Lecture 1: Data Analytics and Practicum

Benjamin Simeon Harvey (Ben)
Adjunct Faculty, CS and EMSE Department
The George Washington University
E-mail: bsharve@gwu.edu
Web: https://bsharvey.github.io/

I. Introductions

Introduction: "Cool Things”

- I work for the National Security Agency as a Cryptologic Computer Scientist
  - EOD data was 10/09 and started in the Computer Science Development Program (CDP)
- Intelligence Community National Counterintelligence and Security Professional Award
  - Data Scientist for the Corporate Disclosures Action Group
- Office of the Director National Intelligence (ODNI) Award for Human Capital
  - NSA Day of Cyber
- I played Division I football and basketball at Mississippi Valley State University
  - The alma mater of “Jerry Rice” the greatest wide receiver of all time

Introduction: My background

- Undergraduate B.S Degree in Computer Science
  - Mississippi Valley State University, Itta-Bena, MS
- Post-baccalaureate in Bioinformatics and Integrative Genomics (BIG) and a fellow at NIH Department of Clinical Research Informatics (DCRI)
  - MIT-Harvard HST, Cambridge, MA
  - NIH Clinical Center, Bethesda, MD
- M.Sc. In Computer Science, and Ph.D. in Computer Science with a focus in Bioinformatics
  - Bowie State University, Bowie, MD
- Research Assistant at Bowie State University's Biomedical Computing Lab
  - Bowie State University, Bowie, MD
- Mentors include computer scientists, enterprise architects, cyber security specialists, system engineers, & biostatiscicians (Vince Carey, Ph.D. Creator of R/Bioconductor)

Introduction (your turn!)

- Name
- Academic background
- Work experience
- What are your goals for this class?
- How do your experiences shape these goals?
- What does "Big Data" mean to you? (why are you here?)
- Something “Cool” about yourself

Agenda

I. Introductions

II. Overview of Syllabus and Course Expectations

III. Course Core Topics and Lectures: Becoming a Unicorn

   I. Big Data Foundations

   II. Big Data and Technology Concepts

   III. Big Data Analysis and Science

   IV. Advanced Big Data Analysis and Science

IV. Real World Examples

V. In-class discussion:

   I. What is Data Science (EDA and Data Science Process)?

   II. What is Big Data?

VI. Big Data Foundations

VII. Lab

II. Syllabus and Course Expectations

Course Outcomes

- Core Topic 1: Fundamentals of Data Science and Big Data
  - **Explain Big Data from a business and technology perspective**, along with an overview of common benefits, challenges, and adoption issues.
    - Data Science Roadmap and Big Data Life Cycle
    - Review of Computer Science and Linear Algebra
- Core Topic 2: Data Science & Big Data Analysis Technology Concepts
  - **Apply contemporary analysis practices, technologies and tools within Big Data environments** through programming and at a conceptual level
  - **Know the common analysis functions** and **features offered by Big Data software solutions**, as well as a high-level understanding of the **back-end components** that enable these functions.
    - Data Engineering and Data Munging
    - Data Visualizations
    - Big Data and Database Storage Technology

Course Outcomes

- Core Topic 3: Big Data Analysis and Science
  - **Knowledge of probability & statistics, modeling, and analysis techniques for data patterns, clusters, classification, and text analytics**, as well as the **identification of outliers** and **errors** that affect the significance and accuracy of predictions made on Big Data datasets.
    - Machine Learning and Classification
    - Unsupervised Learning and Regression
    - Data Analysis 1: AI and NLP, 2: Time Series, and 3: Prob & Stats and Maximum Liklihood
- Core Topic 3: Advanced Big Data Analysis and Science
  - **Apply the learned topic areas and analysis techniques to Big Data** with an **emphasis on how advanced analysis and analytics need to be carried out** individually and collectively in support of the distinct characteristics, **assess requirements and challenges associated with Big Data datasets**.
    - Stochastic Modeling and Advanced Classifiers
    - Graph Analytics and Recommender Systems

Assignments - Portfolio Project (50%)

- In this class you will create a (1) Data Science portfolio and (2) “ToolKit”
  - The goal is to consolidate learning materials, assignments, and Data Science tools within a portfolio for displaying your skillset to colleagues and potential employers and future tool use.
- Assignment 1 – Creating a Portfolio: Intro to GitHub and R/Python and EDA
- Assignment 2 - Statistical Inference
- Assignment 3 - Machine Learning
- Assignment 4 – Use the Data Science Process (DSP) to analyze a selected Big Data set and then visualize, interpret, and communicate your results in your Portfolio. Essentially, document you research project in your portfolio
  - Students will choose a dataset based upon their research interests, things that inspire them, or something we discussed in class.
  - Detail your steps in developing your solution, including how you collected the data, alternative solutions you tried, describing statistical methods you used, and the insights you obtained in your portfolio.

Student Final Project (50%)

- Final projects are to be done in either teams or as an individual
- Each team will present a talk (e.g., PowerPoint or other technique)
- Each team will produce a written document
  - 5 pages long, formatted according to IEEE style
    - Points will be withheld if the document is not formatted appropriately
- Peer evaluations will be factored in the determination of the term project grade.
- Final projects should use at least one of the techniques covered in class and from the assignments
- Final projects will be placed in your final portfolio (as well as Assignment 4)
- This is a significant part of your grade. Choose your problem early.

Student Final Project Format (50%)

Research Proposal

- Overview, and Motivation: Provide an overview of the project goals and the motivation for it. Consider that this will be read by people who did not see your project proposal.
- Related Work: Anything that inspired you, such as a paper, a web site, or something we discussed in class.
- Initial Questions: What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis? - Data: Source, scraping method, cleanup, storage, etc.

Final Paper (also includes Abstract, Overview, Motivation, Related Work, Initial Questions and Data from the Research Proposal):

- Exploratory Data Analysis: What visualizations did you use to look at your data in different ways? What are the different statistical methods you considered? Justify the decisions you made, and show any major changes to your ideas. How did you reach these conclusions?
- Final Analysis: What did you learn about the data? How did you answer the questions? How can you justify your answers?
- Presentation: Present your final results in a compelling and engaging way using text, visualizations, images, and videos on your project web site.

Academic Integrity

- Plagiarism is a direct violation of the GW Code of Academic Integrity which you and I have both agreed to.
- Course policy:
  - 1st violation of academic integrity will result in a warning consisting of a loss of points equal to the value of the plagiarized assignment.
    - A plagiarized homework assignment will be given a score of -5% of your total grade (equivalent to getting zero on two assignments)
    - A plagiarized exam is equivalent to a zero on two exams.
    - A plagiarized final project is equivalent to -40% of your grade and is a practical failure of the course.
  - 2nd violation of academic integrity will result in a complaint filed in the Office of Academic Integrity with a recommended MINIMUM sanction of failure of the course.
  - No exceptions
  - If you are unsure if your work is plagiarized, ask me before you turn it in.
- Plagiarism will not be tolerated.
  - If you feel that you have been incorrectly accused of plagiarism, you may appeal to the GW Committee on Academic Integrity and a hearing will be held.

II. Becoming a Unicorn

Data Analysis Has Been Around for a While

1935: "The Design of Experiments"

R.A. Fisher

1939: "Quality Control"

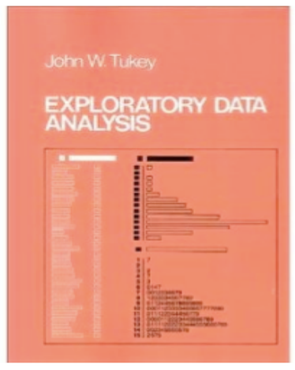
W.E. Demming

1958: "A Business Intelligence System"

Peter Luhn

1997: "Machine Learning"

1977: "Exploratory Data Analysis"

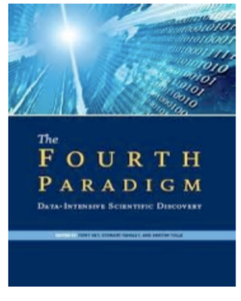


1989: "Business Intelligence"

Howard Dresner

2010:"The Data Deluge"

2007:"The Fourth Paradigm"



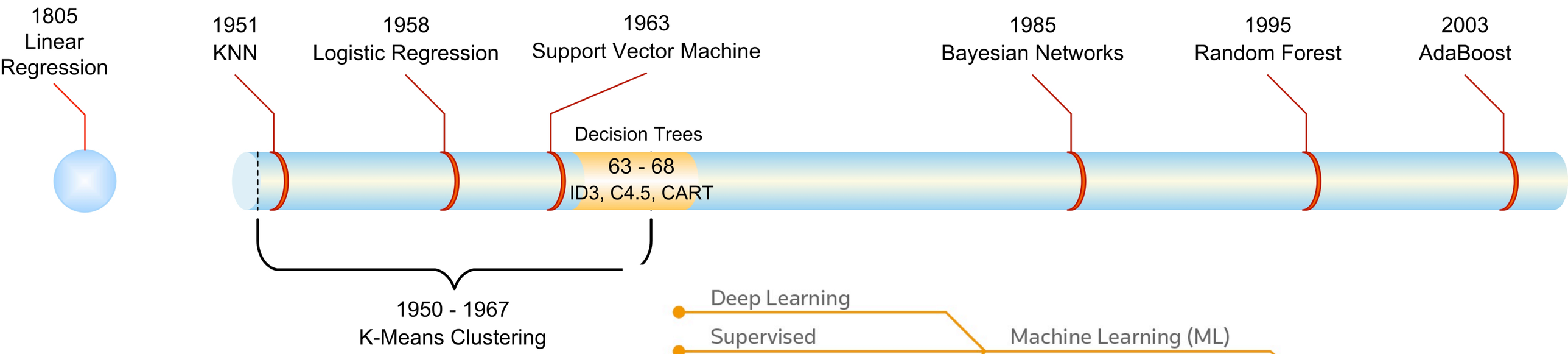
2009:"The Unreasonable Effectiveness of Data"

