What’s the Big Deal About Big Data?

Natasha Balac, Ph.D.

Jun, 2016

Background

- Over 25 Years of Experience in Data Mining
- Ph.D. in Machine Learning – with emphasis on Big Data and Mobile Robots
- Director of Predictive Analytics center of Excellence at the Supercomputer Center at UCSD
- Lecturer
  - UCSD MAS in Data Science and Engineering
  - UCSD Extension Data Mining Certificate
  - Coursera Big Data Specialization

University of California, San Diego UCSD

Student-centered, research-focused, service-oriented public institution

Recognized as one of the top 15 research universities worldwide

Culture of collaboration sparks discoveries that advance society and drive economic impact

UC San Diego's rich academic portfolio includes six undergraduate colleges, five academic divisions and five graduate and professional schools

CalIT2 – Qualcomm Institute

> “Calit2 represents a new mechanism to address large-scale societal issues by bringing together multidisciplinary teams of the best minds.” *Larry Smarr, Director, Calit2*

Spitzer Space Telescope (Infrared)

Hubble Space Telescope (Optical)

Calit2 is taking ideas beyond theory into practice, accelerating innovation and shortening the time to product development and job creation. Where the university traditionally has focused on education and research, Calit2 extends that focus to include development and deployment of prototype infrastructure for testing new solutions in a real-world context.

Data Insight Discovery

- Founded in January 2014 – San Diego, CA
  - Women Owned
- Service Offerings Include
  - Predictive Analytics Services
  - Business Intelligence and Analytics
  - CRM system development
  - Condition Based Maintenance
  - Systems and IoT Integration Services
  - ETL, RDBMS, No-SQL data store development
  - Big Data Technologies
- **Key Strengths include**
  - Numerous successfully deployed Predictive Analytics Projects
  - Over 25 years of experience and expertise

data.
insight.
discovery.

Data Insight Discovery Areas of Expertise

Marketing

- Customer Segmentation
- Campaign Analysis
- Market Mix Modelling
- Dashboards and Reporting
- Promotion Effectiveness
- Price Optimization

Sales

- Sales forecasts
- Incentive Optimization
- Propensity Modelling
- Consumer Insights
- Branding Solutions

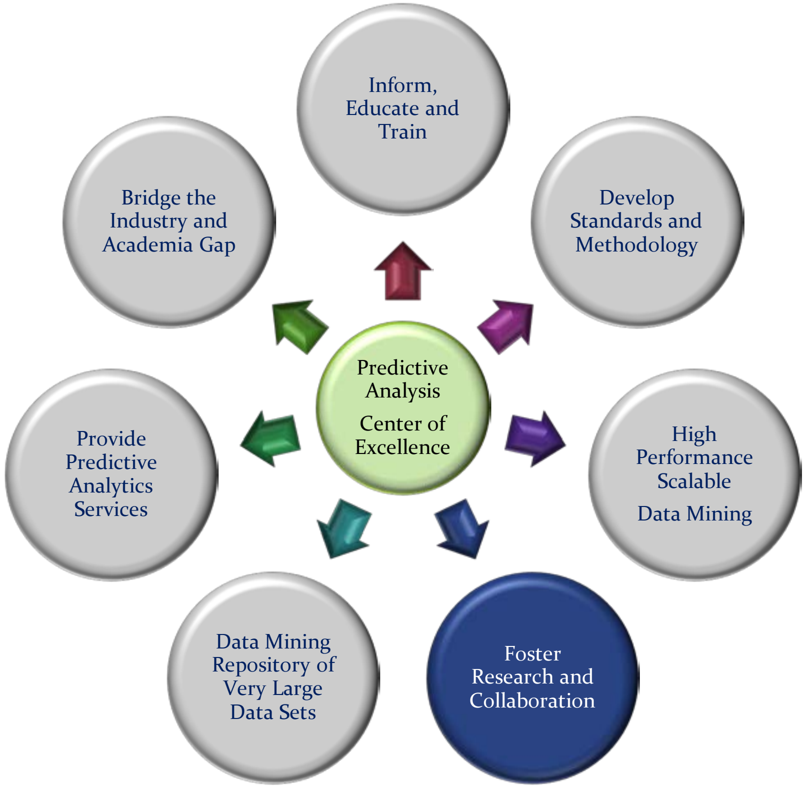
Digital

- Social and Web Analytics
- Sentiment Analysis
- Cross media interaction
- Digital Media ROI
- Anomaly Detection

Condition Based Maintenance

- Sensor Analytics
- Predictive Maintenance
- Intelligent workflow Solution
- Risk modeling and estimation

PACE – Predictive Analytics Center of Excellence
Closing the gap between Government, Industry and Academia



Predictive Analysis Center of Excellence

Inform, Educate and Train

Develop Standards and Methodology

High Performance Scalable Data Mining

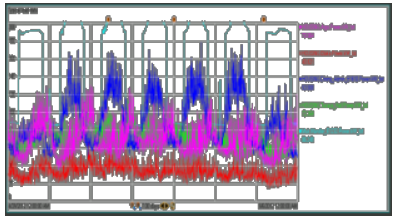
Foster Research and Collaboration

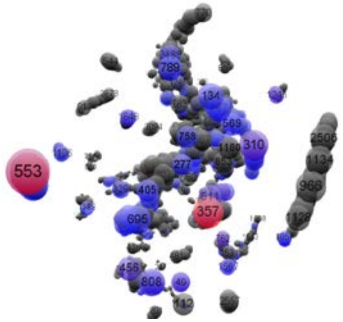
Data Mining Repository of Very Large Data Sets

Provide Predictive Analytics Services

Bridge the Industry and Academia Gap

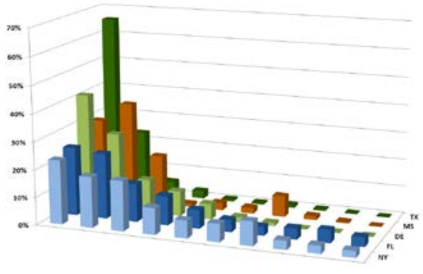
- Fraud Detection
- Modeling behaviors
- Biomedical Informatics
- Smart Grid Analytics
- Solar powered system modeling
- Microgrid anomaly detection
- Battery Storage Analytics
- Sport Analytics
- Genomics
- Population Health
- Nano-engineering

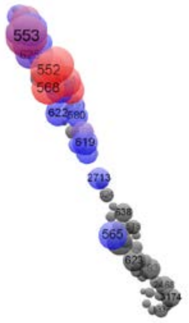


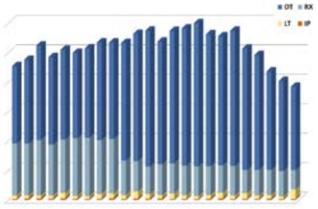


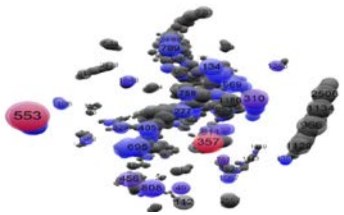
CMS Fraud, Waste and Abuse Detection and Prediction

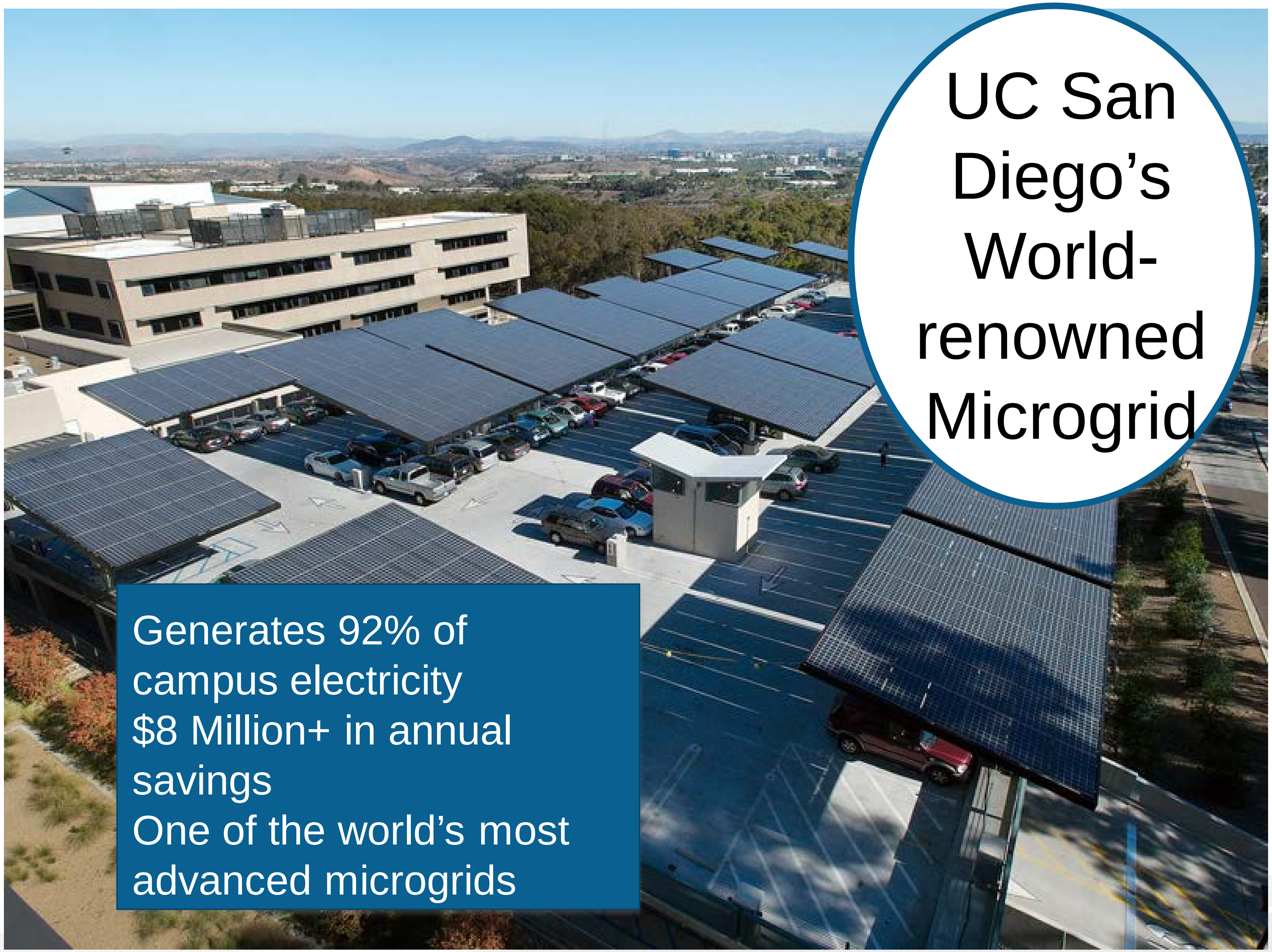
- Descriptive Statistics
  - Claims summary information
  - History and trends
  - Distributions across periods, transactions, etc.
- Exploratory Analysis
  - Profiles of provider transactions
  - Provider similarity according to profiles
  - Visual summaries of large amounts of data
  - Eligibility data link to provider billing
- Predictive analytics
  - Adjustments
  - Equipment, Service Codes
  - Long term vs. short term hospital stay
  - Provider profiles











UC San Diego’s World-renowned Microgrid

Generates 92% of campus electricity
$8 Million+ in annual savings
One of the world’s most advanced microgrids

UCSD Smart Grid

- **UCSD Smart Grid sensor network data set**
  - **45MW peak micro grid; daily population of over 54,000 people**
- Smart Grid data – over 100,000 measurements/sec
  - **Sensor and environmental/weather data**
    - Large amount of complex data streaming from sensor networks
  - **Predictive Analytics throughout the Microgrid**

Collaboration between public, private, and academic organizations working to develop and implement initiatives that will improve the San Diego region’s energy independence, empower consumers to embrace clean technologies, reduce greenhouse gas emissions, and drive economic growth

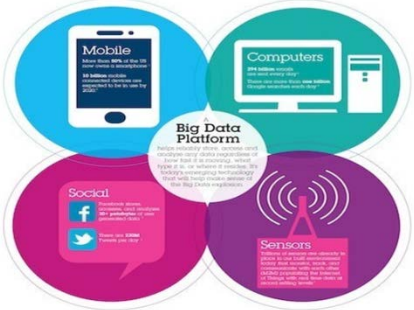
SmartCity
SAN DIEGO

White House Big Data Event: “Data to Knowledge to Action” – Launch Partners Award

What is “Big Data”?

- Big data

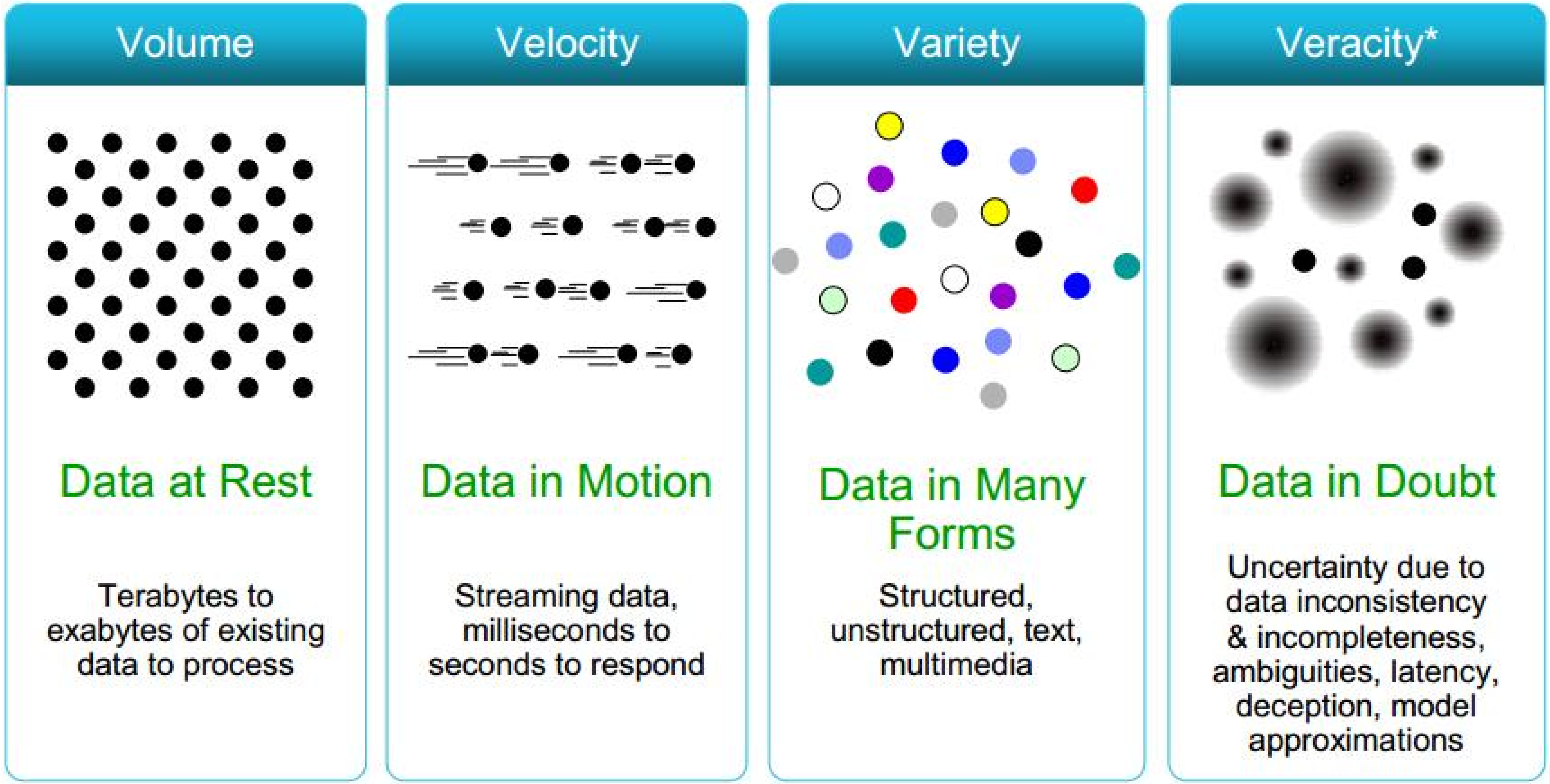
  “Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions”



