RIGACODINGSCHOOL

Data Analysis Platform

Par mums

- Viena no pirmajām programmēšanas skolām Lietuvā un Latvijā;
- Skolas filiāles Viļņā, Klaipēdā, Kauņā un Rīgā;
- Vairāk nekā 1500 absolventi;
- 80% absolventu, kas vēlas strādāt IT jomā;
- Vairāk nekā 50 sadarbības uzņēmumu Rīgā;
- Skolas lektori – ar pieredzi IT sfērā.

VALDIS

- Izglītība: Maģistra grāds datorzinātnēs
- Pieredze programmēšanā: 20+ gadi
- Specialitāte: grafu teorija sociālo tīklu analizēšanā
- Hobiji: prāta spēles, riteņbraukšana, šahs
- RCS Pasniedzu: Python iesācējiem, Datu Analīzes kursus

Data Lake

Brief History of Data Analysis

- ~ 18,000BC – Uganda, Ishango Bone
- ~ 2400BC – Babylon abacus, libraries
- 300BC – 48AD – Library of Alexandria
- ~ 100-200AD Antikythera Mechanism

Brief History of Data Analysis II

- 1663 – London, J.Graunt mortality analysis
- 1865 – banker H. Furnese business intelligence
- 1880-90 US Census Hollerith Machine -> IBM
- 1928 – F. Pfleumer magnetic tape invention

Brief History of Data Analysis III

- 1950s - Flat Files
- 1958 – IBM’s Luhn defines Business Intelligence
- 1960s - CODASYL
- 1970s – Codd’s relational DBs -> SQL
- 1980s – Data Werehouses / Marts
- 2000s – Big Data / noSQL DBs
- 2010s – Rise of accessible ML/DL libraries

BIG DATA LANDSCAPE 2017

INFRASTRUCTURE

HADOOP ON-PREMISE
cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, bluedata, jethro

HADOOP IN THE CLOUD
amazon web services, Microsoft Azure, Google Cloud Platform, IBM InfoSphere BigInsights, TREASURE DATA, Qubole, altiscale, CAZENA, CenturyLink

STREAMING / IN-MEMORY
amazon web services, databricks, confluent, striim, GridGain, METAMARKETS, DATATORRENT, dataArtisans, ORACLE Coherence, hazelcast, TERRACOTTA

NOSQL DATABASES
Google Cloud Platform, amazon DynamoDB, ORACLE, Microsoft Azure, MarkLogic, mongoDB, DATASTAX, AEROSPIKE, Couchbase, redislabs, influxdata

NEWSQL DATABASES
SAP HANA, Clustrix, Pivotal, NUODB, Cockroach LABS, memsql, splice MACHINE, MariaDB, VOLTDB, citusdata, Trafodion, deepdb, paradigm4

GRAPH DBS
neo4j, IBM, ORACLE, OrientDB, InfiniteGraph, Objectivity

MPP DBS
TERADATA, VERTICA, NETEZZA, Actian, kognitio, EXASOL, dremio

CLOUD EDW
amazon web services, Google Cloud Platform, Microsoft Azure, Pivotal, snowflake, Infoworks

DATA TRANSFORMATION
talend, pentaho, alteryx, TRIFACTA, tamr, Paxata, StreamSets, UNIFI

DATA INTEGRATION
informatica, MuleSoft, snapLogic, TEALIUM, Segment, enigma, podium data, alooma, xplenty, ZALONI, import.io, Stitch

DATA GOVERNANCE
informatica, IBM, skyhigh, collibra, Alation, Waterline Data

MGMT / MONITORING
amazon web services, New Relic, APPDYNAMICS, actifio, DATADOG, WAVEFRONT, unravel, splunk, iDoc, Anodot, Procana, Numerify, pagerduty

STORAGE
amazon web services, Google Cloud Platform, Microsoft Azure, ALLUXIO, nimblestorage, Qumulo, COHO, panasas

CLUSTER SERVICES
amazon web services, kubernetes, docker, MESOSPHERE, CoreOS, pepperdata

