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/

Introduction to Big Data

Introduction to Big Data

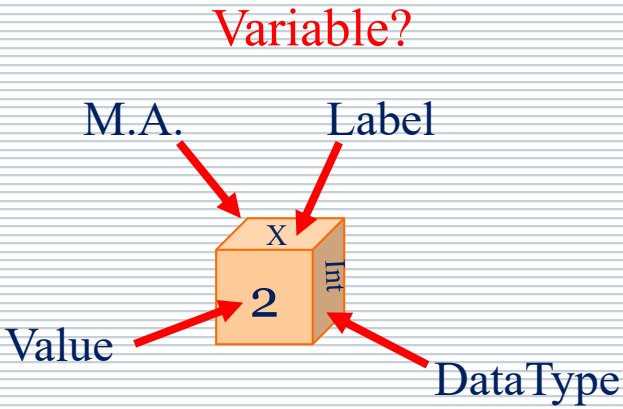
Application

Set of Related **Programs**

Set of Related **Functions**

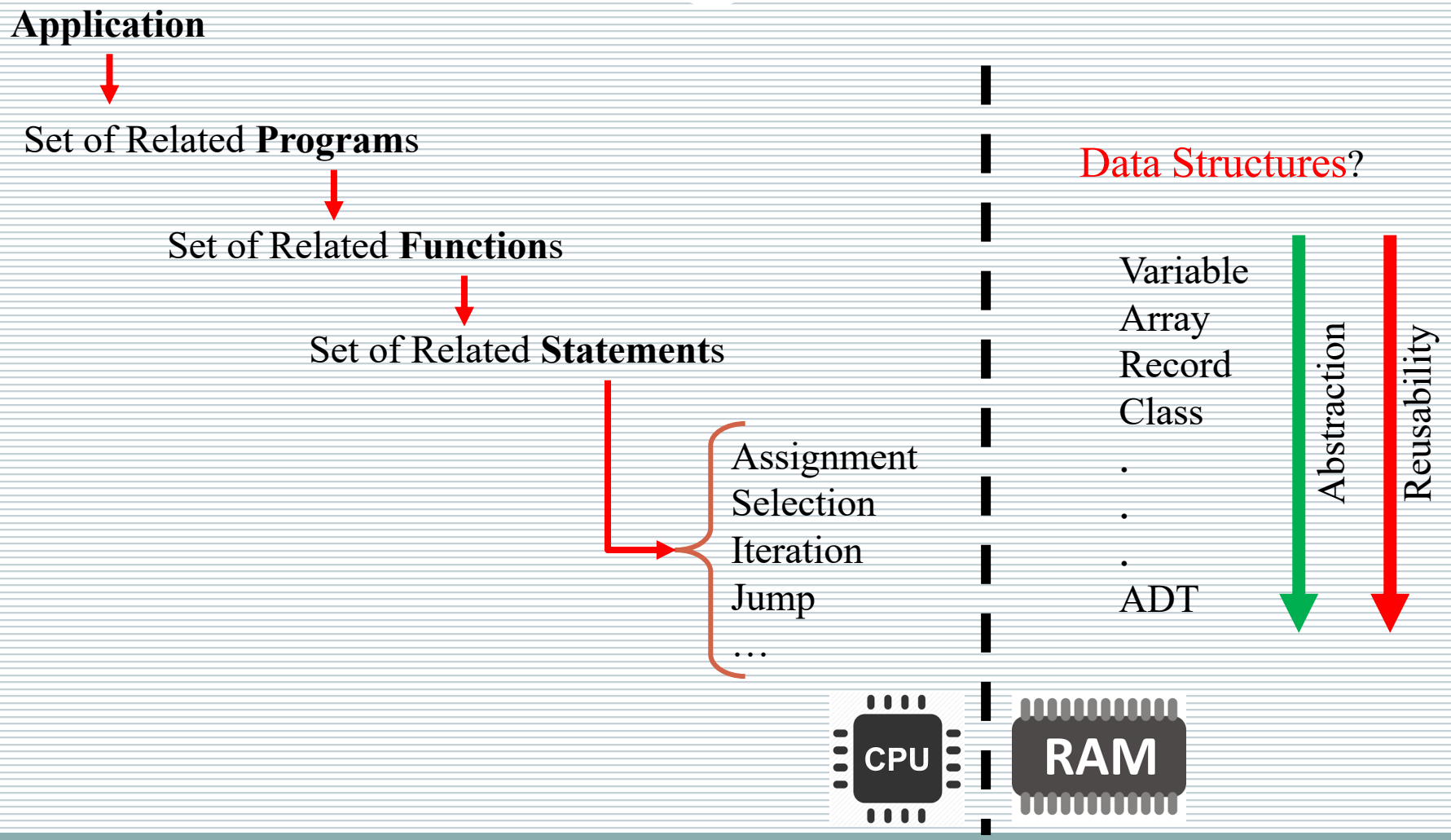
Set of Related **Statements**

- Assignment
- Selection
- Iteration
- Jump
- …

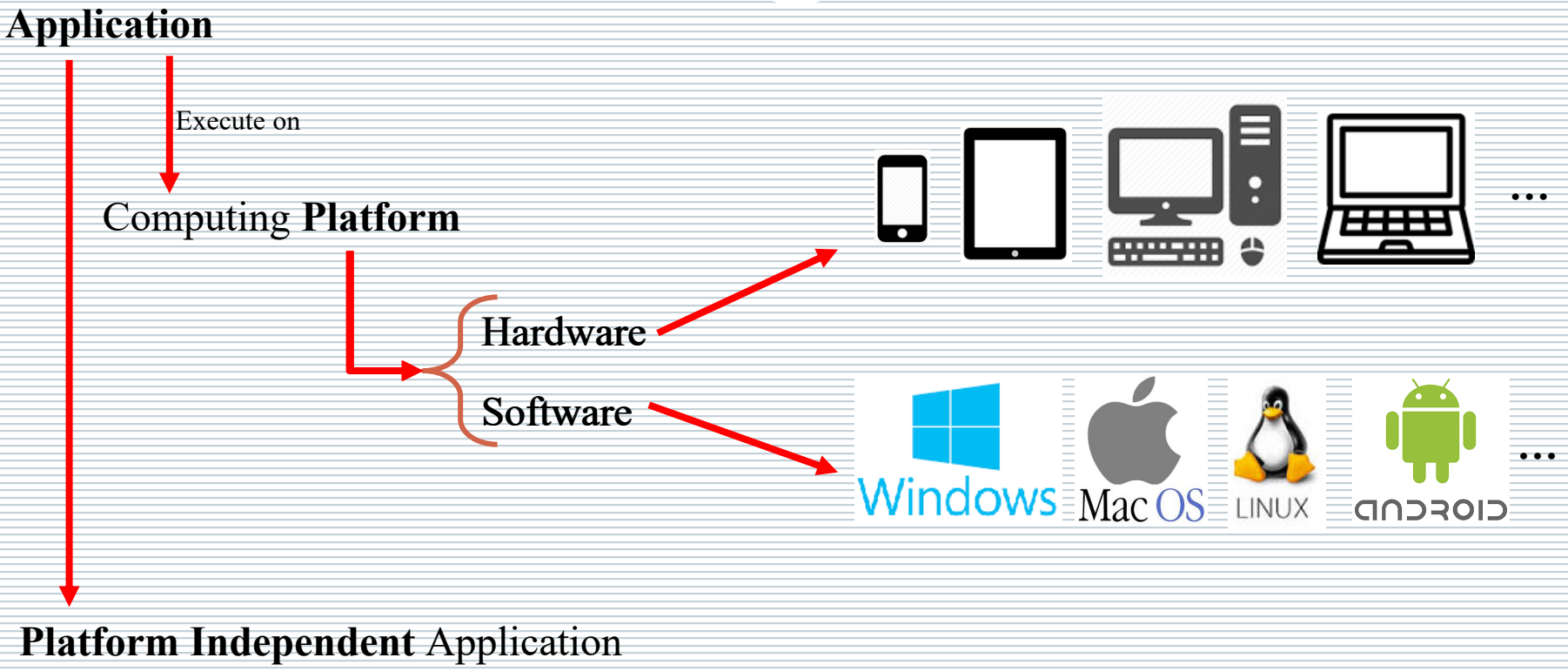


X ← 2

Introduction to Big Data