- Big data is in the eye of the beholder

  Data that exceeds the ability to handle with current resources

4 V’s of Big Data



Volume

Data at Rest

Terabytes to exabytes of existing data to process

Velocity

Data in Motion

Streaming data, milliseconds to seconds to respond

Variety

Data in Many Forms

Structured, unstructured, text, multimedia

Veracity*

Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

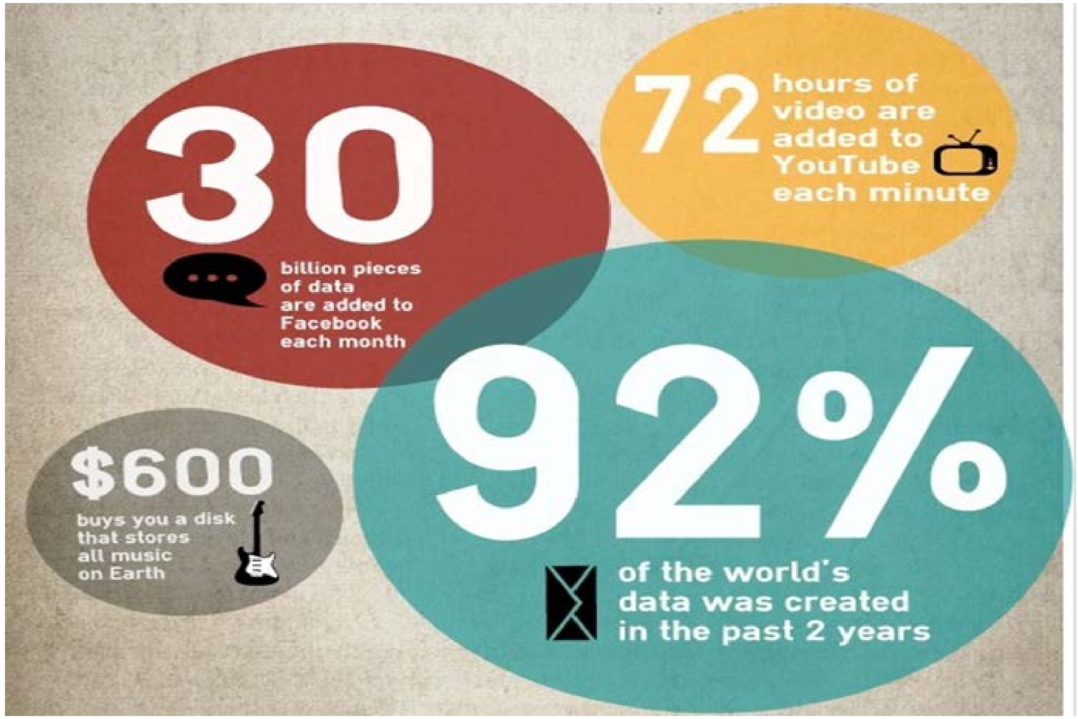
IBM, 2012

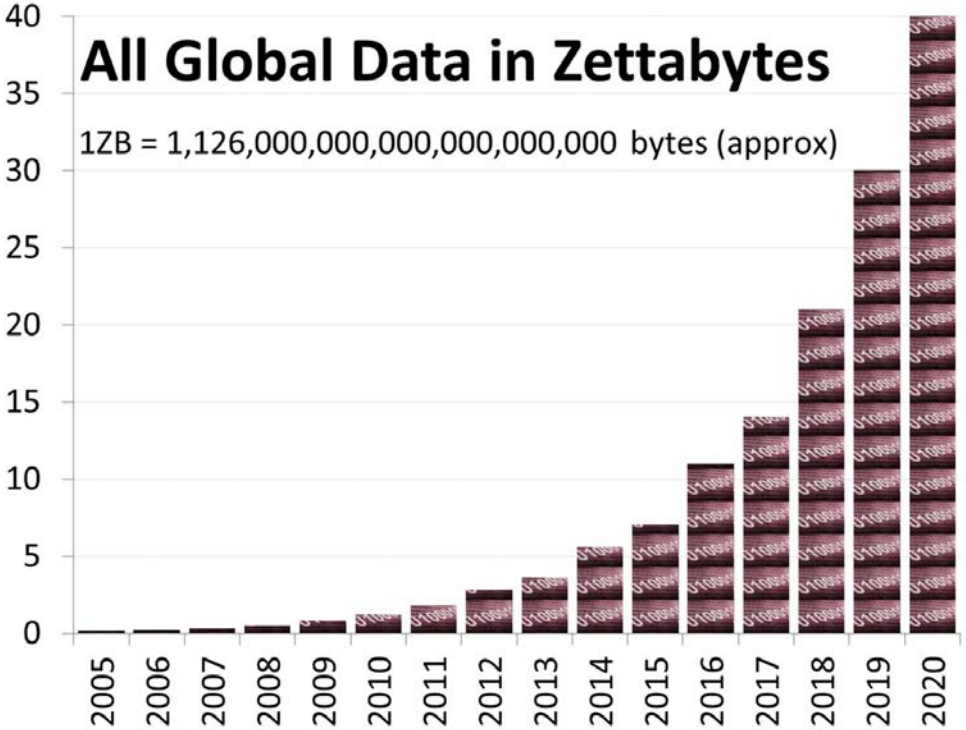
Big Data Definition

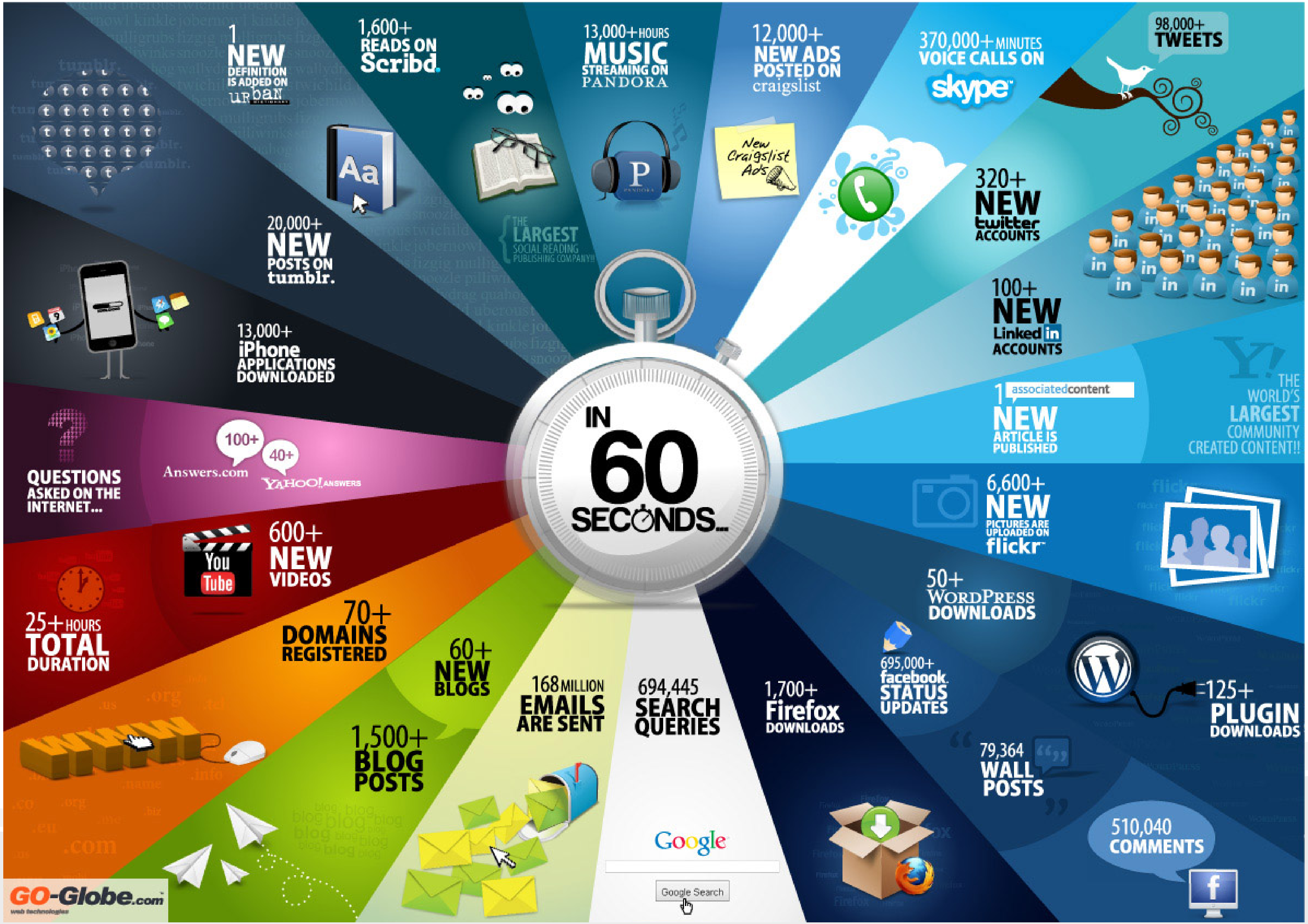
- *Wikipedia*: an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications

- *Oxford English Dictionary:* data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges

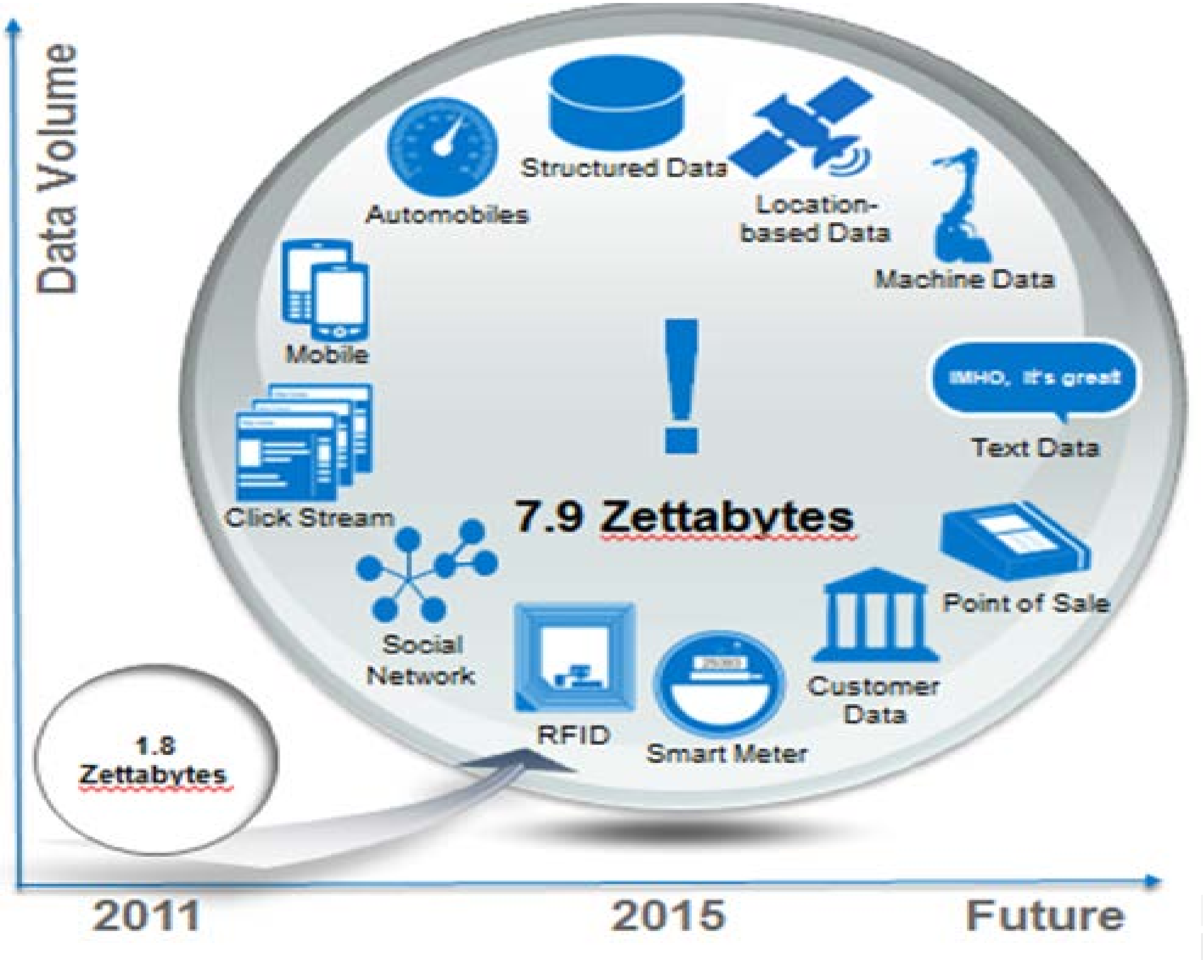
How Big is Big Data?

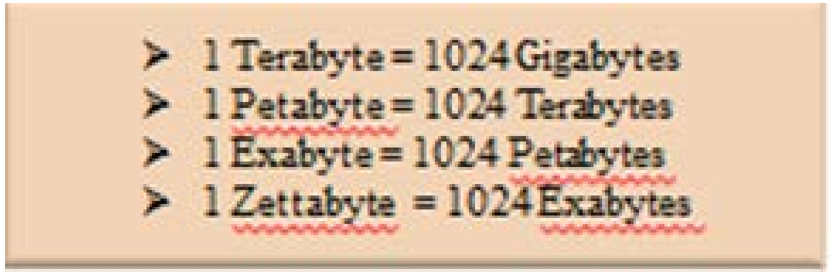

30
billion pieces of data are added to Facebook each month
72 hours of video are added to YouTube each minute
$600 buys you a disk that stores all music on Earth
92% of the world's data was created in the past 2 years




IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban dictionary
1,600+ READS ON Scribd.
THE LARGEST SOCIAL READING PUBLISHING COMPANY!!
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
New Craigslist Ads
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW LinkedIn ACCOUNTS
1 NEW associatedcontent ARTICLE IS PUBLISHED
Y! THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WORDPRESS DOWNLOADS
125+ PLUGIN DOWNLOADS
695,000+ facebook STATUS UPDATES
79,364 WALL POSTS
510,040 COMMENTS
1,700+ Firefox DOWNLOADS
694,445 SEARCH QUERIES
Google
Google Search
168 MILLION EMAILS ARE SENT
60+ NEW BLOGS
1,500+ BLOG POSTS
70+ DOMAINS REGISTERED
.com .org .us .info .name .biz .eu .co
600+ NEW VIDEOS
You Tube
25+ HOURS TOTAL DURATION
100+ Answers.com
40+ YAHOO! ANSWERS
QUESTIONS ASKED ON THE INTERNET...
13,000+ iPhone APPLICATIONS DOWNLOADED
20,000+ NEW POSTS ON tumblr.
Aa
P PANDORA
GO-Globe.com web technologies

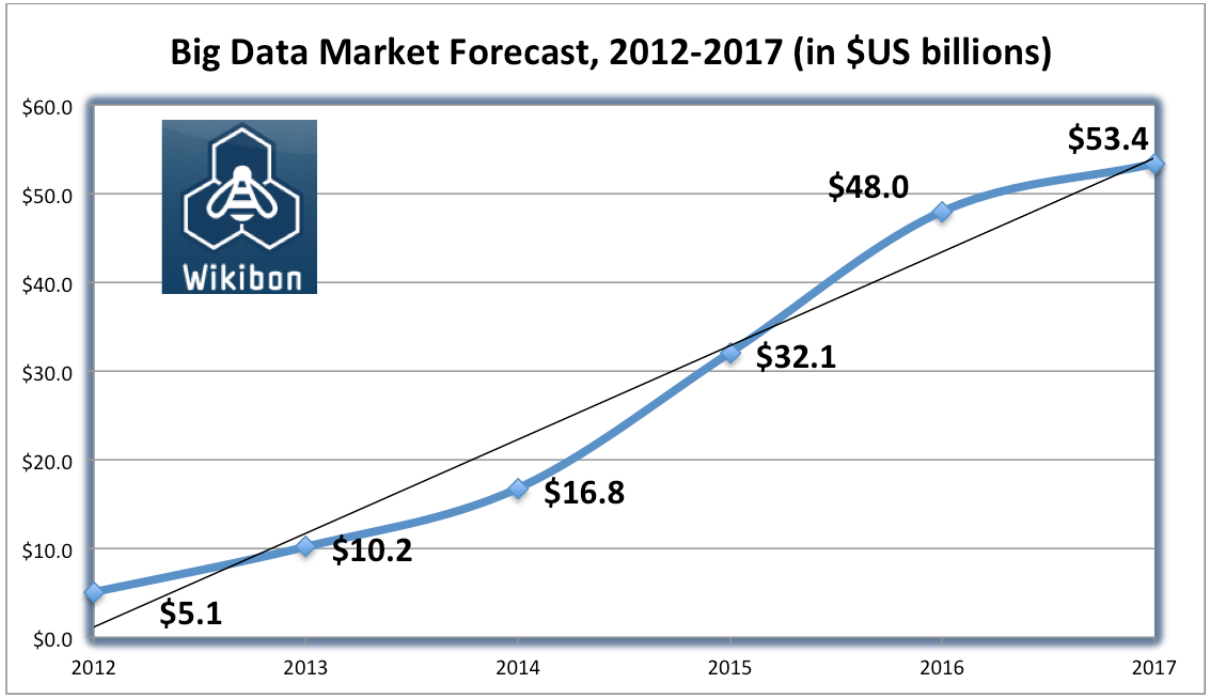
Big Data Growth





- 1 Terabyte = 1024 Gigabytes
- 1 Petabyte = 1024 Terabytes
- 1 Exabyte = 1024 Petabytes
- 1 Zettabyte = 1024 Exabytes

Tip of the Iceberg

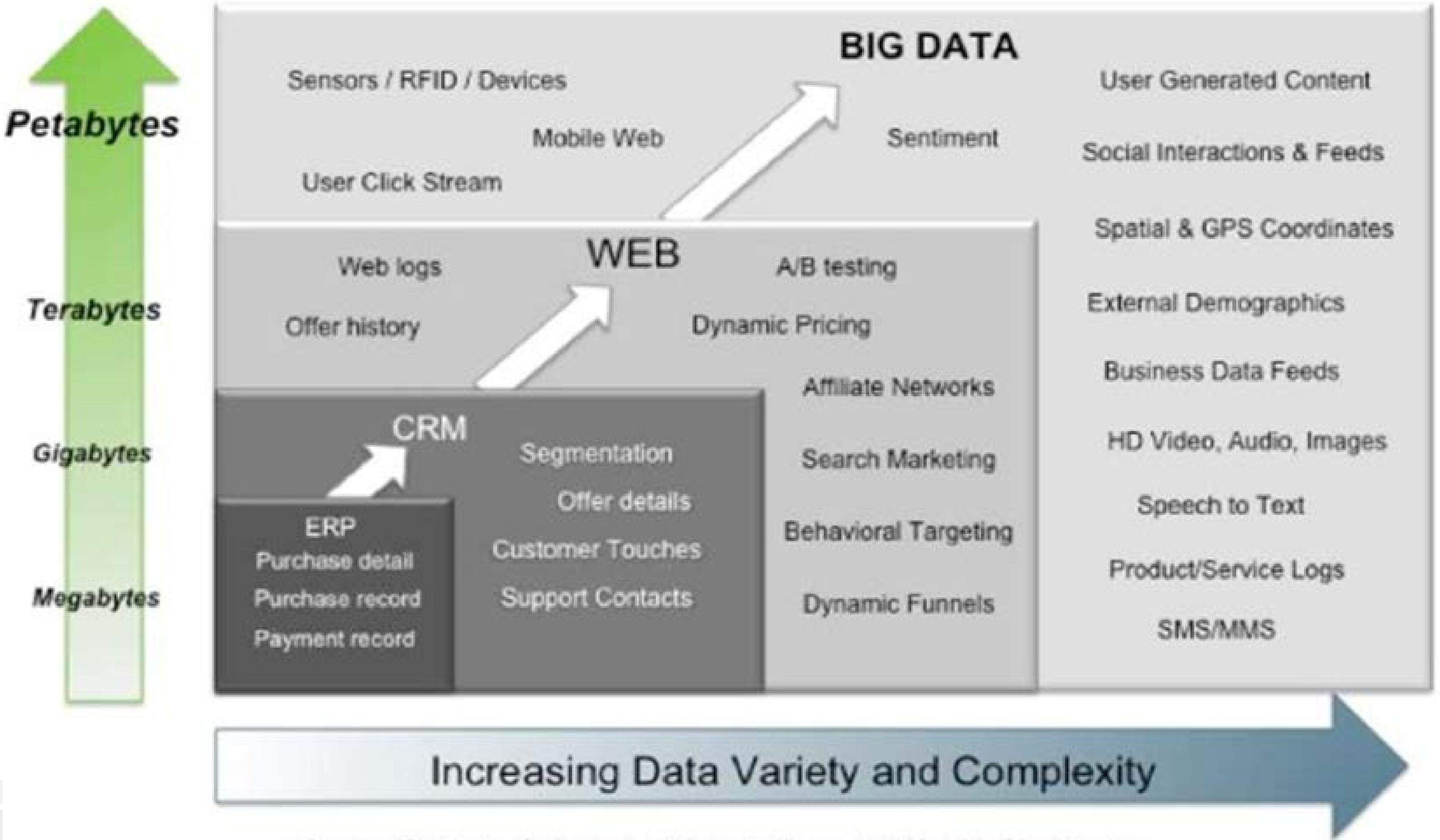


What to do with big data?