APP DEV
Lightbend, Kaon IO, rainforest, CASK

CROWDSOURCING
amazon mechanical turk, Upwork, WorkFusion, CrowdFlower

HARDWARE
Google TPU, ARM, nervana, Graphcore, MYTHIC, cerebras, NVIDIA, Movidius, SCORTEX

ANALYTICS

DATA ANALYST PLATFORMS
Microsoft, pentaho, alteryx, Digital Reasoning, GUAVUS, AYASDI, MATTIVIO, Datameer, Quid, ClearStory, OrigamiLogic, inter|ana, Bottlenose, ARIMO, ENDOR, MODE

DATA SCIENCE PLATFORMS
IBM, KNIME, dataiku, DOMINO, yhat, rapidminer, CONTINUUM ANALYTICS, Alpine, ALGORITHMIA, Angoss

BI PLATFORMS
Microsoft, DOMO, amazon web services, looker, Wave Analytics, atscale, ARCADIA DATA, GoodData, SISENSE

VISUALIZATION
tableau, SAP Lumira, Google Cloud Platform, Qlik, celonis, MAPD, Periscope, ZEPL, CHARTIO, plotly, GOOMDATA

VERTICAL ANALYTICS
PREDIX, C3 IoT, CAPE, UPTAKE, Orbital Insight, TACHYUS, Alluvium, datorama

STATISTICAL COMPUTING
sas, SPSS, MATLAB

DATA SERVICES
Palantir, Mu Sigma, OPERA, DATA SCIENCE, kaggle, Fractal, EXL, DataKind, FF

MACHINE LEARNING
Azure Machine Learning, amazon web services, Google Cloud Platform, H2O.ai, DataRobot, context relevant, VISENZE, bonsai, DATARPM, deepsense.io, nutonian

HORIZONTAL AI
IBM Watson, Cortana, Face++, sentient, Voyager Labs, vicarious, clarifai, Affectiva, CognitiveScale, SENSETIME, Cronocam, PETUUM, Numenta, nara, OSARO, CURIOUS AI, BLUE VISION

SPEECH & NLP
Google Cloud Platform, twilio, amazon alexa, NarrativeScience, semanticmachines, Wolfram Alpha, MindMeld, ARRIA, NUANCE, Talkiq, iDIBON, maluuba, snips, cortical.io, yseop, Gridspace, SoundHound Inc.

SEARCH
Autonomy, elastic, EXALEAD, ORACLE ENDECA, ThoughtSpot, Algolia, Lucidworks, swiftype, MAANA, alphasense, Searchink, SINEQUA

LOG ANALYTICS
splunk, sumologic, loggly, kibana, logz.io

SOCIAL ANALYTICS
Hootsuite, sprinklr, NETBASE, DATASIFT, tracx, synthesio, simplereach, bitly, predata

WEB / MOBILE / COMMERCE ANALYTICS
Google Analytics, mixpanel, AMPLITUDE, sumAll, Airtable, retention science, SIGOPT, granify, custora

CROSS-INFRASTRUCTURE/ANALYTICS
amazon web services, Google Cloud Platform, Microsoft, IBM, SAP, Hewlett Packard Enterprise, sas, 1010data, vmware, TIBCO, TERADATA, ORACLE, NetApp

APPLICATIONS – ENTERPRISE

SALES
einstein, CHORUS, INSIDESALES.COM, conversica, Gong, clari, AVISO, TACT, fuse|machines, TROOPS

MARKETING - B2B
RADIUS, App Annie, EVERSTRING, Lattice, infer, MINTIGO, 6sense, tubular, Reflektion, DataFox, NGAGIO

MARKETING - B2C
Zeta, bloomreach, blueyonder, PERSADO, kahuna, ACTIONIQ, SAILTHRU, QUANTIFIND, mparticle, Amplero

CUSTOMER SERVICE
MEDALLIA, zendesk, CLARABRIDGE, Gainsight, CLICKFOX, NGDATA, DigitalGenius, appuri, AUTOMAT, frame.ai, msg.ai, INTERCOM

HUMAN CAPITAL
HireVue, entelo, hiQ, GIGSTER, textio, RESTLESS BANDIT, Wade&Wendy, Clustree, Stella, pymetrics

