



Applied Research Center for Data Analytics and Web Insights



Big Data Analytics Introduction

Assoc.Prof. Abzeldin ADAMOV

Director, Center for Data Science Research & Training
ADA University
aadamov@ada.edu.az
<http://site.ada.edu.az/~aadamov>



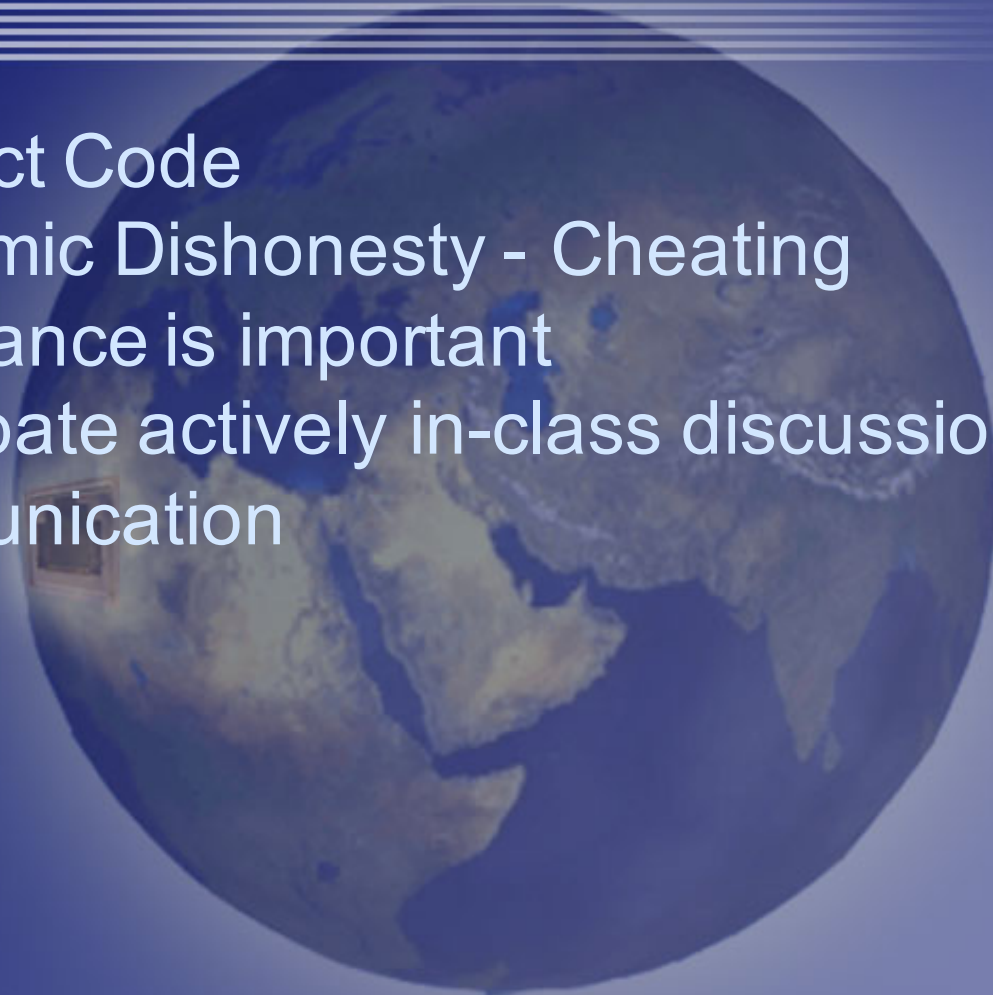
EVALUATION

- Midterm – 25%
 - Labs and Quizzes – 20%
 - Final Project – 15%
 - Final – 35%
 - Attendance – 5%
 - Total – 100 points
-
- Final Project – Research Report – in accordance with IEEE requirements, min 7 pages + Presentation [software implementation - it depends]
 - You are welcome to propose ideas regarding your Final Project topic...



CLASS POLICY

- Conduct Code
- Academic Dishonesty - Cheating
- Attendance is important
- Participate actively in-class discussion topics
- Communication





CLASS RULES





CLASS RULES



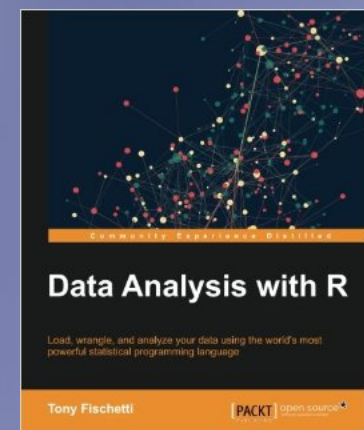
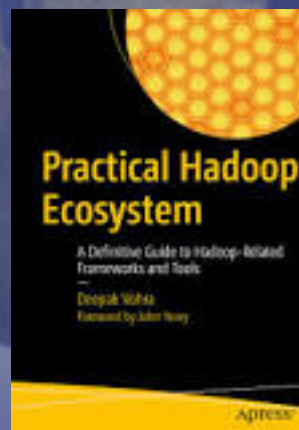
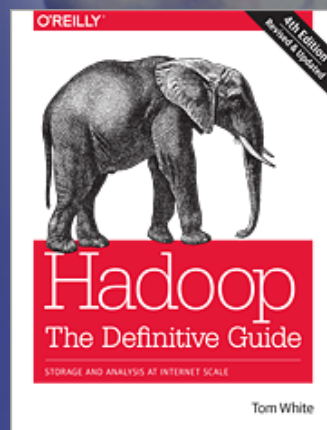
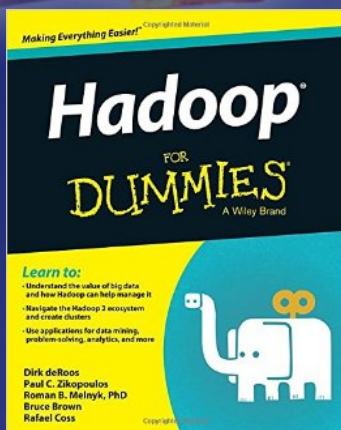
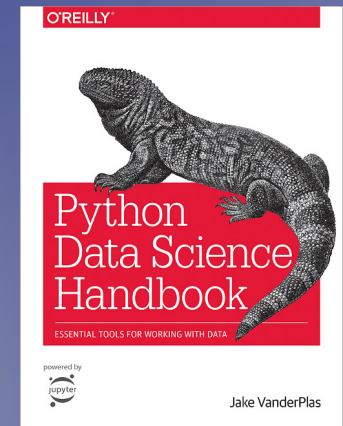
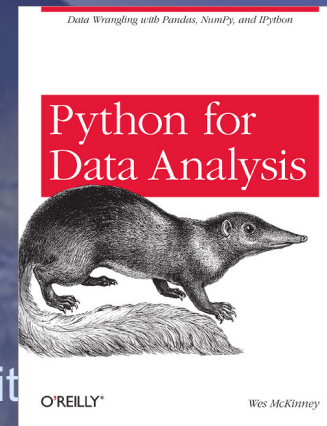


COURSE CONTENT

- Introduction to Big Data Analytics
- New Business Opportunities from Big Data Analytics (Use Cases)
- Setting Virtual Environment for Data Science
- Big Data Platforms, Hadoop Distributed File System (HDFS)
- Hadoop Installation and Configuration (Laboratory)
- Hadoop Ecosystem (Primary components: Hadoop, YARN, MapReduce, HDFS, Pig, Hive, HBase, Zookeeper, Sqoop, Flume)
- Distributed Computing with Hadoop and MapReduce (YARN)
- MapReduce Programming Concept
- Data Mining Techniques and Tools with R / Python / Java
- Data Manipulation and Processing with R / Python
- Big Data Analytics and ML Algorithms R / Python
- Final Projects Presentation
- Text Analytics
- Big Data Security and Privacy

REFERENCES

- Practical Hadoop Ecosystem, Deepak Vohra
- R for Everyone, Jared Lander
- Data Analysis with R, Tony Fischetti
- Python for Data Analysis, Wes McKinney
- Hadoop for Dummies, Dirk deRoos
- Hadoop: The Definitive Guide 4th ed., John White
- Big Data For Dummies, Judith Hurwitz



REFERENCES

- Big Data University www.bigdatauniversity.com
- IBM Academic Initiative <https://developer.ibm.com/academic/>
- EMC Academic Alliance <https://education.emc.com/academicalliance/>
- Hortonworks Academic Program
<http://hortonworks.com/training/hortonworks-university-academic-program/>
- Cloudera Academic Partnership
<http://www.cloudera.com/developers/academic-partnership.html>



CLASS RULES



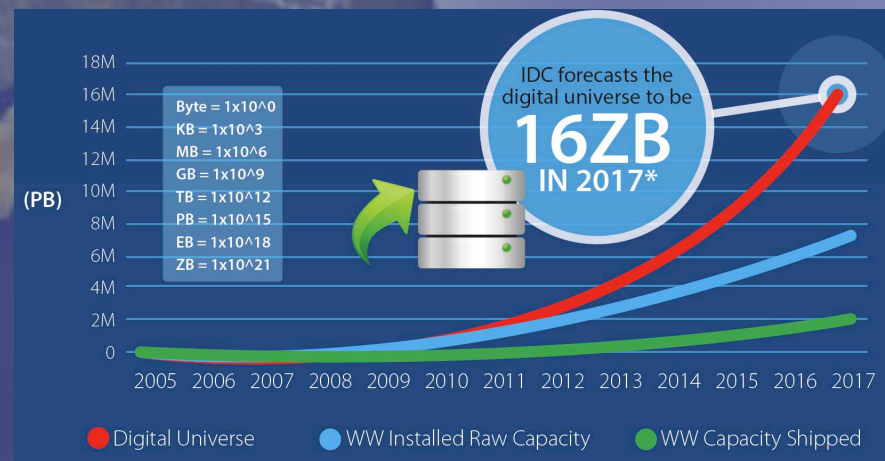


Digital Universe

Volume of Digital Data

- 2003 – 5 exabytes from beginning of civilization
- 2005 – 130 exabytes
- 2008 – 480.000 petabytes (PB)
- 2009 – 800.000 PB
- 2010 – 1200 000 PB or 1.2 zettabyte (ZB)
- 2011 – 1.8 ZB
- 2012 – 2.7 ZB
- 2014 ~ 6.2 ZB
- 2016 ~ 12 ZB
- 2017 ~ 16 ZB
- Expected to reach 44 ZB by 2020

Every day now we create as much information as we did from the dawn of civilization up until 2003





Big Measures for Big Data

- kilobyte (kB) 10^3 2^{10}
- megabyte (MB) 10^6 2^{20}
- gigabyte (GB) 10^9 2^{30}
- terabyte (TB) 10^{12} 2^{40}
- petabyte (PB) 10^{15} 2^{50}
- exabyte (EB) 10^{18} 2^{60}
- zettabyte (ZB) 10^{21} 2^{70}
- yottabyte (YB) 10^{24} 2^{80}
- Brontobyte (BB) 10^{27} 2^{90}

Why Data Grows so Fast?

