Characteristics of Big Data and Dimensions of Scalability

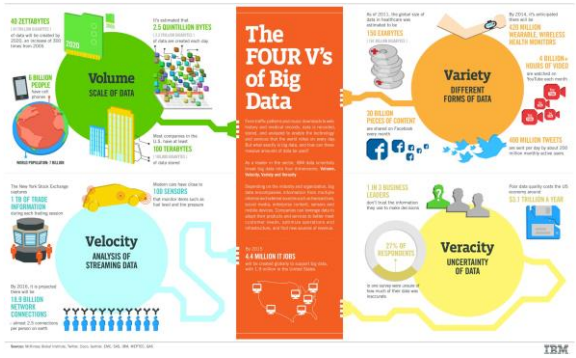
Pejman Rasti

Email: prasti@esaip.org
pejman.rasti@univ-angers.fr

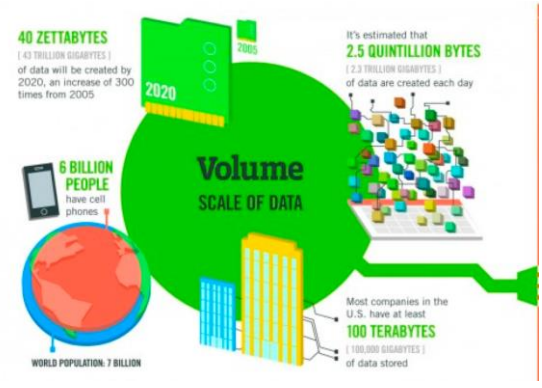
Course Website: Access from your "Moodle" portal

1

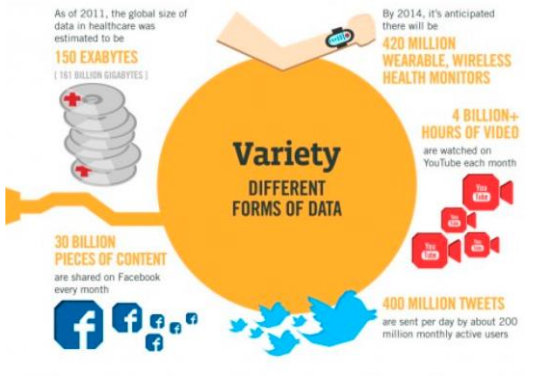
Characteristics of Big Data



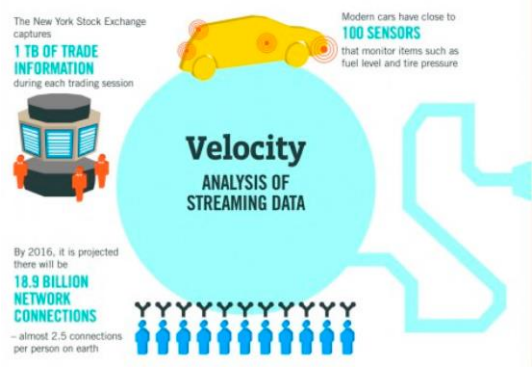
2



3

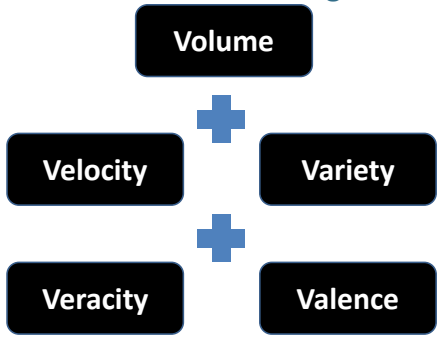


4

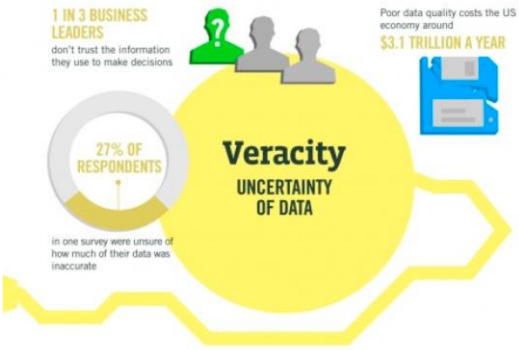


5

Characteristics of Big Data

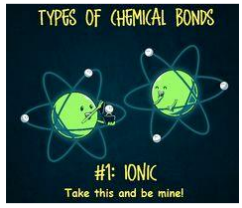


6



7

Valence



8

Characteristics of Big Data

Volume + **Value**

Velocity **Variety**

Veracity **Valence**

9

Volume

Volume = Size

10

Every minute…

204 Million emails

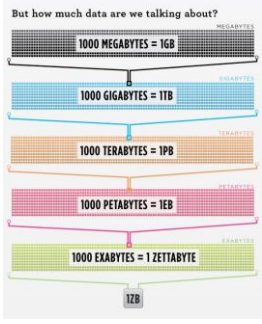
200,000 photos

1.8 Million likes

1.3 Million video views

72 hours of video uploads

11



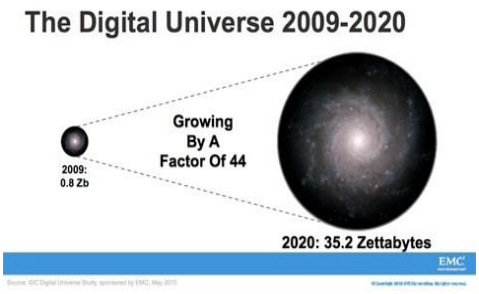
100 MBs ~= couple of volumes of Encyclopedias

A DVD ~= 5 GBs

1 TB ~= 300 hours of good quality video

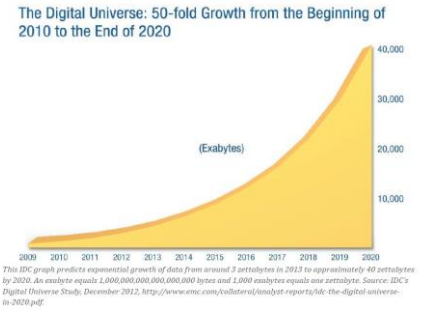
LHC ~= 15 PBs a year

12



13

Exponential data growth!