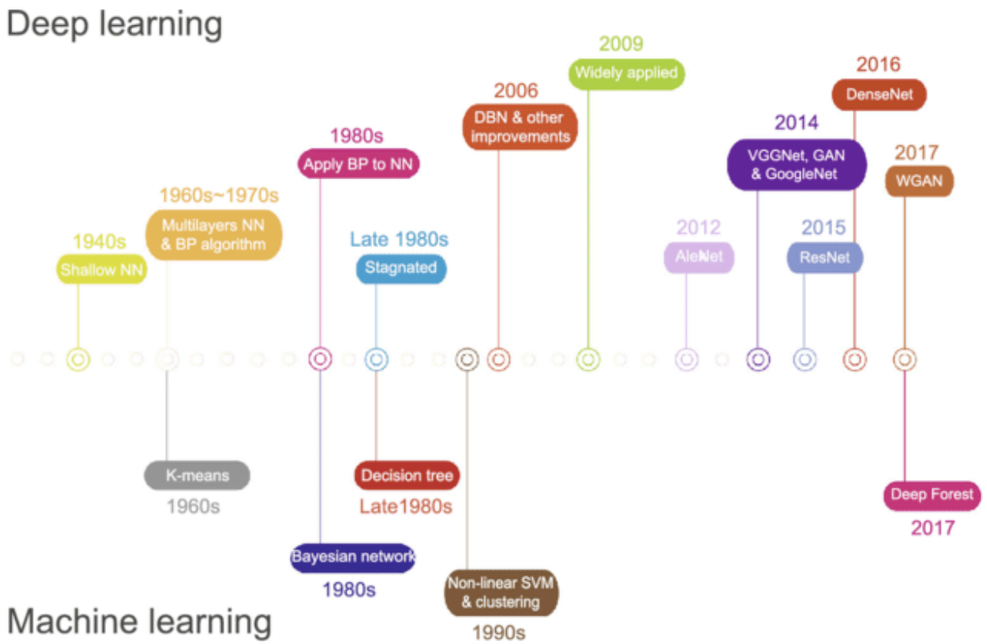
1996: Google

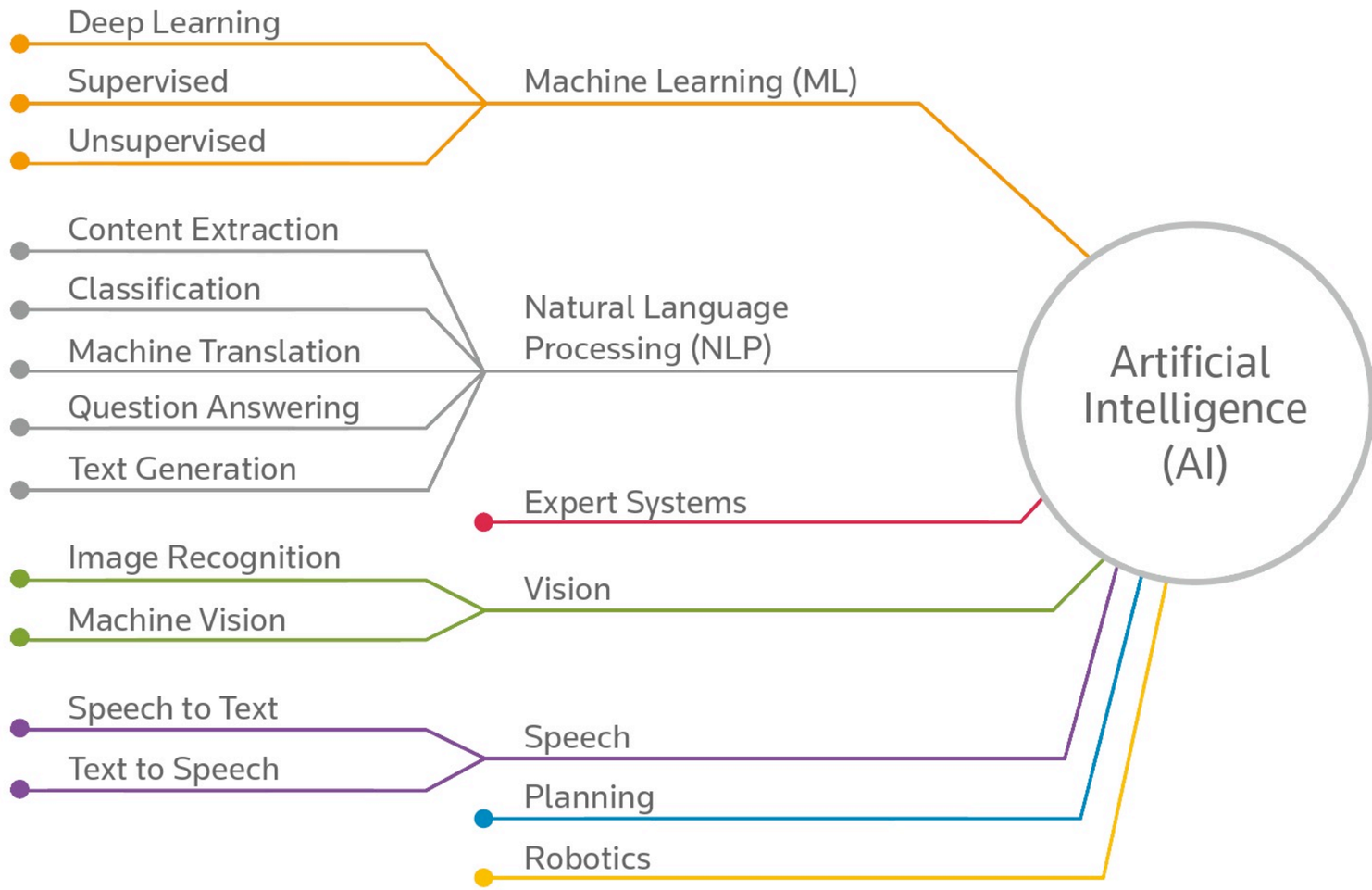
Google™

Abridged Version of Jeff Hammerbacher's timeline for CS 194, 2012

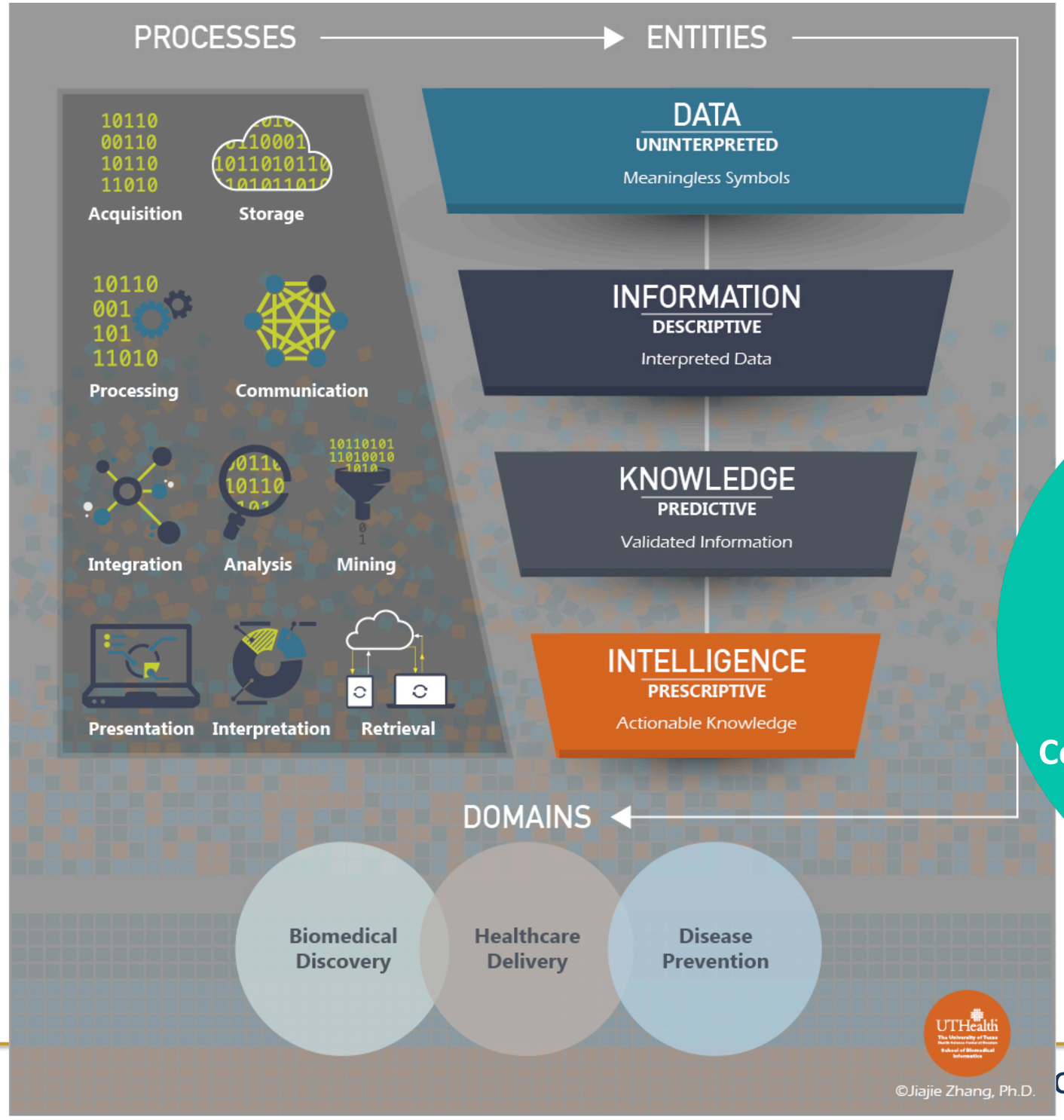
Where are we headed?

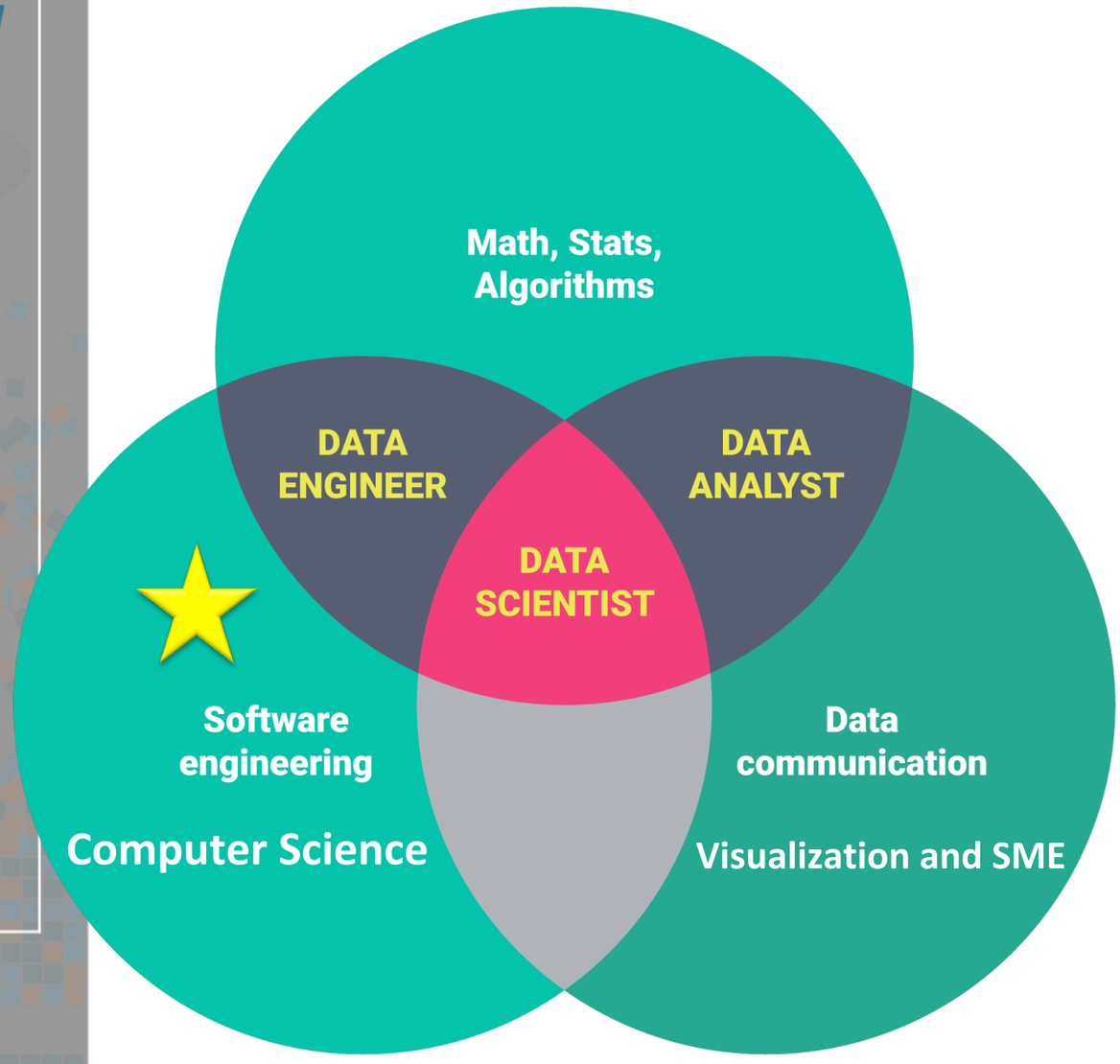






Statistician/Informatician Vs. Data Scientist





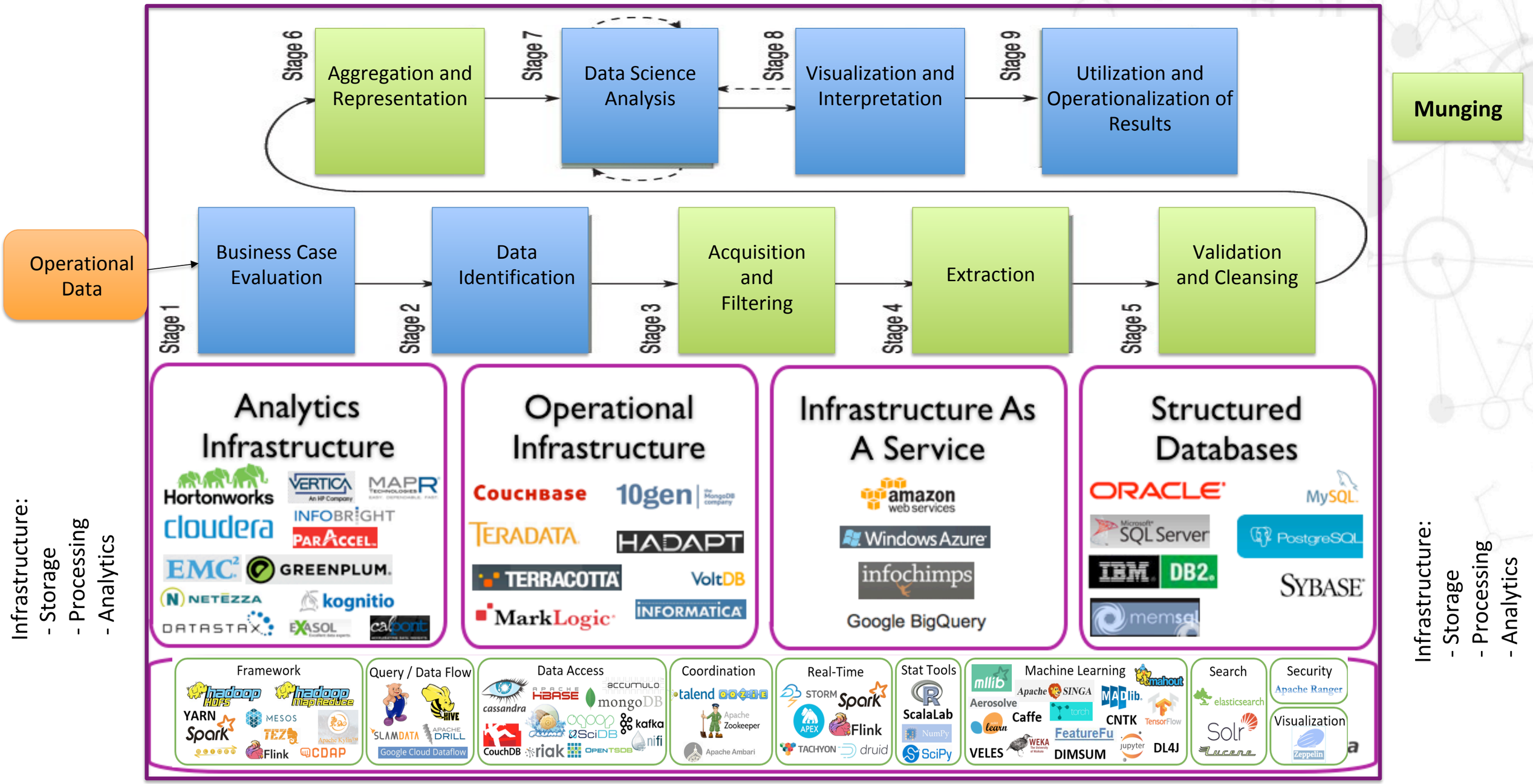
Becoming A Unicorn

- Data science means doing analytics work that, for one reason or another, requires a substantial amount of computer science (CS) and software engineering skills.
- It's very hard to find people who can construct good statistical models, understand software/programming languages/architecture/frameworks, and relate this all in a meaningful way to business problems.
- Integration: A unicorn can take a data science product and integrate it into existing or future architecture.

Steps to Becoming a Unicorn

1. Understand each stage of Data science process and Big Data Analytics life cycle
2. Principles of Programming Languages
3. Understand Software Engineering Life Cycle
4. Understand Data Engineering and Munging
5. Become an expert in Data Analytics, modeling, ML, AI
6. Start exploring advanced analytics
7. Understand Big Data analytics and Architectures

Big Data Analytics Life Cycle



IDC Estimate
40 ZETABYTES
Of Data by 2020

2.5 QUINTILLION BYTES
of Data are Created each Day

Facebook has
1 PETTA BYTES
Of Storage

VOLUME
Scale of Data

Hadron Collider generates
1 PETA BYTES
Of Data are create per year

Estimated
100 TERA BYTES
Of Data per US Company

Estimated
6 BILLION PEOPLE
Have a cell Phone

Big Data

200 BILLIONS PHOTOS

Global Healthcare data
150 EXABYTES
2.4 EXABYTES per year
Growth

VARIETY
Different Forms of Data

100 MILLION VIDEO
600 Years of Video
13 Hours of video uploaded per minute

500 MILLION TWEETS
Per day

1.8 BILLION SMARTPHONES

VELOCITY
Analysis of Data

20 BILLION NETWORK CONNECTIONS
By 2016

NY Stock Exchange generates
1 TERRA BYTES
Of Trade Data per day

Analytics

MAP

REDUCE

RESULT

Poor Data Quality cost businesses
600 BILLION A YEAR

1 IN 3 LEADERS
Don't trust the information they use to make decision

30% OF DATA COLLECTED
By marketers are not usable for real-time decision making

VERACITY
Uncertainty of Data

Poor data across business and the government costs the US economy
3.1 TRILLION DOLLARS
a year

III. Motivational Examples

Sports Big Data Analytics

"Please help us identify opportunities to impact decision making by modeling and understanding field goal percentage. Select an opportunity and tailor your analysis accordingly. You will be provided a dataset on which you should base your work. The accompanying **15 GB file contains 100,000,000 rows**, each corresponding to a shot attempt. There are 9 fields: **shot_x_location, shot_y_location, shooter_velocity_ft_set, shooter_angle, dribbles_before, made, defender_distance, defender_angle, defender_velocity_ft_sec**. Please help by submitting unique solutions. Include and describe a visualization and deliver the code through one of the appropriate cloud storage platforms below."

Technical Analysis of Field Goal Shooting:
Fundamentals and Suggestions

Justin Jacobs

3 December 2016

Abstract

In this document, we analyze a collection of 10,000 field goal attempts made during NBA games. We take a look at the spatial distributions conditional on factors such as player movement leading into the shot, defender movement relative to the shooter, and quality of the shot location given by rate of success against value of the shot. Given this analysis, we can make small recommendations on strategizing optimal field goal attempts. Code is provided in supplementary materials.

