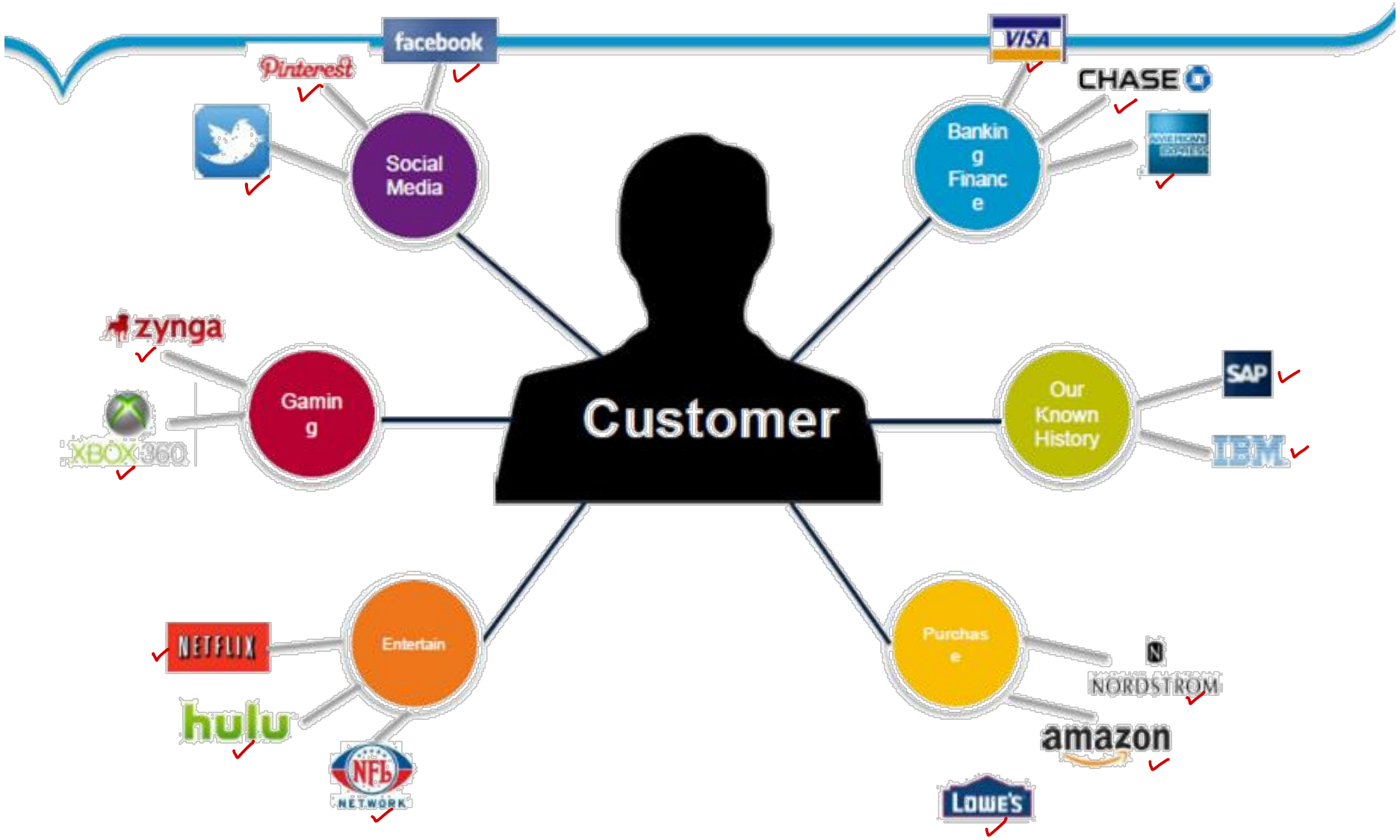
Big Data Analytics - Overview

Trainer: Mr. Nilesh Ghule.

Source of Big Data



History of Big Data

RDBMS
(1970)

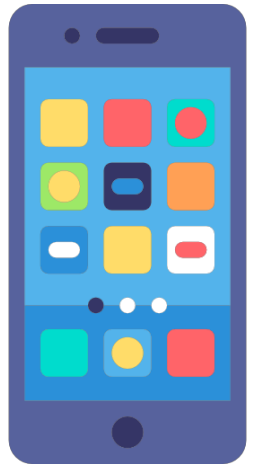
Internet
(1991)

Applets
(1995)

NoSQL
(1998)

2004 #nosql

IoT
(2008)



Phones
(2011)

History of Big Data

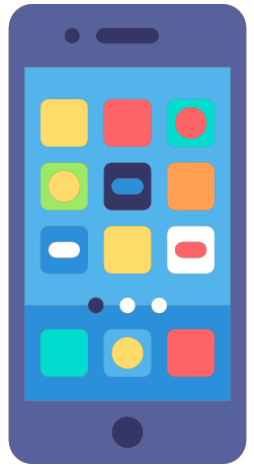
RDBMS
(1970)

Internet
(1991)

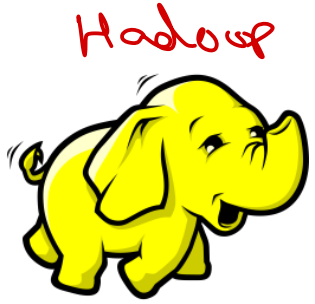
Applets
(1995)

NoSQL
(1998)

IoT
(2008)