This IDC graph predicts exponential growth of data from around 3 zettabytes in 2013 to approximately 40 zettabytes by 2020. An exabyte equals 1,000,000,000,000,000,000 bytes and 1,000 exabytes equals one zettabyte. Source: IDC's Digital Universe Study, December 2012, http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

14

Relevance of Volume for Us?

More data = Better safety

15

16

Challenges: Storage and more…

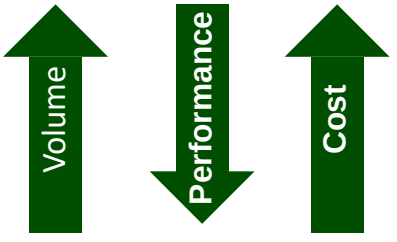
Storage → **Distribution** → **Processing**

Data acquisition

Retrieval

17

Processing Big Data



18

Scalability - Variety

Variety == Complexity

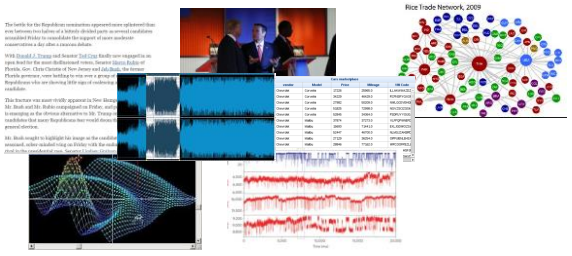
Data were confined onlyto tables

Cars marketplace

vendor	Model	Price	Mileage	VIN Code
Chevrolet	Corvette	17226	25965.0	ILLAKAWAZDZ
Chevrolet	Corvette	34229	46429.0	RCPHSRYGXKR
Chevrolet	Corvette	27982	50209.0	NWLGCEVEHGI
Chevrolet	Corvette	51825	72998.0	NGVZSCUZGSN
Chevrolet	Corvette	52845	34364.0	PSDRUYYOUJG
Chevrolet	Malibu	37874	37273.0	VLFPQPWMEFC
Chevrolet	Malibu	15600	71441.0	EXLJGDWOZSU
Chevrolet	Malibu	52447	46700.0	NLMGJZAKBPC
Chevrolet	Malibu	27129	36254.0	OIPFUIENLEHDH
Chevrolet	Malibu	28846	77182.0	WRCOOPREZLJ
Chevrolet	Malibu	46165	60590.0	HJFTTHQHSFJF
Chevrolet	Malibu	18763	37790.0	JLMHIAFSHUT

19

Today, Data are more heterogeneous



20

Axes of Data Variety

Structural Variety – formats and models	**Semantic Variety – how to interpret and operate on data**
Media Variety – medium in which data get delivered	**Availability Variations – real-time? Intermittent?**

21

Variety within a Type

- Think of an email collection
  - Table-like part

from:	**Banikumar Maiti (GMAIL)** <banikumar.maiti@gmail.com>
to:	Reghu Rajan <reghurajan@gmail.com>
cc:	Amarnath Gupta <aguptasd@gmail.com>
date:	Tue, Feb 2, 2016 at 2:29 PM
subject:	Re: Connecting
mailed-by:	gmail.com
signed-by:	gmail.com

22

Variety within a Type

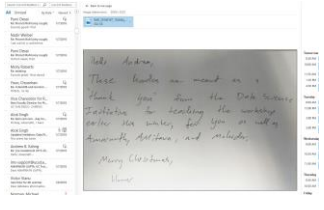
- Think of an email collection
  - Sender, receiver, date… Well-structured
  - Unstructured Text

Dear All,
I would like to congratulate you for putting together a wonderful show.
It was only possible by your hard work.
Dreaming of an UNIQUE show. This credit goes to Zubair. You
dreamed about it and made it happen.

23

Variety within a Type

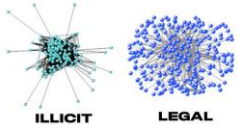
- Think of an email collection
  - Sender, receiver, date… Well-structured
  - Body of the email Text
  - Media



24

Variety within a Type

- Think of an email collection
  - Sender, receiver, date… Well-structured
  - Body of the email Text
  - Attachments Multi-media
  - Who-sends-to-whom



25

Variety within a Type

- Think of an email collection
  - Sender, receiver, date… Well-structured
  - Body of the email Text
  - Attachments Multi-media
  - Who-sends-to-whom Network
  - A current email cannot reference a past email Semantics

26

Variety within a Type

- Think of an email collection
  - Sender, receiver, date… Well-structured
  - Body of the email Text
  - Attachments Multi-media
  - Who-sends-to-whom Network
  - A current email cannot reference a past email Semantics
  - Real-Time? Availability

27

Scalability Issues

- Impact of data variety
  - Harder to ingest
  - Difficult to create common storage
  - Difficult compare and match data across variety
  - Difficult to integrate
  - Management and policy challenges

28

Velocity

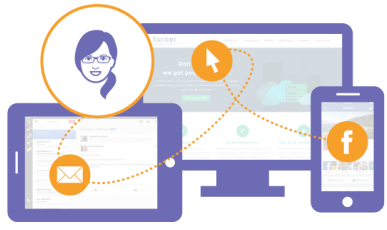
Velocity == Speed

$$\bar{v} = \frac{\Delta x}{\Delta t}$$

Speed of creating data
Speed of storing data
Speed of analyzing data

29

Big Data → **Real-time action**



30

Late decisions

Missing opportunities

31

32

How to decide what to pack ?