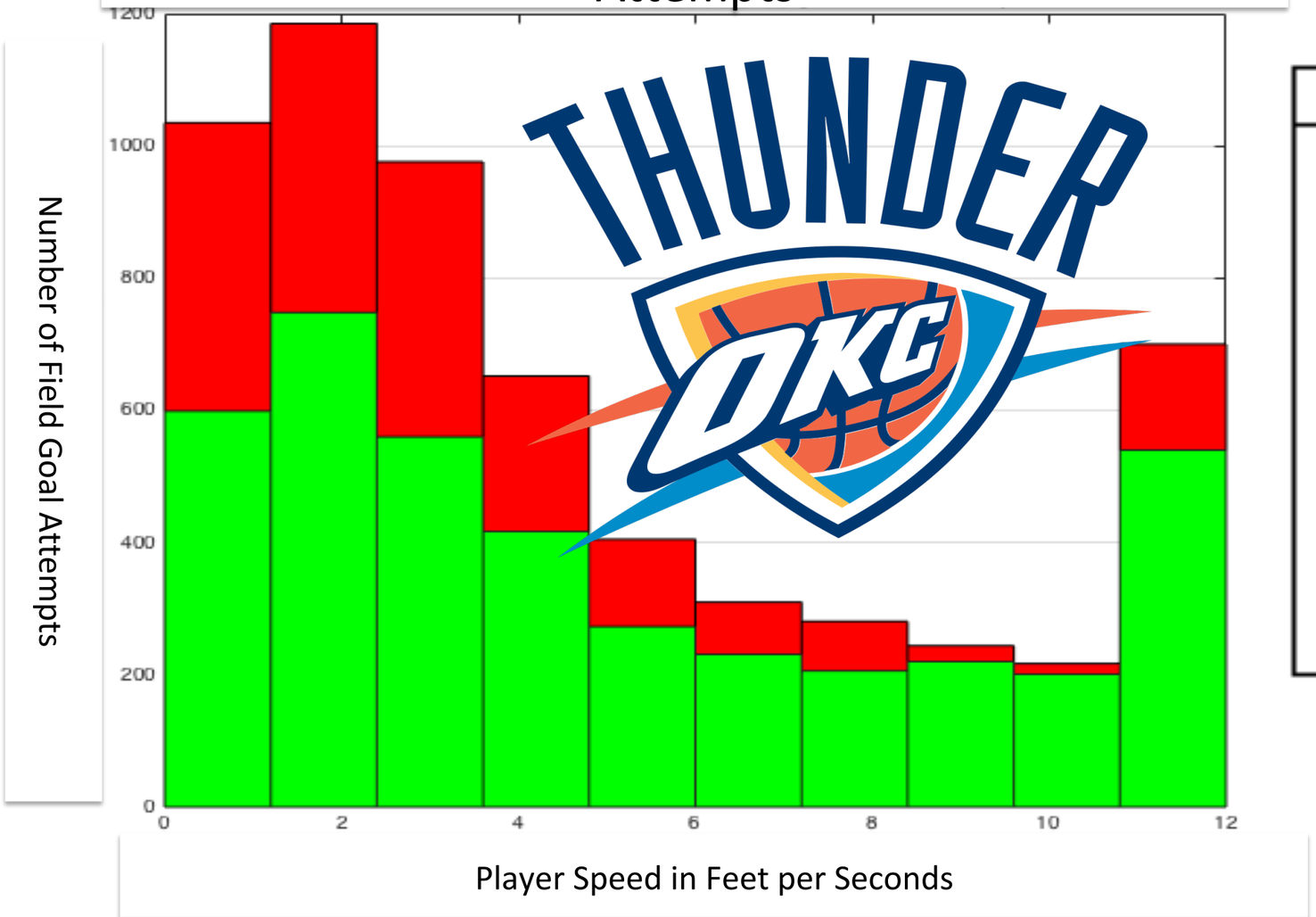
1 Prelude: Understanding the Data

The available dataset is a collection of 10,000 field goal attempts; each with ten measurements:

Variable	Columns	Description
Shooter Location	0-1	The (x, y) position of the shooter during a field goal attempt.
Shooter Velocity	2-3	The speed and direction of the shooter heading into a field goal attempt.
Dribbles	4	The number of dribbles the shooter takes before a field goal attempt.
Result	5	Indicator of a Made or Missed attempt.
Defender Location	6-7	The distance and relative position of the defender from the shooter during a field goal attempt.
Defender Velocity	8-9	The speed and direction of the defender, relative to the shooter, during a field goal attempt.

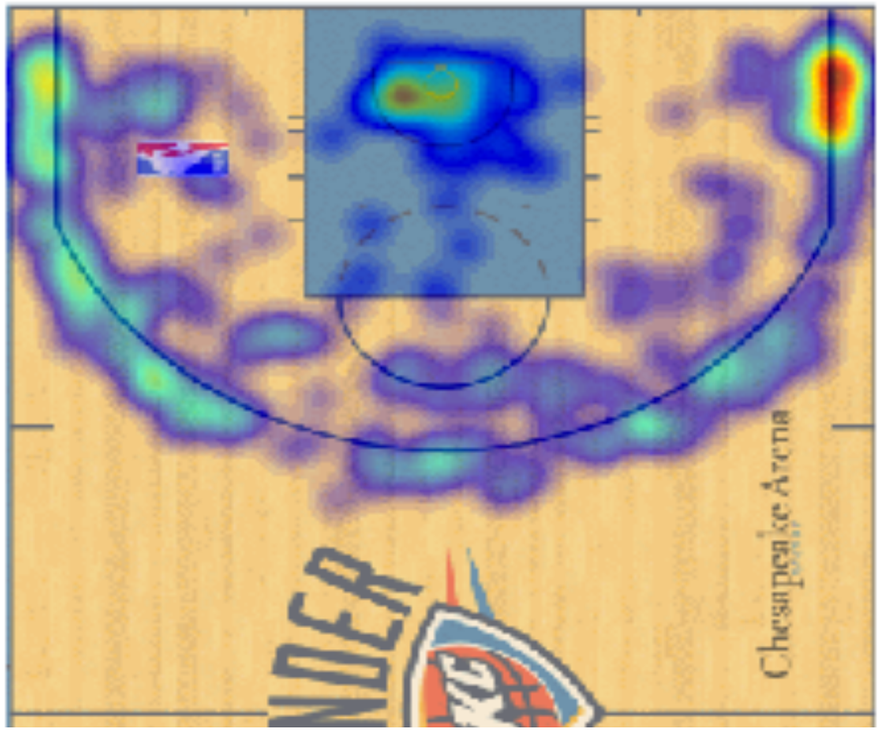
Sports Big Data Analytics

Distribution of Shooter Velocities during Field Goal Attempts

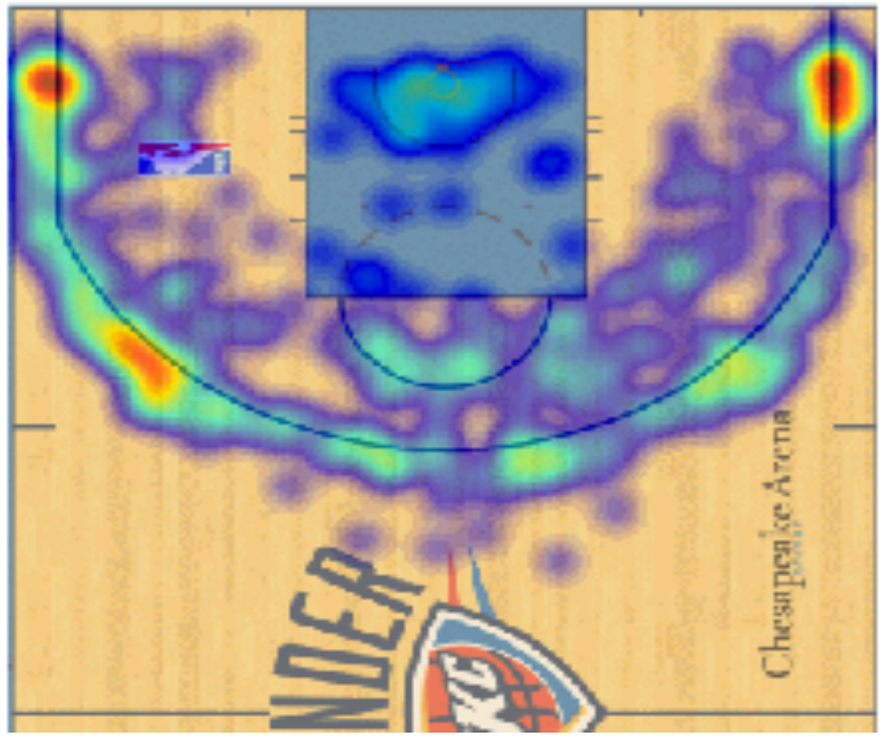


Speed (fps)	FGM	FGA	FGP
0 - 1.2	599	1634	36.6585
1.2 - 2.4	748	1933	38.6963
2.4 - 3.6	560	1536	36.4583
3.6 - 4.8	417	1069	39.0084
4.8 - 6.0	273	678	40.2655
6.0 - 7.2	231	541	42.6987
7.2 - 8.4	206	487	42.2998
8.4 - 9.6	220	464	47.4138
9.6 - 10.8	201	418	48.0861
10.8 - 12.0	54	124	43.5484

Distribution of Spot-Up Catch-and-Shoot Makes



Distribution of Spot-Up Catch-and-Shoot Misses



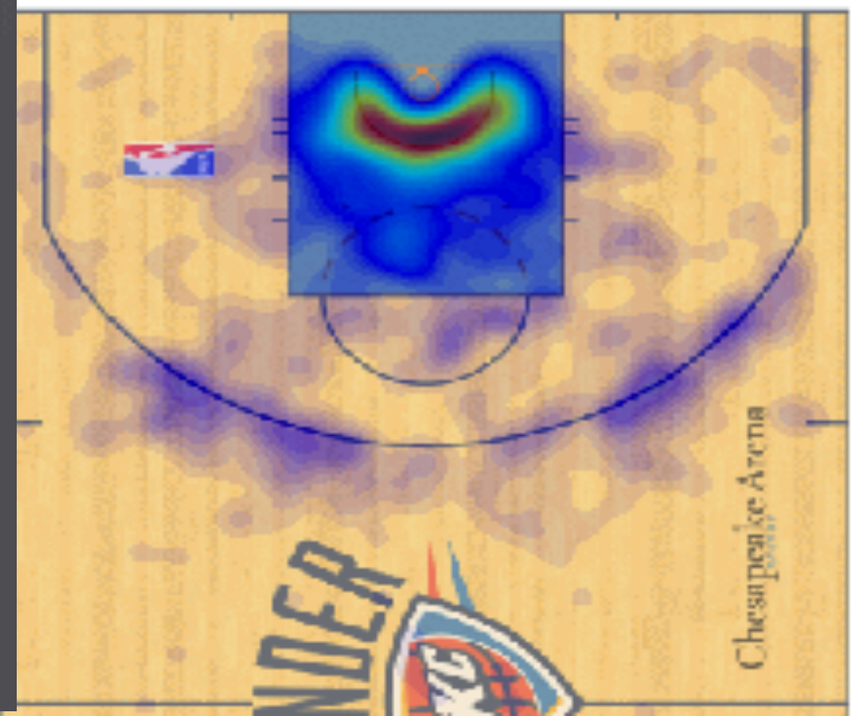
1.	WASHINGTON D.C.	KEVIN DURANT	CARMELO ANTHONY	TY LAWSON	ROY HIBBERT	RUDY GAY
2.	L.A. CLIPPERS	RUSSELL WESTBROOK	JAMES HARDEN	KAWHI LEONARD	PAUL GEORGE	DERRICK WILLIAMS
3.	CHICAGO	DERRICK ROSE	DWYANE WADE	ANTHONY DAVIS	JABARI PARKER	TONY ALLEN
4.	NEW YORK	JOAKIM NOAH	KYRIE IRVING	KENNETH FARIED	KEMBA WALKER	J.R SMITH
5.	DALLAS	CHRIS BOSH	LAMARCUS ALDRIDGE	DERON WILLIAMS	MARCUS SMART	C.J. MILES
6.	ATLANTA	DWIGHT HOWARD	JOSH SMITH	DERRICK FAVORS	LOUIS WILLIAMS	JODIE MEEKS
7.	L.A. LAKERS	DEMAR DEROZAN	TYSON CHANDLER	KLAY THOMPSON	BRANDON JENNINGS	TREVOR ARIZA
8.	INDIANA	ZACH RANDOLPH	MIKE CONLEY	JEFF TEAGUE	GORDON HAYWARD	JOSH MCROBERTS
9.	CLEVELAND	LEBRON JAMES	KEVIN MARTIN	BYRON MULLENS	JARED SULLINGER	KOSTA KOUFOS
10.	PHILADELPHIA	KOBE BRYANT	TYREKE EVANS	KYLE LOWRY	MARKIEFF MORRIS	MARCUS MORRIS
11.	CHARLOTTE	CHRIS PAUL	STEPH CURRY	ISH SMITH	ANTHONY MORROW	ANTAWN JAMISON
16.	ORLANDO	CHANDLER PARSONS	AMAR'E STOUDEMIRE	VINCE CARTER	NICK CALATHES	MARREESE SPEIGHTS
17.	HOUSTON	DEANDRE JORDAN	JIMMY BUTLER	GERALD GREEN	RASHARD LEWIS	KENDRICK PERKINS
18.	PORTLAND	KEVIN LOVE	TERRENCE JONES	TERRENCE ROSS	KYLE SINGLER	LOCAL PLAYER #1
19.	DETROIT	JORDAN CRAWFORD	DRAYMOND GREEN	JAVALE MCGEE	SHANE BATTIER	CHRIS DOUGLAS-ROBERTS
20.	BOSTON	MICHAEL CARTER-WILLIAMS	NOAH VONLEH	NERLENS NOEL	MATT BONNER	SHABAZZ NAPIER
21.	MIAMI	BRANDON KNIGHT	TIM HARDAWAY JR.	STEVE BLAKE	UDONIS HASLEM	ALONZO GEE
22.	MILWAUKEE	WES MATTHEWS	MIKE DUNLEAVY	CARON BUTLER	DEVIN HARRIS	STEVE NOVAK
23.	PHOENIX	CHANNING FRYE	JERRYD BAYLESS	GREG SMITH	RICHARD JEFFERSON	SHANE EDWARDS
24.	MINNESOTA	KRIS HUMPHRIES	JON LEUER	ALAN ANDERSON	NATE WOLTERS	COLE ALDRICH
25.	SACRAMENTO	MATT BARNES	RYAN ANDERSON	JAMES NUNNALLY	ROBIN LOPEZ	BROOK LOPEZ
26.	MEMPHIS	COREY BREWER	THADDEUS YOUNG	SHAWNE WILLIAMS	ELLIOT WILLIAMS	JARVIS VARNADO
27.	OKLAHOMA CITY	BLAKE GRIFFIN	XAVIER HENRY	DANIEL ORTON	EKPE UDUH	LOCAL PLAYER #1
28.	DENVER	REGGIE JACKSON	JASON SMITH	LOUIS AMUNDSON	CHAUNCEY BILLUPS	LOCAL PLAYER #1
29.	UTAH	BRANDON DAVIES	JUSTIN HAMILTON	LOCAL PLAYER #1	LOCAL PLAYER #2	LOCAL PLAYER #3
30.	SAN ANTONIO	ANDRE ROBERSON	LOCAL PLAYER #1	LOCAL PLAYER #2	LOCAL PLAYER #3	LOCAL PLAYER #4