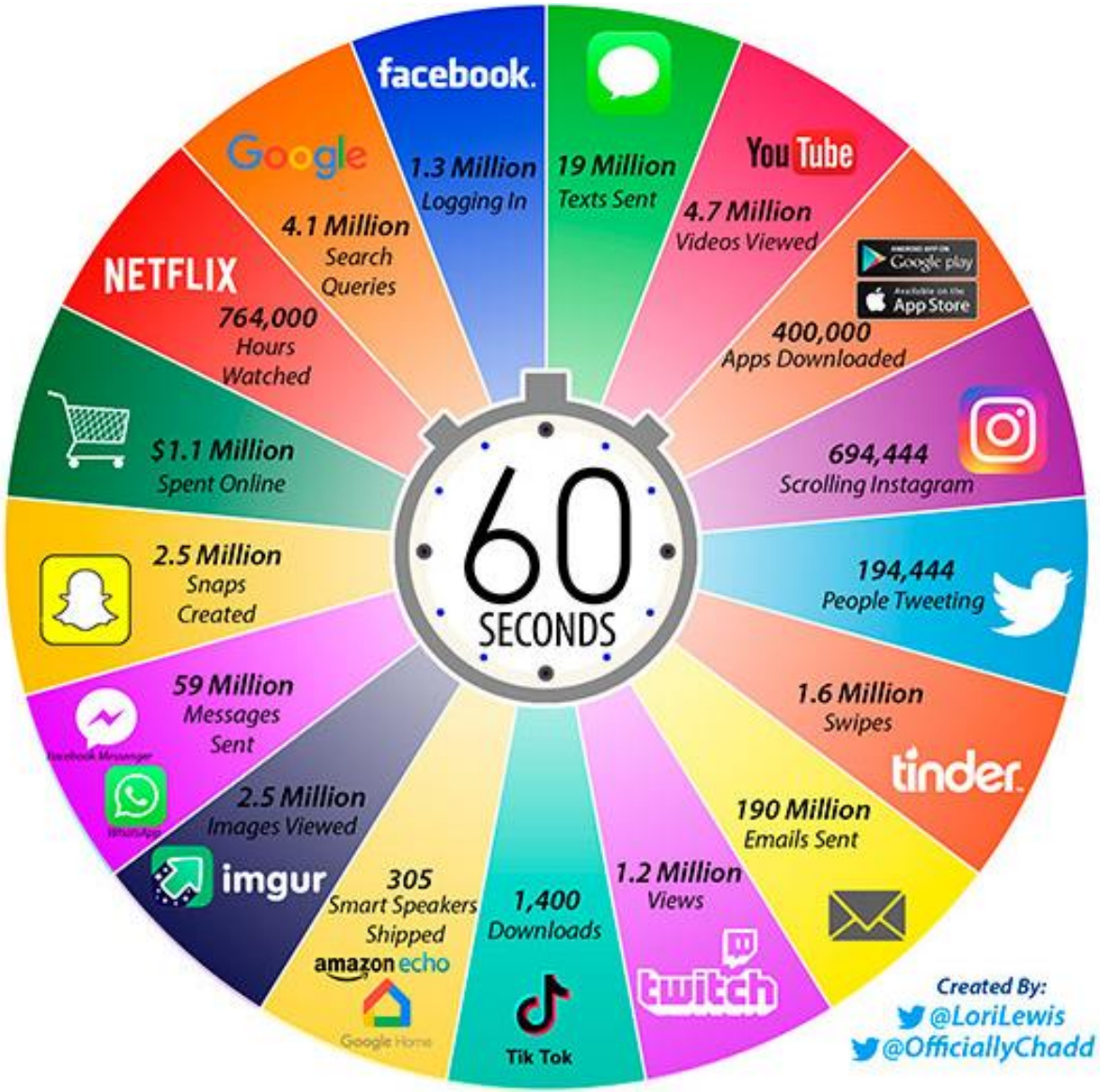
Phones
(2011)



...

