Data Science: state of the art

Valentin Kuznetsov, Cornell University

SOSC 2019

Who am I?

- Theoretical Physicists (neutrino oscillations) at Irkutsk Univ & JINR
- Particle Physicists (tracking, silicon detectors) at CERN
- PhD in Physics (theory + experiment) at JINR
- Computing in HEP at JINR, CERN, Fermilab, Cornell
  - HEP experiments: NOMAD, D0, Cleo-c, CMS
- Data Scientists at Cornell University
  - Data management, data discovery, services
  - BigData, Analytics, Monitoring, Machine Learning

Introduction

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY | SAVE | SHARE | COMMENT 5 | TEXT SIZE | PRINT | $8.95 BUY COPIES

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at

WHAT TO READ NEXT

Big Data: The Management Revolution

5 Essential Principles for Understanding Analytics

Data Scientists Don't Scale

VIEW MORE FROM THE

October 2012 Issue

Josh Wills
@josh_wills

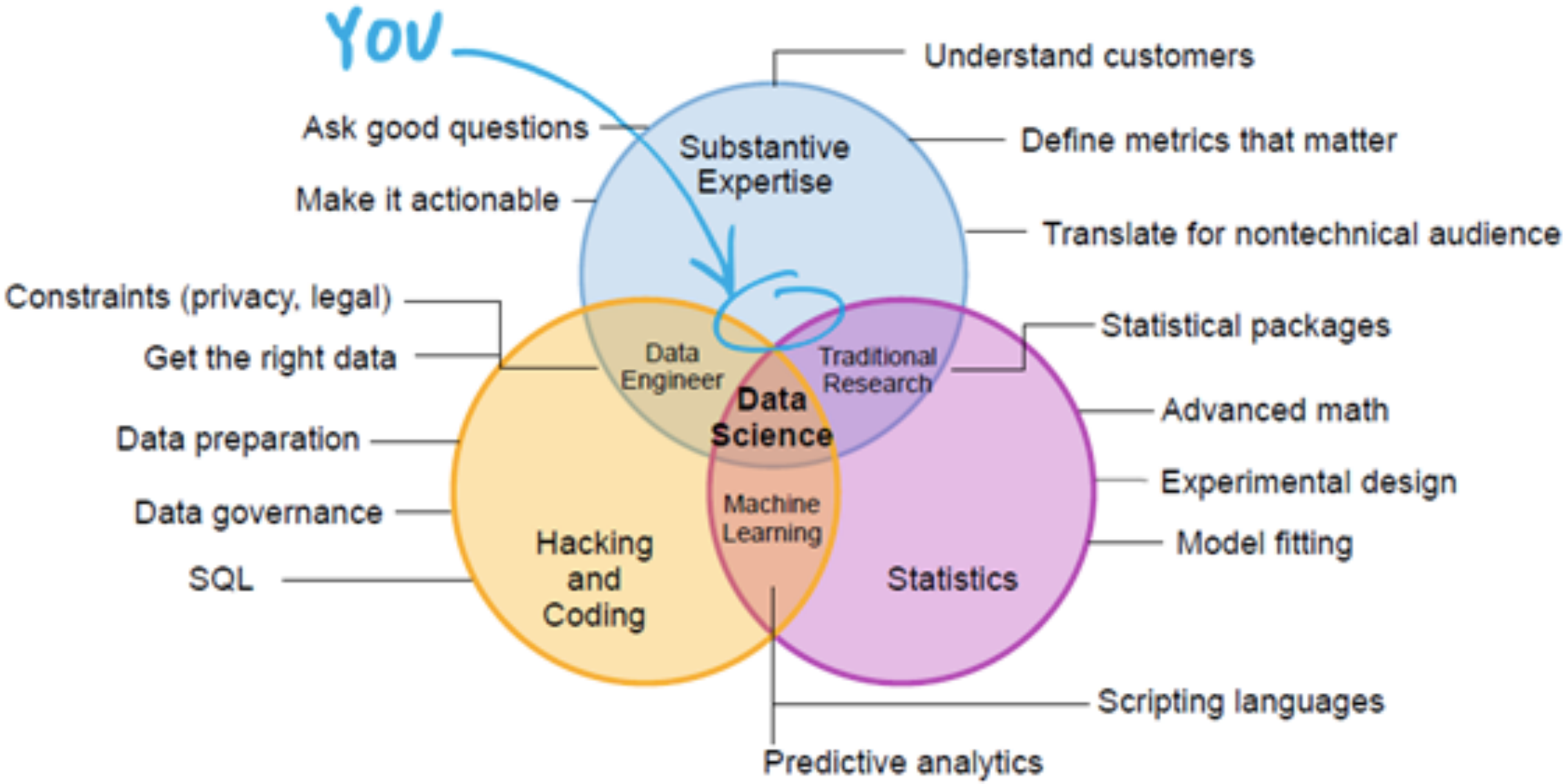
Following

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

RETWEETS 1,381 LIKES 857

6:55 PM - 3 May 2012

Data Science Is Multidisciplinary



Computer Science

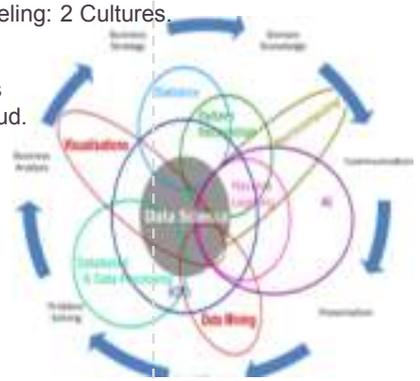
- Liebniz – Binary Logic.
- Babbage, Lovelace
- Boolean Algebra
- Punch cards.
- Turing machines
- Information Theory
- Weiner & Cybernetics
- Von Neumann Architecture.
- First IBM Computers
- DBMS.
- Sort & Search Algorithms – Dijkstra, Kruskal, Shell Sort, …
- Heuristics – Simulated Annealing, …
- Graph Algorithms
- Multigrid methods
- Tree based methods.
- Text/ string search
- 1974 Peter Naur "Concise Survey of Computer Methods", **Data Science, Datalogy**
- Knuth – Art of Computer Programming.
- 1989 First KDD Workshop
- Gregory Piatetsky-Shapiro.
- Database Marketing
- Data Mining, Knowledge Discovery
- "Data science, classification, and related methods."

Data Technology

- Catrography
- Astronomical Charts.
- William Playfair
- Charles Minard
- Florence Nightingale.
- Removable Disk drives
- Relational DBMS.
- John Tukey
- Jacques Bertin.
- Desktop, floppy
- SQL, OOP
- High level languages.
- William Cleveland: Data Science
- Leo Breimann: Statistical Modeling: 2 Cultures.

Visualization

- Edward Tufte.
- Grammar of Graphics
- Word Cloud, Tag Cloud.



Mathematics/ OR

- Calculus
- Logarithms
- Newton-Raphson.
- Optimization Methods
- Fourier and other transforms
- Matrix & Generalizations
- Non-euclidean geometries.
- Applications to Military, manufacturing, Communications.
- Networks
- Assignment Problems
- Automation
- Scheduling

Statistics

- Probability
- Correlation
- Bayes Theorem.
- Regression, Least Squares
- Time Series.
- Theoretical Foundations of Modern Stats
- Hypothesis, DOE
- Mathematical Statistics.
- **1962** John W. Tukey, Future of Data Analysis
- Bayesian Methods
- Time Series Methods (Box Cox, Survival, etc.)
- Stochastic Methods.
- 1976 – SAS Institute
- 1977 The International Association for Statistical Computing (IASC).
- Simulation, Markov
- Computational Statistics.
- Decision Science
- Pattern recognition
- Machine learning.

Pre 1800s | 1800-1900 | 1900-1940 | 1940-1960 | 1960 | 1970 | 1980 | 1990 | 2000 | 2010

DATA & AI LANDSCAPE 2019

INFRASTRUCTURE

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

HADOOP IN THE CLOUD
aws, Microsoft Azure, Google Cloud, SAP Cloud Platform, IBM InfoSphere BigInsights, arm Treasure Data, Qubole, CAZENA

STREAMING / IN-MEMORY
Amazon Kinesis, Google Cloud Dataflow, databricks, SAP Cloud Platform, Oracle Coherence, confluent, striim, hazelcast, GridGain, GIGASPACES, Wallaroo Labs, FASTDATA, kx

NoSQL DATABASES
Google Cloud, aws, ORACLE, Microsoft Azure, mongoDB, MarkLogic, Couchbase, DATASTAX, redislabs, AEROSPIKE, ArangoDB, SCYLLA

NewSQL DATABASES
SAP HANA, Clustrix, Pivotal, nuodb, MariaDB, Cloud Spanner, MEMSQL, influxdata, TIMESCALE, Cockroach Labs, VOLTDB, splice machine, imply, paradigm4, TIDB

GRAPH DBs
neo4j, Amazon Neptune, IBM, ORACLE, OrientDB, InfiniteGraph, Objectivity

MPP DBs
TERADATA, VERTICA, IBM Data Warehouse Systems, Actian, Kognitio, Exasol, dremio, Yellowbrick

CLOUD EDW
aws, Google Cloud, Microsoft Azure, Pivotal, snowflake, Infoworks

SERVERLESS
Microsoft Azure, AWS Lambda, Google Cloud Functions, PULSAR, nuclio, Pivotal Function Service

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

DATA INTEGRATION
SAP Data Services, Informatica, MuleSoft, TEALIUM, snapLogic, enigma, Qlik Data Catalyst, Segment, ATTUNITY, xplenty, ZALONI, import.io, Infoworks, Fivetran, SNOWPLOW, MATILLION

DATA GOVERNANCE
Informatica, IBM, SailPoint, McAfee Skyhigh Security Cloud, collibra, dremio, Alation, Waterline Data, IMMUTA, OKERA, MANTA, data.world

MGMT / MONITORING
aws, New Relic, actifio, rubrik, APPDYNAMICS, DATADOG, dynatrace, WAVEFRONT by vmware, SignalFx, druva, splunk>, Moogsoft, pagerduty, Anodot, unravel, Numerify, SCALYR, VEEAM, zenoss, OpsRamp, MAGNITUDE

STORAGE
aws, Google Cloud, Microsoft Azure, IBM Storage, PURESTORAGE, ALLUXIO, wasabi technologies, nimblestorage, Qumulo, panasas, COHESITY

CLUSTER SVCS
Amazon ECS, Amazon EKS, IBM Spectrum Conductor, IBM Spectrum Symphony, MESOSPHERE, Cluster Manager, packet, Hewlett Packard Enterprise, PENGUIN COMPUTING, Bright Computing, CYCLECLOUD

DATA GENERATION & LABELLING
amazon mechanical turk, Upwork, unity, appen, scale, HIVE, Labelbox, Mighty AI, LIONBRIDGE

AI OPS
ALGORITHMIA, SPELL, comet, Verta.ai, datmo, datatron, Weights & Biases, Determined AI, fiddler

GPU DBs & CLOUD
kinetica, omnisci, SQREAM, blazegraph, brytlyt, BLAZINGDB, PG-Strom, FLOYDHUB, Paperspace

HARDWARE
Google TPU, arm, intel AI, NVIDIA, IBM Power Systems, GRAPHCORE, MYTHIC, Horizon Robotics, Cerebras, Movidius, habana, WAVE COMPUTING, Esperanto, CEREBRAS, HAILO, flexlogix, EFINIX

ANALYTICS & MACHINE INTELLIGENCE

DATA ANALYST PLATFORMS
Microsoft, pentaho, alteryx, Digital Reasoning, guavus, AYASDI, ATTIVIO, Datameer, incorta., interana, MODE, ENDOR, ASG2360.IO, sisu, switchboard, Starburst

DATA SCIENCE PLATFORMS
IBM, databricks, data iku, DOMINO, rapidminer, TIBCO, ANACONDA, sas, Altair Knowledge Works, KNIME, MathWorks

BI PLATFORMS
looker, salesforce einstein analytics, aws, DOMO, ARCADIA DATA, ThoughtSpot, ATSCALE, SISENSE, Qlik, GoodData, Information Builders, birst, MicroStrategy, Keen IO

VISUALIZATION
tableau, Microsoft Power BI, SAP Lumira, Google Cloud, celonis, Periscope Data, zepl, GoMD ATA, plotly, CHARTIO, TOUCAN TOCO

MACHINE LEARNING
Azure Machine Learning, Amazon SageMaker, Google Cloud AutoML Vision, H2O.ai, DataRobot, gamalon, ViSENZE, ELEMENT AI, deepsense.ai

COMPUTER VISION
Microsoft Azure, Amazon Rekognition, Cloud Vision API, clarifai, 商汤, EVER AI, deepomatic, neurala, twenty bn, UBIQUITY6, AiCBE, 依图 | YITU, trax, Clarifai, synthesia, DataGrid

HORIZONTAL AI
IBM Watson, Cortana, Face++ 旷视, sentient Technologies, VoyagerLabs, vicarious, :) Affectiva, PROPHESEE, CognitiveScale, Numenta, PETUUM, SCALED INFERENCE, narrative logics, THE CURIOUS AI COMPANY, OSARO, BLUE VISION, Fortress

SPEECH & NLP
Google Cloud, twilio, amazon alexa, Amazon Translate, narrative science, semantic machines, Wolfram Alpha, Mobvoi, EigenTechnologies, SoundHound Inc., voicera, PRIMER, MindMeld, cortical.io, NUANCE, Gridspace, cogito, snips, Amplify Analytics, SMARTLING, Unbabel, PolyAI

SEARCH
elasticsearch, ORACLE ENDECA, algolia, coveo, Lucidworks, ATTIVIO, swiftype, EXALEAD, alphasense, MAANA, omni:us, SINEQUA

LOG ANALYTICS
splunk>, sumologic, solarwinds loggly, TIMBER, kibana, logz.io

SOCIAL ANALYTICS
Hootsuite, sprinklr, NETBASE, synthesio, tracx, simplereach, bitly, SimilarWeb

WEB / MOBILE / COMMERCE ANALYTICS
Google Analytics, mixpanel, AMPLITUDE, Airtable, RESCI, SIGOPT, granify, custora

APPLICATIONS – ENTERPRISE

SALES
einstein, CHORUS, INSIDESALES.COM, peopleai, conversica, GONG, clari, aviso, tact.ai, TROOPS, fuse|machines, Clearbit, collective[i]

MARKETING - B2B
RADIUS, App Annie, EVERSTRING, Lattice, MINTIGO, 6sense, tubular, RK Reflektion, NGAGIO, KNOTCH, mrp

MARKETING - B2C
zeta, bloomreach, SendGrid, braze, ACTIONIQ, BLUECORE, CONTENTSQUARE, TEALIUM, mparticle, Amplero, amperity, QUANTIFIND, Simon, Lytics, [PERSADO], remesh

CUSTOMER EXPERIENCE / SERVICE
qualtrics, MEDALLIA, SurveyMonkey, UserTesting, CLARABRIDGE, zendesk, Kustomer, freshdesk, INTERCOM, Drift, LIVEPERSON, Gainsight, pendo, HEAP, Amplitude, Watson Assistant, Dialogflow, DigitalGenius, ASAPP, ada, AUTOMAT, afiniti, CaliDesk, netomi, TechSee, frame.ai

ENTERPRISE PRODUCTIVITY
slack, ORACLE, GURU, lumiata, DIFFBOT, clara, talla, Kasisto

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

LEGAL
RAVEL, Seal, Everlaw, DISCO, kira, JUDICATA, eBREVIA, IRONCLAD, PREMONITION, ROSS, casetext

REGTECH & COMPLIANCE
BigID, TESSIAN, text IQ, Comply Advantage

PARTNERSHIPS
CROSSBEAM, INFOSUM, DATA REPUBLIC

FINANCE
Anaplan, zuora, SAP S/4 HANA, TRADESHIFT, mineraltree, SCALEFACTOR, bookkeeper, pilot

BACK OFFICE AUTOMATION & RPA
UiPath, HyperScience, blueprism, VIDADO, AppZen, WorkFusion, workato, Redwood, Catalytic, ANTWORKS, softomotive, KRYON, ALKYMI

