How to manage data-related projects and not fail (too often)?

Jakub Nowacki

whoami

Machine Learning Engineer
Educator

I can code, I do maths

@jsnowacki
https://www.linkedin.com/in/jakubnowacki/
https://github.com/jsnowacki

Why do 87% of data science projects never make it into production?

https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/

85% of big data projects fail, but your developers can help yours succeed

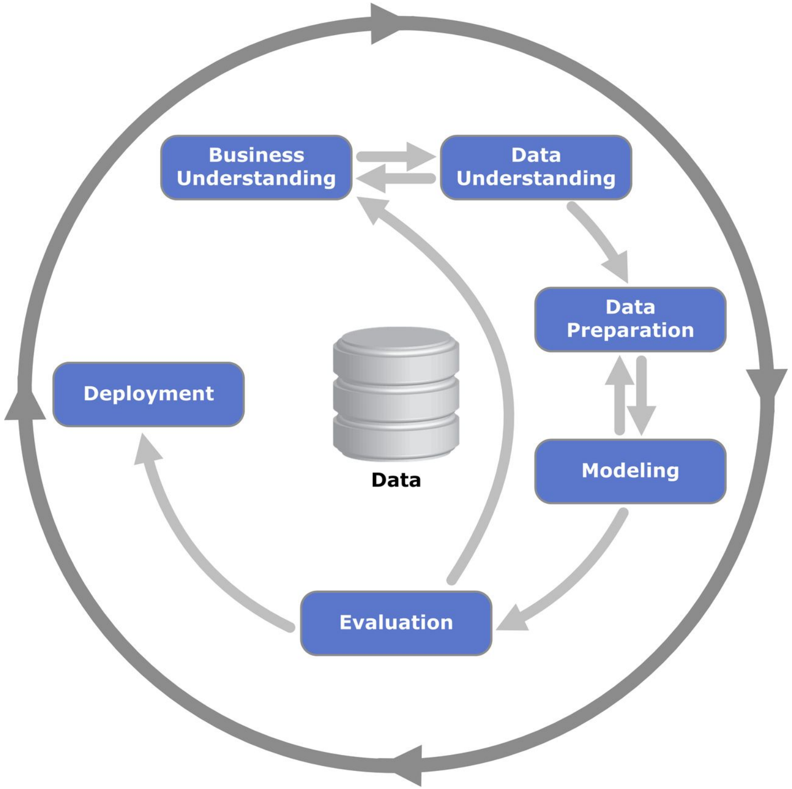
https://www.techrepublic.com/article/85-of-big-data-projects-fail-but-your-developers-can-help-yours-succeed/

Looks familiar...?

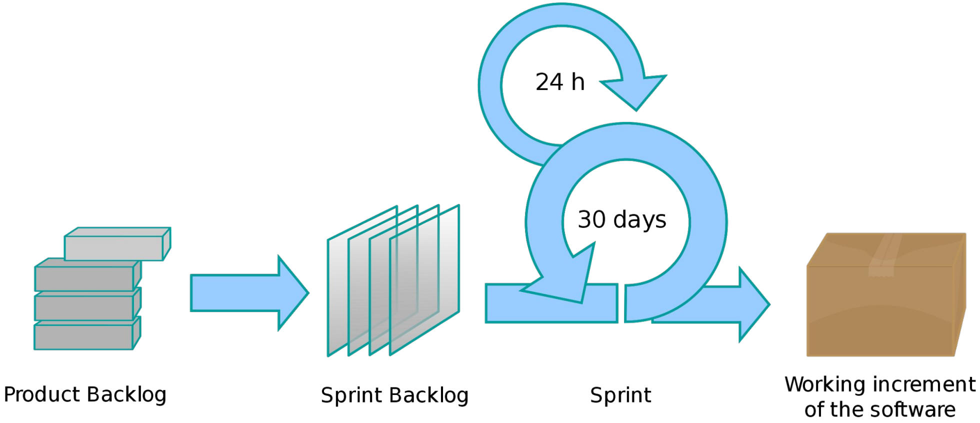
- Why the process differ?
- Who are you anyway?
- What we cannot see?
- What are the costs?
- What to look out for?
- How the team differs?