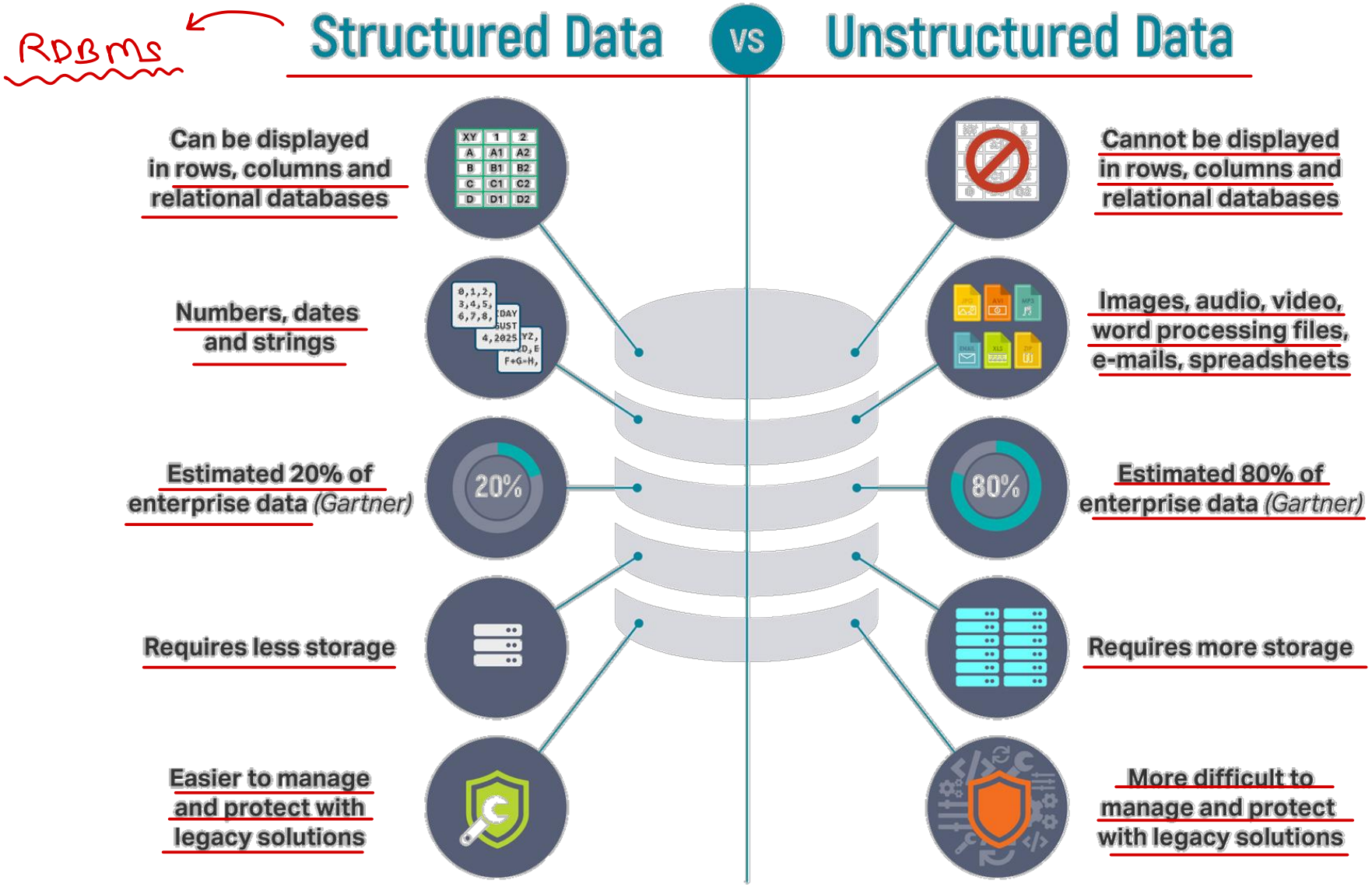
One internet minute

One internet minute

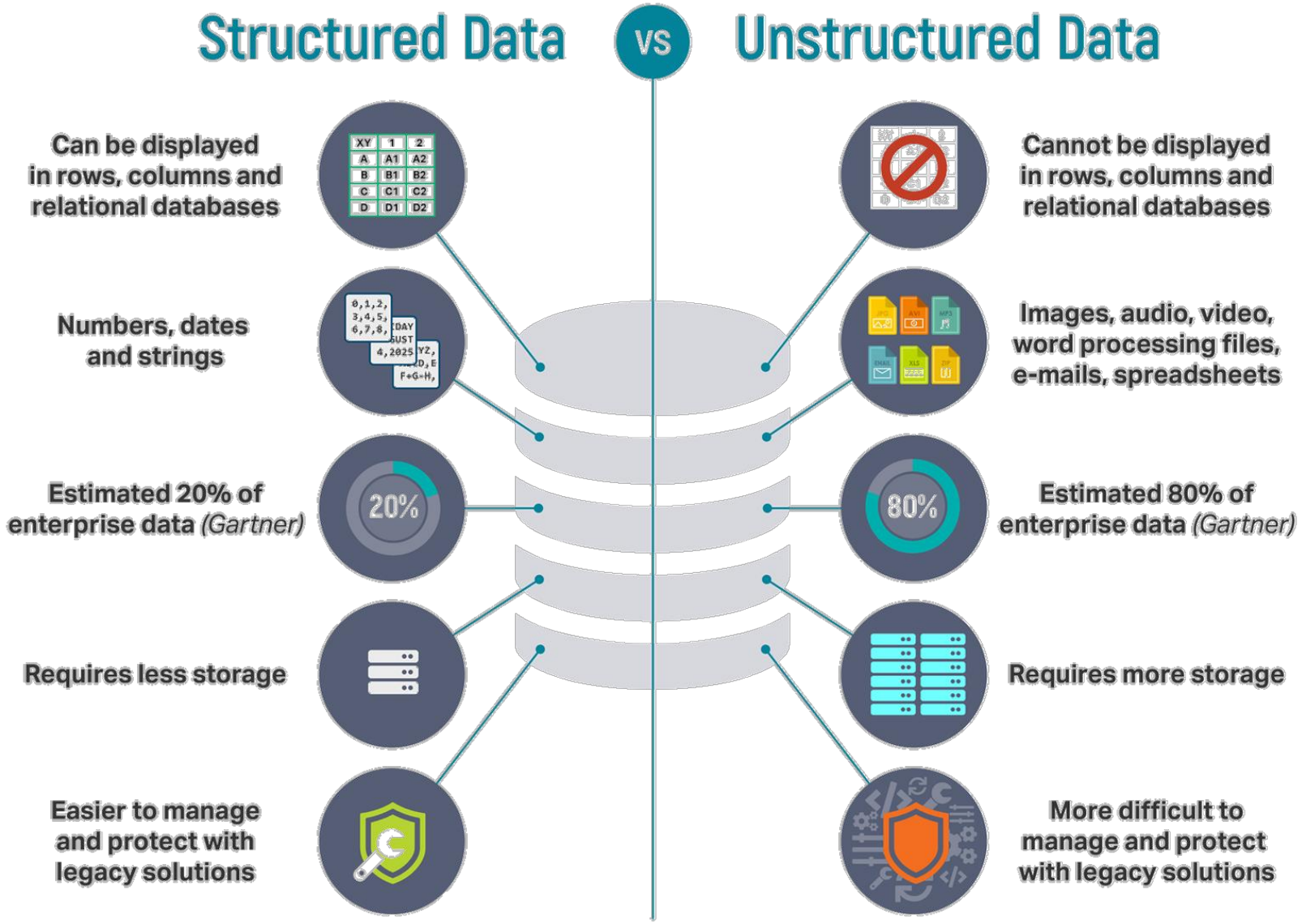


- KB - 2^{10}
- MB - 2^{20}
- GB - 2^{30}
- TB - 2^{40}
- PB - 2^{50}
- XB - 2^{60}
- ZB - 2^{70}
- YB - 2^{80}
- BB - 2^{90}
- … - 2^{100}

Structured vs Unstructured data

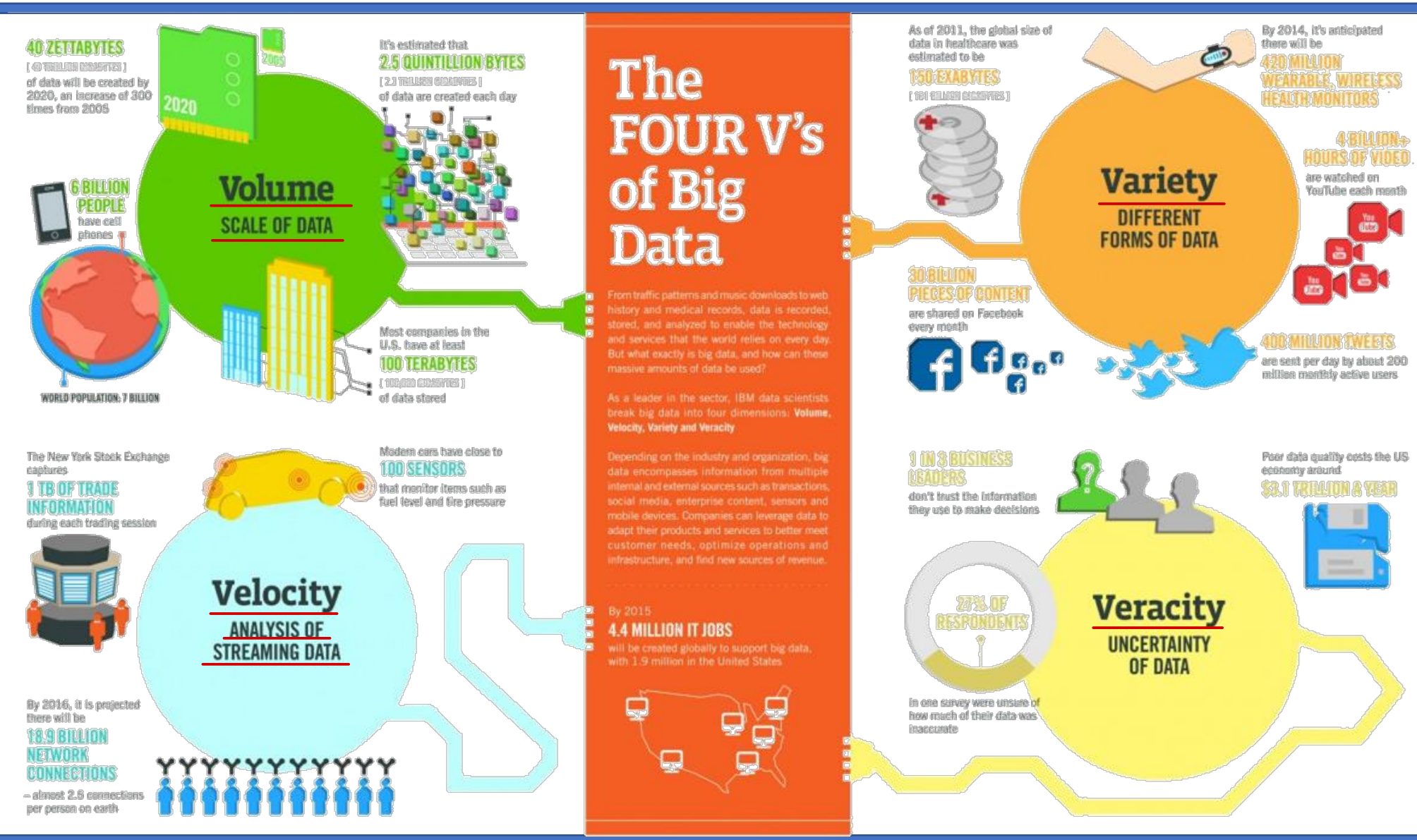


Structured vs Unstructured data vs Semi Structured data



- Semi-structured data
  - Flexible structured data
  - Better storage utilization
  - Usually hierarchical struct
  - Or Key-value format
  - Examples:
    - XML
    - JSON

Big Data characteristics



value

Big Data Spread

Selective Ignorance

- Want to develop a software for NGO to organize Blood donation camp. Which information of donors to be collected?

1. medical condition
2. blood group
3. age
4. phone
5. email
8. last donation date
12. name
9. salary
10. occupation
11. religion
6. gender
7. address
13. qualification

Selective Ignorance

- Blood Donors data to collect?