ERIC SALL's list

- Big Data Exploration
  - To get an overall understanding of what is there
- 360 degree view of the customer
  - Combine both internally available and external information to gain a deeper understanding of the customer
- Monitoring Cyber-security and fraud in real time
- Operational Analysis
  - Leveraging machine generated data to improve business effectiveness
- Data Warehouse Augmentation
  - Enhancing warehouse solution with new information models and architecture

Transactions, Interactions & Observations



Petabytes

Terabytes

Gigabytes

Megabytes

BIG DATA

Sensors / RFID / Devices

Mobile Web

User Click Stream

Sentiment

User Generated Content

Social Interactions & Feeds

Spatial & GPS Coordinates

External Demographics

Business Data Feeds

HD Video, Audio, Images

Speech to Text

Product/Service Logs

SMS/MMS

WEB

Web logs

Offer history

A/B testing

Dynamic Pricing

Affiliate Networks

Search Marketing

Behavioral Targeting

Dynamic Funnels

CRM

Segmentation

Offer details

Customer Touches

Support Contacts

ERP

Purchase detail

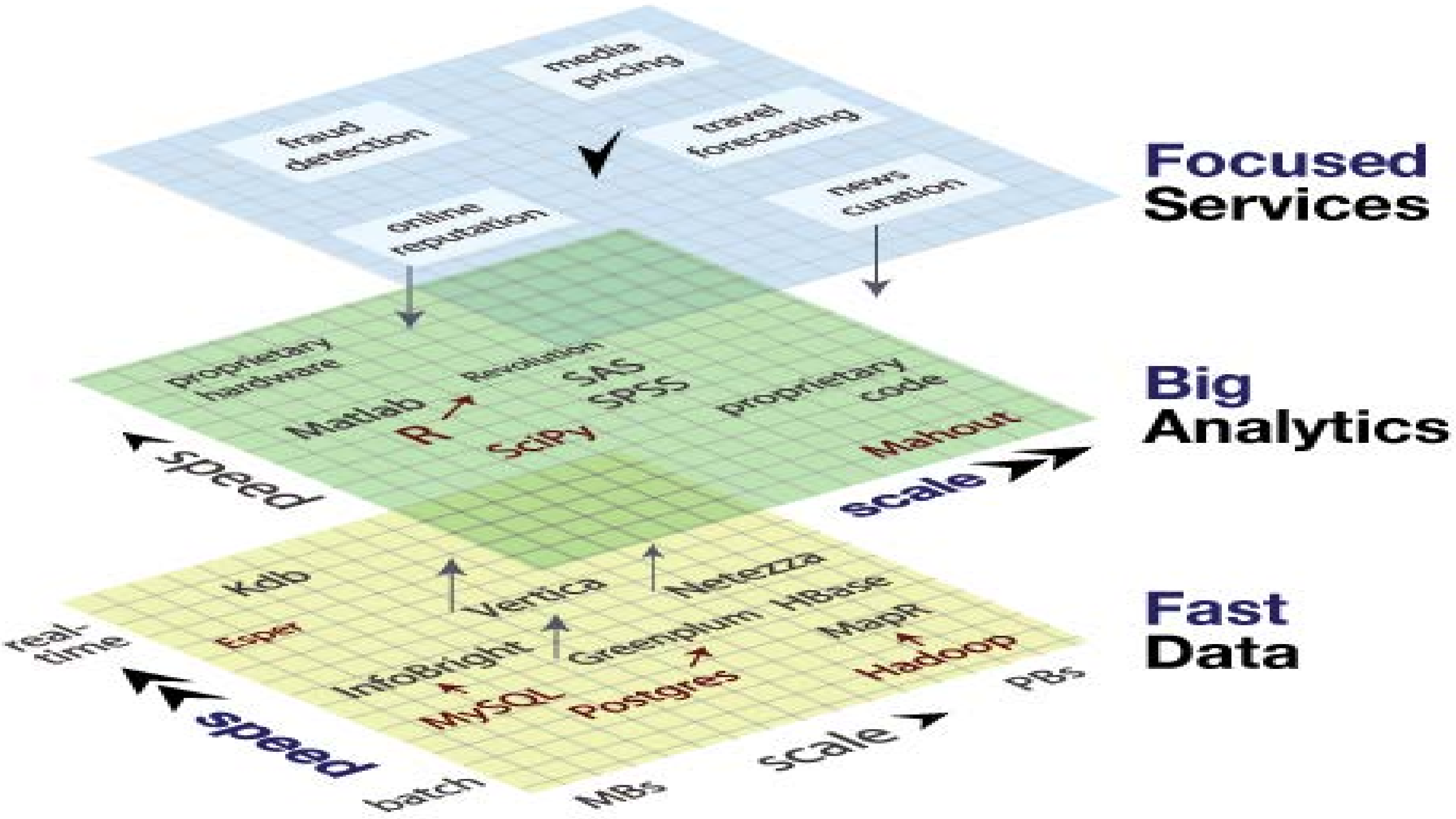
Purchase record

Payment record

Increasing Data Variety and Complexity

Source: *Contents of above graphic created in partnership with Teradata, Inc.*

The Emerging Big Data Stack



Big Data Landscape 2016 (Version 2.0)

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, CAZENA, TREASURE DATA, altiscale, Qubole

Spark
databricks, GridGain, TACHYON NEXUS

Cluster Services
amazon web services, HPCC SYSTEMS, kubernetes, docker, MESOSPHERE, CoreOS, pepperdata, StackIQ

NoSQL Databases
amazon DynamoDB, Google Cloud Platform, ORACLE, Microsoft Azure, MarkLogic, mongoDB, DATASTAX, AEROSPIKE, Couchbase, SequoiaDB, redislabs, influxdata

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

Graph Databases
neo4j, APACHE GIRAPH, OrientDB, InfiniteGraph

MPP Databases
TERADATA, VERTICA An HP Company, NETEZZA an IBM Company, Actian, kognitio, EXASOL, dremio

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

Data Transformation
alteryx, talend, TRIFACTA, tamr, Paxata, StreamSets, Alation

Data Integration
informatica Put potential to work, MuleSoft, snapLogic, BedrockData, xplenty

Management / Monitoring
New Relic, APPDYNAMICS, amazon web services, actifio, Numerify, splunk>, DATADOG, rocana, DRIVEN, Anodot

Security
TANIUM, illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrrl, BlueTalon

Storage
amazon web services, Google Cloud Platform, Microsoft Azure, panasas, nimblestorage, COHO DATA, Qumulo

App Dev
apigee, CASK, Keen IO, Typesafe, DRIVEN

Crowd-sourcing
amazon mechanical turk, CrowdFlower, WorkFusion

Analytics

Analyst Platforms
Palantir, AYASDI, Quid, enigma, Digital Reasoning, ORBITAL INSIGHT

Analytics Platforms
Microsoft, guavus, Datameer, Bottlenose, inter|ana

Data Science Platforms
context relevant, CONTINUUM ANALYTICS, DataRobot, Alpine, plotly, ADATAO, MODE, dataiku, nutonian, DOMINO, sense, yhat, ALGORITHMIA

Visualization
tableau, Google Cloud Platform, Qlik, looker, SISENSE, Roambi, ZOOMDATA, datorama, CHARTIO

BI Platforms
Power BI, amazon web services, DOMO, salesforce Wave Analytics, birst, GoodData, kyvos insights, platfora, atscale, ARCADIA, SISENSE

Statistical Computing
sas, SPSS, MATLAB

Log Analytics
splunk>, sumologic, kibana, CLOUD PHYSICS, loggly

Social Analytics
Hootsuite, NETBASE, DATASIFT, tracx, bitly, synthesio, simplereach

Real-Time
amazon web services, METAMARKETS, striim, confluent, DATATORRENT, dataArtisans

Machine Learning
Azure Machine Learning, amazon web services, H2O.ai, Dato, SKYTREE, rapidminer, DATARPM, deepsense.io, VISENZE, PredictionIO, glowfi.sh

Speech & NLP
NarrativeScience, ARRIA, NUANCE, semanticmachines, Gridspace, api.ai, maluuba, cortical.io, MindMeld, iDiBON, yseop

Horizontal AI
IBM Watson, Cortana, sentient technologies, VIV, nervana SYSTEMS, vicarious, nara, Numenta, HyperScience, SCALED INFERENCE, Descartes Labs, clarifai, DEXTRO, MetaMind, Geometric Intelligence

Search
hp Autonomy, ORACLE ENDECA, EXALEAD, Lucidworks, elastic, ThoughtSpot, MAANA, swiftype, Algolia, SINEQUA

Data Services
Mu Sigma DO THE MATH, OPERA SOLUTIONS, DATASCIENCE, SILICON VALLEY DATA SCIENCE, kaggle, datascope, DataKind

For Business Analysts
OrigamiLogic, ClearStory DATA, CIRRO, import io

Web / Mobile / Commerce
Google Analytics, mixpanel, RJMetrics, BLUECORE, AMPLITUDE, granify, sumAll, Airtable, retention SCIENCE, custora

Applications

Sales & Marketing
RADIUS, Gainsight, bloomreach, Zeta, EVERSTRING, livefyre, blue yonder, Lattice, kahuna, infer, SAILTHRU, persado, AVISO, 6sense, QUANTIFIND, ACTIONIQ, fuse|machines, ENGAGIO

Customer Service
MEDALLIA, ATTENSITY, CLARABRIDGE, CLICKFOX, STELLAService, NGDATA, Preact, DigitalGenius, appuri, Wise.io

Human Capital
gild, Connectifier, textio, entelo, hiQ

Legal
RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

Ad Optimization
AppNexus, MediaMath, criteo, OpenX, rocketfuel, Integral Ad Science, theTradeDesk, Adgorithms, dstillery, LiveIntent, DataXu, Appier, TAPAD

Security
CYLANCE, cybereason, CounterTack, AREA 1 SECURITY, ThreatMetrix, SentinelOne, Recorded Future, Guardian Analytics, FORTSCALE, siftscience, Keybase, feedzai, SIGNIFYD

Vertical AI Applications
x.ai, facebook M, Clara, KASISTO, lumiata

Publisher Tools
Outbrain, Taboola, quantcast, Chartbeat, yieldbot, Yieldmo

Govt / Regulation
Socrata, OPENGOV, FiscalNote, enigma, PREDPOL, mark43, OpenDataSoft

Finance
Affirm, LendingClub, OnDeck, Kreditech, zestfinance, LendUp, Kabbage, tidemark, Payoff, INSIKT, Zuora, Dataminr, Lenddo, KENSHO, AIDYIA, iSENTIUM, Quantopian, sentient technologies

Education/ Learning
KNEWTON, Clever, declara, PANORAMA EDUCATION, knowre

Life Sciences
23andMe, PATHWAY GENOMICS, Counsyl, deep genomics, Recombine, KYRUUS, FLATIRON, zymergen, HealthTap, METABIOTA, ZEPHYR HEALTH, ovia, Ginger.io, transcriptic, Glow, enlitic, AiCure, Atomwise

Industries
OPOWER, eHarmony, RetailNext, duetto, STITCH FIX, WorkFusion, BLUE RIVER, TACHYUS, SwiftKey, Seeq, FarmLogs, HowGood, celect, SIGHT MACHINE, statmuse, BOXEVER

Cross-Infrastructure/Analytics
amazon web services, Google, Microsoft, IBM, SAP, sas, 1010 data, hp Autonomy, VERTICA An HP Company, vmware, TIBCO, TERADATA, ORACLE, NetApp

Open Source

Framework
hadoop HDFS, hadoop MapReduce, YARN, Spark, MESOS, TEZ, Apache Kylin, Flink, CDAP

Query / Data Flow
Pig, HIVE, SLAMDATA, APACHE DRILL, Google Cloud Dataflow

Data Access
cassandra, APACHE HBASE, mongoDB, accumulo, FLUME, SciDB, kafka, nifi, CouchDB, riak, OPENTSDB

Coordination
talend, OOZIE, Apache Zookeeper, Apache Ambari

Real-Time
STORM, Spark, APEX, Flink, TACHYON, druid

Stat Tools
R, Scala, NumPy, SciPy

Machine Learning
mllib, Apache SINGA, mahout, MADlib, Aerosolve, Caffe, torch, CNTK, TensorFlow, scikit learn, WEKA, FeatureFu, jupyter, DL4J, VELES, DIMSUM

Search
elasticsearch, Solr, Lucene

Security
Apache Ranger

Visualization
Zeppelin

Data Sources & APIs

Health
JAWBONE, GARMIN, practicefusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API

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

Financial & Economic Data
Bloomberg, DOW JONES, THOMSON REUTERS, S&P CAPITAL IQ, YODLEE, PREMISE, quandl, xignite, CB INSIGHTS, mattermark, estimize, PLAID

Air / Space / Sea
PLANET LABS, spire, WINDWARD, CRUISE, SKYCATCH, Airware, DroneDeploy

Location / People / Entities
acxiom, Experian, EPSILON, GARMIN, foursquare, InsideView, esri, STREETLINE CONNECTING THE REAL WORLD, CARTODB, factual, PlaceIQ, Crimson Hexagon, placemeter, BASIS TECHNOLOGY, Sense360

Other
qualtrics, panjiva, DATA.GOV

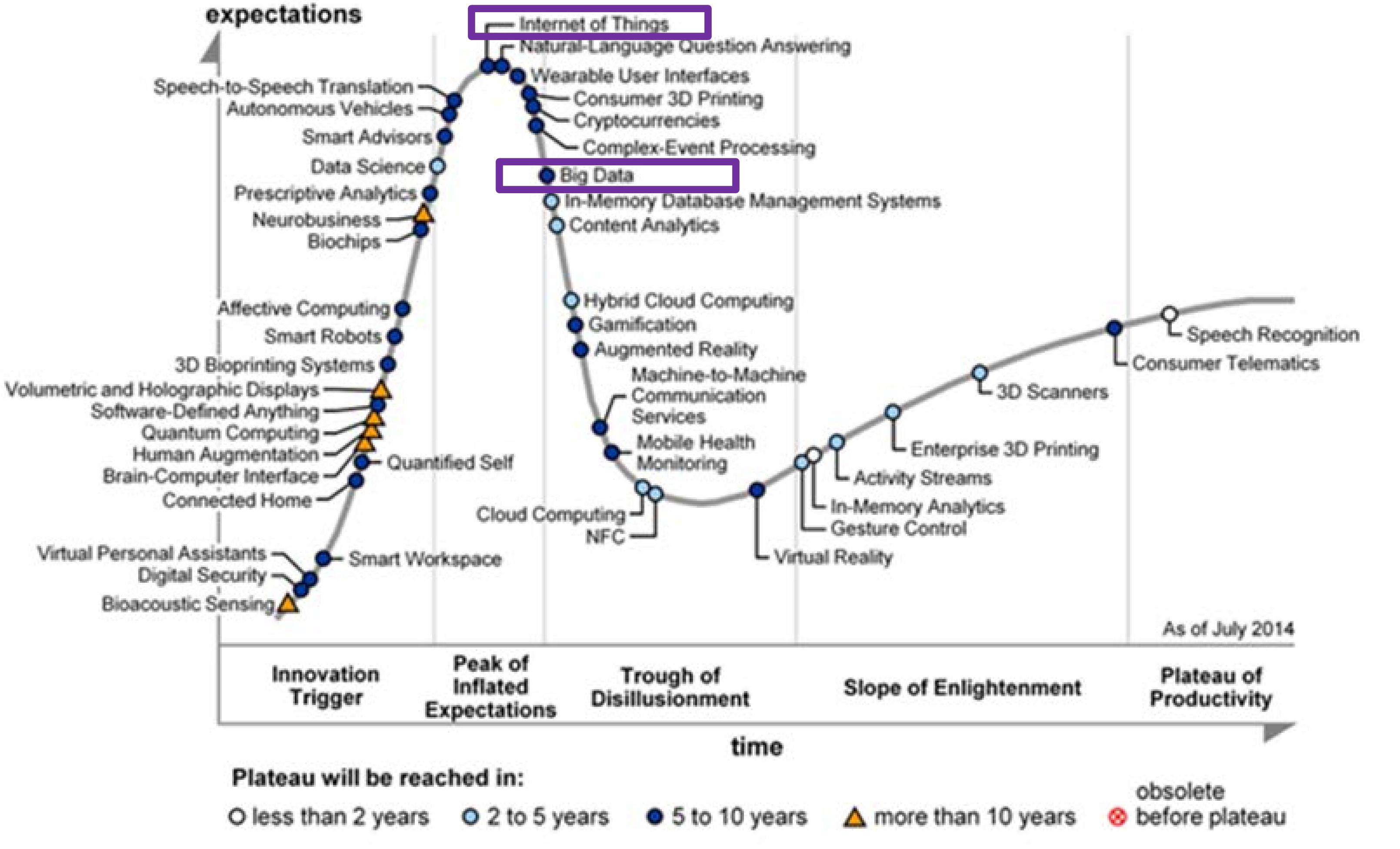
Incubators & Schools
GA, DataCamp, INSIGHT, DataElite, METIS, The Data Incubator

Last Updated 2/12/2016

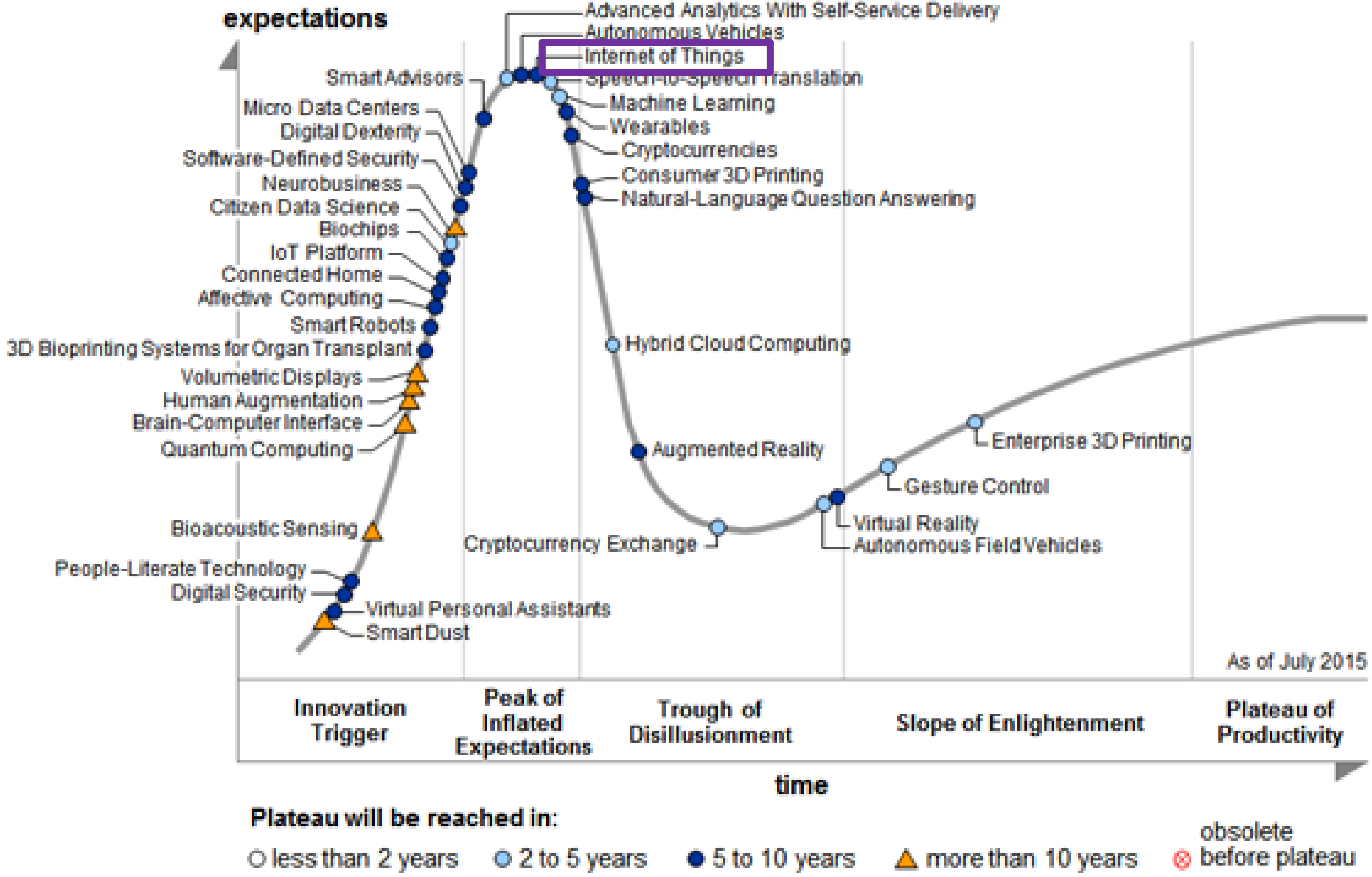
© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK

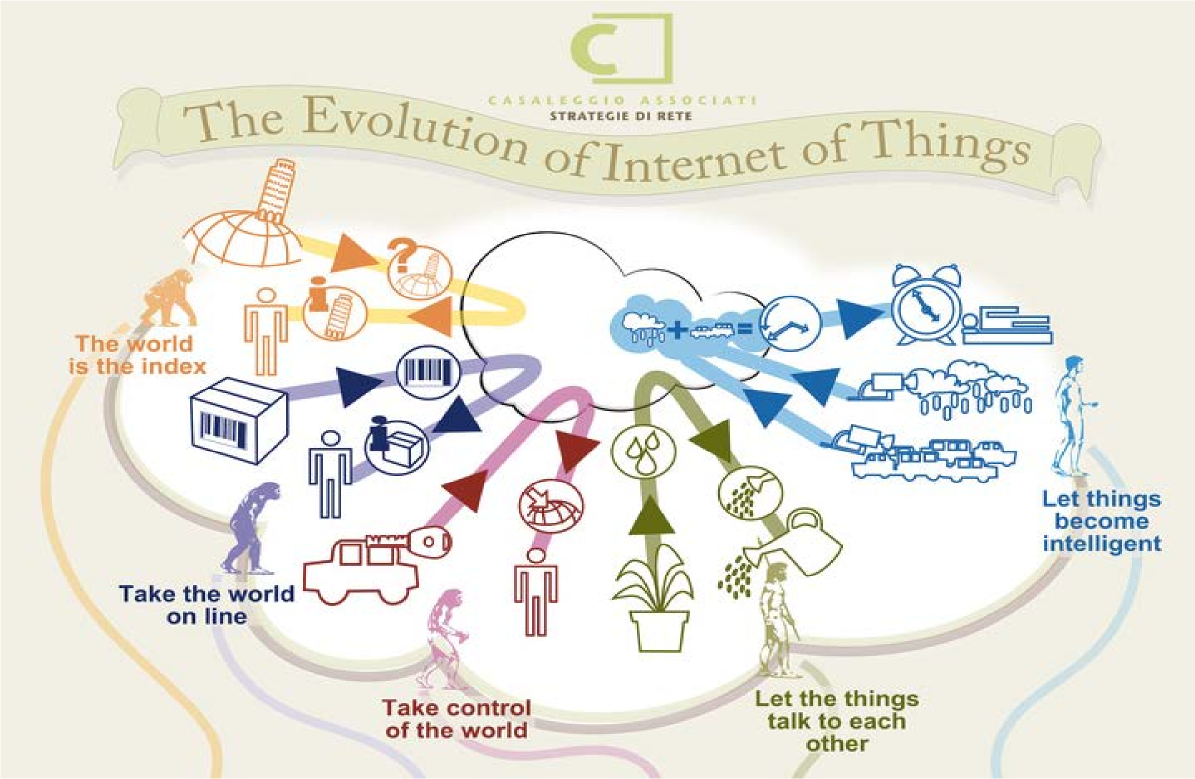
Gartner Hype Cycle for Emerging Technologies 7/14

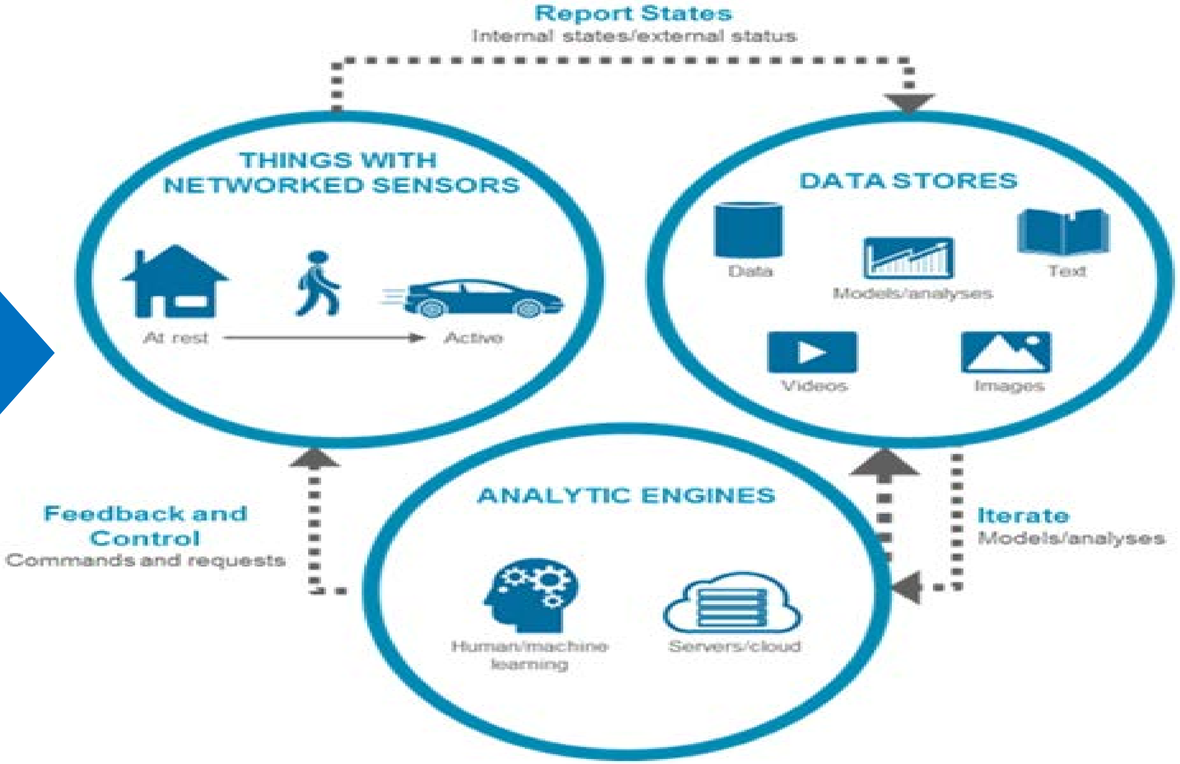


Gartner Hype Cycle for Emerging Technologies 7/15

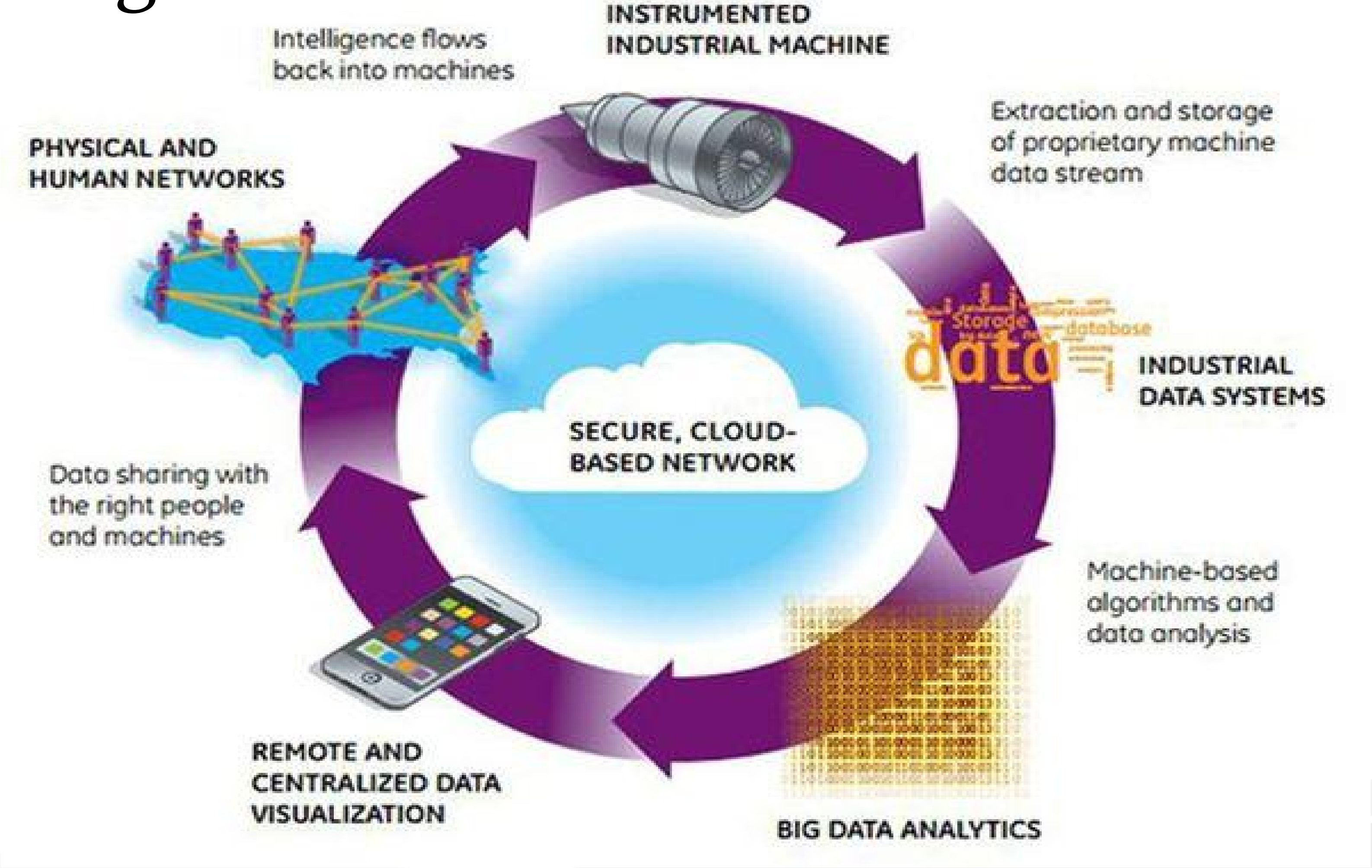


What is IoT?

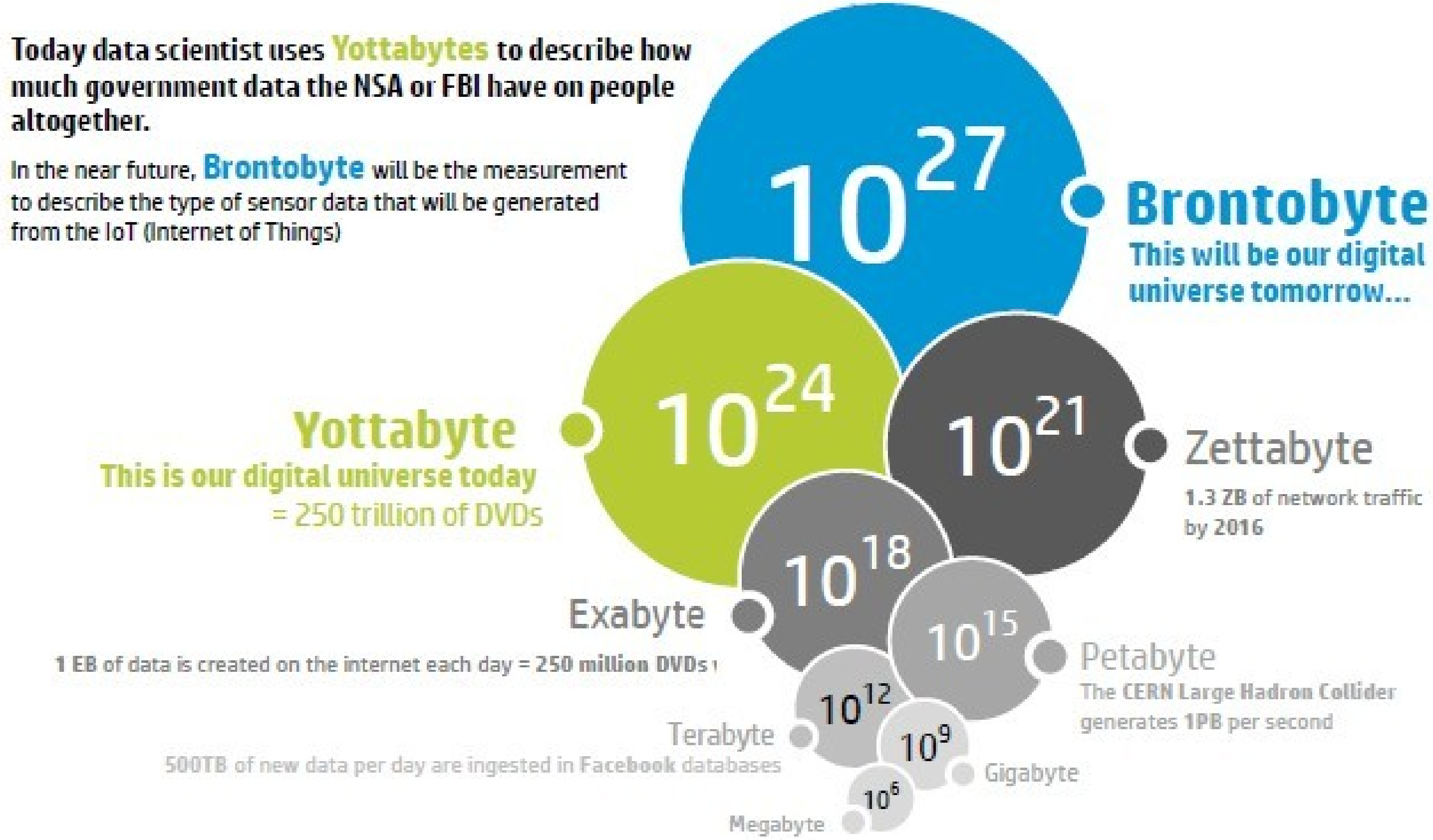

CASALEGGIO ASSOCIATI
STRATEGIE DI RETE
The Evolution of Internet of Things
The world is the index
Take the world on line
Take control of the world
Let the things talk to each other
Let things become intelligent


Report States
Internal states/external status
THINGS WITH NETWORKED SENSORS
At rest
Active
DATA STORES
Data
Models/analyses
Text
Videos
Images
Feedback and Control
Commands and requests
ANALYTIC ENGINES
Human/machine learning
Servers/cloud
Iterate
Models/analyses

IoT and Big Data


Intelligence flows back into machines
INSTRUMENTED INDUSTRIAL MACHINE
Extraction and storage of proprietary machine data stream
PHYSICAL AND HUMAN NETWORKS
Storage
database
data
INDUSTRIAL DATA SYSTEMS
SECURE, CLOUD-BASED NETWORK
Data sharing with the right people and machines
Machine-based algorithms and data analysis
REMOTE AND CENTRALIZED DATA VISUALIZATION
BIG DATA ANALYTICS

Growing Internet of Things Data



Today data scientist uses Yottabytes to describe how much government data the NSA or FBI have on people altogether.

In the near future, Brontobyte will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things)

This will be our digital universe tomorrow...

This is our digital universe today = 250 trillion of DVDs

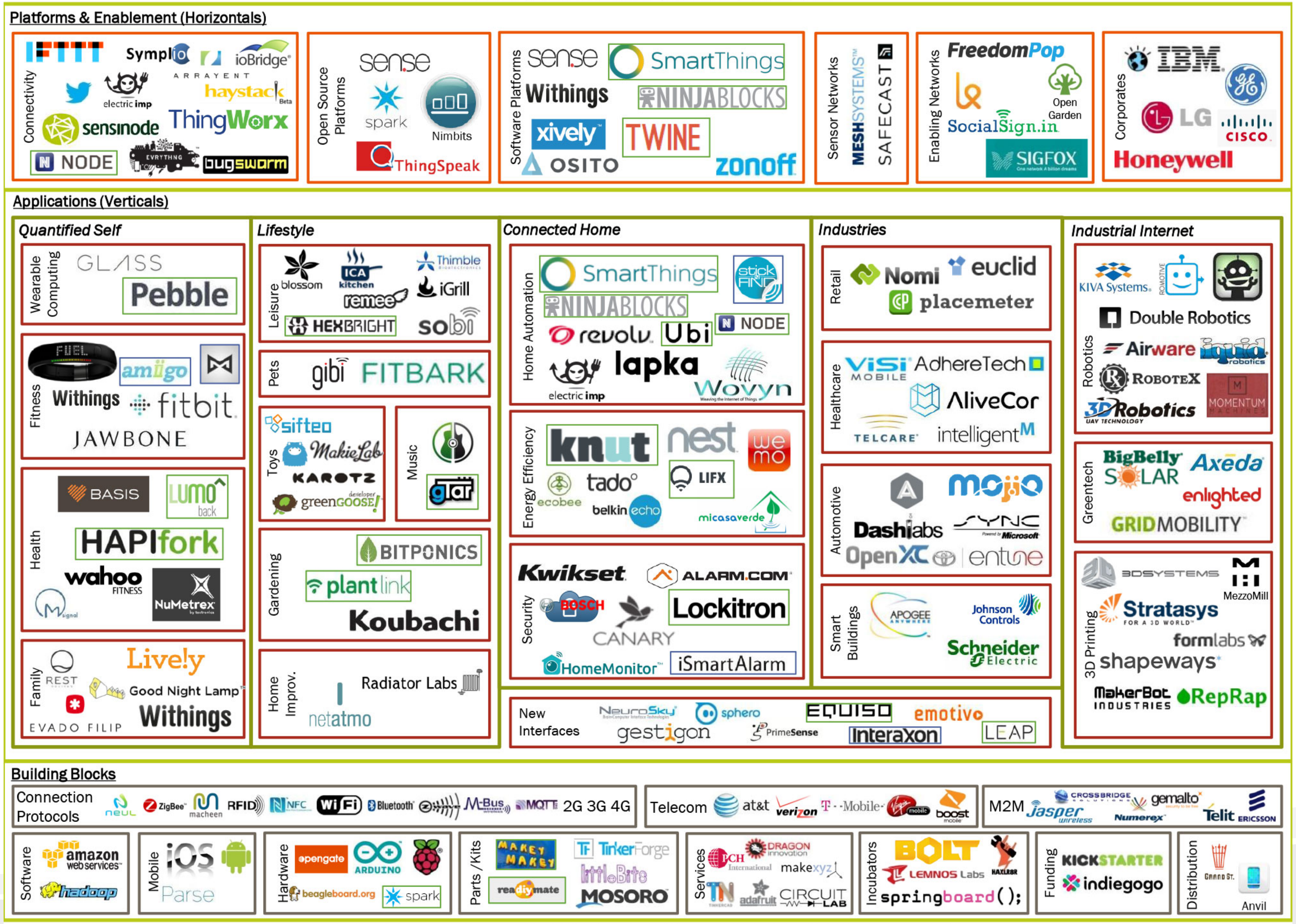
1.3 ZB of network traffic by 2016

1 EB of data is created on the internet each day = 250 million DVDs

The CERN Large Hadron Collider generates 1PB per second

500TB of new data per day are ingested in Facebook databases

INTERNET OF THINGS LANDSCAPE



Platforms & Enablement (Horizontals)

- Connectivity: IFTTT, Symplio, Arrayent, ioBridge, Twitter, electric imp, haystack Beta, sensinode, ThingWorx, NODE, Evrything, bugswarm
- Open Source Platforms: sense, spark, Nimbits, ThingSpeak
- Software Platforms: sense, SmartThings, Withings, NINJABLOCKS, xively, TWINE, OSITO, zonoff
- Sensor Networks: MESHSYSTEMS, SAFECAST
- Enabling Networks: FreedomPop, Open Garden, SocialSign.in, SIGFOX
- Corporates: IBM, GE, LG, Cisco, Honeywell

Applications (Verticals)

Quantified Self

- Wearable Computing: GLASS, Pebble
- Fitness: Fuel, amiigo, Withings, fitbit, JAWBONE
- Health: BASIS, Lumo back, HAPIfork, wahoo fitness, NuMetrex, Signal
- Family: REST, Lively, Good Night Lamp, EVADO FILIP, Withings

Lifestyle

- Leisure: blossom, ICA kitchen, Thimble, iGrill, remee, HEXBRIGHT, sobi
- Pets: gibi, FITBARK
- Toys: sifteo, MakieLab, KAROTZ, greenGOOSE developer
- Music: glar
- Gardening: BITPONICS, plantlink, Koubachi
- Home Improv.: Radiator Labs, netatmo

Connected Home

- Home Automation: SmartThings, stick N find, NINJABLOCKS, revolv, Ubi, NODE, electric imp, lapka, Wovyn
- Energy Efficiency: knut, nest, wemo, ecobee, tado°, LIFX, belkin echo, micasaverde
- Security: Kwikset, ALARM.COM, BOSCH, CANARY, Lockitron, HomeMonitor, iSmartAlarm
- New Interfaces: NeuroSky, sphero, EQUISO, emotivo, gestigon, PrimeSense, InteraXon, LEAP

Industries

- Retail: Nomi, euclid, placemeter
- Healthcare: VISI MOBILE, AdhereTech, AliveCor, TELCARE, intelligentM
- Automotive: Automatic, mojio, Dashlabs, SYNC Powered by Microsoft, OpenXC, entune
- Smart Buildings: APOGEE, Johnson Controls, Schneider Electric

Industrial Internet

- Robotics: KIVA Systems, Romotive, Double Robotics, Airware, liquid robotics, ROBOTEX, 3D Robotics, MOMENTUM MACHINES
- Greentech: BigBelly SOLAR, Axeda, enlighted, GRIDMOBILITY
- 3D Printing: 3D SYSTEMS, MezzoMill, Stratasys, formlabs, shapeways, MakerBot Industries, RepRap

