Intelligent Knowledge Lakes
The Age of Artificial Intelligence and Big Data

University of Malaya
8 January 2020

Dr. Amin Beheshti
Director, AI-enabled Processes (AIP) Research Centre
Director, Data Analytics Research Lab

Web: https://aip-research-center.github.io/

Copyright © DataAnalyticsResearchLab @MQ
https://data-science-group.github.io/
Introduction to Big Data
Introduction to Big Data

Application
  ├── Set of Related Programs
  │     └── Set of Related Functions
  │         └── Set of Related Statements
  └── Assignment
        ├── Selection
        │     └── Iteration
        │         └── Jump
        └── 2

Variable?
  └── M.A.
    └── Label
        └── Value
            └── DataType

Copyright © DataAnalyticsResearchGroup @MQ  https://data-science-group.github.io/
Introduction to Big Data

Application

Set of Related Programs

Set of Related Functions

Set of Related Statements

Assignment
Selection
Iteration
Jump
...

Data Structures?

Variable
Array
Record
Class
.
.
.
ADT

Abstraction

Reusability

Copyright © DataAnalyticsResearchGroup @MQ

https://data-science-group.github.io/
Introduction to Big Data

Application

Execute on Computing Platform

 Hardware
Software

Platform Independent Application

Copyright © DataAnalyticsResearchGroup @MQ  https://data-science-group.github.io/
Introduction to Big Data

Application

Platform Independent Application
(e.g. Web Applications)

File

Binary

Text

Tim Berners-Lee

Copyright © DataAnalyticsResearchGroup @MQ

https://data-science-group.github.io/
Introduction to Big Data

Application

Platform Independent Application (e.g. Web Applications)

Web Services

Program

Keyword Extraction

input

output

Request

Response

World Wide Web

Copyright © DataAnalyticsResearchGroup @MQ

https://data-science-group.github.io/
Introduction to Big Data

Application

- Architecture

Presentation

Logic

Data

3-Tier Architecture

GUI, Command line, ...

- Program

- Data Structures

File

DBMS (Relational and NoSQL)

Copyright © DataAnalyticsResearchGroup @MQ

https://data-science-group.github.io/
What is Data?
What is Data?

Every day, we create **2.5 quintillion** bytes of data.
- posts to social media sites
- sensors used to gather climate information
- digital pictures and videos
- purchase transaction records
- cell phone GPS signals
- ...

- 500 Million Tweets sent each day!
- 5.75 BILLION Facebook likes every day.
- 3.6 Billion Instagram Likes each day.
- 4.3 BILLION Facebook messages posted daily.
- 6 BILLION daily Google Searches!
- ...


Copyright © DataAnalyticsResearchGroup @MQ  
https://data-science-group.github.io/
What is Data?

In computing, data is information that has been translated into a form that is efficient for storage and/or processing.

Preparing the Data for Processing & Analytics
(Organizing and Curating)

Copyright © DataAnalyticsResearchGroup @MQ  https://data-science-group.github.io/
What is Metadata?

We are Tracing everything:
- What is happening?
- Who is doing that?
- Where it is happening?
- When?
- Why?
- How?
- ...

Cross-Cutting Aspects
- Provenance
- Versioning
- Privacy
- Security
- ...

Smart Phones, tracks:
- Our location,
- Our speed,
- What apps we are using,
- What music we listen to,
- ...

Smart TVs, tracks:
- Channels we watch,
- Time and duration,
- Apps we use,
- ...

Smart Watches, tracks:
- Our health signs,
- Our activity,
- Location,
- ...


Copyright © DataAnalyticsResearchGroup @MQ

https://data-science-group.github.io/
What is Metadata?

**Provenance**, a kind of metadata, refers to the documentation of an object’s lifecycle. This documentation (often represented as a graph) should include all the information necessary to reproduce a certain piece of data or the process that led to it.

https://www.w3.org/2005/Incubator/prov/wiki/What_Is_Provenance

Copyright © DataAnalyticsResearchGroup @MQ
https://data-science-group.github.io/
What is Big Data?
What is Big Data?