Data is produced by:

- **People**
 - Social Media, Public Web, Smartphones, ...
- **Organizations (Employer)**
 - OLTP, OLAP, BI, ...
- **Machines**
 - IoT, Satellites, Vehicles, Science, ...

```
for (int j = 0; j < locs[j++] res[j] = buff[j]);
return res;

public void decodeMessage(int[] res) {
    for (int i = 0; i < res.length; i++) {
        res[i] = checkRes(res[i]);
    }
}

decodeMessage(0) {
    0; i < MAX_RES_LEN; i++) buff[i] = 0;
    i = 0;
    while (i < res.length) {
        buff[i++] = res[i];
    }
}

extractMessage(res);

public int[] extractMessage(int[] res) {
    for (int i = 0; i < MAX_RES_LEN; i++) buff[i] = 0;
    while (i < res.length) {
```





Where we were?





Where we are?





Where we were?

Pope Benedict inauguration in **2005**





Where we are?

Pope Francis inauguration in **2013**



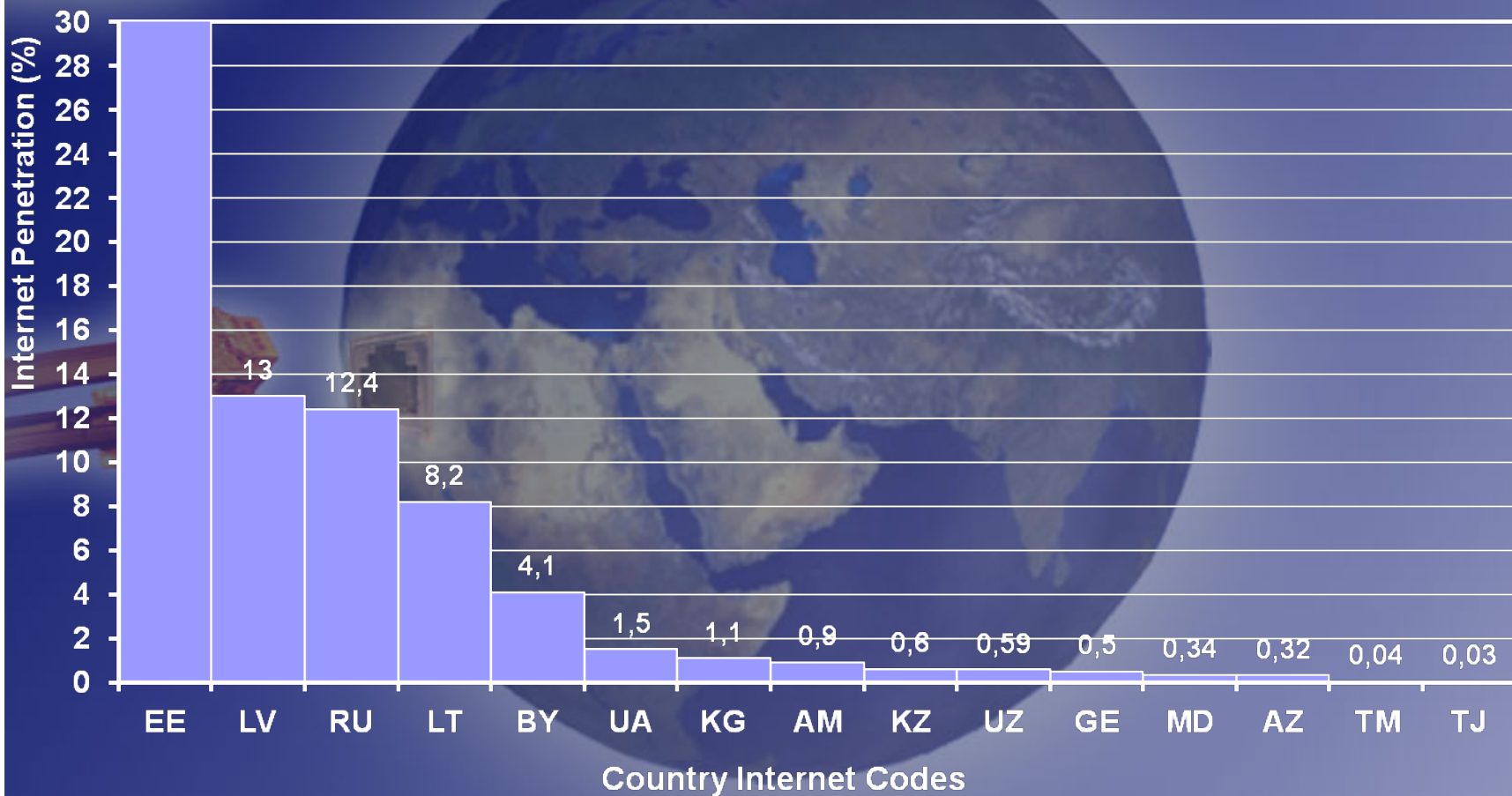


Where we will be?





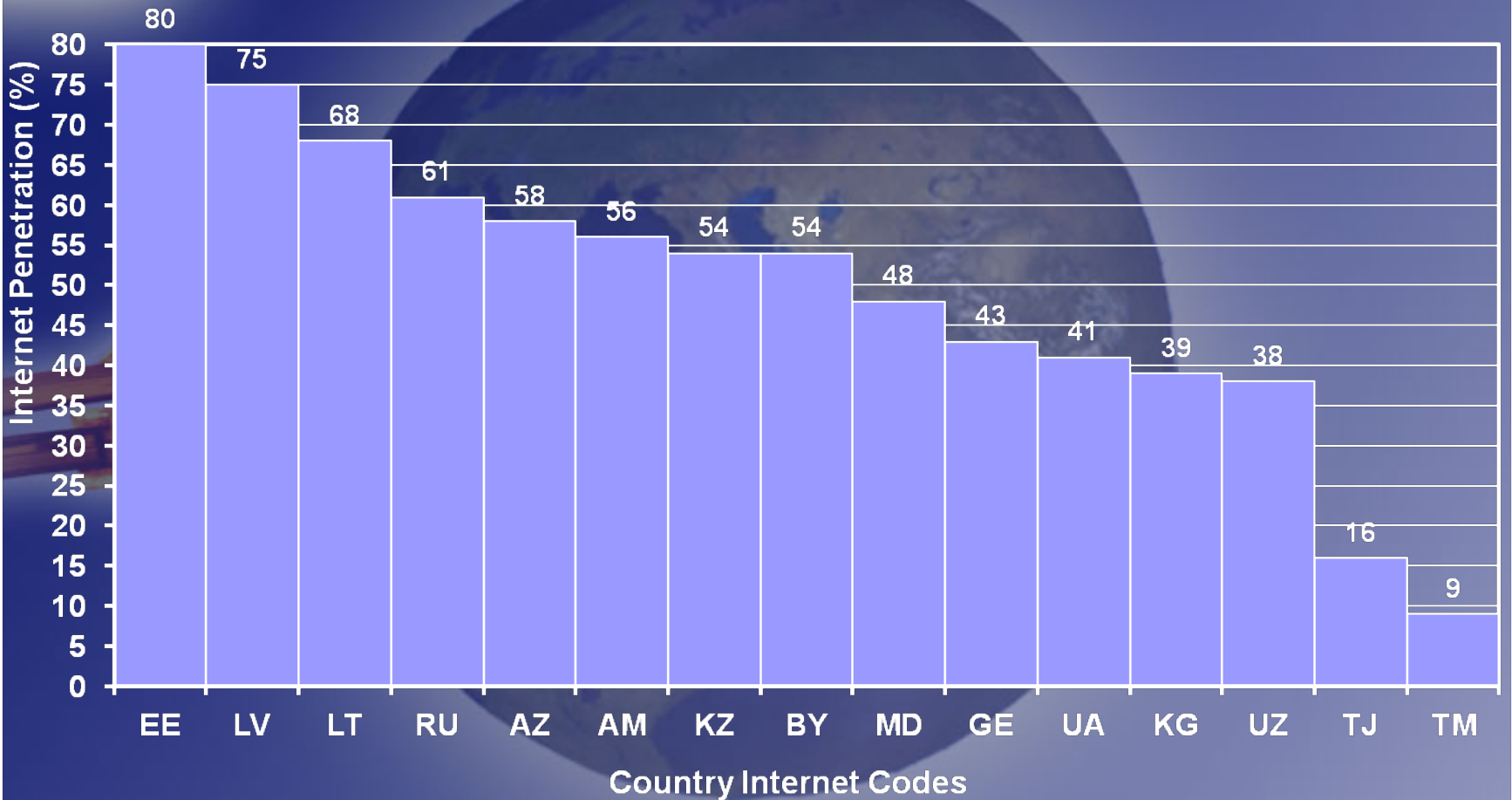
Internet Penetration



Note: Internet stats for December 2001

Average Internet usage in the world 8% - 500 Million - 2001

Foundations of the Web



Note: Internet stats for January 2014

Average Internet usage in the world 42% - 3.0 Billion - 2014

Social Networking

Top 15 Most Popular Social Networking Sites | January 2016



1,310,000,000 - Estimated Unique Monthly Visitors | **2** - Compete Rank



25,500,000 - Estimated Unique Monthly Visitors | **346** - Compete Rank



12,000,000 - Estimated Unique Monthly Visitors | **617** - Compete Rank



284,000,000 - Estimated Unique Monthly Visitors | **24** - Compete Rank



20,500,000 - Estimated Unique Monthly Visitors | **605** - Compete Rank



7,500,000 - Estimated Unique Monthly Visitors | **838** - Compete Rank



343,000,000 - Estimated Unique Monthly Visitors



19,500,000 - Estimated Unique Monthly Visitors | **447** - Compete Rank



5,400,000 - Estimated Unique Monthly Visitors | **122** - Compete Rank



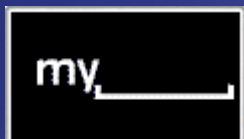
347,000,000 - Estimated Unique Monthly Visitors | **44** - Compete Rank



