Data

DATA SCIENCE SKILLS

Data Types


Geographical
Transport
Cultural
Natural
Scientific
Types of Data
Metrological
Financial
Statistical

- Data Types
- Data Structure
- Data Set
- Data Base

Types of Data

- Classification of Data
  - Boolean
  - Numeric
    - Integer
    - Floating
  - Composite (derived from more than one type)
    - Array
    - Record

Algebraic functions

- >>> 2 + 35
- >>> 7 – 52
- >>> 2*(3+1)8
- >>> 2 + 35
- >>> 7 – 52
- >>> 2*(3+1)8
- >>> 5/22.5
- >>> 2 + 35
- >>> 7 – 52
- >>> 2*(3+1)8

Boolean functions

- Python can evaluate Boolean expressionsBoolean expressions evaluate toTrue or FalseBoolean expressions often involve comparison operators<, >, ==, !=, <=, and >=

Excercises

- Translate the following into Python algebraic or Boolean expressions and then evaluate them:
- The difference between Annie’s age (25) and Ellie’s (21)
- The total of $14.99, $27.95, and $19.83
- The area of a rectangle of length 20 and width 15
- 2 to the 10th power
- The minimum of 3, 1, 8, -2, 5, -3, and 0
- 3 equals 4-2
- The value of 17//5 is 3
- The value of 17%5 is 3
- 284 is even284 is even and 284 is divisible by 3
- 284 is even or 284 is divisible by 3

Excercises

- Translate the following into Python algebraic or Boolean expressions and then evaluate them:
- The difference between Annie’s age (25) and Ellie’s (21)
- The total of $14.99, $27.95, and $19.83
- The area of a rectangle of length 20 and width 15
- 2 to the 10th power
- The minimum of 3, 1, 8, -2, 5, -3, and 0
- 3 equals 4-2
- The value of 17//5 is 3
- The value of 17%5 is 3
- 284 is even284 is even and 284 is divisible by 3
- 284 is even or 284 is divisible by 3

Variables

- Just as in algebra, a value can be assignedto a variable, such as xWhen variable x appears inside an expression, it evaluates to its assigned value
- x = 3

Variables

- (Variable) names can contain these characters:a through z
- A through Z
- the underscore character _
- digits 0 through 9

Strings

In addition to number and Boolean values, Python support string values

'Hello, World!’

"Hello, World!“

A string value is represented as a sequence of characters enclosed within quotes

A string value can be assigned to a variable

String values can be manipulated using string operators and functions

Exercises

```
>>> s1 'good'>>> s2 'bad'>>> s3 'silly'
```

Write Python expressions involving strings s1, s2, and s3 that correspond to:

- 'll' appears in s3
- the blank space does not appear in s1
- the concatenation of s1, s2, and s3
- the blank space appears in the concatenation of s1, s2, and s3
- the concatenation of 10 copies of s3
- the total number of characters in the concatenation of s1, s2, and s3

- Data is any information you are collecting: numbers, statistics, measurements. It can also be words, observations, or other inputs.

	Continuous	Discrete		
	Quantitative data	Qualitative / Categorical / Attribute data		
Measurement	Units (example)	Ordinal (example)	Nominal (example)	Binary (example)
Time of day	Hours, minutes, seconds	1, 2, 3, etc.	N/A	a.m./p.m.
Date				
Cycle time				
Speed				
Brightness				
Temperature				
<Count data>				
Test scores				
Defects				
Defects				
Color				
Location				
Groups				
Anything				

Continuous Data

- has an infinite number of measurements depending on the resolution of the measurement system.
- There are no limits to the gaps between the measurements. It is data that can be expressed on an infinitely divisible scale.

- Examples:
  - Temperature
  - Height
  - … ?

Discrete Data

- Data types that have a finite number of measurements and are based on counts. Data that can be sorted into distinct, countable, and in completely separate categories. The count value can not be divided further on an infinite scale with meaning.

- Example:
  - How many people can comfortably fit into an airplane? It doesn't make sense to say 129.7632213 people. It is either 129 or 130, in this case you would round down to 129. Attribute and discrete do not mean exactly the same when describing data, discrete has more than two outcomes.
  - … ?