Use weather information of last year at this time?

33

How to decide what to pack ?

Use weather information of last month ?

OR

Use weather status of this week or today ?

34

Action

35

Real-time Processing

Batch Processing

Collect Data → Clean Data → Feed in Chunks → Wait → Act

36

Real-time Processing

Real-Time Processing

Instantly capture streaming data → Feed real time to machines → Process Real Time → Act

37

Batch Processing → **Incomplete**

Real-Time Processing → **Fast**

38

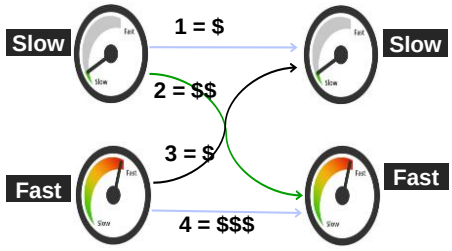
Rate needed for data-driven actions

↓

Rate of generation and processing of data

39

Speed of Data Generation | **Speed of Data Processing**



Which path to choose?

40

Veracity == Quality

Validity

Volatility

42

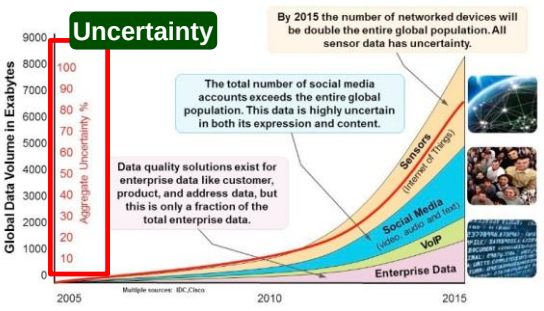
Veracity == Quality

Accuracy of data

Reliability of the data source

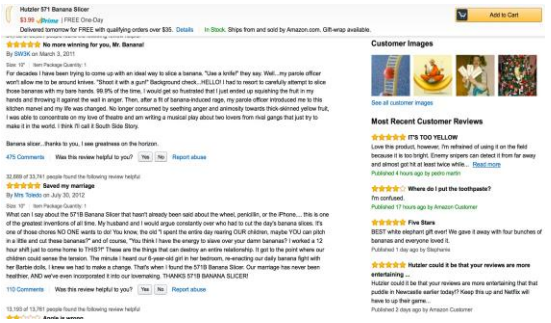
Context within analysis

43

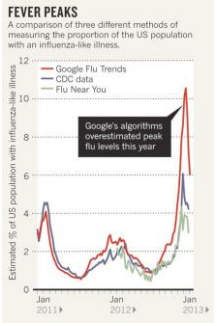


44

When sentiment analysis doesn't work?



45



Uncertainty

Provenance

46

Veracity == Quality

Accuracy of data

Reliability of the data source

Context within analysis

?

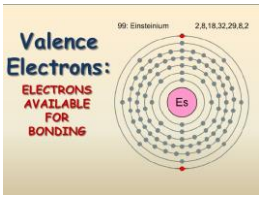
Uncertainty

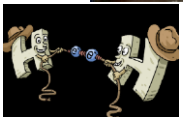
Provenance

47

Valence == Connectedness

Valence – a Concept from Chemistry





48

Valence – Measure of Connectivity

Data Connectivity

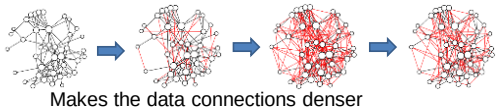
- Two data items are connected when they are related to each other

- Valence
  - Fraction of data items that are connected out of total numbe possible connections

49

Why worry about Valence?

Valence increases over time



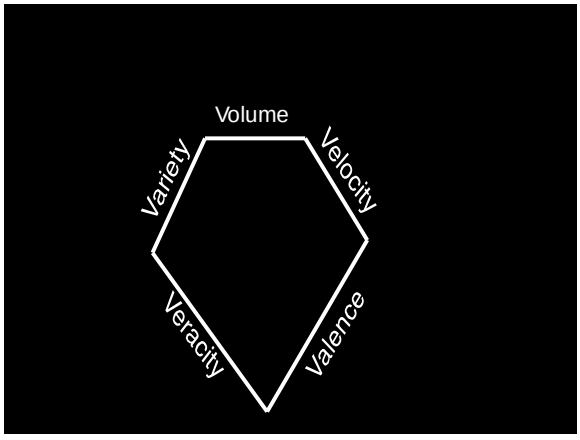
Makes the data connections denser

50

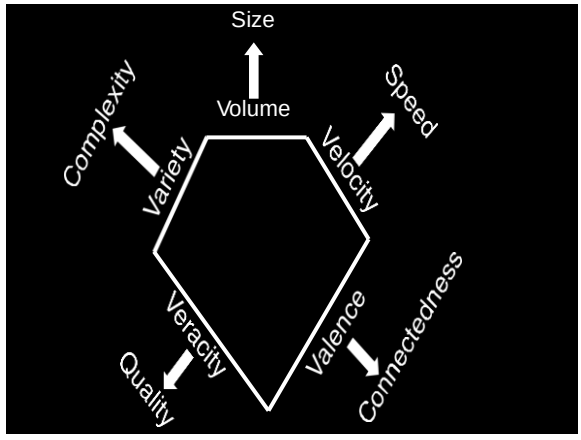
Valence: Challenges

- More complex data exploration algorithms
- Modeling and prediction of valence changes
- Group event detection
- Emergent behavior analysis