LEGAL
RAVEL, Seal, Everlaw, JUDICATA, Brevia, ROSS, casetext

FINANCE
anaplan, Zuora, tidemark, SAP S/4 HANA, TRADESHIFT

ENTERPRISE PRODUCTIVITY
slack, facebook, X.ai, ORACLE, lumiata, diffbot, Clara, talla, butter.ai, KASISTO

BACK OFFICE AUTOMATION
HyperScience, Apptricity, AppZen

SECURITY
TANIUM, CYLANCE, StackPath, DARKTRACE, illumio, CODE42, ThreatMetrix, DataGravity, VECTRA, CipherCloud, Guardian Analytics, cybereason, ANOMALI, siftscience, SIGNIFYD, SentinelOne, SecurityScorecard, BlueTalon, Recorded Future, feedzai, sqrrl, SOCURE, AREA 1 SECURITY, FORTSCALE, Kaybase, sparkcognition

APPLICATIONS – INDUSTRY

ADVERTISING
AppNexus, MediaMath, criteo, xAd, rocketfuel, Integral Ad Science, theTradeDesk, OpenX, MOAT, distillery, Adgorithms, drawbridge, LiveIntent, gumgum, TAPAD, DataXu, DYNAMIC YIELD, databerries, Yieldmo, Clypper

EDUCATION
KNEWTON, Clever, declara, kidaptive, PANORAMA, knowre, gradescope

GOVERNMENT
Socrata, OPENGOV, mark43, FiscalNote, OpenDataSoft

FINANCE - LENDING
OnDeck, affirm, Kreditech, AVANT, zestfinance, INSIKT, TALA, MoneyLion, TrueAccord, trooly, cignifi, aire, Active.Ai

FINANCE - INVESTING
Dataminr, KENSHO, Quantopian, NUMERAI, iSENTIUM, claritymoney, ALGORIZ, AIDYIA, RavenPack

REAL ESTATE
Opendoor, VTS, CREDIFI, reonomy, COMPSTAK

INSURANCE
metromile, Lemonade, CYENCE, Shift Technology, Tractable

HEALTHCARE
FLATIRON, KYRUUS, HealthTap, METABIOTA, Ginger.io, Glow, popvlon, COTA, zebra, ovia, AiCure, enlitic, Qventus, imagia, ARTERYS, BAYLABS, CLOUD MEDX, Kang Health, IMAGEN, freenome, DIAGNOS

LIFE SCIENCES
23andMe, color, iCarbonX, zymergen, BenevolentAI, PATHWAY GENOMICS, ZEPHYR HEALTH, 3Scan, Clear Labs, Citrine, Phosphorus, twoXAR, Atomwise, deep genomics

TRANSPORTATION
UBER, ZOOX, TESLA, nuTonomy, CLEARPATH, nauto, drive.ai, AMOTIVE, pilot.ai, OPTIMUS RIDE, OTTO, nexar, comma.ai, Civil Maps, netradyne, NIO

AGRICULTURE
FARMERS BUSINESS NETWORK, FarmersEdge, FarmLogs, BLUE RIVER, mavrx, TerrAvion, prospera

COMMERCE
instacart, STITCH FIX, RetailNext, HowGood

OTHER
eHarmony, stem, rethink robotics, SIGHT MACHINE, hoppen, BOXEVER, celect, VERDIGRIS, duetto, Unbabel, Jukedeck, Second Spectrum

OPEN SOURCE

FRAMEWORK
hadoop HDFS, Apache Kafka, hadoop MapReduce, YARN, Flink, TEZ, MESOS, Spark, CDAP

QUERY / DATA FLOW
Spark SQL, presto, Pig, HIVE, SLAMDATA, APACHE DRILL, Google Cloud Dataflow

DATA ACCESS
cassandra, nifi, mongoDB, CouchDB, SciDB, OPENTSDB, riak, HBASE, Spanner, accumulo