- Social Data
- Open Data
- Private Data (personal/business)
- Police Investigation (e.g. Boston Marathon Bombing)


Copyright © DataAnalyticsResearchGroup @MQ

https://data-science-group.github.io/
What is Big Data?

Typical properties of the big data:
- wide physical distribution
- diversity of formats
- non-standard data models
- independently-managed
- heterogeneous semantics

Private Data
(personal/business)


Copyright © DataAnalyticsResearchGroup @MQ  https://data-science-group.github.io/
What is Big Data?

- Big data refers to our ability to collect and analyse the ever expanding amounts of data and meta-data that we are generating every second!

- Big data can be seen as a massive number of small data islands from Private (Personal/Business), Open and Social Data.

Organizing, Curating, Analysing and Presenting this data is *challenging* and of high interest.
Organizing Big Data
How to store vast amount of noisy data (varying from structured entities to unstructured documents) being generated on a continuous basis?

The Four V's of Big Data

**Volume**
the vast amounts of data generated every second.

**Variety**
the increasingly different types of data.

**Velocity**
the speed at which new data is generated and moves around.

**Veracity**
the reliability and predictability of imprecise data types.
Big data - Volume
Big data - Volume

Volume, the quantity of data to be stored, is a key characteristic of Big Data.

<table>
<thead>
<tr>
<th>STORAGE CAPACITY UNITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>TERM</td>
</tr>
<tr>
<td>Bit</td>
</tr>
<tr>
<td>Byte</td>
</tr>
<tr>
<td>Kilobyte</td>
</tr>
<tr>
<td>Megabyte</td>
</tr>
<tr>
<td>Gigabyte</td>
</tr>
<tr>
<td>Terabyte</td>
</tr>
<tr>
<td>Petabyte</td>
</tr>
<tr>
<td>Exabyte</td>
</tr>
<tr>
<td>Zettabyte</td>
</tr>
<tr>
<td>Yottabyte</td>
</tr>
</tbody>
</table>

*Note that because bits are binary in nature and are the basis on which all other storage values are based, all values for data storage units are defined in terms of powers of 2. For example, the prefix kilo typically means 1000; however, in data storage, a kilobyte = 2^10 = 1024 bytes.*
Volume, the quantity of data to be stored, is a key characteristic of Big Data.

How to deal with storing large volume of data?

**Scale Up:**
Keep the same number of Systems, but migrating each system to a larger System.

e.g. Changing from a server with 16 CPU cores and 1 TB storage system to a server with 64 CPU cores and a 100 TB storage system.

**Scale Out:**
When the workload exceeds the capacity of a server, the work load is spread out across a number of servers.

This is also referred to as **Clustering**.

**Notice:**
It is cheaper to buy ten 100 TB storage systems than it is to buy a single 1 PB storage system.
Big data - Velocity
**Big data - Velocity**

**Velocity**, refers to the **rate at which new data enters the system** as well as the **rate at which the data must be processed**.

Example:

**Past**

Amazon used to capture only the **data about the final transaction** of a customer making a purchase!

**Present**

Amazon captures **NOT ONLY** the final transaction **BUT ALSO** every click of the mouse in searching, browsing, comparing, as well as the purchase process.

Instead of capturing 1 event it might capture data on more than 30 events.

**30x** increase in the velocity of the data.


Copyright © DataAnalyticsResearchGroup @MQ  https://data-science-group.github.io/
Velocity, refers to the rate at which new data enters the system as well as the rate at which the data must be processed.

The velocity of processing can be broken down into: Stream and Feedback Loop Processing.

Stream Processing, requires analysis of the data stream as it enters the system. (Focus on the INPUT)

Example:
CERN Large Hadron Collider (the largest and most powerful particle accelerator in the world) experiments produce about 600 TB per second of raw data.

All this data can not be processes, accordingly scientists created algorithms to decide ahead of time which data will be kept; and to filter the data down to only about 1 GB per second.
Velocity, refers to the rate at which new data enters the system as well as the rate at which the data must be processed.

