CS590RA: Real-Time Big Data Analytics

References:
Mike Barlow, Real-time big data analytics: Emerging Architecture
Marko Grobelnik, Big Data Tutorial

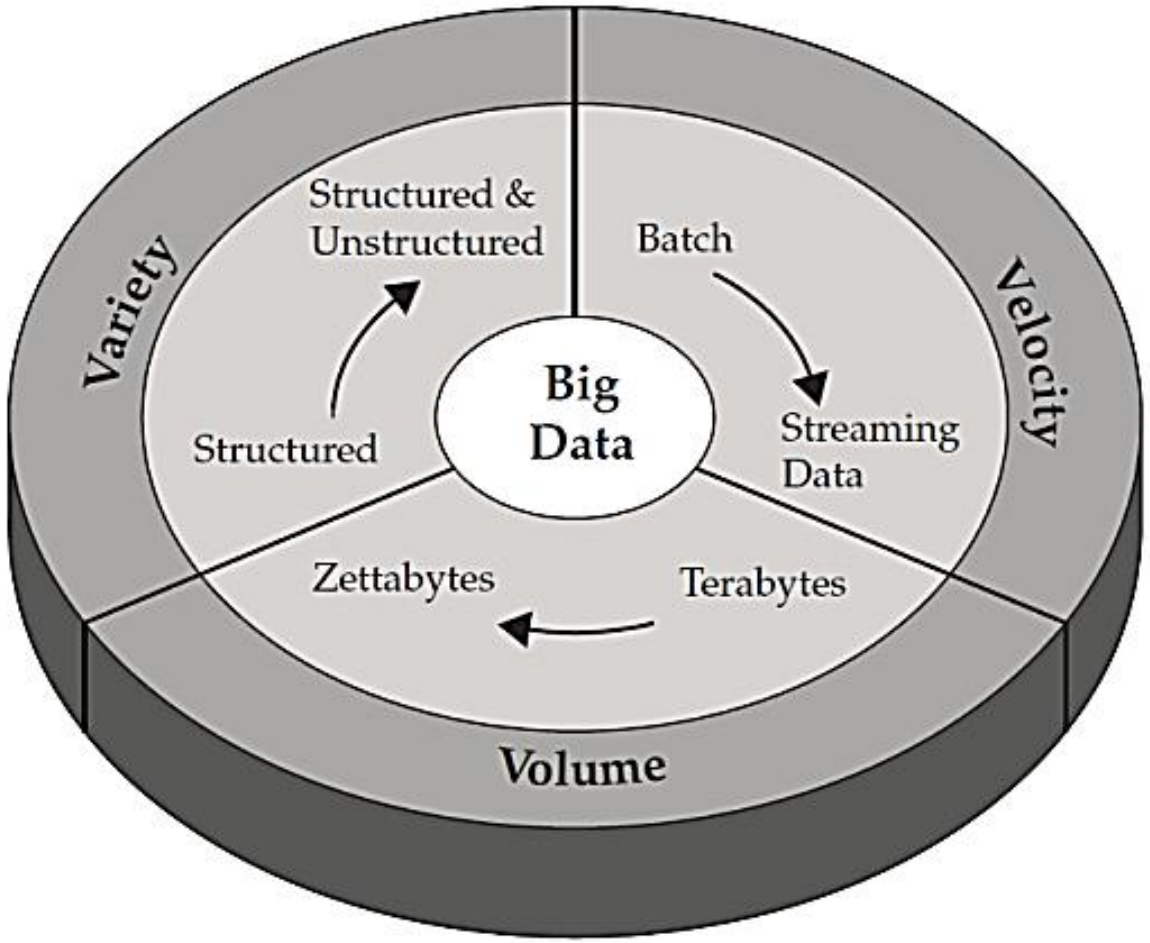
What is Big Data?

- Big Data refers to “data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time.”
- Big Data is a generic term used to describe the **voluminous amount of unstructured, structured and semi-structured data**.