Nominal Data

- The **lowest level** of data classification. A numerical label that represents a qualitative description. These numbers are labels or assignments of numbers that represent a category or classification. This is also referred to a categorical data usually of more than two categories and is a form of discrete data and should apply nonparametric test to analyze. The number assignment does not reflect that one category is better or worse than another.
- Example:
  - Gender
    1 = Male
    2 = Female

Ordinal Data

- The next level higher of data classification than nominal data. Numerical data where number is assigned to represent a qualitative description similar to nominal data. These are measures by only the rank order.
- However, these numbers can be arranged to represent worst to best or vice-versa. Ordinal data is a form of discrete data and should apply non-parametric test to analyze.
- Example:
  - Classifying households as low income, middle-income, and high income
-

Interval Data

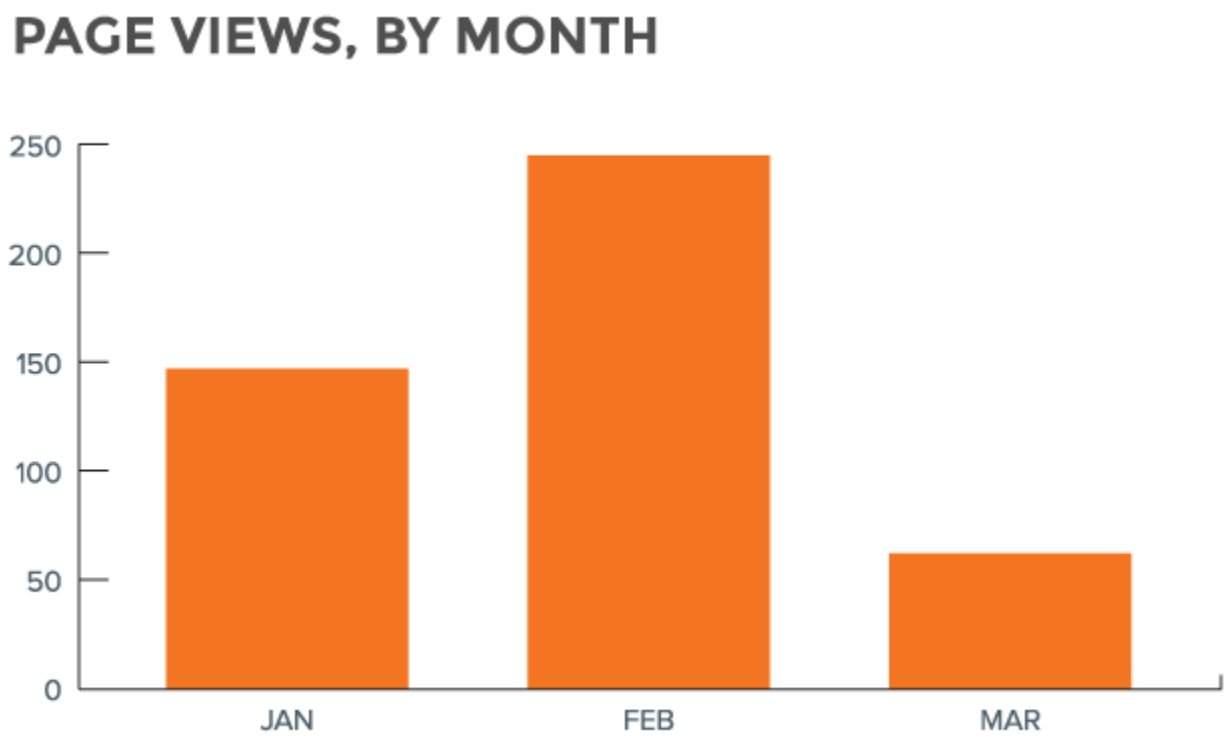
- The next higher level of data classification. Numerical data where the data can be arranged in a order and the differences between the values are meaningful but not necessarily a zero point. These are measures using equal intervals.
- Interval data can be both continuous and discrete. Zero degrees Fahrenheit does not mean it is the lowest point on the scale, it is just another point on the scale.
- Example:
  - Temperature
  - …
-