D.J. AUGUSTIN		
MAURICE HARKLESS		
CHUCK HAYES		

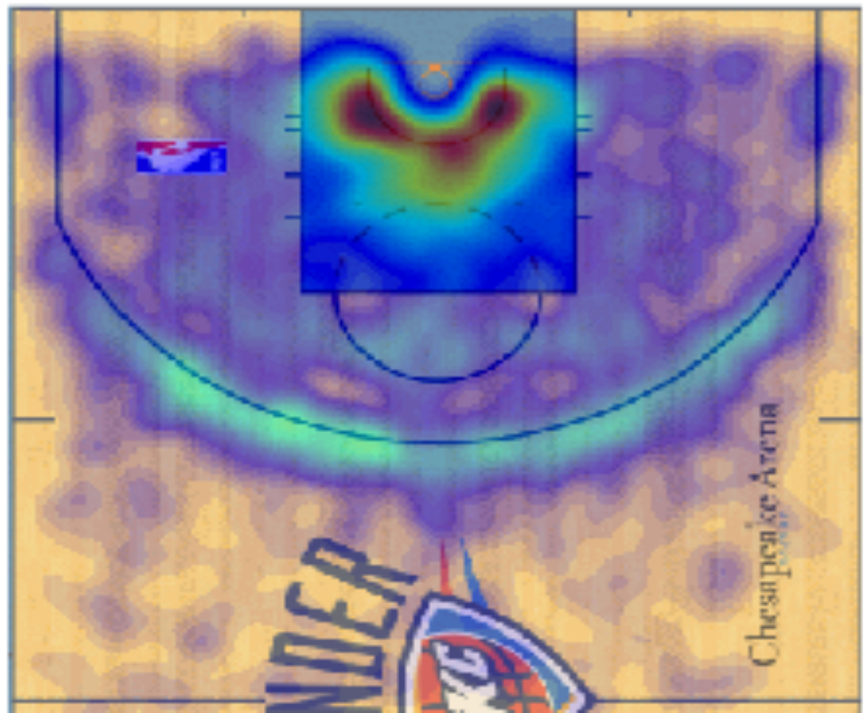
Source: Deadspin

BUSINESS INSIDER

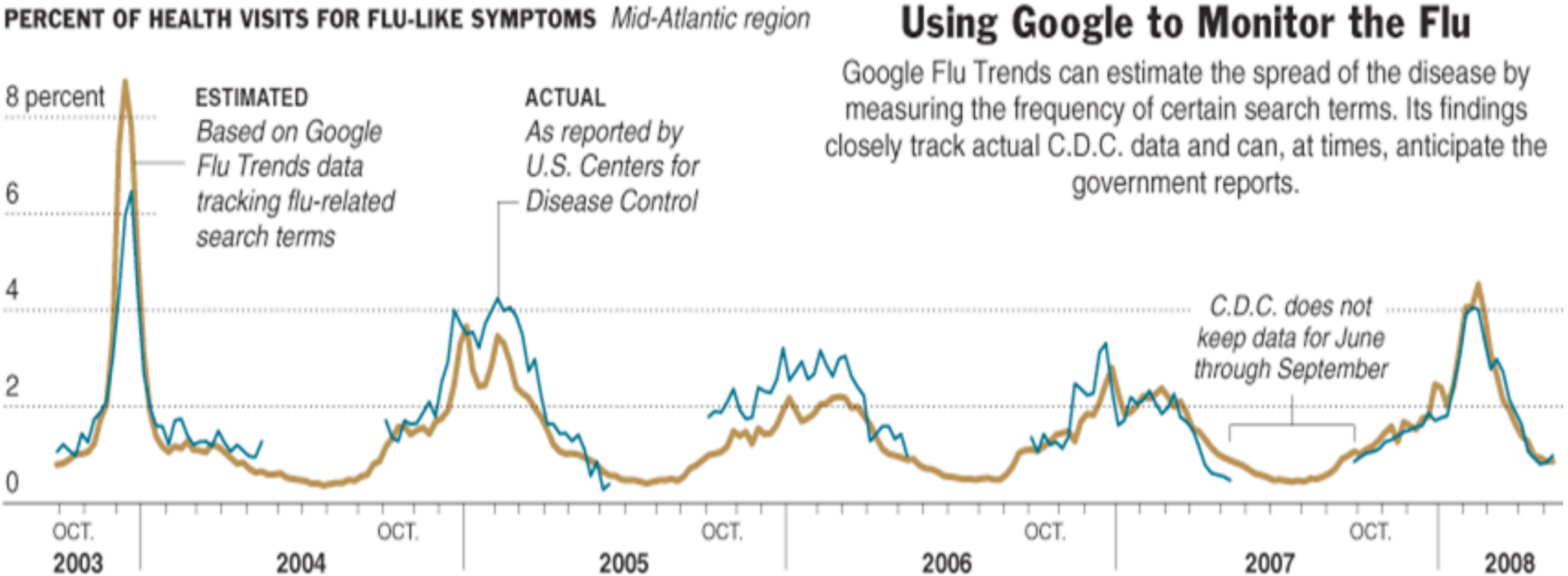
Distribution of Made Pull-Up Jumpers Outside of Three Feet



Distribution of Missed Pull-Up Jumpers Outside of Three Feet



Using Google to Monitor the Flu



Using Google to Monitor the Flu

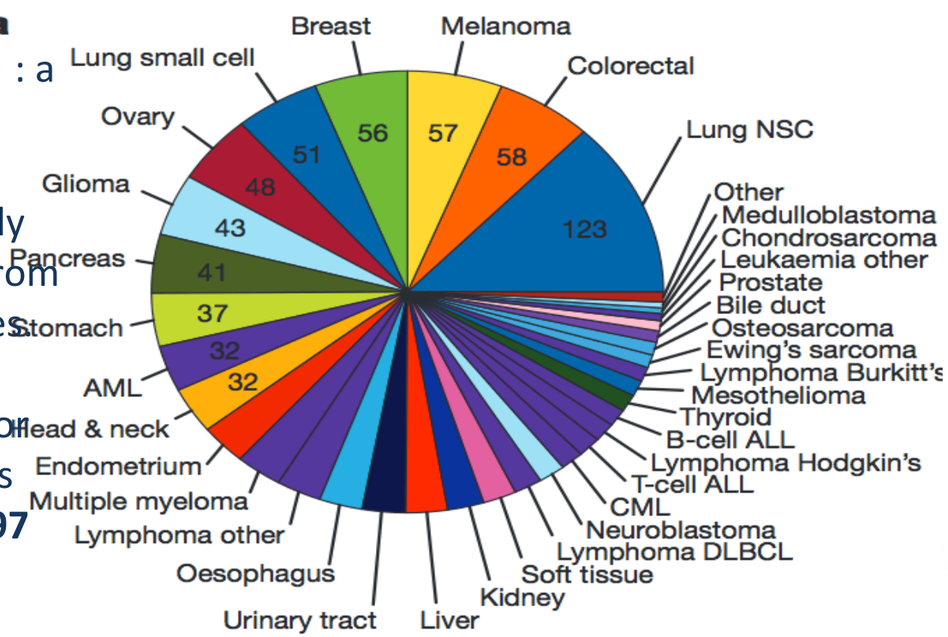
Google Flu Trends can estimate the spread of the disease by measuring the frequency of certain search terms. Its findings closely track actual C.D.C. data and can, at times, anticipate the government reports.

Large-Scale Cancer Genomics (10MB per Sample – 10's of GBs per dataset)

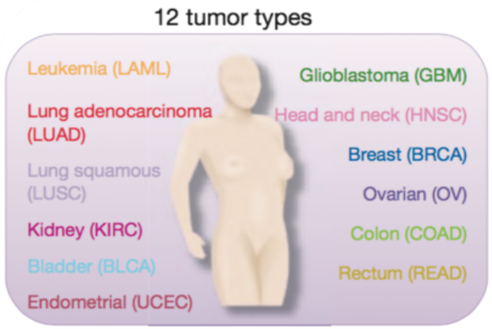
Global Cancer Map (GCM) Data: 218 tumor samples, spanning **14** common tumor types, and **90** normal tissue samples to oligonucleotide microarray gene expression analysis:
16,063 genes

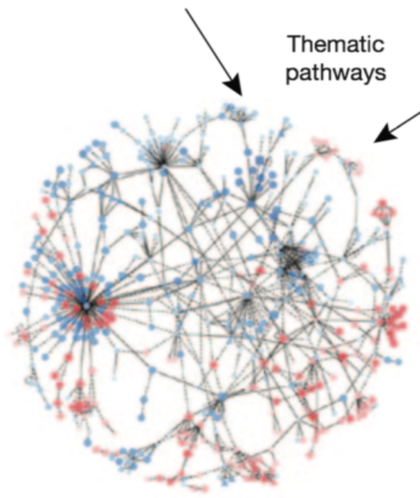
218 Tumor Samples and 16,6063 genes

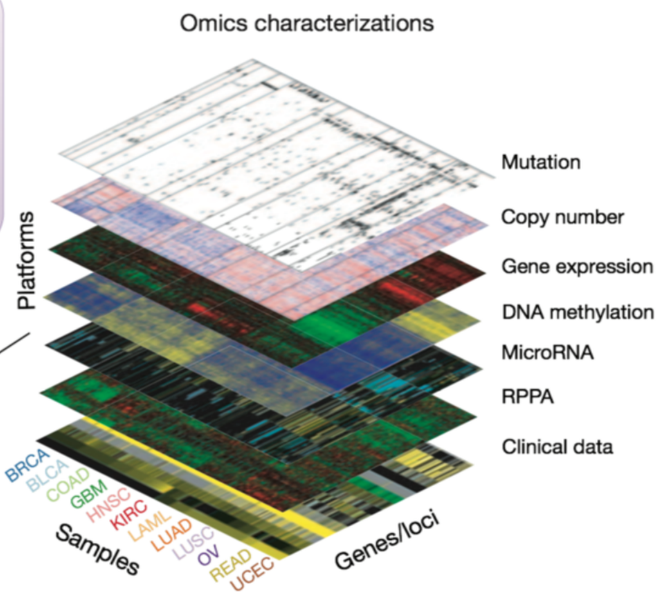
Cancer Cell Line Encyclopedia (CCLE) : a compilation of gene expression, chromosomal copy number and massively parallel sequencing data from **947** human cancer cell lines coupled with pharmacological profiles for **24** anticancer drugs across **479** of the cell lines: **18,897 genes**



The Cancer Genome Atlas: DNA methylation is an epigenetic mark which can be associated with transcriptional inactivity when located in promoter regions. Ovarian cancer study for gene expression and methylation correlated values for **22,000 genes across 598 samples**





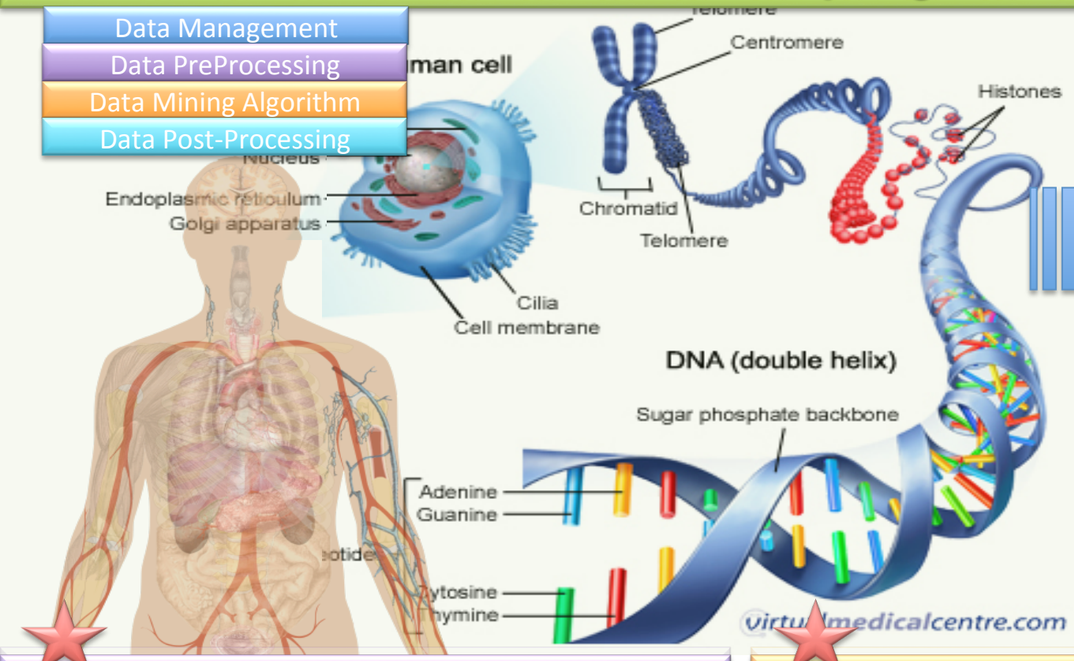


GENOMIC SIGNAL PROCESSING ANALYSIS

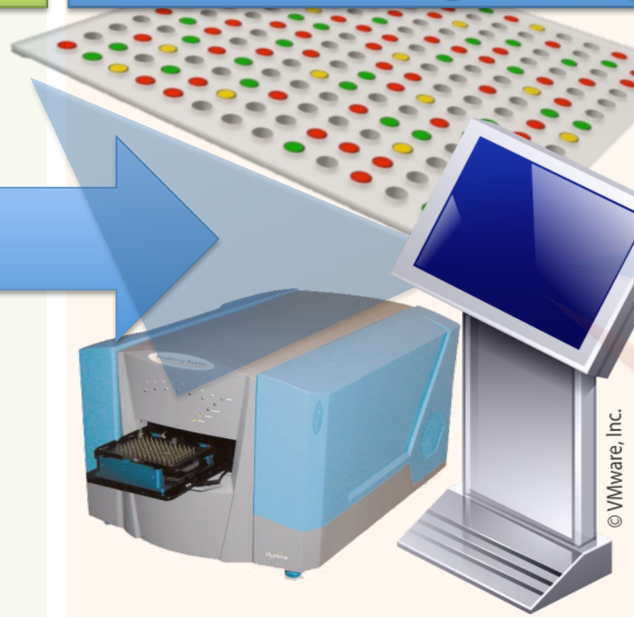
Genomic Signal Processing: is a **genomic data mining process** which involves uncovering patterns, associations, anomalies, and statistically significant structures and events in genetics and genomics data

Data Fusion and Sampling

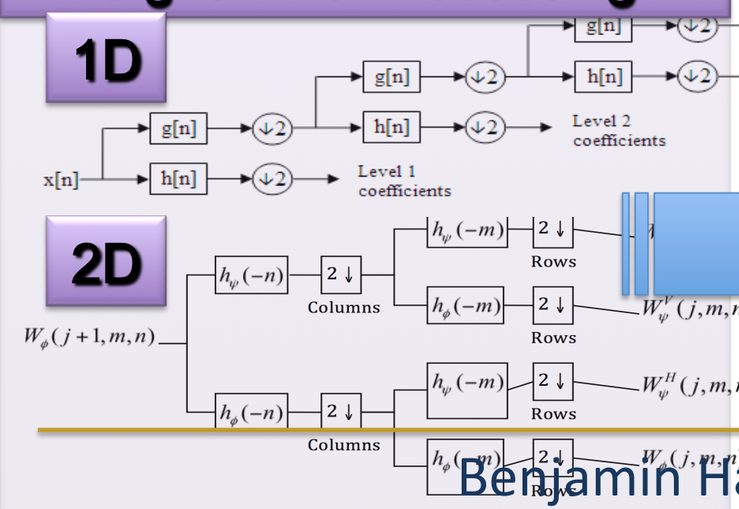
Data Management
Data PreProcessing
Data Mining Algorithm
Data Post-Processing



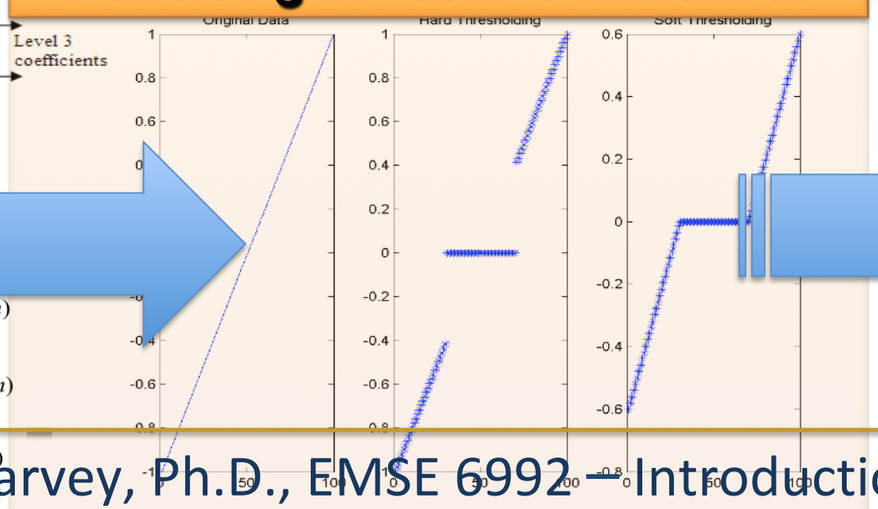
Data Storage, Integration, and Representation