https://twitter.com/historyinpics/status/404077361484038144

The process: CRISP DM vs SCRUM

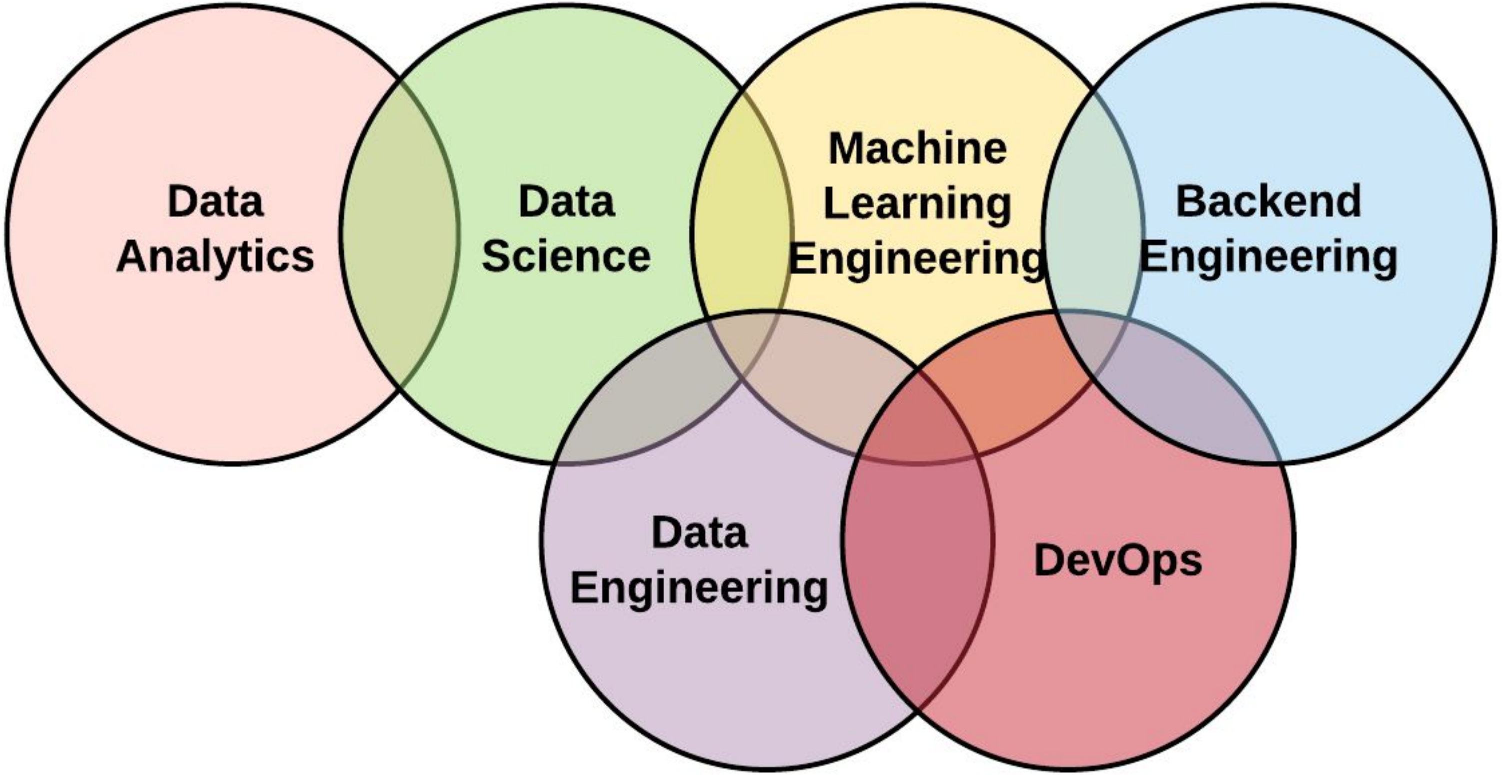


https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining



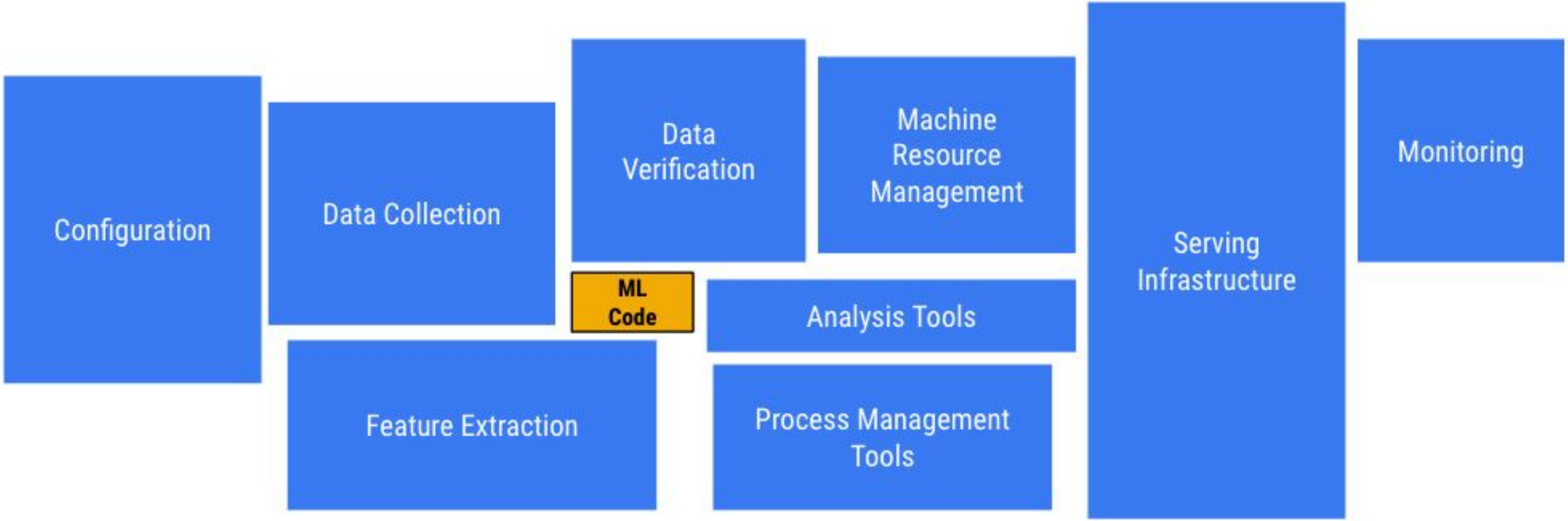
https://en.wikipedia.org/wiki/Scrum_(software_development)

What is actually involved in data project?



https://mindfulmachines.io/blog/2018/3/17/the-flavors-of-data-science-and-data-engineering

Hidden technical debt in ML systems



https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

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, 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, Clustrix, NuoDB, Pivotal, MariaDB, Cloud Spanner, MEMSQL, influxdata, Cockroach Labs, TIMESCALE, VOLTDB, splice, imply, paradigm4, TileDB

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, Segment, ATTUNITY, xplenty, ZALONI, import.io, Infoworks, Fivetran, SNOWPLOW, MATILLION

DATA GOVERNANCE
Informatica, IBM, 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, unravel, Numerify, ScienceLogic, zenoss, OpsRamp, Veeam, MAGNITUDE

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

CLUSTER SVCS
Amazon ECS, IBM, Amazon EKS, IBM Cloud Private, Cluster Server, MESOSPHERE, packet, Pivotal Container Service, Bright Computing, CYCLECLOUD

DATA GENERATION & LABELLING
amazon mechanical turk, upwork, unity, appen, scale, HIVE, Labelbox, Mighty AI, ALEGION, 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, NVIDIA, IBM Power Systems, GRAPHCORE, MYTHIC, Cerebras, Movidius, habana, WAVE COMPUTING, CORNAMI, HAILO, flexlogix, EFINIX

