Introduction Big Data

Pejman Rasti

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

Course Website: Access from your "Moodle" portal

1

Who am I ?

- Assoc. Prof. at ESAIP
- Researcher at the university of Angers

- Research Interests:
  - Artificial Intelligence
  - Deep learning
  - Data analysis

Webpage: http://perso-laris.univ-angers.fr/~rasti/

2

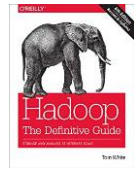
Expectations

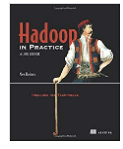
- Please be on time.
- Please pay attention.
- Students are expected and encouraged to ask questions in class!

3

References

- **Main:**
  - White, Tom. Hadoop: The definitive guide. " O'Reilly Media, Inc.", 2017. 4th edition
- **Complementary:**
  - Holmes, Alex. Hadoop in practice. Manning Publications Co., 2012.





4

Course Objective / Requirement

- **Objective:**
  - Understand the Big Data Platform and its Use cases
  - Provide an overview of Apache Hadoop
  - Provide HDFS Concepts and Interfacing with HDFS
  - Understand Map Reduce Jobs
  - Provide hands on Hodoop Eco System
  - Apply analytics on Structured, Unstructured Data.

5

Course Objective / Requirement

- **Requirements:**

– Project: 40%

– Final exam: 60%

– Bonus (Class activity) 1 points (Directly on your final mark)

6

What Could I Learn from This Course?

- **You will be able to:**
- Identify Big Data and its Business Implications.
- List the components of Hadoop and Hadoop Eco-System
- Access and Process Data on Distributed File System
- Manage Job Execution in Hadoop Environment
- Develop Big Data Solutions using Hadoop Eco System

7

What launched the Big Data era?

8

New Opportunities

Data Science
#1 Catalyst for economic growth!
-McKinsey

Changing Times

9

Big data – a growing torrent

McKinsey Report (2013)

10

McKinsey Report (2013)

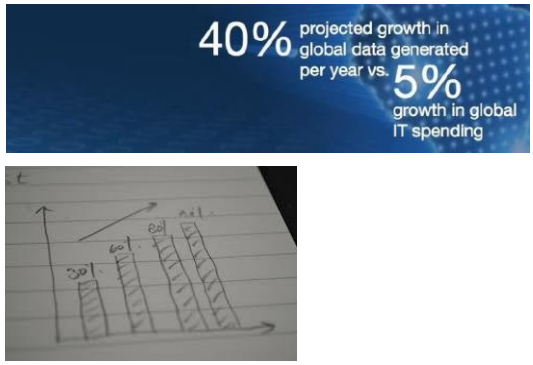
11

McKinsey Report (2013)

12

McKinsey Report (2013)

13



McKinsey Report (2013)