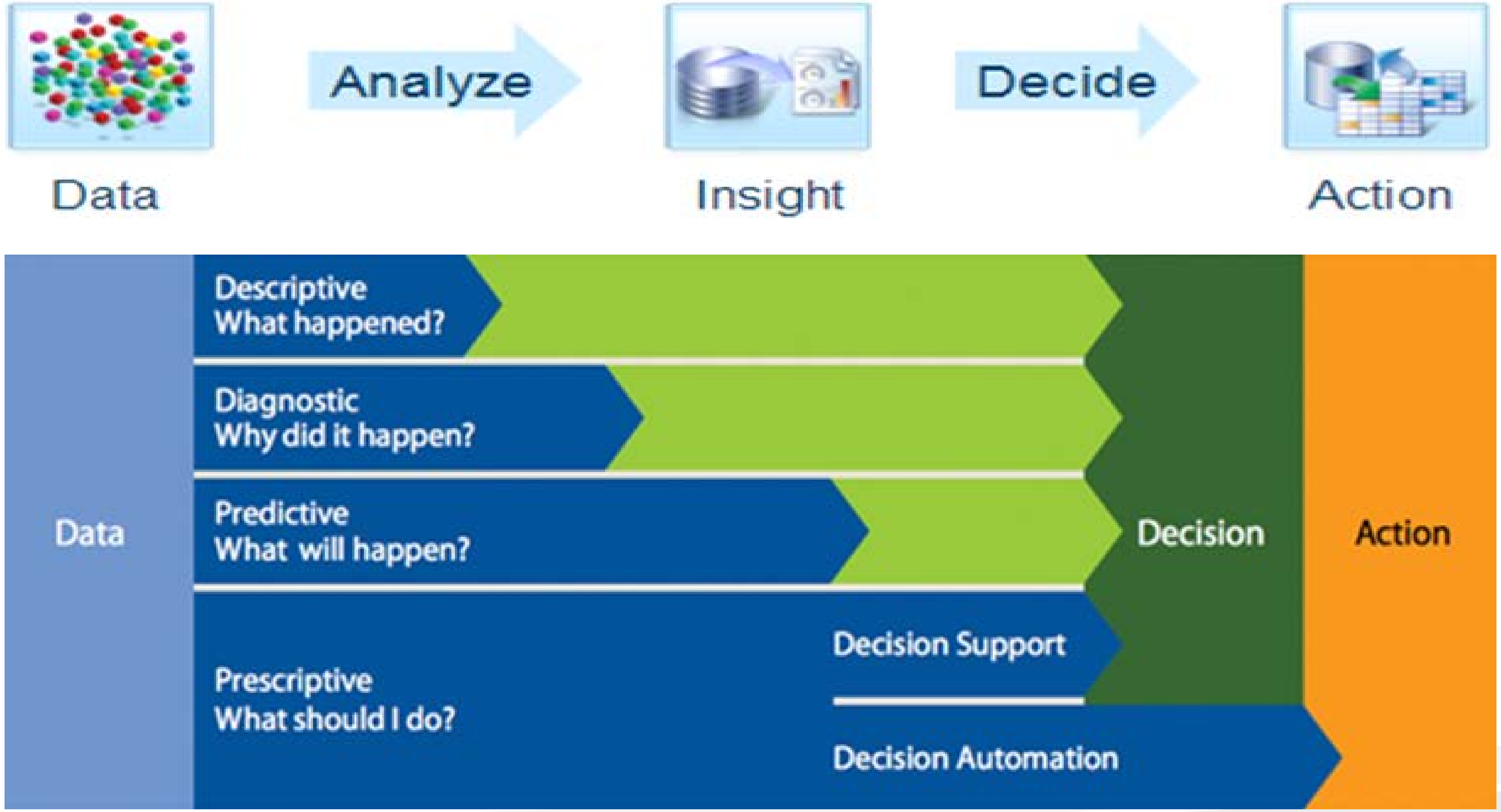
Building Blocks

- Connection Protocols: neul, ZigBee, macheen, RFID, NFC, WiFi, Bluetooth, M-Bus, MQTT, 2G 3G 4G
- Telecom: at&t, verizon, T-Mobile, Virgin mobile, boost mobile
- M2M: Jasper wireless, CROSSBRIDGE, gemalto, Numerex, Telit, ERICSSON
- Software: amazon web services, hadoop
- Mobile: iOS, Android, Parse
- Hardware: opengate, ARDUINO, Raspberry Pi, beagleboard.org, spark
- Parts / Kits: MAKEY MAKEY, TinkerForge, littleBits, readymate, MOSORO
- Services: PCH International, DRAGON innovation, makexyz, TN, adafruit, CIRCUIT LAB
- Incubators: BOLT, LEMNOS Labs, HAXLR8R, springboard();
- Funding: KICKSTARTER, indiegogo
- Distribution: Grand St., Anvil

© Matt Turck (@mattturck), Sutian Dong (@sutiandong) & FirstMark Capital (@firstmarkcap)

What Is The Value In Big Data?

Transforming Data Into Insight For Making Better Decisions



Data

Analyze

Insight

Decide

Action

Data

Descriptive
What happened?

Diagnostic
Why did it happen?

Predictive
What will happen?

Prescriptive
What should I do?

Decision Support

Decision Automation

Decision

Action

Gartner, 2013

Why Data Mining?

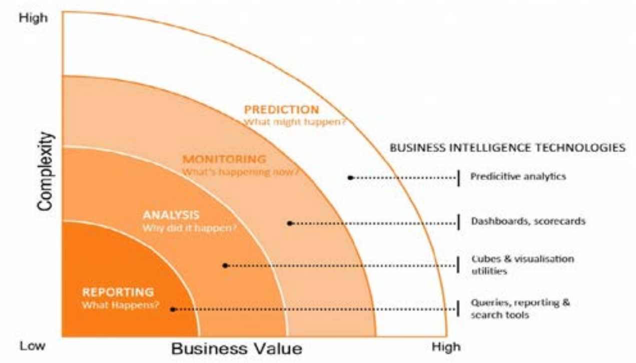
- Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories.

“*We are drowning in data, but starving for knowledge!*”

(John Naisbitt, 1982)

- Data Mining is derived from the necessity to address the Data Explosion

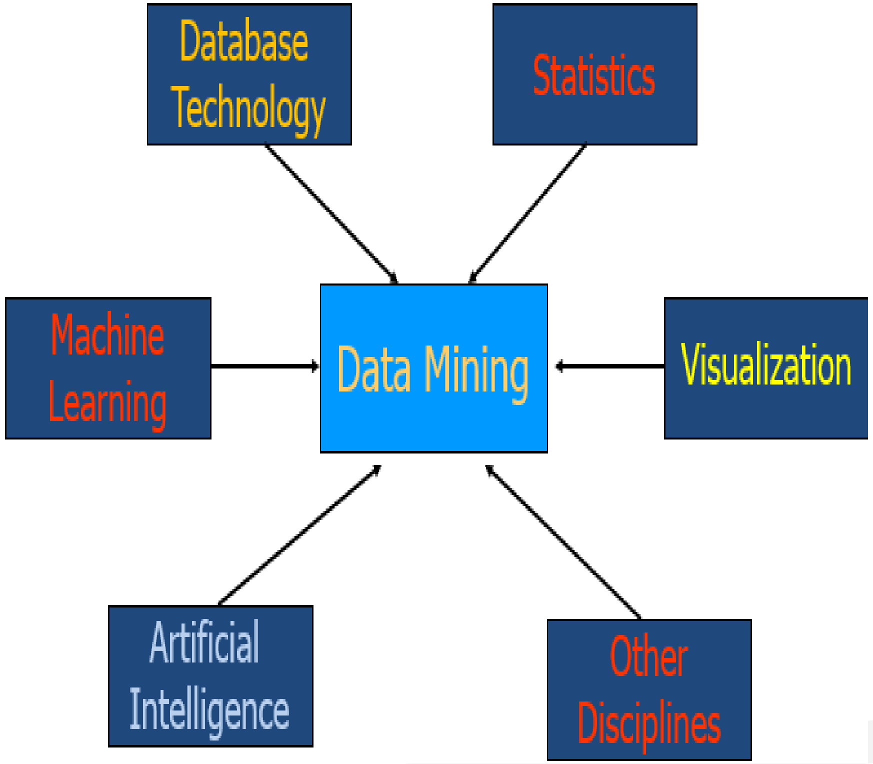
What is Data Mining?



- A set of technologies that uncovers relationships and patterns within large volumes of data that can be used to predict future behavior and events
- Predictive Analytics is technology that learns form experience to predict the future outcomes in order to drive better business decisions
    - Extracting / "Mining"
        - Information/Meaning from data
        - Interesting knowledge (rules, regularities, patterns, constraints) from raw data
        - Implicit, previously unknown and unexpected, potentially extremely useful information from data

Data Mining is Multidisciplinary Field

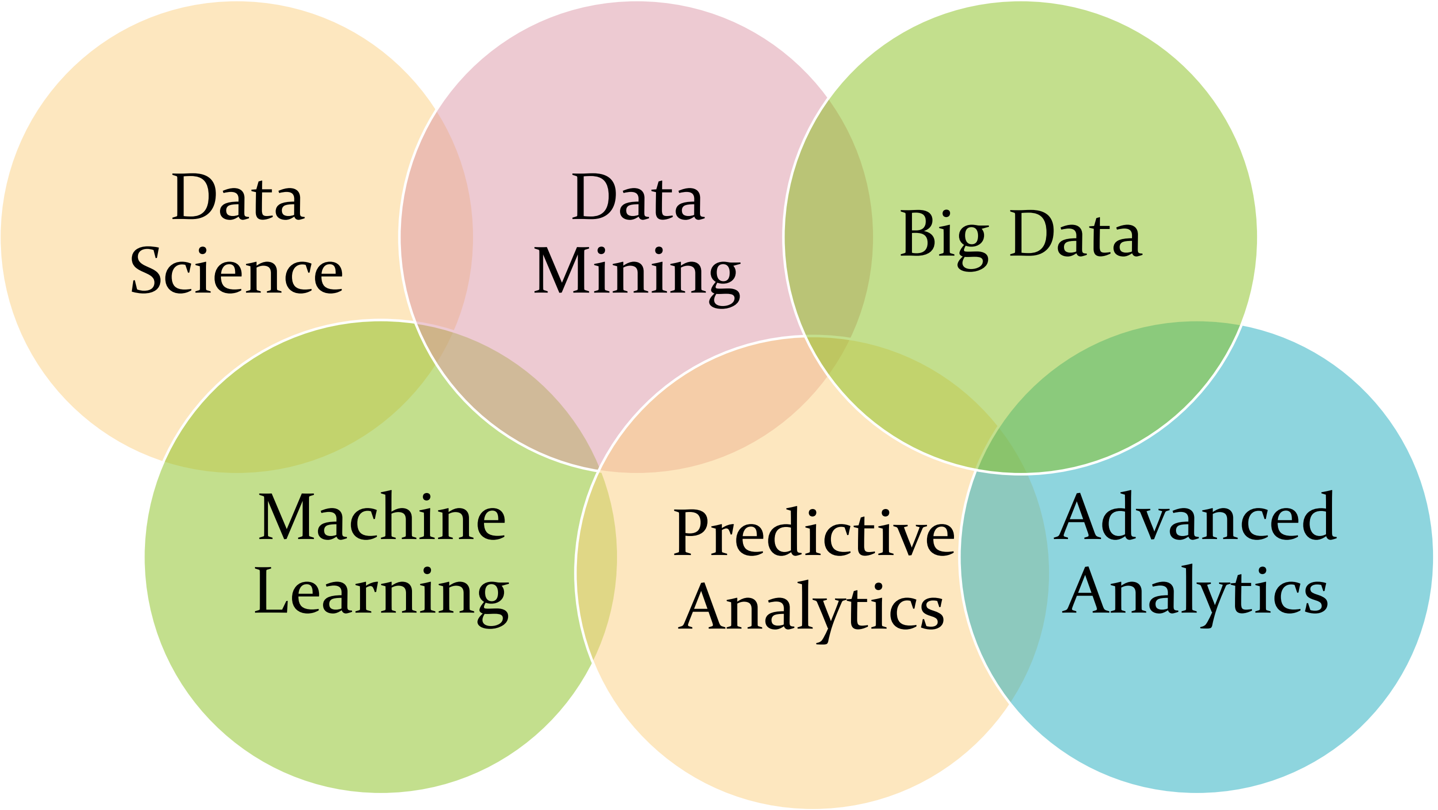
- Database technology
- Artificial Intelligence
- Machine Learning including Neural Networks
- Statistics
- Pattern recognition
- Knowledge-based systems/acquisition
- High-performance computing
- Data visualization
- Other Disciplines



Data Mining is NOT...

- Data Warehousing
- (Deductive) query processing
  - SQL/ Reporting
- Software Agents
- Expert Systems
- Online Analytical Processing (OLAP)
- Statistical Analysis Tool
- Data visualization
- BI – Business Intelligence
- Workflows

Terminology



What’s Involved?

Explore Data

Find Patterns

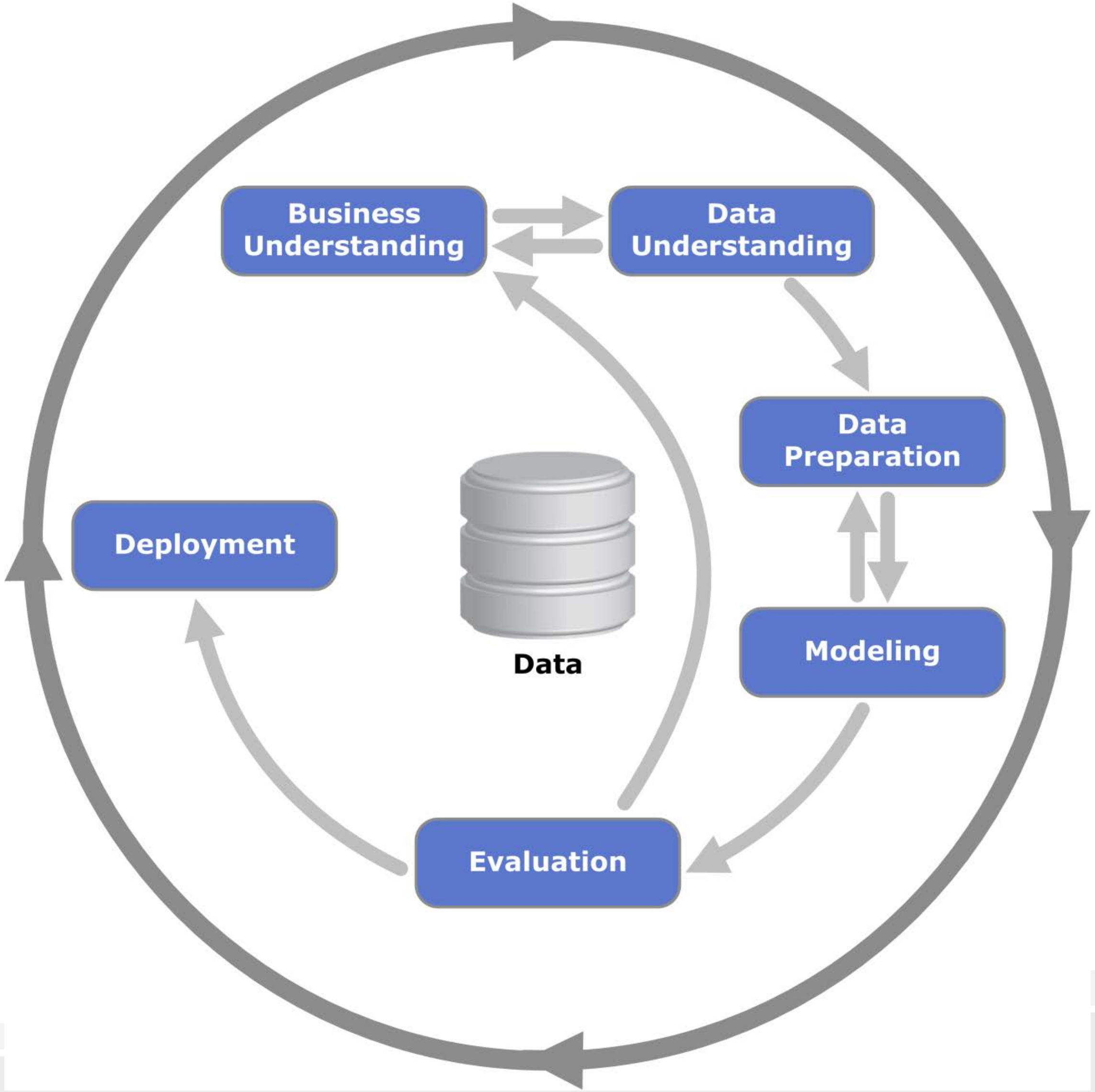
Perform Predictions

- **Learn the application domain**
  - Relevant prior knowledge and goals of application
- **Create a target data set:**
  - Data selection
  - Clean and preprocess the data (may take > 60% of effort!)
    - Data reduction and transformation
    - Find useful features, dimensionality/variable reduction, representation
- **Choose data mining functions**
  - Summarization, classification, regression, association, clustering
- **Apply data mining methods/algorithm(s)**
  - Data mining: search for patterns of interest
- **Evaluation, validation, and knowledge presentation**
  - Visualization, transformation, removing redundant patterns, etc.
- **Use and integrate discovered knowledge**

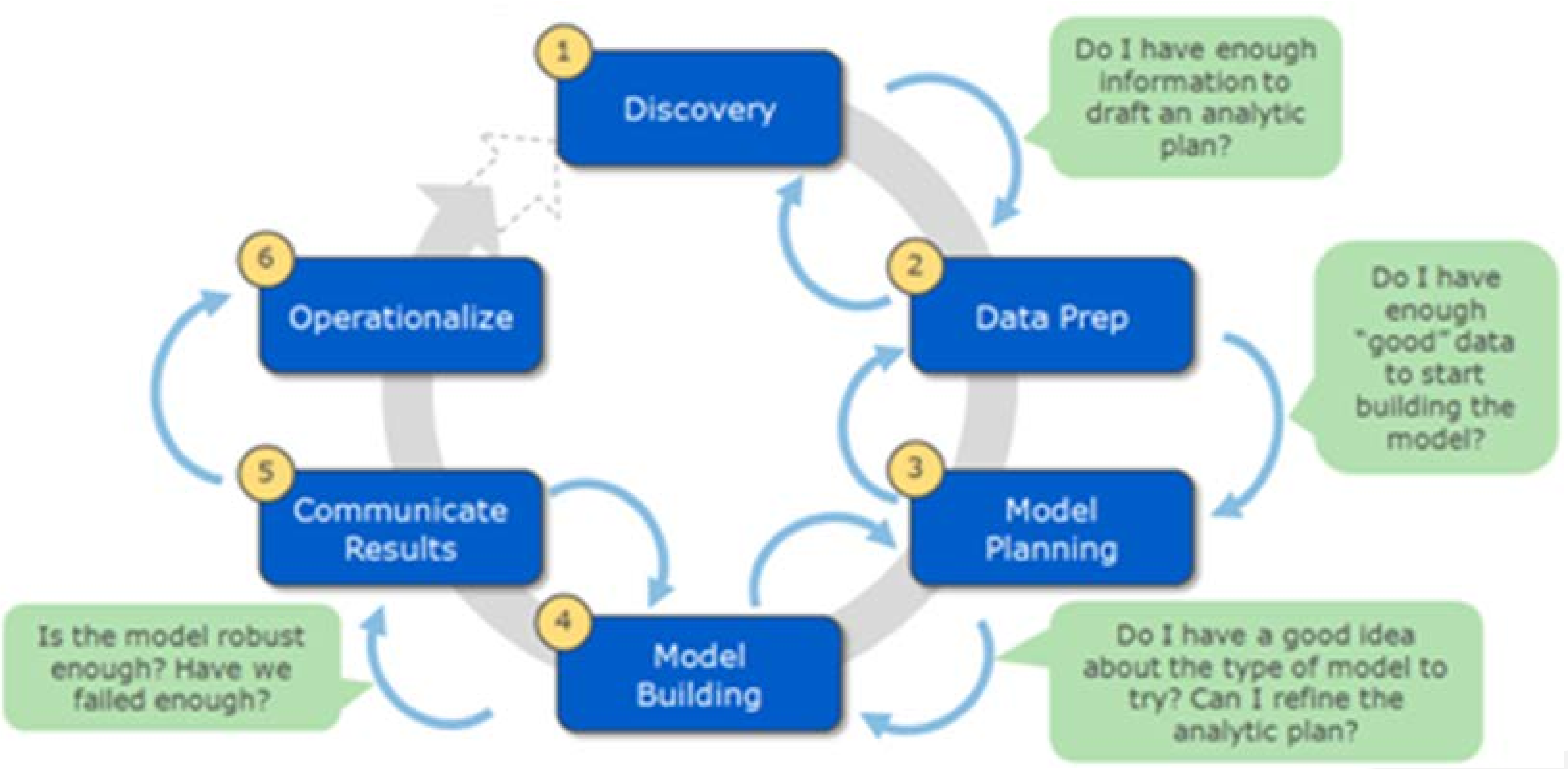
Data Mining Functions

- Exploratory Data Analysis
  - Visualization, Statistical analysis
- Descriptive Modeling/Discovering patterns rules
  - Cluster analysis/segmentation
  - Association/Dependency rules
  - Sequential patterns
- Predictive Modeling
  - Classification and Regression
  - Temporal sequences
  - Deviation detection