17,500,000 - Estimated Unique Monthly Visitors | ***NA*** - Compete Rank



3,000,000 - Estimated Unique Monthly Visitors | **451** - Compete Rank



70,500,000 - Estimated Unique Monthly Visitors | **51** - Compete Rank



12,500,000 - Estimated Unique Monthly Visitors | **127** - Compete Rank



2,500,000 - Estimated Unique Monthly Visitors | **1,596** - Compete Rank



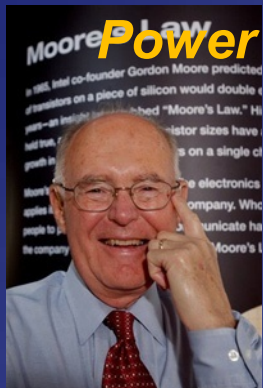
What happens each Second online

- 25 Terabytes transferred through across Internet
- 9 Website created (172 000 per day)
- 1 800 000 SPAM emails sent
- 4 100 Photos posted on Facebook (355 mln per day)
- 5 000 Instagram photos uploaded
- 1 500 Skype calls made
- 4 000 Tweets tweeted
- 10 000 Dropbox files uploaded
- 45 000 Google searches made (3.5 bln per day)
- 92 000 YouTube videos viewed
- 55 000 Facebook likes

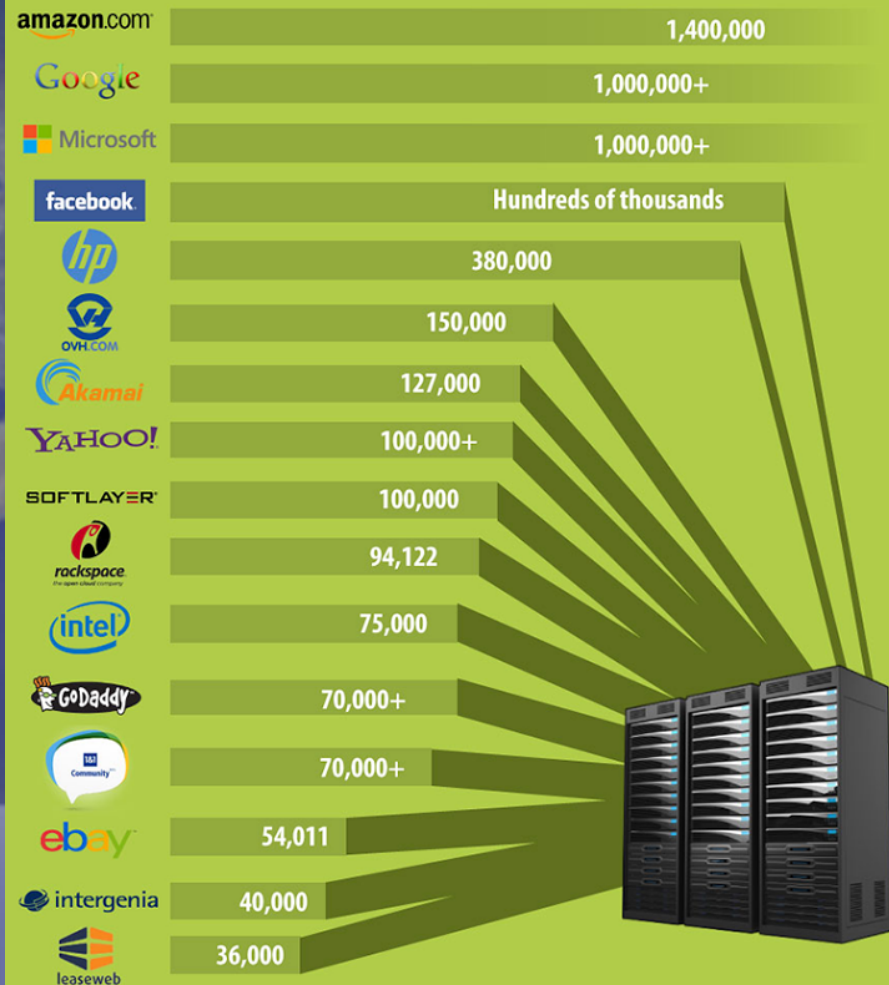


Problem with Moore's Law

- The number of transistors that can be placed on an integrated circuit doubles every 18 months to two years
- It's predicted to reach its limit with existing technology in 2020
- Cutting the size of a transistor to a single atom may defeat that concept
- ***The Digital Universe is growing much more faster than Processing***



Companies by Estimated Number of Servers



What is Big Data?





What is Big Data?

- Gartner: Big Data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making
- Something beyond of our Data Capacity and our Processing Power is Big Data – Big Data is combination of infrastructure, algorithms and visualization used to make sense of user and machine generated data
- Big data is not limited to more data than you can effectively work with on a single computer, its mostly about finding insights not seeing before and answering questions with data.

Why Now?



The Most Valuable Resource

- In its raw form, oil has little value. Once processed and refined, it helps power the World: Ann Winblad
- Data is the new oil: Clive Humby, CNBC



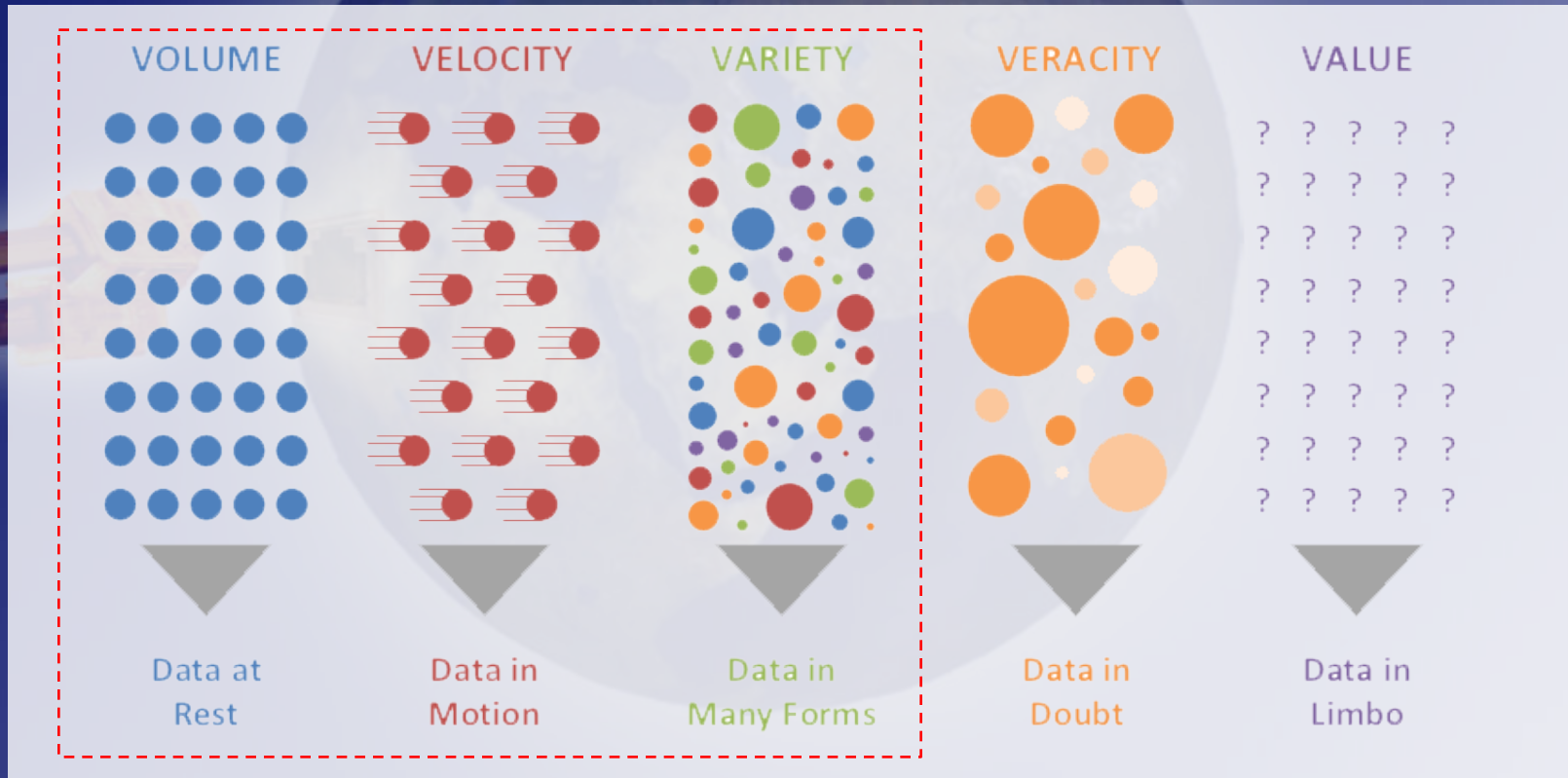


Need for Big Data

- 90% of the Data in the World was generated in the last two years
- Amount of Data doubles each 18 months
- 80% of the Data is unstructured, what makes it difficult to analyze
- Structured formats have limitations of handling large quantities of Data
- Its difficult to integrate the Data distributed across multiple systems
- Most business users do not know what should be analyzed and how
- Potentially valuable Data is dormant and discarded
- It is too expensive to process large amounts of Unstructured Data
- Most of Data has a short lifespan

