrive at their decisions:
  - good: Ideas, Inspiration
  - bad: Deduction, Speculation, Justification

- In general, **one good data visualization** can put any ongoing verbal arguments to rest



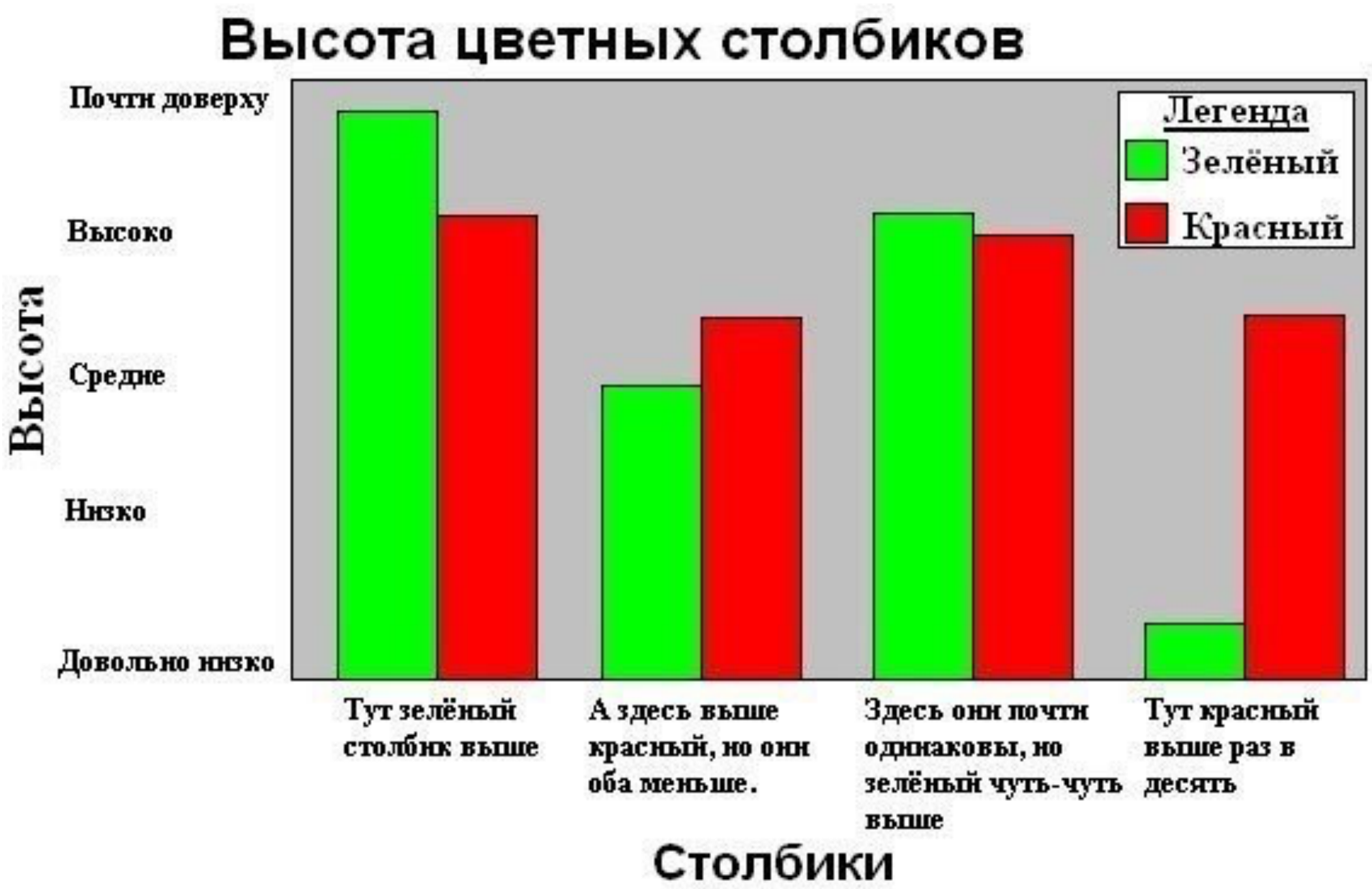
How to get best solution? Visualize it





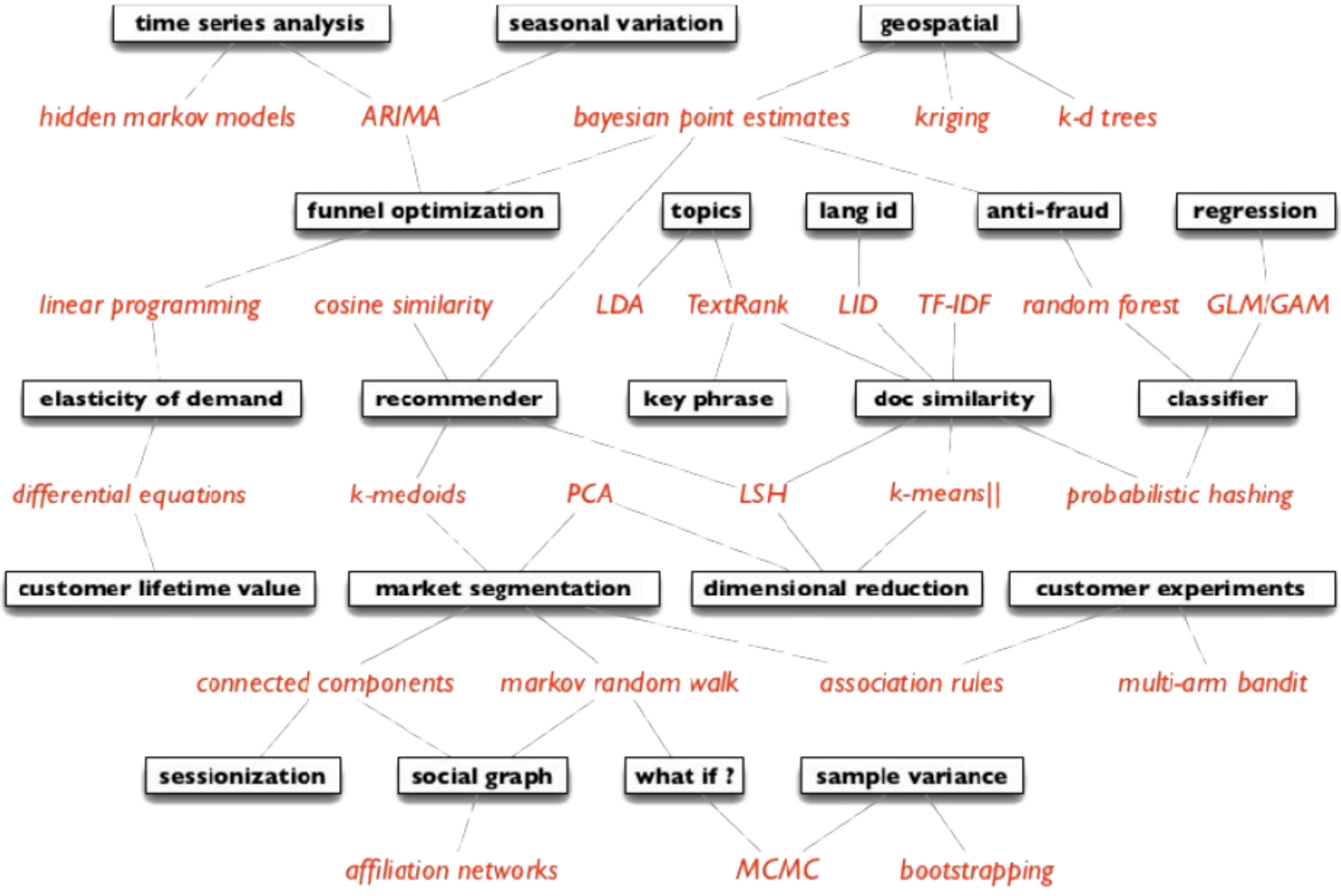
Source: https://habrahabr.ru/company/datalaboratory/blog/311210/