3 Key Characteristics of Big Data

- **Volume**: High volume of data created both inside corporations and outside the corporations via the web, mobile devices, IT infrastructure, and other sources
- **Variety**: Data is in *structured, semi-structured and unstructured format*.
- **Velocity**: Data is generated at a high speed high volume of data needs to be processed within seconds

Characterization of Big-Data: volume, velocity, variety (V3)



Source: Marko Grobelnik, Big Data Tutorial

Zettabyte Age

1 Kilobyte	1,000 bits/byte
1 megabyte	1,000,000
1 gigabyte	1,000,000,000
1 terabyte	1,000,000,000,000
1 petabyte	1,000,000,000,000,000
1 exabyte	1,000,000,000,000,000,000
1 zettabyte	1,000,000,000,000,000,000,000

Data Volume is Growing Exponentially

- Estimated Global Data Volume:
  - 2011: 1.8 ZB
  - 2015: 7.9 ZB
- The world's information doubles every two years
- Over the next 10 years:
  - The number of servers worldwide will grow by 10x
  - Amount of information managed by enterprise data centers will grow by 50x
  - Number of "files" enterprise data center handle will grow by 75x

Source: http://www.emc.com/leadership/programs/digital-universe.htm, which was based on the 2011 IDC Digital Universe Study

6

Internet of Things

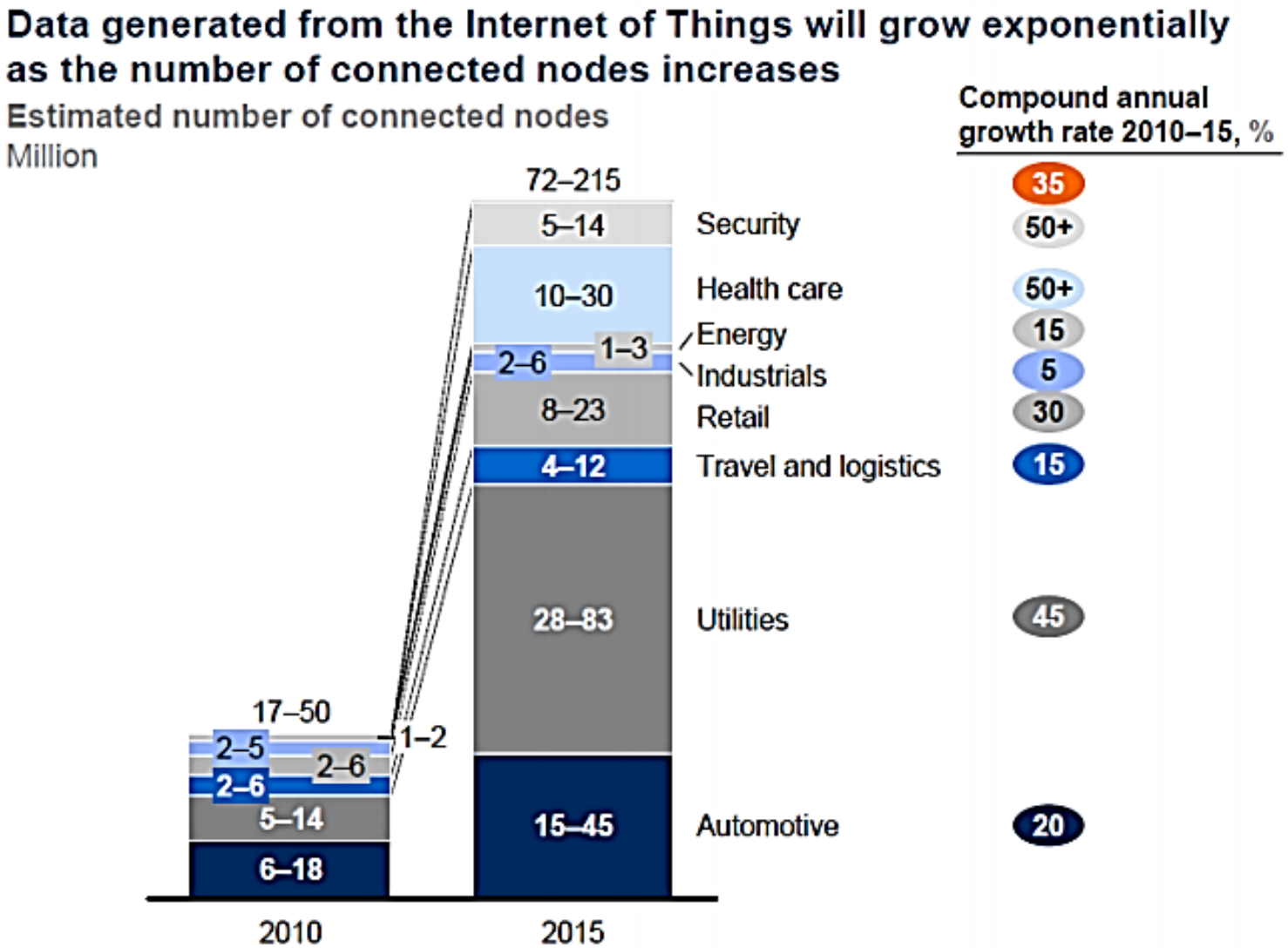
- In the year 2025,
  - People will be wearing sensors at all times,
  - Entire homes will be linked through common gadgets, etc
- Cisco counted
  - 13 billion Internet-connected devices.
  - By 2020, there will be ***50 billion***, tipping phones, chips, sensors, implants, and devices of which we have not yet conceived.