5 Vs of Big Data

5 Vs that best describe the nature of Big Data Problem





Data Concepts/Formats

Structured: 5-10% of all Data Universe

SQL - Databases

Semi-Structured: 5-10%

CSV, XML, JSON, email structure

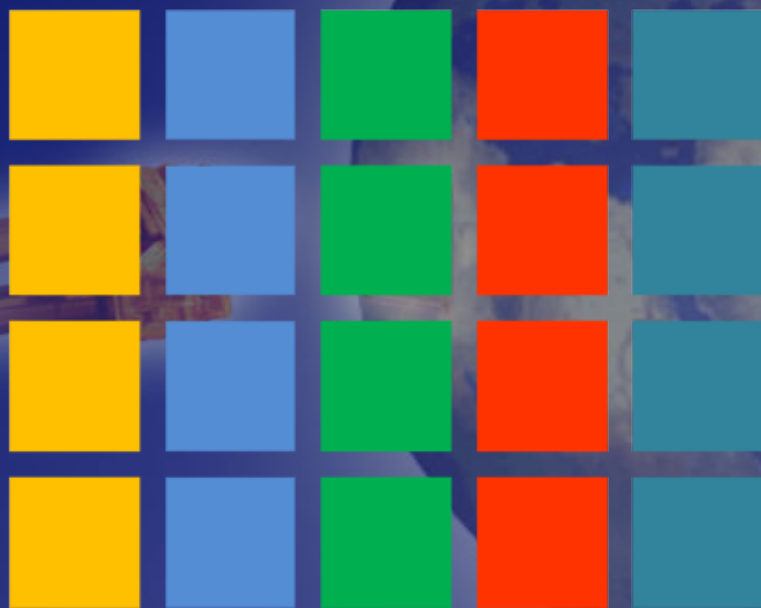
Unstructured: 80-90%

books, journals, documents, metadata, log files, health records, audio, video, images, files, email message, Web page, social media, word-processor document, ...

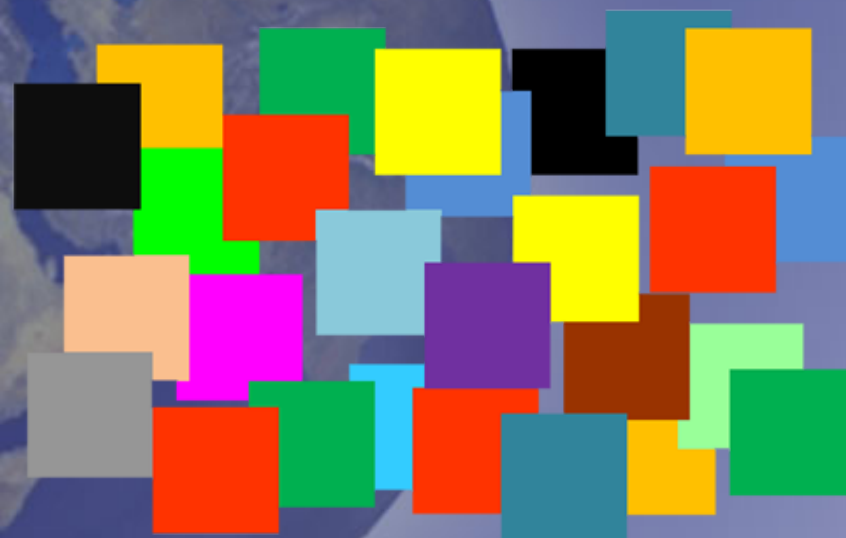


Structured vs Unstructured

Structured Data



Unstructured Data

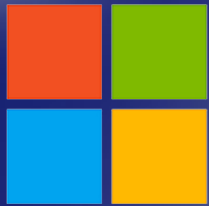


80% of potential benefit from Big Data is expected from unstructured data

Digital Universe study finds that 0.5% of global data is analysed



Big Data Giants and Vendors



Microsoft

facebook.



twitter

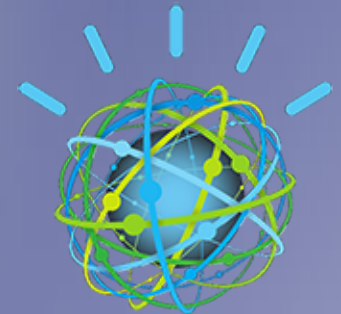
DELL EMC

ORACLE

Google



amazon
web services



IBM Watson

IBM



Hortonworks

cloudera

Big Data Landscape

Vertical Apps



Ad/Media Apps



Business Intelligence



Analytics and Visualization



Log Data Apps



Data As A Service



Analytics Infrastructure



Operational Infrastructure



Infrastructure As A Service



Structured Databases



Technologies



Big Data Landscape 2016 (Version 3.0)

Infrastructure

On-Premise
 cloudera, Hortonworks, Pivotal, IBM InfoSphere, bluedata, jethro

Hadoop in the Cloud
 Amazon Web Services, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, quable

Spark
 databricks, GridGain, TACHYON NEXUS

Cluster Services
 Amazon Web Services, Kubernetes, Docker, Mesosphere, CoreOS, StackIQ

Analytics

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

Analytics Platforms
 Microsoft, GUAVUS, Datameer, Bottlenose, interana

Data Science Platforms
 context relevant, DataRobot, CONTINUUM ANALYTICS, Alpine, MODE, dataiku, yhat, ARIMO, tonian, sense, ALGORITHMIA

Visualization
 tableau, Google Cloud Platform, Olik, looker, Roambi, Sisense, YomDATA, dataroma, CHARTIO

Applications

Sales & Marketing
 RADIUS, Gainsight, bloomreach, Zeta, EVERSTRING, livefyre, blueyonder, Lattice, kahuna, infer, SAILTHRU, persado, AVISO, sense, fuse|machines, QUANTIFIND, ACTIONIQ, ENAGIO

Customer Service
 MEDALLIA, ATTENSTY, CLARABRIDGE, CLICKFOX, STELLASERVICE, NGDATA, Preact, DigitalGenius, appuri, Wiseio

Human Capital
 gild, Connectifier, textic, entelo, hiQ

Legal
 RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

NoSQL Databases
 Amazon DynamoDB, Google Cloud Platform, Microsoft Azure, mongoDB, ORACLE, MarkLogic, DATASTAX, COUCHBASE, KERO SPIKE, Couchbase, SequoiaDB, redislabs, influxdata

NewsQL Databases
 SAP HANA, Clustrix, Pivotal, paradigm4, memsql, nuODB, splice MACHINE, MariaDB, VOLTDDB, citusdata, deepdb, Trafalgar, Cockroach LABS

BI Platforms
 Power BI, Amazon Web Services, DOMO, Wave Analytics, GoodData, platforma, atscale, ACADIA, BISSENSE

Statistical Computing
 SAS, SPSS, MATLAB

Log Analytics
 Splunk, sumologic, kibana, CLOUD PHYSICS, loggly

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

Ad Optimization
 AppNexus, MediaMath, critico, OpenX, rocketfuel, Integral, theTradeDesk, Algorithms, dstillery, LiveIntent, TAPAD, DataXu, Appier, MOAT

Security
 CYLANCE, CounterTack, cyberreason, AREA 1 SECURITY, SentinelOne, Recorded Future, Guardian Analytics, FORTSCALE, sift science, Keybase, feedzai, SICRIFYD

Vertical AI Applications
 facebook, Clara, KASIST@, lumia

Graph Databases
 neo4j, ORACLE, GIRAAPH

MPP Databases
 TERADATA, NETEZZA, Cacton, cognio, SAS, dremio

Cloud EDW
 Amazon Web Services, Google Cloud Platform, Microsoft Azure, Pivotal, snowflake, WATERBURNE DATA, Infoworks

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

Data Integration
 Informatica, MuleSoft, snaplogic, Bedrock Data, xplenty

Real-Time
 Amazon Web Services, METAMARKETS, striim, confluent, DATATORRENT, dataArtisans

Machine Learning
 Azure Machine Learning, H2O, Amazon Web Services, SKYTRIE, rapidminer, DATARPM, deepparso, VISENZE, PredictionIO, glowfish

Speech & NLP
 NarrativeScience, NUANCE, WolframAlpha, semanticmachines, Dato, ARRIA, apiai, cortical.io, maluba, MindMeld, IDIBON, VSCOOP

Horizontal AI
 IBM Watson, Cortana, sentiment, viv, nervana, vicarious, nara, Numenta, HyperScience, SI, Descartes Labs, clarifai, MetaMind

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

Govt / Regulation
 Socrata, OPENGOV, EN FiscalNote, PREPOL, mark43, OpenDataSoft

Finance
 Affirm, LendingClub, OnDeck, Kreditech, zesa finance, LendUp, Kabbage, tdemark, Pafyfi, INSIKT, ZUORA, Dataminr, Lenddo, KENSHO, AIDYA, ISENTIUM, Quantopian, sentient

Management / Monitoring
 New Relic, APPDYNAMICS, Amazon Web Services, acitifo, NUMERIFY, splunk, DATADOG, DRIVEN, Anodot

Security
 TANIUM, illumio, CODE42, DataGravity, CIPHERCLOUD, VECTRA, sqrl, BlueTalon

