Intro to Big Data

ITU ASP COE TRAINING ON
“Developing the ICT ecosystem to harness IoTs”

Marco Zennaro, PhD
13-15 December 2016
Bangkok, Thailand
History of Big Data

• **300 BC – 48 AD** The Library of Alexandria is the world’s largest data storage center – until it is destroyed by the Romans.

• **180,000 books, for a total of 20 Gb**
1663 John Graunt conducts the first recorded statistical-analysis experiments in an attempt to curb the spread of the bubonic plague in Europe.
History of Big Data

- **1881** Herman Hollerith creates the Hollerith Tabulating Machine which uses punch cards to vastly reduce the workload of the US Census. He is one of the founders of IBM.
- **1926** Nikola Tesla predicts that in the future, a man will be able to access and analyze vast amounts of data using a device small enough to fit in his pocket.
History of Big Data

- **1965** The US Government plans the world’s first data center to store 742 million tax returns and 175 million sets of fingerprints on magnetic tape.

- **1989** Early use of term Big Data in magazine article by fiction author Erik Larson – commenting on advertisers’ use of data to target customers.
History of Big Data

• 2010 Eric Schmidt, executive chairman of Google, tells a conference that as much data is now being created every two days, as was created from the beginning of human civilization to the year 2003.
Definition

“Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set.” – Wikipedia
Three Characteristics of Big Data

Volume  Velocity  Variety
Volume

• The sheer size of data in terms of storage and access.
• For example: unstructured data from social media in form of posts, video, audio with relational data such as comments, discussions, likes, etc.
• The Large Hadron Collider (LHC) will generate 60 terabytes of data per day, 25 petabytes annually

• Wallmart generates 2.5 petabytes per hour
Volume
Sensor data from a cross-country flight

20 TB × 2 × 6 × 28,537 × 365

20 terabytes of information per engine every hour
twin-engine Boeing 737
six-hour, cross-country flight from New York to Los Angeles
# of commercial flights in the sky in the United States on any given day
days in a year

= 2,499,841,200 TB
Velocity

• The speed of incoming data and the time it takes to process it.
• With the advent of IoT, streaming data is driving the need to process and analyze data in near-real time.
The New York Stock Exchange captures 1 TB of trade information during each trading session.

By 2016, it is projected there will be 18.9 billion network connections - almost 2.5 connections per person on earth.

**Velocity**

**Analysis of Streaming Data**

Modern cars have close to 100 sensors that monitor items such as fuel level and tire pressure.
Variety

• The type of files and format of data as well as sources.

• Data can be **structured**, such as a traditional database (pre-formatted data collected over time), or **unstructured** (unrelated data from unstructured sources such as social media, email, etc)
Big Data – Supporting Trends

• **Moore's Law**: an observation that the number of transistors on integrated circuits doubles every two years.
Big Data – Supporting Trends

• Kryder's Law: the density of storage is increasing and the cost decreasing at a rate faster than Moore's Law
Big Data Value Chain

- **Collection** – Structured, unstructured and semi-structured data from multiple sources
- **Ingestion** – loading vast amounts of data onto a single data store
- **Discovery & Cleansing** – understanding format and content; clean up and formatting
- **Integration** – linking, entity extraction, entity resolution, indexing and data fusion
- **Analysis** – Intelligence, statistics, predictive and text analytics, machine learning
- **Delivery** – querying, visualization, real time delivery on enterprise-class availability

Source O’Reilly Strata 2012
Big Data – Tools

- **Hadoop** is often used at the server level to organise the cluster along with a NoSQL database for data storage.