SECURITY
TANIUM, CYLANCE, zscaler, StackPath, illumio, CODE42, CipherCloud, DARKTRACE, ANOMALI, ThreatMetrix, VECTRA, CLOUDFLARE, Guardian Analytics, DATAVISOR, siftscience, pindrop, exabeam, SICNIFY, SentinelOne, SecurityScorecard, SOCURE, VadeSecure, bitglass, BlueTalon, Recorded Future, feedzai, CyberX, BITSIGHT, AREA 1 SECURITY, sparkcognition, IronNet Cybersecurity, FORTER, riskrecon, JASK, BLUEHEXAGON, Semmle, OBSIDIAN, AXONIUS, SHIELD IQ, Armorblox

APPLICATIONS – INDUSTRY

ADVERTISING
AppNexus, MediaMath, criteo, xAd, Integral Ad Science, ORACLE MOAT, OpenX, datorama, theTradeDesk, Adgorithms, dstillery, LiveIntent, TAPAD, dataxu, gumgum, Appier, teemo, yieldmo, mobilewalla

EDUCATION
Liulishuo, 猿辅导, KNEWTON, Clever, Declara, kidaptive, PANORAMA, knowre, gradescope

REAL ESTATE
REDFIN, Opendoor, VTS, CREDiFi, GEOPHY, reonomy, COMPSTAK, SPACEMAKER, SKYLINE

GOV'T
OPENGOV, mark43, FiscalNote, LiveStories, Passport, SmartProcure, STREETLIGHTDATA, OpenDataSoft

INTELLIGENCE
Palantir, Dataminr, Quid, PRIMER, FORGE

FINANCE - INVESTING
KENSHO, Quantopian, ADDEPAR, NUMERAI, iSENTIUM, ALGORIZ, RavenPack, PAGAYA

FINANCE - LENDING
ondeck, affirm, 拍拍贷, ppdai.com, JIANPU.AI, Kreditech, AVANT, aura, finance, Upstart, LEARBANC, upgrade, 100Credit, WeLab, Wecash闪现白骑, TrueAccord, MoneyLion, Active.AI, aire, cignifi

INSURANCE
Metromile, Lemonade, CYENCE, Hippo, Shift Technology, ROOT, zesty.ai, TRACTABLE, ARTURO, CAPE

HEALTHCARE
flatiron, Clover, KYRUS, HealthTap, METABIOTA, Ginger.io, Glow, babylon, 3DMed, zebra, PathAI, ovia, TEMPUS, patientslikeme, AiCure, insitro, LinkDoc, SIG TUPLE, Gi.ai, CITIZEN, notable, komodoHealth, RECURSION, prognos, enlitic, imagia, BlackThorn, BAYLABS, Qventus, ARTERYS, IMAGEN, CLOUD MEDX, PAIGE, DATAVANT, innoviccer, DIAGNOSIS PROJECT, LearnTaoS

LIFE SCIENCES
Z3andMe, color, iCarbonX, verily, BenevolentAI, WuXiNextCODE, zymergen, Pathway Genomics, SCAN, Clear Labs, freenome, NANOPORE, DNAnexus, Phosphorus, CITRINE, twoXAR, Atomwise, deep genomics, SOPHIA GENETICS, OWKIN

TRANSPORTATION
UBER, TESLA, WAYMO, ZOOX, CLEARPATH, cruise, nuro, ARGO, nuTonomy, drive.ai, CAMBRIDGE MOBILE TELEMATICS, Aurora, nauto, AMOTIVE, G7, PILOT.AI, NIO, OPTIMUS RIDE, moovit, Ike, nexar, Kodiak, comma.ai, netradyne, German Autolabs, Civil Maps, cognata, thinci, INRIX

AGRICULTURE
FARMERS BUSINESS NETWORK, Granular, JOHN DEERE, BLUE RIVER TECHNOLOGY, FarmersEdge, AgroStar, FarmLogs, TARANIS, GAMAYA, TerrAvion, prospera

COMMERCE
instacart, FAIRE, STITCH FIX, Dia & Co, RetailNext, HowGood, heuritech

INDUSTRIAL
AVEVA, SIEMENS, Schneider, PREDIX, UPTAKE, SIGHT MACHINE, KARGO, OSIsoft, SCORTEX, KONUX, fero labs, TACHYUS

OTHER
eharmony, stem, Amper, ByteDance, hopper, celect, SOJERN, BOXEVER, VERDIGRIS, duetto, Jukedeck, Second Spectrum, Electric, ZINIER, Spoke, Moveworks

CROSS-INFRASTRUCTURE/ANALYTICS
aws, Google Cloud, Microsoft, IBM, SAP, Hewlett Packard Enterprise, sas, 1010DATA, vmware, TIBCO, TERADATA, ORACLE, NetApp, syncsort, MAPR, cloudera

OPEN SOURCE

FRAMEWORKS
Hadoop, Spark, Hadoop MapReduce, Flink, YARN, TEZ, MESOS, kubernetes, docker, CDAP, Red Hat, High Availability, HELIX

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

DATA ACCESS & DATABASES
cassandra, mongoDB, redis, CouchDB, Cockroach Labs, druid, OPENTSDB, SciDB, riak, HBASE, Cloud Spanner, ACCUMULO

ORCHESTRATION & MGMT
talend, Apache Zookeeper, OOZIE, Apache Ambari, Apache Airflow, MESOS, etcd, Kong

STREAMING & MESSAGING
Spark, nifi, Flink, beam, kafka, STORM, Apache RocketMQ

STAT TOOLS & LANGUAGES
python, R, Scala, NumPy, Pyro, Studio, SciPy, julia

AI OPS & INFRA
Pachyderm, mlflow, Kubeflow, mleap, DVC, SELDON, Polyaxon, seldon

AI / MACHINE LEARNING / DEEP LEARNING
TensorFlow, Keras, torch, MADlib, learn, Caffe, Microsoft Cognitive Toolkit, OpenAI, DMTK, theano, PaddlePaddle, Apache SINGA, DIMSUM, FeatureFu, mxnet, VELES, Chainer, Michelangelo, ONNX, WEKA, LUDWIG, PyTorch, neon, DSSTNE, mllib, DL4J, MAHOUT, Aerosolve, fast.ai, mlr, OpenML

SEARCH
elasticsearch, Solr, Lucene

LOGGING & MONITORING
elasticsearch, kibana, SENTRY, logstash, Prometheus, fluentbit, fluentd, Grafana, Timber.io VECTOR

VISUALIZATION
matplotlib, TensorBoard, seaborn, Bokeh

COLLABORATION
BeakerX, jupyter, Zeppelin, ANACONDA

SECURITY
Apache Ranger, KNOX, Sentry, ACCUMULO

DATA SOURCES & APIs

HEALTH
validic, practice fusion, fitbit, GARMIN, HUMAN API, kinsa, MIMIC

IOT
GE Digital, UPTAKE, thingworx, helium, samsara, estimote

FINANCIAL & ECONOMIC DATA
Bloomberg, THOMSON REUTERS, DOW JONES, S&P CAPITAL IQ, CB INSIGHTS, PLAID, SECOND MEASURE, ENVESTNET YODLEE, THE WORLD BANK, estimize, PREMISE, Quandl, Eagle Alpha, StockTwits, xignite, Thinknum, earnest, predata

AIR / SPACE / SEA
Orbital Insight, planet., SKYDATA, AIROBOTICS, spire, PRECISIONHAWK, Descartes Labs, kespry, UNDERSTORY, tellusLabs, WINDWARD, DroneDeploy, MarineTraffic, LOFT ORBITAL, RS Metrics

PEOPLE / ENTITIES
acxiom, experian, EPSILON, InsideView, Crimson Hexagon, BASIS TECHNOLOGY, Quantcast, SAFEGRAPH

LOCATION INTELLIGENCE
FOURSQUARE, MapAnything, mapbox, sense360, pitney bowes, HEXAGON GEOSPATIAL, PlaceIQ, esri, factual, CARTO, Mapillary, Streetline, cuebiq, Radar, OpenStreetMap

OTHER
DATA.GOV, IMAGENET, wiki-links, LABELME, Berkeley DeepDrive, CROWDFLOWER, APOLLOSCAPE, CRUX, HIGHCHART, jgrafiti.io

DATA RESOURCES

DATA SERVICES
OPERA, Mu Sigma, SILICON VALLEY DATA SCIENCE, fractal, kaggle, DataKind, EXL, INNOPLEXUS

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

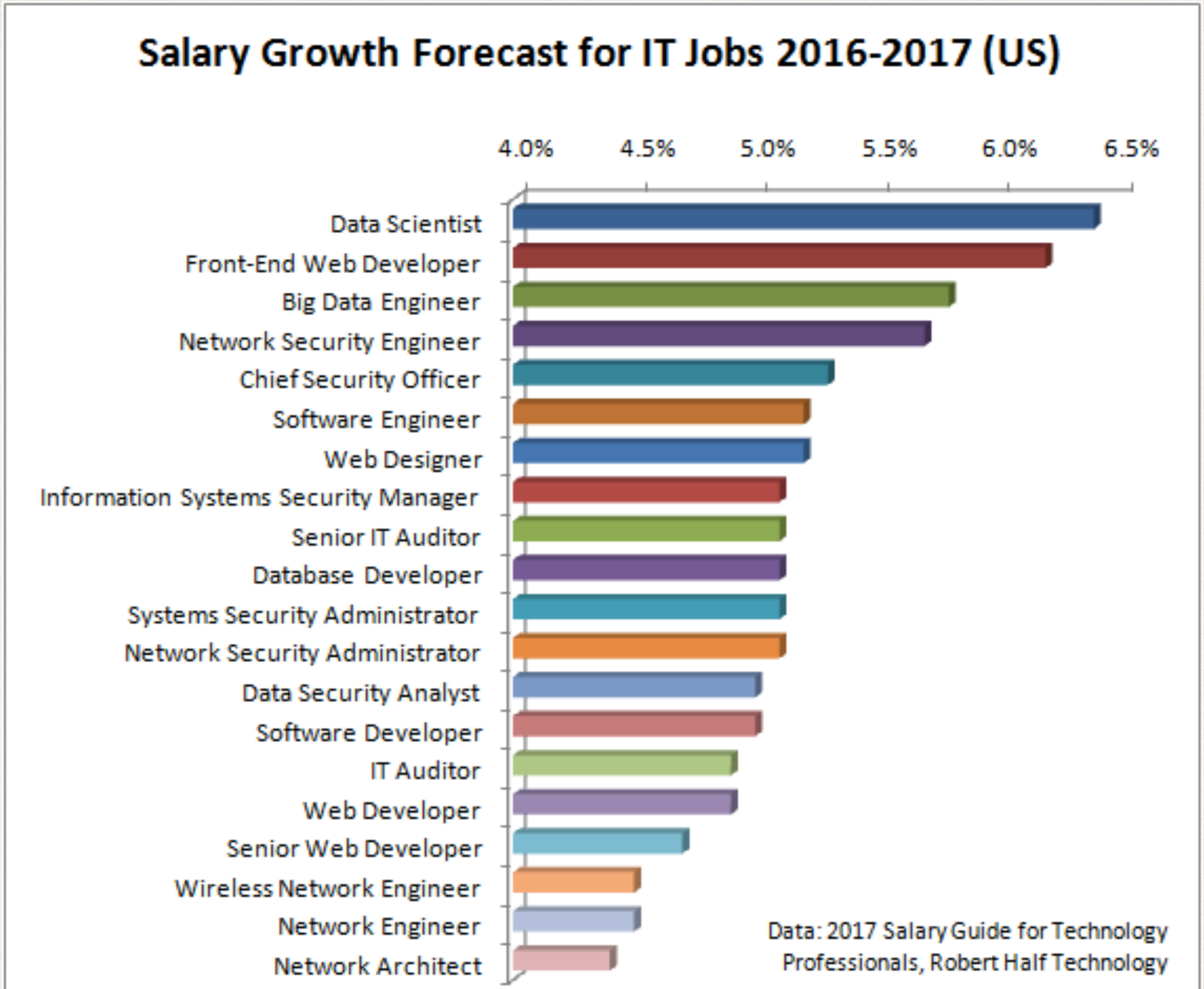
RESEARCH
OpenAI, facebook research, MILA, MIRI, VECTOR INSTITUTE, CSAIL, DFKI, OTA, AI2, ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

You need to know (my bare minimum)

- Math, statistics, algorithms, be able to read scientific paper
- Programming languages: C/C++, Python, R, Go, etc.
- Shell scripting and unix tools: bash, sed, awk, etc.
- How to build/install packages/tools
  - from source code: make, autoconf, environment, tar, etc.
  - from package management tools: rpm, yum, apt, dpkg, pip, anaconda, and/or build your favorite Linux distribution
- Versioning tools: git, gitlab, bitbucket, etc.
- Compilers, linkers, structure of libraries, object files, etc.
- Statistical and visualization tools: R, MatLab, Pandas, NumPy, SciPy, matplotlib, etc.
- ML tools: Scikit-Learn, R, TensorFlow, Keras, xgboost, etc.

You need to know, cont'd

- Platforms: AWS, Microsoft Azure, Google Cloud, etc.
- Cloud infrastructures: Docker, Kubernetes, etc.
- BigData tools: Hadoop, Spark, HDFS, HDF5, etc.
- Databases: ORACLE, MySQL, SQLite, NoSQL, GraphDB, MongoDB, CouchDB, etc.
- Monitoring: ElasticSearch, Kibana, Grafana, Prometheus, etc.
- Streaming: Spark, Kafka, Storm, etc.
- Collaboration: Jupyter, Zeppelin, Anaconda, SWAN, etc.
- Search: ElasticSearch, Lucene, Solr, etc.
- Lexical analysis & NLP: lexer, tokenizer, scanner, etc.
- Read, write, and ask questions about everything


Salary Growth Forecast for IT Jobs 2016-2017 (US)
4.0%
4.5%
5.0%
5.5%
6.0%
6.5%
Data Scientist
Front-End Web Developer
Big Data Engineer
Network Security Engineer
Chief Security Officer
Software Engineer
Web Designer
Information Systems Security Manager
Senior IT Auditor
Database Developer
Systems Security Administrator
Network Security Administrator
Data Security Analyst
Software Developer
IT Auditor
Web Developer
Senior Web Developer
Wireless Network Engineer
Network Engineer
Network Architect
Data: 2017 Salary Guide for Technology Professionals, Robert Half Technology