Storage
 Amazon Web Services, Google Cloud Platform, Microsoft Azure, panasas, nimblestorage, COHO DATA, Qumulo

App Dev
 apigee, CASK, Typesafe, DRIVEN

Crowd-sourcing
 Amazon Mechanical Turk, CrowdFlower, WorkFusion

Search
 HP, ANTOLOGY, ORACLE ENDECA, EXALEAD, Lucidworks, elastic, ThoughtSpot, MAANA, swifttype, Algolia, SINEQUA

Data Services
 MU SIGMA DO THE MATH, OPERA, EXL, DATA SCIENCE, DATA SCIENCE, kaggle, datascopie, DataKind

For Business Analysts
 OrigamiLogic, ClearStory, CIRRO, import io

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

Education / Learning
 KNEWTON, Clever, Declara, PANORAMA, knowtre

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

Industries
 OPPOWER, eHarmony, RetailNext, duetto, STITCH FIX, WorkFusion, BLUE RIVER, TACHYUS, SwiftKey, Seeq, FarmLogs, HowGood, select, SIGHT MACHINE, statmuse, BOBEVER

Cross-Infrastructure/Analytics

Amazon Web Services, Google, Microsoft, IBM, SAP, SAS, HP, Autonomy, VERTICA, VMware, TIBCO, TERADATA, ORACLE, NetApp

Open Source

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

Query / Data Flow
 SLAMDATA, Apache Drill, Hive, Google Cloud Dataflow

Data Access
 Cassandra, HBASE, mongoDB, CouchDB, riak, SCIDB, nifi, OPENTSOB

Coordination
 Apache Zookeeper, Apache Ambari

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

Stat Tools
 ScalaLab, Numpy, SciPy

Machine Learning
 mlilb, Aerosolve, Apache SINGA, MADlib, Caffe, CNTK, TensorFlow, FeatureFu, jupyter, DL4J, VELES, WEKA, DIMSUM

Search
 Elasticsearch, Solr, Lucene

Security
 Apache Ranger, Zeppelin

Data Sources & APIs

Health
 Apple, JAWBONE, GARMIN, practice fusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API

IOT
 UPTAKE, ThingWorx, helium, samsara, AUGURY, estimate

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

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

Location / People / Entities
 axioma, Experian, EPSILON, InsideView, GARMIN, foursquare, STREETLINE, esri, Crimson Hexagon, CARTODB, Factual, PlaceIQ, CIRCULATE, placemeter, BASIS, Sense

Other
 qualtrics, panjiva, DATA.GOV

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



Data Mining



Data Mining Synonyms



- Data Mining
- Knowledge Mining from Data
- Knowledge Extraction
- Information Harvesting
- Data/Pattern Analysis
- Data Archaeology
- Data Dredging
- Data Fishing
- Knowledge Discovery from Data (KDD)





Steps/Stages of Data Mining?

Data mining is the process of automatically discovering useful information in large data repositories.





Big Data and Data Science

“War is ninety percent information.” – Napoleon Bonaparte



What Big Data is and isn't?

BIG DATA



Google's First Data Centers



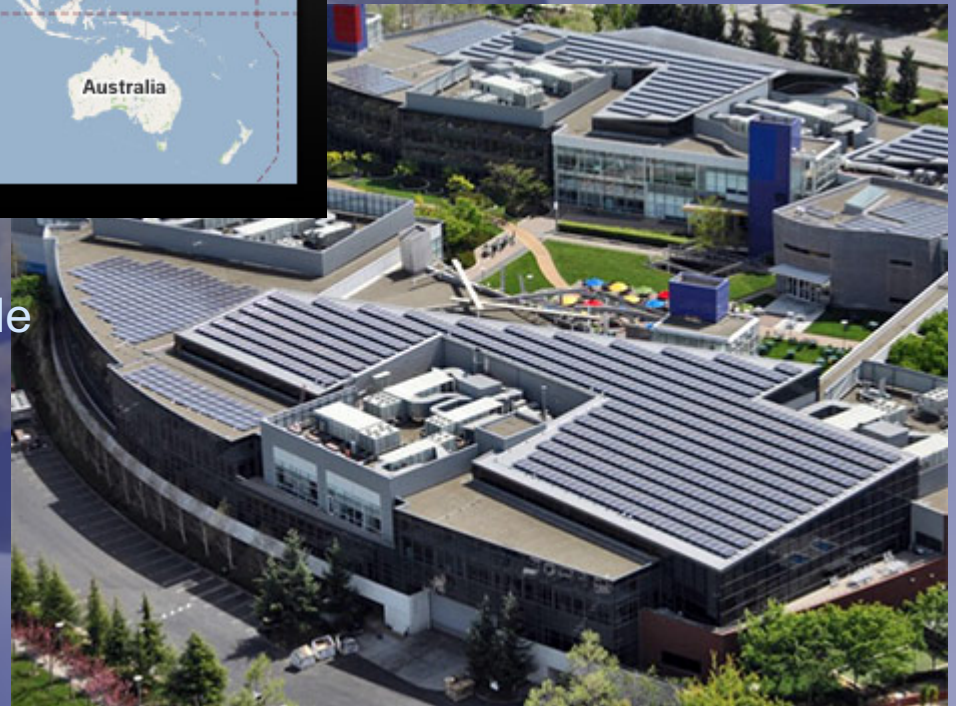
Google's first data center

Google New Data Centers



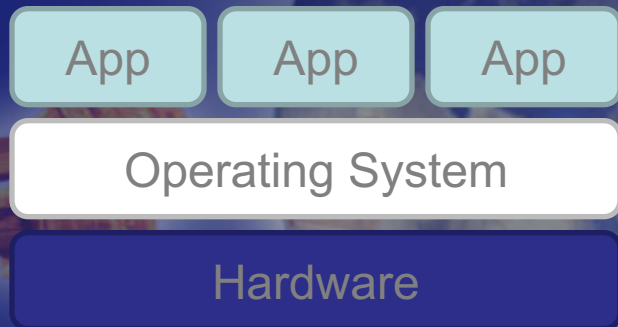
Map of Google Data Centers Worldwide

450,000 servers range upwards of 20 megawatts, which cost on the order of US\$2 million per month in electricity charges.

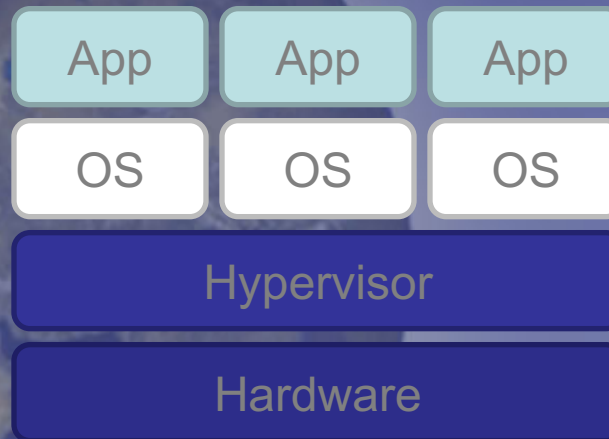




Virtualization as an Infrastructure



Traditional Stack



Virtualized Stack



Everything as a Service

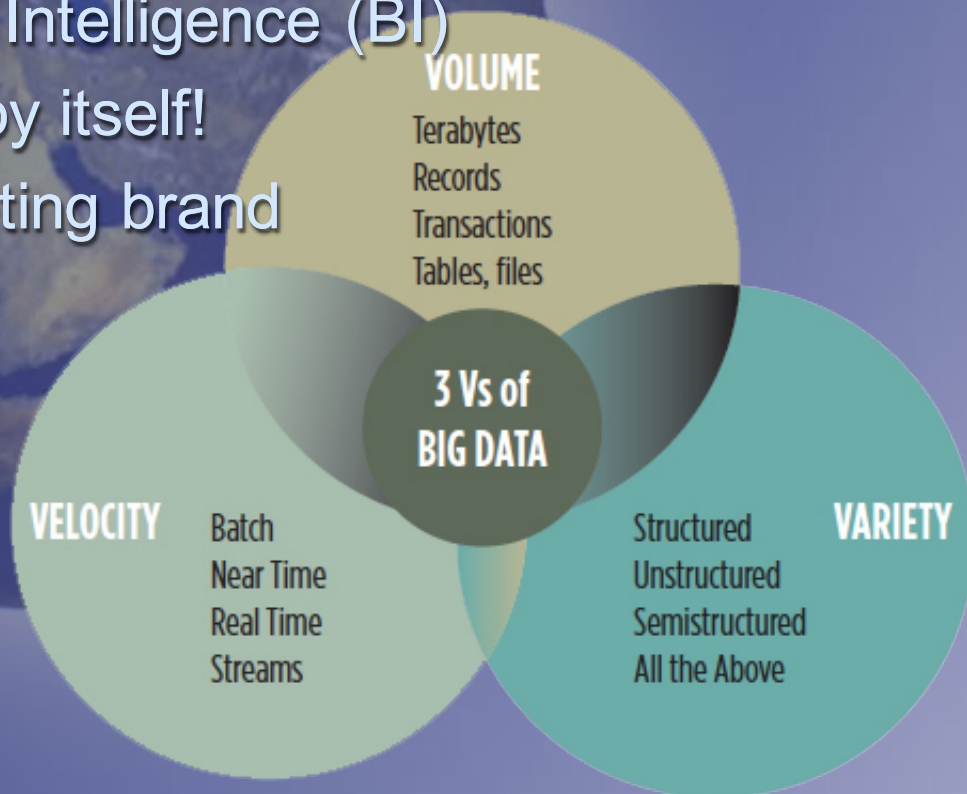
- Utility computing = Infrastructure as a Service (**IaaS**)
 - Why buy machines when you can rent cycles?
 - Examples: Amazon's EC2 (Elastic Compute Cloud), Rackspace, Microsoft Azure
- Platform as a Service (**PaaS**)
 - Give me nice API and take care of the maintenance, upgrades, ...
 - Example: Google App Engine
- Software as a Service (**SaaS**)
 - Just run it for me!
 - Example: Gmail, Salesforce
- Data as a Service (**DaaS**)
- Security as a Service (**SECaaS**)



What Big Data is and isn't?

