Infrastructure for Machine Learning Applications

Natalie Pistunovich
@NataliePis

Captcha Challenge

Captcha Challenge

1. Inspect the model
2. Load the model
3. Attempt logging in with the PIN:
   i. Open a cookie jar
   ii. Get the CAPTCHA image
   iii. Predict CAPTCHA using ML
   iv. Guess the PIN + CAPTCHA
      a. if false CAPTCHA,
         fall back to (ii)

Captcha Challenge

Read all about it at the December 28 2017
Gophers Academy Advents Blog post

https://github.com/Pisush/break-captcha-tensorflow

Using Machine Learning: Go + TensorFlow

@GopherConRU
#GopherConRU
@NataliePis

What is ML?

Machine learning is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.

It is seen as a subset of artificial intelligence.

- Wikipedia

What is AI?

@theAlexFreeman

In computer science, **artificial intelligence** (**AI**) … is intelligence demonstrated by machines, in contrast to the **natural intelligence** displayed by humans.

- Wikipedia

The term **artificial intelligence** is often used to describe machines (or computers) that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving".

- Wikipedia

How to ML?

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. Choose a model
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

How to ML

1. **Define the problem**
2. Gather data
3. Prepare data
4. Choose a model
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

How to ML

1. Define the problem
2. **Gather data**
   - relevant to the task
3. Prepare data
4. Choose a model
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

How to ML

1. Define the problem
2. Gather data
3. **Prepare data**
4. Choose a model
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

- clean and pre-process
- randomize
- split: train/test

How to ML

1. Define the problem
2. Gather data
3. **Prepare data**
4. Choose a model
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

clean and pre-process
randomize
split: train/test
75/25

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. **Choose a model**
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

- learning task
- input type
- possible number of categories

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. Choose a model
5. **Train the model**
   - assign random values
   - predict the train data
   - adjust weights
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. Choose a model
5. Train the model
6. **Evaluate the model**
   - check test data metrics
7. Tune the hyperparameters
8. Predict

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. Choose a model
5. Train the model
6. Evaluate the model
7. **Tune the hyperparameters**
   - or, fine tune
8. Predict

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. Choose a model
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. **Predict**

TensorFlow

TensorFlow

TensorFlow is an open-source software
for Machine Intelligence,
used mainly for
Machine Learning applications
such as neural networks.

TensorFlow

TensorFlow is an open-source software for Machine Intelligence,
used mainly for machine learning applications such as neural networks.

A tensor is a generalization
of vectors and matrices to
potentially higher dimensions

1. data type
2. shape
   - number of dimensions
   - number of values / dimension

TensorFlow is an open-source software for Machine Intelligence,
used mainly for machine learning applications such as neural networks.

A tensor is a generalization
of vectors and matrices to
potentially higher dimensions

1. data type
2. shape
   - number of dimensions
   - number of values / dimension

The flow part comes to describe:
- the graph (model) is a set of nodes (operations)
- the data (tensors) "flows" through those nodes,
  undergoing mathematical manipulation

You can look at, and evaluate, any node of the graph

TensorFlow

- Community driven

- Becoming friendly for developers
  - AutoML: automates ML models design
  - TF Hub: repo for modules
  - Black-box tools built on top of TF

TensorFlow

- Community driven

- Becoming friendly for developers
  - **AutoML: automates ML models design**
  - TF Hub: repo for modules
  - Black-box tools built on top of TF

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. Choose a model
5. Train the model
6. Evaluate the model
7. **Tune the hyperparameters**
   - or, fine tune
8. Predict

TensorFlow

- Community driven

- Becoming friendly for developers
  - AutoML: automates ML models design
  - **TF Hub: repo for modules**
  - Black-box tools built on top of TF

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. Choose a model
5. **Train the model**
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

- assign random values
- predict the train data
- adjust weights

TensorFlow

- Community driven

- Becoming friendly for developers
  - AutoML: automates ML models design
  - TF Hub: repo for modules
  - **Black-box tools built on top of TF**

So, Infrastructure?