Video:

https://www.youtube.com/watch?v=sfEbMV295Kk

http://www.pcmag.com/article2/0,2817,2458060,00.asp

Data available from “Internet of Things”

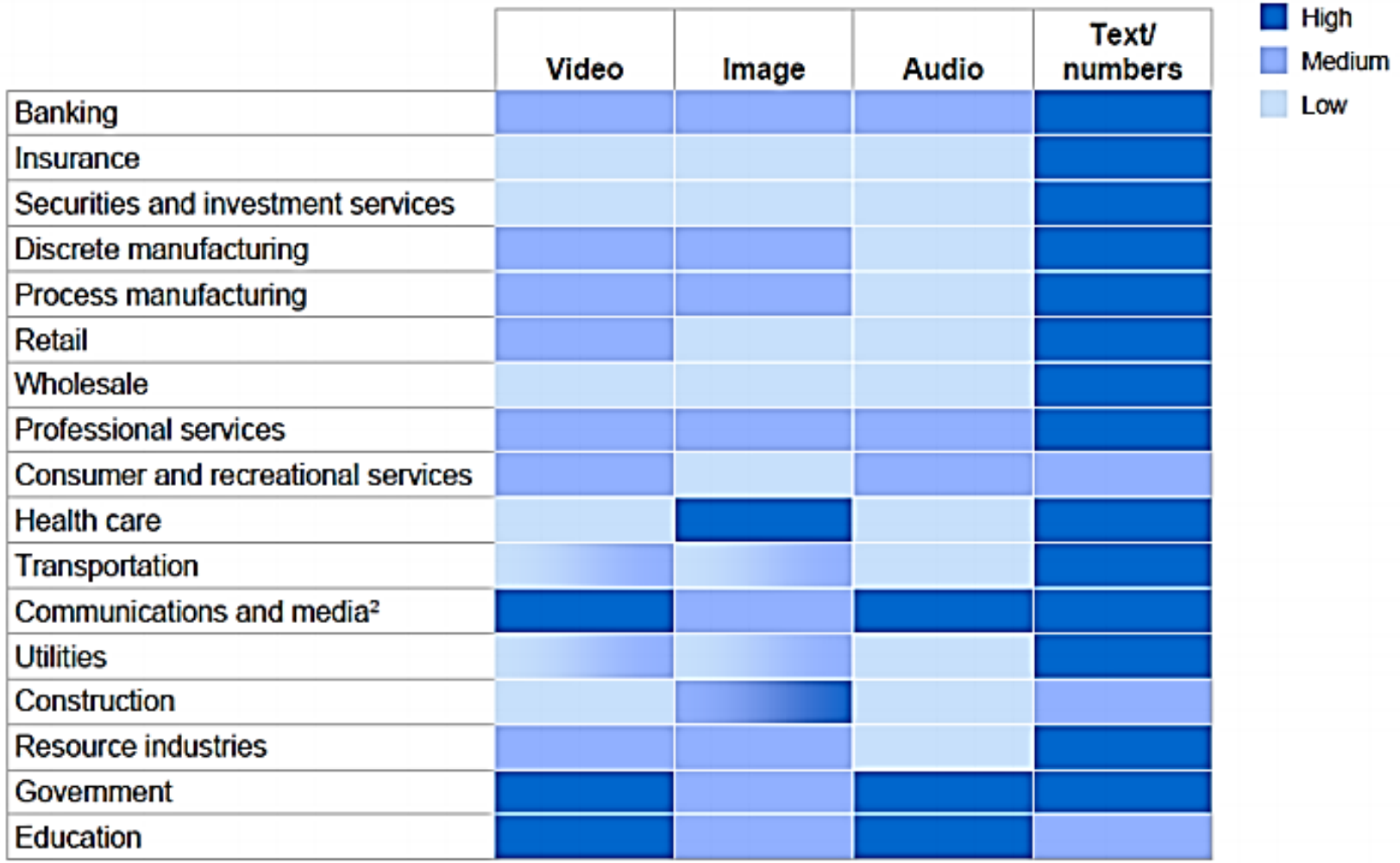


NOTE: Numbers may not sum due to rounding.

SOURCE: Analyst interviews; McKinsey Global Institute analysis

Type of available data

The type of data generated and stored varies by sector[1]



1 We compiled this heat map using units of data (in files or minutes of video) rather than bytes.
2 Video and audio are high in some subsectors.
SOURCE: McKinsey Global Institute analysis

A lot happens in the Digital World in 60 seconds…

- **2 million searches are processed by Google**