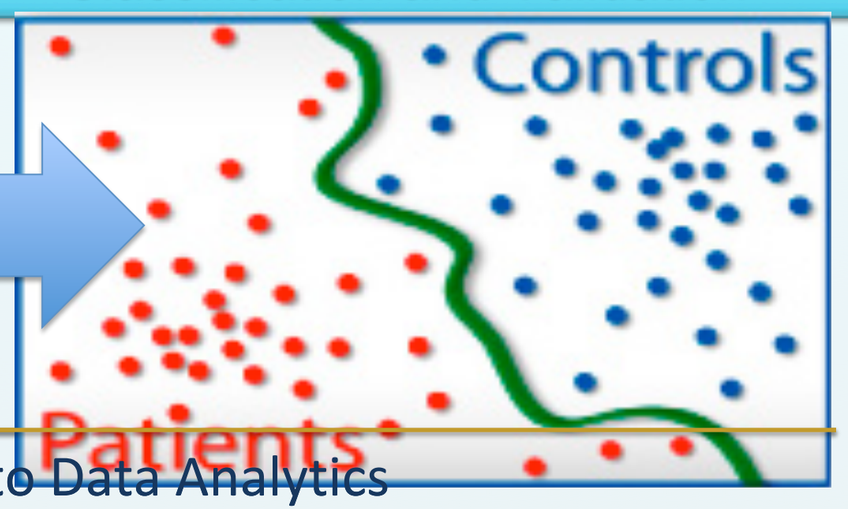
Signal Pre-Processing



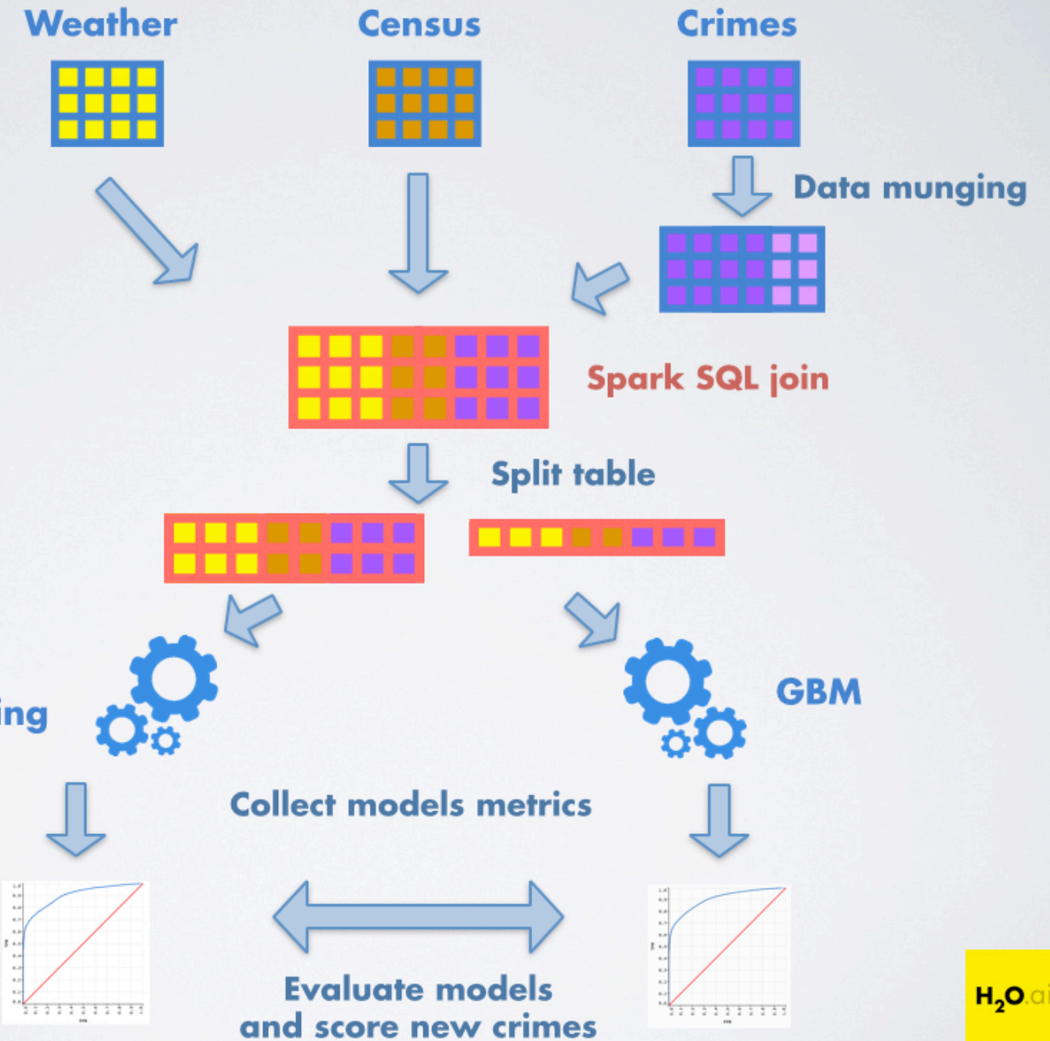
Denoising & Feature Extraction

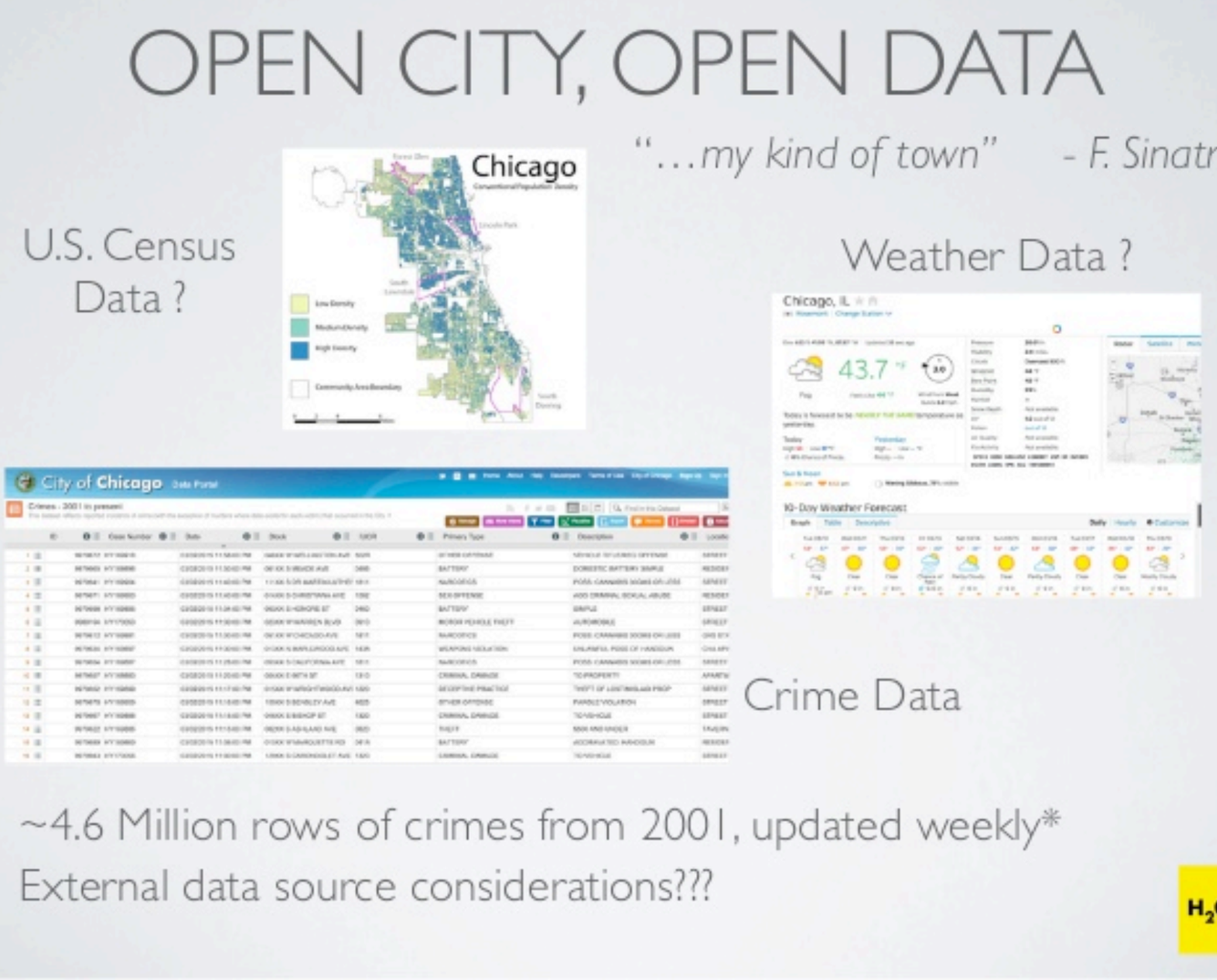


Classification and Validation



H_2O ai and Crime Prediction





deep learning
machine learning
H2O.ai
Popular Posts

H2O.ai Raises $20M For Its Open Source Machine Learning Platform

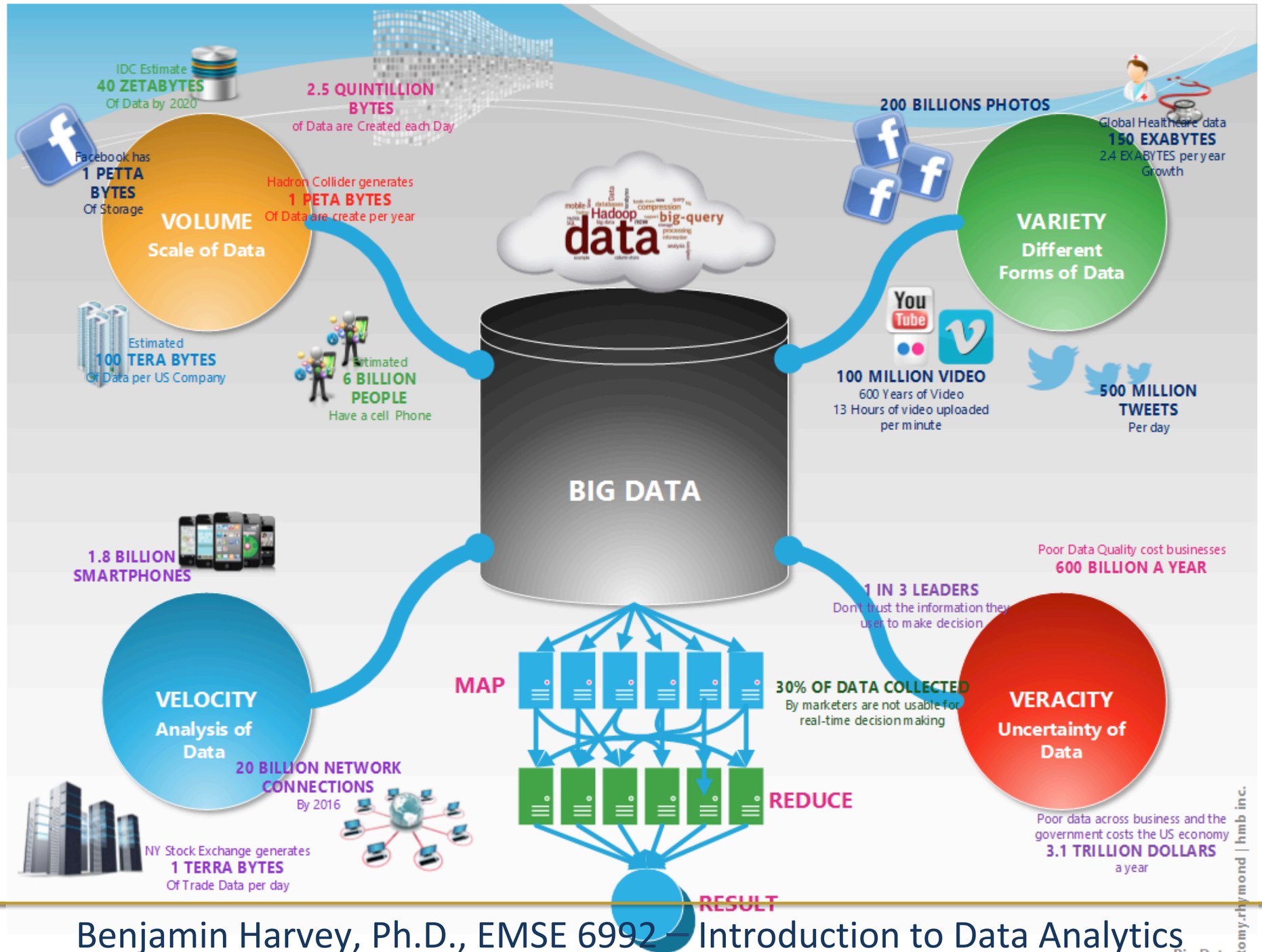
*Posted Nov 9, 2015 by **Frederic Lardinois** (@fredericl)*

Next Story

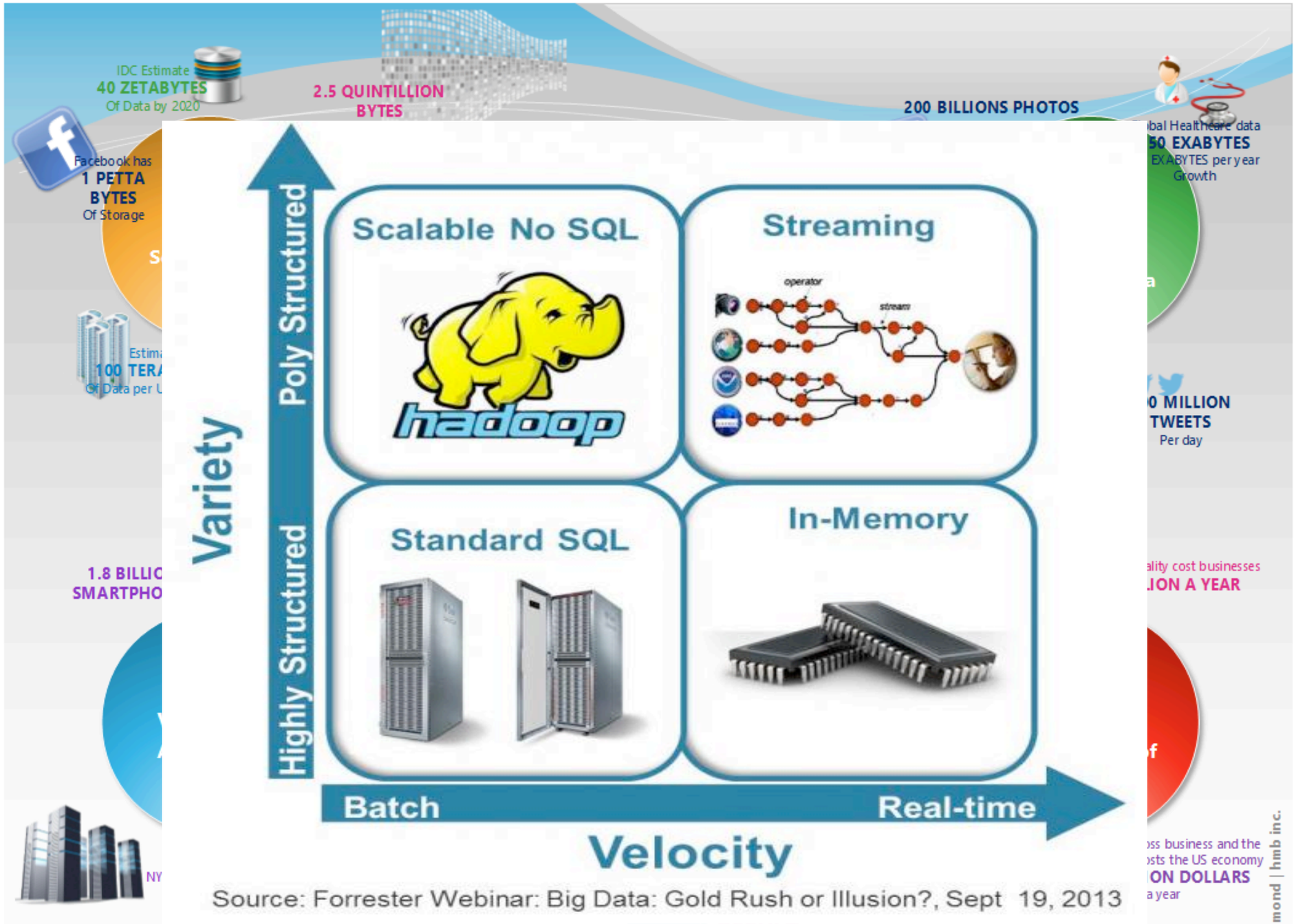
Mark Cuban

IV. What is Big Data?

What is Big Data?

