

Data Science Toolbox:

How bad data science can destroy
a good business

Leonid Danilchenko

About me

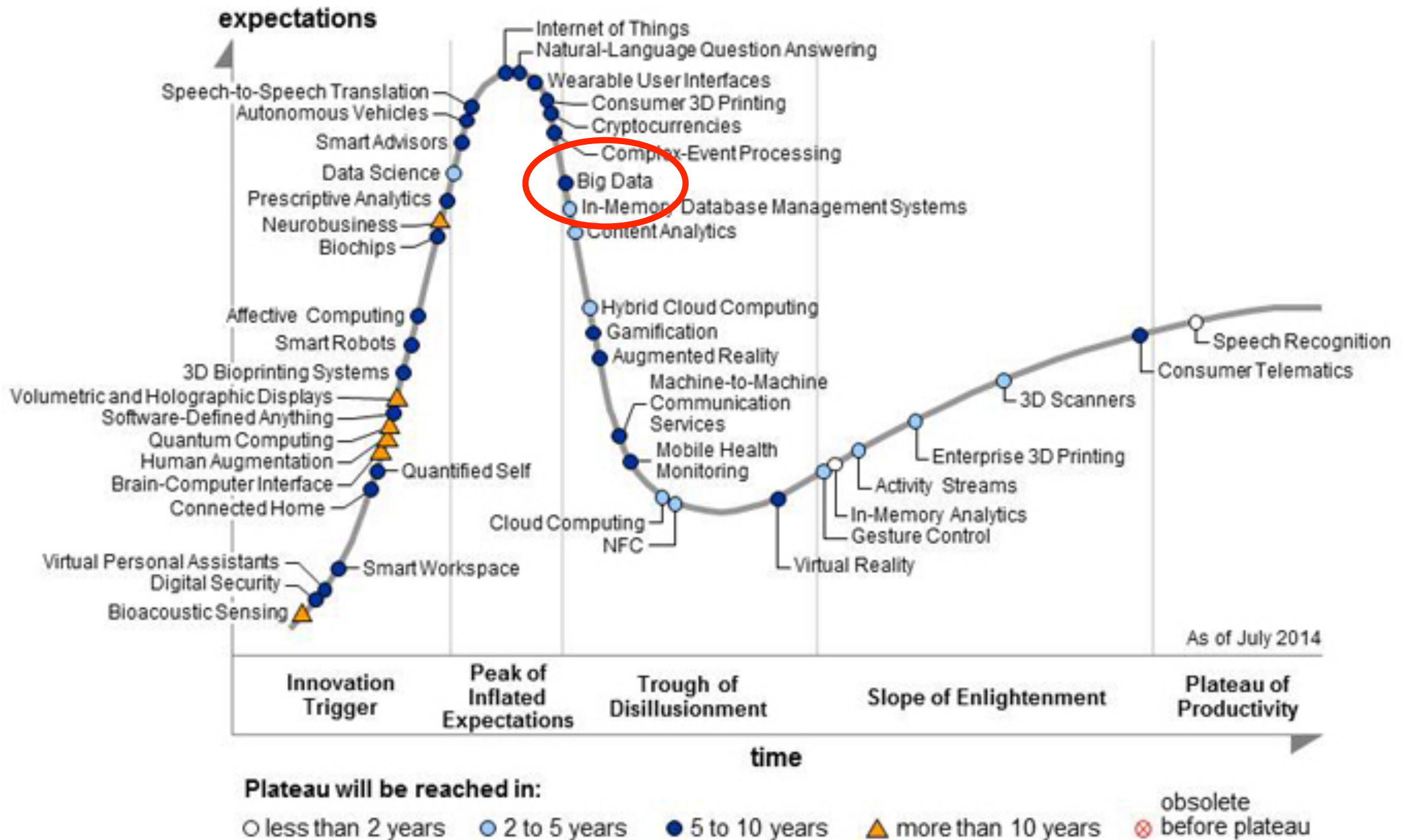


Center of Mathematical Finance
2015/2016



Optimum Media Direction
Media Research

Gartner's 2015 Hype Cycle



Gartner's 2016 Hype Cycle



Source: Gartner (July 2016)

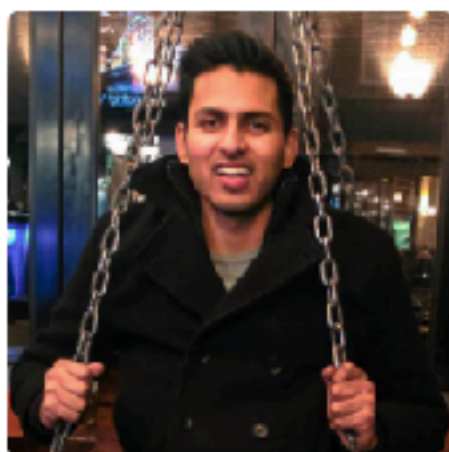




Халява:

- 1 Atom — Текстовый редактор 21 века. Вообще он бесплатный, но в списке присутствует.
- 2 AWS Education — 75-150\$ на сервера амазона (!!!)
- 3 Crowdfunder — 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-репозиторий (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)

GitHub



Soumith Chintala

soumith



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/1511.08434>

★ 316 ● Lua

[cudnn.torch](#)

Torch 7 FFI bindings for NVIDIA CuDNN

★ 231 ● Lua

[imagenet-multiGPU.torch](#)

an imagenet example in torch.