ANALYTICS & MACHINE INTELLIGENCE

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

DATA SCIENCE PLATFORMS
IBM, databricks, dataiku, 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, Power BI, SAP Lumira, Google Cloud, celonis, Periscope Data, zepl, GOODDATA, 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, Google Cloud Vision API, clarifai, 商汤, EVER AI, deepomatic, neurala, twentybn, UBIQUITY6, ANDEE, 依图 YITU, trax, ALARBER, synthesia, DataGrid

HORIZONTAL AI
IBM Watson, Cortana, Face++ 旷视, sentient, Voyager Labs, vicarious, Affectiva, PROPHESEE, CognitiveScale, neurala, twentybn, Numenta, PETUUM, narrative science, CURIOUS AI, OSARO, BLUE VISION, Pefectron, Fortress IQ

SPEECH & NLP
Google Cloud, twilio, amazon alexa, Amazon Translate, narrative science, semantic machines, HEALTHGRADES, Mobvoi, EigenTechnologies, SoundHound Inc., volcera, PRIMER, MindMeld, Mindfold, cogito, snips, NUANCE, 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, traackr, simplereach, bitly, SimilarWeb

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

APPLICATIONS – ENTERPRISE

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

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

MARKETING - B2C
zeta, bloomreach, SendGrid, braze, ACTIONIQ, BLUECORE, CONTENTSQUARE, TEALIUM, mparticle, Amplero, amperity, QUANTIFIND, Simon, Liftics, 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, CallDesk, netomi, TechSee, frame.ai

ENTERPRISE PRODUCTIVITY
slack, x.ai, ORACLE, GURU, lumiata, DIFFBOT, clara, talla, Kasisto

HUMAN CAPITAL
HireVue, pymetrics, hiQ, GIGSTER, Pilot, mya, Allyo, textio, Wade&Wendy, Stella, Cluster, entelo, RESTLESS BANDIT, ideal, 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, botkeeper, pilot

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

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

APPLICATIONS – INDUSTRY

ADVERTISING
AppNexus, MediaMath, criteo, xAd, Integral, ORACLE MOAT, OpenX, datorama, theTradeDesk, Adgorithms, dstillery, LiveIntent, TAPAD, dataxu, gumgum, Appier, tremor video, yieldmo, mobvista

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, STREETLIGHT DATA, OpenDataSoft

INTELLIGENCE
Palantir, Dataminr, Quid, PRIMER, FORGE

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

FINANCE - LENDING
ondeck, affirm, 拍拍贷, JIANPU.AI, Kreditech, AVANT, TALA, ZestFinance, Upstart, aura, CLEARBANC, upgrade, 100Credit, WeLab, Wecash闪银, TrueAccord, MoneyLion, Active.Ai, aire, cignifi

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

HEALTHCARE
flatiron, Clover, ZYMERGEN, HealthTap, METABIOTA, Ginger.io, Glow, babylon, 3DMed, zebra, PathAI, ovia, TEMPUS, patientslikeme, AiCure, insitro, LinkDoc, butterfly, citizen, notable, komodo health, RECURSION, prognos, enlitic, imagia, BlackThorn, BAYLABS, Qventus, ARTERYS, IMAGEN, CLOUDMEDX, PAIGE, DATAVANT, innovaccer, DIAGNOSS PROJECT, LeanTaaS

LIFE SCIENCES
23andMe, color, iCarbonX, BenevolentAI, verily, WuXiNextCODE, Zymergen, Numerate, 3SCAN, 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, Civil Maps, cognata, thinci, INRIX

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

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

INDUSTRIAL
AVEVA, SIEMENS, Schneider Electric, PREDIX, UPTAKE, SIGHT MACHINE, FOGHORN, OSIsoft, SCORTEX, KONUX, 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 HDFS, Spark, Hadoop MapReduce, Flink, YARN, TEZ, MESOS, kubernetes, docker, CDAP, RedHat, HELIX