- Name
- Blood group
- Medical history
- Last blood donation date
- Gender
- Email
- Mobile
- Address

- Blood Donors data to collect?

- Qualification
- Salary
- Occupation
- Religion
- Caste

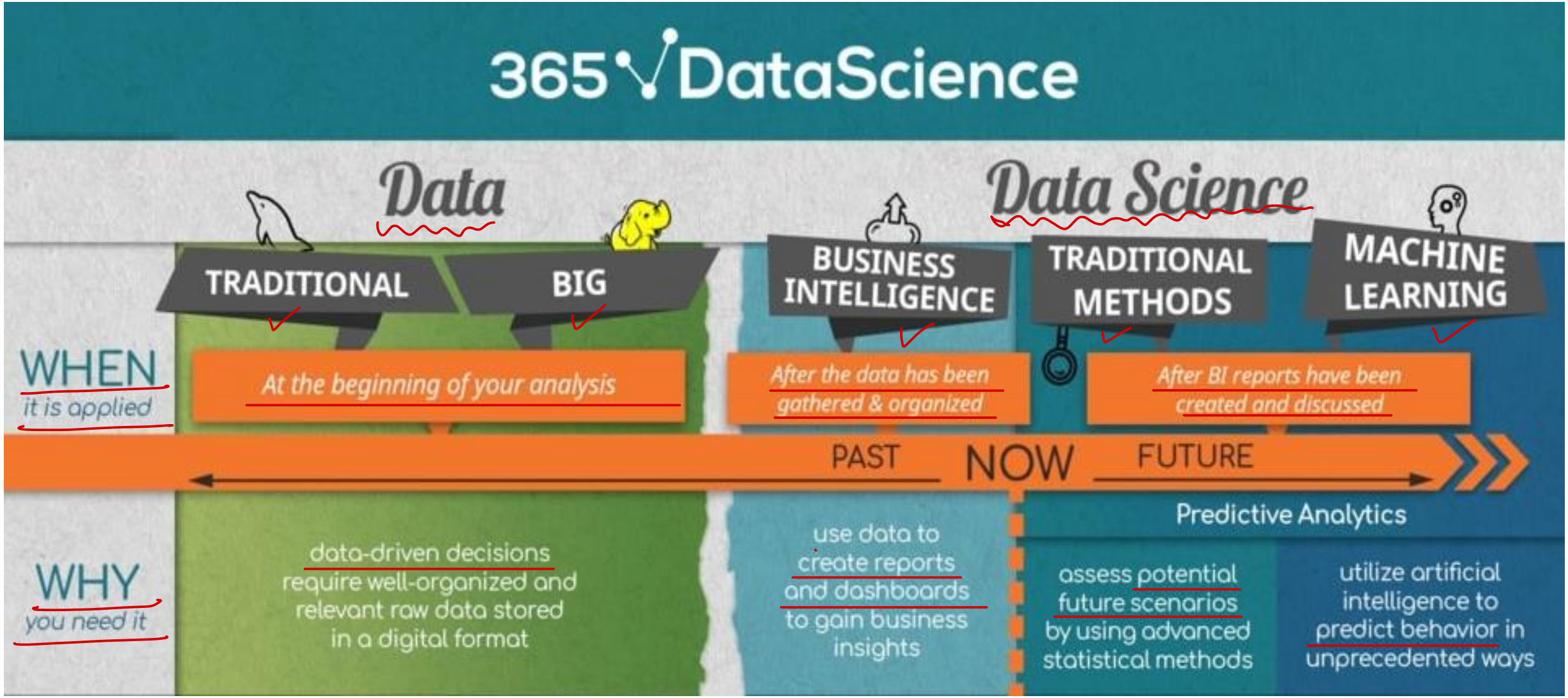
Big Data & IoT

- Home Automation
  - ON/OFF appliance
  - Report state of appliance periodically.
- Huge Storage
  - High velocity, High volume & High variety.
- Process data
  - Analyze Time/Day of max/min usage.
  - Decide state of appliance - minimize electricity.
  - Invent new products, new marketing schemes.