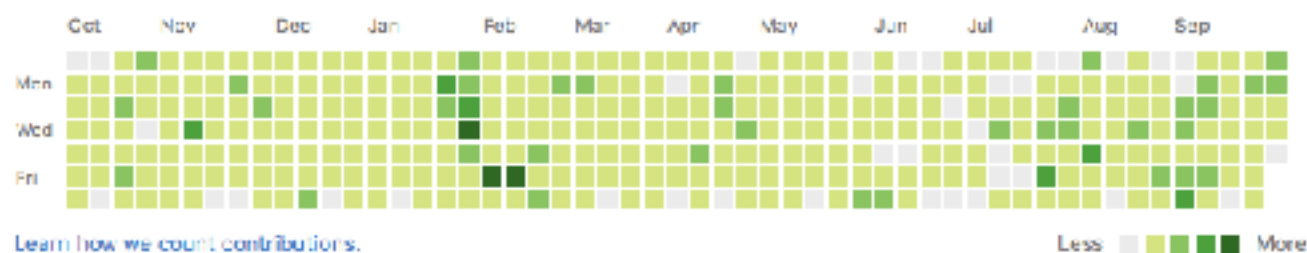
★ 212 ● Lua

[torch-android](#)

Torch-7 for Android

★ 148 ● CMake

3,050 contributions in the last year





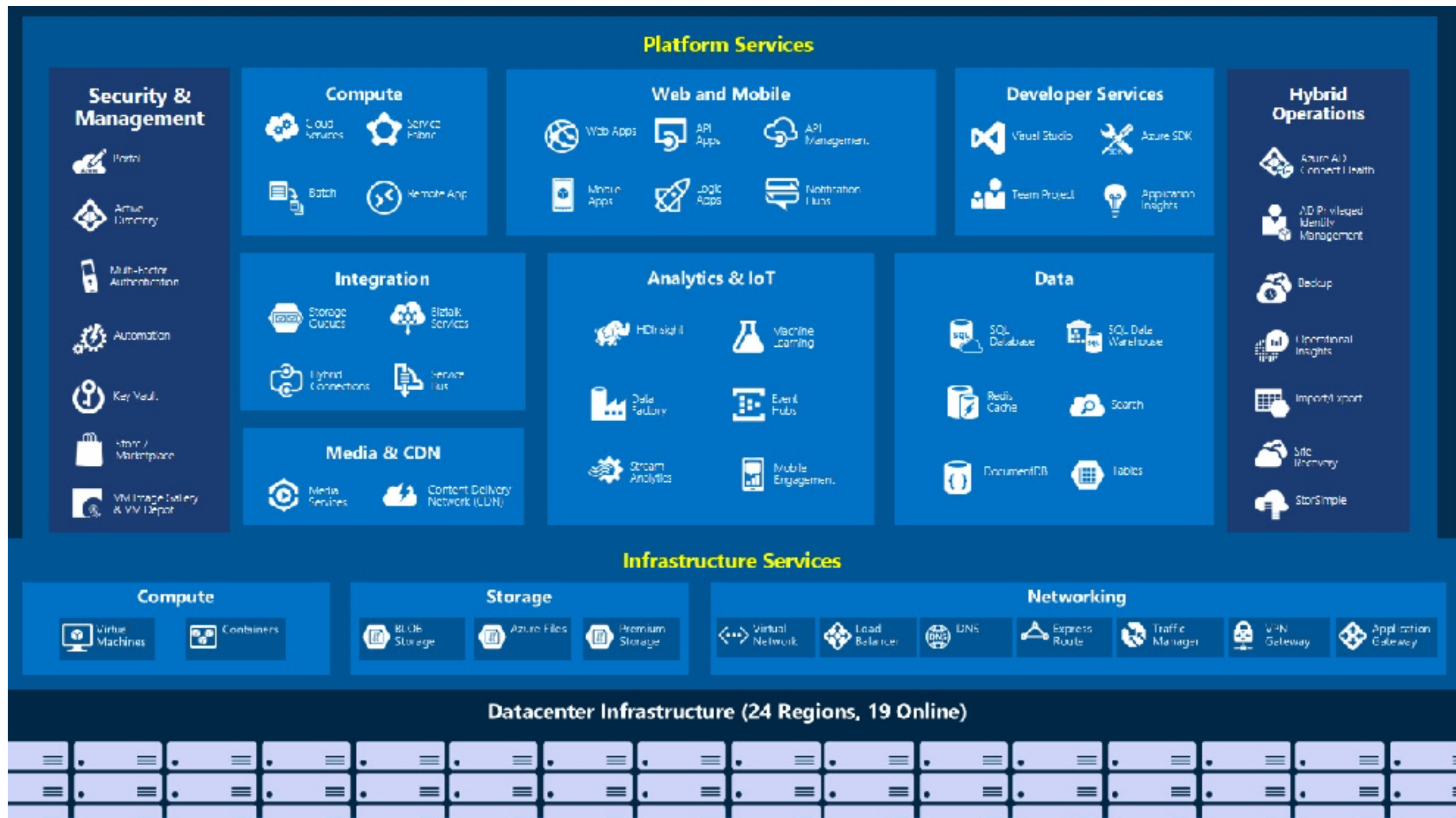
Халява:

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

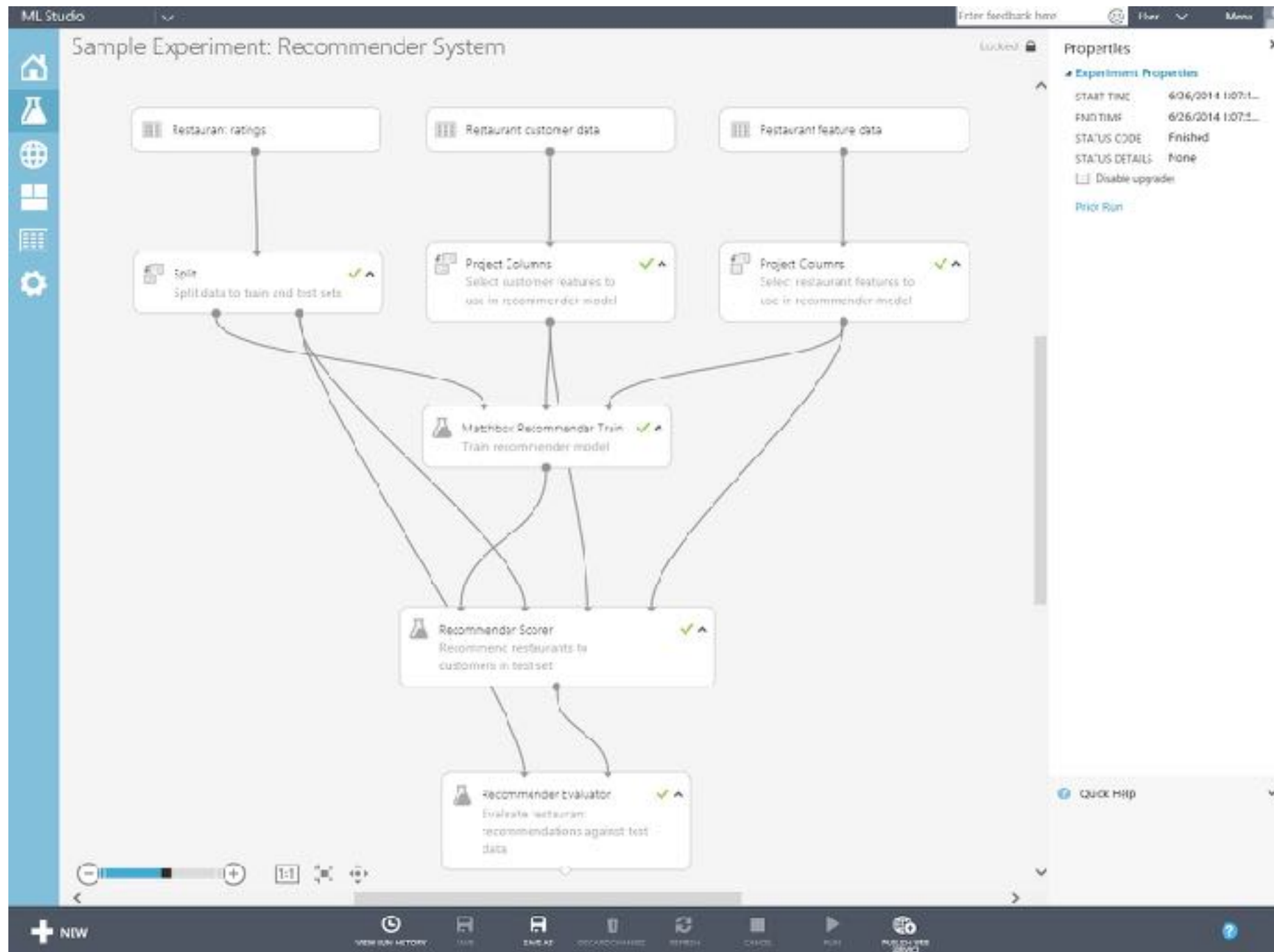


Microsoft Azure Infrastructure



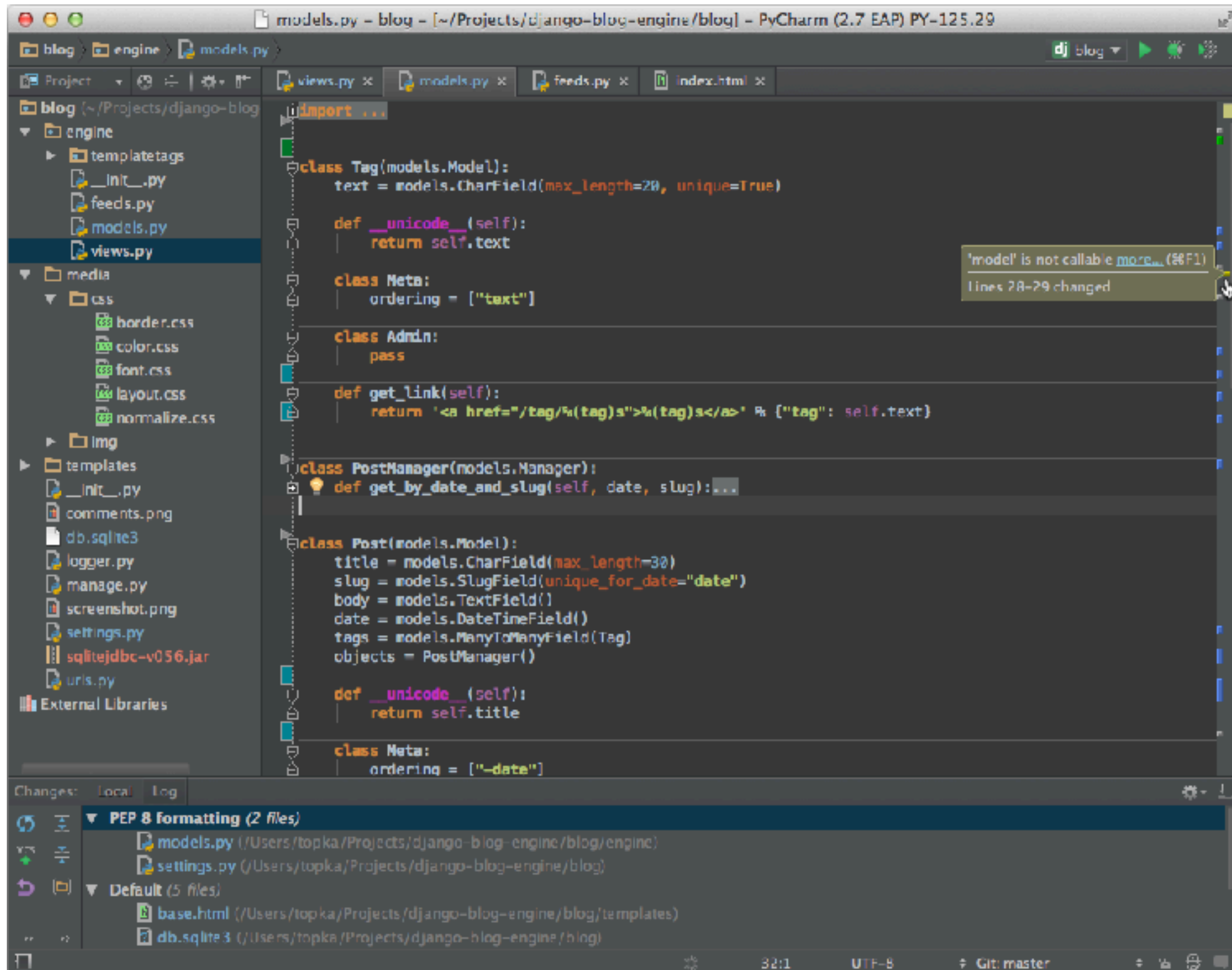


Microsoft Azure ML Studio





PyCharm



Big Data Landscape 2016

Infrastructure

Hadoop On-Premise
cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, splice, bluedata, jethro