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

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

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

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

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

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

AI / MACHINE LEARNING / DEEP LEARNING
TensorFlow, Keras, torch, MLlib, Caffe, Microsoft Cognitive Toolkit, OpenAI, DMTK, theano, PaddlePaddle, Apache SINGA, DIMSUM, FeatureFu, mxnet, VELES, Chainer, Michelangelo, ONNX, WEKA, H2O, 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, VECTOR

VISUALIZATION
matplotlib, TensorBoard, seaborn, Bokeh

COLLABORATION
BeakerX, jupyter, Zeppelin, ANACONDA

SECURITY
Apache Ranger, KNOX, Sentry, accumulo

DATA SOURCES & APIs

HEALTH
Apple, 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, INVESTNET YODLEE, THE WORLD BANK, estimize, PREMISE, Quandl, Eagle Alpha, StockTwits, xignite, Thinknum, earnest, predata

AIR / SPACE / SEA
Orbital Insight, planet, SKYCATCH, AIRBOTICS, spire, PRECISIONHAWK, Descartes Labs, kespry, UNDERSTORY, tellus labs, WINDWARD, DroneDeploy, MarineTraffic, LOFT ORBITAL, RS Metrics

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

LOCATION INTELLIGENCE
FOURSQUARE, MapAnything, mapbox, sense360, pitney bowes, HEXAGON, Placed, esri, factual, CARTO, Mapillary, Streetlire, cuebiq, Radar, OpenStreetMap

OTHER
DATA.GOV, IMAGENET, wiki-links, LabelMe, APOLLOSCAPE, CRUX, HIGHCHARTS, grafiti.io

DATA RESOURCES

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

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

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

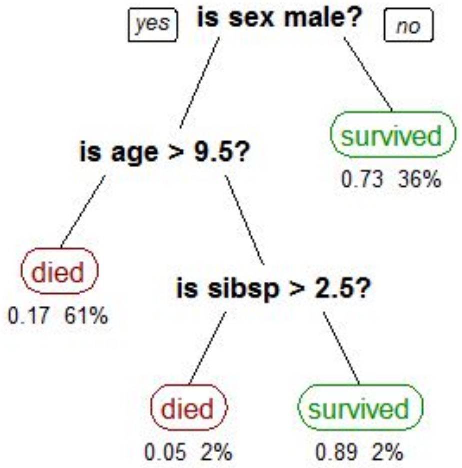
July 16, 2019 - FINAL 2019 VERSION

mattturck.com/data2019

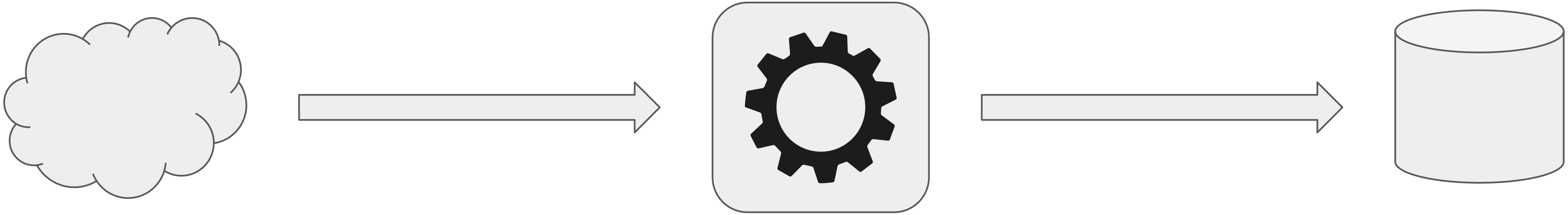
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Man vs machine: let's automate!

