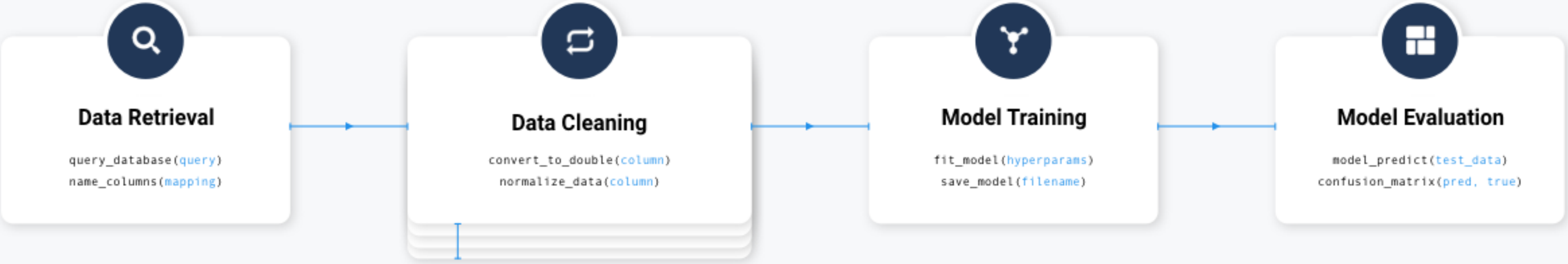
Telkomsel
by Telkom Indonesia

SPILL THE TECH WORKSHOP

Data Science Pipeline

Randy Galawana
Data Engineer

Data Science Pipeline in General



Data Retrieval

```
query_database(query)
name_columns(mapping)
```

Data Cleaning

```
convert_to_double(column)
normalize_data(column)
```

Model Training

```
fit_model(hyperparams)
save_model(filename)
```

Model Evaluation

```
model_predict(test_data)
confusion_matrix(pred, true)
```

Data Engineer Domain

- Big Data
- Different Tech in all use case
- Data Warehouse and Data Lake
- Advance Data Transformation
- Data Quality Checks

Data Scientist Domain

- Sample Data
- Using most common tech / playground
- Statistic and Visualization
- Simple Data Cleaning
- Simple Data Transformation

Data Retrieval using Pandas

Retrieve data from known data sources

Database (JDBC / ODBC)

Using pandas read_sql function with sqlalchemy connection, example in notebook. Support most of ODBC and JDBC database

IO File

Using pandas io read https://pandas.pydata.org/docs/user_guide/io.html

Support almost file extension and format

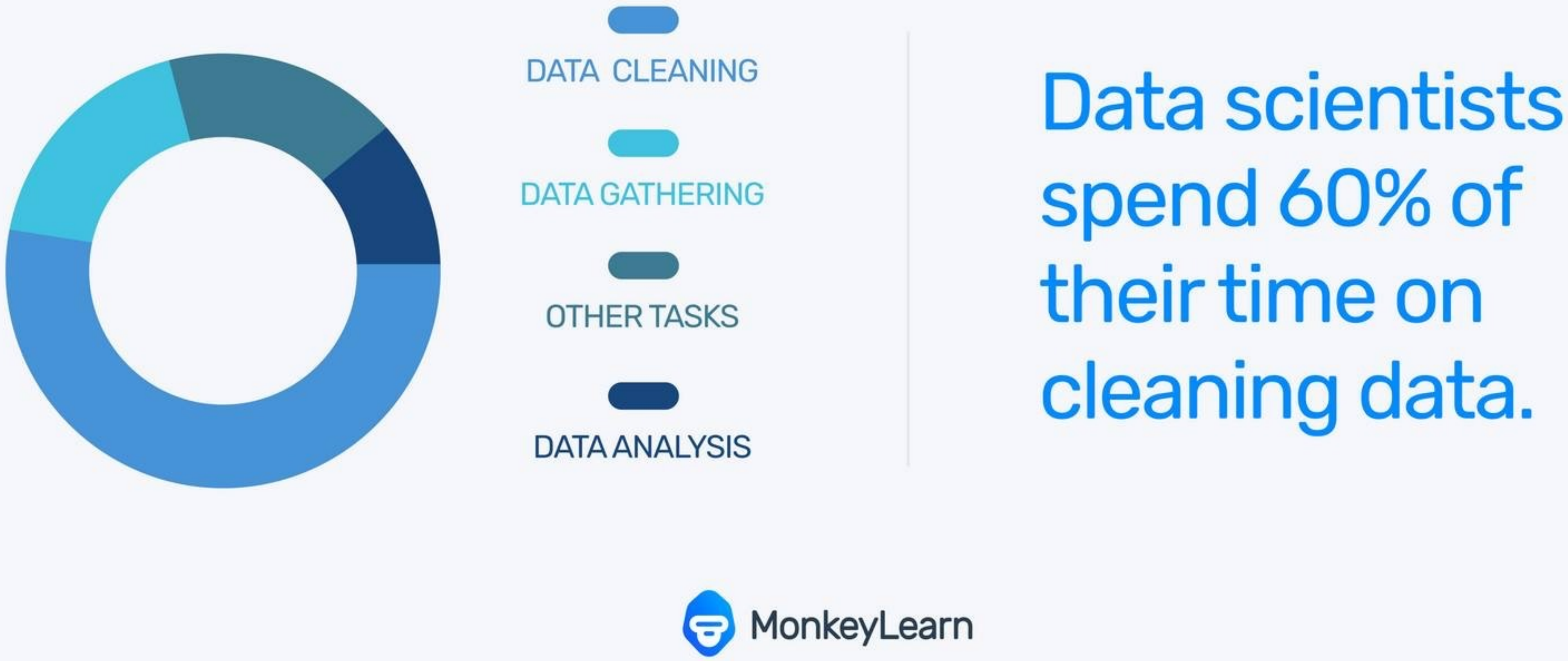
API’s

Using pandas io read https://pandas.pydata.org/docs/user_guide/io.html

Support almost file extension and format

Data Cleaning

Why Data Cleaning need to include in Pipeline



Data scientists spend 60% of their time on cleaning data.