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

Abstract

Machine learning offers a fantastically powerful toolkit for building useful complex prediction systems quickly. This paper argues it is dangerous to think of these quick wins as coming for free. Using the software engineering framework of *technical debt*, we find it is common to incur massive ongoing maintenance costs in real-world ML systems. We explore several ML-specific risk factors to account for in system design. These include boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, configuration issues, changes in the external world, and a variety of system-level anti-patterns.

1 Introduction

As the machine learning (ML) community continues to accumulate years of experience with live systems, a wide-spread and uncomfortable trend has emerged: developing and deploying ML systems is relatively fast and cheap, but maintaining them over time is difficult and expensive.

Infrastructure

There's a lot more to machine learning than just implementing an ML algorithm.

Infrastructure

The ML code is at the heart of a real-world production system, but it accounts for **5% or less** of the overall code of that system.

Infrastructure

The ML code is at the heart of a real-world production system, but it accounts for **5% or less** of the overall code of that system.

hadoop

APACHE
hadoop
TM

APACHE
hadoop
TM

APACHE
hadoop
TM

Distributed File-System

MapReduce
Distributed
File-System

Hadoop
Common
MapReduce
Distributed
File-System

Hadoop
Common
YARN
MapReduce
Distributed
File-System

https://github.com/colinmarc/hdfs

Data Lake Platform

Data Lake Platform

“If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption –

the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples.”

- James Dixon

Data Lake Platform

Google BigQuery

APACHE hadoop

APACHE Spark™

Azure Data Lake

Database

mongoDB

AEROSPIKE

Cockroach DB

Couchbase

Functions as a Service (aka Serverless)

AWS Lambda

Azure Functions

Google Cloud Functions

OPENFAAS

Data Governance

MANTA

Alation

CONTIAMO

dremio

Monitoring

pagerduty

New Relic

DATADOG

splunk>

Putting it All Together: Bare Bones Infrastructure

Parquet
JSON
CSV

Parquet
JSON
CSV
Data Governance

Parquet
JSON
CSV
Data Governance
Data Lake Platform
Database

Parquet
JSON
CSV
Data Governance
Data Lake Platform
Database
Data Governance

Parquet
JSON
CSV
Data Governance
Data Lake Platform
Database
Data Governance
Model Training

Zooming in on Training

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. Choose a model
5. **Train the model**
   - assign random values
   - predict the train data
   - adjust weights
6. Evaluate the model
7. Tune the hyperparameters
8. Predict

Types of Training

Static model (Offline) vs. Dynamic model (Online)

Static Model

+	-
Easier to build and test.	Can only predict things we know about. Update latency likely measured in hours or days.

Dynamic Model

+

Adapted to changing data, hence more likely to make better predictions.

-

Compute intensive, latency sensitive, may limit model complexity.

Monitoring needs are more intensive: outputs and performance.

Parquet
JSON
CSV
Data Governance
Data Lake Platform
Database
Data Governance
Model Training

Parquet
JSON
CSV
Data Governance
Data Lake Platform
Database
Data Governance
Model Training
FaaS

Parquet
JSON
CSV
Data Governance
Data Lake Platform
Database
Data Governance
Model Training
FaaS
Output

Zooming in on Inference

How to ML

1. Define the problem
2. Gather data
3. Prepare data
4. Choose a model
5. Train the model
6. Evaluate the model
7. Tune the hyperparameters
8. **Predict**

Types of Inference

Online Inference vs. Offline Inference

Online

+

Can make a prediction on any new item as it comes in — great for long tail.

-

Compute intensive, latency sensitive — may limit model complexity.

Monitoring needs are more intensive.

Offline

+

Don't worry much about cost of inference.

Likely to use batch quota.

Can do post-verification of predictions on data before pushing.

-

Can only predict things we know about.