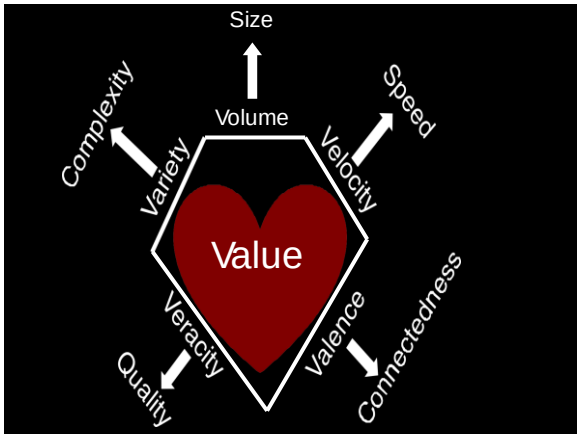
51



52



53



54

Eglence Inc. Big Data Case:

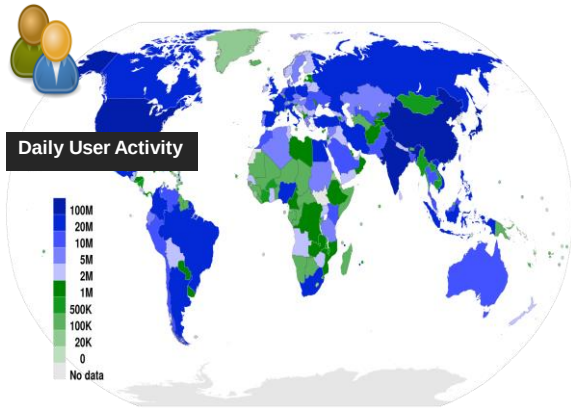
Catch The Pink Flamingo

55

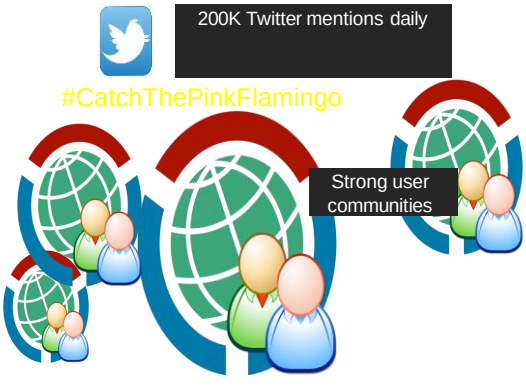
56

57

58



59



60

61

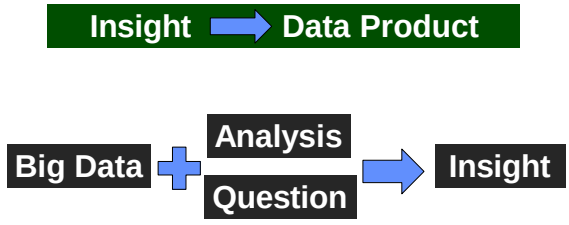
Data Source	
Machine	• User activity logs
People	• Twitter conversations
Organization	• User demographic info • Game stats

62

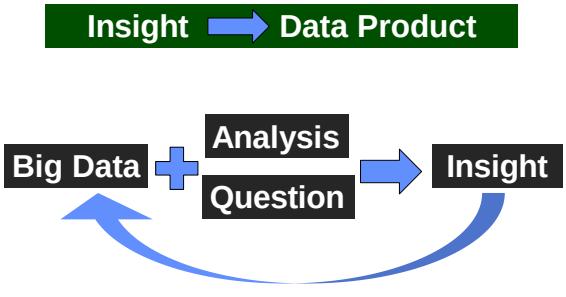
Dimension	
Volume	• Big daily workload and associated data on players and game stats
Variety	• Multiple types of data
Velocity	• Real-time analysis of usage activity
Veracity	• Demographic info not accurate
Valence	• Connections between players