MonkeyLearn

Data Cleaning using Pandas

Most common Data cleaning using pandas and python

Handling Null Values / Missing

Imputer

- Simple Imputer
- Mode, Mean Values
- Hot Deck

Drop Na

- Thresholding

Standardization and Normalization

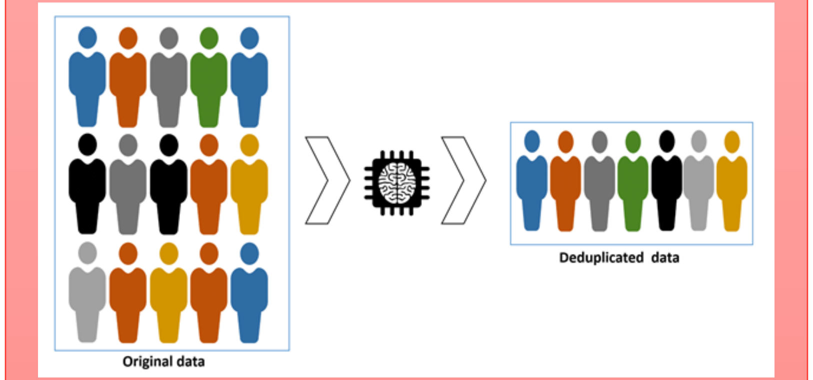
Category Encoder

- One hot encoding
- Label encoder, etc

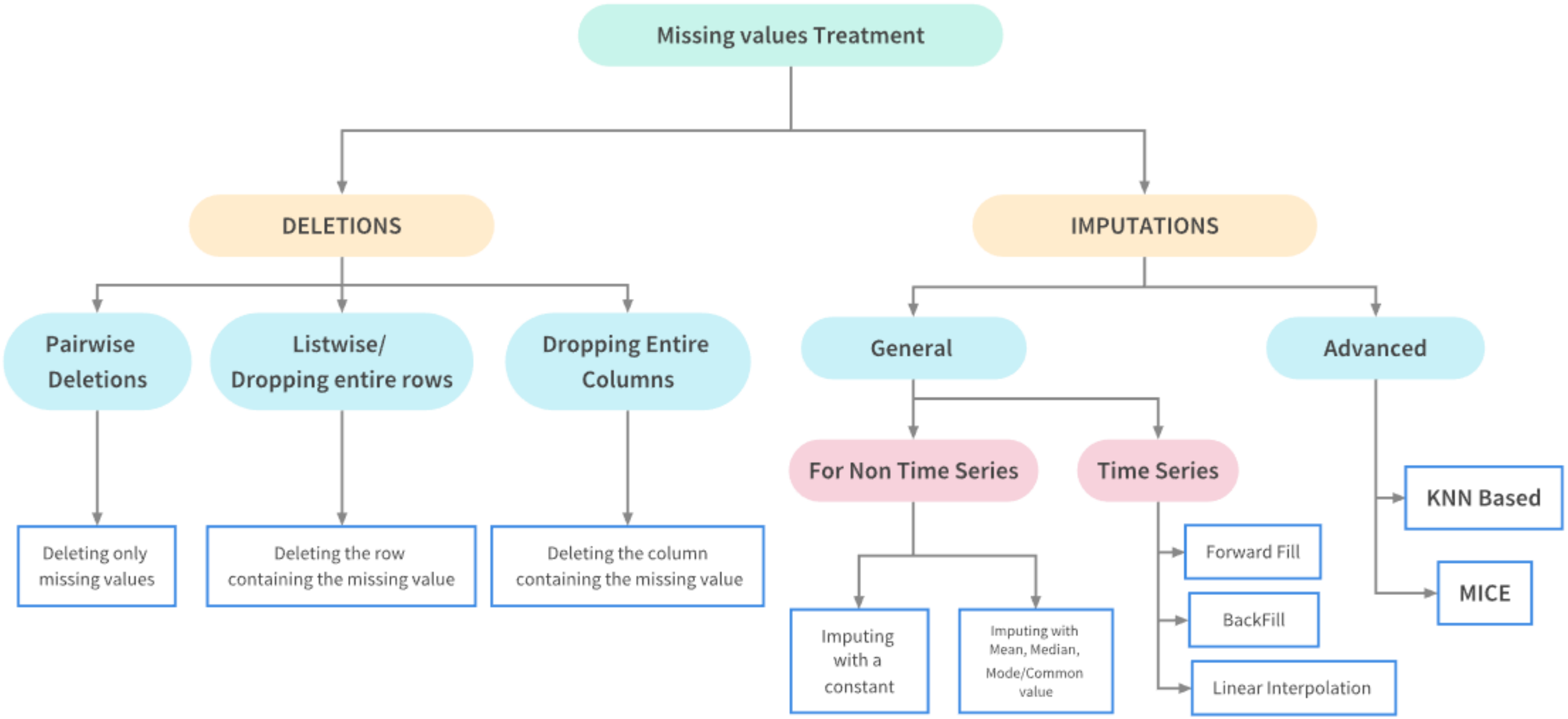
Scaler

- MinMaxScaler
- StandartScaler, etc

De-Duplication



Missing Value Treatment



Missing values Treatment

DELETIONS

Pairwise Deletions

Deleting only missing values

Listwise/ Dropping entire rows

Deleting the row containing the missing value

Dropping Entire Columns

Deleting the column containing the missing value

IMPUTATIONS

General

For Non Time Series

Imputing with a constant

Imputing with Mean, Median, Mode/Common value

Time Series