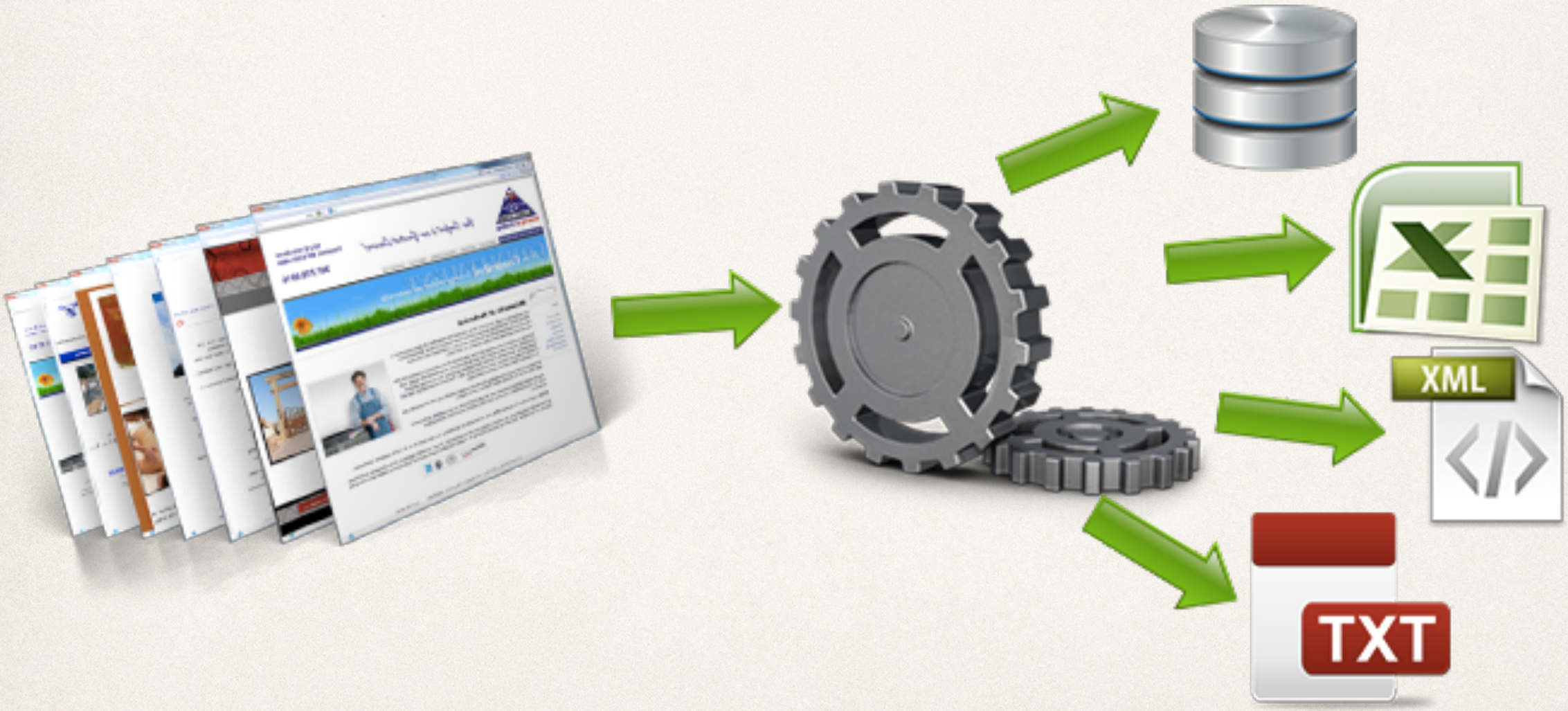
Engineering challenges of 21st century

National Academy of Engineering

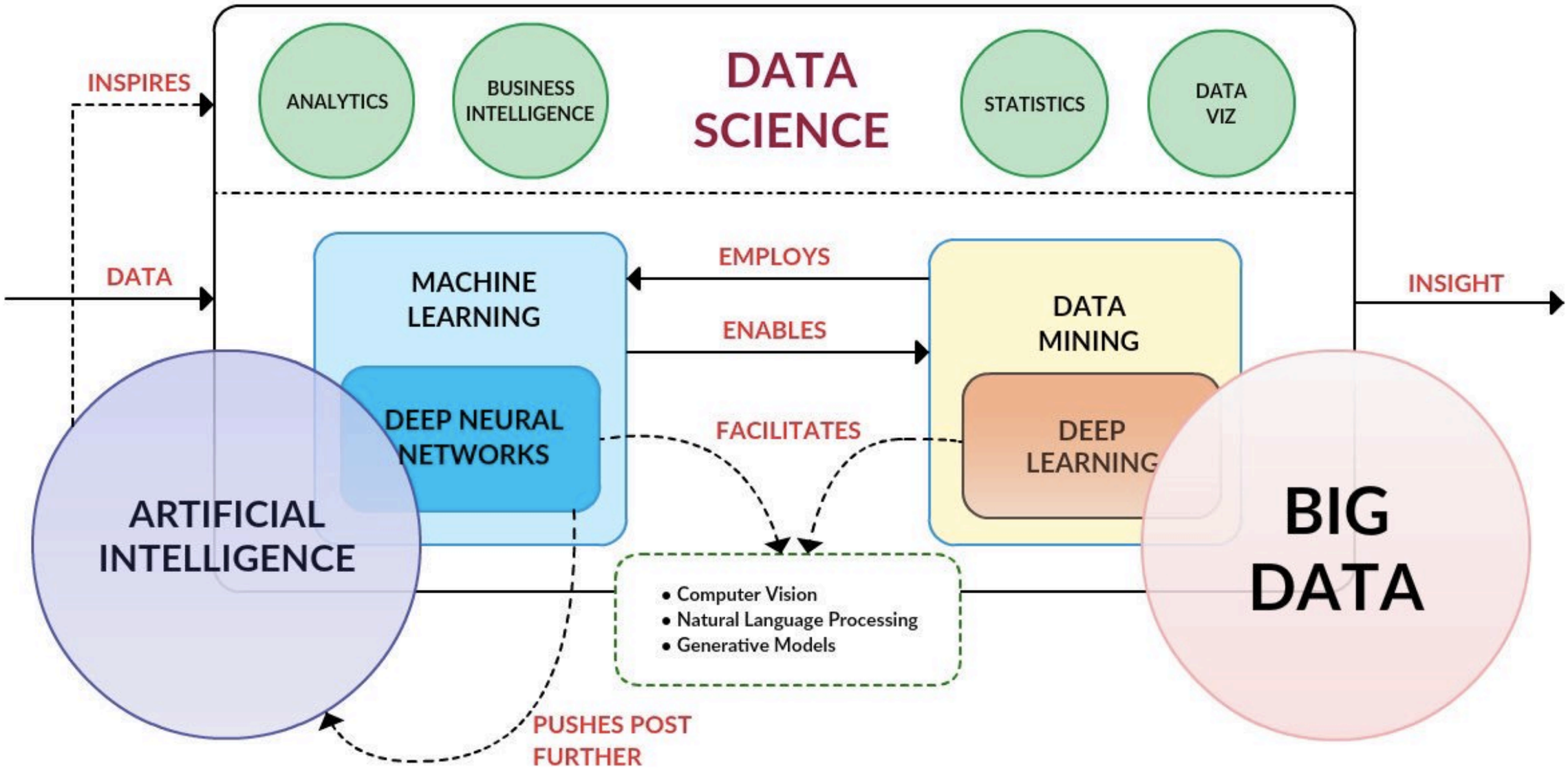
- Advance personalized learning
- Make solar energy economical
- Enhance virtual reality
- Reverse-engineer the brain
- Engineer better medicine
- Advance Health informatics
- Restore and improve urban infrastructure
- Provide access to clean water
- Secure Cyberspace
- Prevent Nuclear terror
- Manage the Nitrogen cycle
- Develop carbon sequestration methods
- Engineer tools of scientific discovery

Problem statement

Acquire the data



Process the data



Understand the data

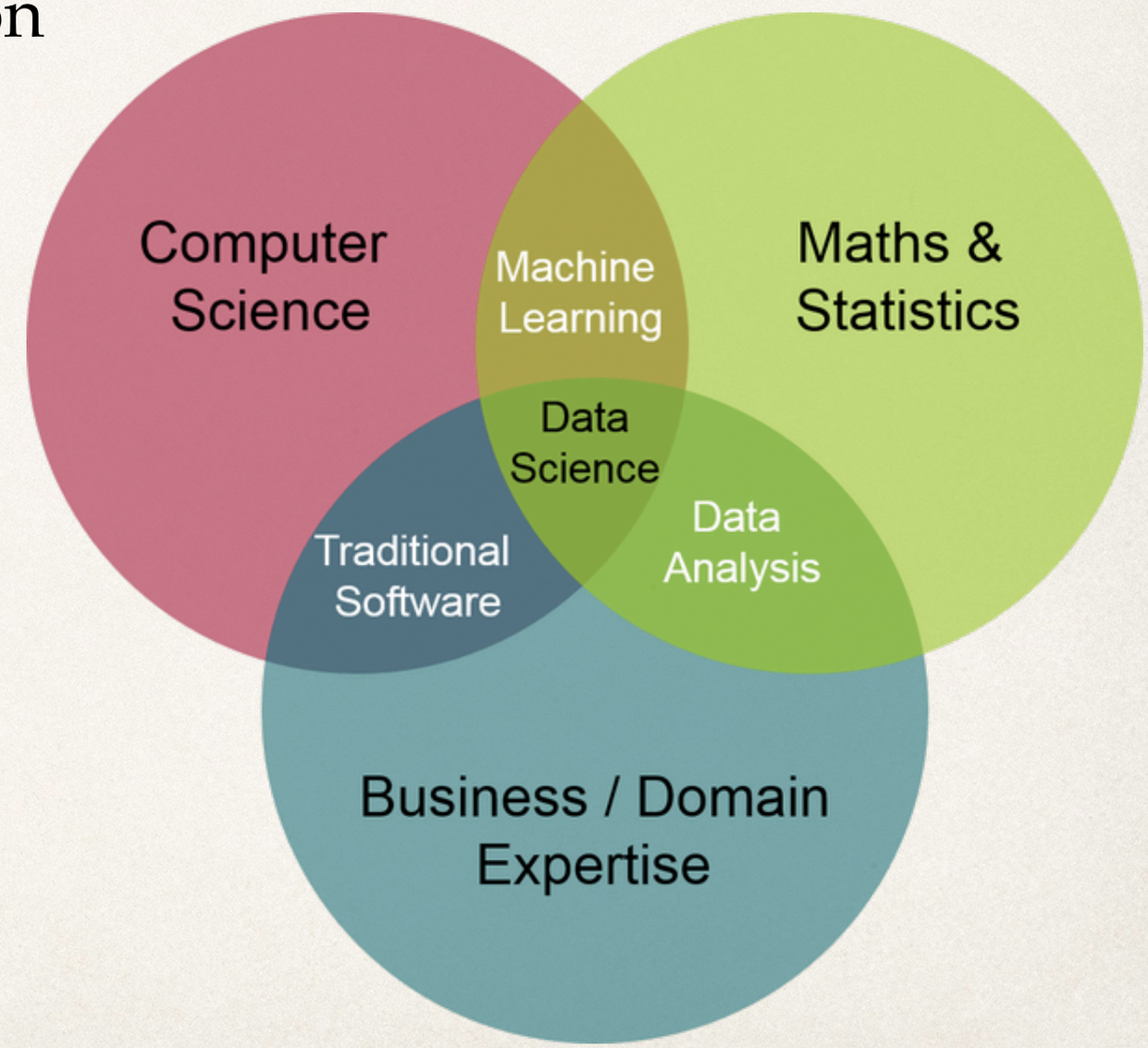
Actions

- Increase revenue
- Reduce operational costs
- Understand behavior
- Find anomalies
- Identify strategy
- Discover new features

Data Science

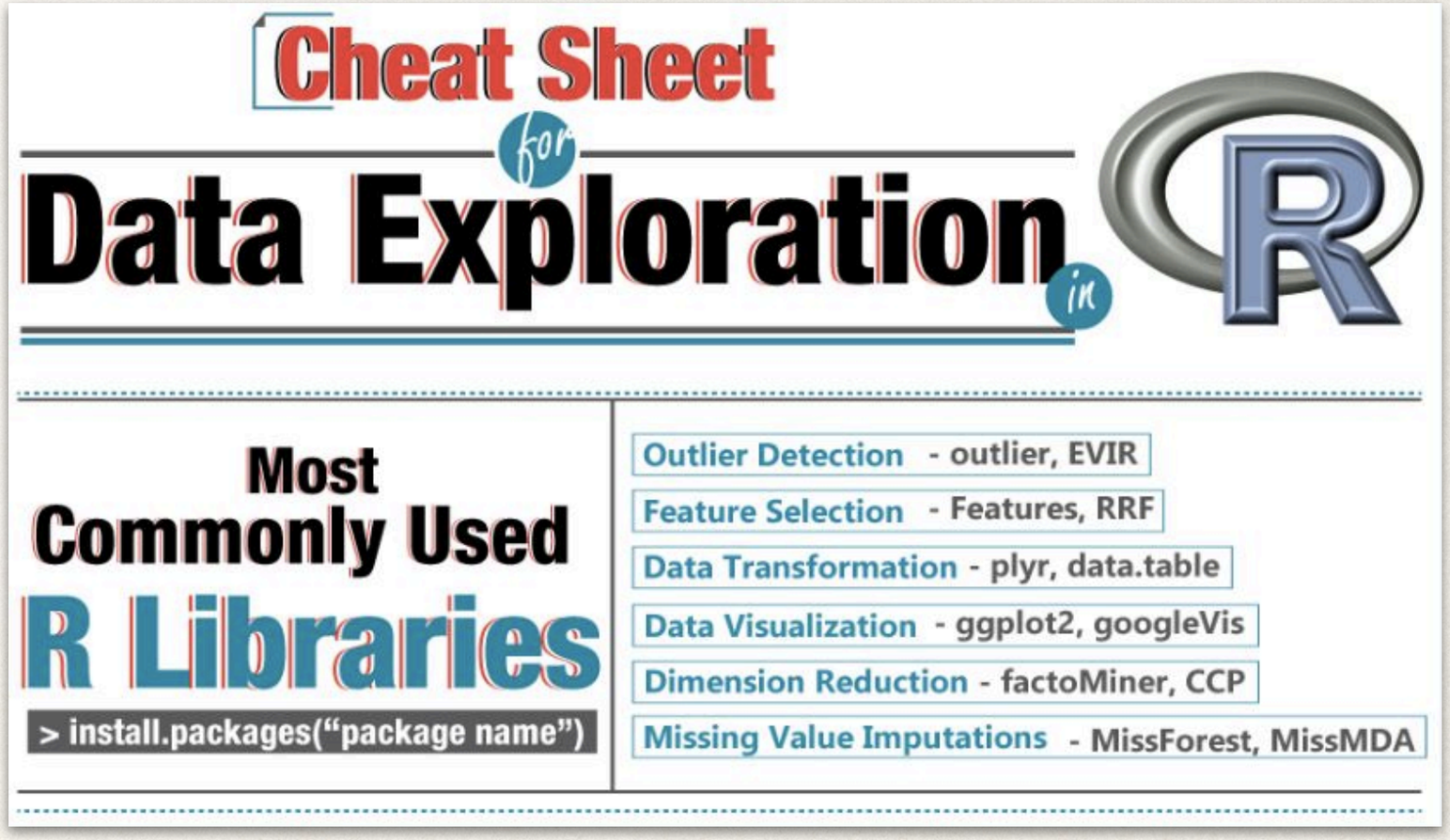
Scope of Data Science

- Data Exploration and Preparation
- Data Representation and Transformation
- Computing with Data
- Data Modeling
- Data Visualization and Presentation
- Science about Data Science



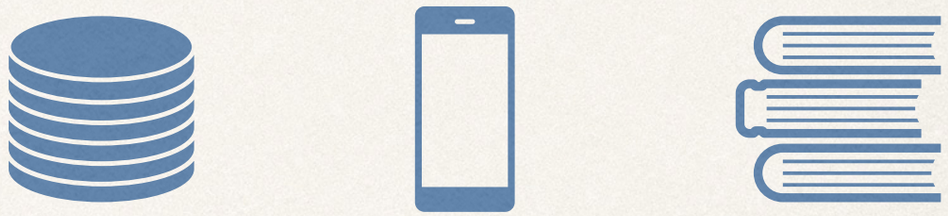
Data Exploration and Preparation

- Exploration
  - understand your data and perform Exploratory Data Analysis
- Preparation
  - perform data cleaning and understand anomalies and various artifacts



Data Representation and Transformation

- Data sources
  - data today comes from variety of sources, from home made txt files to SQL and noSQL databases, data streams, etc. Data Scientists need to know the structure, transformation, and algorithms to deal with modern data



- Data Transformation includes data cleaning, pruning, normalization, standardization, etc.