Hadoop in the Cloud
amazon, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, dubble, xplenty

Spark
databricks, GridGain, TACHYON, NEXUS

Cluster Services
amazon web services, kubernetes, mesosphere, docker, Core OS, paperkit, StockIQ

Analytics

Analyst Platforms
Palantir, AYASDI, Quid, enigma, Jira, Jira, ORBITAL INSIGHT

Analytics Platforms
Microsoft, guavus, Datameer, interana

Data Science Platforms
cccontest, relevant, DataRobot, Alpine, ADATA, MODE, plotly, dataiku, DCIMINO, genio, yhat, ALGORITHMIA

Visualization
tableau, Google Data Studio, Roambi, Qlik, CHARTIO

Applications

Sales & Marketing
RADIUS, Gainsight, bloomreach, Zeta, livefyre, blueyonder, @kehuna, Lattice, SAILTHRU, persado, infer, sense, AVISO, ACTIONIQ, QUANTIFIND, ENGADIO

Customer Service
MEDALLIA, ATTENITY, STELLA Service, NGDATA, Digital Genies, base machines

Human Capital
gild, Connectifier, textio, entelo, hiQ

Legal
RAVEL, JUDICATA, Everlaw, Brevia, ROSS INTELLIGENCE

NoSQL Databases
amazon DynamoDB, Google Cloud Platform, Microsoft Azure, mongoDB, MarkLogic, COUCHBASE, CERO SPIKE, Couchbase, SequoiaDB, redislabs, influxdata

SQL Databases
SAP, Clustrix, Pivotal, paradigm4, memsql, nuodb, MariaDB, VOLTDB, citusdata, deepdb, Trafalgar, Cockroach LABS

BI Platforms
Power BI, amazon web services, DOMO, Wave Analytics, GoodData, birist, platfora, Looker, abt scale

Statistical Computing
sas, SPSS, MATLAB

Log Analytics
splunk, sumologic, loggly

Social Analytics
NETBASE, DATASIFT, track, bitly, synthetio, simplereach

Ad Optimization
MediaMath, Integral, OpenX, theTradeDesk, Adaptics, LiveIntent, distillery, DataXu, Clupier, TAPAD

Security
CYLANCE, CounterTack, CYBERBASTION, ARCA SECURITY, SentinelOne, Recorded Future, Guardian Analytics, FORTSCALE, sift science, Keybase, feedzai, SIGNIFY5

Vertical AI Applications
facebook, Clara, KASIST, lumina

Graph Databases
neo4j, OrientDB, InfoGraphix

MPP Databases
TERADATA, VERTICA, Netezza, Kognitio, dremio

Cloud EDW
amazon web services, Google Cloud Platform, Microsoft Azure, Pivotal, snowflake, VANTAGE DATA, InfoWorks

Data Transformation
alteryx, TRIFACTA, tamr, Pakata, StreamSets, Alation

Data Integration
informatica, Purview, MuleSoft, snaplogic, BedrockData

Real-Time
amazon web services, KETAMARKETS, confluent, dataArtisans

Machine Learning
Google Machine Learning, H2O, Dato, SKYTRIL, rapidminer, YISEZE, PredictionID, ginsulab

Speech & NLP
NarrativeScience, apilai, NUANCE, semantic machines, Cortana, VIV, Numenta, SI, clarifai

Horizontal AI
IBM Watson, Cortana, VIV, Numenta, SI, clarifai

Publisher Tools
outbrain, mixpanel, Chartbeat, yieldbot, Yieldmo

Govt/Regulation
Socrata, OPENGOV, EN, FiscalNote, PRICED, mark43, OpenDataSoft

Finance
Affirm, LendingClub, OnDeck, Kreditech, Kabbage, tidemark, INSIGHT, ZUORA, Dataminr, Lenddo, KENSHO, AIDYA, ISENTIUM, Quantopian, sentient

Management / Monitoring
New Relic, APPDYNAMICS, amazon web services, Numerify, splunk, safaric, roccano, Anavox

Security
TANUUM, illumio, CODE 42, DataGravity, CoreCloud, VECTRA, sqrl, bluealan

Storage
amazon web services, Google Cloud Platform, Microsoft Azure, parasetech, nimblestorage, Clumulo

App Dev
apigee, CASK, Typesafe, CONCLUSION

Crowd-sourcing
amazon mechanical turk, Crowdflower, WorkFusion

Search
hp, Oracle, EXALEAD, Lucidworks, elastic, ThoughtSpot, MAANA, swifttype, Algolia, SINGOLIA

Data Services
UO, OPERA, Mu Sigma, DATA SCIENCE, kaggle, DataKind

For Business Analysts
Origami Logic, ClearStory, CIRRO, Import IO

SMB / Commerce
Google Analytics, AMPITUDE, RJMetrics, BLUECORE, sumail, granify, Airtale, retention, custora

Education/ Learning
KNEWTON, Clever, Gleara, PANORAMA, knowre

Life Sciences
23andMe, Counsyl, Rncombine, CYRUS, FLATIRON, oozymergen, HealthTap, METABIOTA, ZEPHYR, ovla, Gingerio, transcriptic, Glow, enlho, AICure, 23andMe

Industries
OPower, eHarmony, RetailNext, duetto, STITCH FIX, WorkFusion, BLUESRIVER, TACHYUS, Seeq, FarmLogs, SwiftKey, select, Best Machine, statmuse, BOXEVER

Cross-Infrastructure/Analytics

amazon web services, Google, Microsoft, IBM, SAP, SAS, hp, Juniper, vmware, talend, TIBCO, TERADATA, ORACLE, NetApp

Open Source

Framework
Hadoop, HADOOP, YARN, Spark, MESOS, TEZ, Flink, CDAP

Query / Data Flow
SLAMDATA, ARACHNE DRILL, Google Cloud Dataflow

Data Access
cassandra, HBASE, mongoDB, CouchDB, riak, kafka, nifi

Coordination
talend, Apache Zookeeper, Apache Ambari

Real-Time
STORM, Spark, Flink, druid

Stat Tools
R, Scala, SciPy

Machine Learning
mlib, Aerosolve, Caffe, FeatureFu, DIMSUM, mllib, Apache SINGA, MADlib, CNTK, TensorFlow, DL4J