- **70 new domains are registered**
- **347 blog posts are created on WordPress**
- **11,000 searches happen on LinkedIn**
- **278 thousand tweets get published**
- **20 million look at Flickr photos**
- **1.8 million people like something on Facebook**

The numbers are indeed baffling - but these are the signs of our times, really.

Source: http://pinoyanalyticsguy.blogspot.com/2013/08/new-york-minute-online-what-happens-in.html

Source: Online in 60 Seconds [Infographic] is an infographic that was produced by Qmee

Key enablers for the growth of “Big Data”

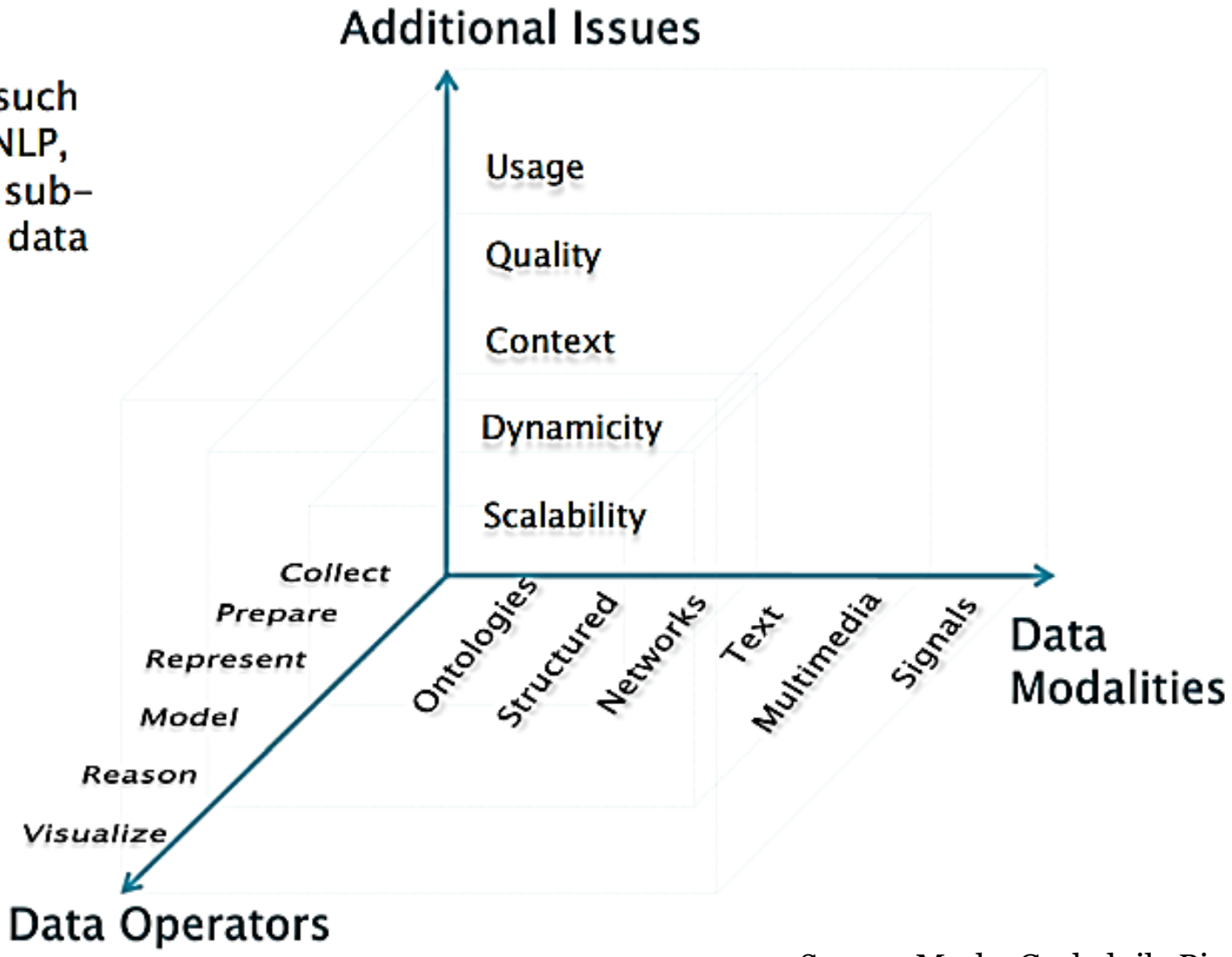
- Increase of storage capacities
- Increase of processing power
- Availability of data