Forward Fill

BackFill

Linear Interpolation

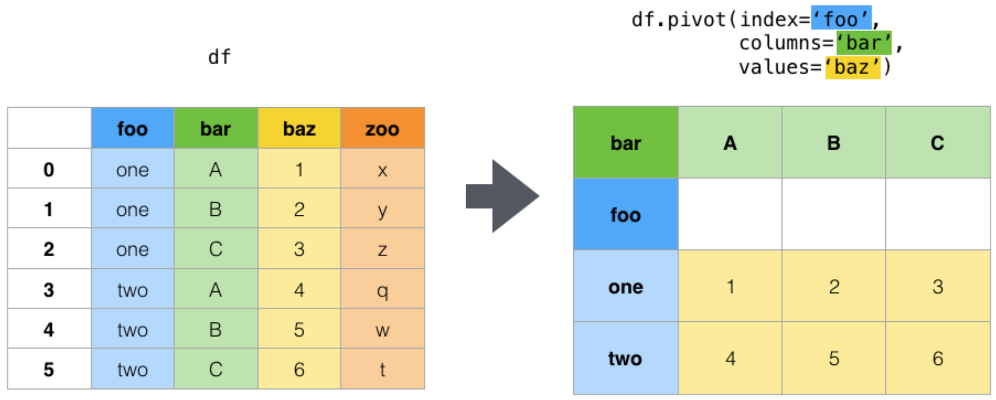
Advanced

KNN Based

MICE

Data Transformation using pandas

Pivot



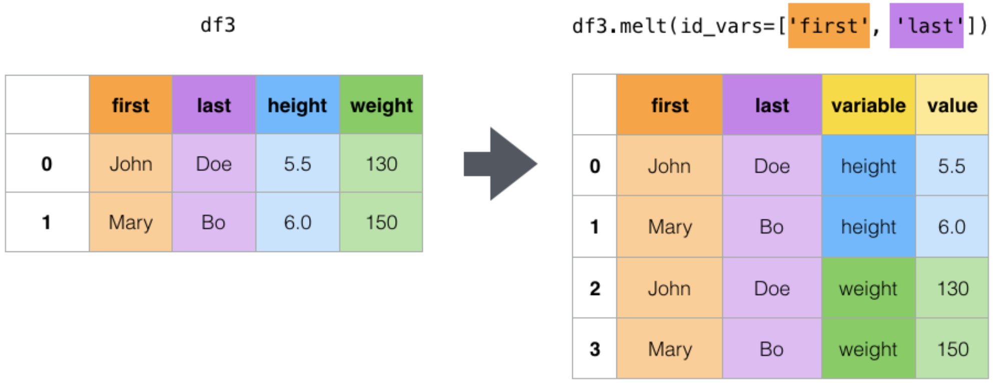
df

	foo	bar	baz	zoo
0	one	A	1	x
1	one	B	2	y
2	one	C	3	z
3	two	A	4	q
4	two	B	5	w
5	two	C	6	t

```
df.pivot(index='foo',
         columns='bar',
         values='baz')
```

bar	A	B	C
foo			
one	1	2	3
two	4	5	6

Melt



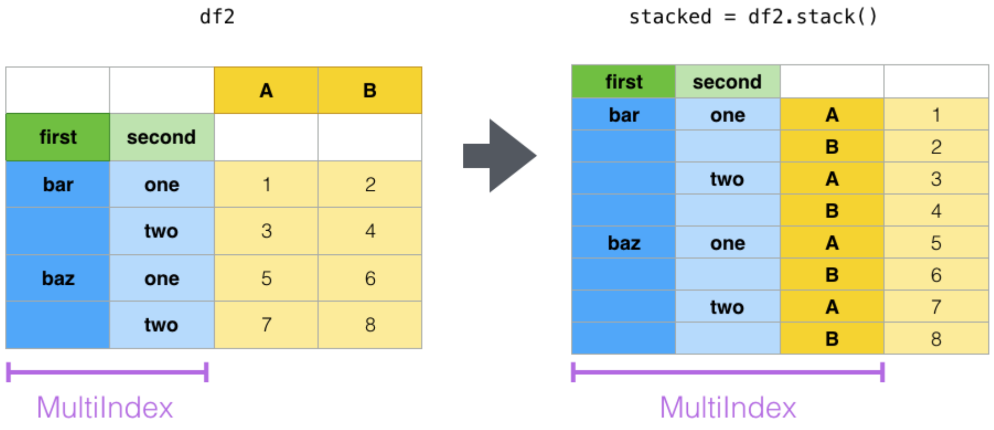
df3

	first	last	height	weight
0	John	Doe	5.5	130
1	Mary	Bo	6.0	150

```
df3.melt(id_vars=['first', 'last'])
```

	first	last	variable	value
0	John	Doe	height	5.5
1	Mary	Bo	height	6.0
2	John	Doe	weight	130
3	Mary	Bo	weight	150

Stack



df2

		A	B
first	second		
bar	one	1	2
	two	3	4
baz	one	5	6
	two	7	8

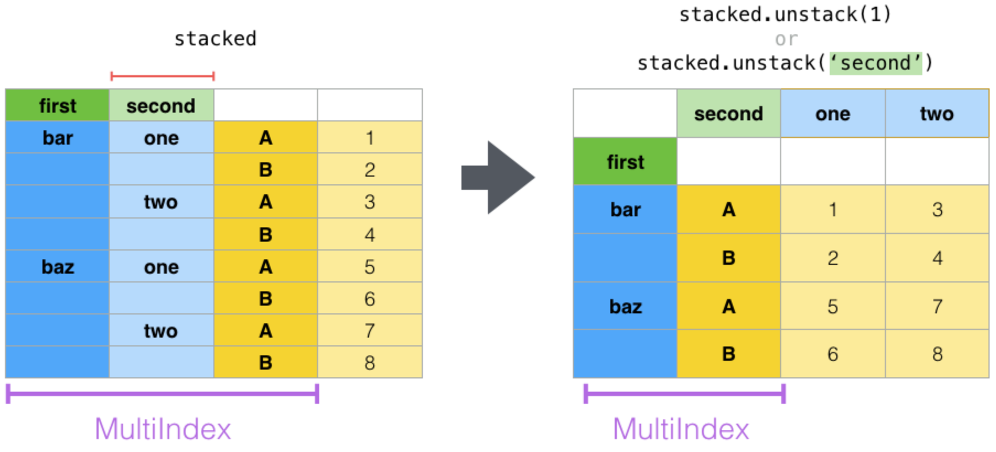
MultiIndex

```
stacked = df2.stack()
```