Search
elasticsearch, Solr, Lucene

Security
Apache Ranger

Visualization
zeppelin

Data Sources & APIs

Health
Apple, JAWBONE, GARMIN, practicefusion, fitbit, Withings, VALIDIC, nclatmo, kinsa, Human API

IOT
UPTAKE, ThingWorx, helium, somero

Financial & Economic Data
Bloomberg, DOW JONES, YODLEE, PREMISE, SEP CAPITAL IQ, Quandl, xignite, CB Insights, mattermark, Destimize, FLIND

Air / Space / Sea
PLANET Labs, spire, WINDWARD, CRUISE, Airware, Transporexp

Location/People/Entities
GARMIN, foursquare, InsideView, esri, STREETLINE, CARTOON, factual, PlaceIQ, Camsen Hexagon, placemeter, BASIS, Sensus

Other
qualtrics, panjiva, DATA.GOV

Incubators & Schools
GA, DataCamp, INSIGHT, DataElite, METIC, The Data Incubator

Data Science Toolkit:

Visualization:

ggplot2, matplotlib, D3, GraphViz

Modeling:

Python, R, Scala, C++, Java

Reporting:

PowerBI, Shiny, Tableau, Zeppelin

Deep Learning:

Theano, CUDA, TensorFlow

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



КОГДА ЭКСПЕРИМЕНТИРУЮТ...

НО, СКОРЕЕ ВСЕГО, ОНИ СКАЖУТ...



twisteddoodles.com

[Some] Data Science Principles:

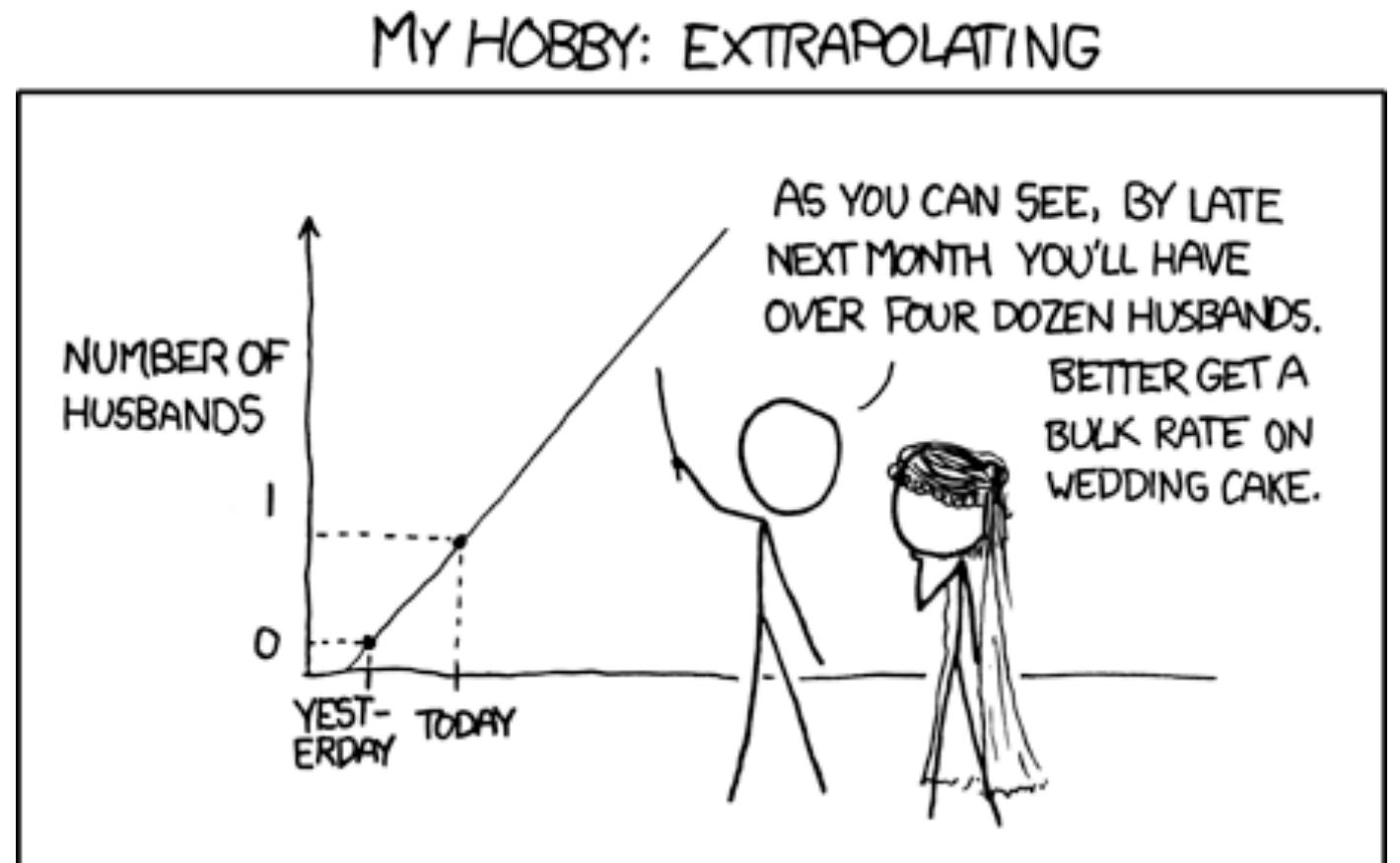
- approximately **80% of the costs** for data-related project get spent on data preparation - data is always dirty. Deal with it
- most valuable skill:
 - learn to use programmable tools that prepare data
 - learn to generate compelling data visualizations
 - learn to estimate the confidence for reported results
 - learn automate work, making analysis repeatable