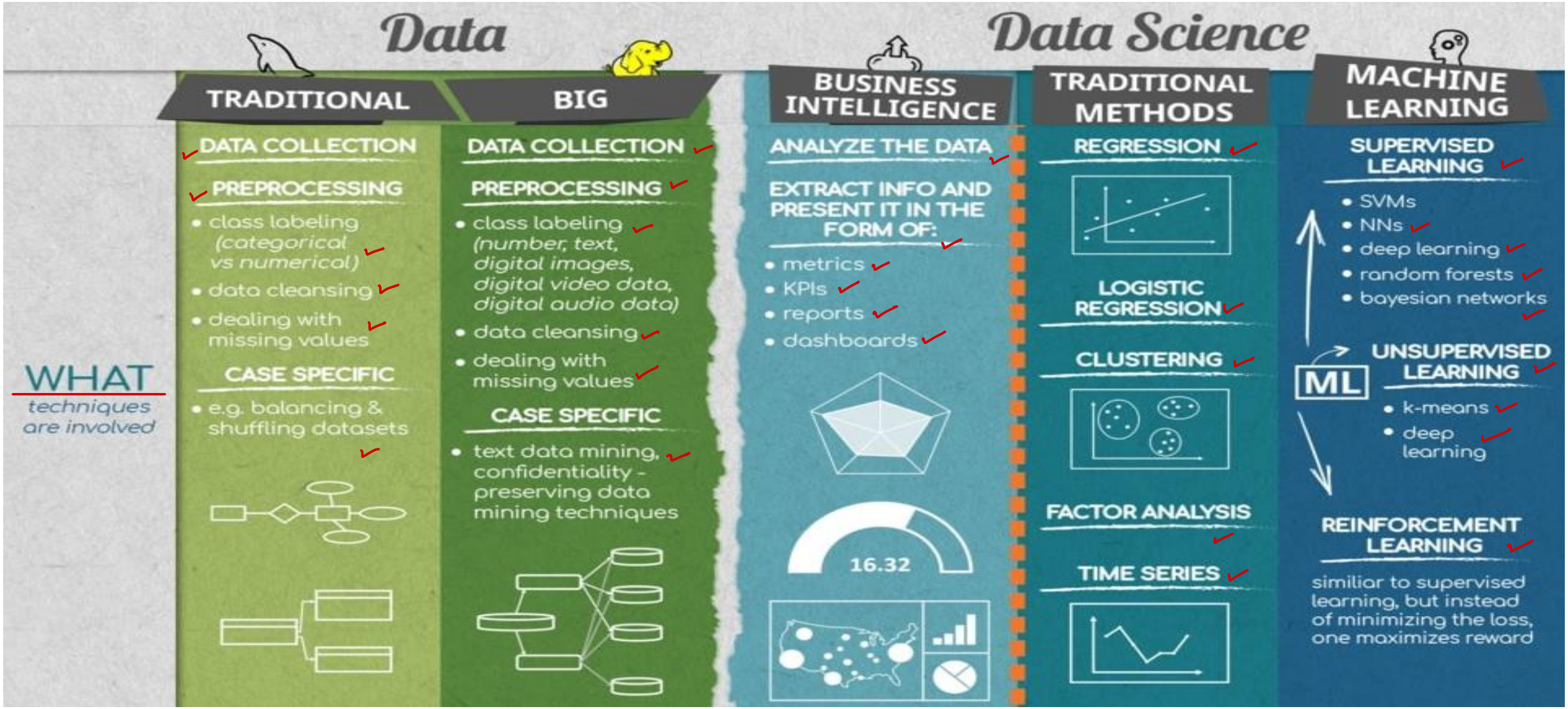
Big Data & Covid-19

- Covid tracker apps
  - GPS & Bluetooth data from apps
  - Tracking patients
  - Contact tracing
  - Tracking quarantined people
- Covid testing devices
  - Digital thermometer
  - Swab testing devices
  - Pathology testing
- Covid treatment
  - Treatment details
  - Patient progress
- Covid Vaccine trials
  - Simulating drugs impact
  - Tracking effects & side-effects

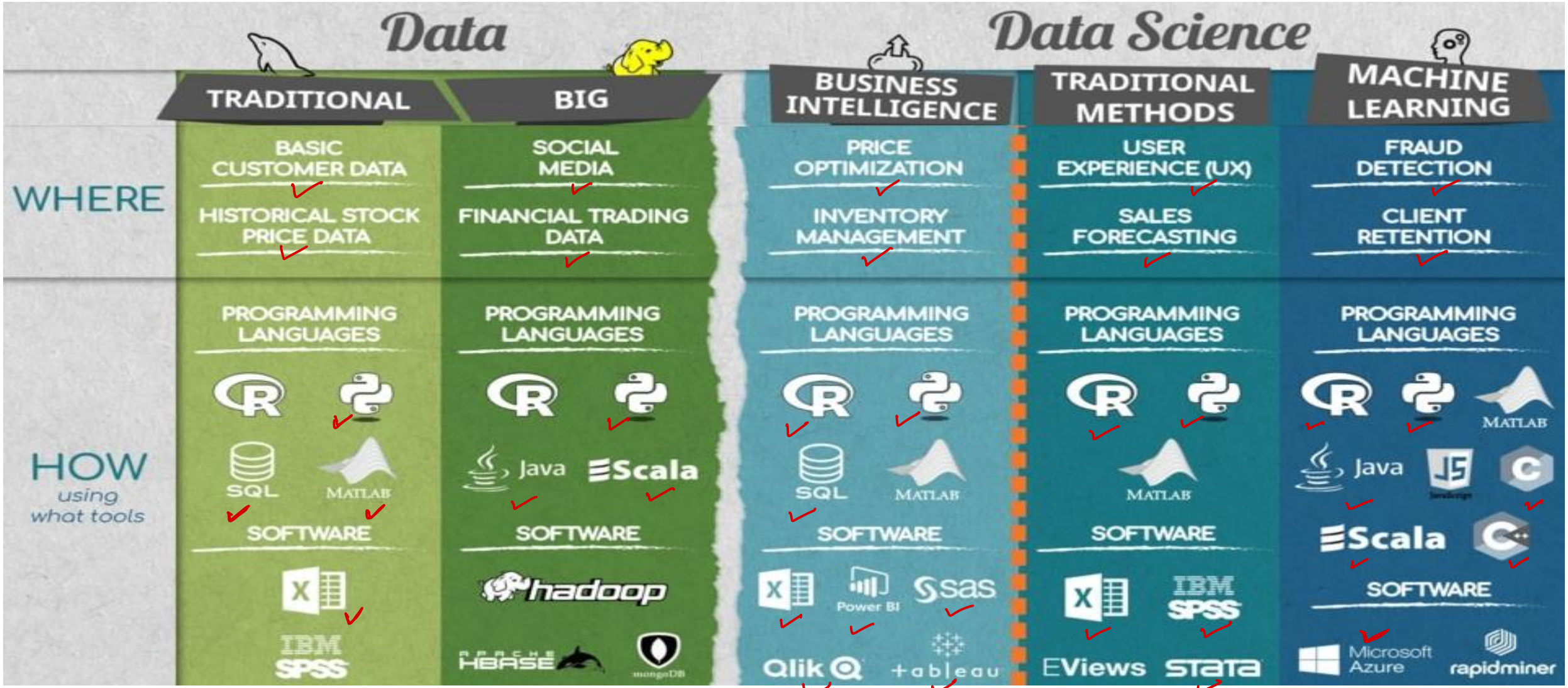
Jargon of Big Data Analytics & Data science