14

Cloud Computing

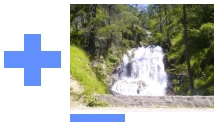
15

Computing anywhere and anytime

On-Demand Computing

16

Computing anywhere and anytime



dynamic and scalable data analysis

17

Data Torrent

Computing Anytime, Anywhere

Big Data Era

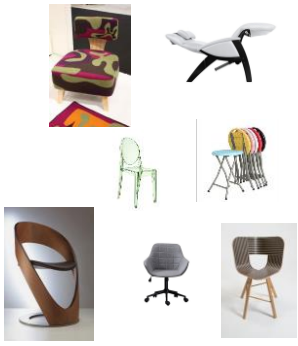
18

What makes Big Data valuable

Big Data

Better Models

Higher Precision



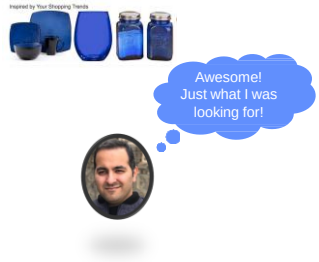
19

What makes Big Data valuable

Big Data

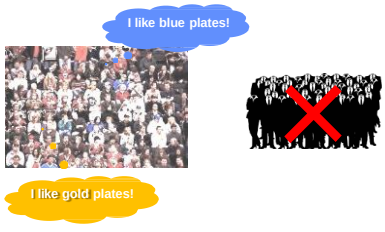
Better Models

Higher Precision



20

What makes Big Data valuable



21

Personalized Marketing

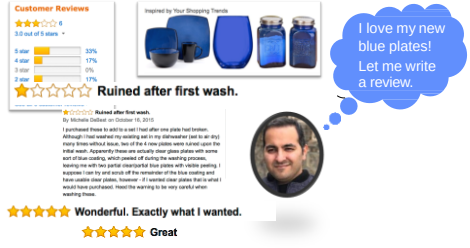
22

Recommendation Engines

amazon NETFLIX

23

Sentiment Analysis



Natural Language Processing

24

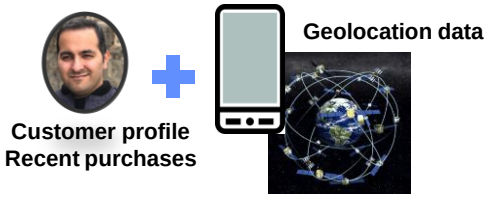
Mobile Advertising

25

Mobile Advertising

26

Mobile Advertising

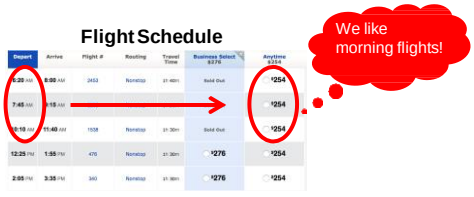


27

Consumer Growth to Guide Product Growth

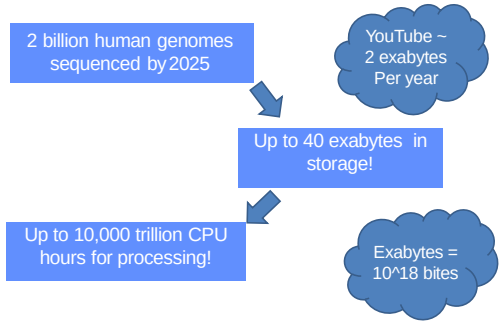
28

Consumer Growth to Guide Product Growth



29

Biomedical Applications



30

Personalized Cancer Treatment

31

Personalized Cancer Treatment

Perform large scale analysis of patient genes and tumor growth

Customized treatment

32

Big Data-Driven Cities

- **Use city wide sensor datato**
  - **Lower energy costs, pollution**
  - **Improve services, traffic, safety, …**

33

How are other applications using Big Data?

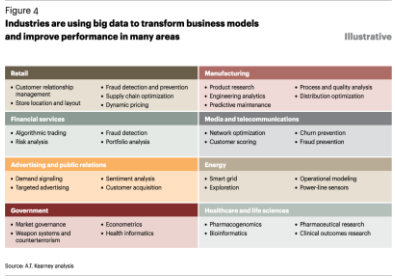


Figure 4
Industries are using big data to transform business models and improve performance in many areas
Illustrative

Retail		**Manufacturing**	
• Customer relationship management • Store location and layout	• Fraud detection and prevention • Supply-chain optimization • Dynamic pricing	• Product research • Engineering analytics • Predictive maintenance	• Process and quality analysis • Distribution optimization
Financial services		**Media and telecommunications**	
• Algorithmic trading • Risk analysis	• Fraud detection • Portfolio analysis	• Network optimization • Customer scoring	• Churn prevention • Fraud prevention
Advertising and public relations		**Energy**	
• Demand signaling • Targeted advertising	• Sentiment analysis • Customer acquisition	• Smart grid • Exploration	• Operational modeling • Power-line sensors
Government		**Healthcare and life sciences**	
• Market governance • Weapon systems and counterterrorism	• Econometrics • Health informatics	• Pharmacogenomics • Bioinformatics	• Pharmaceutical research • Clinical outcomes research

Source: A.T. Kearney analysis

34

Smart and personalized business!

Big Data

Better Models

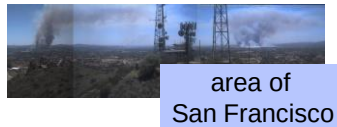
Higher Precision

Personalized Marketing
Personalized Medicine
Smart Cities
< and more! >

35

Saving Lives with Big Data Wildfire Prediction

- **San Diego County, May 14, 2014**



14 fires burning

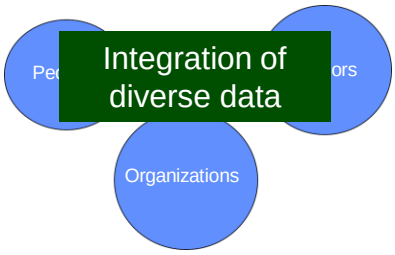
area of San Francisco

6 injured + 1 death

$60 million USD

36

Why can Big Data help?