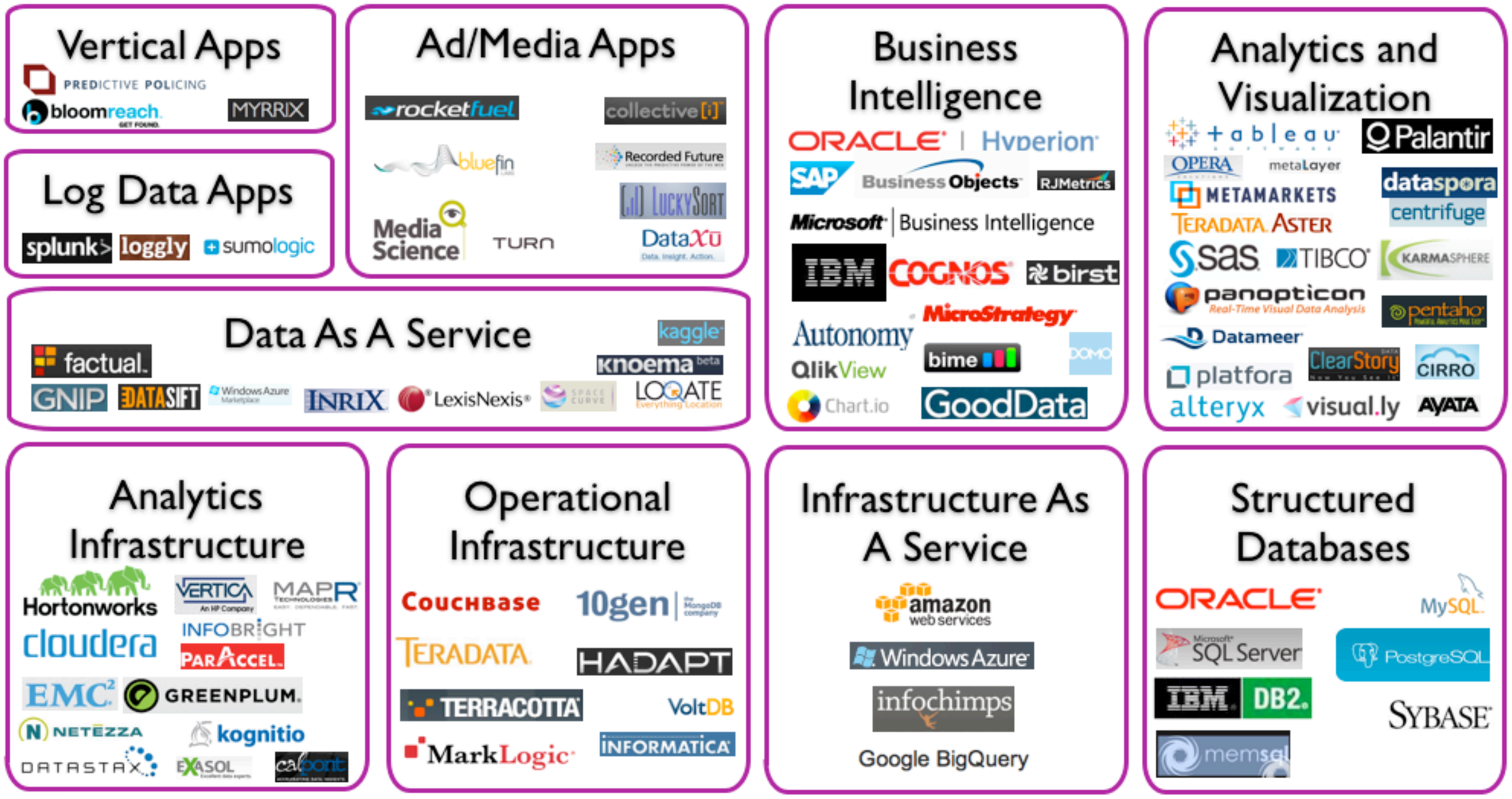


What's **NOT** Big Data?



60 Second Data Analysis

Big Data Landscape - 2012



dave@vcdave.com
blogs.forbes.com/davefeinleib

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, CAZENA, TREASURE DATA, altiscale, Qubole

Spark
databricks, GridGain, TACHYON NEXUS

Cluster Services
amazon web services, kubernetes, HPCC SYSTEMS, 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, GIRAPH, OrientDB, InfiniteGraph

MPP Databases
TERADATA, VERTICA, NETEZZA, 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, Prometheus, 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, interana

Data Science Platforms
context relevant, CONTINUUM ANALYTICS, DataRobot, Alpine DATA LABS, plotly, ARIMO, MODE, dataiku, Nutonian, DOMINO, sense, yhat, ALGORITHMIA

Visualization
+ a b l e a u, Google Cloud Platform, Qlik, looker, SISENSE, Roambi, GOOMDATA, datorama, CHARTIO

BI Platforms
Power BI, amazon web services, DOMO, salesforce, Wave Analytics, GoodData, birst, kyvos insights, platfora, atscale, ARCADIA DATA, 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, SKYTREE, Dato, rapidminer, DATARPM, deepsense.io, ViSENZE, PredictionIO, glowfish

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

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

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

Data Services
Mu Sigma, OPERA SOLUTIONS, DATASCIENCE, EXL analytics, 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, blueyonder, 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, theTradeDesk, Adgorithms, dstillery, LiveIntent, TAPAD, DataXu, Appier, MOAT

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

Vertical AI Applications
facebook, 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, zest finance, LendUp, Kabbage, tidemark, Payoff, INSIKT, uora, 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, 1010data, hp, Autonomy, VERTICA An HP Company, vmware, TIBCO, TERADATA, ORACLE, NetApp

Open Source

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

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

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

Coordination
talend, Google, Apache Zookeeper, Apache Ambari

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

Stat Tools
ScalaLab, NumPy, SciPy

Machine Learning
mllib, mahout, Apache SINGA, Aerosolve, MADlib, Caffe, learn, torch, CNTK, TensorFlow, FeatureFu, WEKA, 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, YODLEE, PREMISE, S&P CAPITAL IQ, quandl, xignite, CB INSIGHTS, mattermark, StockTwits, estimize, PLAID

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

Location / People / Entities
acxiom, Experian, EPSILON, InsideView, GARMIN, foursquare, STREETLINE, esri, Crimson Hexagon, CARTODB, factual., PlaceIQ, CIRCULATE, placemeter, BASIS, Sense 360

Other
qualtrics, panjiva, DATA.GOV

Incubators & Schools

GA, PLURALSIGHT, DataCamp, INSIGHT, DataElite, METIS, The Data Incubator

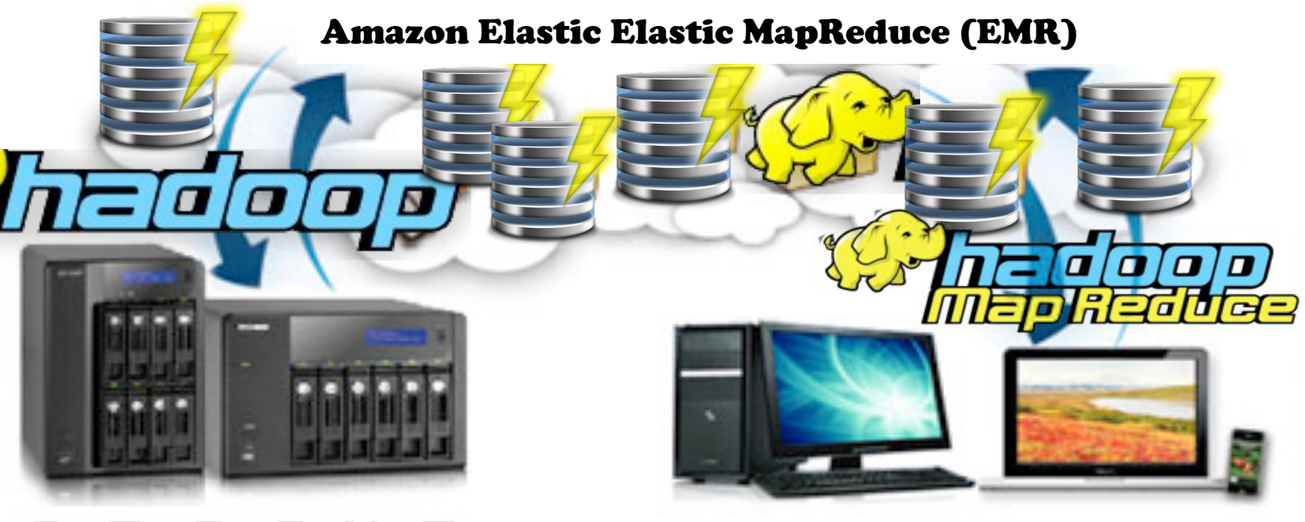
Last Updated 3/23/2016

FIRSTMARK

DATA CLOUD

amazon web services™

Amazon Elastic Elastic MapReduce (EMR)



APACHE HBASE

STORAGE CLOUD

amazon web services™

GW

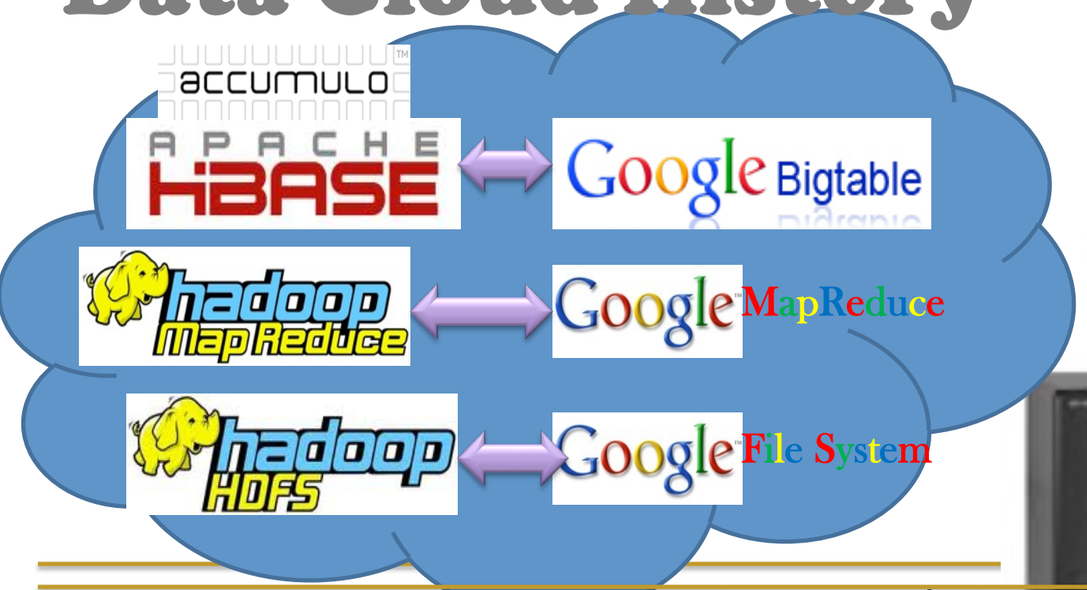
Amazon Simple Storage Service (S3)

Utility CLOUD

amazon web services™

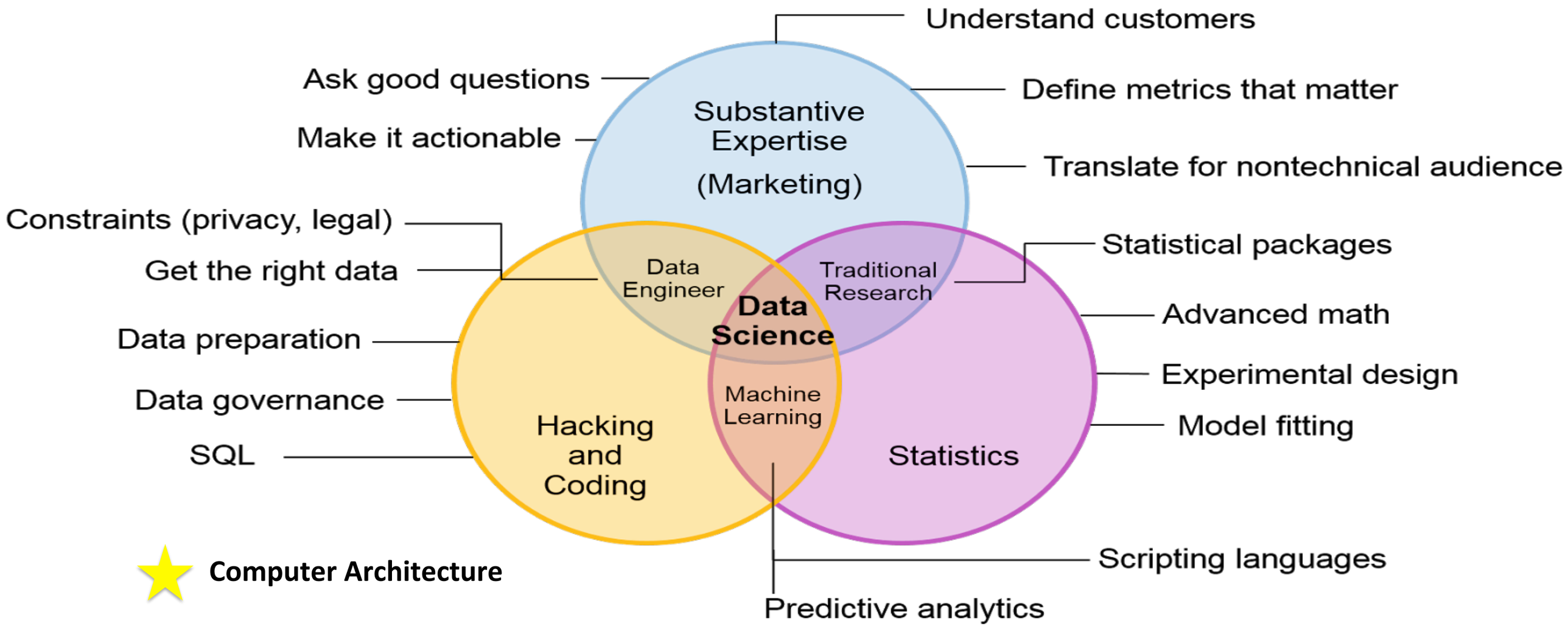
Amazon Simple Storage Service (S3)