Bad case :c


Высота цветных столбиков
Почти доверху
Высоко
Средне
Низко
Довольно низко
Высота
Легенда
Зелёный
Красный
Тут зелёный столбик выше
А здесь выше красный, но они оба меньше.
Здесь они почти одинаковы, но зелёный чуть-чуть выше
Тут красный выше раз в десять
Столбики

Type of machine learning algorithms

Some great algorithms:



Data Models vs. Algorithmic Models

Data Modeling

Y ◄ (X, random noise, parameters)

We understand the world:

- Type of regression
- Standard statistic methods
- How well my data model works
- and other

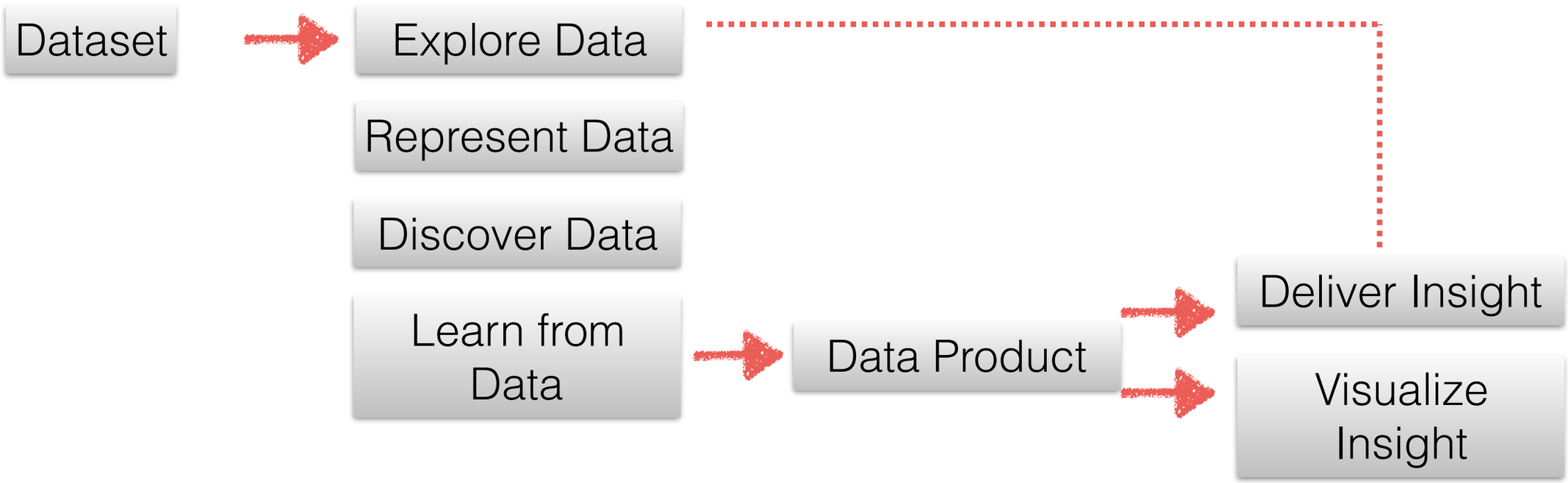
Algorithmic Modeling

Y ◄ (Black Box) ◄ X

We don't understand the world:

- Hard structure algorithms
- Ensemble, XGBoost, Deep Learning
- Clasterization (Why that?)

Data Science Process:



Description & Inference
Data & Algorithm Models
Networks & Graphs
Regression & Prediction
Classification & Clustering
Experiment & Iteration

Modeling
Simulation
Optimization
Objectives
Levers

Predictive
Immediate Impact
Business Value
Easy to Explain

A Data Product is..

Data Jiu-Jitsu: ability to turn data into data product that generate business value

In God we Trust, all other bring data