Computing + Internet = Big Data

- Big Data is not new technology
- Big Data is not just about size
- Big Data is not Business Intelligence (BI)
- Big Data is not Solution by itself!
- Big Data is mostly marketing brand





Interdisciplinary Subfields of Computer Science

- Artificial Intelligence,
- Machine Learning,
- Statistics,
- Applied Mathematics,
- Text Mining,
- Database Systems,
- Business Intelligence,
- Computational Linguistics,
- Natural Language Processing (NLP),
- Information Theory And Information Technology,
- Signal Processing,
- Probability Models,
- Statistical Learning,

- Data Mining,
- Data Engineering,
- Pattern Recognition and Learning,
- Information Visualization,
- Regular Expressions,
- Predictive Analytics,
- Uncertainty Modeling,
- Data Warehousing,
- Data Compression,
- Computer Programming,
- High Performance Computing,
- Distributed Systems,
- Information Extraction,
- Cloud Computing,
- Computer Vision



Subfields involved in Unstructured Data Management

Platform & Data Management

- Data Warehousing,
- Data Compression,
- High Performance Computing,
- Distributed Systems,
- Cloud Computing,
- Data Mining,
- Data Engineering,
- Information Extraction,
- Computer Vision
- Database Systems,
- Business Intelligence,
- Information Theory

AI & NLP

- Artificial Intelligence,
- Machine Learning,
- Text Mining,
- Computational Linguistics,
- Natural Language Processing (NLP),
- Regular Expressions
- Statistical Learning,
- Pattern Recognition and Learning,
- Information Visualization,
- Predictive Analytics,
- Uncertainty Modeling,
- Sentiment Analysis



Jobs Derived from Big Data

- Chief Data Officer,
- Big Data Solution Architect,
- Big Data Platform Engineer,
- Big Data Analyst,
- Big Data Analytics Business Consultant,
- Big Data Software Designer,
- Big Data Consultant,
- Hadoop Architects,
- Consultant Hadoop Developer,
- Senior Analytics Manager,
- Data & Reporting Analyst,
- Analytics Analyst (Big Data)





Jobs Opportunities

- According to International Data Corporation (IDC), the Big Data and Analytics market reached \$128 billion worldwide in 2015, Growth 6 times faster than IT.





Demand for Data Scientists

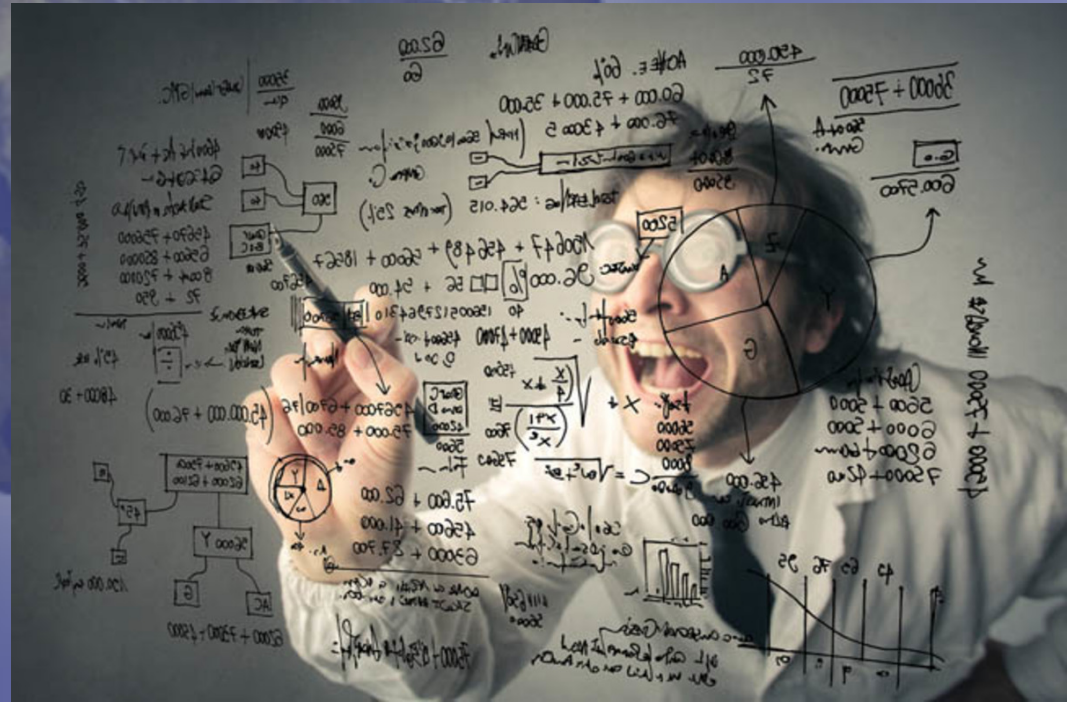
Job Trends from Indeed.com

— "Data Scientist"



Jobs Opportunities

- According to McKinsey Global Institute, 2018 the United States alone could face a shortage of 140,000 – 190,000 people with Big Data Analytics/Development skills as well as 1.5 million managers with the know-how to use analysis of Big Data to make effective decisions.





Big Data and Industry 4.0

"The growth of data streams and data analysis will generate significant developments in the 21st century. Big data will play a key role in shaping the design of products, systems and services by making them smarter, better connected, more efficient and widely accessible. Data will become a powerful economic engine." – Industrial Engineer magazine

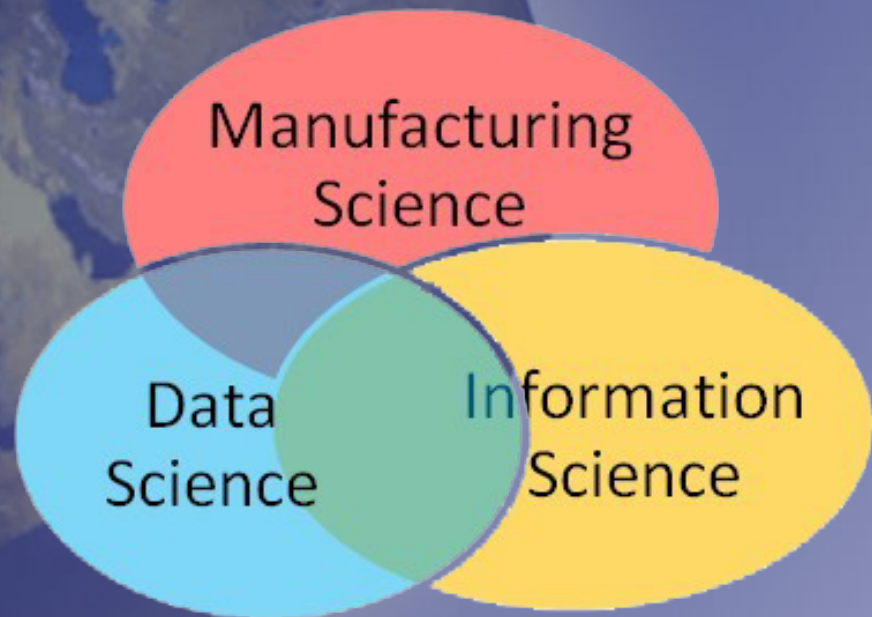




'5Vs' in Industry Engineering

Big Data as a holistic approach to manage, process in Industrial Engineering and analyze

- Volume
- Velocity
- Variety
- Veracity
- Value



Transformation through Industry 1.0 to 4.0



1st

2nd

3rd

4th

Mechanization,
water power, steam
power

Mass production,
assembly line,
electricity

Computer and
automation

Cyber Physical
Systems

Transformation through Industry 1.0 to 4.0





Components of Industry 4.0

- **Technology**

- Cloud Computing
- Big Data
- Industrial IoT
- Wireless

- **Collaboration**

- Integrated Industries
- Social Innovation

- **Processes**

- Sustainable Manufacturing
- Lifecycle Assessments
- Internet of Services



Data-Driven Decision Making (DDD)

Data alone won't change the world. It's the people that use data to make better decisions.



Data-driven decision making (DDD) refers to the practice of basing decisions on the analysis of data rather than purely on intuition.

Natural Language Processing (NLP)



- Natural Language Processing (NLP)
- Computational Linguistics (CL)
- Machine Translation (MT)



Natural Language Processing (NLP)

- Multilingual NLP
- Text Mining in Multimedia Networks
- Mining Text Streams
- Text Mining in Social Media
- Cross-Lingual Mining of Text Data
- Contextual analysis of text data

Some of available NLP tools: NLTK, Apache OpenNLP, MontyLingua, VisualText, etc.

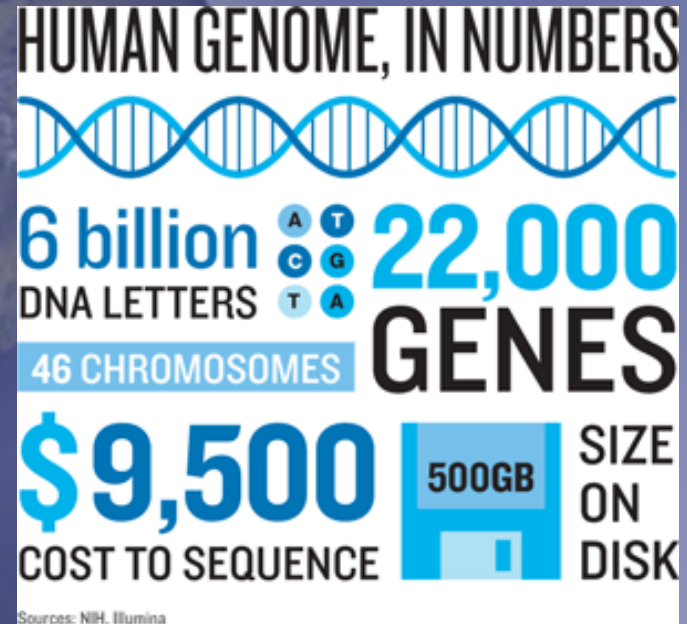
Data Science in Chemistry

In the era of Big Data Medicinal Chemists and Bio-Informatics are exposed to an enormous amount of data.