Exercise

	Continuous	Discrete		
	Quantitative data	Qualitative / Categorical / Attribute data		
Measurement	Units (example)	Ordinal (example)	Nominal (example)	Binary (example)
Time of day	Hours, minutes, seconds	1, 2, 3, etc.	N/A	a.m./p.m.
Date	Month, date, year	Jan., Feb., Mar., etc.	N/A	Before / After
Cycle time	Hours, minutes, seconds, month, date, year	10, 20, 30, etc.	N/A	Before / After
Speed	Miles per hour/centimeters per second	10, 20, 30, etc.	N/A	Fast / Slow
Brightness	Lumens	Light, medium, dark	N/A	On / Off
Temperature	Degrees C or F	10, 20, 30, etc.	N/A	Hot / Cold
<Count data>	Number of things	10, 20, 30, etc.	N/A	Large / Small
Test scores	Percent, number correct	F, D, C, B, A	N/A	Pass / Fail
Defects	N/A	Number of cracks	N/A	Good / Bad
Defects	N/A	N/A	Oversized, missing	Good / Bad
Color	N/A	N/A	Red, blue, green	N/A
Location	N/A	N/A	East, West, South	Domestic / International
Groups	N/A	N/A	HR, Legal, IT	Exempt / Non-exempt
Anything	Percent	10, 20, 30, etc.	N/A	Above / Below

- **Cross-Sectional:** The sample of elements is measured only once. This shows you a snapshot of variables at a point in time (e.g., market survey).

- **Longitudinal:** The data sample is measured repeatedly over time (e.g., stock prices, monthly sales data).



What Makes a Data Set?

- A data set is comprised of variables; each individual data point—the thing that is measured or counted—is a **variable**. Each variable can be examined on its own or in relation to other variables to reveal insights, including:
- Mean:

What Makes a Data Set?

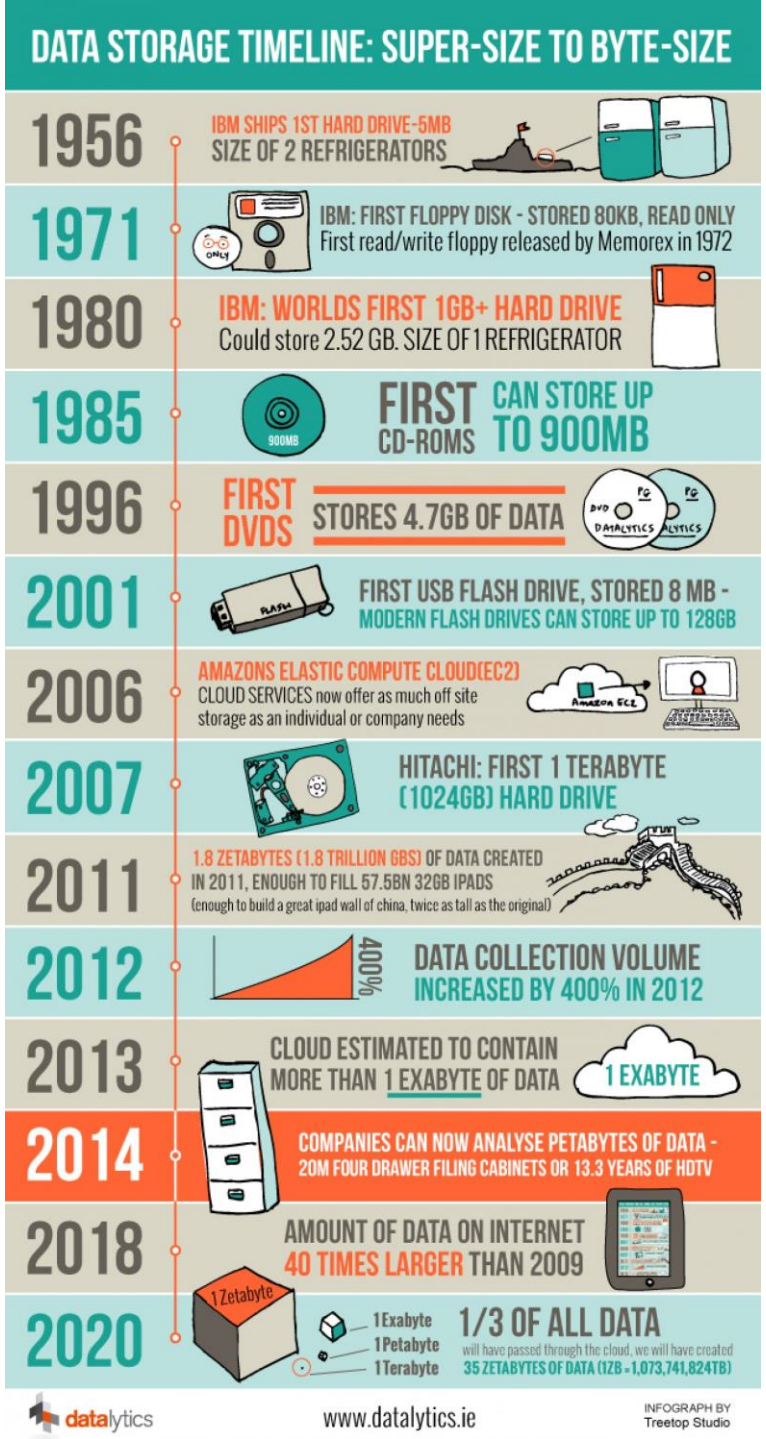
- Range:

What Makes a Data Set?