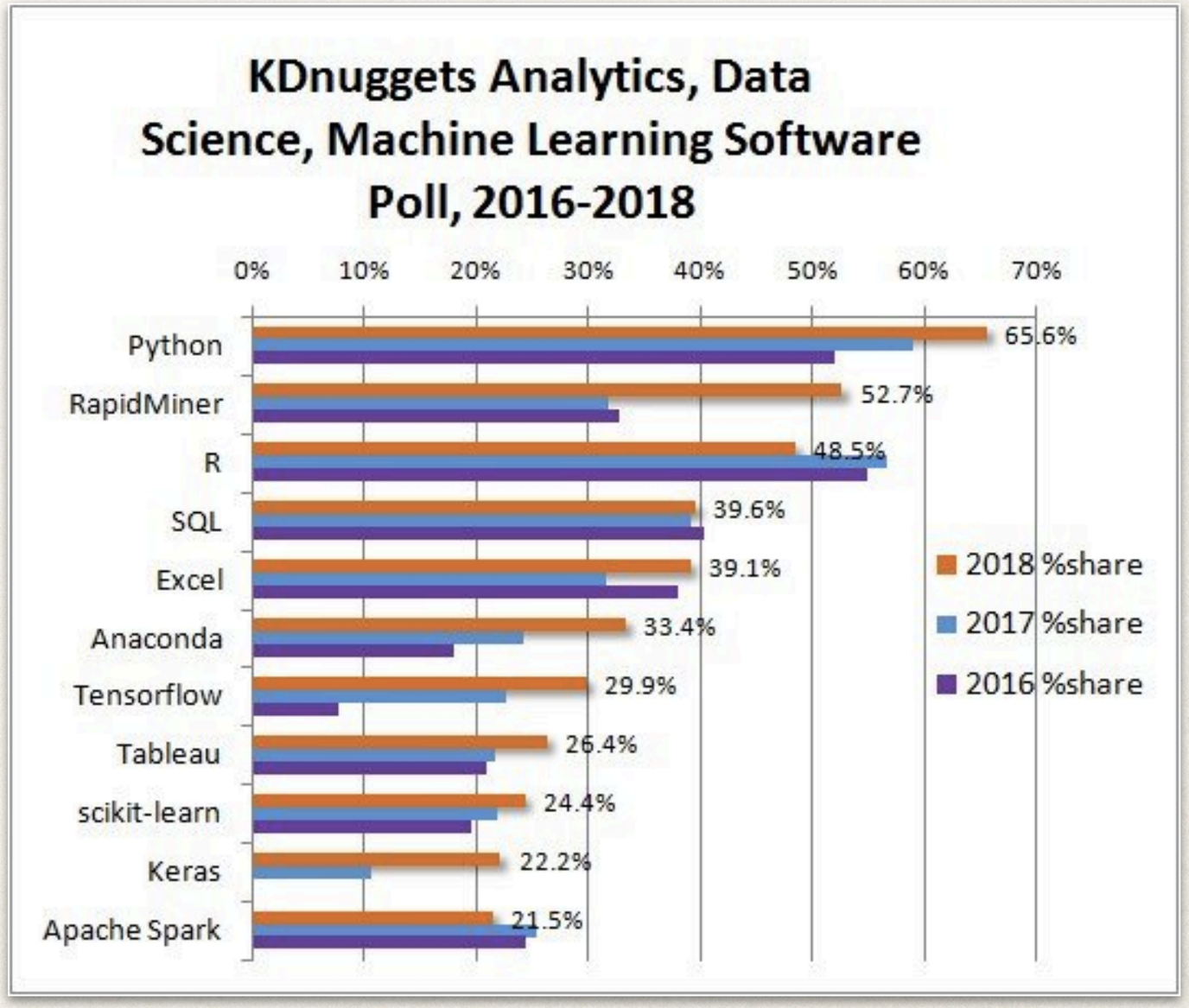
$$x' = \frac{x - \bar{x}}{\sigma} \qquad x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Data Representation
  - use different mathematical structures for data representation

Categorical vs Numerical
One-hot encoding
Leave-one out encoding
Word embedding

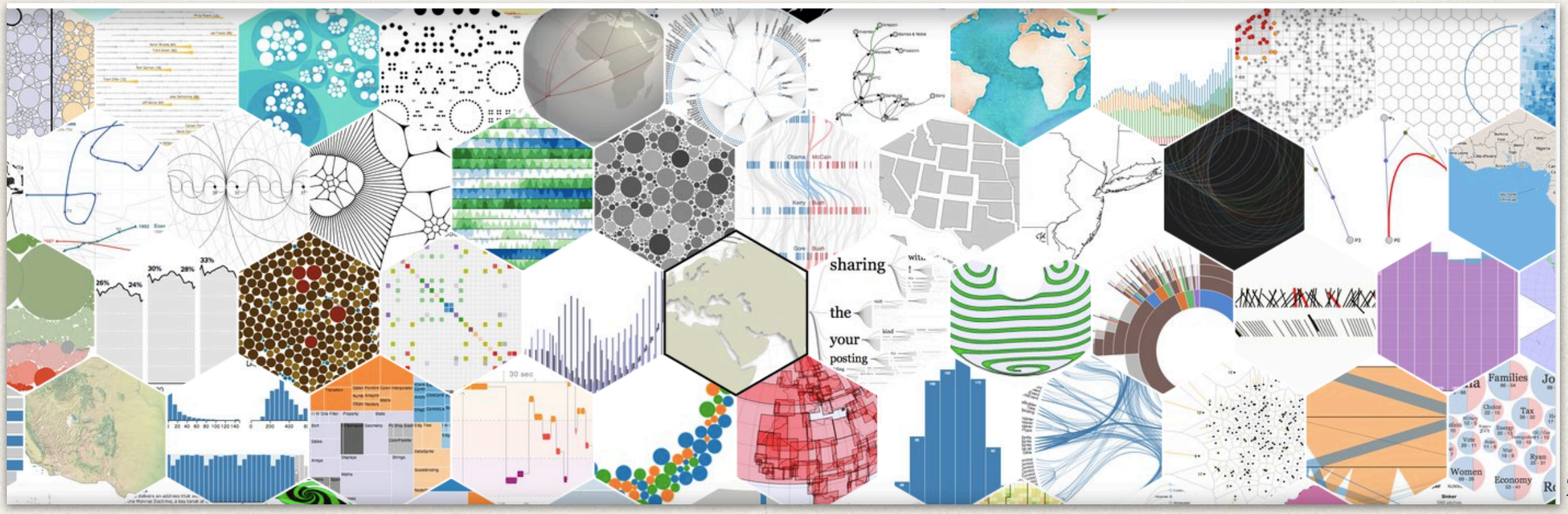
Computing with data

- Every Data Scientists should know several programming languages
  - scripting, specialized languages, general purpose languages
  - be fluent with cluster and cloud computing to run jobs over massive datasets
  - organize workflows, from scripting to reproducible notebooks



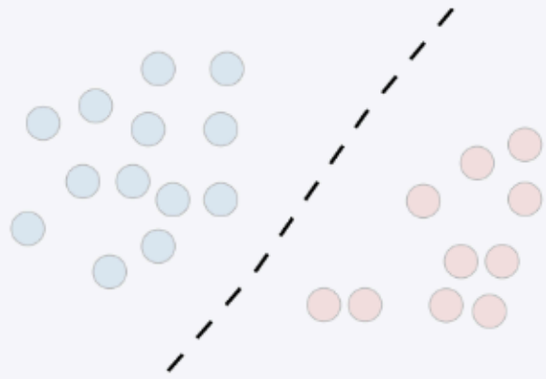
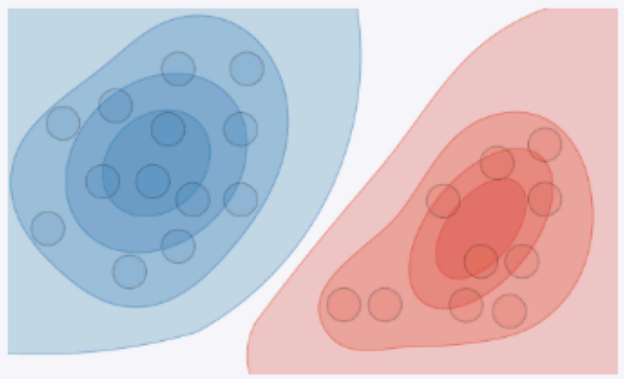
Data Visualization and Presentation

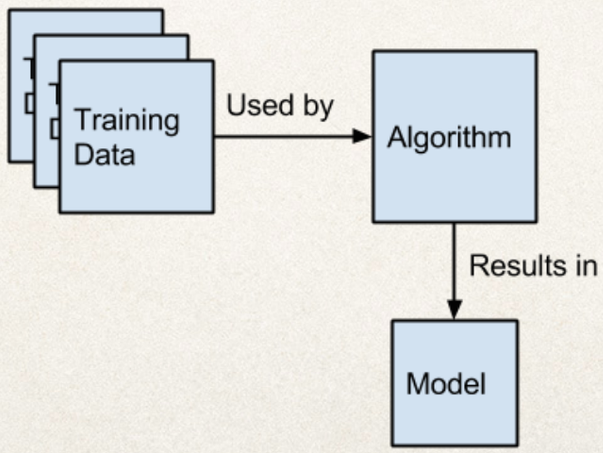
- Be able to tell the story
  - use histograms, scatterplots, time series plots, heat maps, etc.
  - use variety of visualization frameworks suitable for different needs

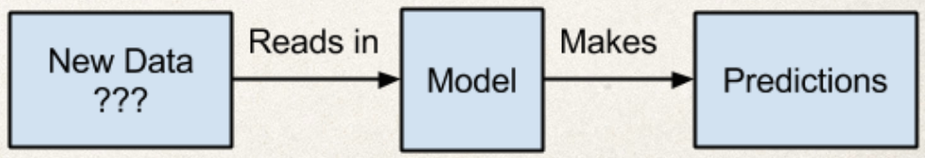


Data Modeling

- **Generative modeling**, in which one proposes a stochastic model that could have generated the data (domain of **Academic statistics**)
- **Predictive modeling**, in which one constructs methods which predict well over some given data universe (domain of **Machine Learning**)

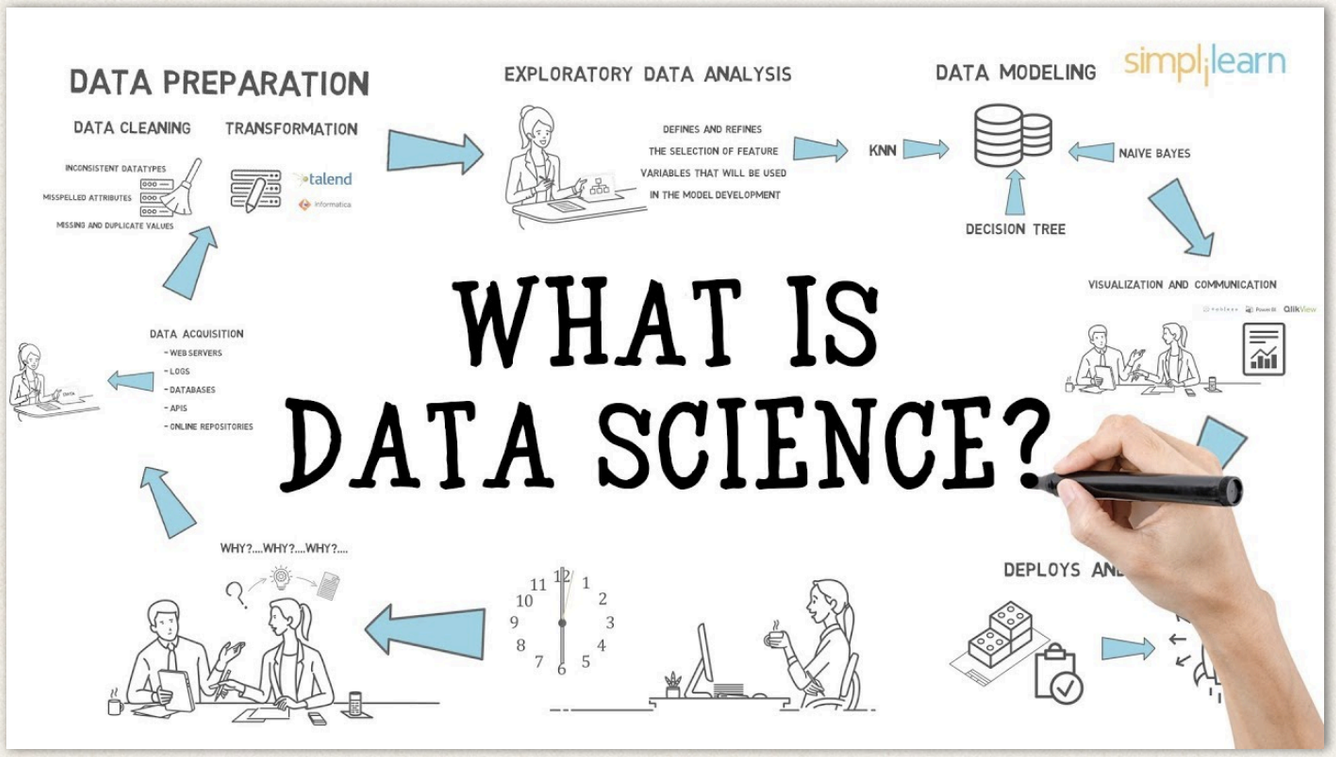
	Discriminative model	Generative model
Goal	Directly estimate $P(y\|x)$	Estimate $P(x\|y)$ to then deduce $P(y\|x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes





Science about Data Science

- Data scientists are doing science about data science when they identify commonly occurring analysis or processing workflows



Does increase in AUC bring any value?

Simple vs complex model

Operational cost

Efficiency

F ( )

Common Task Framework (CTF)

- How to choose best model/approach
  - Provide publicly available datasets
  - A set of enrolled competitors whose common task is to infer a class prediction rule from the training data
    - A scoring referee, to which competitors can submit their training rule. The referee runs the prediction rule against a test dataset not available to competitors, and objectively and automatically reports the score achieved by the submitted rule.
- kaggle.com is an example of CTF
- DIY solutions:
  - split data into training, validation and test set
  - use different models
  - perform cross-validation, ensembles, etc.

Scientific analyses

- A broad collection of technical activities is not a science
  - to do science we must have continually evolving, evidence based approach
- Meta-data analysis: study data analyses on a given topic, does published analyses succeeded and can be improved (reproduced)
- Cross-study analysis: use common datasets to validate studies of different groups by fitting proposed models on different set of datasets and validating them on another set and measure concordance of predictive results
  - study individual models across datasets
  - study individual datasets across models
- Cross-workflow analysis: study effect of different analysis workflows on prediction outcome

Data, Algorithms, Techniques

Engineering Effort for Effective ML

- From "Hidden Technical Debt in Machine Learning Systems", D. Sculley at al. (Google), paper at NIPS 2015