COORDINATION
talend, Apache Zookeeper, Apache Ambari

STREAMING
Spark, APEX, Flink, beam, kafka, druid, STORM

STAT TOOLS
python, R, ScalaLab, NumPy, SciPy

AI / MACHINE LEARNING / DEEP LEARNING
TensorFlow, theano, torch, MADlib, learn, Caffe, CNTK, DMTK, Keras, PaddlePaddle, Apache SINGA, OpenAI, neon, FeatureFu, mxnet, VELES, WEKA, DSSTNE, mllib, DL4J, Mahout, Chainer, DIMSUM, Aerosolve

SEARCH
elasticsearch, Solr, Lucene

LOG ANALYSIS
elasticsearch, kibana, logstash

VISUALIZATION
BEAKER, Rodeo

COLLABORATION
jupyter, Zeppelin, ANACONDA

SECURITY
Apache Ranger, KNOX, Sentry

DATA SOURCES & APIS

HEALTH
JAWBONE, VALIDIC, practicefusion, fitbit, GARMIN, Human API, kinsa

IOT
GE Digital, UPTAKE, ThingWorx, helium, samsara, AUGURY, estimote

FINANCIAL & ECONOMIC DATA
Bloomberg, THOMSON REUTERS, DOW JONES, S&P CAPITAL IQ, CB INSIGHTS, xignite, quandl, YODLEE, PREMISE, estimize, Eagle Alpha, StockTwits, SECOND MEASURE, Thinknum, PLAID, mattermark

AIR / SPACE / SEA
PLANET LABS, Airware, spire, SKYCATCH, AIROBOTICS, Descartes Labs, UNDERSTORY, WINDWARD, TELLUSLABS, DroneDeploy, MarineTraffic

PEOPLE / ENTITIES
acxiom, Experian, EPSILON, InsideView, Crimson Hexagon, BASIS, quantcast, SAFEGRAPH

LOCATION INTELLIGENCE
FOURSQUARE, MapAnything, Sense360, PlaceIQ, esri, factual, CARTO, Mapillary, STREETLINE

OTHER
qualtrics, DATA.GOV, data.world, panjiva, enigma

DATA RESOURCES

INCUBATORS & SCHOOLS
PLURALSIGHT, GA, galvanize, DataCamp, DataElite, INSIGHT, The Data Incubator, METIS

RESEARCH
facebook research, OpenAI, MIRI, CSAIL, DFKI, AI2, ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

V2 – Last updated 5/3/2017

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark (@firstmarkcap)

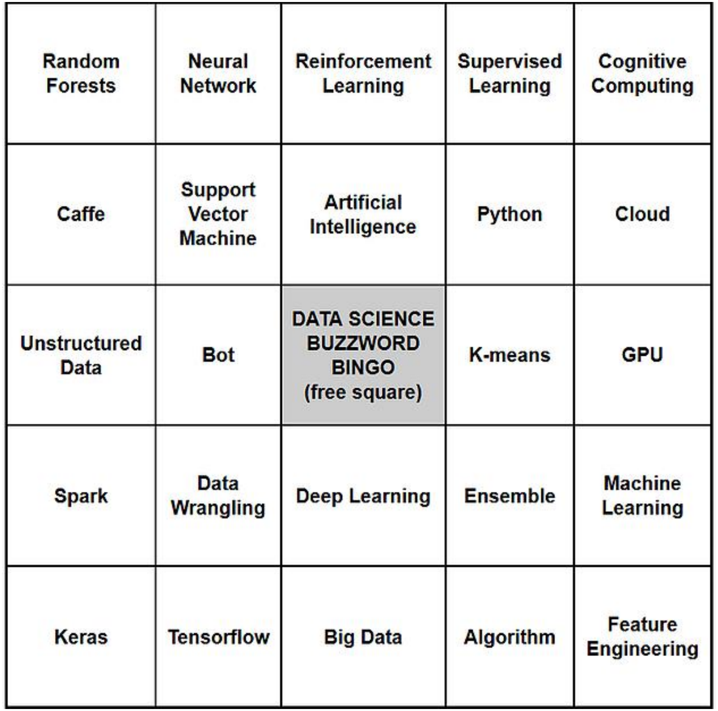
mattturck.com/bigdata2017

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Buzzword bingo

- Big Data
- Data Mining (datizrace)
- Machine Learning – subset of AI
- Data Science – statistics