The velocity of processing can be broken down into: Stream and Feedback Loop Processing.

Feedback Loop Processing, refers to the analysis of the data to produce actionable results. (Focus on the OUTPUT)
Big data - Variety
Variety, refers to the vast array of formats and structures in which the data may be captured: structured, unstructured and semi-structured.

Structured Data, is data that has been organized to fit a predefined data model.

Unstructured Data, is data that is not organized to fit into a predefined data model.

Semi-structured Data, combines elements of both Structured and Unstructured.
Data Persistence
Data Persistence

Various related Data Islands:
From open to private and social data.

Various Technologies to persist the big data:
From Relational to NoSQL
Data Persistence

- How to store vast amount of noisy data (varying from structured entities to unstructured documents) being generated on a continuous basis?

- What technology to use for persisting the data (from Relational to NoSQL databases)?

RDBMS vs NoSQL

- Key-value Store
- Document Store
- Graph Database

Structured Data
- Scale Up
- Atomic Transactions
- Impedance Mismatch

Semi-/Un-Structured Data
- Scale Out
- Eventual Consistency
- Relaxed..
Introduction to Data Lakes

DBMS and Relational DBMS
What Is a Database Management System (DBMS) ?

A collection of files that store the data

A big C program written by someone else that accesses and updates those files for you!
What is a Relational DBMS?

A Relational Database Management System (RDBMS) is a database management system (DBMS) based on the Relational Model invented by Edgar F. Codd at IBM's San Jose Research Laboratory.

In the Relational Model, all data must be stored in relations (tables), and each relation consists of rows and columns.

Where are RDBMS used?
- Backend for traditional “database” applications
- Backend for large Websites
- Backend for Web services
NoSQL
Not only SQL!
NoSQL, is a new generation of database management systems that is not based on the traditional Relational Database Model.

<table>
<thead>
<tr>
<th>NoSQL DATABASES</th>
<th>EXAMPLE DATABASES</th>
<th>DEVELOPER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key-value database</td>
<td>Dynamo</td>
<td>Amazon</td>
</tr>
<tr>
<td></td>
<td>Riak</td>
<td>Basho</td>
</tr>
<tr>
<td></td>
<td>Redis</td>
<td>Redis Labs</td>
</tr>
<tr>
<td></td>
<td>Voldemort</td>
<td>LinkedIn</td>
</tr>
<tr>
<td>Document databases</td>
<td>MongoDB</td>
<td>MongoDB, Inc.</td>
</tr>
<tr>
<td></td>
<td>CouchDB</td>
<td>Apache</td>
</tr>
<tr>
<td></td>
<td>OrientDB</td>
<td>OrientDB Ltd.</td>
</tr>
<tr>
<td></td>
<td>RavenDB</td>
<td>Hibernating Rhinos</td>
</tr>
<tr>
<td>Column-oriented databases</td>
<td>HBase</td>
<td>Apache</td>
</tr>
<tr>
<td></td>
<td>Cassandra</td>
<td>Apache (originally Facebook)</td>
</tr>
<tr>
<td></td>
<td>Hypertable</td>
<td>Hypertable, Inc.</td>
</tr>
<tr>
<td>Graph databases</td>
<td>Neo4J</td>
<td>Neo4j</td>
</tr>
<tr>
<td></td>
<td>ArangoDB</td>
<td>ArangoDB, LLC</td>
</tr>
<tr>
<td></td>
<td>GraphBase</td>
<td>FactNexus</td>
</tr>
</tbody>
</table>
NoSQL – Key-Value Databases

Key-Value Databases, are conceptually the simplest of the NoSQL data models.

Data will be stored as a collection of Key-Value pairs.

An identifier for a value

The value can be anything such as:
- Text
- Document (XML/JSON)
- Image
- etc.

The Database does not attempt to understand the content of the value!

(It is the role of the application to analyse and understand the content)
Document Databases, are conceptually similar to Key-Value Databases, and can be considered as a sub-type of Key-Value Databases.

The Value component can only contain Documents!

The Document can be in any encoded format, such as:
- XML
- JSON
- BSON (Binary JSON)
- etc.