37

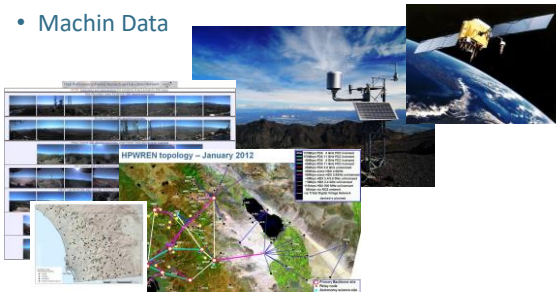
Integration of diverse streams

see new things
develop predictive analytics

38

Diverse Data Sources

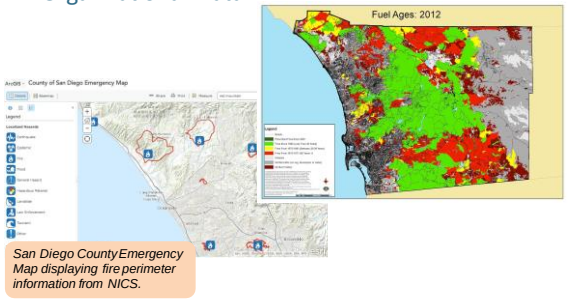
- Machin Data



39

Diverse Data Sources

- Organizational Data



San Diego County Emergency Map displaying fire perimeter information from NICS.

40

Diverse Data Sources

- People

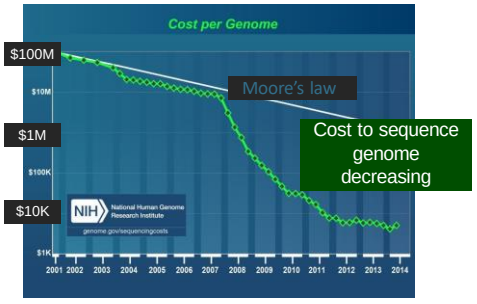
41

42

43

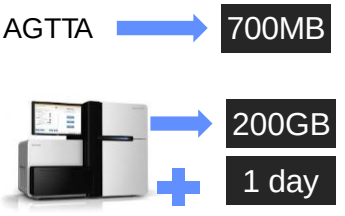
44

Saving Lives with Big Data Precision Medicine and Health Informatics



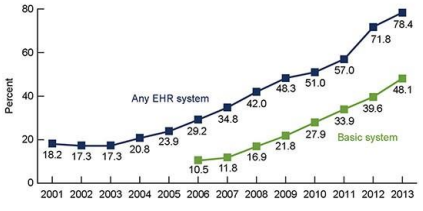
45

Genome Data Storage



46

Health Records →→ Digital

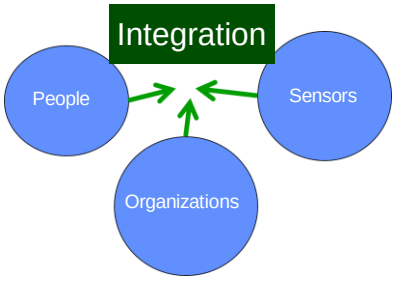


47

120 Terabytes in 2013
2X more than in 2011

48

Why can Big Data help?



50

Sensor Data

101100010 ↠

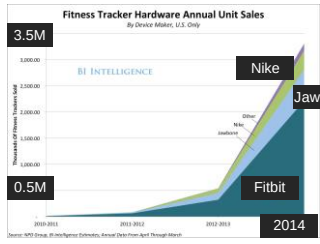
51

Sensor Data

More sensors, More places
Data ↠Storage & Analysis

52

Fitness Device Industry



53

Data Generated?

2-5 GB / day

54

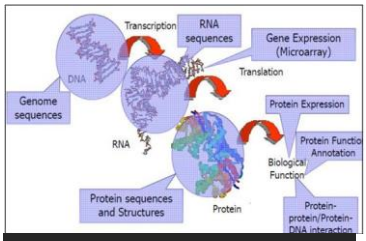
Save health care costs?

?
?
?