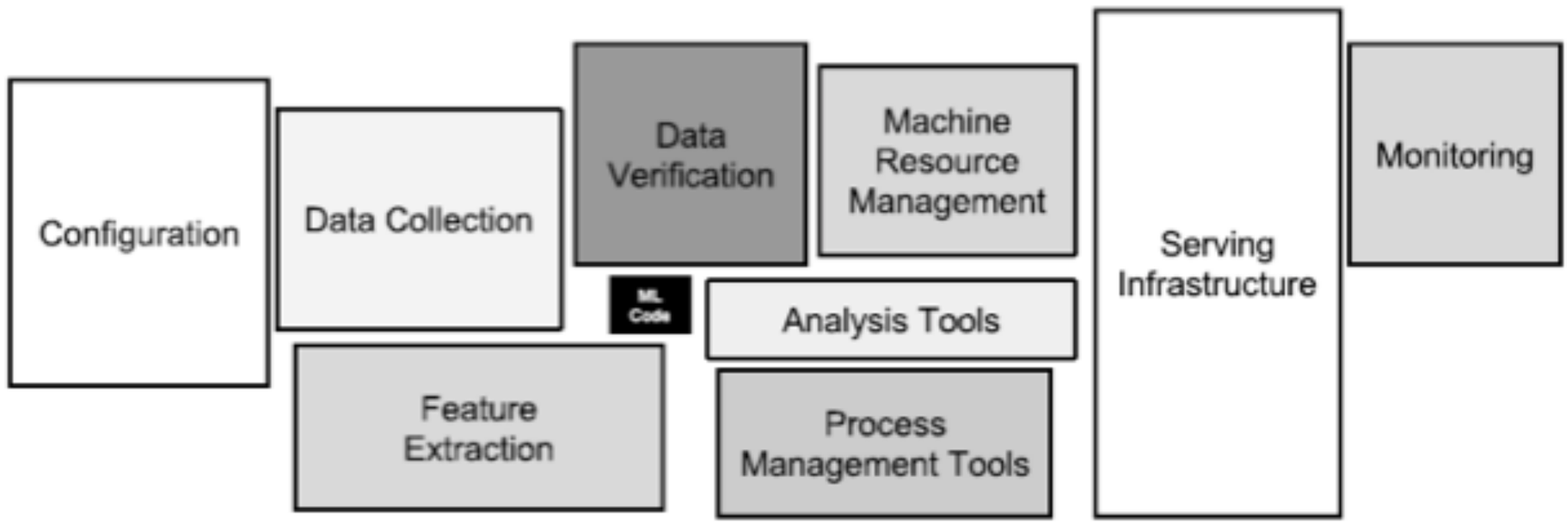
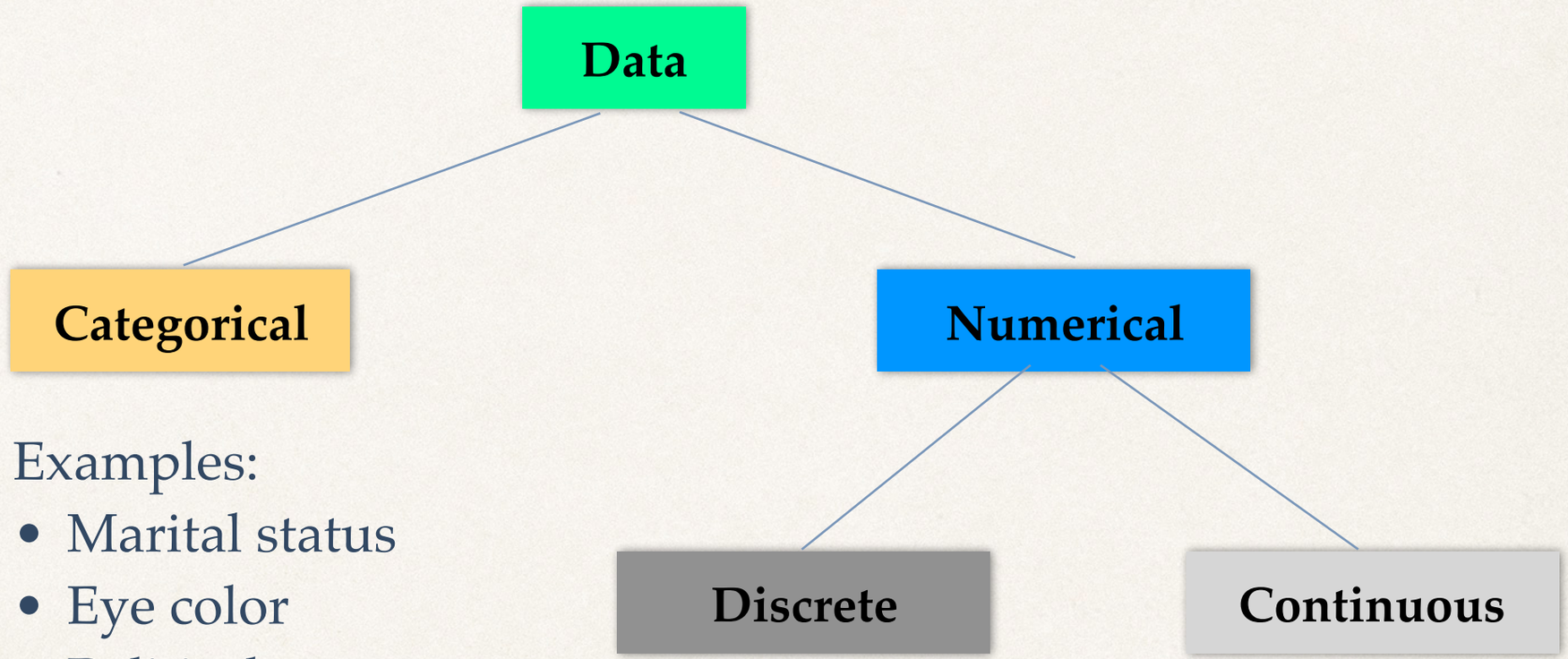


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Data pre-processing

- Most of the time will be spend in this step
- Data clean-up, data transformation, feature engineering
  - data transformation
    - scaling and normalization
    - encoding, aggregation features, log-transformation (to remove outliers)
  - data visualization, exploration
  - data augmentation, imputing, bucketing, binning, feature interactions
  - dimensionality reduction
- **Your programming skills will be required here: R, Python, Databases, etc.**

Types of data



Examples:

- Marital status
- Eye color
- Political party

Examples:

- # children
- month of the year

Examples:

- weight
- house prices

How to better represent dates: categorical or numerical?

Data transformation

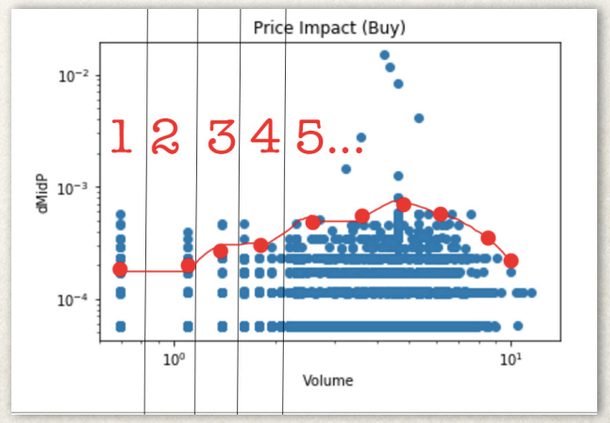
- Data transformation and aggregation: log, sum of values
- Scaling: a technique to scale data to a given range [0,1] or any other range

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Normalization/Standardization: a technique to scale data to mean with zero and and unit-variance

$$x' = \frac{x - \bar{x}}{\sigma}$$

- Augmentation: a technique to create additional data based on input sample which slightly differ from it, e.g. image rotation, flip, scale, crop, etc.
- Bucketing/Binning: a technique to place similar values into buckets/bins



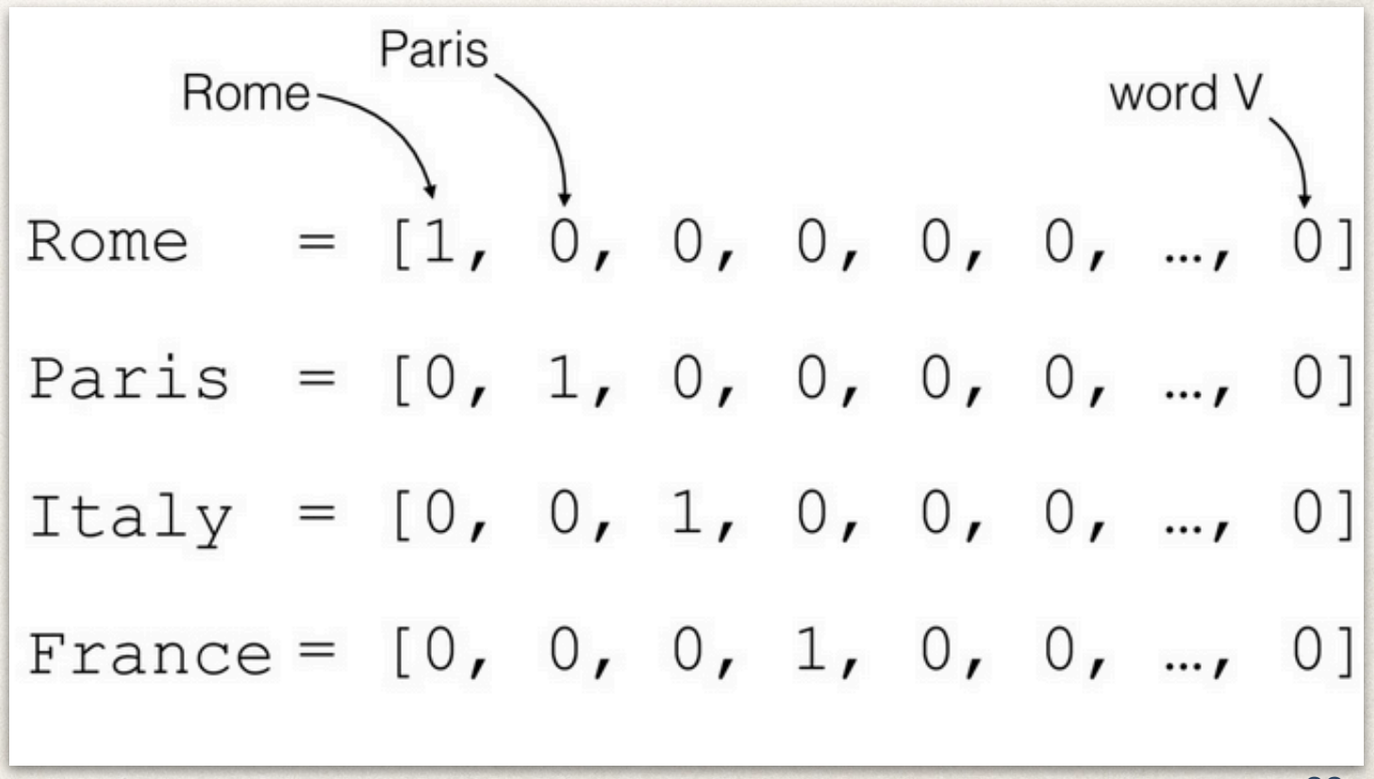
One-hot-encoding

Ref

- It is a technique to handle "categorical" data
- It represents categorical column as vector of words

"One-Hot" refers to a state in electrical engineering where all of the bits in a circuit are 0, except a single bit with a value of 1 (it is said to be "hot").

- You need to define word vector for full set of data (train + test datasets)
- Issues with NULL or missing data
  - delete rows with missing data
  - impute data for missing values



Leave-one-out encoding

[Ref]

- Use mean of all values within the same category except given row
- Add random noise
- Replace categorical value with leave-one-out times noise
- The test categorical values always represented as mean and no noise
- This technique may complement one-hot encoding

Split	UserID	Y	mean_y	random	new_Y
Train	A1	0			
Train	A1	1			
Train	A1	1			
Train	A1	0			
Test	A1	-			
Test	A1	-			
Train	A2	0			

We'll show how to add new categorial encoding features:

mean_y: average value for given user

random: random factor for given user

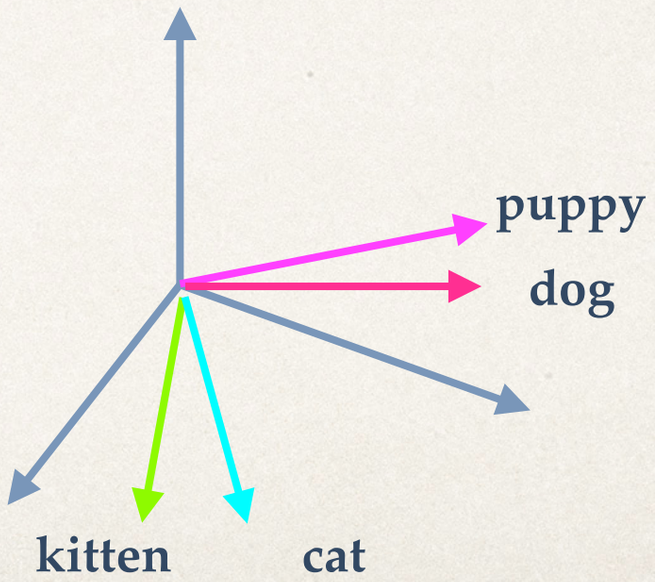
new_Y: new encoding feature for given user

Word embedding

Ref

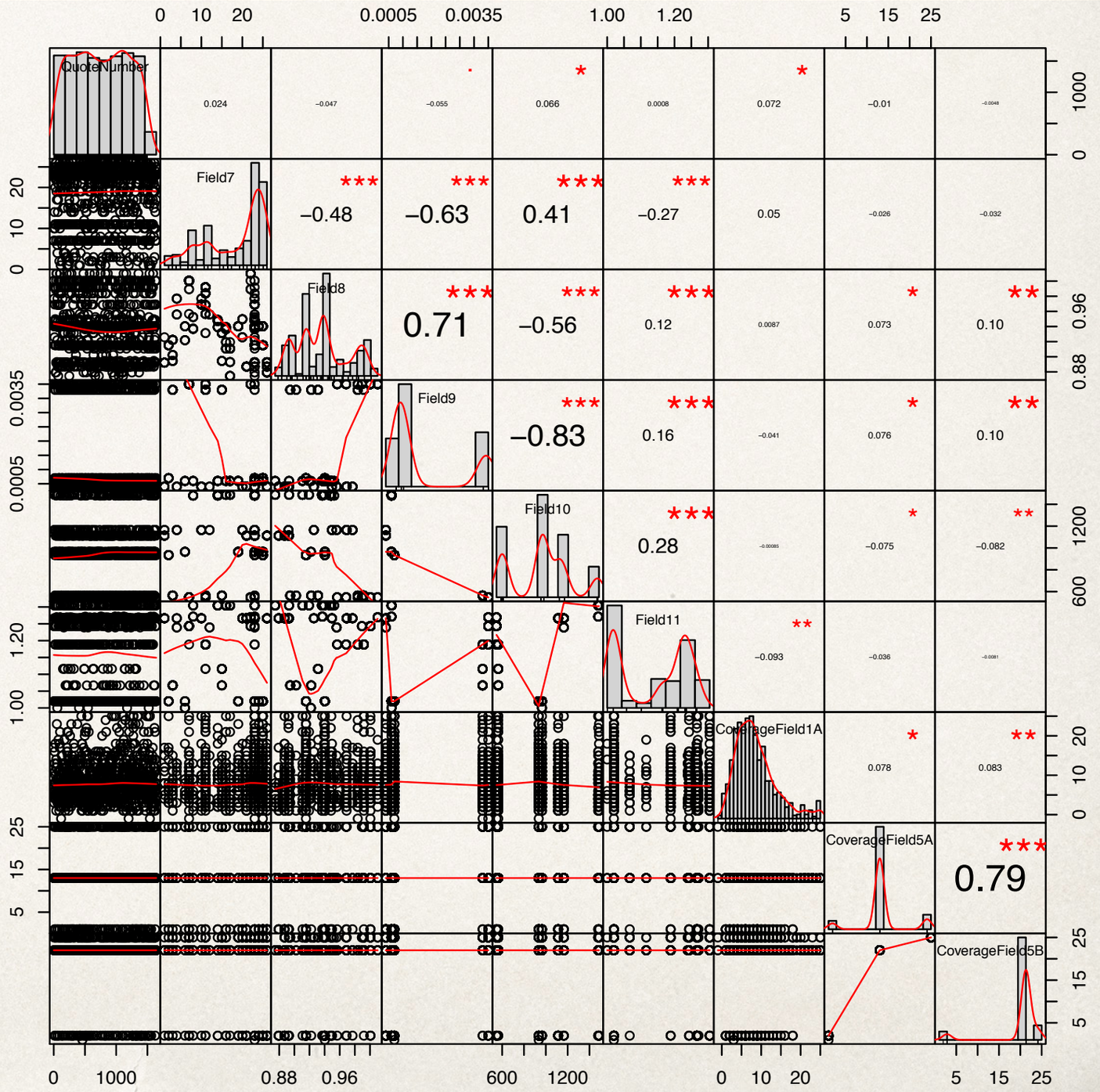
- A way to capture multi-dimensional relationships between categories
  - e.g. Sun and Sat may have similar effect while other days may be treated independently
  - you define a dimension of word vector up-front
  - it projects categorical variables into another phase space, e.g. days may be sunny or rainy, season or off season; all of these features are hidden from original data representation
- Use NN or other ML algorithms to train the model to find best representation of embedded variables