first	second		
bar	one	A	1
		B	2
	two	A	3
		B	4
baz	one	A	5
		B	6
	two	A	7
		B	8

MultiIndex

Unstack(1)



stacked

first	second		
bar	one	A	1
		B	2
	two	A	3
		B	4
baz	one	A	5
		B	6
	two	A	7
		B	8

MultiIndex

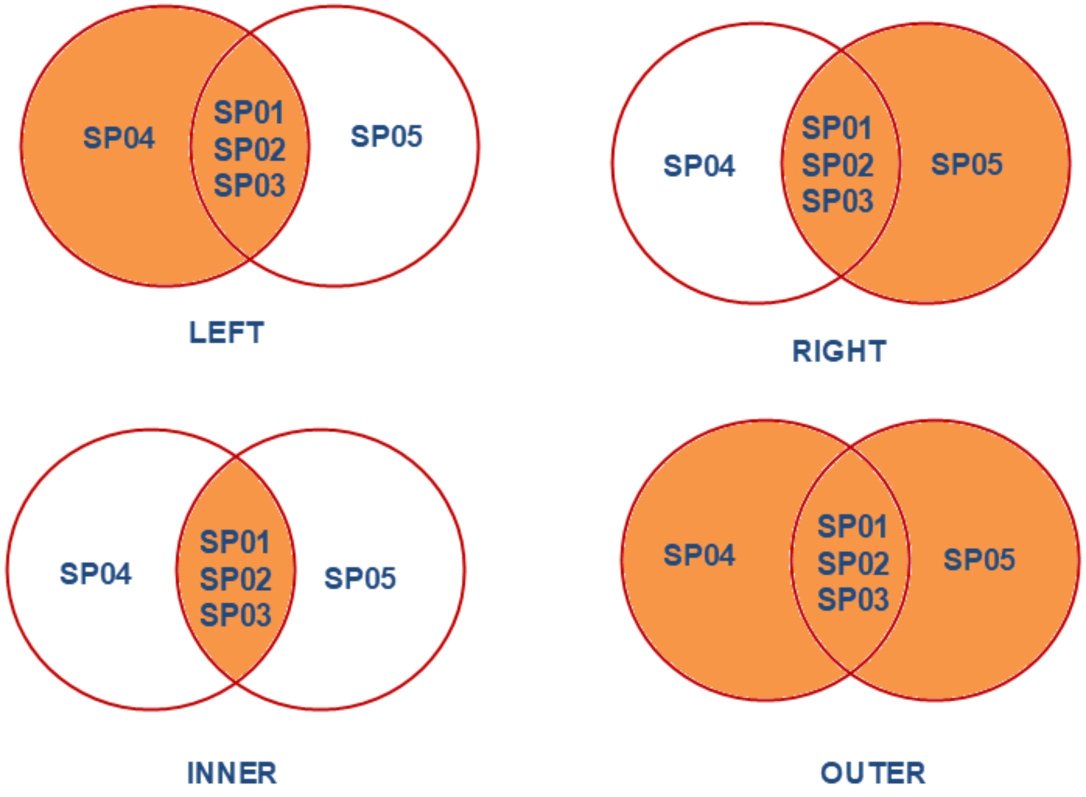
```
stacked.unstack(1)
or
stacked.unstack('second')
```

	second	one	two
first			
bar	A	1	3
	B	2	4
baz	A	5	7
	B	6	8

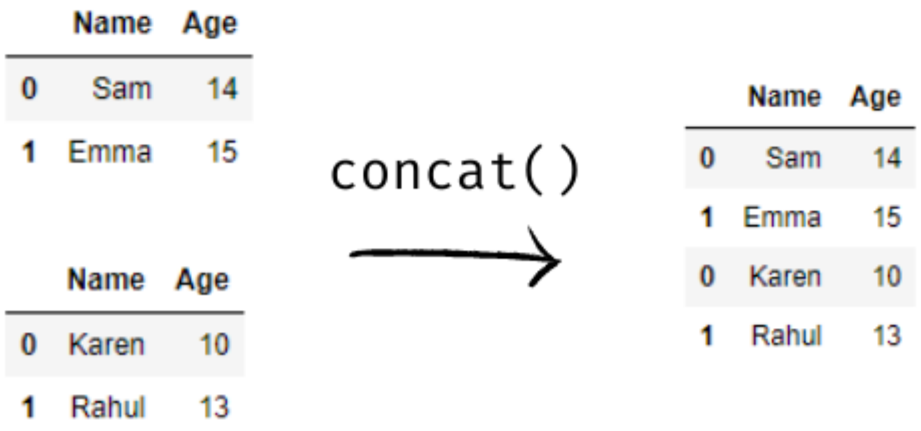
MultiIndex

Data Transformation using pandas

Join / Merge



Concat



	Name	Age
0	Sam	14
1	Emma	15

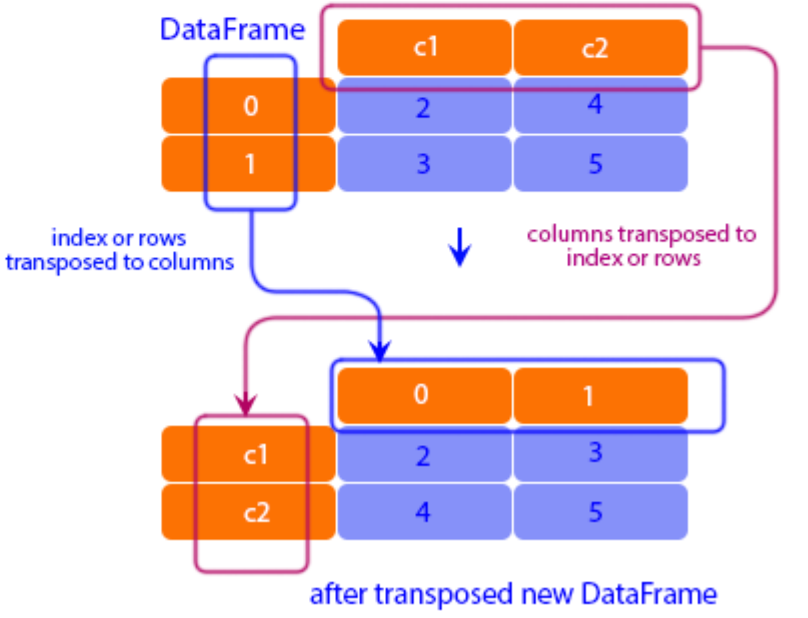
	Name	Age
0	Karen	10
1	Rahul	13

concat()