63

64



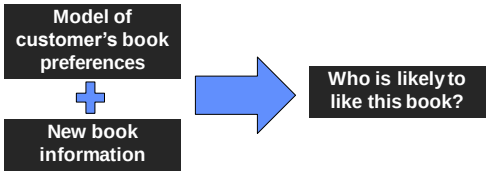
65



66

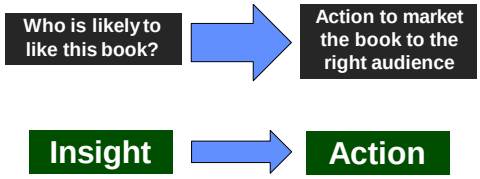
67

Find Potential Audience for a Book



68

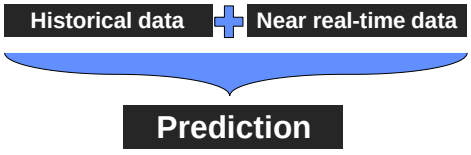
Market a New Book



Insight → **Action**

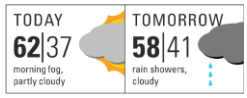
69

Actionable Information



70

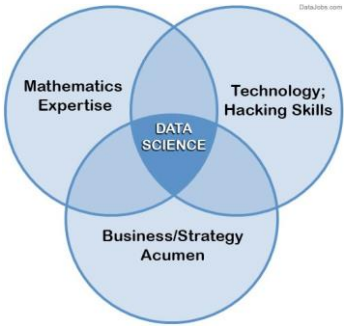
Prediction



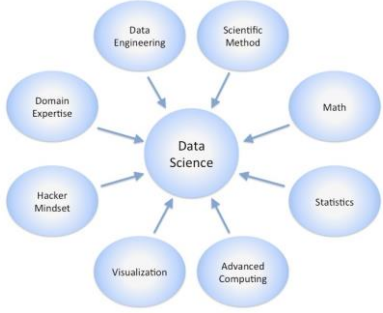
Action

71

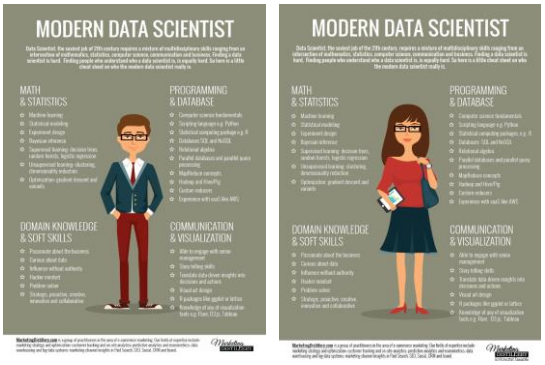
Data Science is Team Work!



72

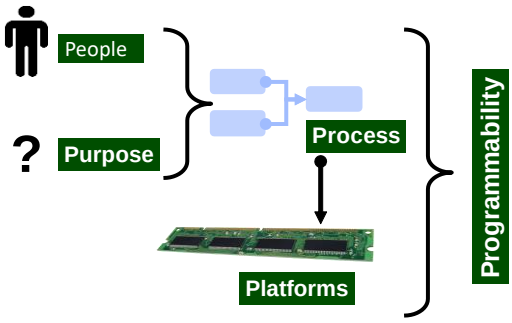


73

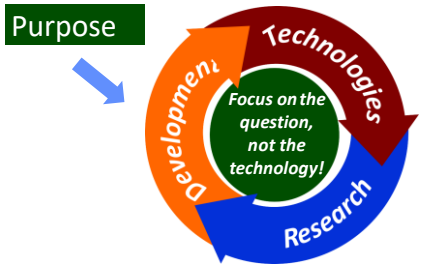


74

Five P's of Data Science



75

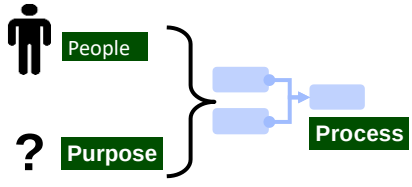


76

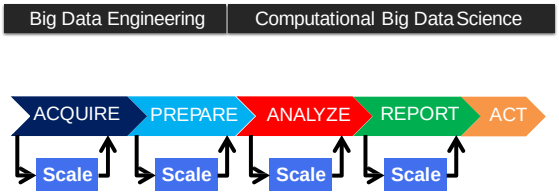
77

Let's not dive into the techniques yet! What is the problem at large? How do you see yourself solving it?

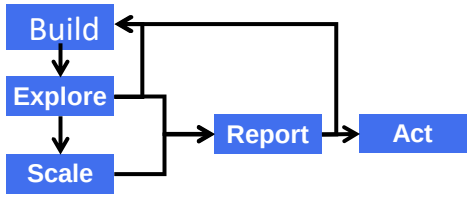
78



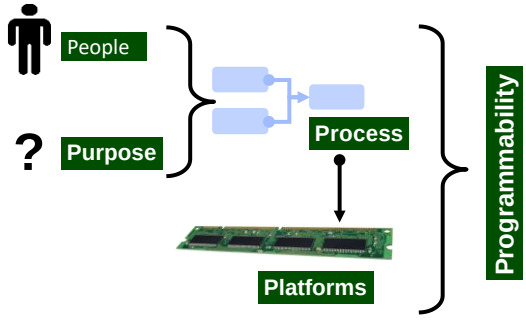
79



80



81

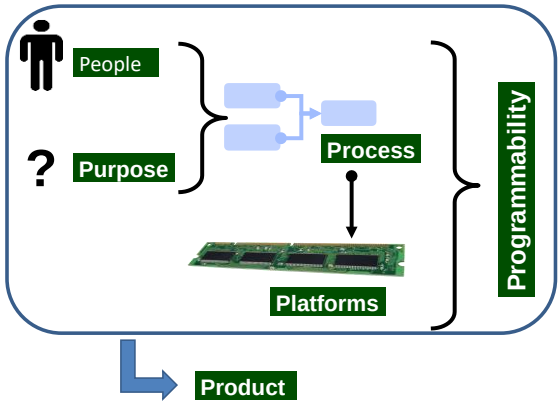


82

Process: Build metrics for accountability

- Cost
- Timeline
- Planning of deliverables
- Expectations
- Purpose

83



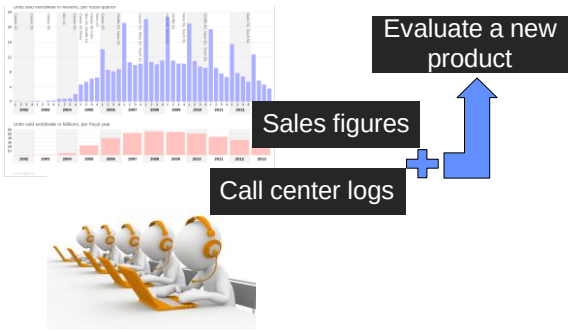
84

Define the Problem

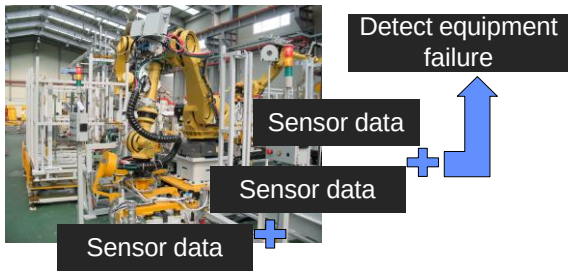
"A problem well defined is a problem half solved."