Parquet
JSON
CSV
Data Governance
Data Lake Platform
Database
Data Governance
Model Training
FaaS
Output

Parquet
JSON
CSV
Data Governance
Data Lake Platform
Database
Data Governance
Model Training
FaaS
Output
Monitoring

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, Cockroach Labs, TIMESCALE, VOLTDB, splice machine, imply, paradigm4, Trafodion, TiDB

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

MPP DBs
TERADATA, VERTICA, IBM Data Warehouse Systems, Action, 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, Alation, IMMUTA, MANTA, dremio, Waterline Data, OKERA, data.world

MGMT / MONITORING
aws, New Relic, actifio, rubrik, APPDYNAMICS, DATADOG, dynatrace, WAVEFRONT by vmware, SignalFx, druva, splunk, Moogsoft, pagerduty, Anodot, 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, Mesosphere, packet, Hewlett Packard Enterprise, PENGUIN COMPUTING, BrightComputing, CYCLECLOUD

DATA GENERATION & LABELLING
amazon mechanical turk, Upwork, appen, unity, scale, HIVE, Labelbox, Mighty AI, AI.REVERIE, 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, CORNAMI, NALLA, flexlogix, EFINIX

ANALYTICS & MACHINE INTELLIGENCE

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

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

BI PLATFORMS
looker, 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, GOODATA, 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, ADEE, 依图, YITU, trax, CLARIFAI, synthesia, DataGrid

HORIZONTAL AI
IBM Watson, Cortana, Face++ 旷视, sentient, Voyager Labs, vicarious, Affectiva, PROPHESEE, CognitiveScale, Numenta, PETUUM, SCALED INFERENCE, naralogics, THE CURIOUS AI COMPANY, OSARO, BLUE VISION, Pelican, Fortress IQ

SPEECH & NLP
Google Cloud, twilio, amazon alexa, Amazon Translate, narrative science, semantic machines, WolframAlpha, Mobvoi, EigenTechnologies, SoundHound Inc., volcera, PRIMER, MindMeld, cortical.io, NUANCE, Grid.ai, Grialspace, cogito, snips, SMARTLING, Unbabel, PolyAI, PRISM

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

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

APPLICATIONS – ENTERPRISE

SALES
einstein, CHORUS, INSIDESALES.COM, peopleai, conversica, clari, aviso, tact.ai, TROOPS, fuse|machines, Clearbit, collectiveIQ

MARKETING - B2B
RADIUS, App Annie, EVERSTRING, Lattice, MINTIGO, sense, tubular, NGAGIO, Reflektion, 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, CaWDesk, netomi, TechSee, frame.ai

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

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

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

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

PARTNERSHIPS
GROSSBEAM, INFOSUM, DATA REPUBLIC

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

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

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

APPLICATIONS – INDUSTRY

ADVERTISING
AppNexus, MediaMath, criteo, xAd, Integral Ad Science, ORACLE MOAT, OpenX, datorama, dstillery, LiveIntent, Adgorithms, TRADE DESK, 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, PDAM.COM, LIANPU.AI, Kreditech, AVANT, TALA, Zest Finance, Upstart, aura, LEARBANC, upgrade, 100Credit, WeLab, Wecash闪现奇异, 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 Medical Vision, PathAI, ovia health, TEMPUS, patientslikeme, AiCure, insitro, LinkDoc, SIG TUPLE, Y.ai, citizen, notable, komodo health, RECURSION PHARMA, prognos, enlitic, imagen, BlackThorn THERAPEUTICS, BAYLABS, Qventus, ARTERYS, IMAGEN, CLOUD MEDX, PAIGE, DATAVANT, innovaccer, HUMAN DIAGNOSIS PROJECT, LeanTaaS

LIFE SCIENCES
23andMe, color, CARBON BITBIO, BenevolentAI, verily, WuXiNextCODE, ZIMMER BIOMET, DATAWAY GENOMICS, AiCure, ESCRIPT, Clear Labs, freenome, NANOPORE, DNAnexus, Phosphorus, CITRINE INFORMATICS, twoXAR, Atomwise, deep genomics, SOPHIA GENETICS, OWKIN

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

AGRICULTURE
FARMERS, Granular, JOHN DEERE, BLUE RIVER TECHNOLOGY, FarmersEdge, AgroStar, FarmLogs, TARANIS, GAMAYA, TerraMotus, prospera