	Name	Age
0	Sam	14
1	Emma	15
0	Karen	10
1	Rahul	13

Data Transformation using pandas

Transpose

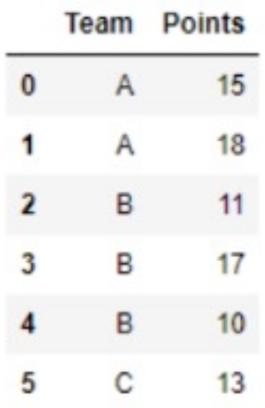

DataFrame
index or rows transposed to columns
columns transposed to index or rows
after transposed new DataFrame

	c1	c2
0	2	4
1	3	5

	0	1
c1	2	3
c2	4	5

©w3resource.com

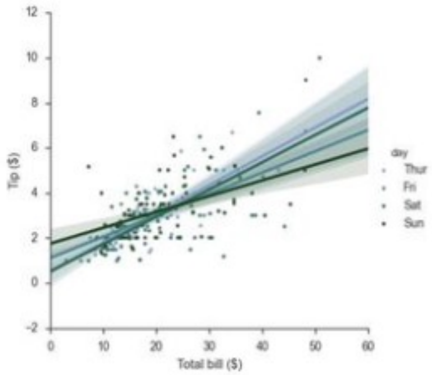
Group by / Aggregate

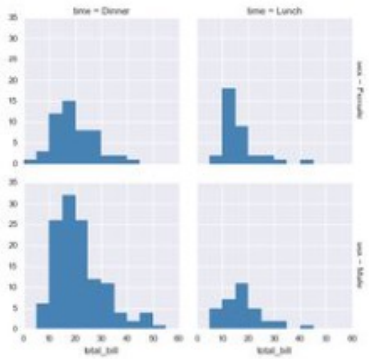


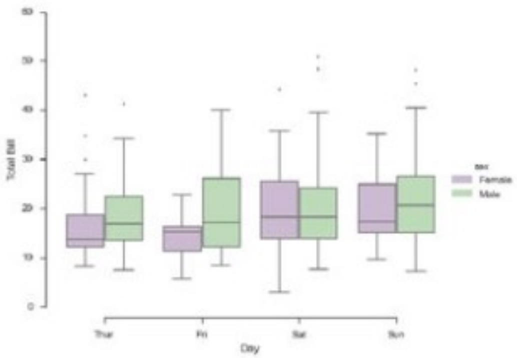
	Team	Points
0	A	15
1	A	18
2	B	11
3	B	17
4	B	10
5	C	13

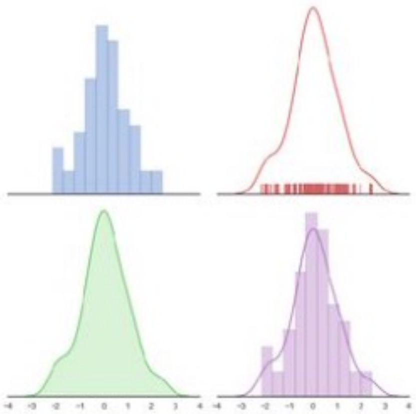
```
Team
A      2
B      3
C      1
dtype: int64
```