Charles F. Kettering

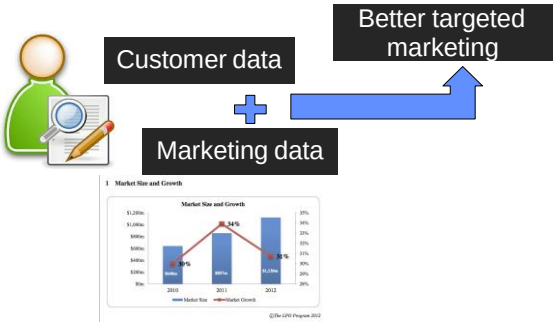
85



86

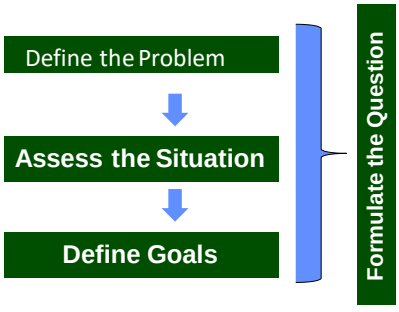


87



88

89



90

Steps in the Data Science Process

91

ACQUIRE PREPARE ANALYZE REPORT ACT

Step 1: Acquire Data

- Identify data sets
- Retrieve data
- Query data

92

ACQUIRE PREPARE ANALYZE REPORT ACT

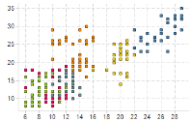
Step 2: Prepare Data

- Step 2-A: Explore
- Step 2-B: Pre-process

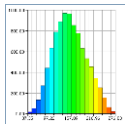
93

ACQUIRE PREPARE ANALYZE REPORT ACT

Step 2-A: Explore Data



Understand nature of data

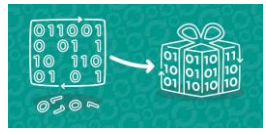


Preliminary analysis

94

ACQUIRE PREPARE ANALYZE REPORT ACT

Step 2-B: Pre-process Data



Clean Integrate Package

95

ACQUIRE PREPARE ANALYZE REPORT ACT

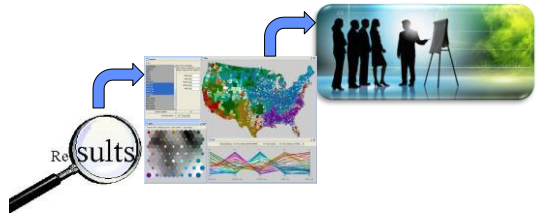
Step 3: Analyze Data

- Select analytical techniques
- Build models

96

ACQUIRE PREPARE ANALYZE REPORT ACT

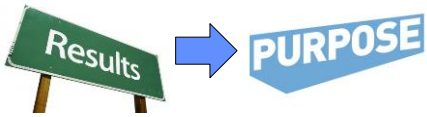
Step 4: Communicate Results



97

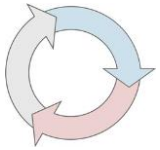
ACQUIRE PREPARE ANALYZE REPORT ACT

Step 5: Apply Results



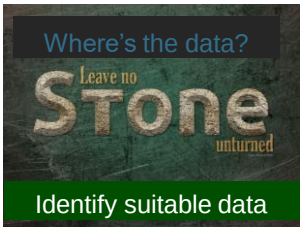
98

ACQUIRE PREPARE ANALYZE REPORT ACT



Iterative process

99



Identify suitable data

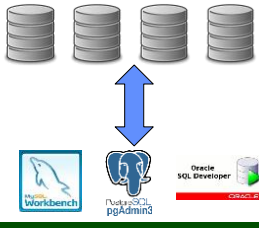
Acquire all available data

100