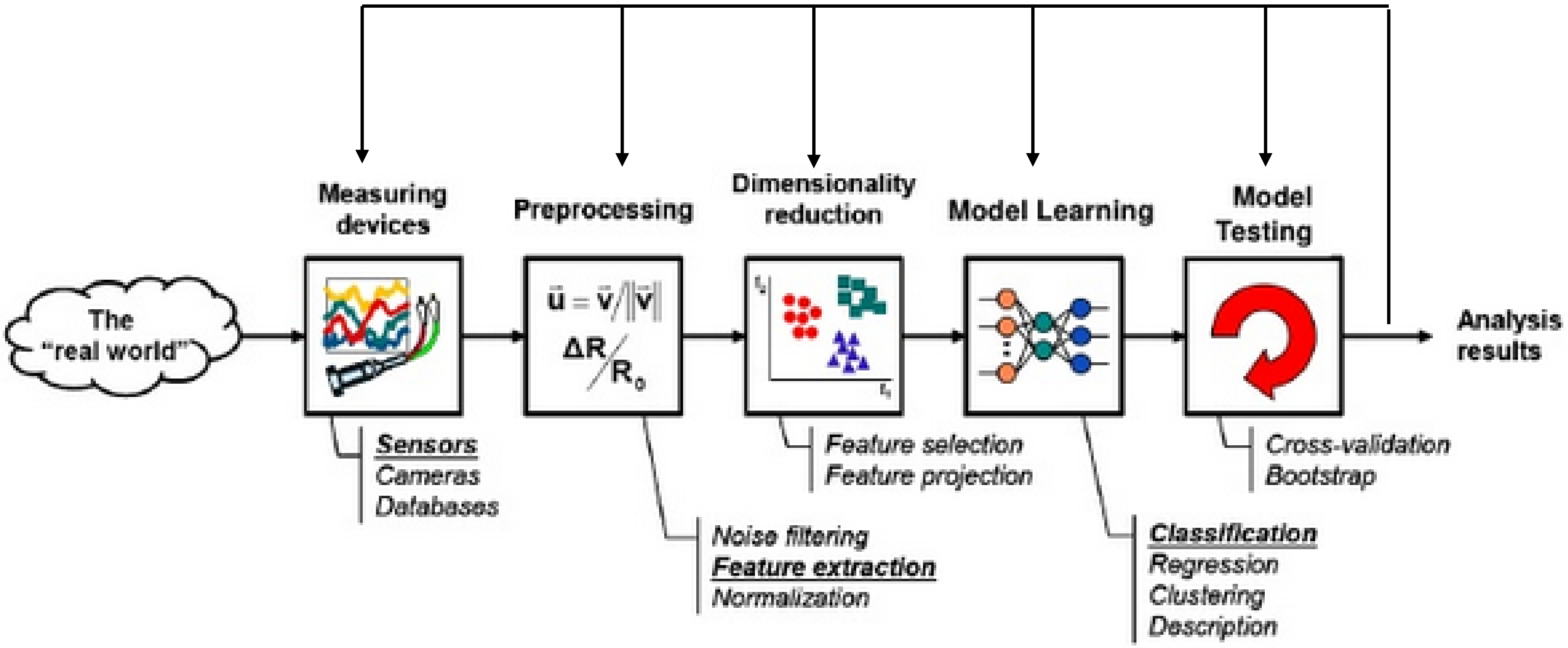
CRISP-DM Iterative Process



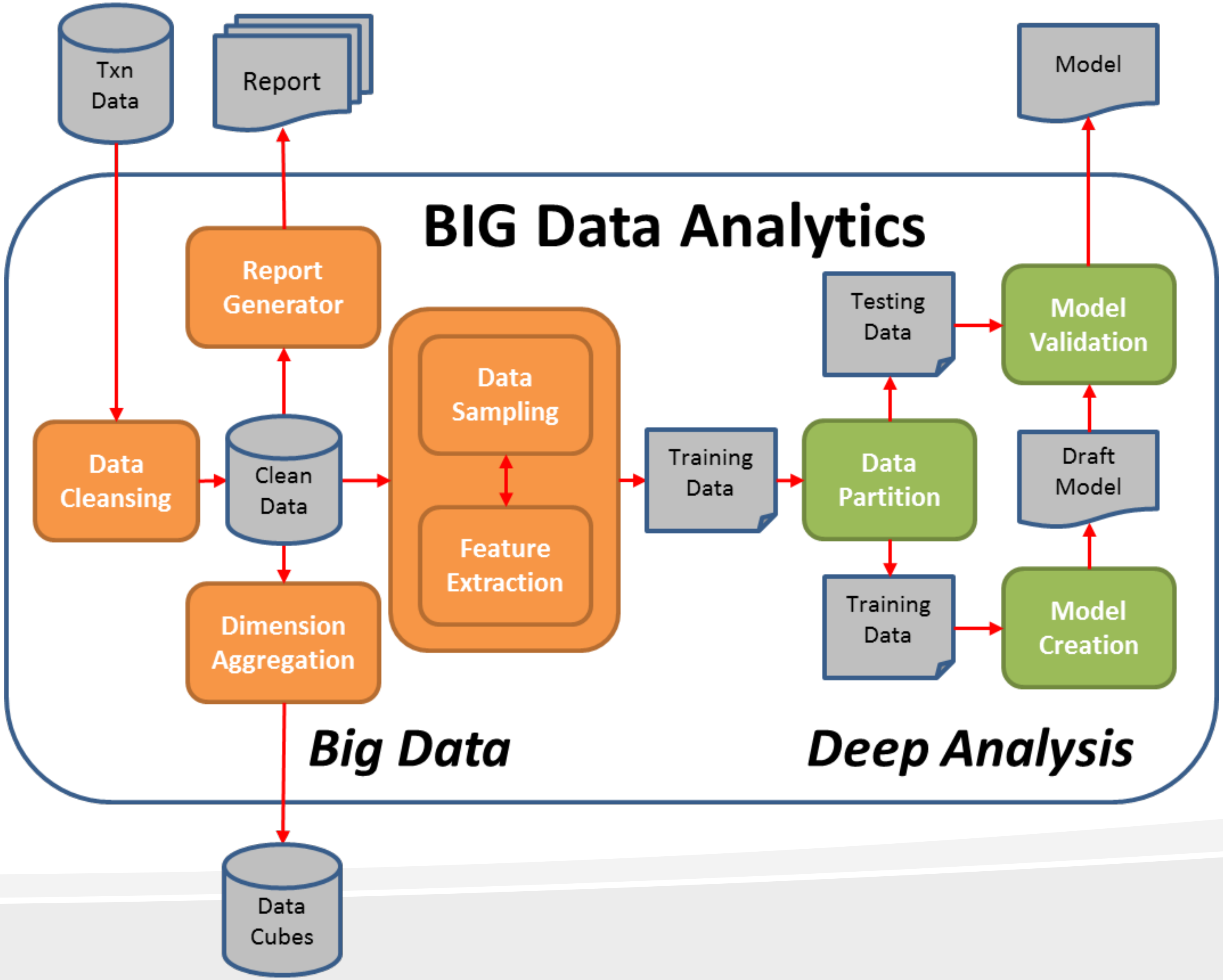
Predictive Analytics Process


1
Discovery
Do I have enough information to draft an analytic plan?
2
Data Prep
Do I have enough "good" data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
Is the model robust enough? Have we failed enough?
5
Communicate Results
6
Operationalize

Data Mining Process

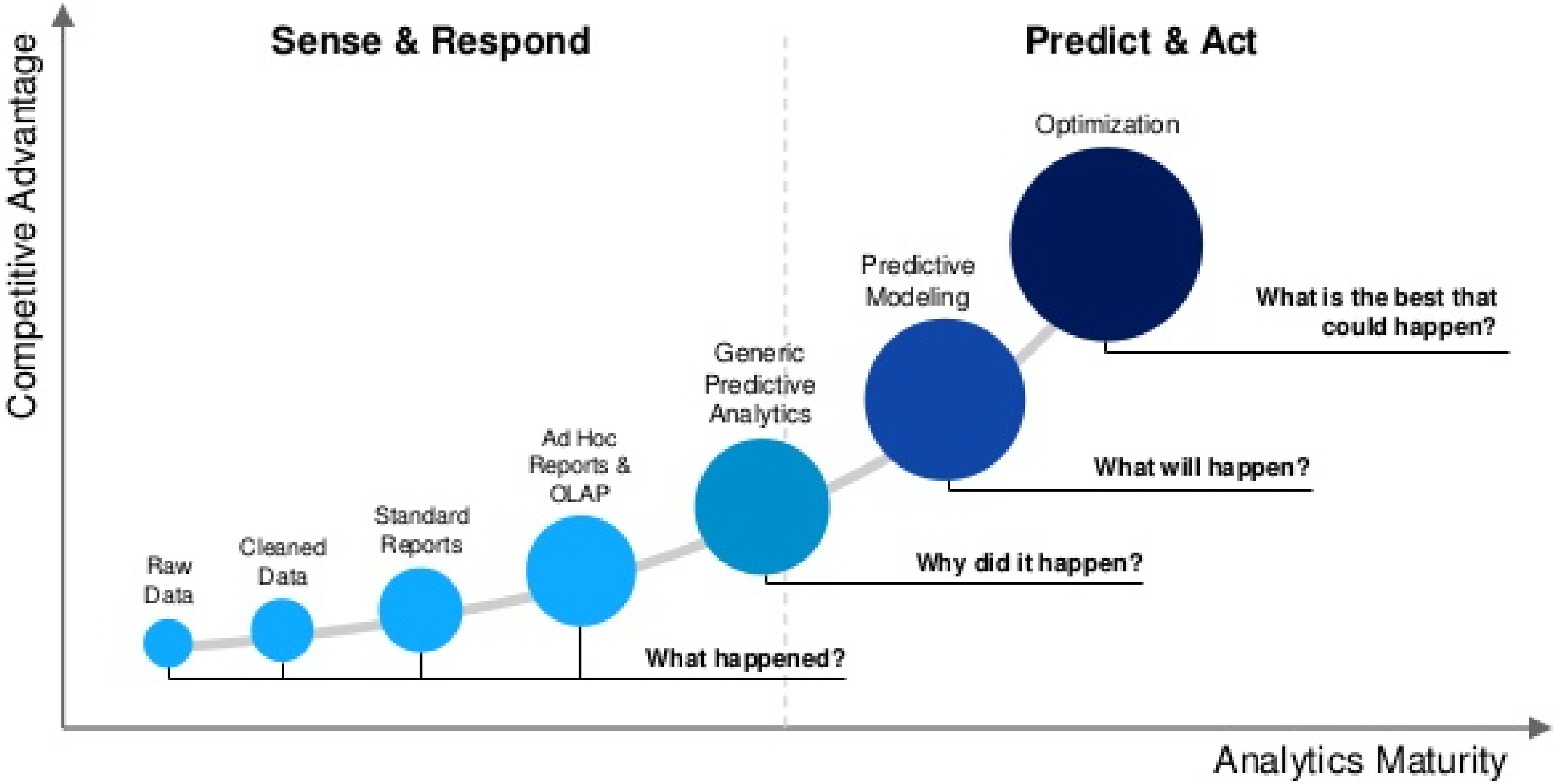


Big Data and Predictive Analytics Processing



Horickyblogspot.com

Analytics Maturity Levels



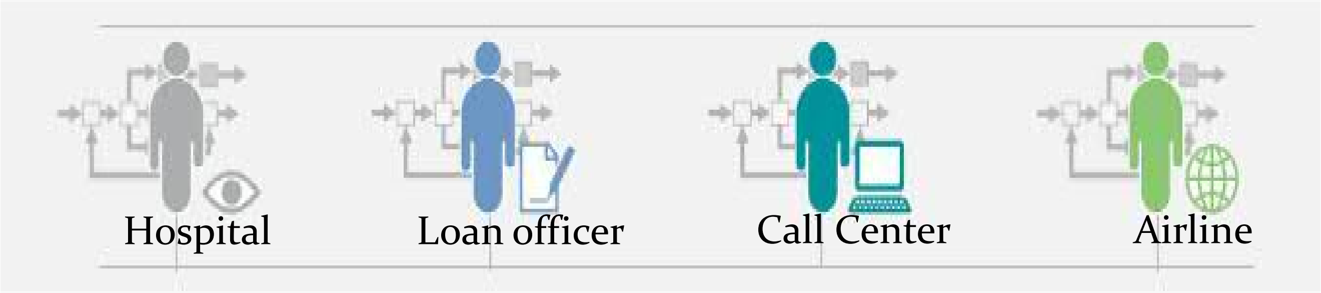
Numerous Applications

Improve ability to classify and treat cancer, tumors, diseases

Adjust credit scores as transactions are occurring to account for risk fluctuations

Apply inferred customer social relationships to prevent churn

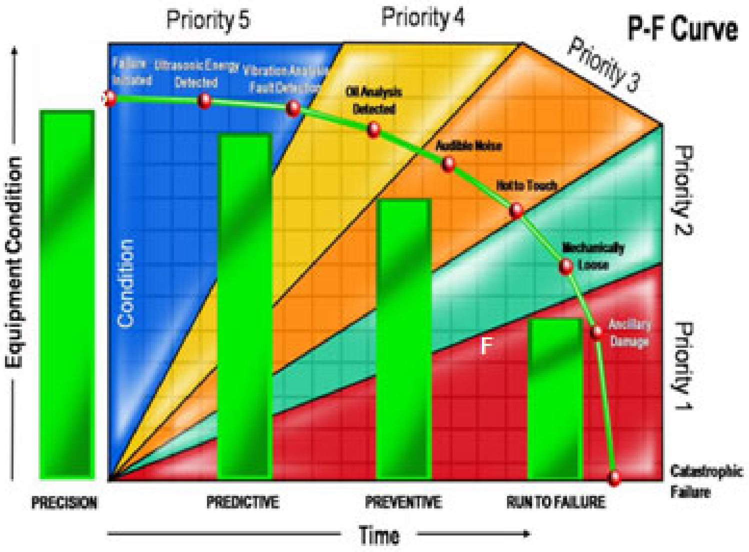
Increase revenue and customer satisfaction by discovering passengers who are likely to miss their flight



Condition Based Maintenance (CBM)

- ❑ Consumes online data - conditions are constantly monitored and their signatures evaluated
- ❑ Focuses on sensors and communications
- ❑ Machines are networked collecting large volumes of data
- ❑ What to do with all this data?
- ❑ We need to turn data into insight

- ❑ Patterns of past behavior can be “learned” to provide deeper insight
  - ❑ How much change or degradation has occurred since the last round
  - ❑ What is the chance of failure for each machine?
  - ❑ What is the current most likely failure to happen?
  - ❑ What is the impact of the failure?