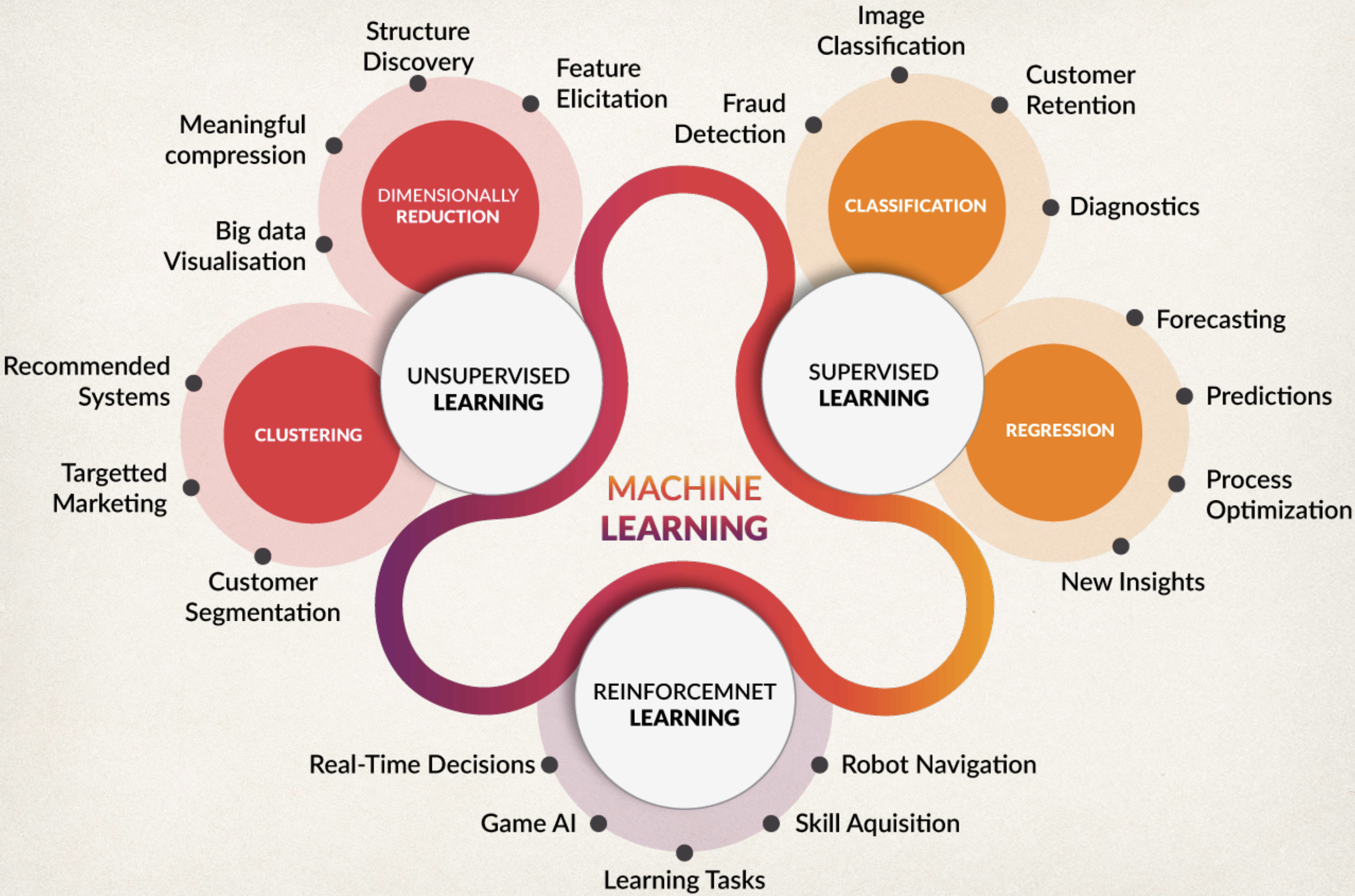
real\hidden	Dog Age Cat
puppy	[0.9, 1.0, 0.0]
dog	[1.0, 0.2, 0.0]
kitten	[0.0, 0.1, 0.9]
cat	[0.0, 1.0, 1.0]

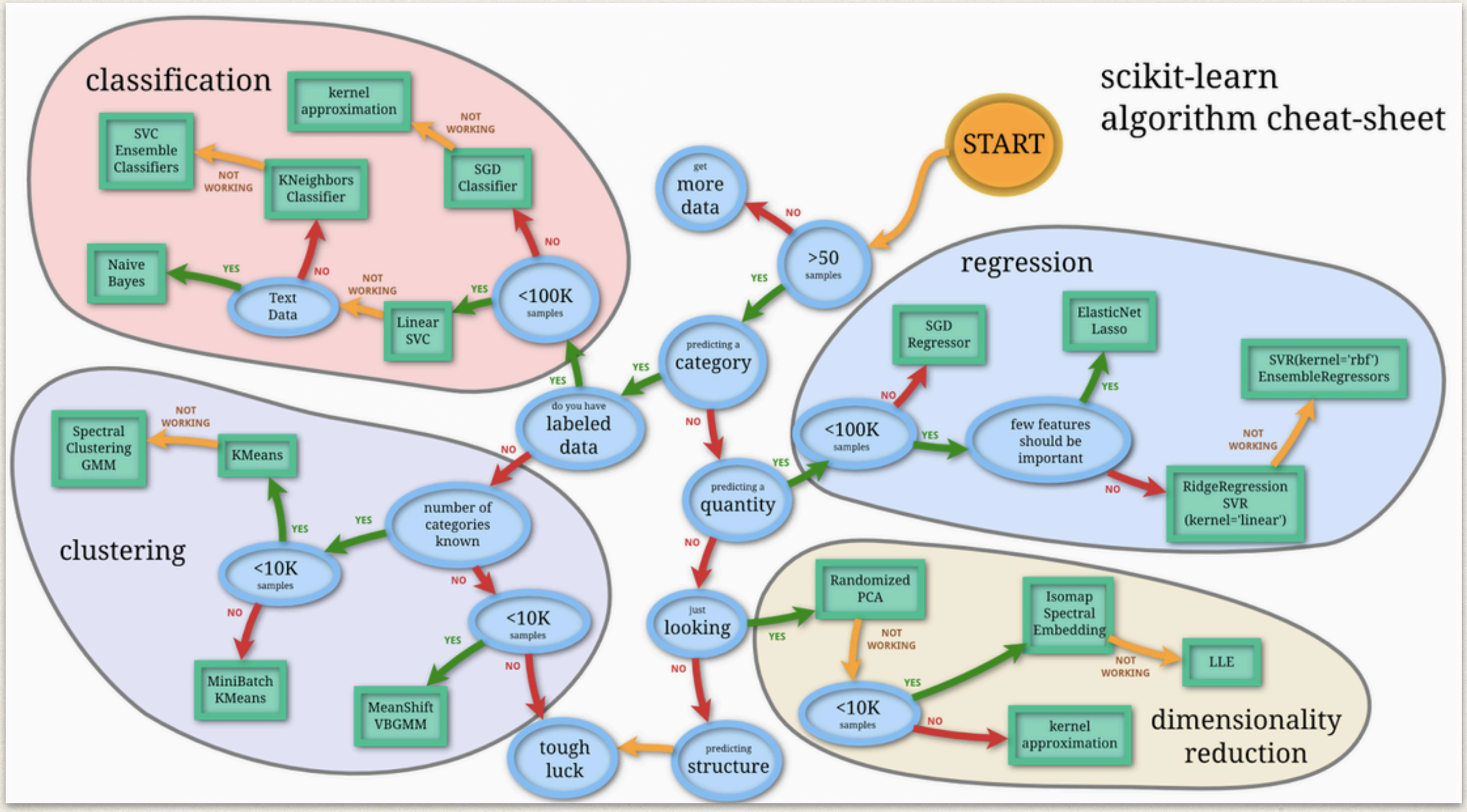


Data visualization

- Graphical representation may reveal important features of the data
  - find correlations, identify range, etc.
- Identify features which may require transformations, e.g. see outliers or skewness in data
- It helps to identify a strategy how to deal with different features

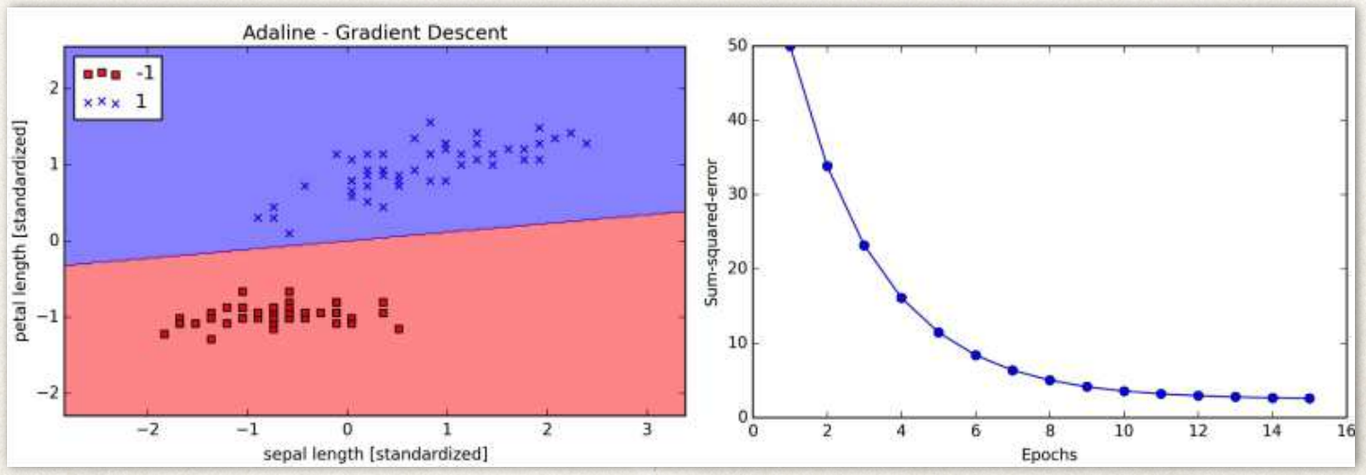



MACHINE LEARNING
UNSUPERVISED LEARNING
DIMENSIONALLY REDUCTION
Structure Discovery
Feature Elicitation
Meaningful compression
Big data Visualisation
CLUSTERING
Recommended Systems
Targetted Marketing
Customer Segmentation
SUPERVISED LEARNING
CLASSIFICATION
Image Classification
Customer Retention
Fraud Detection
Diagnostics
REGRESSION
Forecasting
Predictions
Process Optimization
New Insights
REINFORCEMNET LEARNING
Real-Time Decisions
Robot Navigation
Game AI
Skill Aquisition
Learning Tasks


scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
NOT WORKING
SVC Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
NO
Text Data
NOT WORKING
Linear SVC
YES
<100K samples
NO
get more data
NO
>50 samples
YES
regression
SGD Regressor
ElasticNet Lasso
SVR(kernel='rbf') EnsembleRegressors
NO
<100K samples
YES
few features should be important
YES
NOT WORKING
NO
RidgeRegression SVR (kernel='linear')
predicting a category
YES
YES
do you have labeled data
NO
NO
predicting a quantity
YES
NO
clustering
Spectral Clustering GMM
NOT WORKING
KMeans
YES
number of categories known
YES
<10K samples
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift VBGMM
NO
tough luck
just looking
NO
predicting structure
YES
Randomized PCA
NOT WORKING
YES
<10K samples
NO
Isomap Spectral Embedding
NOT WORKING
LLE
kernel approximation
dimensionality reduction

ML algorithm

- Inputs: X, e.g. timestamp, price, color, size, etc.
- Features: $\mathbb{X}$, transformed inputs
- Labels: y (stay vs leave)
- Weights: $\mathbf{W}$ (matrix)
- Activation function: ϕ (step function, e.g. sigmoid)
- Predictions: $z = \phi(\mathbf{W}^T \mathbb{X})$ yields (-1,1)
- Cost function: $J(\mathbf{W})$, e.g. $\sum(y_i - z_i)^2/2$
- Algorithm: minimizes cost function & find best separation



TYPE	NAME	DESCRIPTION	ADVANTAGES	DISADVANTAGES
Linear	 Linear regression	**The "best fit" line** through all data points. Predictions are numerical.	Easy to understand -- you clearly see what the biggest drivers of the model are.	**X** Sometimes too simple to capture complex relationships between variables. **X** Tendency for the model to "overfit".
Linear	 Logistic regression	The adaptation of **linear regression** to problems of classification (e.g., yes/no questions, groups, etc.)	Also easy to understand.	**X** Sometimes too simple to capture complex relationships between variables. **X** Tendency for the model to "overfit".
Tree-based	 Decision tree	A graph that uses a **branching method** to match all possible outcomes of a decision.	Easy to understand and implement.	**X** Not often used on its own for prediction because it's also often too simple and not powerful enough for complex data.
Tree-based	 Random Forest	Takes the average of many decision trees, each of which is made with a sample of the data. Each tree is weaker than a full decision tree, but **by combining them we get better overall performance.**	A sort of "wisdom of the crowd". Tends to result in very high quality models. Fast to train.	**X** Can be slow to output predictions relative to other algorithms. **X** Not easy to understand predictions.
Tree-based	 Gradient Boosting	Uses even weaker decision trees, that are increasingly **focused on "hard" examples.**	High-performing.	**X** A small change in the feature set or training set can create radical changes in the model. **X** Not easy to understand predictions.
Neural networks	 Neural networks	Mimics the behavior of the brain. Neural networks are interconnected neurons that pass messages to each other. Deep learning uses several **layers of neural networks put one after the other.**	Can handle extremely complex tasks - no other algorithm comes close in image recognition.	**X** Very, very slow to train, because they have so many layers. Require a lot of power. **X** Almost impossible to understand predictions.

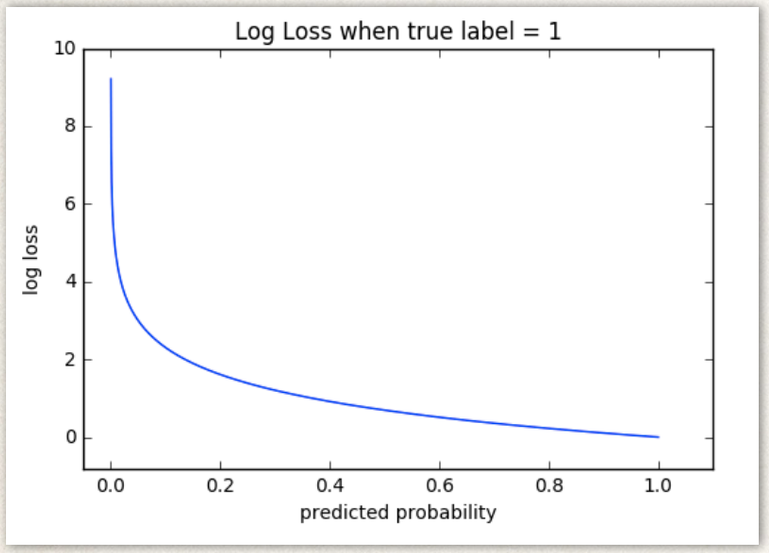
Loss functions

[Ref 1]

[Ref 2]

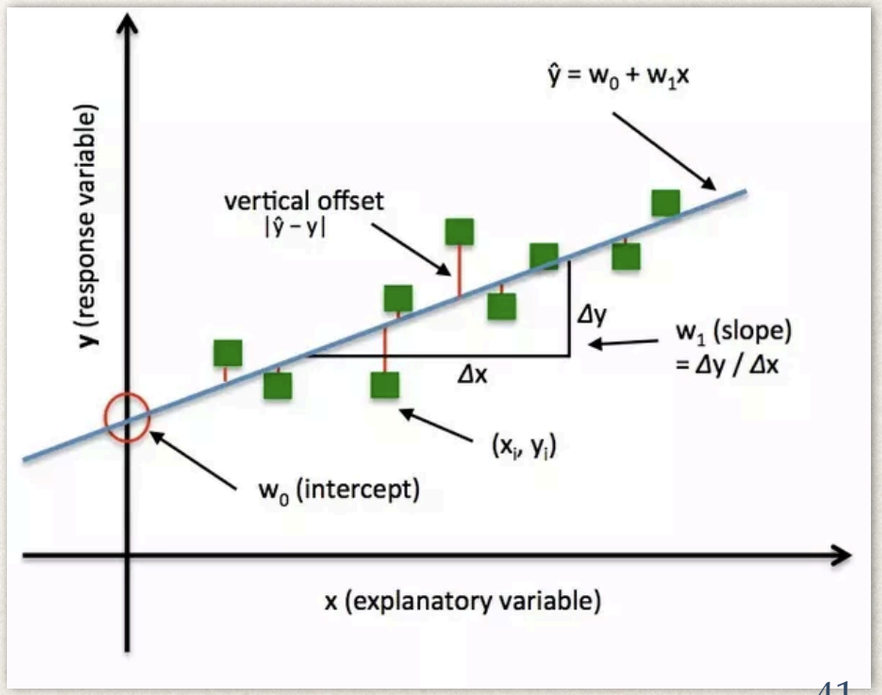
Cross-Entropy Loss

$$-(y\log(p)+(1-y)\log(1-p))$$



- Classification
  - Log Loss
  - Focal Loss
  - Relative Entropy
  - Exponential Loss
  - Hinge Loss
- Regression
  - MSE
  - MAE
  - Huber Loss
  - Log cosh Loss
  - Quantile Loss

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}.$$



Regularization

[Ref 1]

[Ref 2]