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

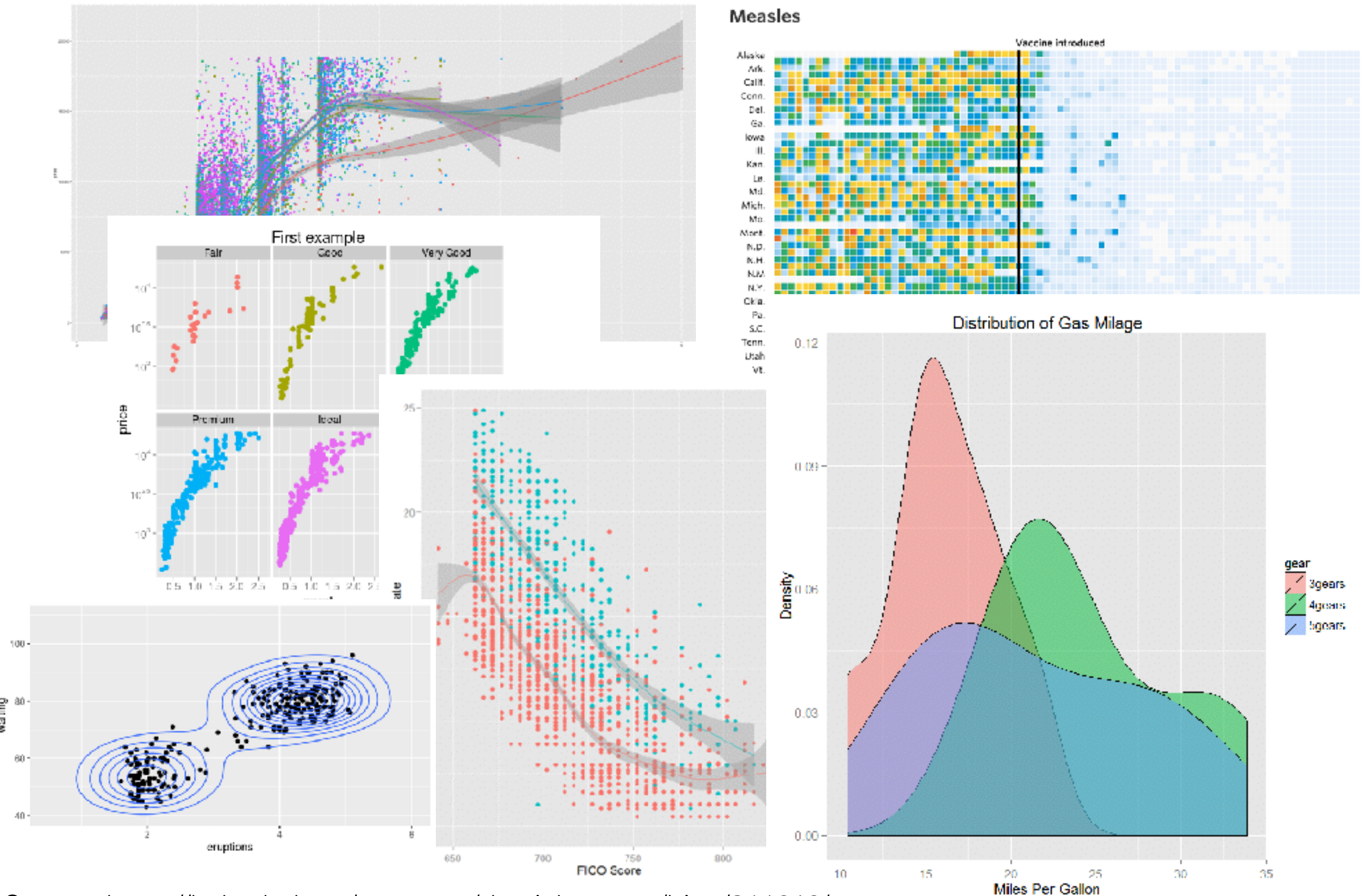
Some points:

- 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

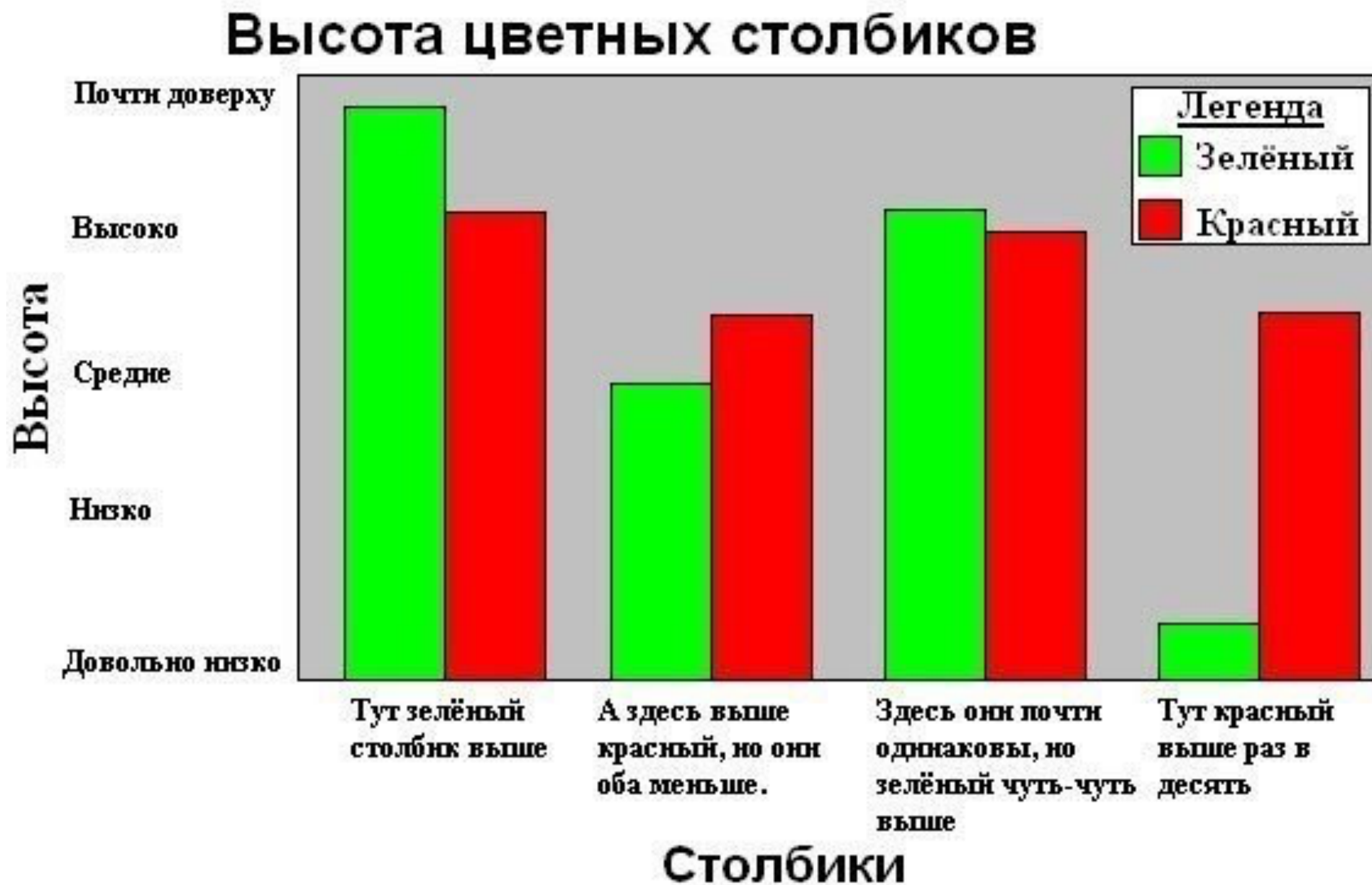
- In general, **one good data visualization** can put any ongoing verbal arguments to rest