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

INDUSTRIAL
AVEVA, SIEMENS, Schneider, PREDIX, UPTAKE, SIGHT MACHINE, X-ARRAY, OSIsoft, SCORTEX, KONUX, TACHYUS, Fero Labs

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

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, SLAMDATA, APACHE DRILL, GraphQL, Google Cloud Dataflow, Flink

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

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

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

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

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

AI / MACHINE LEARNING / DEEP LEARNING
TensorFlow, Keras, torch, MAPLib, learn, Caffe, Microsoft Cognitive Toolkit, OpenAI, DMTK, theano, PaddlePaddle, Apache SINGA, DIMSUM, FeatureFu, mxnet, VELES, Chainer, Michelangelo, ONNX, WEKA, Weka, 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, D3.js

COLLABORATION
BeakerX, jupyter, Zeppelin, ANACONDA

SECURITY
Apache Ranger, KNOX, Sentry, accumulo

DATA SOURCES & APIs

HEALTH
validic, practicefusion, 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., SKYCATCH, AIROBOTICS, spire, PRECISIONHAWK, Descartes Labs, kespry, UNDERSTORY, tellusLabs, WINDWARD, DroneDeploy, MarineTraffic, LOFT ORBITAL, RS Metrics

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

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

OTHER
DATA.GOV, IMAGENET, wiki-links, LabelMe, Berkeley DeepDrive, 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

© Matt Turck (@mattturck), Lisa Xu (@lisaxu92), & FirstMark (@firstmarkcap)

mattturck.com/data2019

FIRSTMARK EARLY STAGE VENTURE CAPITAL

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 HOME OF REDIS, AEROSPIKE, ArangoDB, SCYLLA

NewSQL DATABASES

SAP HANA, Clustrix, NUODB, Pivotal, MariaDB, Cloud Spanner, MEMSQL, influxdata, Cockroach LABS, TIMESCALE, VOLTDB, splice MACHINE, imply, paradigm4, Trafodion, TIDB

GRAPH DBs

neo4j, Amazon Neptune, IBM, ORACLE, OrientDB, GIRAPH, 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 A Hitachi Group Company, alteryx, TRIFACTA, tamr, Paxata, StreamSets, UNIFI

DATA INTEGRATION

SAP Data Services, Informatica, MuleSoft, TEALIUM, snaplogic, enigma, Qlik Data Catalyst, Segment, ATTUNITY, xplenty, ZALONI THE DATA LAKE COMPANY, 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, dynatrace, WAVEFRONT by vmware, DATADOG, SignalFx, druva, Anodot, splunk>, Moogsoft, pagerduty, unravel, Numerify, SCALYR, ScienceLogic, veeam, zenoss Own IT, OpsRamp, MAGNITUDE

STORAGE

aws, Google Cloud, Microsoft Azure, IBM Storage, PURESTORAGE, ALLUXIO, wasabi hot cloud storage, nimblestorage, Qumulo, panasas, COHESITY

CLUSTER SVCS

Amazon ECS, IBM, Amazon EKS, IBM Power Systems, Microsoft Azure, Cluster Server, MESOSPHERE, packet, pepperdata, Hewlett Packard Enterprise, PENGUIN COMPUTING, Bright Computing, CYCLECLOUD

DATA GENERATION & LABELLING

amazon mechanical turk Artificial Artificial Intelligence, Upwork, unity, appen, scale, HIVE, Labelbox, Mighty Ai, ALLEGEDIE, LIONBRIDGE

AI OPS

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

GPU DBs & CLOUD

kinetica, omnisci, SQREAM TECHNOLOGIES, 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 TECHNOLOGIES, CORNAMI, HAILO, flexlogix, EFINIX

CROSS-INFRASTRUCTURE/ANALYTICS

aws, Google Cloud, Microsoft, IBM, SAP, Hewlett Packard Enterprise, sas, 1010DATA, vmware, TIBCO, TERADATA, ORACLE, NetApp, syncsort, MAPR, cloudera, C3.ai

OPEN SOURCE

FRAMEWORKS

hadoop HDFS, Spark, hadoop MapReduce, Pulsar, Flink, YARN, TEZ, Apache Kylin, MESOS, kubernetes, docker, CDAP, Red Hat, HighAvailability, 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, Sqoop, OPENTSDB, SciDB, riak, APACHE HBASE, Cloud Spanner, accumulo