55

Organization Data

- Scientific Data and Knowledge-bases



National Center for Biotechnology Information

Experimental Data
+
Computed Data

56

Organization Data

- Scientific Data and Knowledge-bases

GENEONTOLOGY
Unifying Biology

Unified Medical Language System

57

People Data

- Mobile Health APP

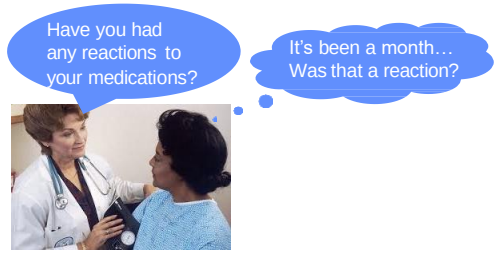
Webby
2015 PEOPLE'S VOICE

>100,000 health apps (iTunes & Google Play)

By 2017 →→ $26 billion market?

58

The impact of novel people-generated data



59

Today --- Self-Reported Data Social Media

60

Integration →→
Personalization →→
Precision

61

Big Data – Where does it come from?

62

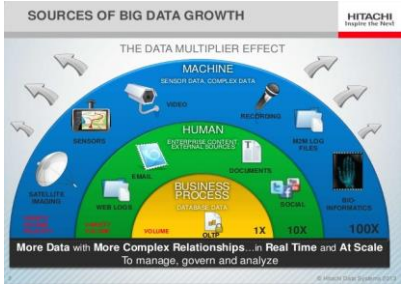
Big Data generated by Machine

Big Plane → Big Data???

Half a terabyte of data !

63

Machine data is the largest source of big data!



64

Machine data is the largest source of big data!

Sensing → Smart

65

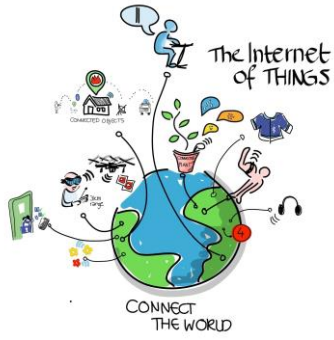
What makes a smart device smart?

Connect to other devices / networks

Collect and analyze data autonomously

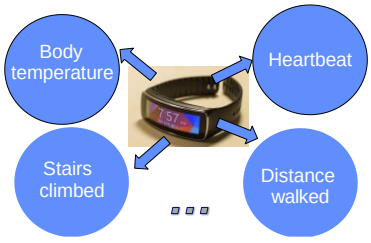
Provide environmental context

66



67

Example Smart Device: Activity Tracker



68

Example Smart Device: Activity Tracker

69

Example Smart Device: Activity Tracker

70

Increasing number of machines that sense

Data collected by each device

Machines →→ Biggest Source

71

Big Data generated by Machines: Why it's useful

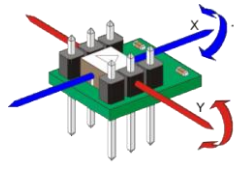
Big Plane → Big Data???

Half a terabyte of data !

72

What produces data?

Accelerometers → turbulence



73

What produces data?

Sensors:
temperature, pressure, etc.
→ turbulence

74

Real-time problem detection at 40,000 ft. with machine generated data!

In-Situ Processing

75

Then (RDBMS)	Now (In-Situ)
Data moved to computational space	Bring computation to data

Real-time Notification Enables Real-time Actions

76

Design Differently!

77

Culture shift to real-time operations

- **Customer relations**
- **Fraud detection**
- **System monitoring/control**

78

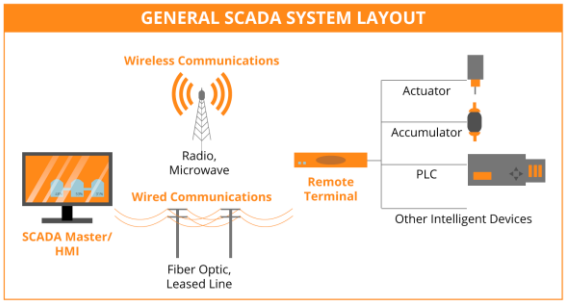
Increased use of scalable computing

amazon web services™

Microsoft Azure

79

Supervisory Control and Data Acquisition (SCADA)



Remote monitoring / control industrial processes

80

Supervisory Control and Data Acquisition (SCADA)

Reduce waste, improve efficiency

Identify trends, patterns, and anomalies

81

Big Data generated by People: The Unstructured Challenge

82

A huge growth and volume of data!

Daily facebook data > All US Academic Libraries

2 PBs vs. 30+ PBs

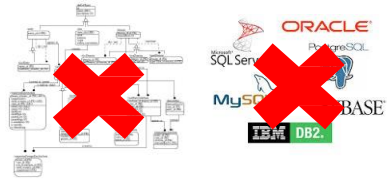
83

A huge growth and volume of data!