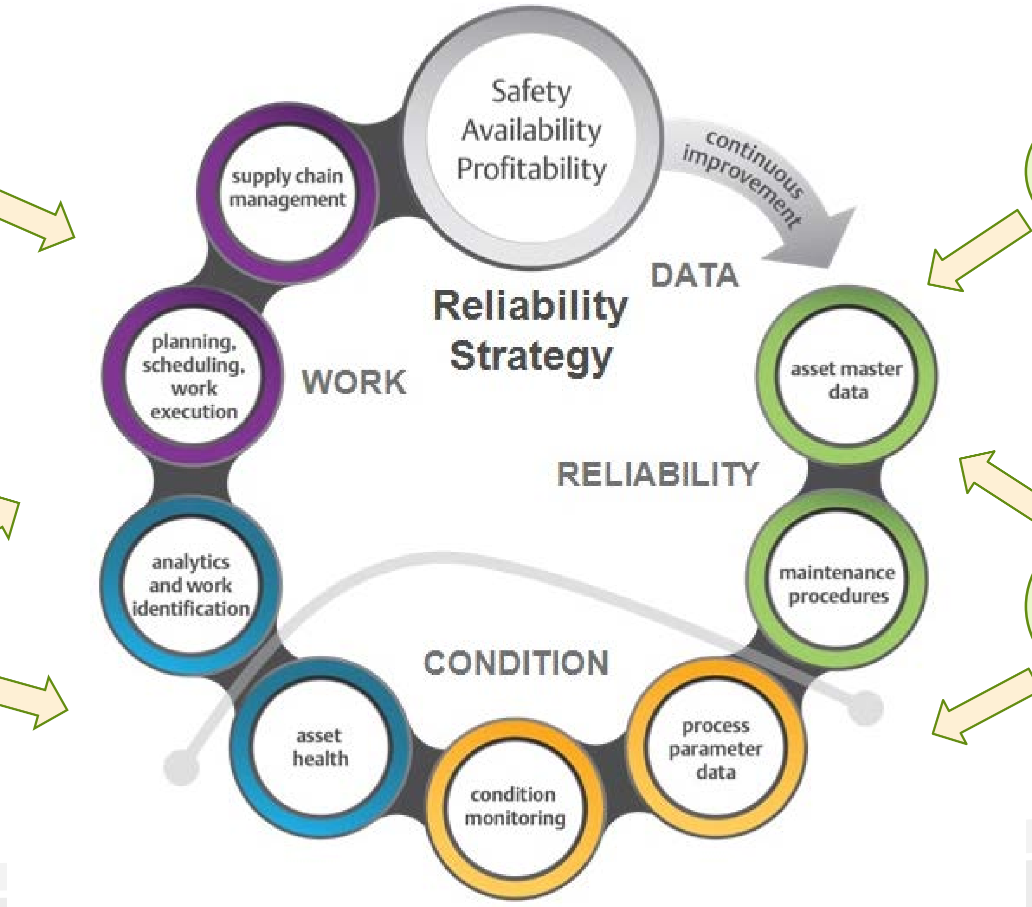


PF Curve

Reliability Value Chain

Optimization
Prescriptive
Analytics

Predictive
Analytics

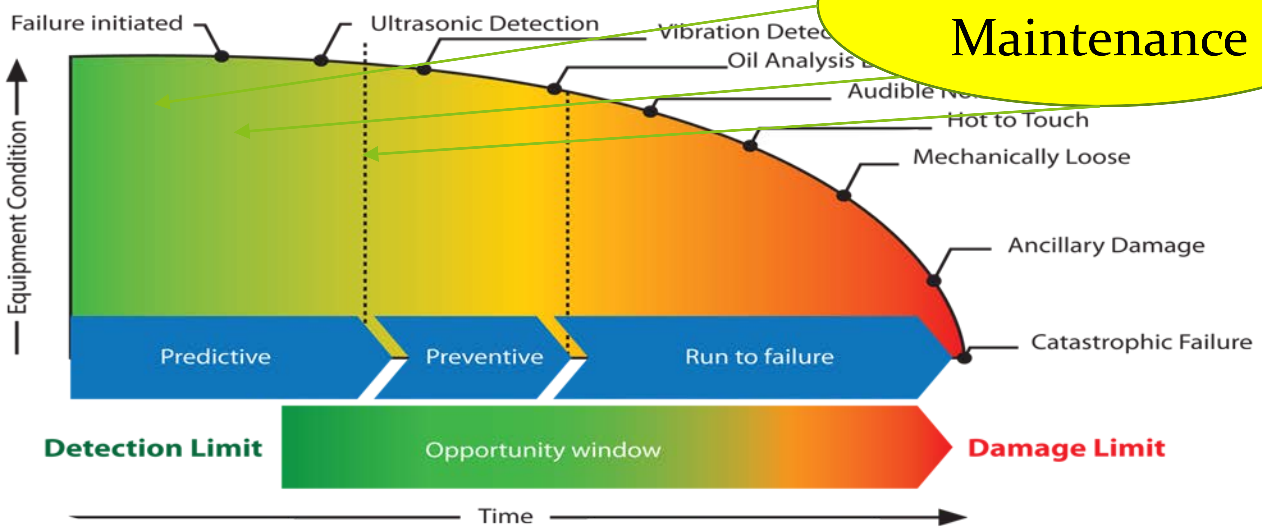


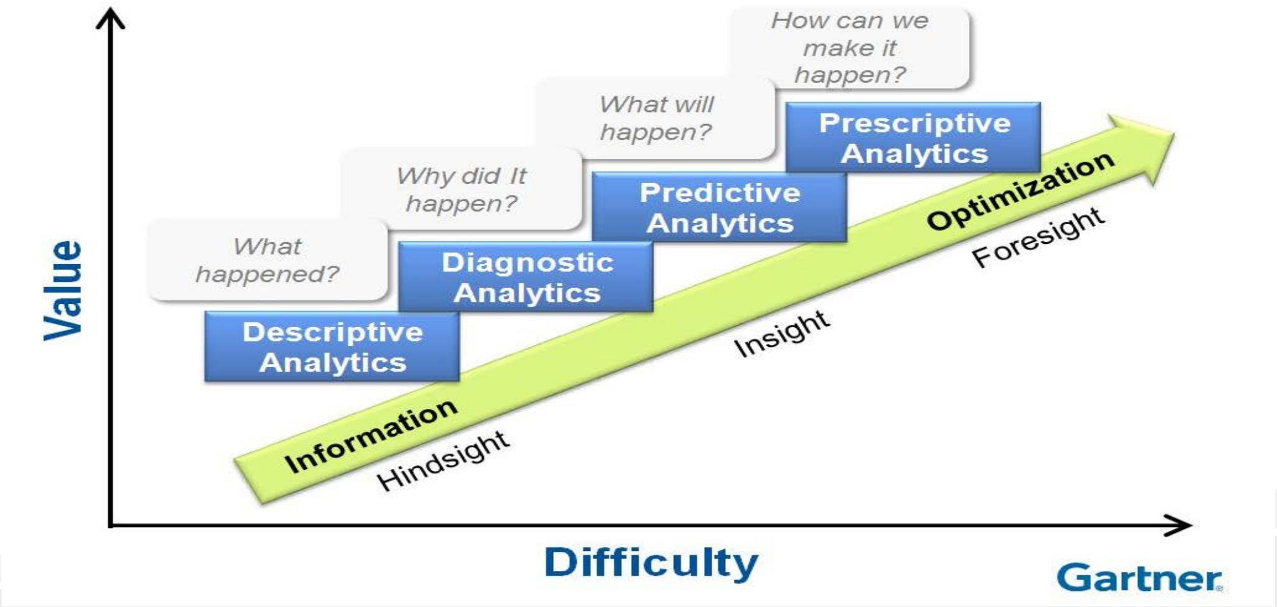
Sensor
Fusion
Big Data

Predictive
Analytics

Unexpected Failures

Predictive Maintenance

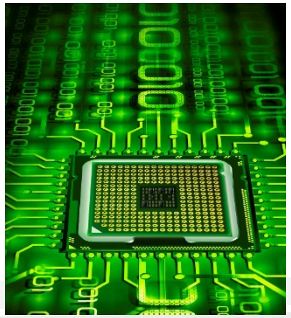




Intelligent CBM Across Industries

Semiconductor industry

- Capital process equipment
- Anomaly and event detection
- Predictive Maintenance
- Proactive decisions

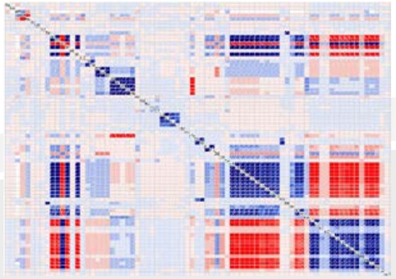


Network Data

- Improved flow
- Asset optimization
- Proactive decisions
- Resource allocation
- Data driven risk prediction

Manufacturing

- Improved process
- Failure prediction
- Quality control
- Prevents damages
- Lowers production cost



Intelligent CBM for Utilities Projects

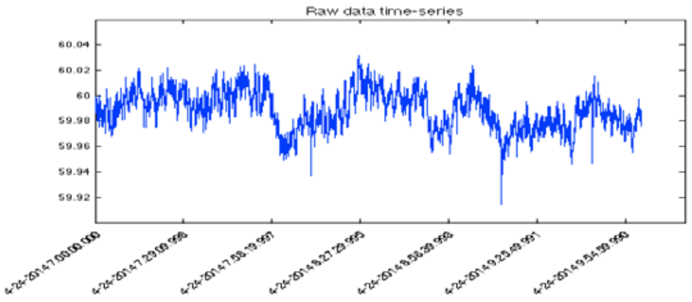
Smart Grid

- Challenge
  - Rising energy demand
  - Aging infrastructure
  - Grid Stability
- Goals
  - Situational awareness
  - Event Detection
  - Power outage prevention

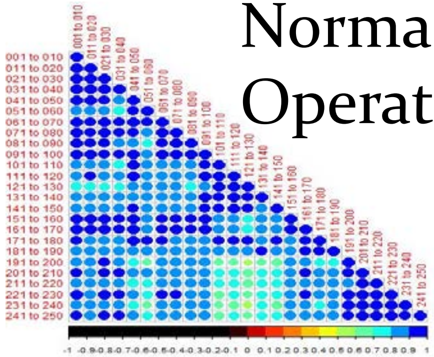
Gas Pipelines

- Challenge:
  - Increased compliance regulations
  - Rising demand
  - Aging infrastructure
- Goals
  - Improved safety
  - Asset optimization
  - Situational awareness
  - Proactive decisions
  - Resource allocation
  - Data driven risk prediction

PMU Data Analysis

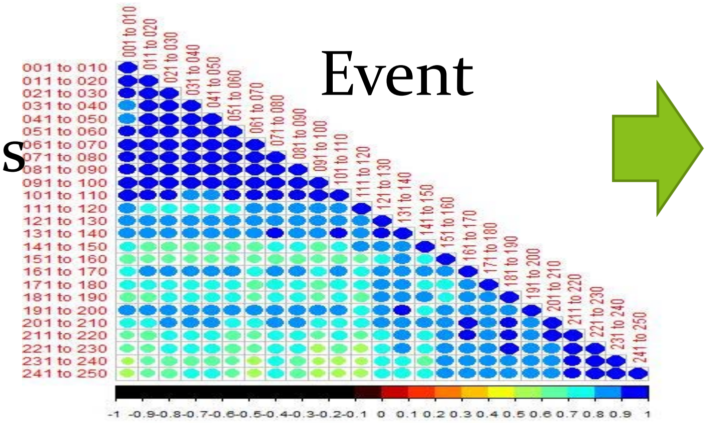


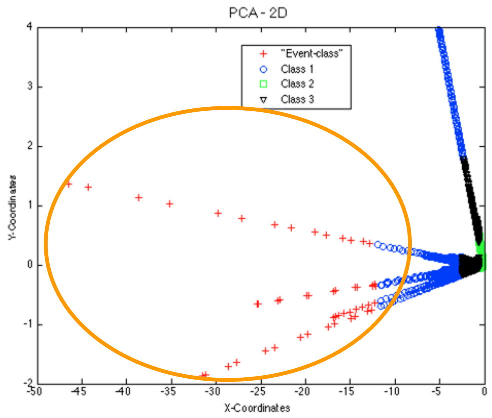
- Frequency, Magnitude, and Angle for two PMU's
- Collected 30 times a second
- Very sensitive and noisy measurements
- Goal: detect and predict event

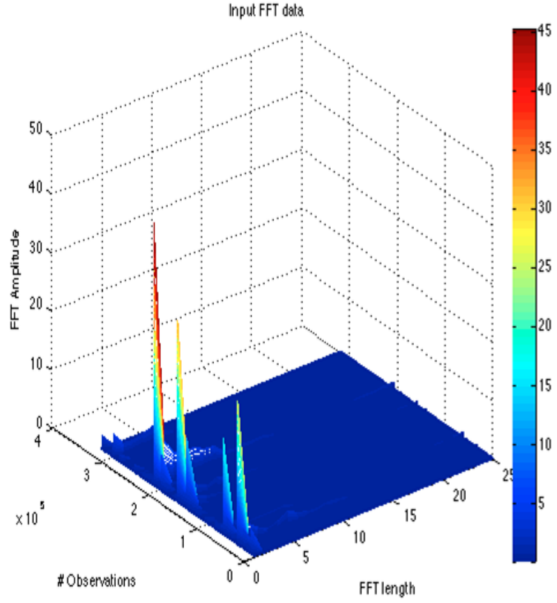


Normal Operations

Event



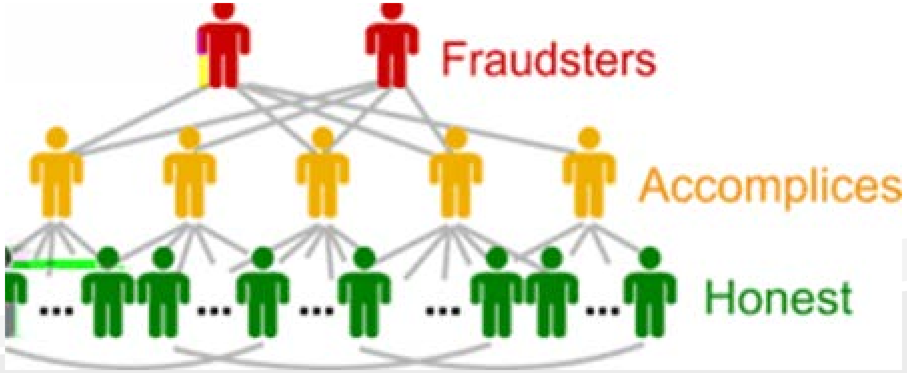




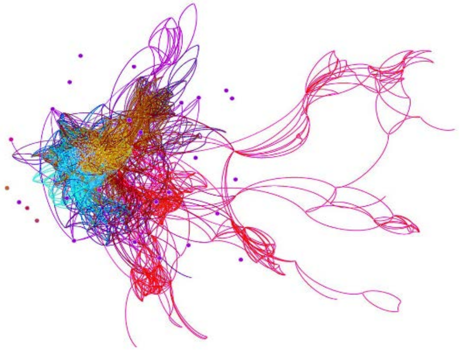
Outliers detection

Big Data for Cyber Security

- Identify risks and anticipate problems before they occur
- Traditional security mechanisms leverage rule, pattern, signature and algorithm-based approaches to detect threats
- Analyze changes in behavior and predict risks and breaches before they happen
  - Malicious Code Detection
  - Network intrusion detection
  - Anomaly detection
  - Data Stream Mining



Big Data – Big Data Science

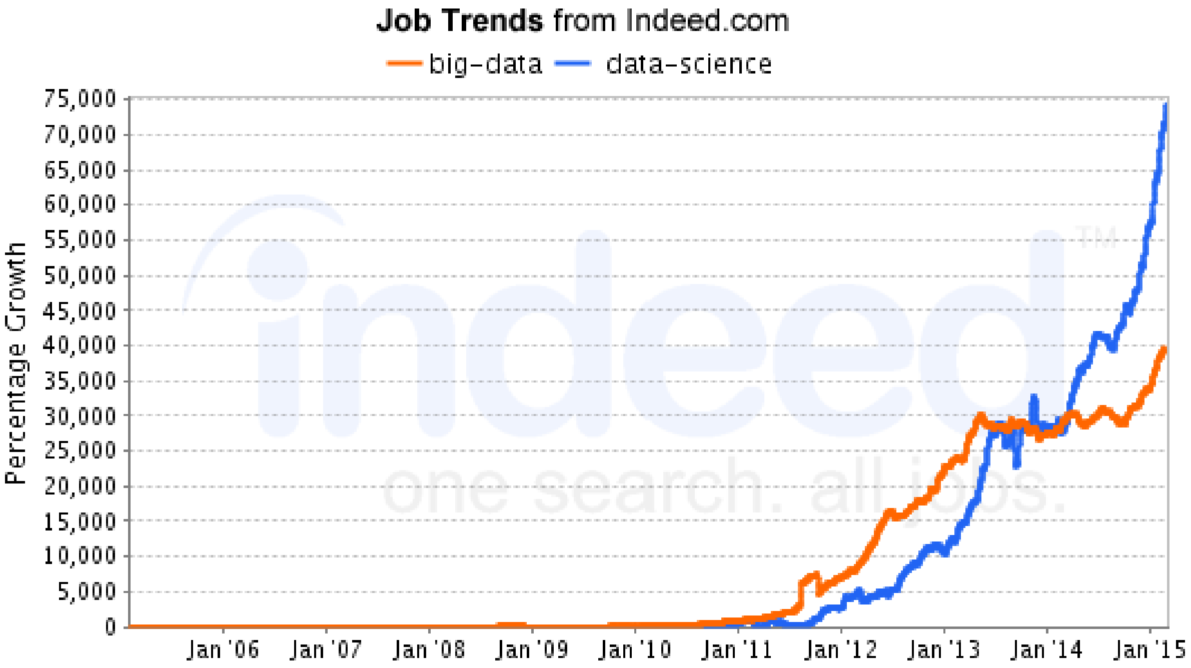


- “Data Scientist”
  - **The “Hot new gig in town”**
    - O’Reilly report
  - **Data Scientist: The Sexiest Job of the 21st Century**
    - Harvard Business Review, October 2012
    - The next sexy job in next 10 years will be statistician” – Hal Varian, Google Chief Economist
    - Geek Chic – Wall Street Journal – new cool kids on campus
  - The future belongs to the companies and people that turn data into products
- *“The human expertise to capture and analyze big data is both the most expensive and the most constraining factor for most organizations pursuing big data initiatives”* – Thomas Davenport

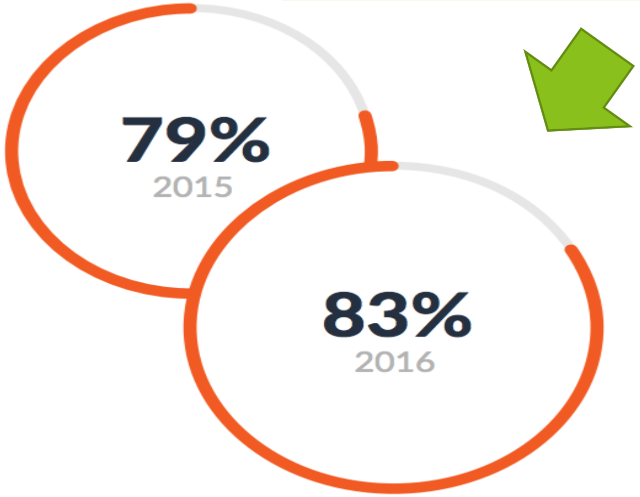
Data scientist: The hot new gig in tech

- Article in ***Fortune*** - "*nerdy-cool job that companies are scrambling to fill: data scientist*"
- Gartner in 2012 said there would be a shortage of 100,000 data scientists in the United States by 2020.
- *McKinsey Global Institute* "Big data Report" in 2011
  - By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions - demand that's 60 percent greater than supply
- In 2014, the consulting firm Accenture found that more than 90 percent of its clients planned to hire people with data science expertise
- Gartner says the current demand for data scientist exceeds the current supply by factor of three

Data Science Job Growth



Not enough Data Scientist



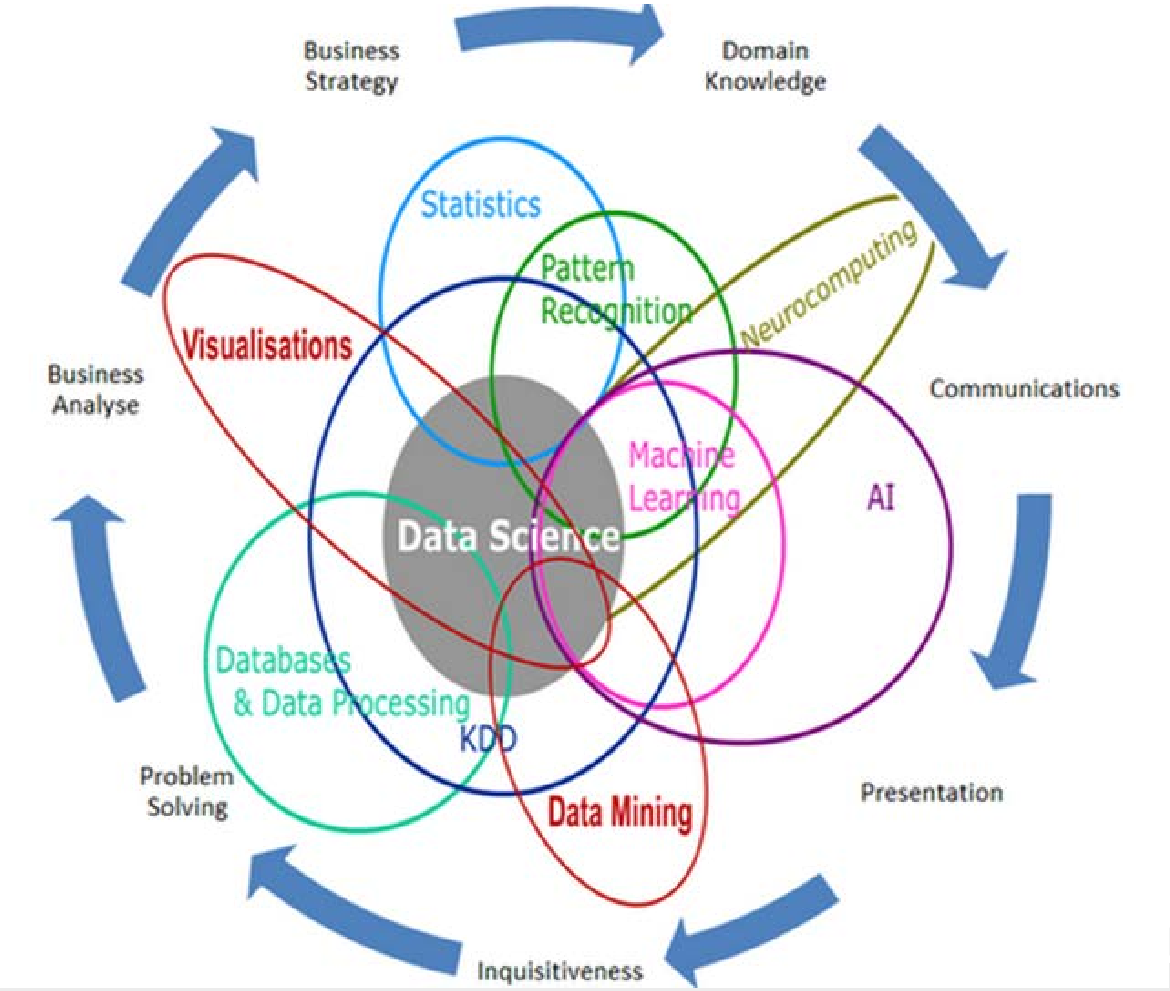
Crowdflower survey report:
A full 83% of respondents said there weren't enough data scientists to go around