Data comes from many places…

…with many ways to access it

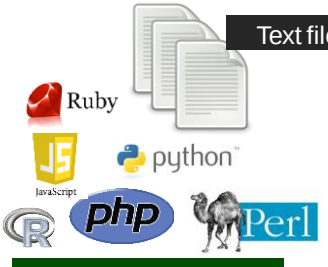
101



Traditional databases

SQL and query browsers

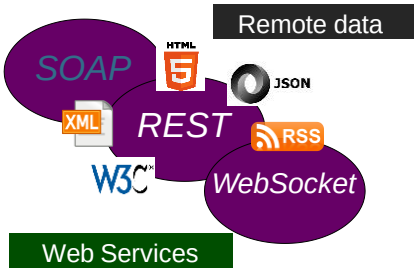
102



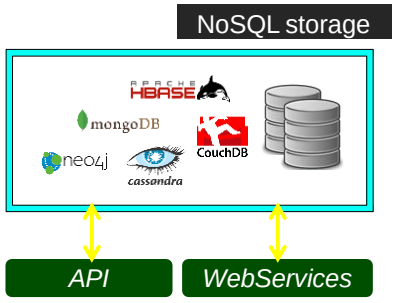
Text files

Scripting languages

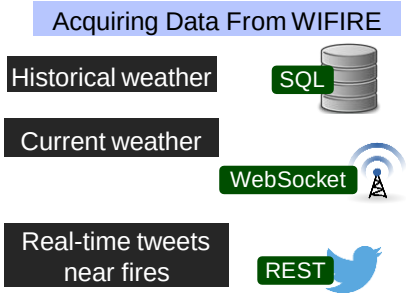
103



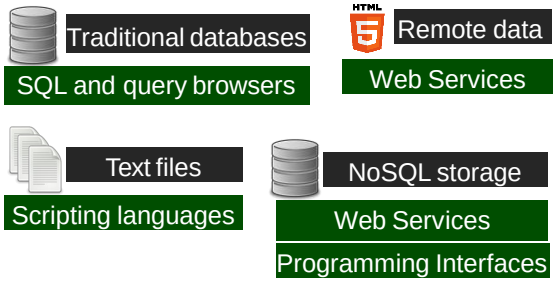
104



105

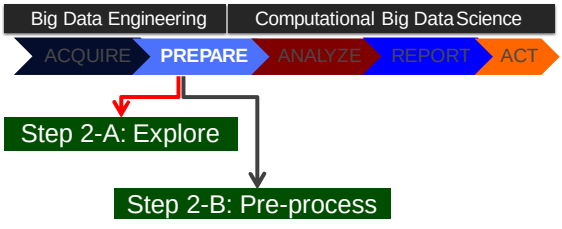


106



107

Exploring Data

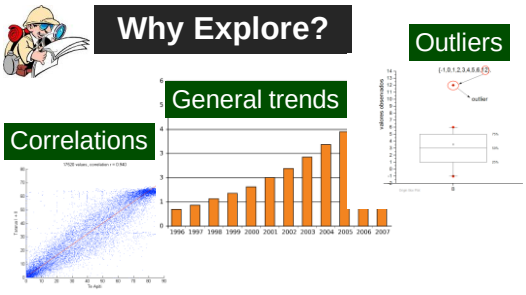


108

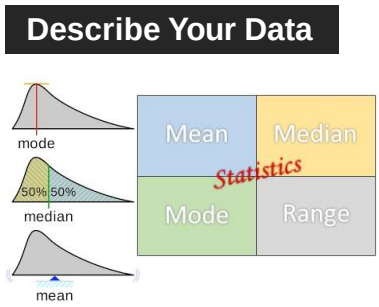
Why Explore?

Goal: Understand your data

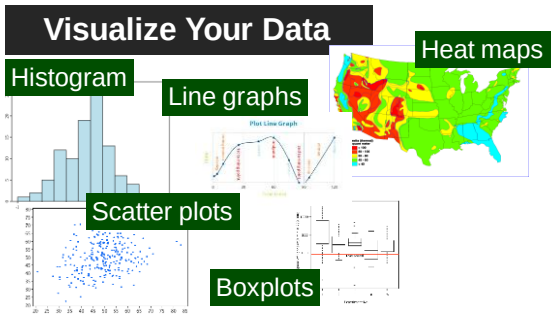
109



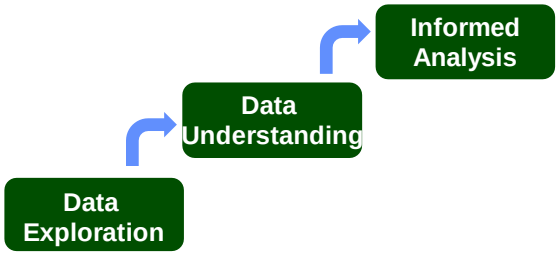
110



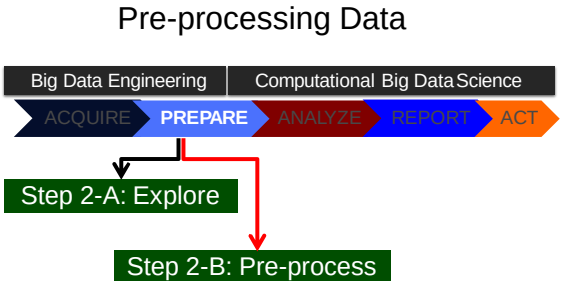
111



112



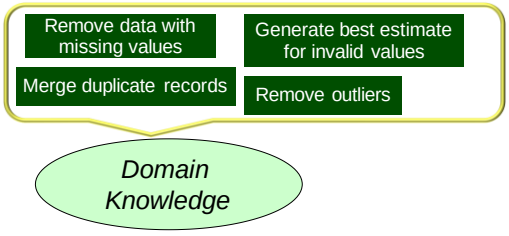
113



114

115

Addressing Data Quality Issues

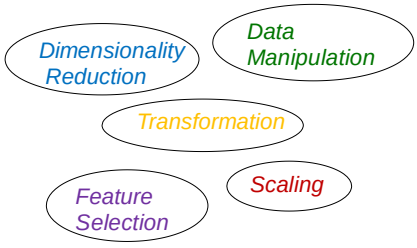


116

Getting Data in Shape

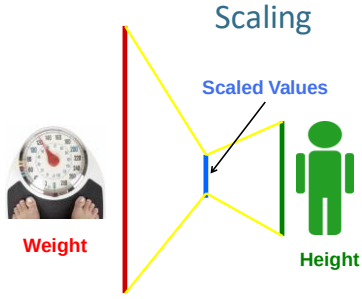
117

Data Munging



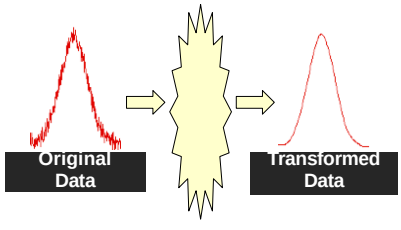
118

Scaling



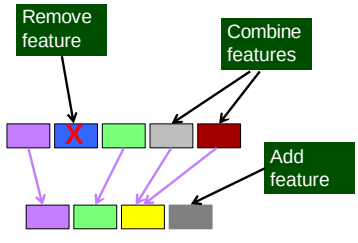
119

Transformation



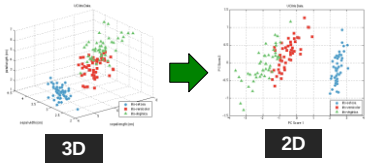
120

Feature Selection



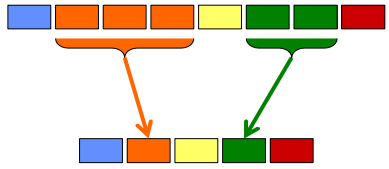
121

Dimensionality Reduction



122

Data Manipulation



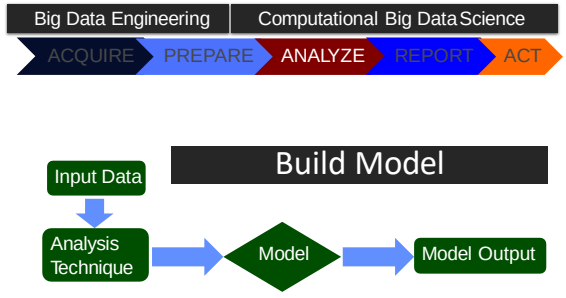
123

Always Remember!