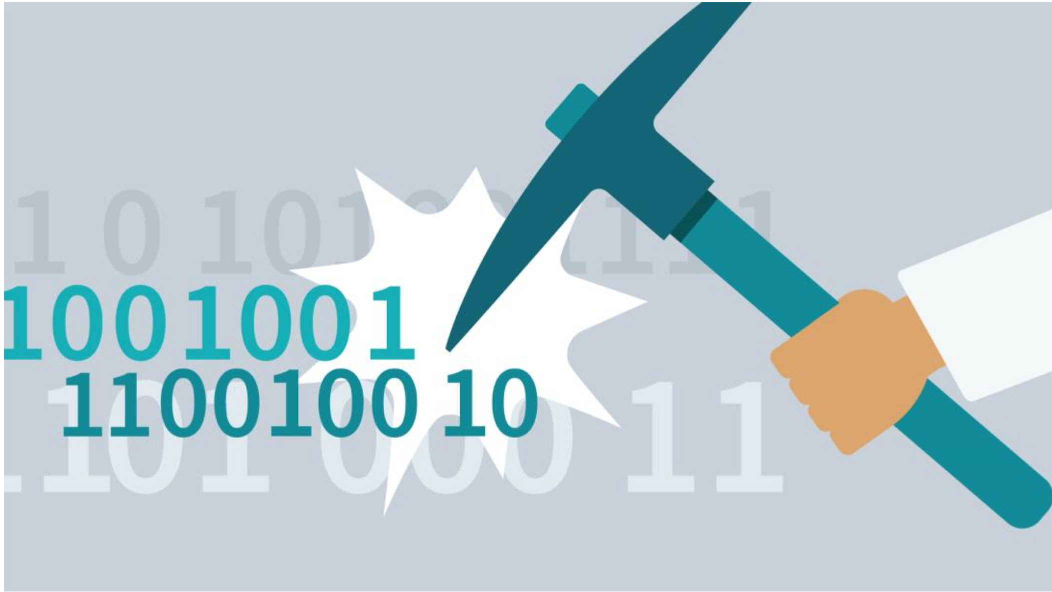
- Big Data or Pokemon
- https://pixelastic.github.io/pokemonorbigdata/



Random Forests	Neural Network	Reinforcement Learning	Supervised Learning	Cognitive Computing
Caffe	Support Vector Machine	Artificial Intelligence	Python	Cloud
Unstructured Data	Bot	DATA SCIENCE BUZZWORD BINGO (free square)	K-means	GPU
Spark	Data Wrangling	Deep Learning	Ensemble	Machine Learning
Keras	Tensorflow	Big Data	Algorithm	Feature Engineering

Data Mining

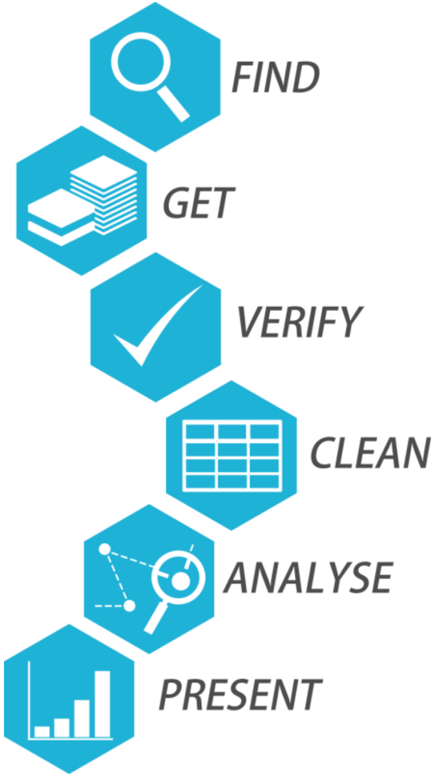
- Anomalies
- Classification
- Clusters
- Dimension Reduction
- Regression
- Relationship finding
- Summarization / Visualization