By 2018 shortage of 140-190,000 predictive analysts and 1.5M managers / analysts in the US

Data Miners: Past and Present

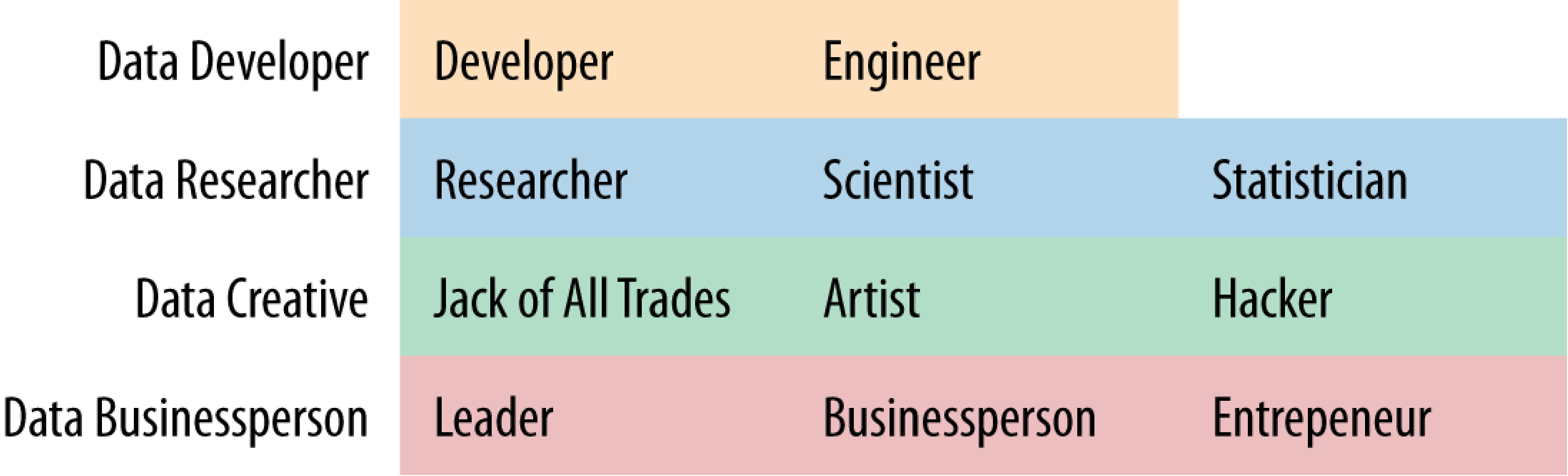
- Traditional approaches have been for DM experts: “White-coat PhD statisticians”
  - DM tools also fairly expensive
- Today: approach is designed for those with *some* Database/Analytics skills
  - DM built into DB, easy to use GUI, Workflows
  - Many jobs available from Statistical analyst to Data Scientist!
- Data Science: The Art of mathematically sophisticated data engineers delivering insights from data into business decisions and systems

Data Scientist Skill and Characteristics



- Intellectual curiosity, Intuition
  - Find needle in a haystack
  - Ask the right questions – value to the business
- Communication and engagements
- Presentation skills
  - Let the data speak but tell a story
  - Story teller – drive business value not just data insights
- Creativity
  - Guide further investigation
- Business Savvy
  - Discovering patterns that identify risks and opportunities
  - Measure

Data Scientist Self-ID

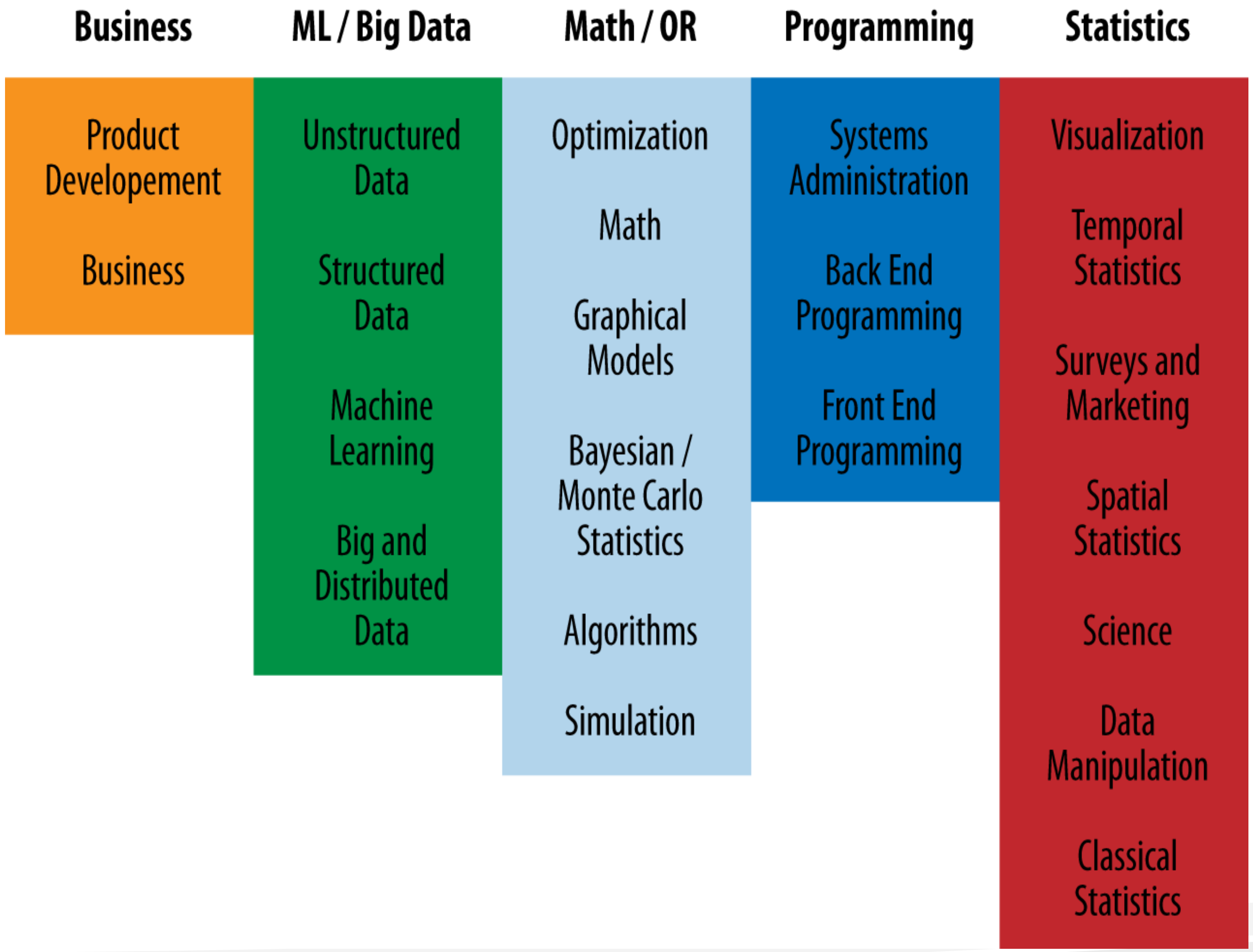


Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

O’Reilly Strata Survey suggested Self-ID Group, along with the self-ID categories most strongly associated with each Group

Strata survey, 2013

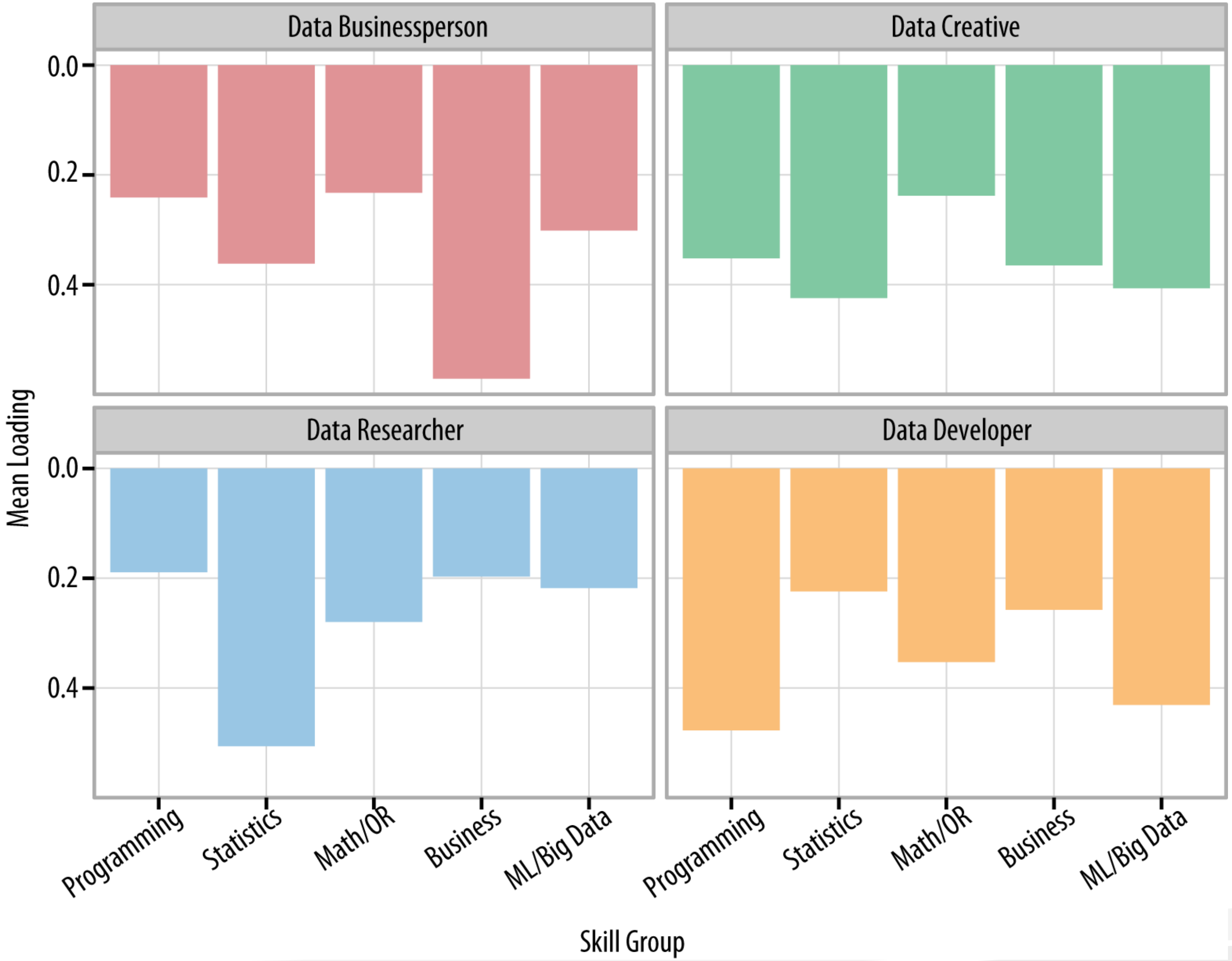
Strata Survey Skills



Business	ML / Big Data	Math / OR	Programming	Statistics
Product Developement	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian / Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics

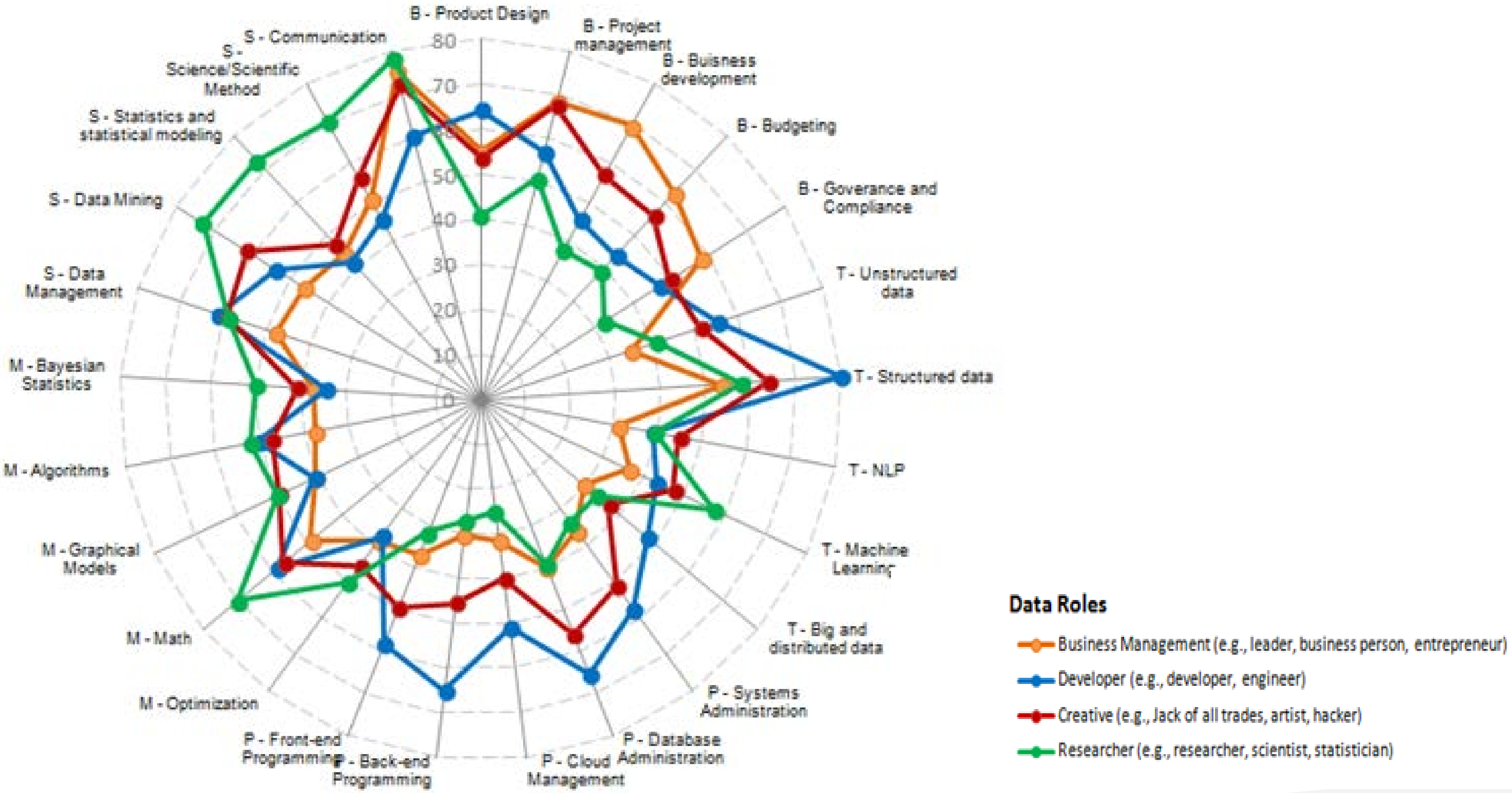
Strata survey, 2013

Strata Survey Skills



Strata survey, 2013

Strata Survey Skills

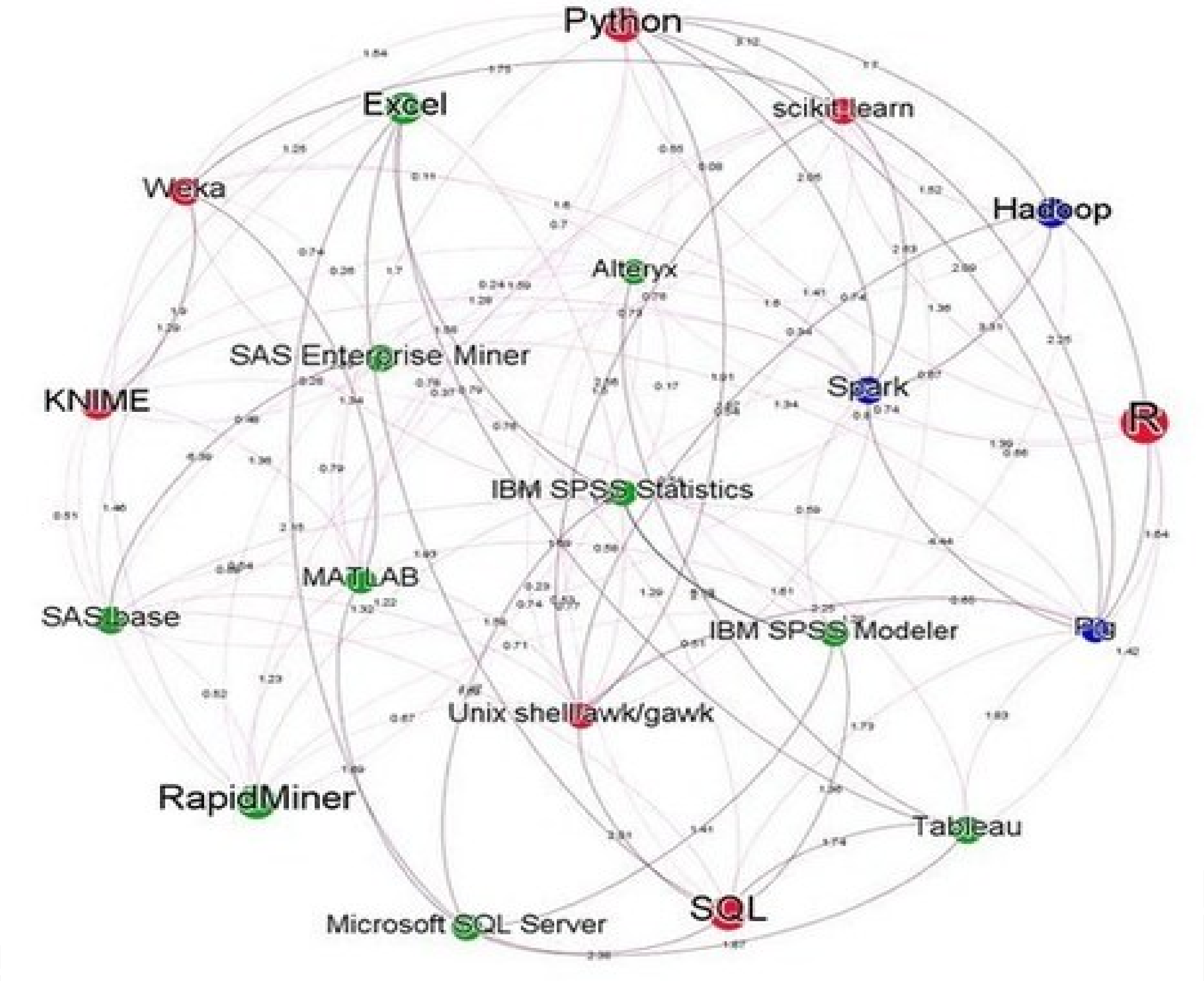


Learning and Training Opportunities

- Locally: UCSD MAS, UCSD Extension Certificate, PACE Boot Camps, etc.
- Introduction to Data Science Example
  - Part 1: Data Manipulation at scale
    - Databases and the relational algebra
    - Parallel databases, parallel query processing, in-database analytics, MapReduce, Hadoop, relationship to databases, algorithms, extensions, languages
    - Key-value stores and NoSQL; Entity resolution, record linkage
  - Part 2: Analytics, Predictive Analytics, Text mining
  - Part 3: Communicating Results
    - Visualization, data products, visual data analytics
    - Provenance, privacy, ethics, governance

https://www.coursera.org/course/datasci

World the Data Science Tools



Scoop.it

Successful Data Analytics Project Guidelines

- Don’t start a Big Data project without understanding the value – what is my ROI?
- Don’t ignore the wider Enterprise story
- A big data project is not just a technology project – it is about business change
- It pays to build the right team – Data Scientist
- Build the support structure from the start
- Build for tomorrow
- Break down the silos - involve the organization

Thank you!

natashabalac@gmail.com