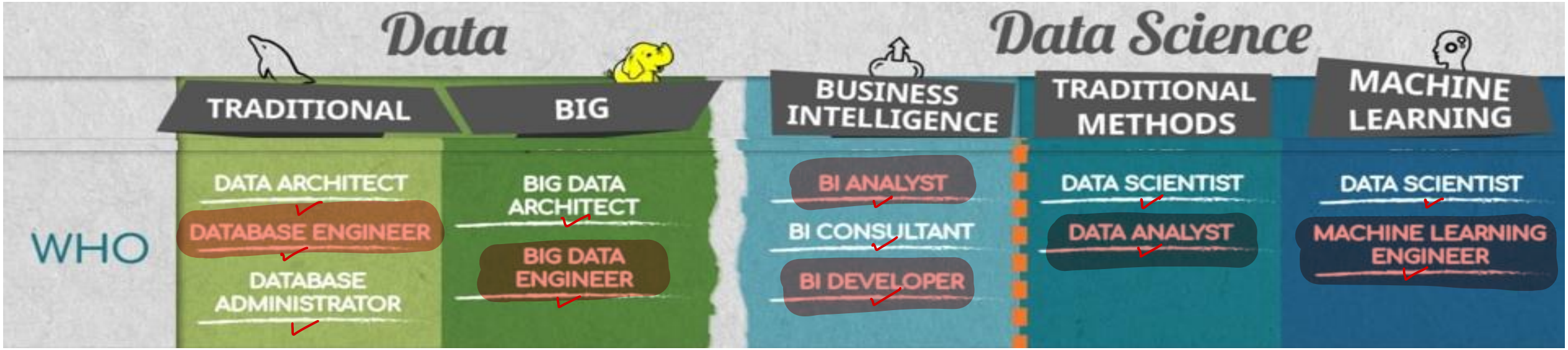
Jargon of Big Data Analytics & Data science



Jargon of Big Data Analytics & Data science



Jargon of Big Data Analytics & Data science



- In 2018, Big data analytics market was 169 billion dollar worth.
- By 2022, more than 274 billion dollar industry.
- Apart from technologies mentioned above few more terms are relevant
  - Artificial intelligence
  - Cloud computing
  - Parallel computing

Jargon of Big Data Analytics & Data science

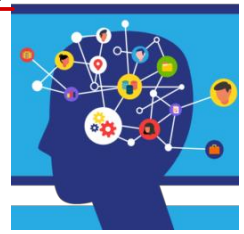
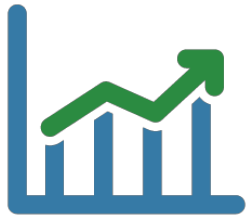
- Data analysis vs Data analytics
- Business analysis vs Business analytics
- Data analysis vs Data visualization
- Data analysis vs Data mining
- Artificial intelligence vs Machine learning vs Deep learning
- Data engineering vs Data science

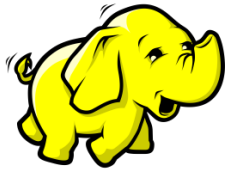
Big Data & Analytics Spectrum

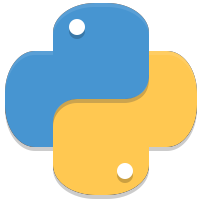
- Data storage
  - RDBMS & NoSQL databases
  - Data warehouse
  - S3, DFS, …
- Data Analysis & visualizations
  - Data Visualizations
  - Business reports
- Artificial Intelligence, Data Science & Data mining
  - Mathematics, Statistics & Computer algorithms
  - Machine learning & Deep learning
  - R Programming, Python
- Data Engineering
  - Hadoop, Hive, Spark, Kafka, BigTable, …
  - Parallel processing
  - Java, Scala, Python.
- Infrastructure
  - Linux, Cloud Computing









Big Data domains & opportunities

- Domains: Health-care, Retails, Trading/Share market, Finance, Security, Fraud, Search engines, Log Analysis, Telecom, Traffic Control, Manufacturing and lot more.
- Big Data is all about :- Think, Collect, Manage, Analyze, Summarize, Visualize, Discover Knowledge and Take Decisions.
- Job profiles:
  - Business Analyst/Intelligence
  - Database engineer / DWH
  - Big Data engineer
  - IT operations
  - AI/ML engineer
  - Data Scientist
  - Big Data Architect






- The sexiest job in the 21st century require a mixture of multidisciplinary abilities and suitable candidates must be prepared to learn and develop constantly.