Data Science in Medicine

Data alone won't change the world. It's the people that use data to make better decisions.

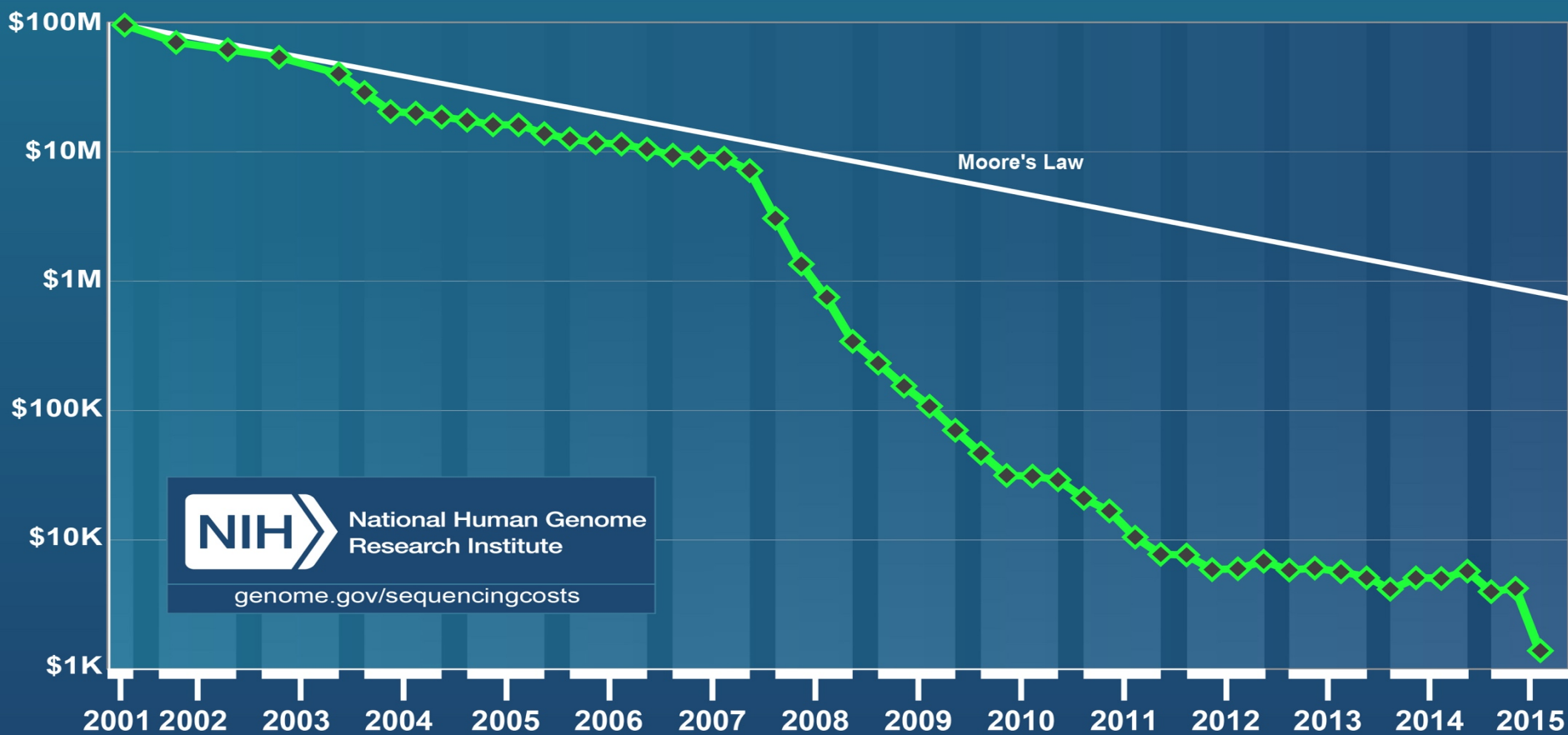


Cost to sequence a human genome in 2008 was almost \$10 million, get down less than \$1,000 today.



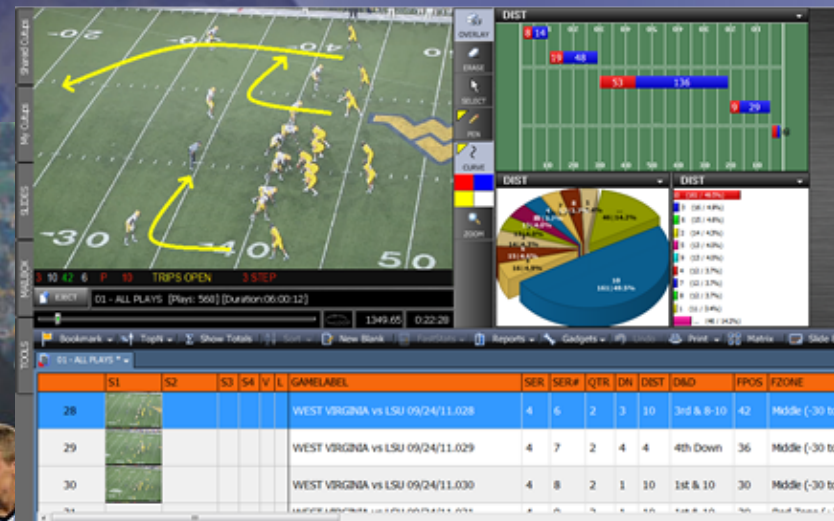
Data Science in Medicine

Cost per Genome



Data Science in Sports

Big Data & Data Analytics Help Germany Score the World Cup



	S1	S2	S3	V	L	GAME/REL	SER	SER#	QTR	DR	DIST	D&D	FPOS	FZONE
28						WEST VIRGINIA vs LSU 09/24/11.028	4	6	2	3	10	3rd & 8-10	42	Middle (-30 to
29						WEST VIRGINIA vs LSU 09/24/11.029	4	7	2	4	4	4th Down	36	Middle (-30 to
30						WEST VIRGINIA vs LSU 09/24/11.030	4	8	2	1	10	1st & 10	30	Middle (-30 to



How Industries can benefit from Data Mining and Analysis?

- FINANCE
- BANKING
- RETAIL
- TELECOM
- ENERGY
- MARKETING
- GOVERNMENT
- HEALTHCARE
- SECURITY





Data Science Application

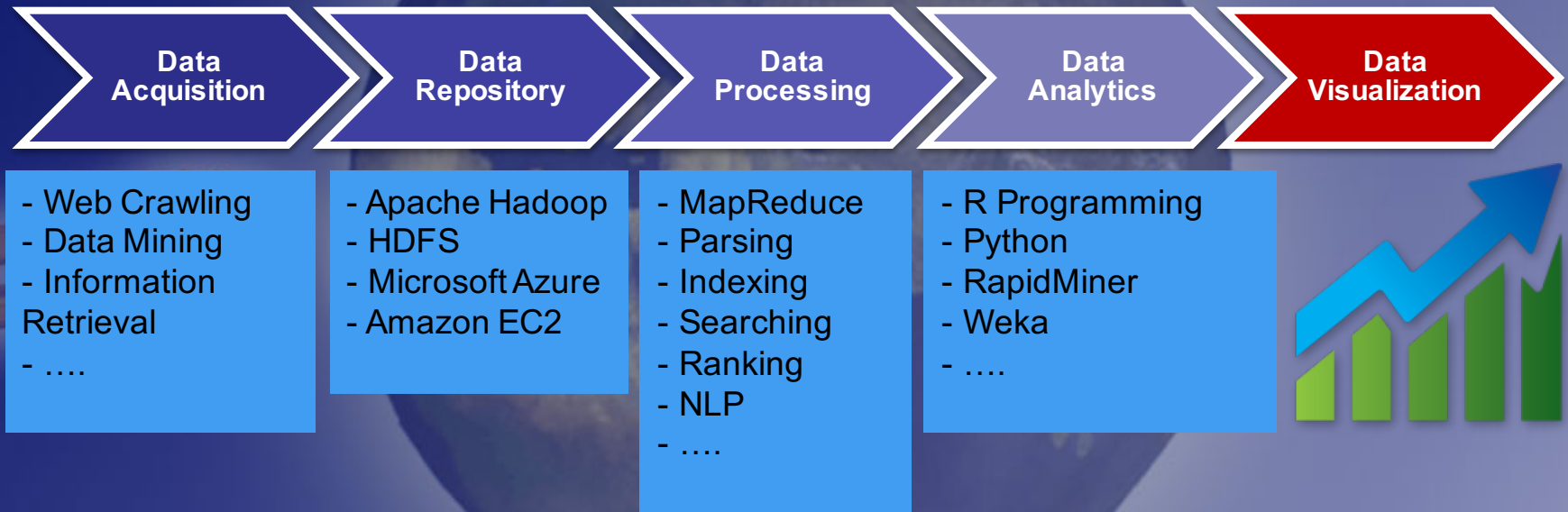
Data-driven decision making (DDD)

- Direct Marketing,
- Online Advertising,
- Credit Scoring and Risk Management
- Help Desk Management
- Fraud Detection
- Search Ranking
- Product Recommendation
- Predicting Unusual Behavior
- Customer Retention in Telecom