Unlike Key-Value Databases, the Document Database do attempt to understand the content of the Value!
Column-Oriented Database, stores the data in blocks by column instead of by rows.
NoSQL – Column-Oriented Databases

Column-Oriented Database, stores the data in **blocks by column** instead of by rows.

This type of database:
- works very well for databases that are primarily used to run queries over few columns but many rows, as is done in many reporting systems and data warehouses.
- Would be inefficient for processing transactions since Insert, Update and Delete activities would be very disk intensive.

Example: HBase, HyperTable, Cassandra.

Developed by Facebook; one of the most popular Column-Oriented DBs.
As a NoSQL DB based on Graph theory to store data about relationship-rich environments, Graph Database models relationships (edges) among objects called nodes. Modelling and storing data about relationships is the focus of Graph Databases. Interest in Graph Databases originated in the area of social networks.

http://www.cse.unsw.edu.au/~sbeheshti/EDBT16/
Search NoSQL Documents:

- **Elasticsearch** can be used to search all kinds of documents.
- Elasticsearch uses **Lucene** (an indexing and search library) and tries to make all its features available through the JSON and APIs.

Database Service:

- Dozens of new DBs! how do we choose which DB to use?
- **Solution:**
  - Manage multiple database technologies and weave them together at the app layer.
  - Make this service accessible through a single API
Database as a Service

- Database as a service (DBaaS) is a cloud computing service model that provides users with some form of access to a database without the need for setting up physical hardware, installing software or configuring for performance.

- Dozens of new DBs! how do we choose which DB to use?

- Solution: **Data Lakes**
  - Manage multiple database technologies and weave them together at the app layer.
  - Make this service accessible through a single API to support CRUD and Query data.
Data Lake
A **Data Lake** is a storage repository that holds a vast amount of raw data in its native format, including structured, semi-structured, and unstructured data.

Beheshti et al., CoreDB: a Data Lake Service, CIKM 2017; [https://github.com/unsw-cse-soc/CoreDB](https://github.com/unsw-cse-soc/CoreDB)

Copyright © DataAnalyticsResearchGroup @MQ [https://data-science-group.github.io/](https://data-science-group.github.io/)
## Data Lake vs. Data Warehouse

<table>
<thead>
<tr>
<th>Data Lake</th>
<th>Data Warehouse</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Raw Data</strong></td>
<td><strong>Data</strong></td>
</tr>
<tr>
<td><strong>Structured, Semi-Structured, Unstructured</strong></td>
<td><strong>Variety</strong></td>
</tr>
<tr>
<td><strong>Schema-on-Read</strong></td>
<td><strong>Processing</strong></td>
</tr>
<tr>
<td><strong>Designed for Low Cost Storage</strong></td>
<td><strong>Volume</strong></td>
</tr>
<tr>
<td><strong>Highly Agile, Configured and re-Configure as needed</strong></td>
<td><strong>Agility</strong></td>
</tr>
<tr>
<td><strong>Maturing</strong></td>
<td><strong>Security</strong></td>
</tr>
<tr>
<td><strong>Data Scientists and Analysts</strong></td>
<td><strong>Users</strong></td>
</tr>
</tbody>
</table>

A **data warehouse** is a system used for reporting and data analysis, and is a central repository of integrated data from one or more disparate sources.
Curating Big Data
Data curation has been defined as the active and on-going management of data through its lifecycle of interest and usefulness.

Data Curation is the process of transforming raw data into Contextualized Data.

Data curation includes all the tasks needed for principled and controlled data creation, maintenance, and management, together with the capacity to add value to data.

Freitas et al., "Big data curation". In New Horizons for a Data-Driven Economy, 2016.

Copyright © DataAnalyticsResearchGroup @MQ  https://data-science-group.github.io/
Curating Big Data

Big Data Curation involves:

- **Identifying** relevant data sources,
- **Ingesting** data and knowledge,
- **Cleaning**, 
- **Integration**, 
- **Transformation** (Normalization and aggregation), 
- **Adding Value** (Preparing Raw Data for Analytics):
  - **Extraction**
  - **Enrichment**, 
  - **Linking**, 
  - **Summarization**.