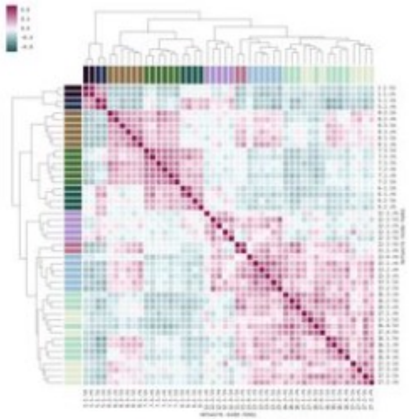
EDA and Data Viz using Seaborn

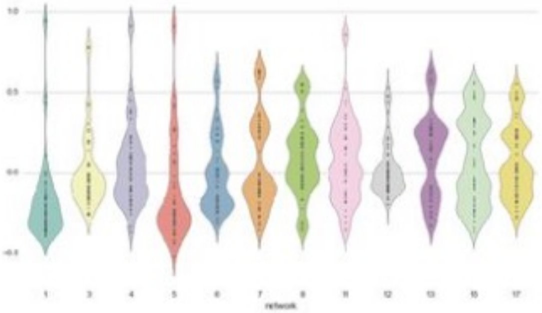








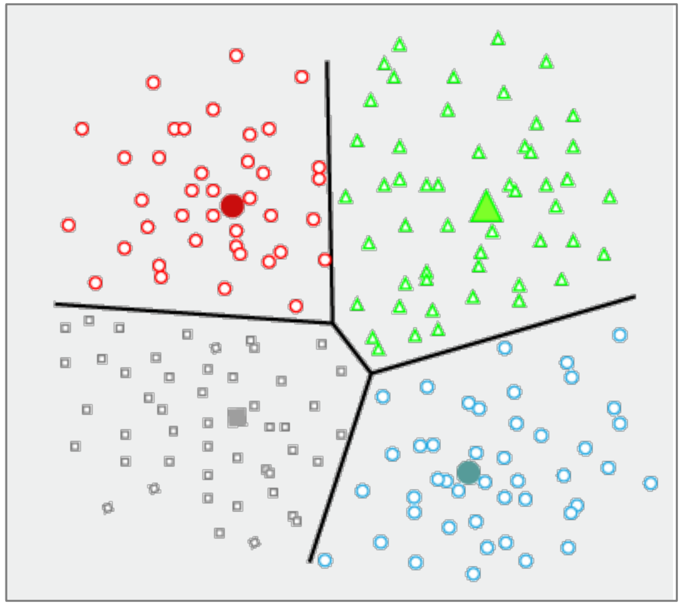




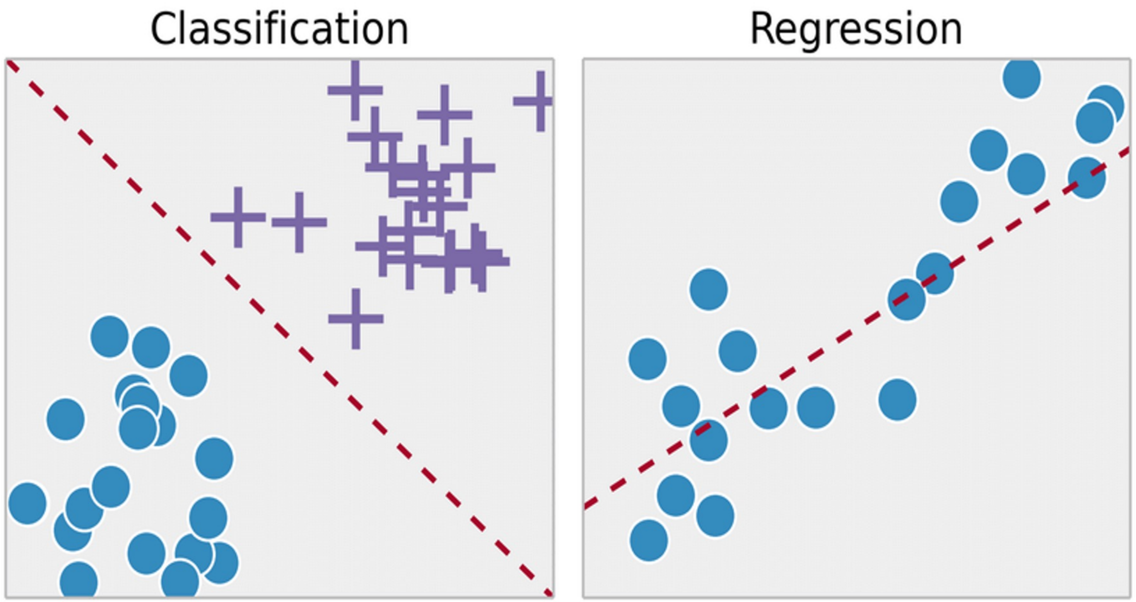
Machine Learning

Unsupervised Learning

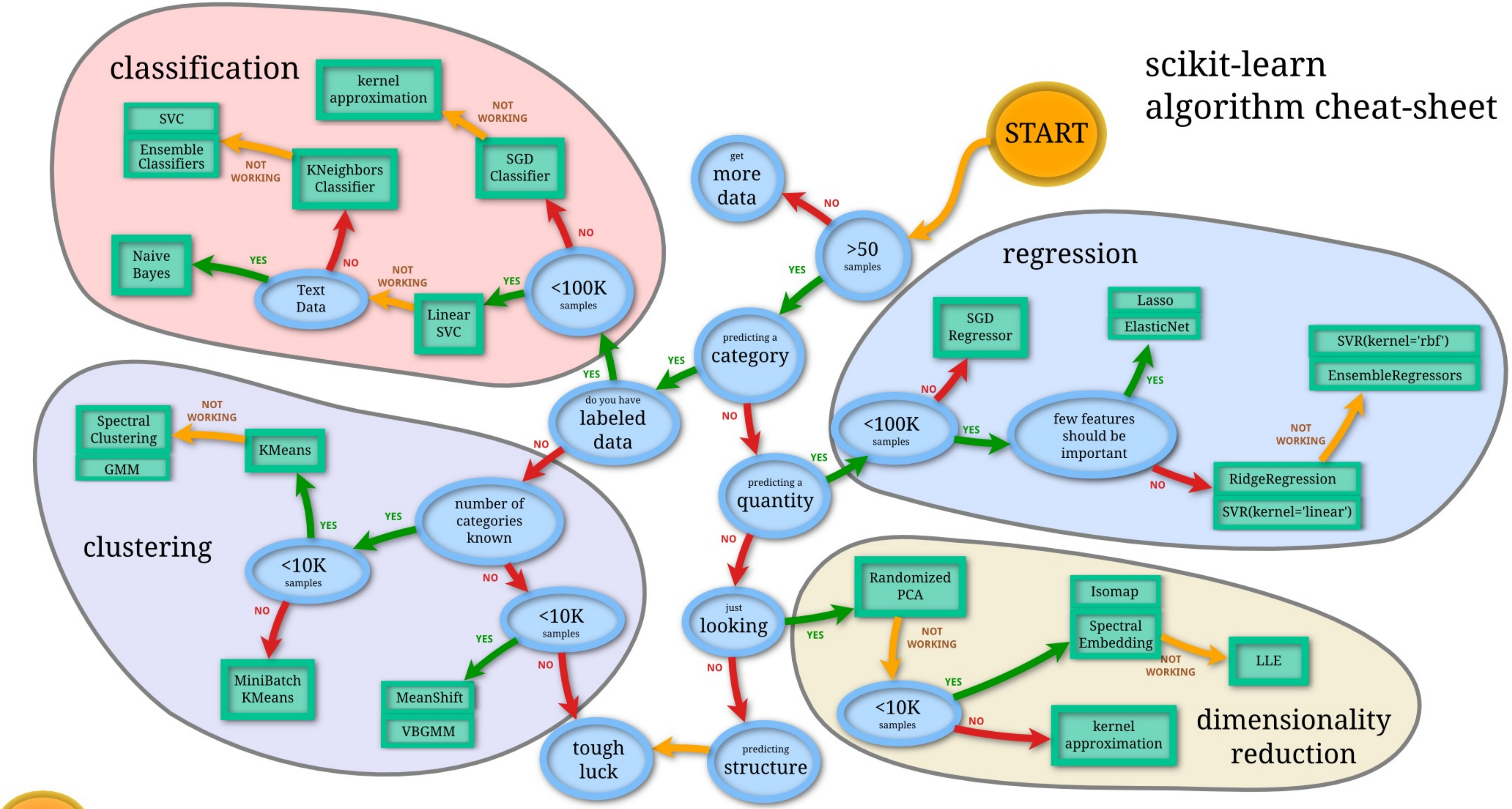
Clustering



Supervised Learning

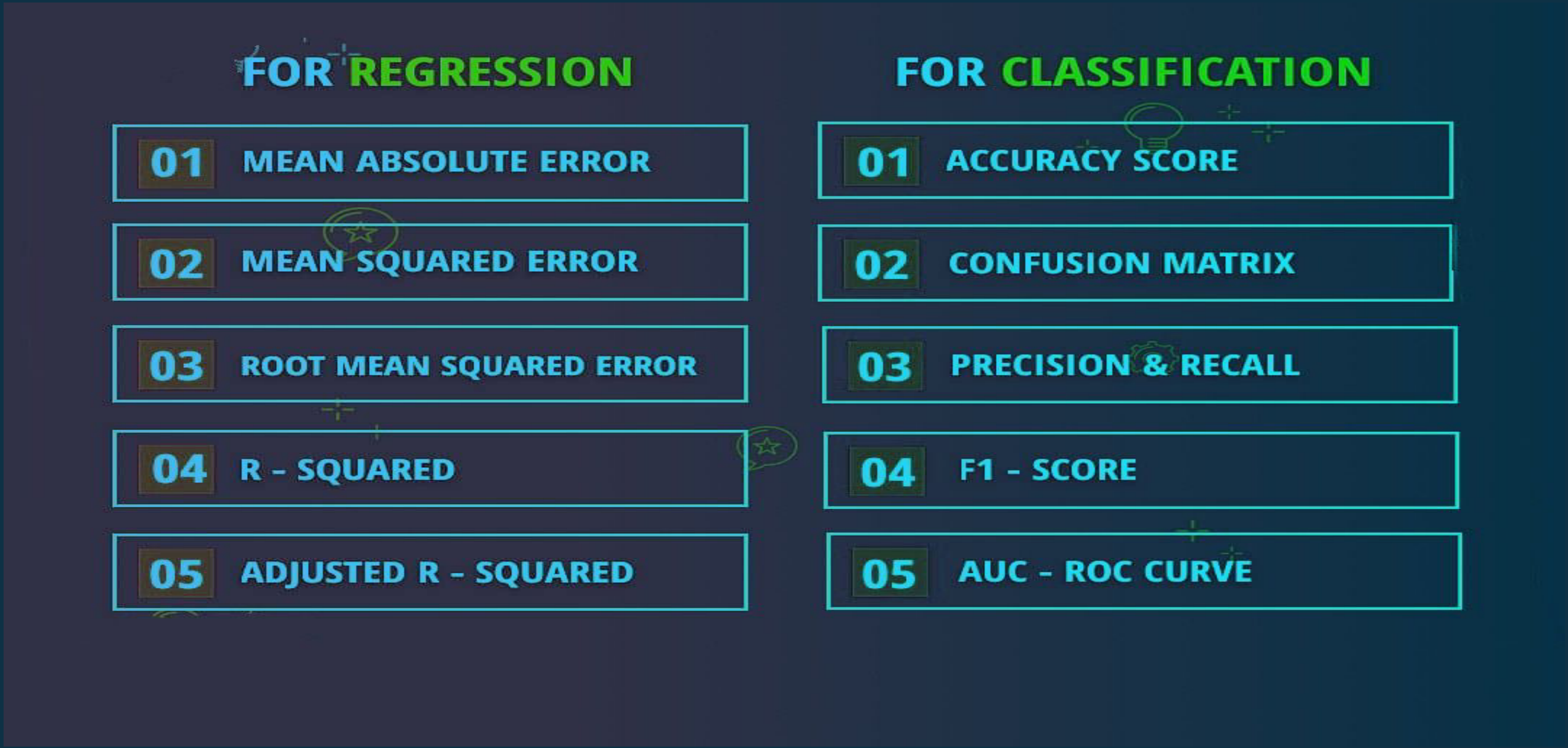


Machine Learning Using Scikit Learn



Back

Machine Learning Model Evaluation



FOR REGRESSION

01 MEAN ABSOLUTE ERROR

02 MEAN SQUARED ERROR

03 ROOT MEAN SQUARED ERROR

04 R – SQUARED

05 ADJUSTED R – SQUARED

FOR CLASSIFICATION

01 ACCURACY SCORE

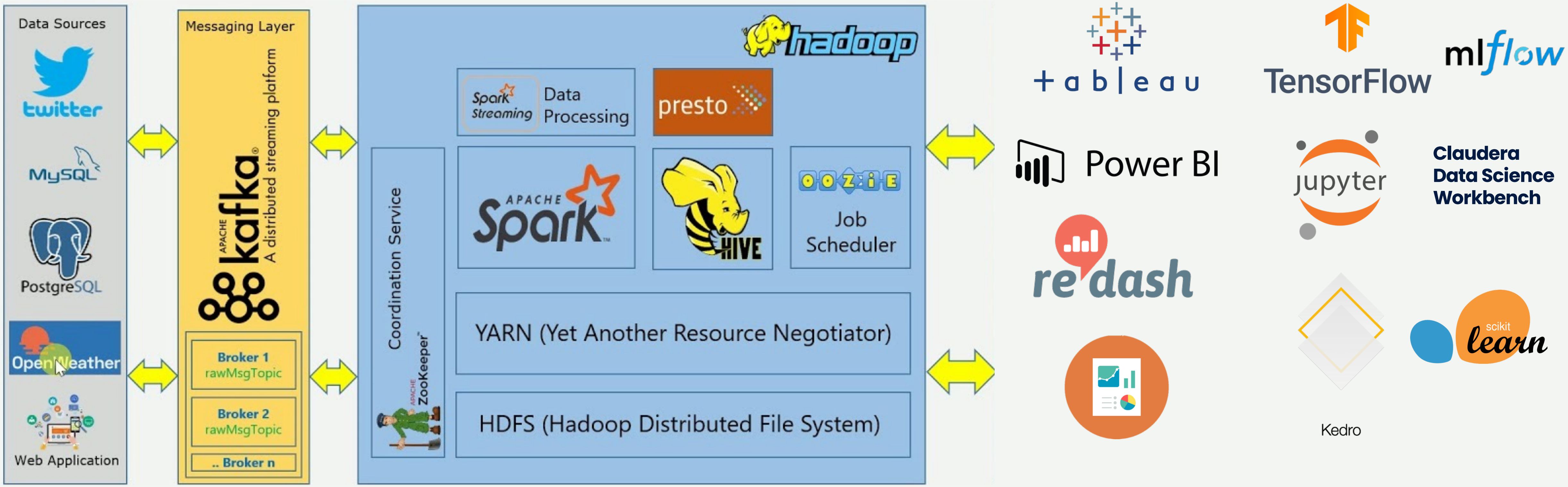
02 CONFUSION MATRIX

03 PRECISION & RECALL

04 F1 – SCORE

05 AUC – ROC CURVE

Sample Big Data ML Pipeline Stack



- Data Ingestion : Kafka, Apache NIFI, Sqoop
- Orchestrator : Apache NIFI, Airflow, Oozie
- Big Data Lake : Hadoop (HDFS, Hive)
- Data Transformation : Apache Spark (PySpark), Hive SQL, Apache Flink

- Data Visualization : Tablaeu Server, ReDash, PowerBI, Dashboard web
- Machine Learning : CDSW, SparkML, scikit-learn, kedro, MLFlow
- Data Quality : Apache Griffin

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, Cloud Spanner, MariaDB, MEMSQL, influxdata, Cockroach Labs, TIMESCALE, VOLTDB, splice, 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, WAVEFRONT by vmware, dynatrace, SignalFx, druva, DATADOG, splunk, Moogsoft, pagerduty, unravel, Numerify, ScienceLogic, SCALYR, VEEAM, zenoss, OpsRamp, MAGNITUDE

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

CLUSTER SVCS
Amazon ECS, IBM, Amazon EKS, Cluster Server, MESOSPHERE, packet, Rancher, kubernetes, Bright Computing, CYCLECLOUD

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

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

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

HARDWARE