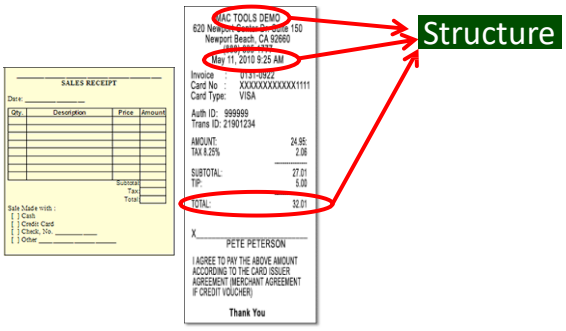
Company	Data Processed Daily
eBay	100 Petabytes (PB)
Google	100 PB
Facebook	30+ PB
Twitter	100 Terabytes(=.1PB)
Spotify	64 Terabytes

84

The Unstructured Data Challenge

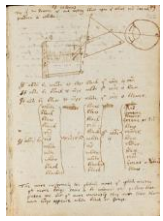


85



86

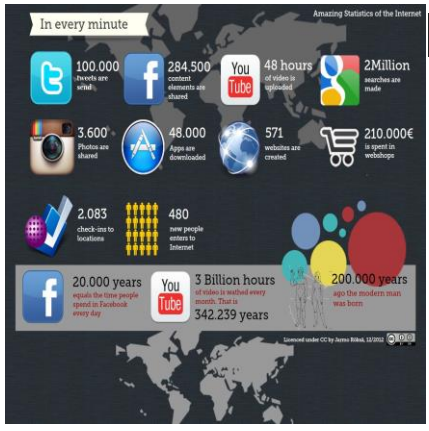
80%-90% of entire data is unstructured!



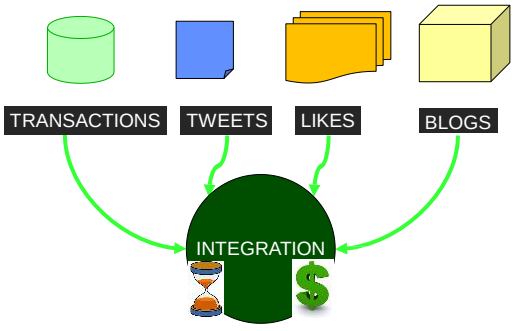
87

80%-90% of entire data is unstructured!

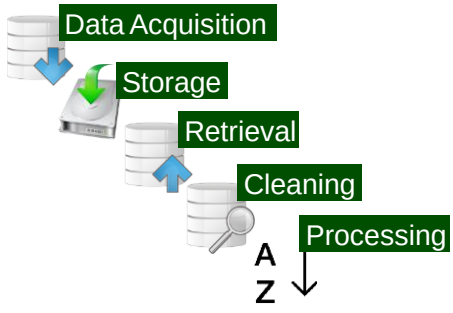
88



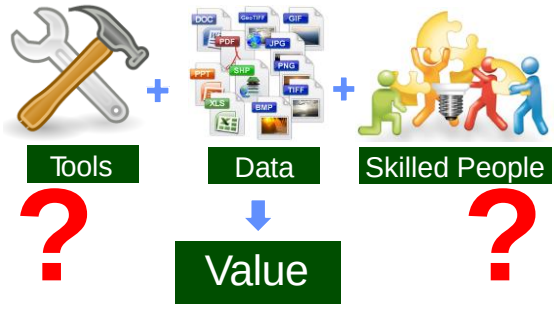
89



90

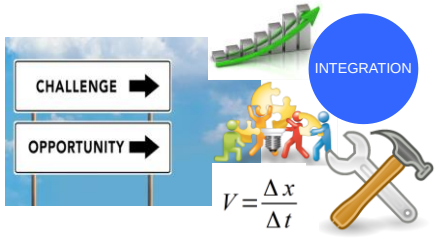


91



92

Big Data generated by People: How is it being used?