- One of the major aspects of training the model is overfitting, when ML model tries too hard to capture the noise in your training dataset
- **Regularization** term is an addition to loss function which helps generalize the model. It helps to learn simpler model, induce models to be sparse, introduce group structure into learning problem $\min_f \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda R(f)$
  - **L1** or **Lasso regularization** adds penalty which is a sums of the absolute values of weights $Min(\sum_{i=1}^{n}(y_i - w_i x_i)^2 + p\sum_{i=1}^{n}|w_i|)$ **MSE+L1**
  - **L2** or **Ridge regularization** adds penalty which is a sums of the squared values of weights $Min(\sum_{i=1}^{n}(y_i - w_i x_i)^2 + p\sum_{i=1}^{n}(w_i)^2)$ **MSE+L2**
- **Dropout** is a term introduced in NN context where hidden nodes are dropped randomly and allow model to generalize better
- **Early Stopping** is time regularization technique which stop training based on given criteria

Data Science recipe

- Understand your data: preprocessing, cleaning, augmentation, one-hot-encoding
- Categorize the problem: classification, regression, clustering, dimensionality reduction
- Choose the language and toolkit: R, Python, Hadoop+Spark, ML providers
- Choose the right technique: trees, bagging, stacking, boosting, (rank | weight) averaging, NNets
- Start coding using your favorite ML framework and visualization tools

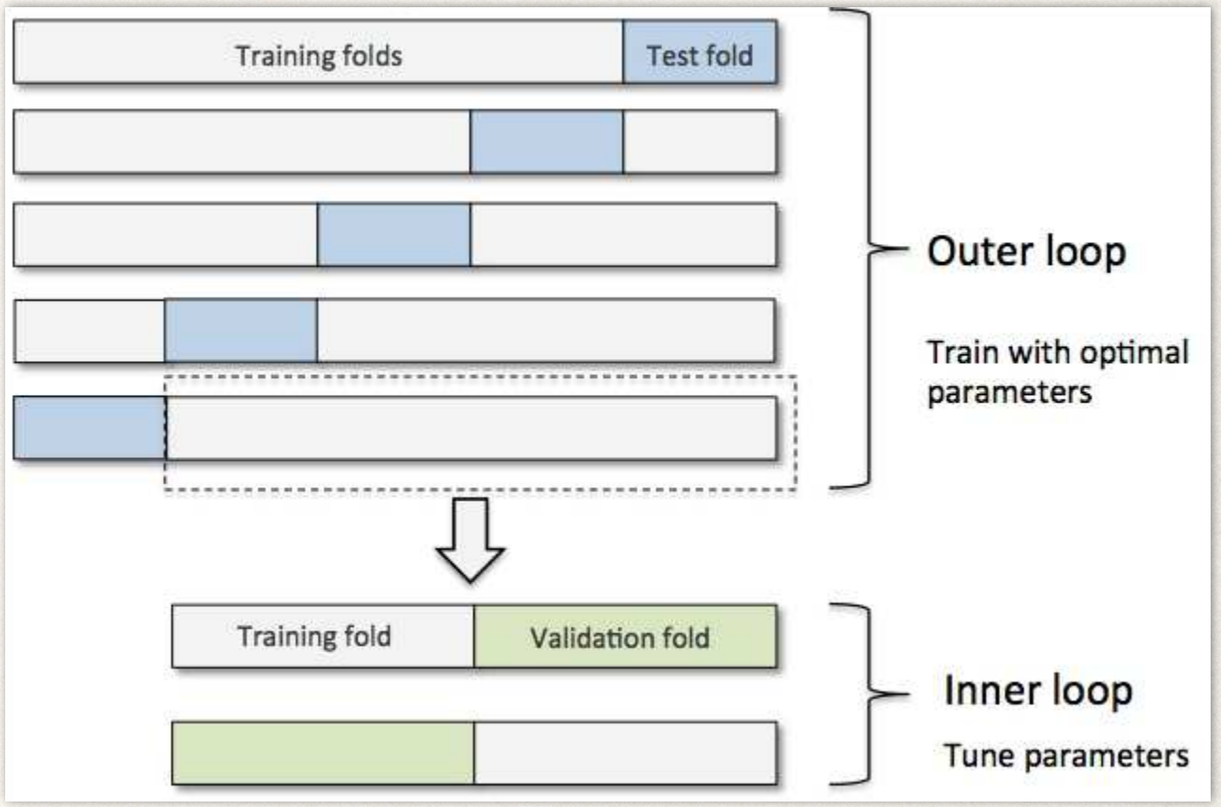
Techniques

[Ref1]

Cross-validation

[Ref2]

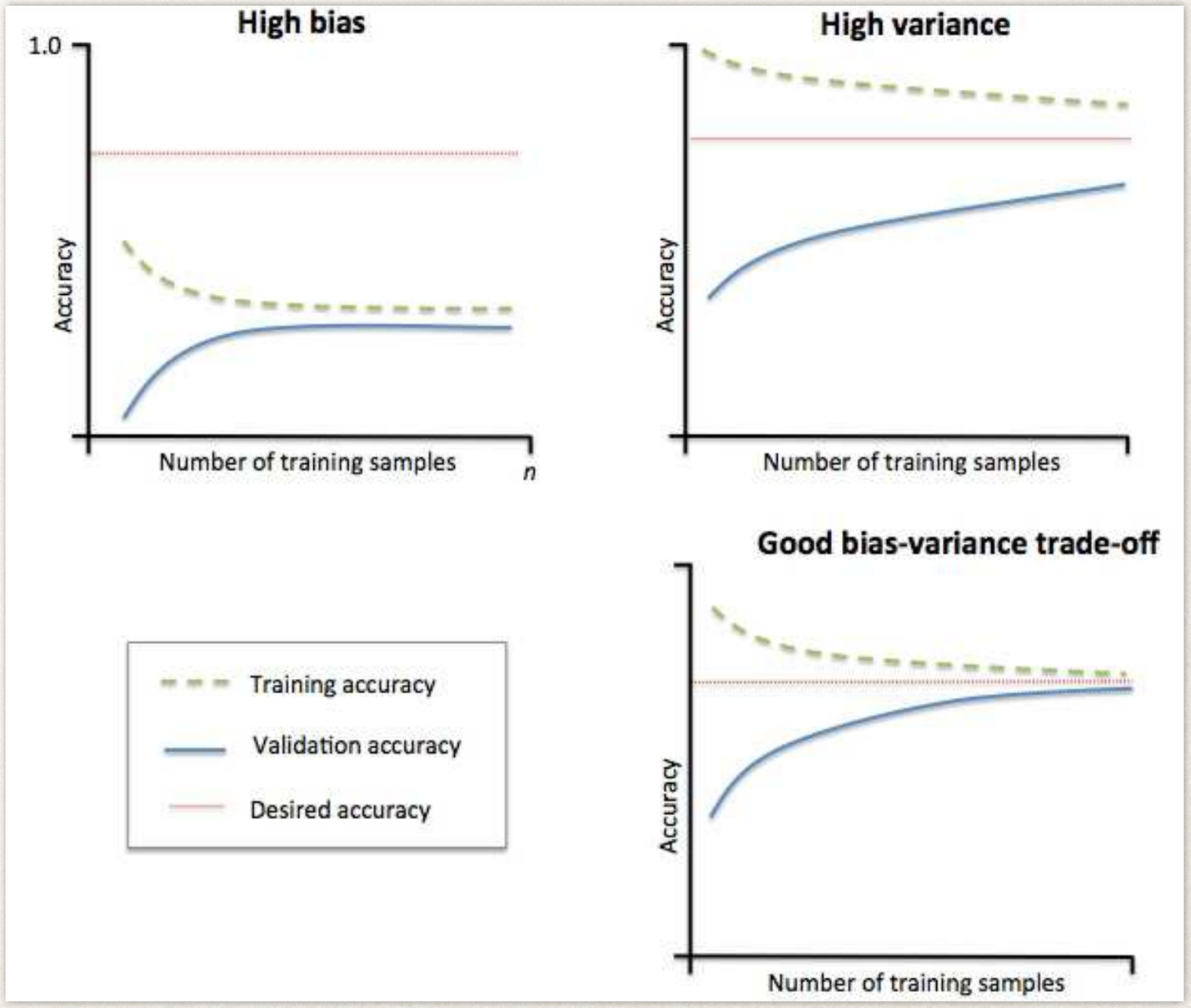
- In **K Fold cross validation**, the data is divided into k subsets
- Hold-out method: each time, one of the k subsets is used as the test set/ validation set and the other k-1 subsets are put together to form a training set
- The error estimation is averaged over all k trials to get total effectiveness of our model
- Every data point gets to be in a validation set exactly once, and gets to be in a training set *k-1* times
- This significantly reduces bias as we are using most of the data for fitting, and also significantly reduces variance as most of the data is also being used in validation set



- Stratified K-Fold Cross Validation deals with imbalanced data, each fold contains the same percentage of samples of each class
- Leave-P-Out Cross Validation: leaves p data points out of train set, i.e. if we have n data points, then n-p points used for training and p points for validation

Bias and Variance problems

- Plotting model training and validation accuracies detects high variance and high bias
- High bias shows that model under fit the training data
  - increase the number of parameters in a model via additional feature engineering
- High-variance indicates of model overfitting problems
  - to combat collect more data and reduce complexity of the model



Ensembles

[Ref 1]

[Ref 2]

[Ref 3]

All models are wrong, but some are useful (George Box)

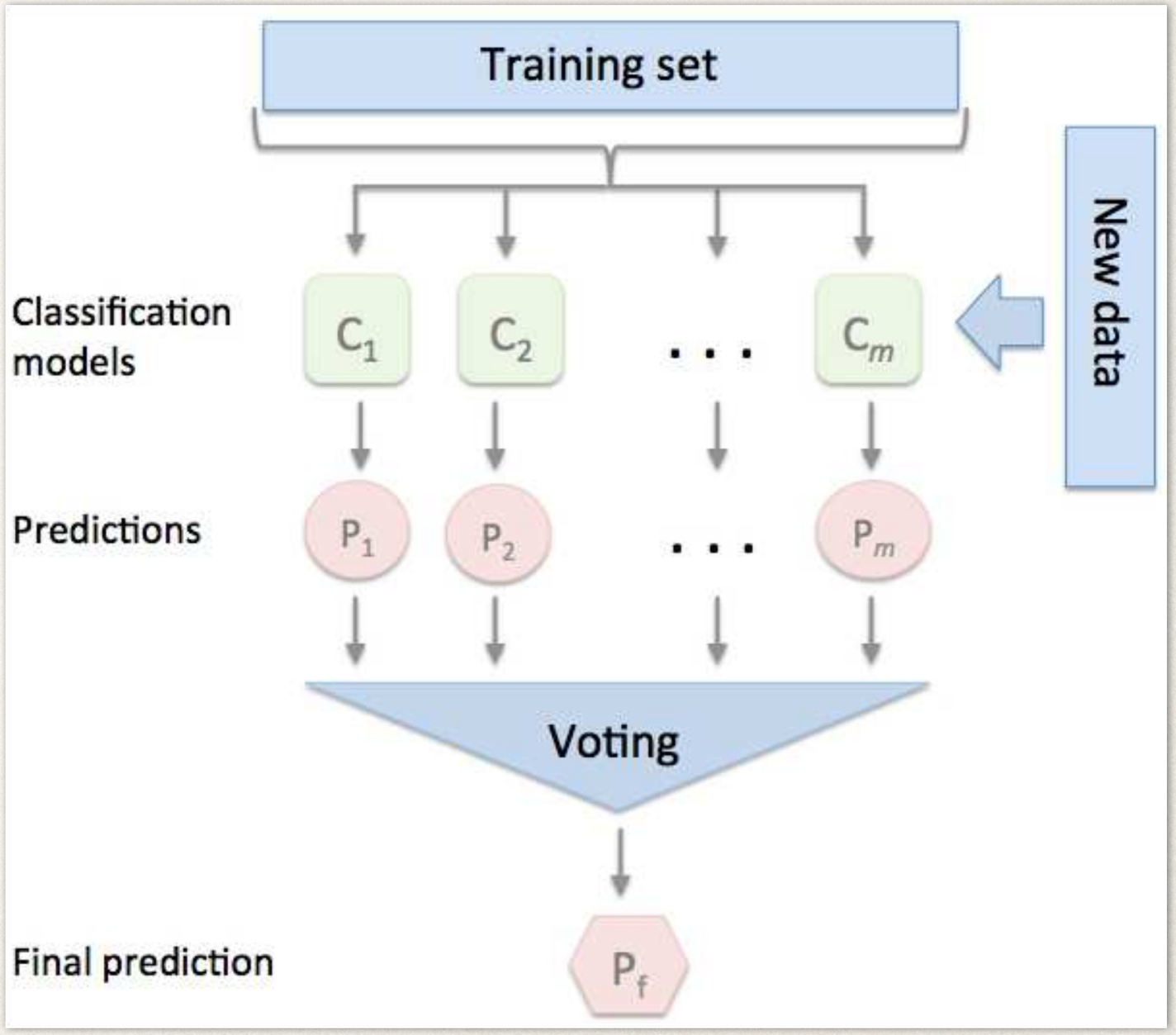
Sometimes intentionally built weak models are good blending candidates

- Bagging
  - building multiple models (typically of the same type) from different subsamples of the training dataset
- Boosting
  - building multiple models (typically of the same type) each of which learns to fix the predictions errors of a prior model in the chain
- Stacking
  - building multiple models (typically of the different types) and supervisor model that learns how to best combine the predictions of the primary model
- Weighting | Blending
  - combine multiple models into single prediction using different weight functions

Diversity is a key: use different un-correlated models, e.g. GBM, RF, SVM, NN

Ensembles: voting

- Unaminity
- Majority
- Plurality



better to fight over-fitting

better to get lower errors

Bagging vs Boosting

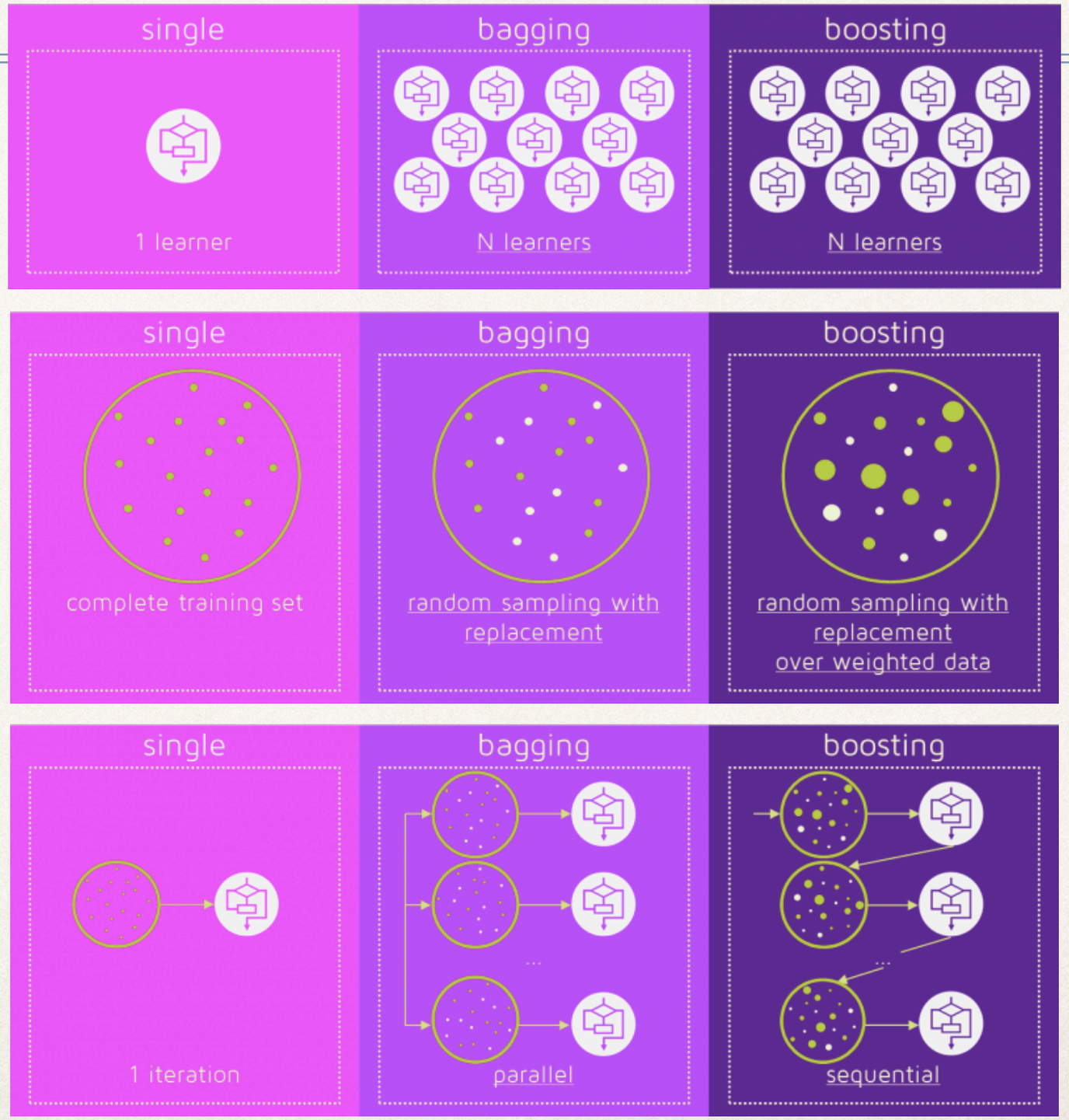
Ref

Similarities

Both are ensemble methods to get N learners from one

Generate several training sets by random sampling

Make final decision by averaging N learners or taking majority of them



Differences

build independently for Bagging, and Boosting tries to add new models that do well where previous models fail