https://en.wikipedia.org/wiki/Factory



https://en.wikipedia.org/wiki/Decision_tree_learning

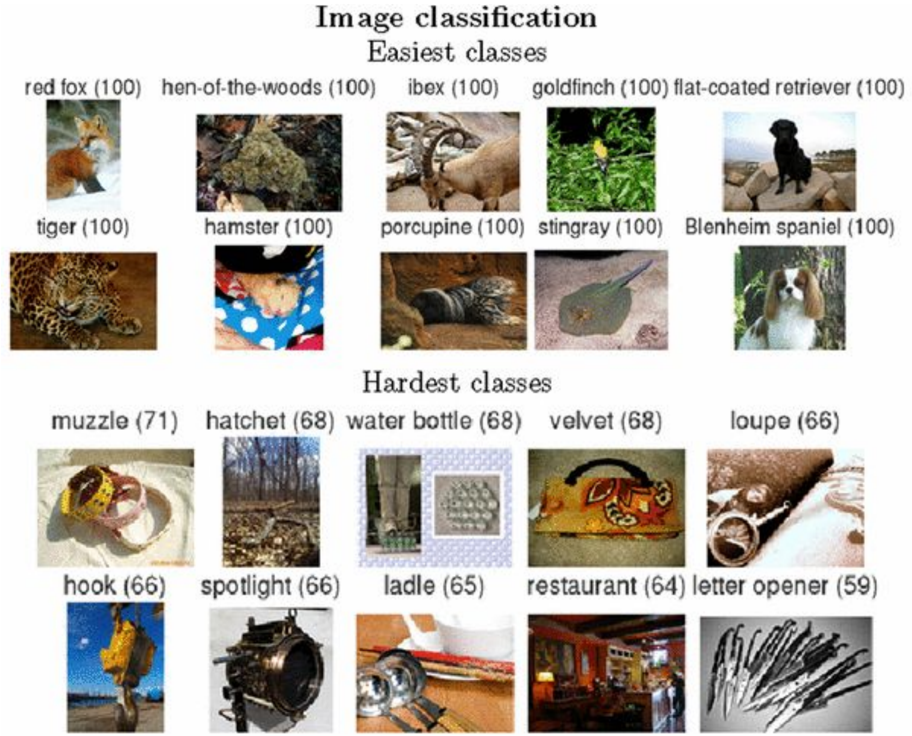


It ain't cheap!

“We train XLNet-Large on 512 TPU v3 chips for 500K steps with an Adam optimizer, linear learning rate decay and a batch size of 2048, which takes about 2.5 days.”

https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/

There is no data!



WIKIPEDIA
The Free Encyclopedia

https://link.springer.com/article/10.1007/s11263-015-0816-y

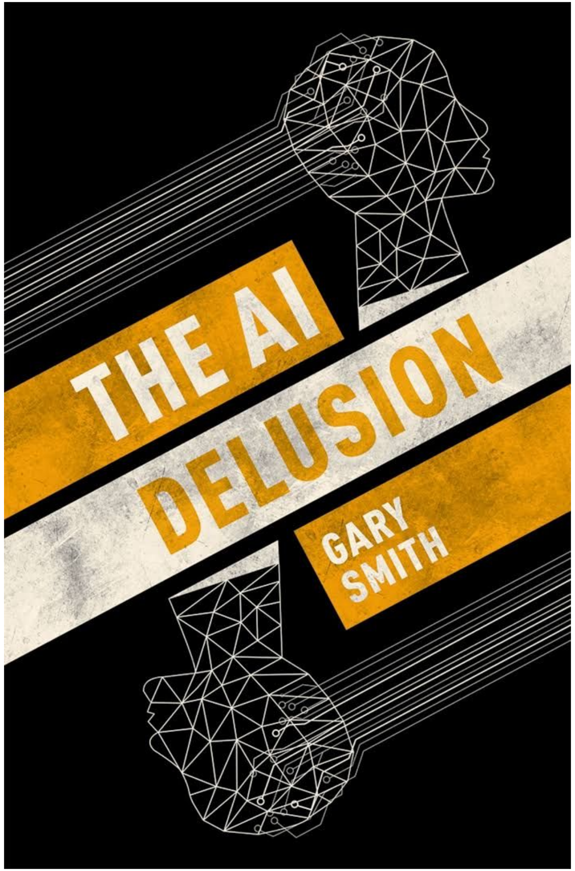
Try it first!