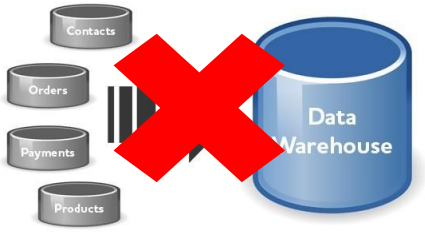


93

94

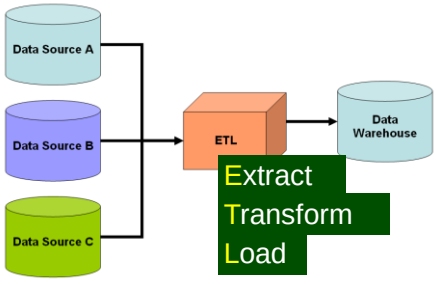
$V = \frac{\Delta x}{\Delta t}$

95



96

Traditional Data Warehouse



97

NoSQL Data Storage in the Cloud

- Beyond relational databases!

Organize data to suit the problem and objectives!

98

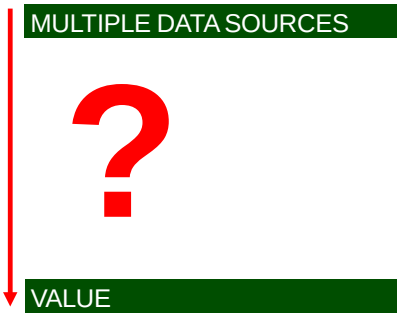
Connections between data

Neo4j

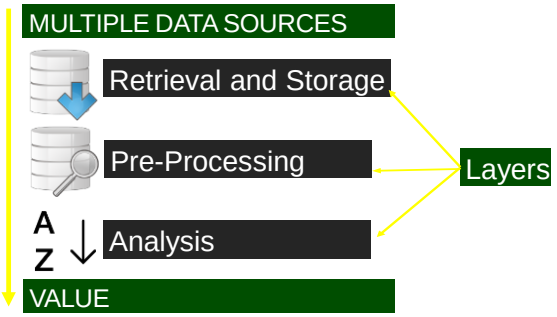
Key-Value pairs

Cassandra

99

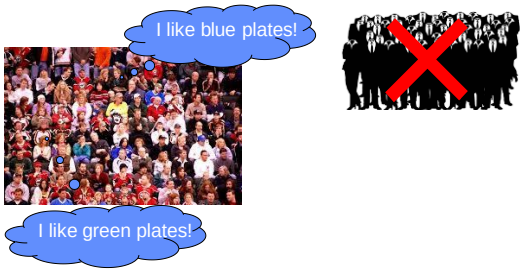


100



101

Value in Action
Sentiment Analysis



102

Guess, how much Twitter data companies analyze everyday to measure "sentiment" around their products?

Answer: 12 Terabytes/Day

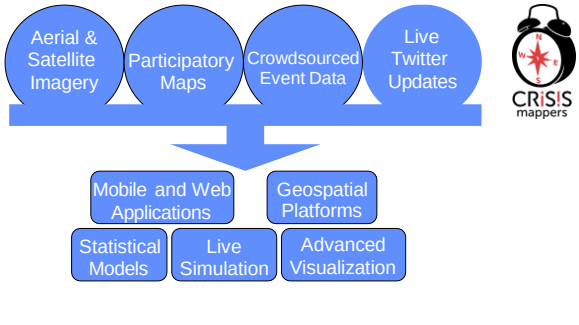
You would need to listen continuously for ~ 2 years to finish 1 TB of music !

103

104

Collective Disaster Response

105



106

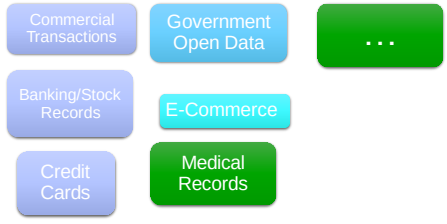
107

Can you list 3 things you can do from Big Data analysis

108

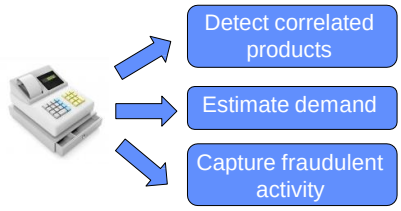
Big Data generated by Organizations: Structured but often siloed