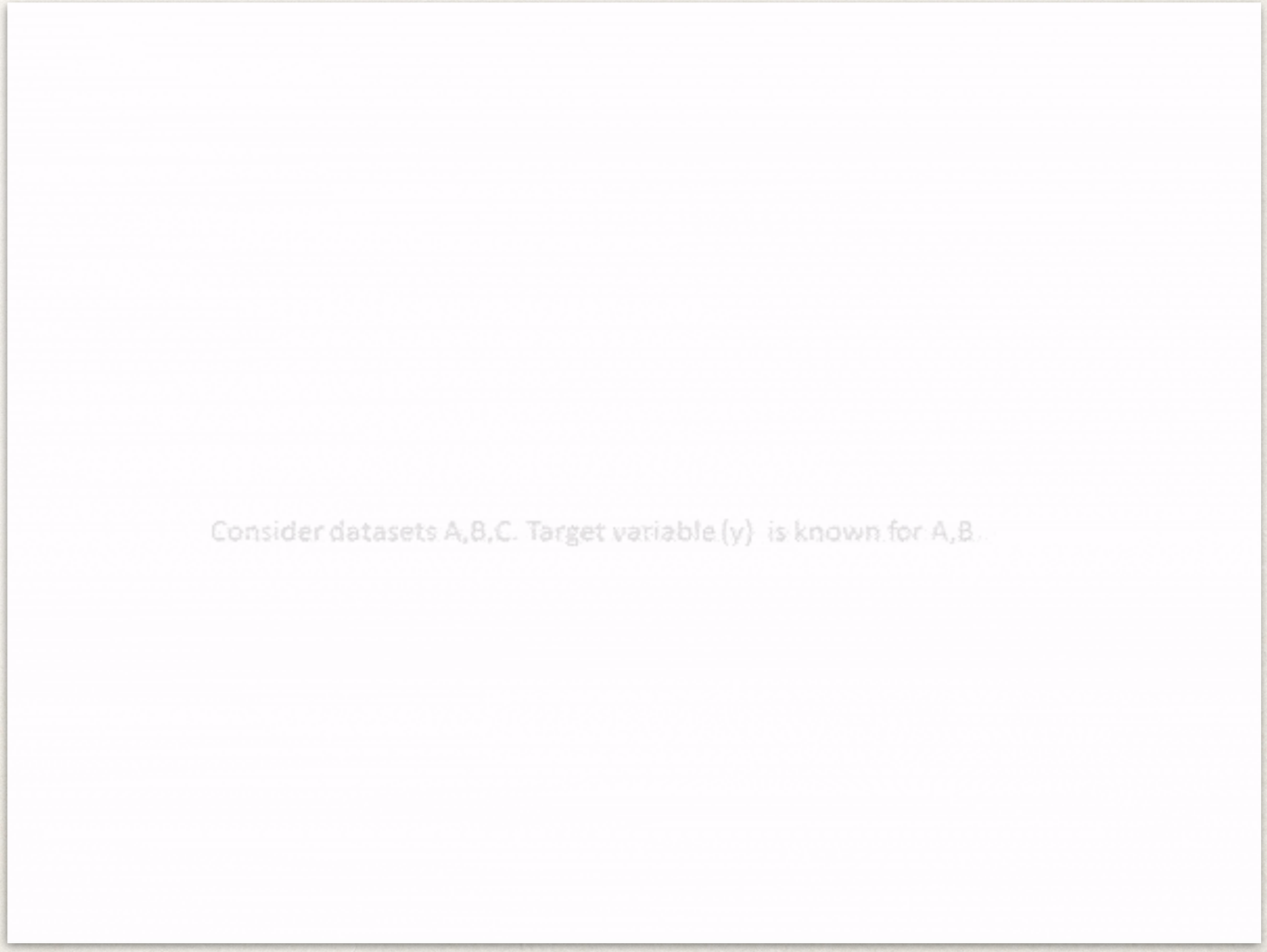
Boosting weights the data to scale in favor of most difficult cases

Bagging: equally weighted average
Boosting: weighted average, more weight to those who perform better on training set

Stacking

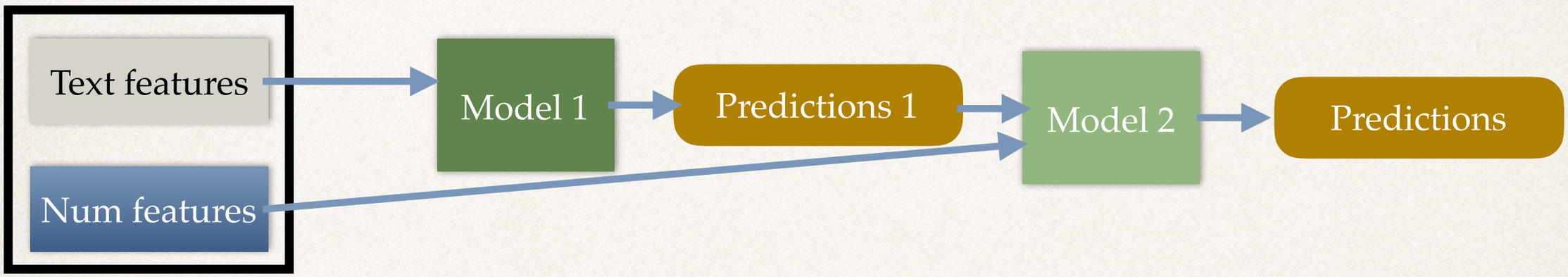
Ref

- Stacking (also called meta-ensembling) is a model ensembling technique used to combine information from multiple predictive models to generate a new model
- Usually outperform individual models used in ensemble, e.g. GBM+RF+NN
- Most effective when base models are independent
- May be applied at multiple level, e.g. stacking first set, then second set, etc.



Technical tricks

- Use one set of features (text) for simple model 1, and use numerical features and model1 prediction for model 2, etc.



- Use chained models: build stand-alone model for G, then used in next model, e.g. F=>G=>B=>A
- Feature engineering:
  - one-hot-encoding, leave-one-out, word embedding and add them to original data set
  - split days into years, months, dates and threat them as categorical variables
  - aggregate values, e.g. sum all numerical values in a row and/or use its mean/median
  - handle missing values, e.g. apply mean across column or even apply additional training to find their values

Tools and frameworks

python

jupyter

pandas

$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$

matplotlib

NumPy

ANACONDA

DASK

Keras

TensorFlow

PYTORCH

theano

Caffe

fast.ai

scikit learn

Classification
Regression
Clustering
Dimensionality reduction
Model selection
Preprocessing

R

DataFrame
data.table
ggplot
xgboost
NeuralNetwork
Trees, Bagging

H2O.ai

torch

WEKA The University of Waikato

STACK NET

VOWPAL WABBIT

dmlc mxnet

dmlc XGBoost

ML for “standard” use-cases

- In most cases you may rely on R or Python eco-system. In Python scikit-learn is de-facto standard, in R all ML tools are available through 3rd party packages via install.packages(<pkg>)
- Majority of DataScientists in kaggle competition use xgboost, the distributed gradient boosting library (both R and Python APIs are available) based on parallel tree boosting algorithm (aka GBDR, GBM)
- Less known libraries are:
  - Weka is Waikato Environment for Knowledge Analysis is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand (GUI environment)
  - StackNet is a computational, scalable and analytical Meta modeling framework (developed by top-level kaggle competitor Kaza-Nova and used in many competition to won first places). Written in Java and uses uses Wolpert's stacked generalization to improve accuracy of ML models. The network is built iteratively one layer at a time (using stacked generalization), each of which uses the final target as its target.
  - h2o Open Source Fast Scalable Machine Learning Platform For Smarter Applications (Deep Learning, Gradient Boosting, Random Forest, Generalized Linear Modeling (Logistic Regression, Elastic Net), K-Means, PCA, Stacked Ensembles, Automatic Machine Learning (AutoML)

Neural network frameworks

- Torch is an open source machine learning library, a scientific computing framework, and a script language based on the Lua programming language.
- Theano is a numerical computation library for Python that allows you to define, optimize, and evaluate mathematical expressions involving multi-dimensional arrays efficiently. In Theano, computations are expressed using a NumPy-esque syntax and compiled to run efficiently on either CPU or GPU architectures.
- Caffe is a deep learning framework (C++ and Python) made with expression, speed, and modularity in mind.
- TensorFlow is an open-source software library (C++, Python, Go) for data-flow programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks.
- PyTorch is a deep learning framework for fast, flexible experimentation. It is Tensors and Dynamic neural networks in Python with strong GPU acceleration.
- Keras is a high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK, or Theano
- fast.ai library simplify training ML/DL learning using best DataScientists and ML practitioner practices.
- Apache MXNet framework (Python and R) is a modern deep learning framework
- onnx.ai is an Open Neural Network exchange format which allows to import and export Neural Network models from/to different frameworks

Visualization of Neural Networks

- TensorFlow playground: provides an intuitive web based interface to train Neural Networks for a given dataset
- ConvNetJS is a Javascript library for training Deep Learning models (Neural Networks) entirely in your browser
- LSTMVis - visual analysis for Recurrent Neural Networks
- Netron is a visualizer for Deep Learning and machine learning models
- Ann-visualizer, is a python library for visualizing Artificial Neural Networks
- Keras-vis is a high-level toolkit for visualizing and debugging your trained keras neural net models
- VisualDL is an open-source cross-framework web dashboard that richly visualizes the performance and data flowing through your neural network training

ML for Big Data

- Some datasets can't be trained with standard ML tools since they are too big to fit into memory, therefore you can't use "standard" tools like scikit-learn or R
- Gradient Boosting Algorithm (GBM) is a ML technique which produces a prediction model in a form of ensemble of weak prediction models, typically decision trees
  - Boosting is an ensemble technique in which the predictors are not made independently, but sequentially. Therefore a large dataset can be learned in "chunks" with GBM
- Vowpal Wabbit is online learning algorithm designed to deal with tera-features datasets
- Spark ML Big Data platform (MLlib), Spark is a technique to deal and process large datasets using Hadoop platform which now has a set of ML algorithms available as a part of platform

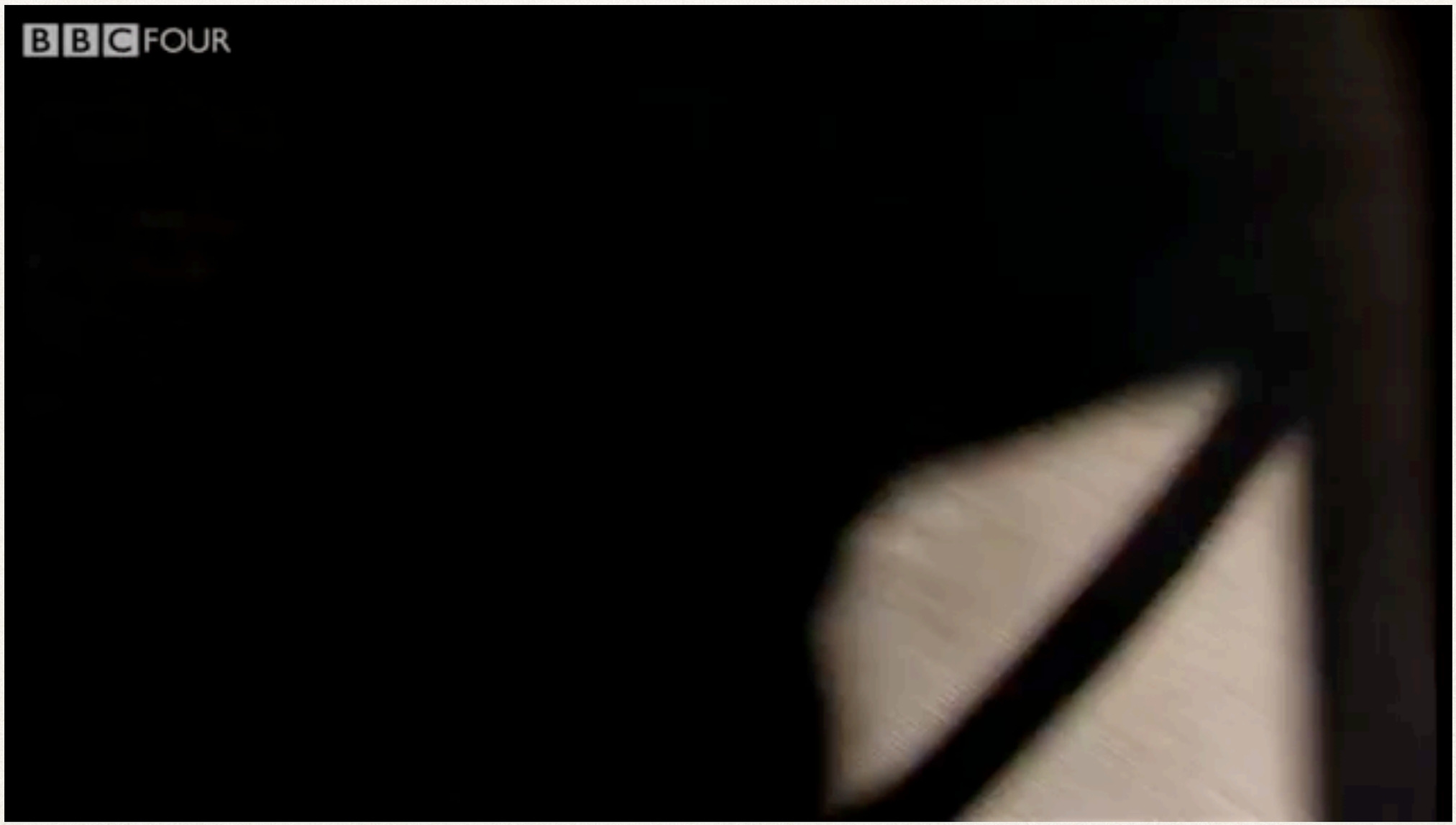
Courses

- kaggle.com is a place to do data science projects, it is your **ULTIMATE** source of knowledge in DataScience, ML, DL and AI
- fast.ai provides cutting edge about deep learning
- Google TensorFlow Development Summit new ideas and practical implication of TF
- Machine Learning A-Z: Hands on Python & R In Data Science covers machine learning workflows
- Scala and Spark for Big Data and Machine Learning covers Big Data technology
- Building Neural Network from scratch: github and blog
- Machine Learning courses ranked by user reviews

Resources

- How to get started with ML
- Choosing the right ML algorithm
- Colah's blog
- Stacking Made Easy
- Gradient Descend Optimization
- ML, Python and Math Cheat Sheets
- Data Science interview questions
- Neural Network zoo
- Large Scale Deep-Learning with TensorFlow
- Learning Machine Learning
- Cheat Sheet for AI, ML, NN, BigData Salary history and career path of a Data Scientist

The Story



https://www.youtube.com/watch?feature=player_embedded&v=jbkSRLYSojo