Copyright © DataAnalyticsResearchGroup @MQ  https://data-science-group.github.io/
Identifying relevant data sources

With more data repositories constantly being published every day, choosing appropriate data sources for a specific analyst GOAL becomes very important.

Private Personal/Business Data:

Social Data:

Open Data:

Copyright © DataAnalyticsResearchGroup @MQ  https://data-science-group.github.io/
Ingesting data and knowledge

**Data ingestion** is the process of obtaining and importing data for immediate use or storage in a database.

Data can be **streamed** in real time or ingested in **batches**.

When data is ingested in real time, each data item is imported as it is emitted by the source.

When data is ingested in batches, data items are imported in **discrete chunks** at periodic intervals of time.

**Hortonworks Dataflow (HDF)**
Makes Big Data Ingest Easy!

https://hortonworks.com/solutions/data-ingestion/

Copyright © DataAnalyticsResearchGroup @MQ  https://data-science-group.github.io/
Ingesting data and knowledge

Internet of things (IoT) is the network of physical objects augmented with Internet-enabled computing devices to enable those objects sense the real world.

As the number of **IoT devices** grows, both volume and variance of data sources are expanding rapidly.
Big Data Cleaning

Data Cleaning:
- also known as data cleansing and data scrubbing.
- is the process of amending or removing data in a database that is incorrect, incomplete, improperly formatted, or duplicated.
- is the number one problem in data warehousing

Other data problems which requires data cleaning
- duplicate records,
- incomplete data,
- inconsistent data
Data integration, combines data from multiple sources.

Issues during data integration:

- **Schema integration**
  - Integrate metadata (about the data) from different sources
  - Entity identification problem: identify real world entities from multiple data sources. E.g. Change of Name issue.

- **Detecting and resolving data value conflicts**
  - For the same real world entity, attribute values from different sources are different, e.g., different scales (metric vs. British units)

- **Removing duplicates and redundant data**
  - An attribute can be derived from another table (annual revenue)
  - Inconsistencies in attribute naming. E.g., A.lastName vs. B.failyName (same attribute?)


Copyright © DataAnalyticsResearchGroup @MQ https://data-science-group.github.io/
**Big Data Transformation**

**Data Transformation** is usually used to smooth the noisy data, summarize, generalize, or normalize the data scale falls within a small, specified range.

- **Smoothing**: remove noise from data (binning, clustering, regression)

- **Normalization**: scaled to fall within a small, specified range such as –1.0 to 1.0 or 0.0 to 1.0

- **Attribute/feature construction**
  - New attributes constructed / added from the given ones

- **Aggregation**: summarization or aggregation operations apply to data

- **Generalization**: concept hierarchy climbing
  - Low level/ primitive/raw data are replace by higher level concepts

(Granularity !)


Copyright © DataAnalyticsResearchGroup @MQ https://data-science-group.github.io/
Curation: Tasks for Adding Value

- Extraction,
- Enrichment,
- Linking,
- Summarization.

(Preparing Raw Data for Analytics)

Beheshti et al., "DataSynapse: A Social Data Curation Foundry". Distributed and Parallel Databases (DAPD) Journal, 2018

Copyright © DataAnalyticsResearchGroup @MQ  https://data-science-group.github.io/
Extraction – Featurized Item

Beheshti et al., "DataSynapse: A Social Data Curation Foundry", Distributed and Parallel Databases (DAPD) Journal, 2018

Copyright © DataAnalyticsResearchGroup @MQ  https://data-science-group.github.io/
Extraction – Featurized Item

Raw Tweet

Contextualized Tweet

http://www.cse.unsw.edu.au/~sbeheshti/WWW17/
Beheshti et al., "On Automating Basic Data Curation Tasks", WWW, 2017
Copyright © DataAnalyticsResearchGroup @MQ
Motivating Example:
A typical scenario for analyzing Urban Social Issues from Twitter as it relates to the Government Budget, to highlight how DataSynapse significantly improves the quality of extracted knowledge compared to the classical curation pipeline (in the absence of feature extraction, enrichment and domain-linking contextualization).