-Ronald Van Loon

BIG DATA LANDSCAPE, VERSION 3.0

Exited: Acquisition or IPO

Infrastructure

NoSQL Databases: redis, FOUNDATIONDB, DATASTAX, mongoDB, Couchbase, basho, AEROSPIKE, HYPERTABLE, scality, CLOUDANT, OhmData, Neo4j, sones

Hadoop On Prem: HADAPT, cloudera, splice machine, Zettaset, MAPR, amazon, Microsoft, Hortonworks, Pivotal

Cloud: MORTAR, infochimps, bole, JETHRO DATA, altiscale, amazon

NewSQL Databases: MarkLogic, TRANSLATTICE, Rain, paradigm4, memsql, deepdb, citusdata, SkySQL, nuodb, Clustrix, VoltDB, SQLFire

Cluster Services: LexisNexis, HPCC Systems, mesosphere, Acunu

MPP Databases: TERADATA, ParStream, InfiniDB, kognitio, VERTICA, NETEZZA, SQL Server, Pivotal, PARACCEL

Management / Monitoring: GUTEFY, TRIGHT, metafor, New Relic., StackIQ, tidemark, appnomic, AppFirst, oceansync, boundary, DATADOG

Graph Databases: Neo4j, GIRAPH, aster data, InfiniteGraph

Data Transformation: TRIFACTA, Paxata, DataTamer, KALIDO, revelytix, SYNCSORT

Security: DATAGUISE, codefortytwo, Stormpath, IMPERVA

Storage: Cleversafe, panasas, nimblestorage, Coraid

Crowd-sourcing: microtask, CrowdFlower, servio, amazon mechanicalturk, mobileworks

App Dev.: CONCURRENT, CONTINUITY, wibidata

Analytics

Analytics Platforms: databricks, QuantCell, PERVASIVE, guavus, Datameer, KARMASPHERE, collective[i], PRECOG, dataspora

For Business Analysts: STAT WING, CIRRO, TREPAREL, OrigamiLogic, ClearStory, DataGravity

Data Science Platforms /Tools: domino, nutonian, Alpine, Sense, CONTINUUM, MORTAR, plotly, yhat, MODE

BI Platforms: birst, JASPERSOFT, bime, pentaho, GoodData, SiSense, platfora

Unstructured Data: BASIS, ATTIVIO, GENERAL SENTIMENT, semantria, crimson hexagon, DIGITAL REASONING, ai-one, Quid, Palantir

Data Visualization: +ableau, ZoomData, visual.ly, Roambi, Chart.io, libratio, qunb, Quantum4D, ACTUATE, Kitenga, looker, Ayasdi, ISS, DataHero, GECKOBOARD

Machine Learning: SKYTREE, bigml, VOTANIMIE ANALYTICS, wise.io, context relevant

IBM WATSON reactor labs AI, vicarious

Social Analytics: simple reach, GNIP, bitly, synthesio, Dataminr, bluefin, Statilliser, DATASIFT, tracx, bottlenose

Location/ People/ Events: RADIUS, Fliptop, LOGATE, Locu, Placed

Analytics Services: BIG DATA, THINK BIG, VALIANCE, DATA SCIENCE, accenture, OPERA, uO

Big Data Search: hp, LucidWorks, ontology

Statistical Computing: Prior Knowledge, REVOLUTION, SAS, MATLAB, SPSS

Log Analytics: splunk>, LOUD PHYSICS, loggly, sumologic, Kibana

Crowd-Sourced: kaggle, DataKind

Real-Time: NETANARKETS, amato, causata

SMB: RJMetrics, retention, GoSquared, sumall, custora

Applications

Ad Optimization: aggregate knowledge, rocketfuel, TAPAD, ai Match, MediaMath, thetradedesk, 33across, exelate, DataXu, dstillery, m6d

Publisher Tools: Chartbeat, Yieldex, yieldbot

Marketing: LATTICE ENGINES, Sailthru, spinnakr, gainsight, Kontera, RelateIQ, Telkpart, persado, bloomreach, CLICKFOX, Pursway

Finance: Lenddo, BILLGUARD, cignifi, wongai, zestfinance, LendUp, KENSHO, OnDeck

Human Capital: evolv, entelo, gild.

Legal: JUDICATA, Lex Machina, RAVEL

Government / Regulation: mark43, enigma, Socrata, FiscalNote, FREE STOP, PREDPOL

Security: SONIFY, siftscience, FORTSCALE, feedzai

Education /Learning: KNEWTON, declara, PANORAMA, Clever

Health: Recombine, 23andMe, Ginger.io, µBiome, FLATIRON, Counsyl

Industries: tubular, NEXT BIG SOUND, OPOWER, SIGHT MACHINE, THE CLIMATE CORPORATION, Fire

Cross Infrastructure / Analytics

SAP, sas, IBM, Google, Microsoft, vmware, amazon, 1010data, talend, TERADATA, hp, NetApp

Open Source

Framework: Spark, Hadoop YARN, HDFS

Query / Data Flow: Pig

Data Access: Cassandra, SciDB, ORACLE, HBASE, CouchDB, mongoDB, riak, Sqoop

Coordination / Workflow: ZooKeeper, talend, Oozie

Real-Time: Storm

Stat Tools: SciPy, R, {P}

Machine Learning: MLlib, Mahout

Cloud Deploy

Search: elasticsearch, Solr, LUCENE.net

Data Sources

Data Mkts: Windows Azure Marketplace, bluekai, knoema, DataMarket, factual

Data Sources: DATA.GOV, premise, YODLEE, xignite, VALIDIC, plaid, quandl, STANDARD TREASURY, human/api

Sensor Data: kinsa, SKYCATCH, STREETLINE, fitbit, RunKeeper, JAWBONE, LumaSense, Withings, BASIS, estimote

Incubators & Schools: zipfian, GA, INSIGHT, DataElite

Thank you!

Nilesh Ghule <nilesh@sunbeaminfo.com>