Data Cloud History

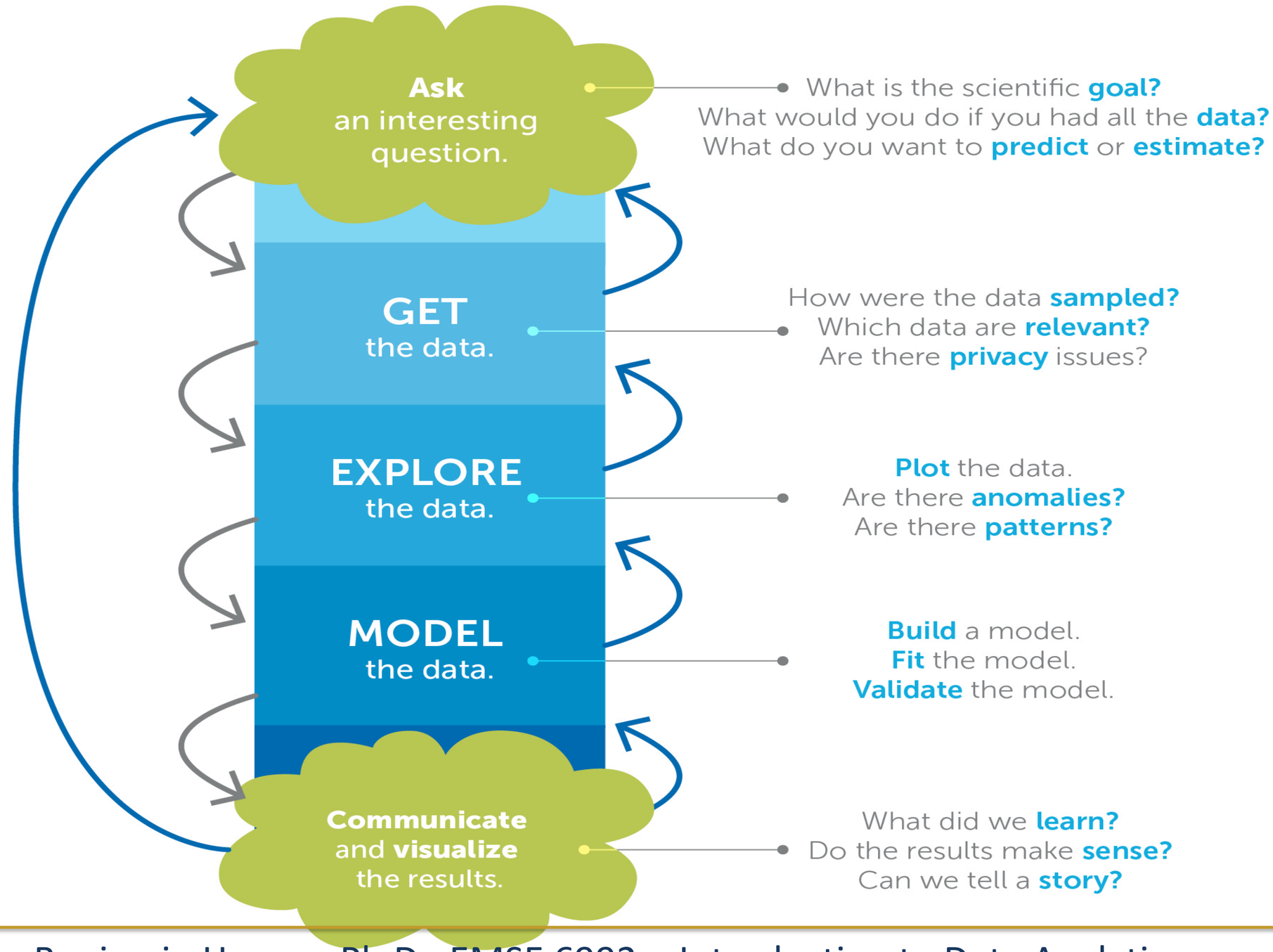


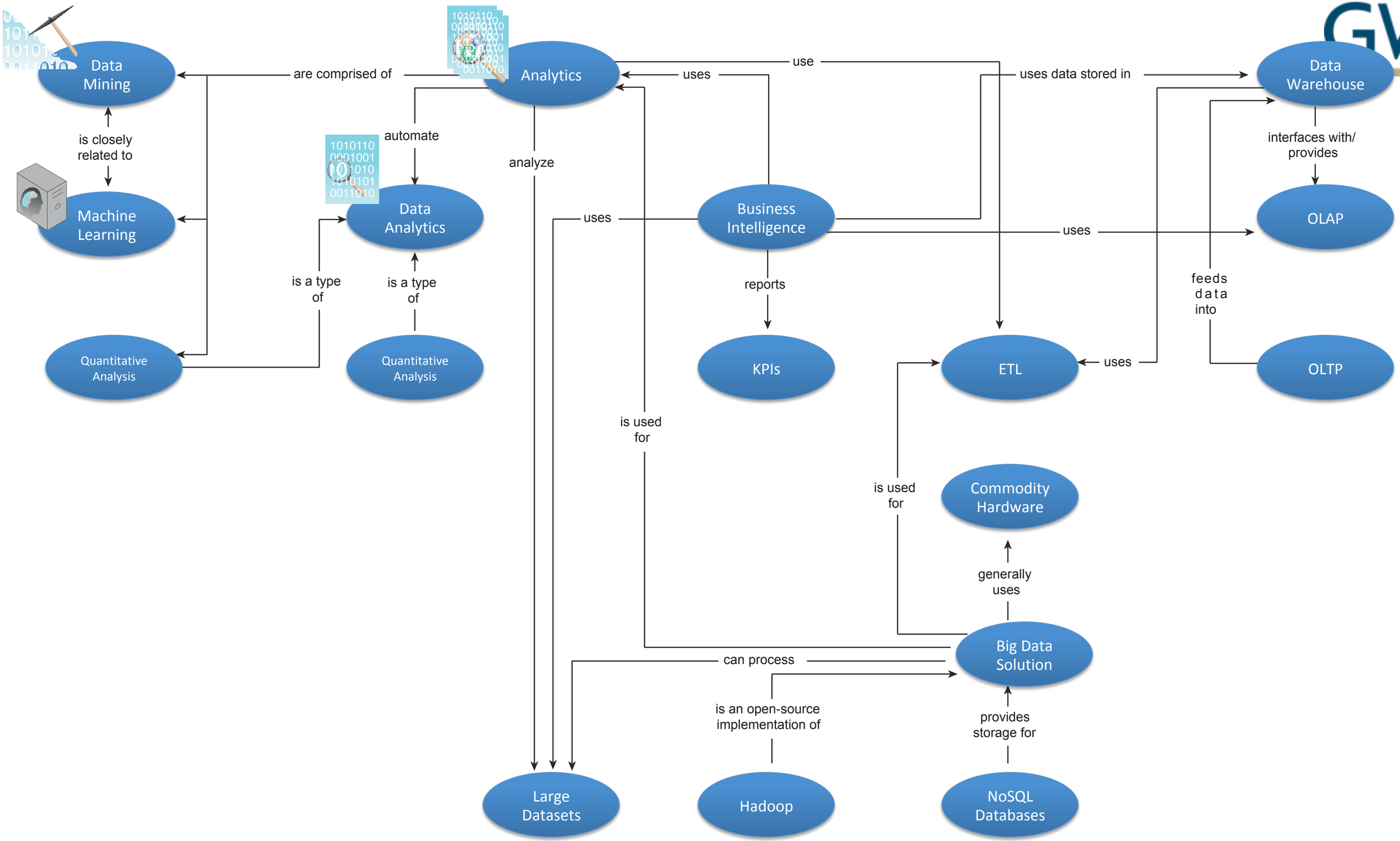
IV. What is Data Science

Data Science Venn Diagram



Data Science Process



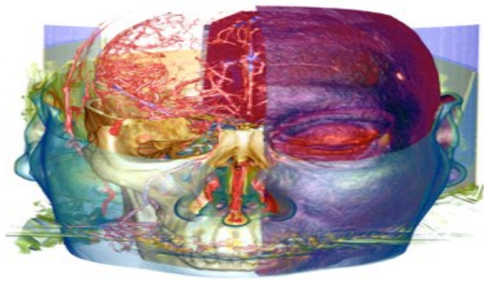

Data Mining
are comprised of
Analytics
use
uses
uses data stored in
Data Warehouse
is closely related to
Machine Learning
automate
analyze
Data Analytics
Business Intelligence
uses
uses
interfaces with/ provides
OLAP
is a type of
is a type of
reports
feeds data into
Quantitative Analysis
Quantitative Analysis
KPIs
ETL
uses
OLTP
is used for
is used for
Commodity Hardware
generally uses
Big Data Solution
can process
is an open-source implementation of
provides storage for
Large Datasets
Hadoop
NoSQL Databases

Discussion

- What is Data Science to you?
- Provide an example.
- What is ”Big Data” to you?
- Provide an example.

Summary: What is Data Science? What is Big Data?

Data Science: To gain insights through computation, statistics, and visualization




- In this course, you will learn four core topics that have been associated with Data Science
  - Data Science and Big Data Foundations, Big Data and Technology Concepts, Big Data Analysis and Science, Advanced Big Data Analysis and Science
- The definition of Big Data is subjective and really depends of the following:
  - Volume, Velocity, Veracity, Variety of data
  - Current CPU capacity

IV. Foundations of Big Data

Agenda

Concepts and Terminology

- Datasets
- Data Analysis
- Types of Data Analytics
  - Descriptive Analytics, Diagnostic Analytics, Predictive Analytics, Prescriptive Analytics
- Business Intelligence (BI)
- Key Performance Indicators (KPI)

Big Data Characteristics

- Volume, Velocity, Variety, Veracity, Value

Different Types of Data

- Structured Data
- Unstructured Data
- Semi-structured Data
- Metadata

Lab

Concepts and Terminology

Data Analysis

- Data analysis is the process of examining data to find facts, relationships, patterns, insights and/or trends. The overall goal of data analysis is to support better decision making.

Data Analytics

- Data analytics is a discipline that includes the management of the complete data lifecycle, which encompasses collecting, cleansing, organizing, storing, analyzing and governing data.

Big Data Analytics

- The lifecycle generally involves identifying, procuring, preparing and analyzing large amounts of raw, unstructured data to extract meaningful information that can serve as an input for identifying patterns, enriching existing enterprise data and performing large-scale searches.

Data analytics enable data-driven decision-making with scientific backing so that decisions can be based on factual data and not simply on past experience or intuition alone.

- descriptive analytics
- diagnostic analytics
- predictive analytics
- prescriptive analytics

Concepts and Terminology

Descriptive Analytics

- Descriptive analytics are carried out to answer questions about events that have already occurred.

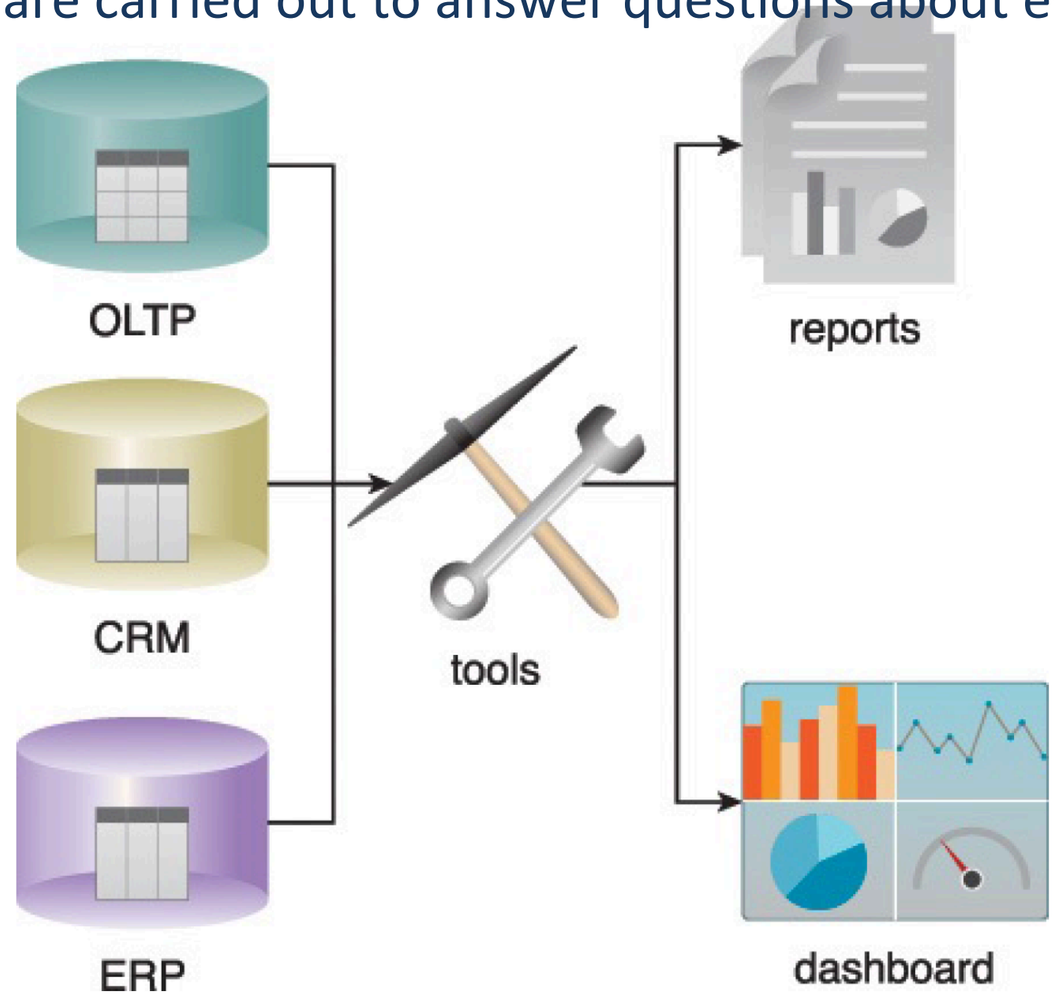


Figure 1.5 The operational systems, pictured left, are queried via descriptive analytics tools to generate reports or dashboards, pictured right.

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

Concepts and Terminology

Diagnostic Analytics

- Diagnostic analytics aim to determine the cause of a phenomenon that occurred in the past using questions that focus on the reason behind the event.
- The goal of this type of analytics is to determine what information is related to the phenomenon in order to enable answering questions that seek to determine why something has occurred.

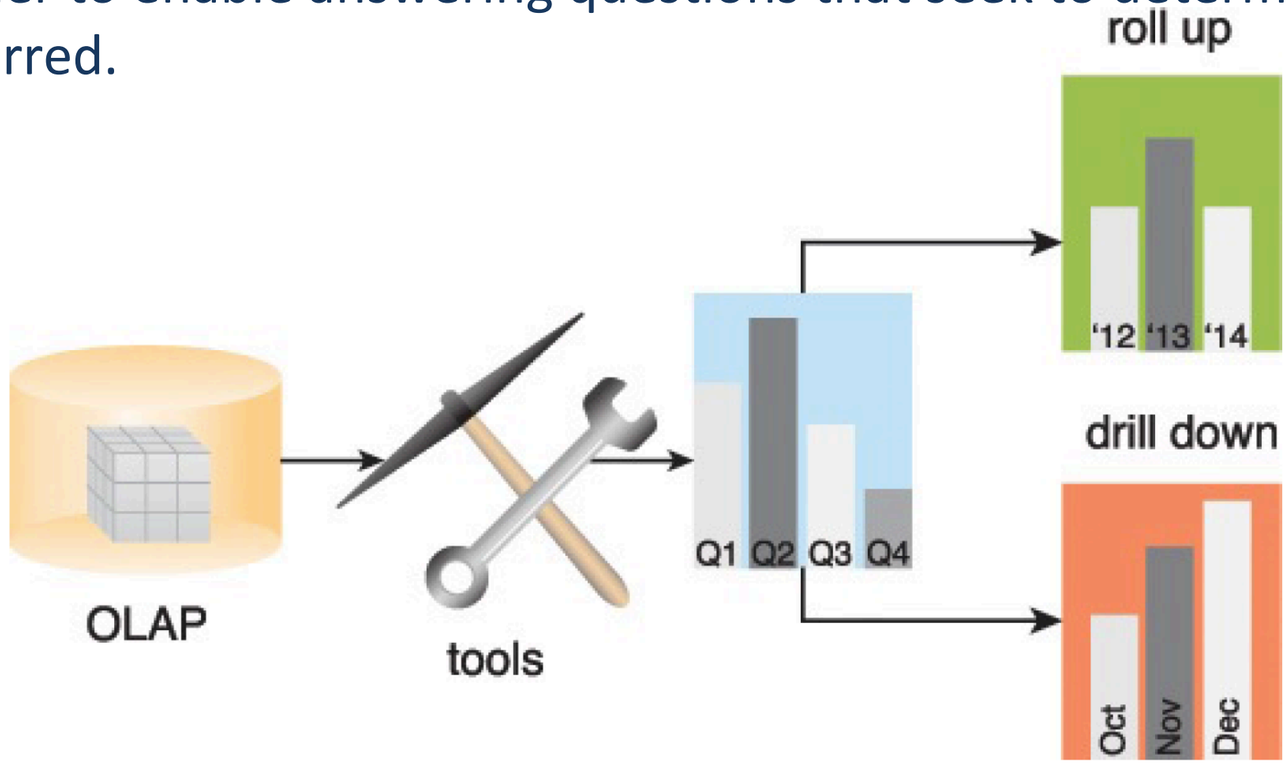


Figure 1.6 Diagnostic analytics can result in data that is suitable for performing drill-down and roll-up analysis.