Beheshti et al., "DataSynapse: A Social Data Curation Foundry". Distributed and Parallel Databases (DAPD) Journal, 2018

Copyright © DataAnalyticsResearchGroup @MQ

https://data-science-group.github.io/
Knowledge Lake
A Knowledge Lake, i.e. a contextualized Data Lake, is a centralized repository containing virtually inexhaustible amounts of both data and contextualized data that is readily made available to perform analytical activities.

Beheshti et al., CoreKG: a Knowledge Lake Service (VLDB'18), https://github.com/unsw-cse-soc/CoreKG

Copyright © DataAnalyticsResearchGroup @MQ https://data-science-group.github.io/
Motivating Scenario
Motivating Scenario

Police Investigation for Missing Person

BPM 2018:
• "iProcess: Enabling IoT Platforms in Data-Driven Knowledge-Intensive Processes"

ICSOC 2018:
• "iCOP: IoT-enabled Policing Processes"
• "iSheets: A Spreadsheet-based Machine Learning Development Platform for Data-driven Process Analytics"
Enabling IoT in Policing

IoT will enable us connect people, things and businesses!

**Goal:**
- Fully connected enterprise
- Integrating IoT and Business Data
- Harness real-time intelligence
- Build New Business Models
- Support **Knowledge-Intensive Data-Driven** Processes

**Challenge:**
- To enable analysts *ingest* data from IoT devices, *extract* knowledge from this data and *link* them to process (execution) data.
Data-Driven Knowledge-Intensive Processes


Copyright © DataAnalyticsResearchGroup @MQ

https://data-science-group.github.io/
Motivating Scenario: Missing Person!

In Australia, more than 38,000 people are reported missing each year.


In USA, on any given day, there are as many as 100,000 active missing person's cases.

https://nij.gov
**Solution: iProcess**

**iProcess:** is a scalable and extensible IoT-Enabled Process Data Analytics Pipeline to enable analysts ingest data from IoT devices, extract knowledge from this data and link them to process execution data.

Beheshti et al., "iProcess: Enabling IoT Platforms in Data-Driven Knowledge-Intensive Processes", 16th conference on Business Process Management (BPM), Sydney, Australia, 2018

Copyright © DataAnalyticsResearchGroup @MQ  
https://data-science-group.github.io/
Solution: iProcess

Beheshti, Benatallah, et al., CoreDB: a Data Lake Service (CIKM’17) — https://github.com/unsw-cse-soc/CoreDB

Copyright © DataAnalyticsResearchGroup @MQ — https://data-science-group.github.io/
Solution: iProcess

Knowledge Lake as a Service

SPARQL
Full-Text Search

SQL Query

Knowledge Lake

Data Lake

NoSQL Databases
- MongoDB
- CouchDB
- HBase
- Hive

Relational Databases
- Oracle
- MySQL
- PostgreSQL
- SQL Server

CoreKG: a Knowledge Lake Service (VLDB'18)
https://github.com/unsw-cse-soc/CoreKG

Beheshti, Benatallah, et al., CoreKG: a Knowledge Lake Service (VLDB’18) https://github.com/unsw-cse-soc/CoreKG

Beheshti, Benatallah, et al., DataSynapse: A Social Data Curation Foundry (DAPD Journal, 2018)

Copyright © DataAnalyticsResearchGroup @MQ
https://data-science-group.github.io/
Solution: iProcess

OLAP, is an approach to answering multi-dimensional analytical queries swiftly.

Problem:
- extension of existing OLAP techniques to analysis of graphs is not straightforward.
- key business insights remain hidden in the interactions among objects.