Is this sufficient?

What about ***Machine Intelligence***?

What matters when dealing with data?

- Research areas (such as IR, KDD, ML, NLP, SemWeb, …) are sub-cubes within the data cube



Source: Marko Grobelnik, Big Data Tutorial

What's Real-time Big Data Analytics?

- It's about the ability to make better decisions and take meaningful actions at the right time.
  - E.g., detecting fraud while someone is swiping a credit card,
  - E.g., triggering an offer while a shopper is standing on a checkout line,
  - E.g., placing an ad on a website while someone is reading a specific article.
- Real-time Big data Analytics (RTBDA)
  - combining and analyzing data to take the right action, at the right time, and at the right place.
  - "*Machines begin to think and respond more like humans*" by Michael Minelli, co-author of *Big Data, Big Analytics*.

RTBDA: Applications

…an example: recommendation @Bloomberg.com

- Good recommendations can make a big difference when keeping a user on a web site
  - …the key is how rich context model a system is using to select information for a user
  - Bad recommendations <1% users, good ones >5% users click

Contextual personalized recommendations generated in ~20ms

Source: Marko Grobelnik, Big Data Tutorial

Each click on the web site is enriched and indexed using:

- Domain
- Sub-domain
- Page URL
- URL sub-directories

- Page Meta Tags
- Page Title
- Page Content
- Named Entities

- Has Query
- Referrer Query

- Referring Domain
- Referring URL
- Outgoing URL

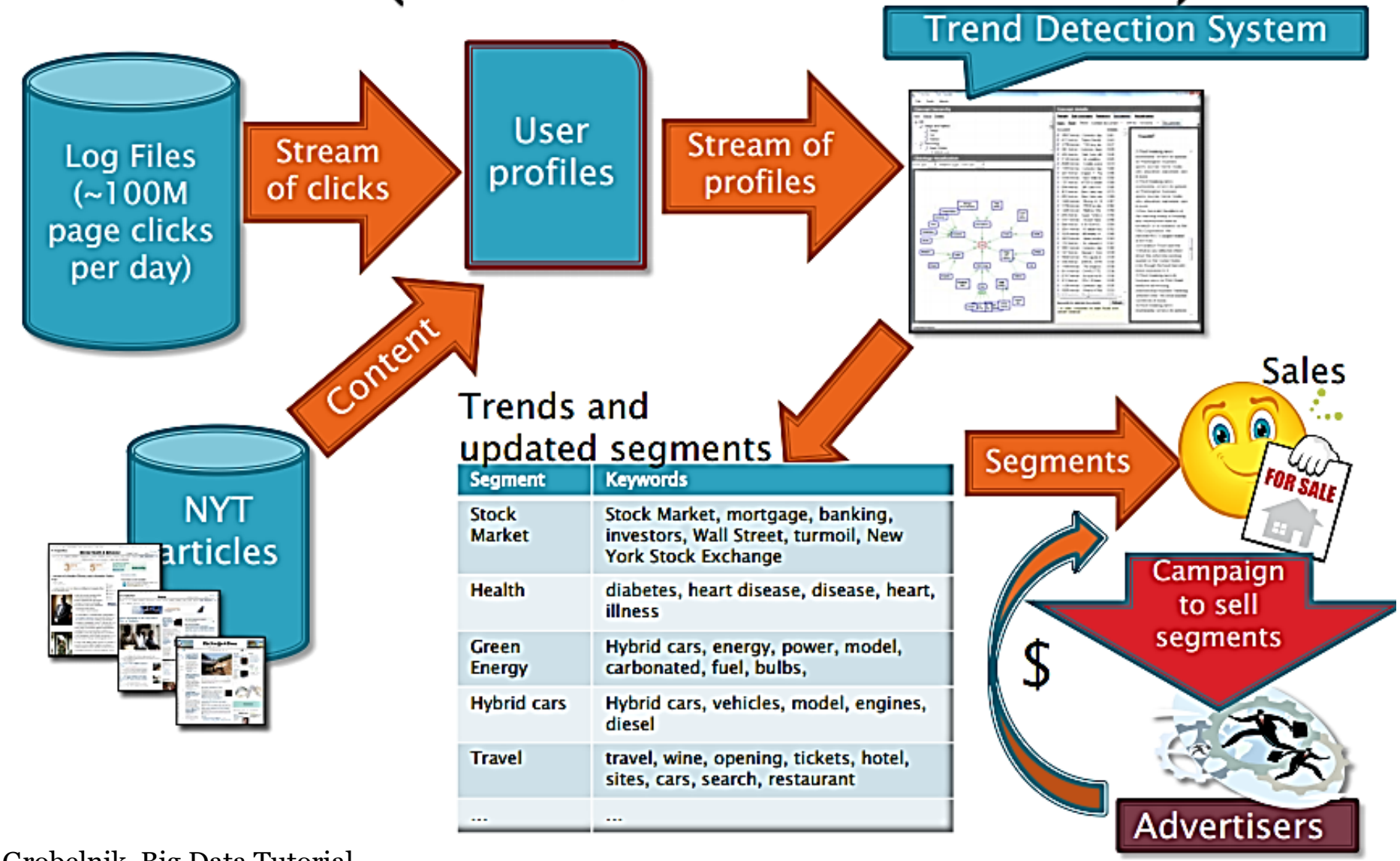
- GeoIP Country
- GeoIP State
- GeoIP City

- Absolute Date
- Day of the Week
- Day period
- Hour of the day
- User Agent

- Zip Code
- State
- Income
- Age
- Gender
- Country
- Job Title
- Job Industry

Source: Marko Grobelnik, Big Data Tutorial

Application: Online Advertising for NYTimes (microtrends detection)



Segment	Keywords
Stock Market	Stock Market, mortgage, banking, investors, Wall Street, turmoil, New York Stock Exchange
Health	diabetes, heart disease, disease, heart, illness
Green Energy	Hybrid cars, energy, power, model, carbonated, fuel, bulbs,
Hybrid cars	Hybrid cars, vehicles, model, engines, diesel
Travel	travel, wine, opening, tickets, hotel, sites, cars, search, restaurant
...	...