- **NoSQL** are databases that use looser consistency models than relational databases. Performance gains via simplification using key value stores.
# Examples of data generated by IoT

<table>
<thead>
<tr>
<th>Level</th>
<th>Individual</th>
<th>Community</th>
<th>Society</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IoT</strong></td>
<td>Smart phones, Wearables</td>
<td>Connected Cars, Health devices, Smart homes</td>
<td>Smart Cities, Smart Grids</td>
</tr>
<tr>
<td><strong>Examples</strong></td>
<td>GPS, Fitbits, Visa PayWave, Mastercard PayPass, Employee passes</td>
<td>Intelligent Transport Systems, Event Data Recorders (EDRs), Blood pressure monitors, remote burglar/heating systems</td>
<td>Smart metering, Smart water meters, Traffic monitoring</td>
</tr>
<tr>
<td><strong>Data</strong></td>
<td>Mobile money, Fitness data, GPS location-based data</td>
<td>Speed, distance, airbag, crash locations/alerts, Heart rate, blood pressure, Diet, remote heating data</td>
<td>Electricity/water consumption &amp; billing, Traffic flow data</td>
</tr>
<tr>
<td><strong>Intended Audience</strong></td>
<td>Individual person, Immediate friends/family, banks, employers</td>
<td>GP, health authorities, health &amp; car insurance, police, social networks</td>
<td>Authorities/regulators, Utility companies, Other citizens</td>
</tr>
</tbody>
</table>

Source: Adapted from the ITU Draft GSR Discussion Paper 2015, “Regulation and the Internet of Things”,
Use case: Traffic

- **Collect** traffic data and transportation data from sensors
- **Build** a model of traffic patterns
- **Predict** the traffic and congestions
- **Act**: divert traffic, adjust toll, adjust traffic lights
Source: REDtone IOT
Number of tweets per day in Britain that contained the word "spider". Retweets, replies, and tweets about Spider-Man have been excluded.
The number of tweets about spiders in Britain is correlated with the local mean temperature:

Number of British spider tweets per day (top), shown against the mean temperature in Central England obtained from the UK Met Office (bottom).
United States Flu Activity

Influenza estimate

- Google Flu Trends estimate
- United States data

United States: Influenza-like illness (ILI) data provided publicly by the [U.S. Centers for Disease Control](https://www.cdc.gov).
Switzerland Flu Activity

Influenza estimate

- Google Flu Trends estimate
- Switzerland data

Switzerland: Influenza-like illness (ILI) data provided publicly by the European Influenza Surveillance Network of the European Centre for Disease Prevention and Control.
Per capita consumption of mozzarella cheese (US) correlates with Civil engineering doctorates awarded (US)

Correlation: 96%  Sources: USDA & National Science Foundation  tylervigen.com

<table>
<thead>
<tr>
<th>Year</th>
<th>Per capita consumption of mozzarella cheese (US)</th>
<th>Civil engineering doctorates awarded (US)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>9.3</td>
<td>480</td>
</tr>
<tr>
<td>2001</td>
<td>9.7</td>
<td>501</td>
</tr>
<tr>
<td>2002</td>
<td>9.7</td>
<td>540</td>
</tr>
<tr>
<td>2003</td>
<td>9.7</td>
<td>552</td>
</tr>
<tr>
<td>2004</td>
<td>9.9</td>
<td>547</td>
</tr>
<tr>
<td>2005</td>
<td>10.2</td>
<td>622</td>
</tr>
<tr>
<td>2006</td>
<td>10.5</td>
<td>655</td>
</tr>
<tr>
<td>2007</td>
<td>11</td>
<td>701</td>
</tr>
<tr>
<td>2008</td>
<td>10.6</td>
<td>712</td>
</tr>
<tr>
<td>2009</td>
<td>10.6</td>
<td>708</td>
</tr>
</tbody>
</table>

Correlation: 0.958648

Source: [http://tylervigen.com/](http://tylervigen.com/)
Tweets about the price of rice (per month)

Food Price Inflation

Source: UN Global Pulse
The Internet of Things Ecosystem

- Internet Network
- Gateway
- IoT Devices
- Data Storage
- Analytics

Analysis
Command/RFI
Demo

http://discover-iot.eu-gb.mybluemix.net/#/play/device/smartphone
Thank You