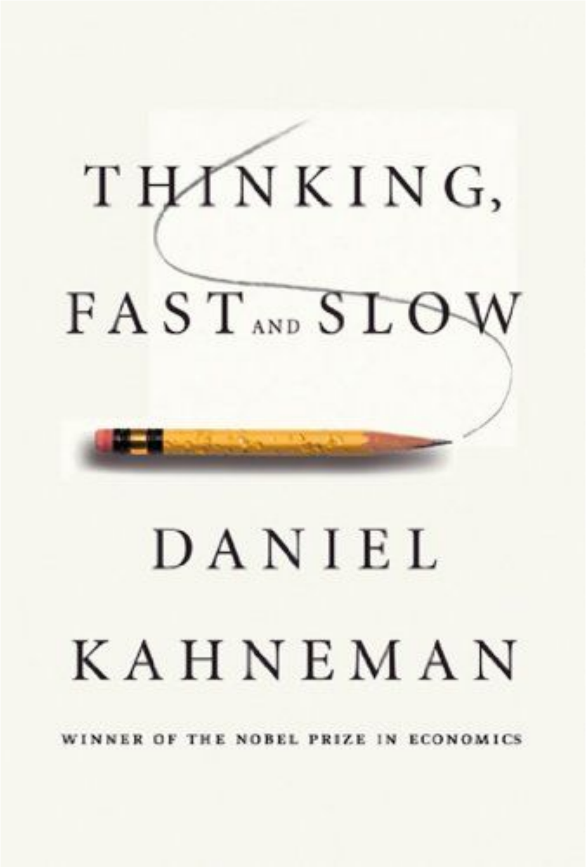
Google Cloud

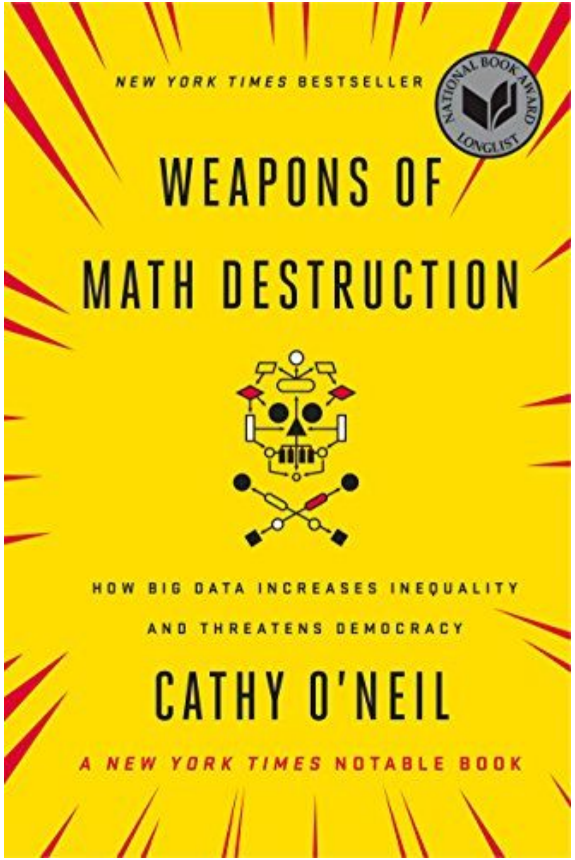
aws

Azure

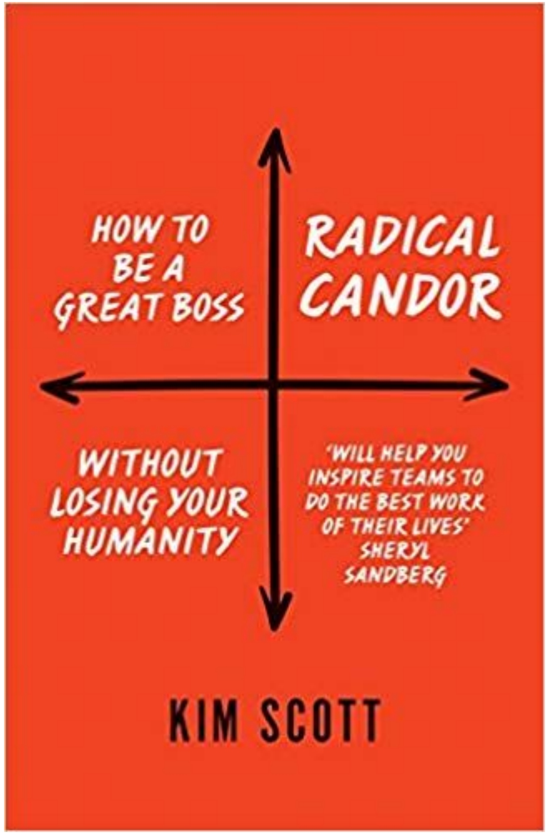
Be careful!


THE AI
DELUSION
GARY SMITH

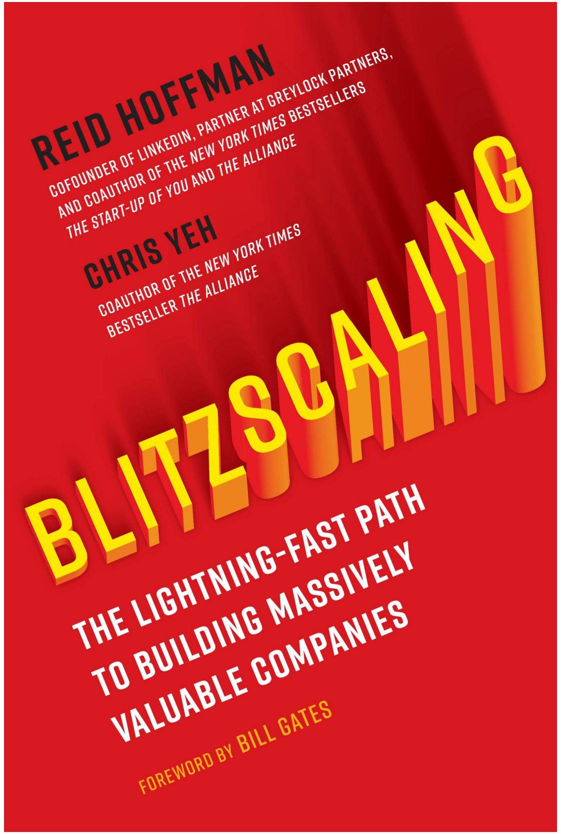

THINKING,
FAST AND SLOW
DANIEL
KAHNEMAN
WINNER OF THE NOBEL PRIZE IN ECONOMICS


NEW YORK TIMES BESTSELLER
NATIONAL BOOK AWARD LONGLIST
WEAPONS OF
MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY
AND THREATENS DEMOCRACY
CATHY O'NEIL
A NEW YORK TIMES NOTABLE BOOK

Different type of teams


HOW TO BE A GREAT BOSS
RADICAL CANDOR
WITHOUT LOSING YOUR HUMANITY
'WILL HELP YOU INSPIRE TEAMS TO DO THE BEST WORK OF THEIR LIVES' SHERYL SANDBERG
KIM SCOTT

CREATIVITY, INC.
OVERCOMING THE UNSEEN FORCES THAT STAND IN THE WAY OF TRUE INSPIRATION
ED CATMULL
President of PIXAR ANIMATION and DISNEY ANIMATION
WITH AMY WALLACE


REID HOFFMAN
COFOUNDER OF LINKEDIN, PARTNER AT GREYLOCK PARTNERS, AND COAUTHOR OF THE NEW YORK TIMES BESTSELLERS THE START-UP OF YOU AND THE ALLIANCE
CHRIS YEH
COAUTHOR OF THE NEW YORK TIMES BESTSELLER THE ALLIANCE
BLITZSCALING
THE LIGHTNING-FAST PATH TO BUILDING MASSIVELY VALUABLE COMPANIES
FOREWORD BY BILL GATES

Thank you!

Questions?