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

Concepts and Terminology

Predictive Analytics

- Predictive analytics are carried out in an attempt to determine the outcome of an event that might occur in the future. Information is enhanced with meaning to generate knowledge that conveys how that information is related.

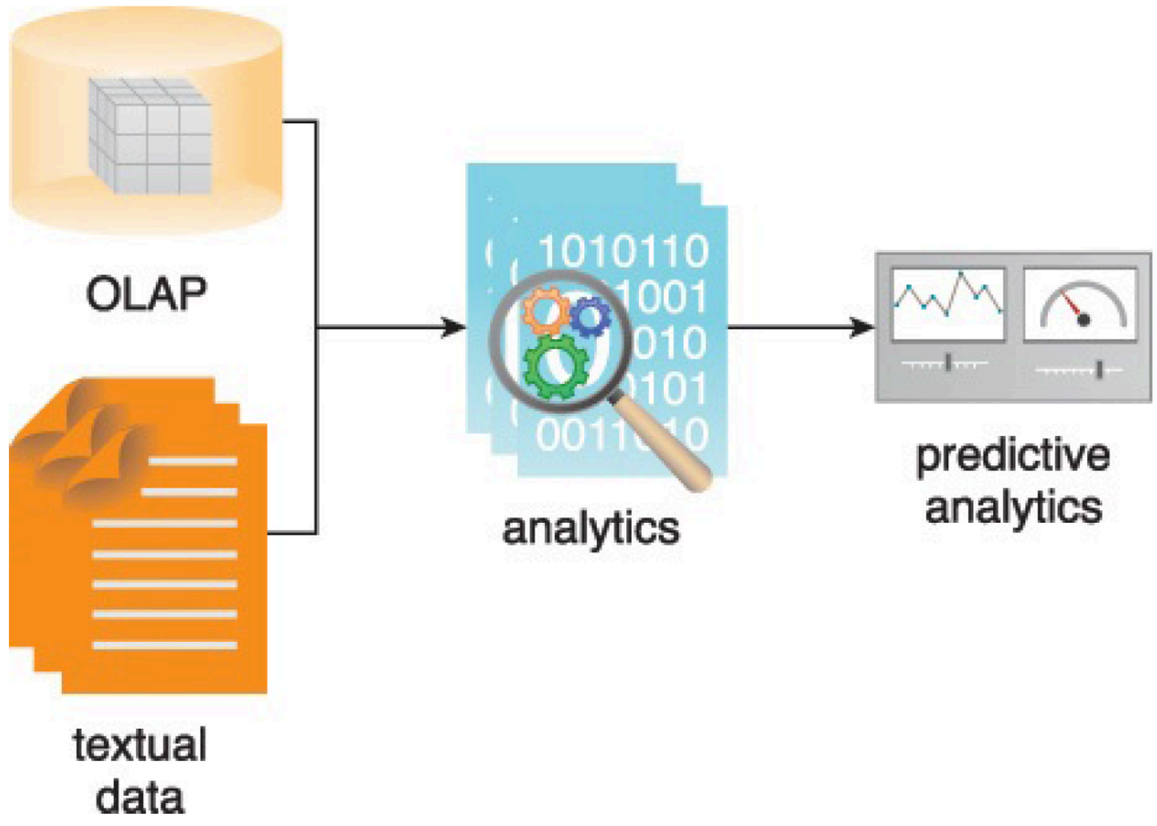


Figure 1.7 Predictive analytics tools can provide user-friendly front-end interfaces.

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

Concepts and Terminology

Prescriptive Analytics

- Prescriptive analytics build upon the results of predictive analytics by prescribing actions that should be taken (best option and why).

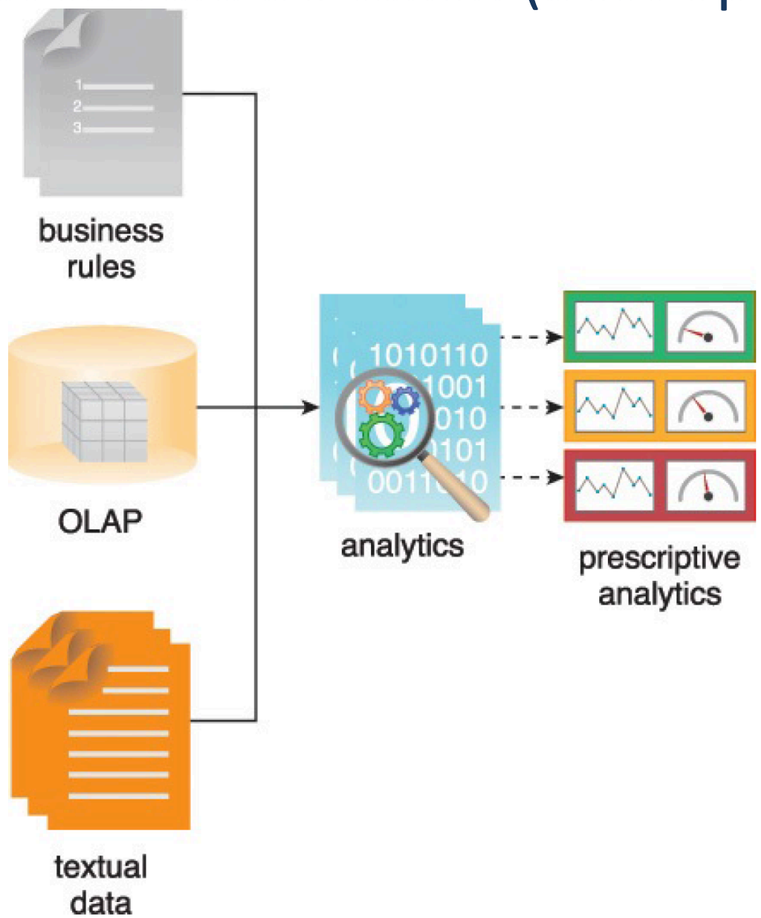


Figure 1.8 Prescriptive analytics involves the use of business rules and internal and/or external data to perform an in-depth analysis.

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

Concepts and Terminology

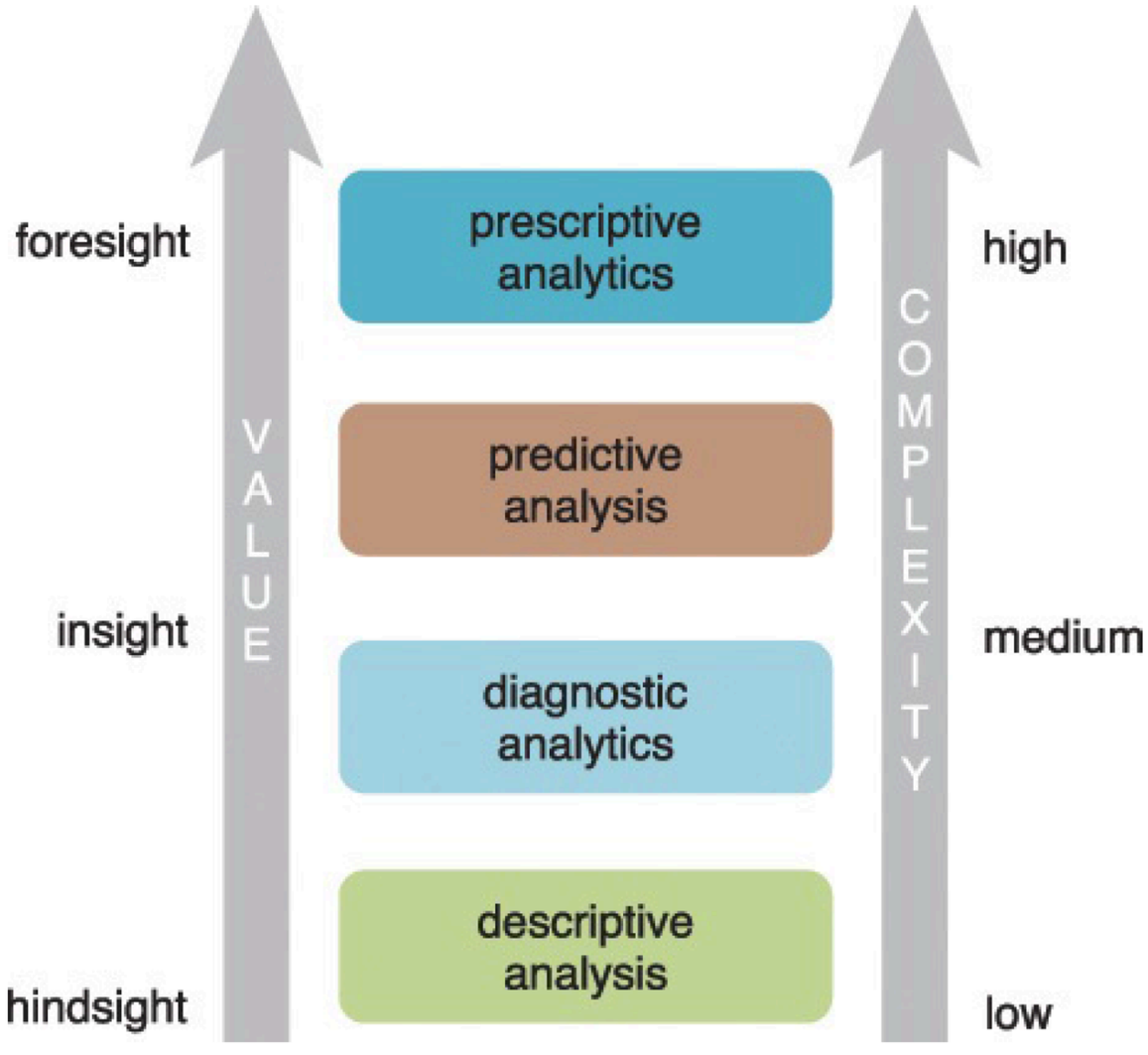


Figure 1.4 Value and complexity increase from descriptive to prescriptive analytics.

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

Concepts and Terminology

Key Performance Indicators (KPI)

- A KPI is a metric that can be used to gauge success within a particular business context.
- KPIs are linked with an enterprise's overall strategic goals and objectives.
- They are often used to identify business performance problems and demonstrate regulatory compliance.

Business Intelligence (BI)

- BI enables an organization to gain insight into the performance of an enterprise by analyzing data generated by its business processes and information systems.
- The results of the analysis can be used by management to steer the business in an effort tc formance

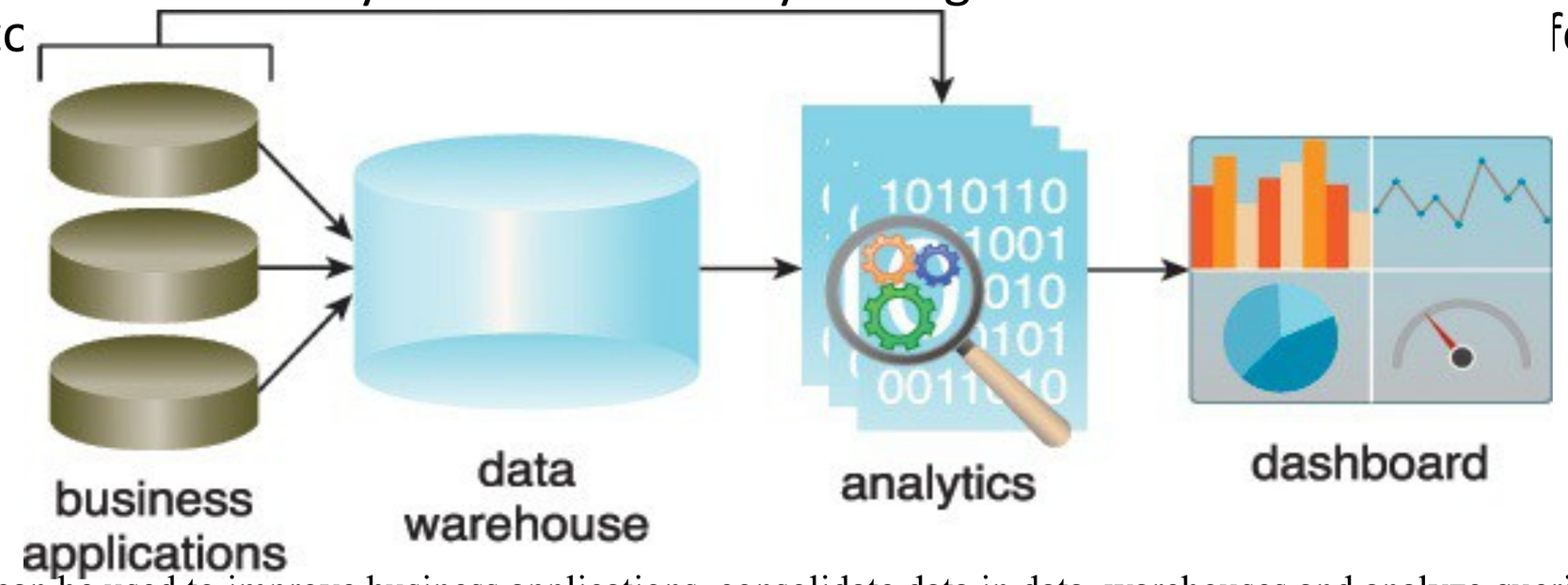


Figure 1.9 BI can be used to improve business applications, consolidate data in data warehouses and analyze queries via a dashboard.
Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

Concepts and Terminology

Structured Data

- Structured data conforms to a data model or schema and is often stored in tabular form. It is used to capture relationships between different entities and is therefore most often stored in a relational database.

Semi-structured Data

- Semi-structured data has a defined level of structure and consistency, but is not relational in nature. Instead, semi-structured data is hierarchical or graph-based.

Unstructured Data

- Data that does not conform to a data model or data schema is known as unstructured data.
- Metadata provides information about a dataset's characteristics and structure.

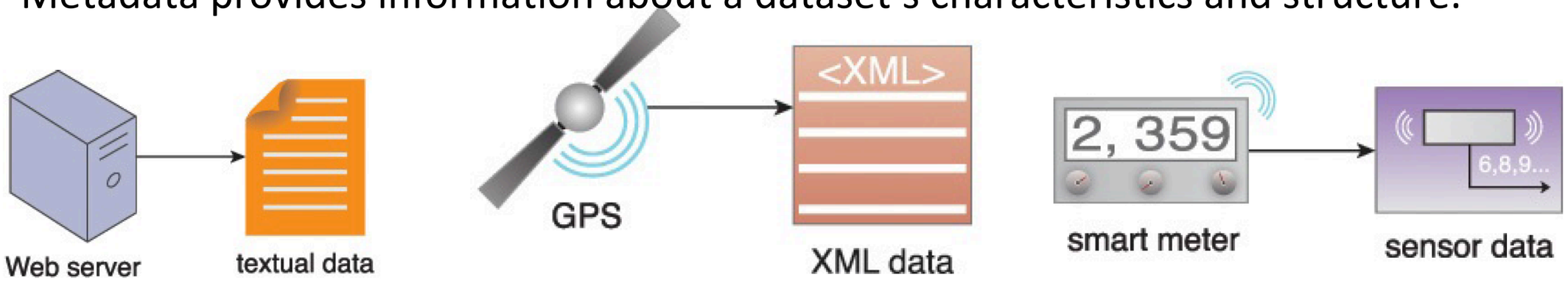


Figure 1.17 Examples of machine-generated data include web logs, sensor data, telemetry data, smart meter data and appliance usage data

Erl, Thomas, Wajid Khattak, and Paul Buhler. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press, 2016.

Case Study: Ensure to Insure

Identifying Types of Data

- The IT team members go through a categorization exercise of the various datasets that could be identified:
  - Categorize them as either: Structured, Unstructured and Semi-Structured

Identifying Data Characteristics

- The IT team members want to gauge different datasets that are generated inside ETI's boundary as well as any other data generated outside ETI's boundary that may be of interest to the company in the context of volume, velocity, variety, veracity and value characteristics.
- The team members take each characteristic in turn and discuss how different datasets manifest that characteristic.

Do you think the new Big Data strategy will work?