How organizations produce data

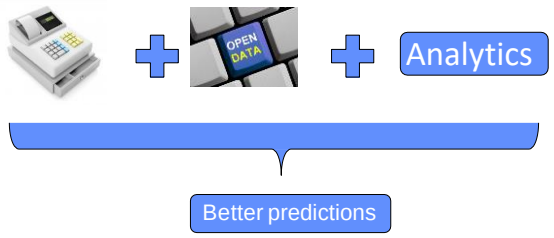


109

Sale transaction data

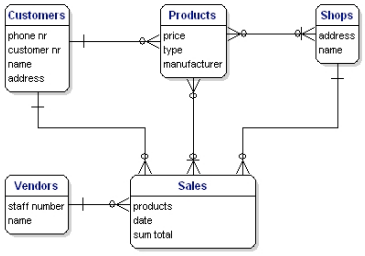


110



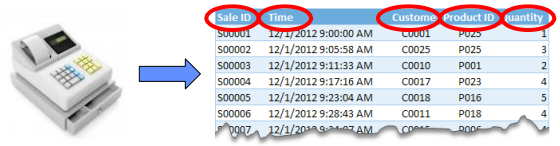
111

Highly structured data



112

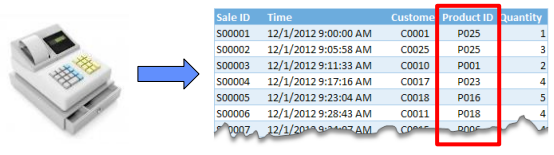
Sales Transaction Records



Sale ID	Time	Custome	Product ID	Quantity
S00001	12/1/2012 9:00:00 AM	C0001	P025	1
S00002	12/1/2012 9:05:58 AM	C0025	P025	3
S00003	12/1/2012 9:11:33 AM	C0010	P001	2
S00004	12/1/2012 9:17:16 AM	C0017	P023	4
S00005	12/1/2012 9:23:04 AM	C0018	P016	5
S00006	12/1/2012 9:28:43 AM	C0011	P018	4

113

Sales Transaction Records

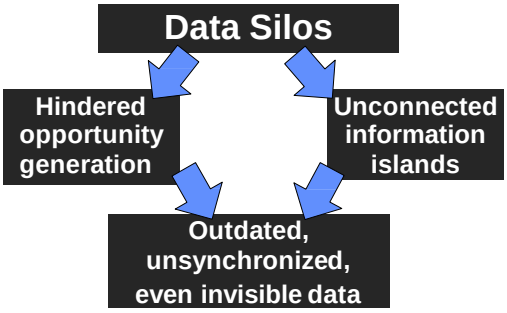


Sale ID	Time	Custome	Product ID	Quantity
S00001	12/1/2012 9:00:00 AM	C0001	P025	1
S00002	12/1/2012 9:05:58 AM	C0025	P025	3
S00003	12/1/2012 9:11:33 AM	C0010	P001	2
S00004	12/1/2012 9:17:16 AM	C0017	P023	4
S00005	12/1/2012 9:23:04 AM	C0018	P016	5
S00006	12/1/2012 9:28:43 AM	C0011	P018	4

114

Data silos within an organization!

116



117

Course Objective / Requirement

- **Requirements:**
  - Project: 40%
  - Final exam: 60%
  - Bonus (Class activity) 1 points (Directly on your final mark)

118

Organization-Generated Data: Benefits come from combining with other types

- Real-World Examples
  - The UPS success

16 Million Shipments per day

40 Million tracking request

UPS is estimated to have 16 PBs of data about its operations

119

Can you guess how much money UPS can save by reducing each driver's route by just 1 mile ?

50 Million Dollars!

120

Large operational data

+

Optimization algorithms

Savings!

Route Optimization

121

- Real-World Examples
  - The Walmart success

250 Million customers

10,000 stores

2.5 petabytes data collected every 60 minutes !

122

- Twitter data
- local events
- local weather
- in-store purchases
- online clicks