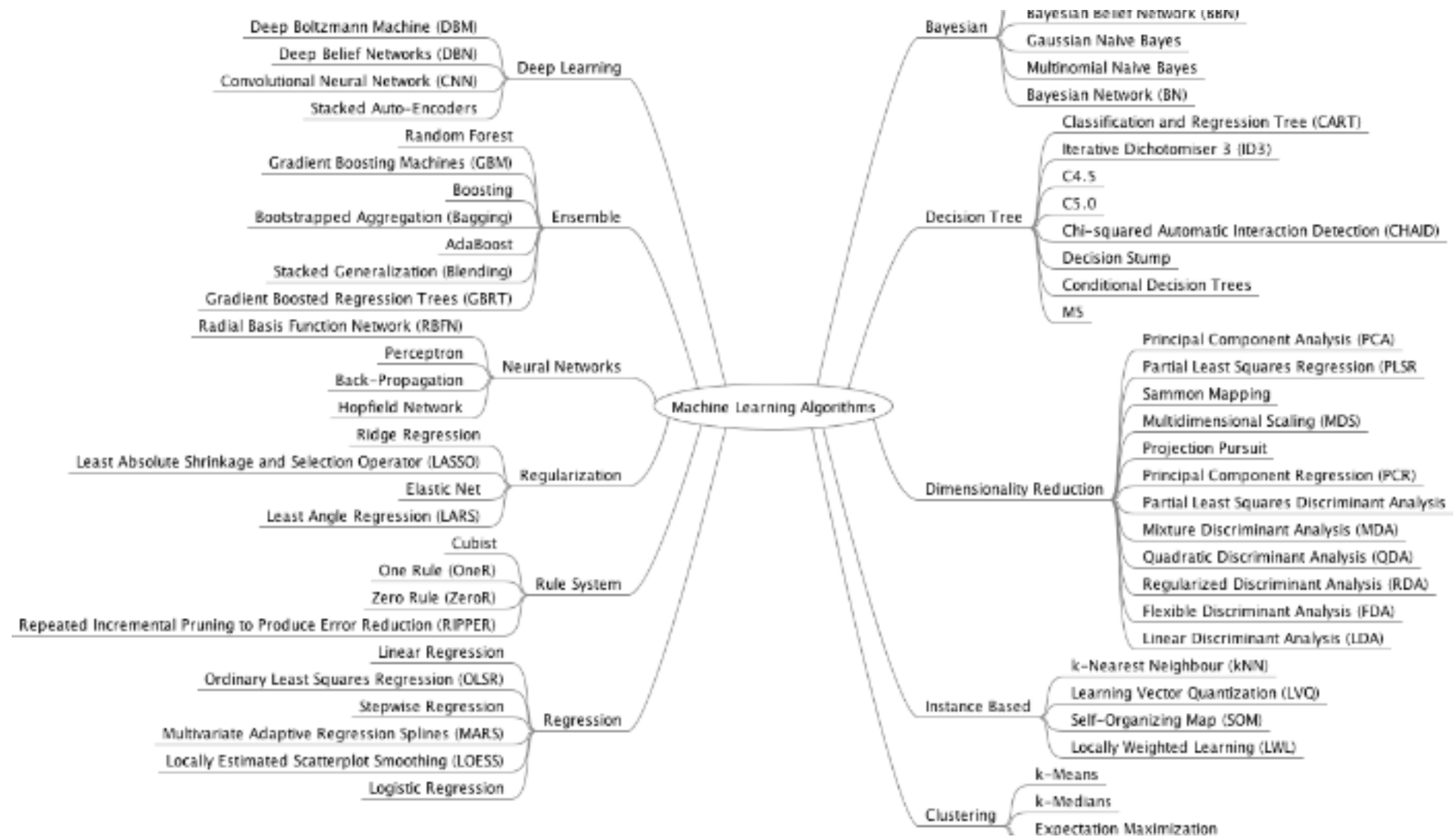
How to get best solution? Visualize it



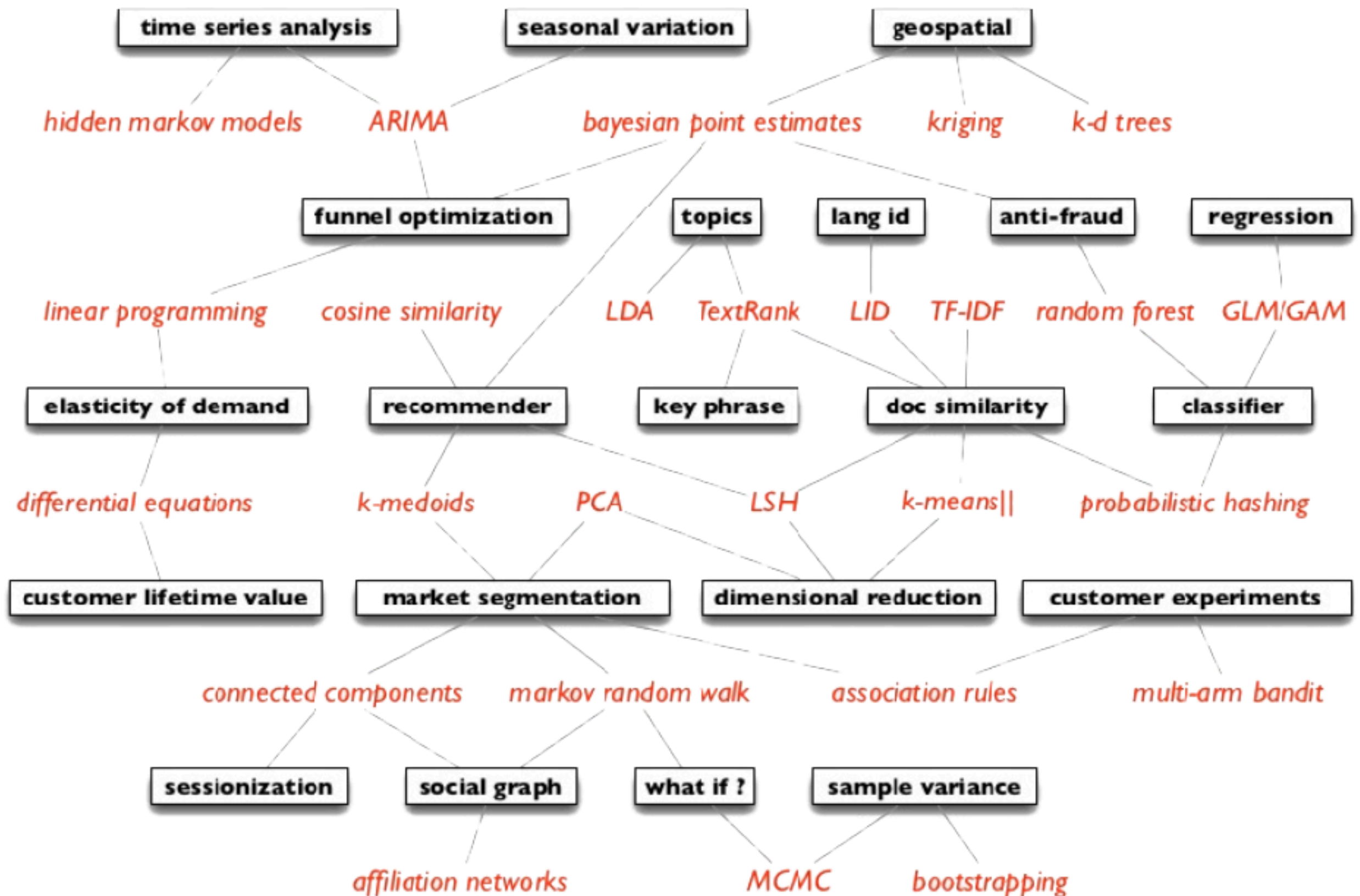
Bad case :c



Type of machine learning algorithms



Some great algorithms:



Data Models vs. Algorithmic Models

Data Modeling

$Y \leftarrow (X, \text{random noise, parameters})$

We understand the world:

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

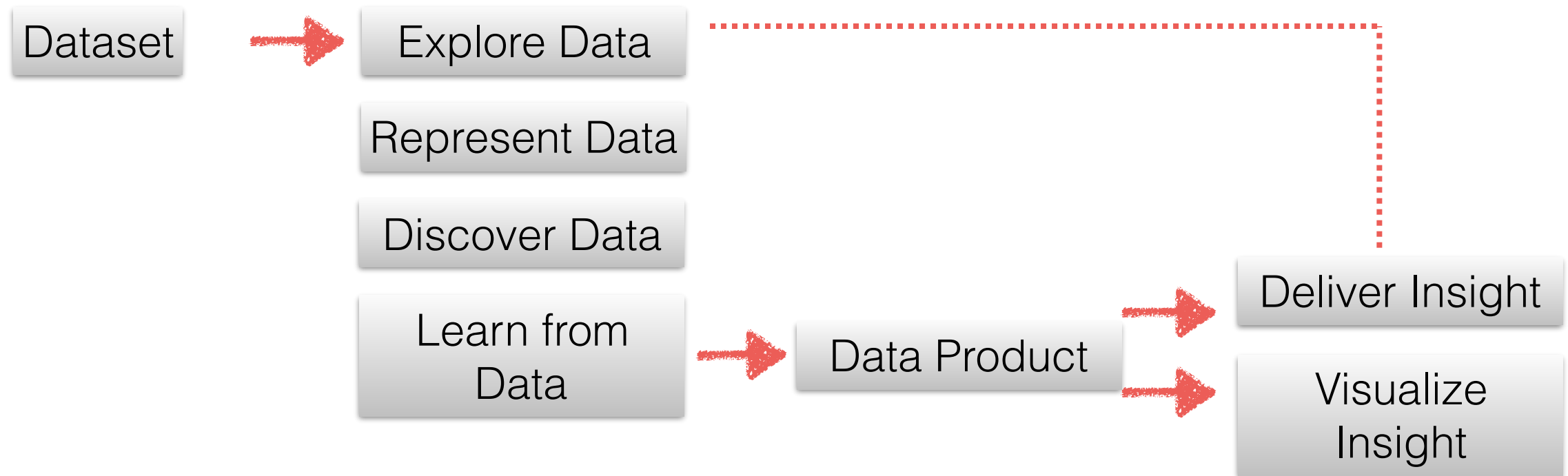
Algorithmic Modeling

$Y \leftarrow (\text{Black Box}) \leftarrow X$

We don't understand the world:

- Hard structure algorithms
- Ensemble, XGBoost, Deep Learning
- Clusterization (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