Google TPU, arm, intel AI, NVIDIA, GRAPHCORE, IBM Power Systems, MYTHIC, cerebras, Movidius, habana, WAVE COMPUTING, CERNAMI, 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, Google Cloud, celonis, Periscope Data, zepl, GOOD DATA, 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, ADEE, 依图 YITU, trax, synthesia, DataGrid

HORIZONTAL AI
IBM Watson, Cortana, Face++ 旷视, sentient, Voyager Labs, vicarious, Affectiva, PROPHESEE, CognitiveScale, Numenta, PETUUM, SparkCognition, narrative science, nara logics, CURIOUS AI, OSARO, BLUE VISION, Pefinite, Fortress

SPEECH & NLP
Google Cloud, twilio, amazon alexa, Amazon Translate, narrative science, semantic machines, Mobvoi, EigenTechnologies, PRIMER, SoundHound Inc., voicera, MindMeld, cogito, snips, 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, 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, Persado, ACTIONIQ, BLUECORE, CONTENTSQUARE, TEALIUM, mparticle, Amplero, amperity, QUANTIFIND, Simon, Liftigniter, 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, frame.ai

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

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

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, softomotive, KRYON, ALKYMI

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

APPLICATIONS – INDUSTRY

ADVERTISING
AppNexus, MediaMath, criteo, Xad, Integral Ad Science, ORACLE MOAT, OpenX, datorama, theTradeDesk, Algorithmia, dstillery, LiveIntent, TAPAD, dataxu, gumgum, Appier, teemo, yieldmo, mobvoi

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

REAL ESTATE
REDFIN, Opendoor, VTS, CREDIFI, reonomy, SmartProcure, STREETLIGHTDATA, OpenDataSoft

GOV'T
OPENGOV, mark43, FiscalNote, LiveStories, Passport, GEOPHY, COMPSTAK, SPACEMAKER, SKYLINE

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闪银奇异, TrueAccord, MoneyLion, Active.Ai, aire, cignifi

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

HEALTHCARE