Solution:
- On-Line Analytical Processing on Graphs


Copyright © DataAnalyticsResearchGroup @MQ

https://data-science-group.github.io/
Solution: iProcess


Copyright © DataAnalyticsResearchGroup @MQ https://data-science-group.github.io/
Solution: iProcess

ML as a Service

Machine Learning

Supervised Learning
- Regression
  - Linear Regression
  - Non-Linear Regression
- Classification
  - Decision Tree
  - Random Forest
  - Support Vector Machine
  - K-Nearest Neighbors
  - Naive Bayes
  - Ensemble Learning
    - Random Forest
    - AdaBoost
    - Gradient Boosting

Unsupervised Learning
- Clustering
  - K-means
  - DBSCAN
  - AGNES
  - CLARA
  - Outlier Detection
  - Dimensionality Reduction
  - Association Rule Learning

Copyright © DataAnalyticsResearchGroup @MQ

https://data-science-group.github.io/
Solution: iProcess

Formula Bar:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>group(&quot;location&quot;, &quot;time&quot;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brisbane</td>
<td>N/A</td>
</tr>
<tr>
<td>Canberra</td>
<td>N/A</td>
</tr>
<tr>
<td>Melbourne</td>
<td>Tweets</td>
</tr>
<tr>
<td>Perth</td>
<td>N/A</td>
</tr>
<tr>
<td>Sydney</td>
<td>Tweets</td>
</tr>
</tbody>
</table>

classify.by(topic)()

<table>
<thead>
<tr>
<th>Sport</th>
<th>Politics</th>
<th>Business</th>
<th>Education</th>
<th>Entertainment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sydney-2016-04-05</td>
<td>N/A</td>
<td>Tweets</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Sydney-2016-04-06</td>
<td>N/A</td>
<td>Tweets</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Sydney-2016-04-07</td>
<td>N/A</td>
<td>Tweets</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

classify.bySentiment()

<table>
<thead>
<tr>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweets</td>
<td>Tweets</td>
</tr>
</tbody>
</table>

ML as a Service

- Classification
  - SVM
  - NN
- Logistic Regression
- C5.0
- MLP
- Trained Topic Classifier
- Trained Sentiment Classifier

- Clustering
- Association Learning
- Operations
  - Select
  - Group
  - Sort
  - Partition
  - Addition

Copyright © DataAnalyticsResearchGroup @MQ  
https://data-science-group.github.io/
Solution: iProcess

Motivating Scenario
Missing People!

Information Collection
(Use iCoPs to take a photo, record the interview with witnesses, etc)

IoT Devices
(Use services to ingest data from IoT devices)

Information Extraction

Knowledge Graph

Realtime Dashboard
Select Entities to see Evidences and Facts

Summarization
Entities and Facts

ICSOC 2018:
• "iCOP: IoT-enabled Policing Processes"
• "iSheets: A Spreadsheet-based Machine Learning Development Platform for Data-driven Process Analytics"

Copyright © DataAnalyticsResearchGroup @MQ
https://data-science-group.github.io/
Solution: iProcess

Motivating Scenario

Missing People!

ICSOC 2018:
- "iCOP: IoT-enabled Policing Processes"
- "iSheets: A Spreadsheet-based Machine Learning Development Platform for Data-driven Process Analytics"

Copyright © DataAnalyticsResearchGroup @MQ

https://data-science-group.github.io/
Intelligent Knowledge Lakes!
Other Scenarios
Beheshti et al., "iProcess: Enabling IoT Platforms in Data-Driven Knowledge-Intensive Processes", BPM, 2018

Copyright © DataAnalyticsResearchGroup @MQ  https://data-science-group.github.io/
AI-Enabled Policing
(Applications)

Beheshti et al., "personality2vec: Enabling the Analysis of Behavioral Disorders in Social Networks", 13th ACM International WSDM Conference (WSDM), Houston, Texas, USA, 2020. (ERA Rank: A*)

Copyright © DataAnalyticsResearchLab @MQ  https://data-science-group.github.io/
AI-enabled Processes (AIP) Research Centre

Web: https://aip-research-center.github.io/
email: aip@mq.edu.au
Address: Level 3, Becton-Dickinson (BD) Building, 4 Research Park Drive, Macquarie University, Sydney, Australia.

Copyright © DataAnalyticsResearchLab @MQ https://data-science-group.github.io/
Thank You!