ORCHESTRATION & MGMT

talend, Apache Zookeeper, OOZIE, Apache Ambari, Apache Airflow, Apache MESOS, etcd, Kong

STREAMING & MESSAGING

Spark, APEX, nifi, Flink, beam, kafka, STORM, Apache RocketMQ

STAT TOOLS & LANGUAGES

python, R, Scala, NumPy, PYRO, Studio, SciPy, julia, tidyverse

AI OPS & INFRA

Pachyderm, mlflow, Kubeflow, mleap, DVC, SELDON, Polyaxon, snorkel

AI / MACHINE LEARNING / DEEP LEARNING

TensorFlow, Keras, torch, MADlib, scikit learn, Caffe, Microsoft Cognitive Toolkit, OpenAI, DMTK, theano, PaddlePaddle, Apache SINGA, DIMSUM, FeatureFu, mxnet, VELES, Chainer, Michelangelo, ONNX, Keystone, 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, ggplot2, Bokeh

COLLABORATION

BeakerX, jupyter, Zeppelin, ANACONDA

SECURITY

Apache Ranger, KNOX, Sentry, accumulo

End to End
Off the Shelf Solution

TensorFlow
Extended
Example
Valida tor
ExampleGen
StatisticsGen
SchemaGen
TRAINING &
EVAL DATA
Transform
Trainer
Evalua tor
Model
Valida tor
Pusher
TENSORFLOW
SERVING
TENSORFLOW
JS
TENSORFLOW
LITE

TensorFlow
Extended
Example
Valida tor
ExampleGen
StatisticsGen
SchemaGen
TRAINING &
EVAL DATA
Transform
Trainer
Evalua tor
Model
Valida tor
Pusher
TENSORFLOW
SERVING
TENSORFLOW
JS
TENSORFLOW
LITE

TensorFlow
Extended
Example
Valida tor
ExampleGen
StatisticsGen
SchemaGen
TRAINING &
EVAL DATA
Transform
Trainer
Evalua tor
Model
Valida tor
Pusher
TENSORFLOW
SERVING
TENSORFLOW
JS
TENSORFLOW
LITE

A Concrete Examples

Gmail

Search mail

New Message

Recipients

Subject

Sans Serif

Send

Gmail
Search mail
New Message
Recipients
Subject
-Sender email
-Receiver email
-Email title
-Email content
-Header
-Footer
Sans Serif
Send

- Sender email
- Receiver email
- Email title
- Email content
- Header
- Footer

- Sender email
- Receiver email
- Email title
- Email content
- Header
- Footer

AEROSPIKE

Presence of a recognised header	1
Presence of an official signature	0
Email structure	0.7
Language	0.78
Frequency of “special price”	0
Frequency of “prince”	0.9
Grammatical correctness	0.61

- Sender email
- Receiver email
- Email title
- Email content
- Header
- Footer
AEROSPIKE
Input Feature Vector

- Sender email
- Receiver email
- Email title
- Email content
- Header
- Footer
AEROSPIKE
Input Feature Vector
AWS
Lambda

- Sender email
- Receiver email
- Email title
- Email content
- Header
- Footer
AEROSPIKE
Input Feature Vector
In-house ML model
Online served
AWS
Lambda

- Sender email
- Receiver email
- Email title
- Email content
- Header
- Footer
AEROSPIKE
Input Feature Vector
In-house ML model
Online served
AWS Lambda
Spam / not spam

Recap

- What's AI? What's ML? How to ML?
- TensorfFlow
- Training and Inference: online vs. Offline
- Bare bones architecture components, the shelf full platforms
- A concrete architecture example

To Summarize

- ML code is about **5% or less** of the overall code of a system
- Input data is a big and tricky part of an ML system
- Bare bones flow:

  process data → store data → train model →

  → serve predictions → monitor

@NataliePis

@GopherConRU

#GopherConRU

Thank You!

@NataliePis
@GopherConRU
#GopherConRU