flatiron, Clover, KYRUUS, HealthTap, Ginger.io, Glow, babylon, 3DMed, zebra, PathAI, OVIA, TEMPUS, patientslikeme, AiCure, insitro, LinkDoc, SIGTUPLE, ciitizen, notable, komodo health, RECURSION, prognos, enlitic, imagia, BlackThorn, BAYLABS, Qventus, ARTERYS, IMAGEN, CLOUD MEDX, PAIGE, DATAVANT, innovaccer, DIAGNOSIS PROJECT, LeanTaaS

LIFE SCIENCES
23andMe, color, CarbonX, BenevolentAI, verily, WuXiNextCODE, Zymergen, 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, 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 Electric, PREDIX, UPTAKE, SIGHT MACHINE, KONUX, OSIsoft, SCORTEX, 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, Flink, YARN, TEZ, MESOS, kubernetes, docker, CDAP, Red Hat, HELIX

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

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

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

STREAMING & MESSAGING
Spark, 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

AI / MACHINE LEARNING / DEEP LEARNING
TensorFlow, Keras, torch, MADlib, H2O, 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, 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, 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, Place IQ, esri, factual, CARTO, Mapillary, Streetline, cuebiq, Radar, OpenStreetMap

OTHER
DATA.GOV, IMAGENET, wiki-links, Labelme, Common Crawl, APPLOSCAPE, 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

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

mattturck.com/data2019

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Prerequisites

What need to be installed before Hands on

- Anaconda Jupyter Notebook
- Python v3.6 +

Or use **google collabs**

https://colab.research.google.com/

Get Sample Code

Download sample code from

https://github.com/galawana/tselworkshop1

(Download Zip and extract or git clone repository)

Demo and Hands On

Upload extracted folder into **drive.google.com**

Thank You