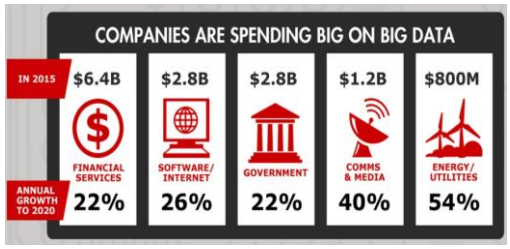
Launch new products

Improve predictive analytics

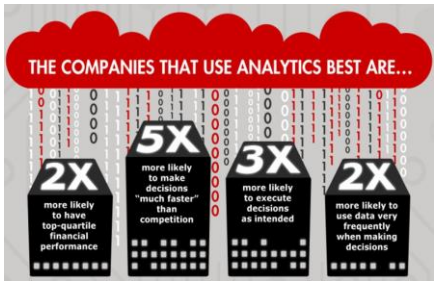
Customize recommendations

123

Future Trends for Organizations



124



125

Efficient Operations

Higher Sales

Improved Safety

Customer Satisfaction

Better Profit Margins

Improved Product Placement

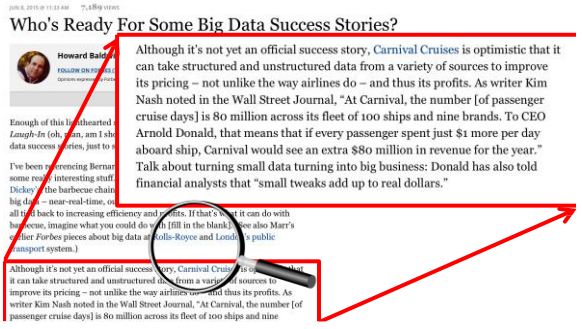
126

Integrating diverse data

- Getting Value from Big Data

Value comes from integrating different types of data sources

127


Who's Ready For Some Big Data Success Stories?

Although it's not yet an official success story, Carnival Cruises is optimistic that it can take structured and unstructured data from a variety of sources to improve its pricing – not unlike the way airlines do – and thus its profits. As writer Kim Nash noted in the Wall Street Journal, "At Carnival, the number [of passenger cruise days] is 80 million across its fleet of 100 ships and nine brands. To CEO Arnold Donald, that means that if every passenger spent just $1 more per day aboard ship, Carnival would see an extra $80 million in revenue for the year." Talk about turning small data turning into big business: Donald has also told financial analysts that "small tweaks add up to real dollars."

128

Structured
+
Unstructured Data

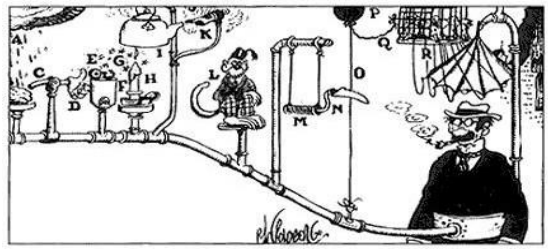
Although it's not yet an official success story, Carnival Cruises is optimistic that it can take structured and unstructured data from a variety of sources to improve its pricing – not unlike the way airlines do – and thus its profits. As writer Kim Nash noted in the Wall Street Journal, "At Carnival, the number [of passenger cruise days] is 80 million across its fleet of 100 ships and nine brands. To CEO Arnold Donald, that means that if every passenger spent just $1 more per day aboard ship, Carnival would see an extra $80 million in revenue for the year." Talk about turning small data turning into big business: Donald has also told financial analysts that "small tweaks add up to real dollars."

Price optimization

Increased revenue!

129

Insert Big Data Integration Here



130

Data Integration →→Knowledge

Key:
Turning complex data into something usable

131

Data Integration Process

Discovering

Accessing

Monitoring

→

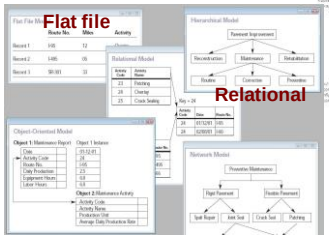
Modeling

Transforming

132

Why do we need Data integration?

Data comes in all shapes and sizes!



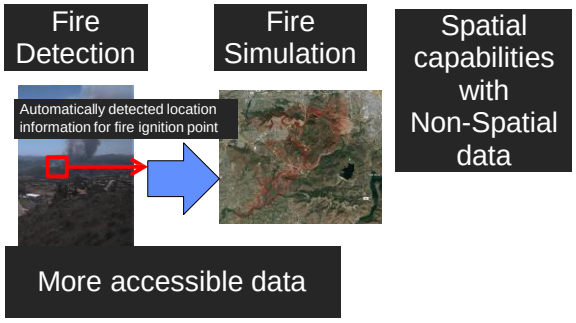
133

Data Integration

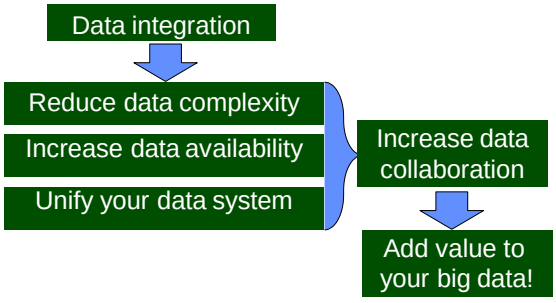
↓

Richer Data

134



135



136

THANK YOU

Assoc. Prof. Pejman Rasti

137