Big Data Management Life-Cycle



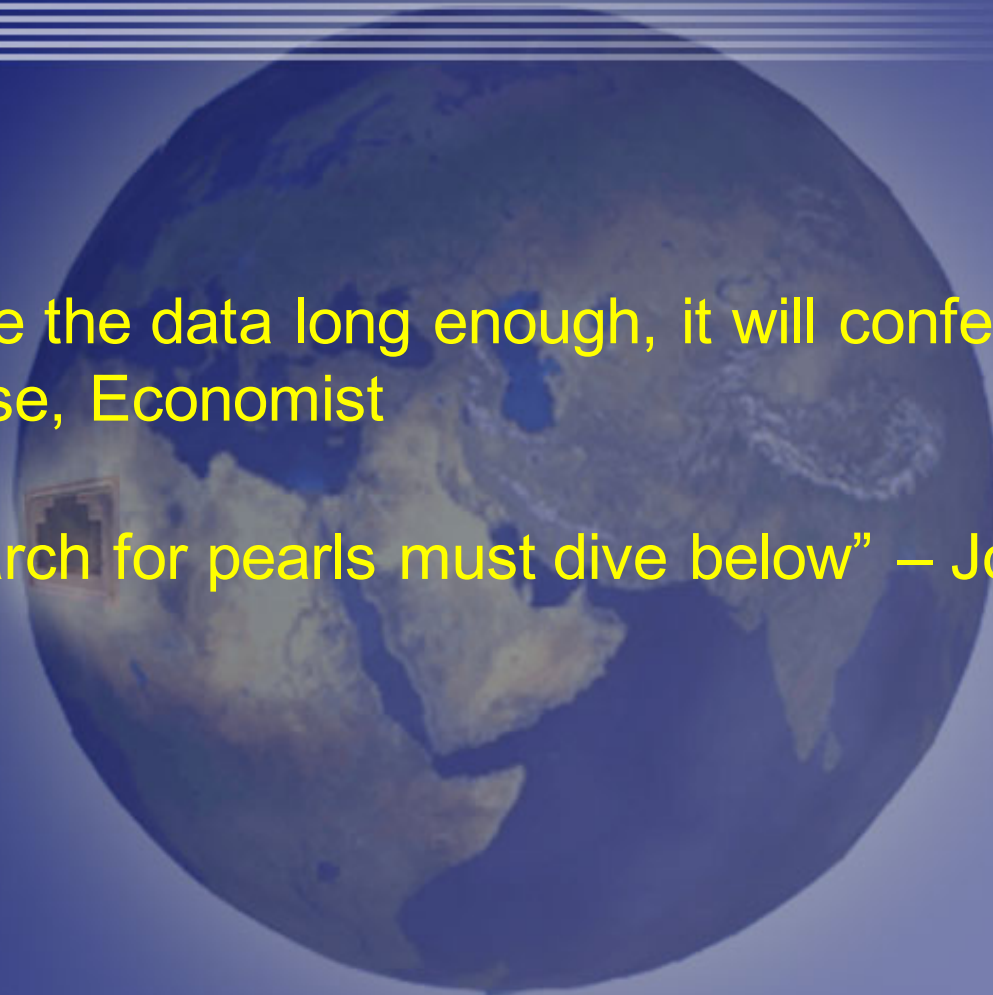
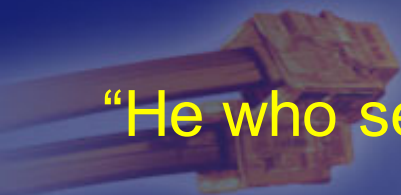
Big Data Management involves Data Science and Data Engineering areas for implementing Data Mining Techniques



Quotes on Big Data

“If you torture the data long enough, it will confess.” –
Ronald Coase, Economist

“He who search for pearls must dive below” – John Dryden





Big Data Day Baku 2015

- “Transforming Big Data into Big Value”
- 15 May 2015, www.cedawi.org/BDDDB2015





Big Data Day Baku 2016

- “Transforming Big Data into Big Value”
- 17 May 2016, www.cedawi.org/BDDDB2016



Big Data Day Baku 2016

- “Transforming Big Data into Big Value”
- 17 May 2016, www.cedawi.org/BDDDB2016



	<p>Assoc.Prof., Abzedin Adamov, Computer Engineering Dep. at the Qafqaz University Founding Director, Applied Research Center for Data Analytics and Web Insights (CeDAWI) Speech Title: <i>Quality Corpus Generation for NLP and Linguistic Research</i></p> <p>in</p>
	<p>Mark Torr, Senior Product Marketing Manager, IOT Incubation, Microsoft, Germany Speech Title: <i>Delivering Value from the Internet of Your Things and Big Data</i></p> <p>in</p>
	<p>Pavel Shkjudov, Advanced Technologies Leader Europe Cognitive Solutions Team – IBM Watson Group, Russia Speech Title: <i>Cognitive Computing – Undiscovered Territory.</i></p> <p>in</p>
	<p>Prof. Eshref Adali, Istanbul Technical University, Turkey Speech Title: <i>Natural Language Processing and Semantic Analysis</i></p> <p>in</p>
	<p>Jamal Shahverdiyev, Head of IT department at ATL Group Speech Title: <i>Cloud Computing and Hadoop on top of OpenStack</i></p> <p>in</p>
	<p>Roman Chesov, CEO, FLEXBBY Solutions, Russia Speech Title: <i>Containerized Architecture for Software as a Service Applications Development</i></p> <p>in</p>
	<p>Togrul Jajarov, Data Analyst State Agency for Public Services and Social Innovations (ASAN), Azerbaijan Speech Title: <i>Effective Data Management using R</i></p> <p>in</p>

#BDDDB2016

ORGANIZED BY: **CeDAWI**
Center for Data Analytics and Web Insights

SUPPORTED BY: **barama**

SUPPORTED BY: **QU TECHNOPARK**

SUPPORTED BY: **region 8 IEEE**

SUPPORTED BY: **IEEE computer society Azerbaijan Chapter**

SUPPORTED BY: **IEEE COMMUNICATIONS SOCIETY Azerbaijan Chapter**

info@cedawi.org

BDDDB2016
transforming Big Data into Big Value...

QAFQAZ UNIVERSITY
17 MAY 2016
www.CeDAWI.org/BDDDB2016
www.fb.com/CeDAWI

PRIMARY SPONSOR: **Azercell**

PARTNERS: **IBM** **ORACLE** **Microsoft**

MEDIA PARTNERS: **INFOCITY** **itrend** **Day.Az** **Milli.Az**

AZERNEWS **spicyblog.az** **METBUAT.AZ** **StarStar.AZ** **Publi.AZ** **REPORT** **APRESS.AZ**



Computing Facilities at CeDSRT

3RD INTERNATIONAL FORUM **BIG DATA DAY** - BAKU 2017 -

ORGANIZED BY:

CEDSRT

Center for Data Science Research and Training



CeDAWI

Center for Data Analytics and Web Insights

MEDIA PARTNER:

INFOCITY

technics & technology magazine

OFFICIAL PARTNER:

bp



30 MAY, 2017

ADA UNIVERSITY

WWW.CEDAWI.ORG/BDDDB2017



INNOVATION PARTNER:



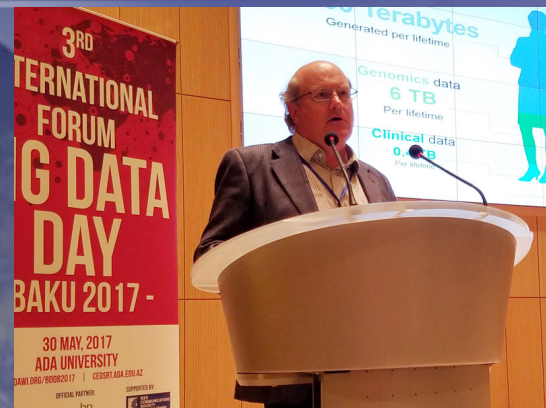
INDUSTRIAL PARTNER:

ORACLE

SUPPORTED BY:



Big Data Day Baku 2017





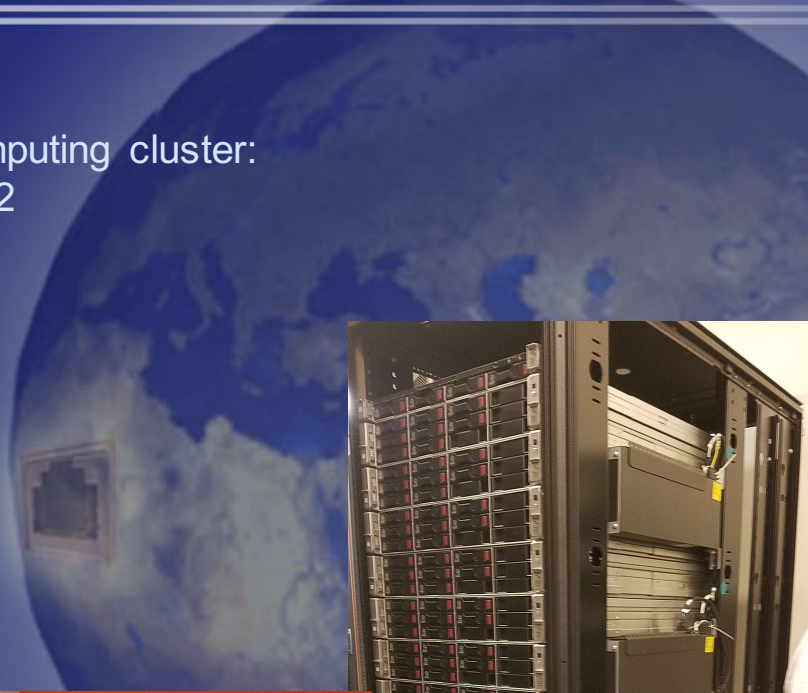
Computing Facilities at CeDSRT

Characteristics of computing cluster:

Processing Cores: 102

Memory: 1,568 TB

Storage: 136 TB





References

- Awesome Data Science (Infographics)
<https://github.com/bulutyazilim/awesome-datascience>





Thank you



info@CeDAWI.org



www.CeDAWI.org



fb.com/CeDAWI