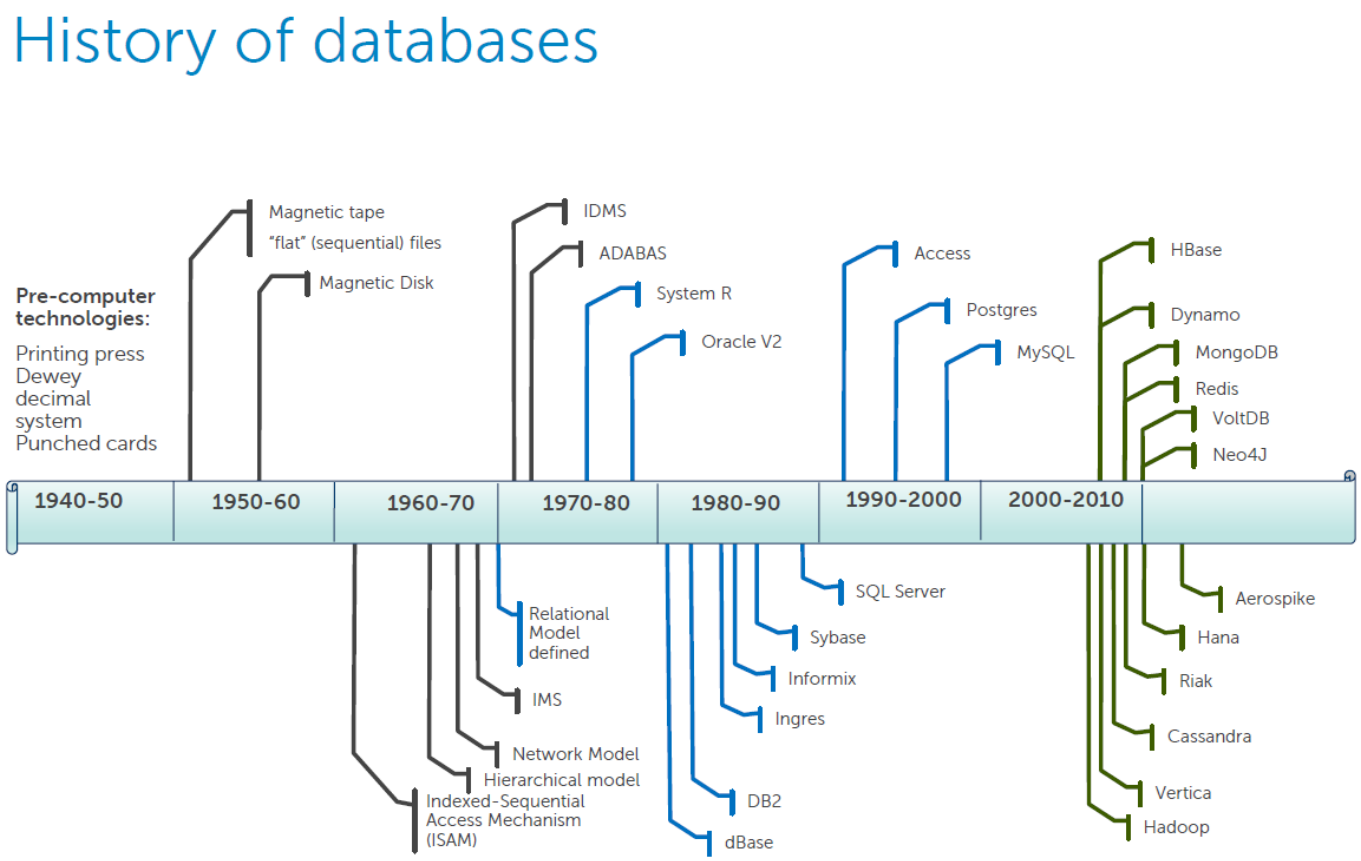
- Range:

Data Storage

- Files
  - CSV
  - XML
  - JSON
- Place
  - HDD
  - Memory
  - Locally
  - Cloud
  - Database



History of databases



BIG DATA LANDSCAPE 2017

INFRASTRUCTURE

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

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

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

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

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

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

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

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

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

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

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

MGMT / MONITORING
amazon web services, New Relic, APPDYNAMICS, actifio, Google Cloud Platform, WAVEFRONT, DATADOG, unravel, splunk>, LOOGE, Anodot, rocana, Numerify, pagerduty

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

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

APP DEV
Lightbend, Apache IO, rainforest, CASK

CROWDSOURCING
amazon mechanical turk, Upwork, WorkFusion, CrowdFlower

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

ANALYTICS

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

DATA SCIENCE PLATFORMS
IBM, KNIME, data iku, DOMINO, yhat, rapidminer, CONTINUUM ANALYTICS, Alpine DATA LABS, ALGORITHMIA, Angoss

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

VISUALIZATION
tableau, SAP, Google Cloud Platform, Qlik, MAPD, celonis, Periscope DATA, ZEPL, CHARTIO, plotly, GOODDATA

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

STATISTICAL COMPUTING
sas, SPSS, MATLAB

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

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

HORIZONTAL AI
IBM Watson, Cortana, Face++ 旷视, sentient technologies, Voyager Labs, vicarious, clarifai, :) Affectiva, SENSETIME, cognitiveScale, Numenta, chronocam, PETUUM, SI, nara, OSARO, THE CURIOUS AI COMPANY, BLUE VISION

SPEECH & NLP
Google Cloud Platform, twilio, amazon alexa, NarrativeScience, WolframAlpha, semantic machines, Microsoft, MindMeld, NUANCE, ARRIA, iDIBON, TalkIQ, cortical.io, maluuba, snips, yseop, Gridspace, SoundHound Inc.

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

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

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

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

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

APPLICATIONS – ENTERPRISE

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

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

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

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

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

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

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

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

BACK OFFICE AUTOMATION
HyperScience, captricity, AppZen

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

APPLICATIONS – INDUSTRY

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

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

GOVERNMENT
Socrata, OPENGOV, mark43, FN FiscalNote, OpenDataSoft

FINANCE - LENDING
OnDeck, Affirm, Kreditech, AVANT, finance, INSIKT, TALA, MoneyLion, TrueAccord, trooly, cignifi, aire, Active.AI

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

REAL ESTATE
Opendoor, VTS, CREDIFI, reonomy, COMPSTAK

INSURANCE
Metromile, Lemonade, CYENCE, Shift Technology, Tractable

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

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

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

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

COMMERCE
instacart, STITCH FIX, RetailNext, HowGood

OTHER
eHarmony, stem, rethink robotics, AMPER MUSIC, hoppin', SIGHT MACHINE, BOXEVER, celect, VERDIGRIS, duetto, Unbabel, Jukedeck, Second Spectrum

OPEN SOURCE

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

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

DATA ACCESS
nifi, mongoDB, cassandra, sqoop, CAFFO, SciDB, CouchDB, OPENTSDB, riak, HBASE, Spanner, ACCUMULO

COORDINATION
talend, Apache Zookeeper, OOZIE, Apache Ambari

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

STAT TOOLS
python, R, ScalaLab, NumPy, SciPy

AI / MACHINE LEARNING / DEEP LEARNING
theano, torch, MADlib, learn, TensorFlow, Caffe, Microsoft CNTK, DMTK, Keras, PaddlePaddle, Apache SINGA, OpenAI, mxnet, WEKA, neon, FeatureFu, Chainer, VELES, DIMSUM, DSSTNE, mllib, DL4J DEEPLEARNING4J, Intel nervana, Aerosolve

SEARCH
elasticsearch, Solr, Lucene

LOG ANALYSIS
elasticsearch, kibana, logstash

VISUALIZATION
BEAKER, Rodeo

COLLABORATION
jupyter, Zeppelin, ANACONDA

SECURITY
Apache Ranger, KNOX, Sentry

DATA SOURCES & APIS

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

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

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

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

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

LOCATION INTELLIGENCE
FOURSQUARE, MapAnything, Sense 360, PlaceIQ, esri, factual, CARTO, Mapillary, STREETLINE

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

DATA RESOURCES

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

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

V2 – Last updated 5/3/2017

mattturck.com/bigdata2017

FIRSTMARK EARLY STAGE VENTURE CAPITAL