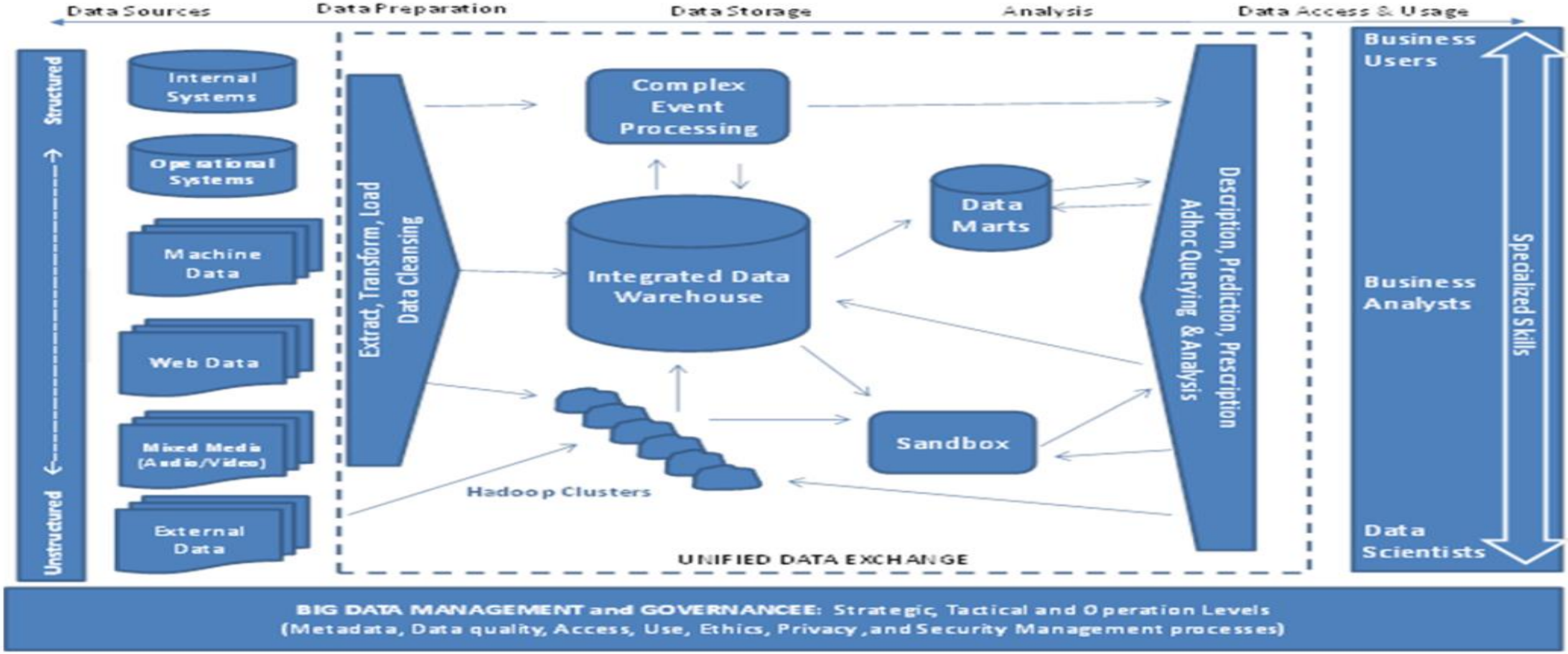
Building a Pipeline

- Data Security
- Cleanup
- Organising Database
- Analysis
- Visualization – Dashboard

- Emphasis on Analysis less on Infrastructure



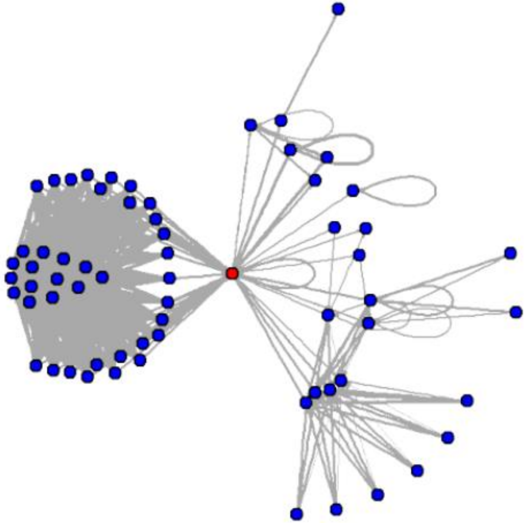
Full Analysis Framework



Sandbox for solutions

Common Projects

- Recommendation System / Churn Prediction
- Web Comments Sentiment Analysis
- Network Analysis (such as transactions on blockchain)