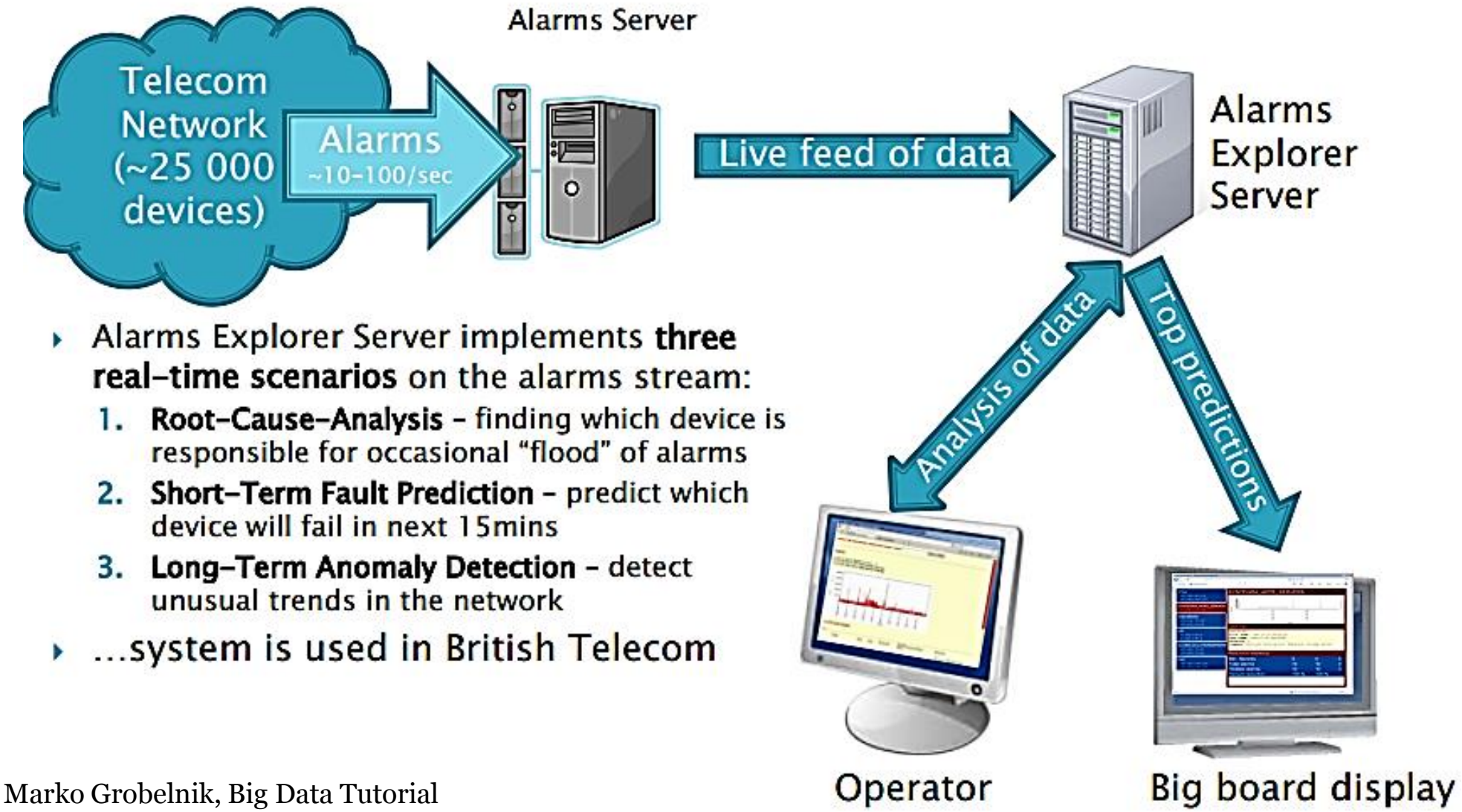
Source: Marko Grobelnik, Big Data Tutorial

Figures for one day of NYTimes

- 50Gb of uncompressed log files
- 10Gb of compressed log files
- 0.5Gb of processed log files
- 50-100M clicks
- 4-6M unique users
- 7000 unique pages with more then 100 hits
- Index size 2Gb
- Pre-processing & indexing time
  - ~10min on workstation (4 cores & 32Gb)
  - ~1hour on EC2 (2 cores & 16Gb)

Source: Marko Grobelnik, Big Data Tutorial

Applications: Telecommunication Network Monitoring



- Alarms Explorer Server implements **three real-time scenarios** on the alarms stream:
  1. **Root-Cause-Analysis** - finding which device is responsible for occasional "flood" of alarms
  2. **Short-Term Fault Prediction** - predict which device will fail in next 15mins
  3. **Long-Term Anomaly Detection** - detect unusual trends in the network
- ...system is used in British Telecom

Source: Marko Grobelnik, Big Data Tutorial