Garbage in = Garbage out

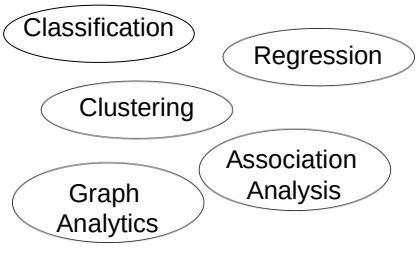
Data preparation is very important for meaningful analysis!

124



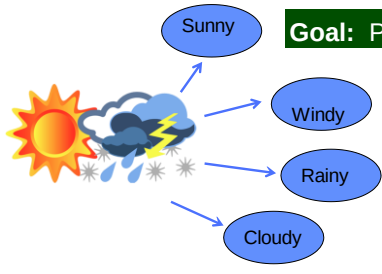
125

Categories of Analysis Techniques



126

Classification



127

Regression

Goal: Predict numeric value

128

Clustering

Goal: Organize similar items into groups

129

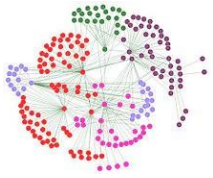
Association Analysis

Goal: Find rules to capture associations between items

130

Graph Analytics

Goal: Use graph structures to find connections between entities



131

Modeling

Select technique

Build model

Validate model

132

Evaluation of Results

133

Classification & Regression

Predicted Value VS Correct Value

134

Clustering

135

Association Analysis & Graph Analytics

Investigate

Validate

136

Determine Next Steps

Repeat analysis?

Take deeper dive?

Act on results?

137

Select technique

Build model

Evaluate

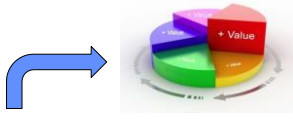
Classification
Regression
Clustering
Association Analysis
Graph Analytics

138

Big Data Engineering | Computational Big Data Science

ACQUIRE PREPARE ANALYZE REPORT ACT

What to Present

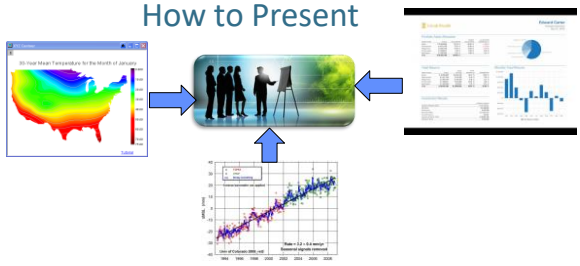


139

What to Present

140

How to Present



141

Visualization Tools

R python D3

Google Developers CHARTS tableau

Timeline JS
Beautifully crafted timelines that are easy and intuitive to use.

142

Present

with

using

tableau

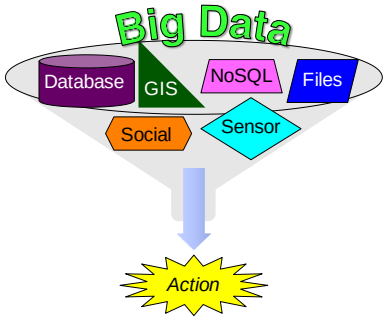
Google Developers CHARTS

Power BI

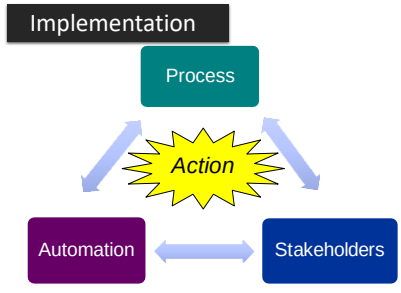
143

Big Data Engineering | Computational Big Data Science

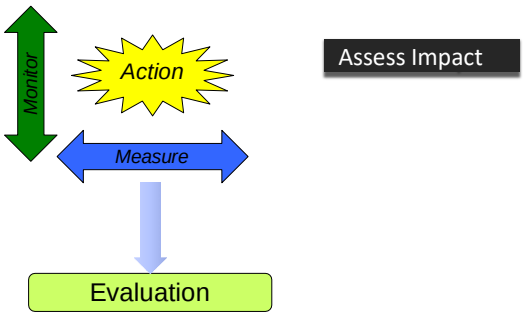
ACQUIRE PREPARE ANALYZE REPORT ACT



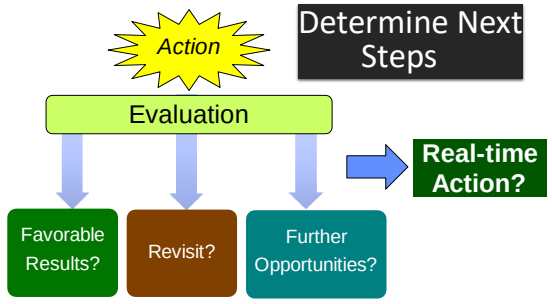
144



145



146



147