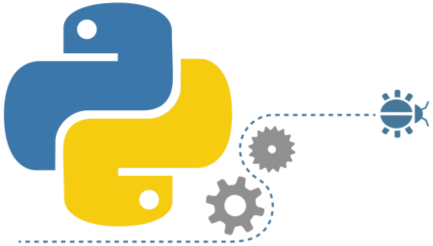


Avoiding Pitfalls

- Overfitting
- Data Dredging / p-hacking
- https://xkcd.com/882/

Python Programming Language

- #3 in TIOBE language index
- #1 in Popularity of Programming Languages index
- General purpose / readable
- 2018 Nobel Economist Paul Romer recommended!

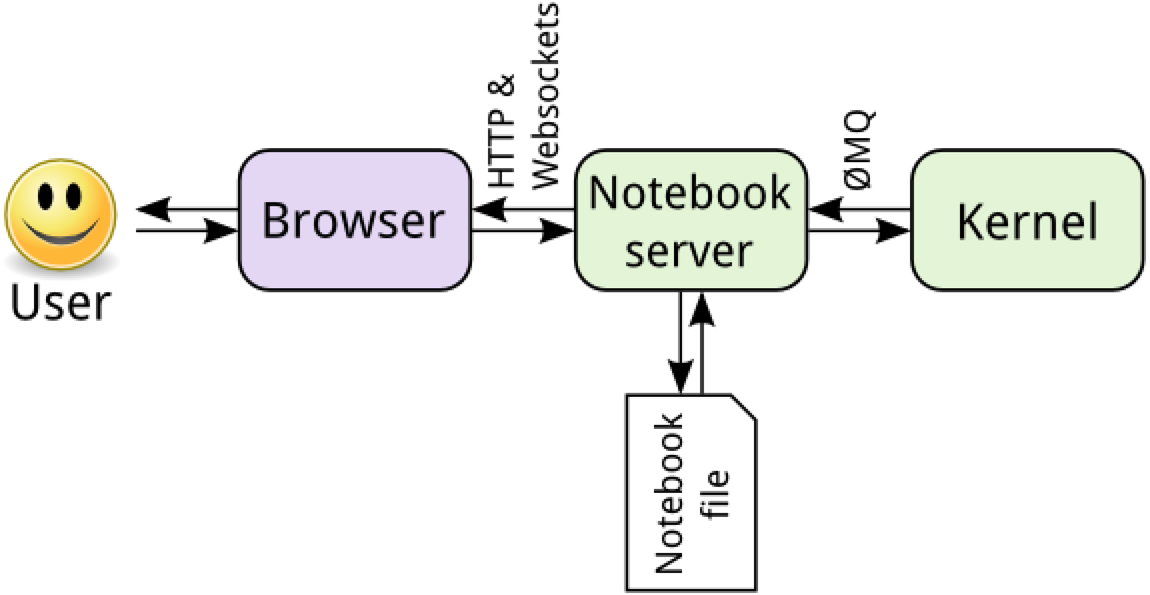


Resources

- Anaconda - **The Most Popular Python Data Science Platform**

- (includes Jupyter) : https://www.anaconda.com/download/
- Github: https://github.com/ValRCS/RigaComm_DataAnalysis

How Jupyter Works



Jupyter Hosted in the Cloud

- Microsoft - https://notebooks.azure.com/
- Google - https://colab.research.google.com/notebooks/welcome.ipynb
- IBM - https://dataplatform.cloud.ibm.com/
- Anaconda Cloud : https://anaconda.org/
- MyBinder – https://mybinder.org/
- Self Hosted or just run locally by installing Anaconda