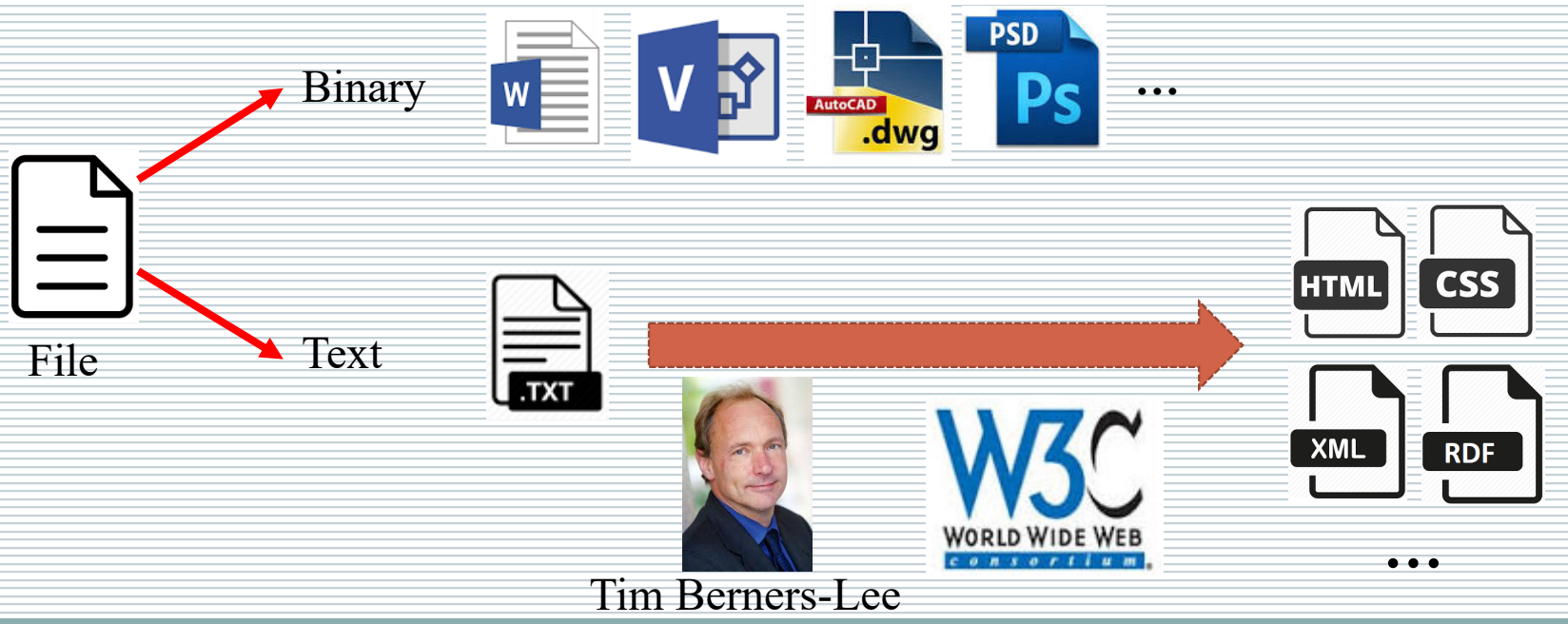
Introduction to Big Data



Introduction to Big Data

Application

Platform Independent Application
(e.g. Web Applications)

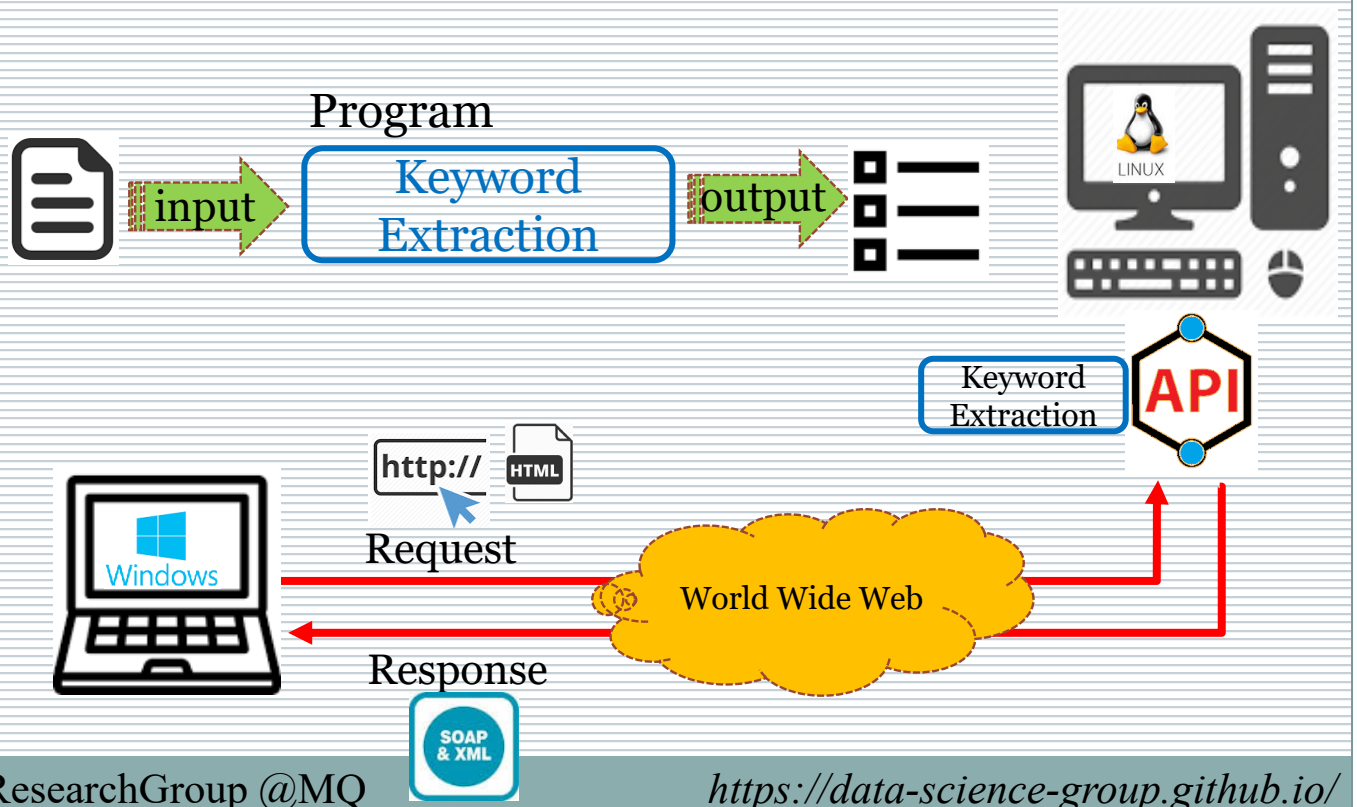


Introduction to Big Data

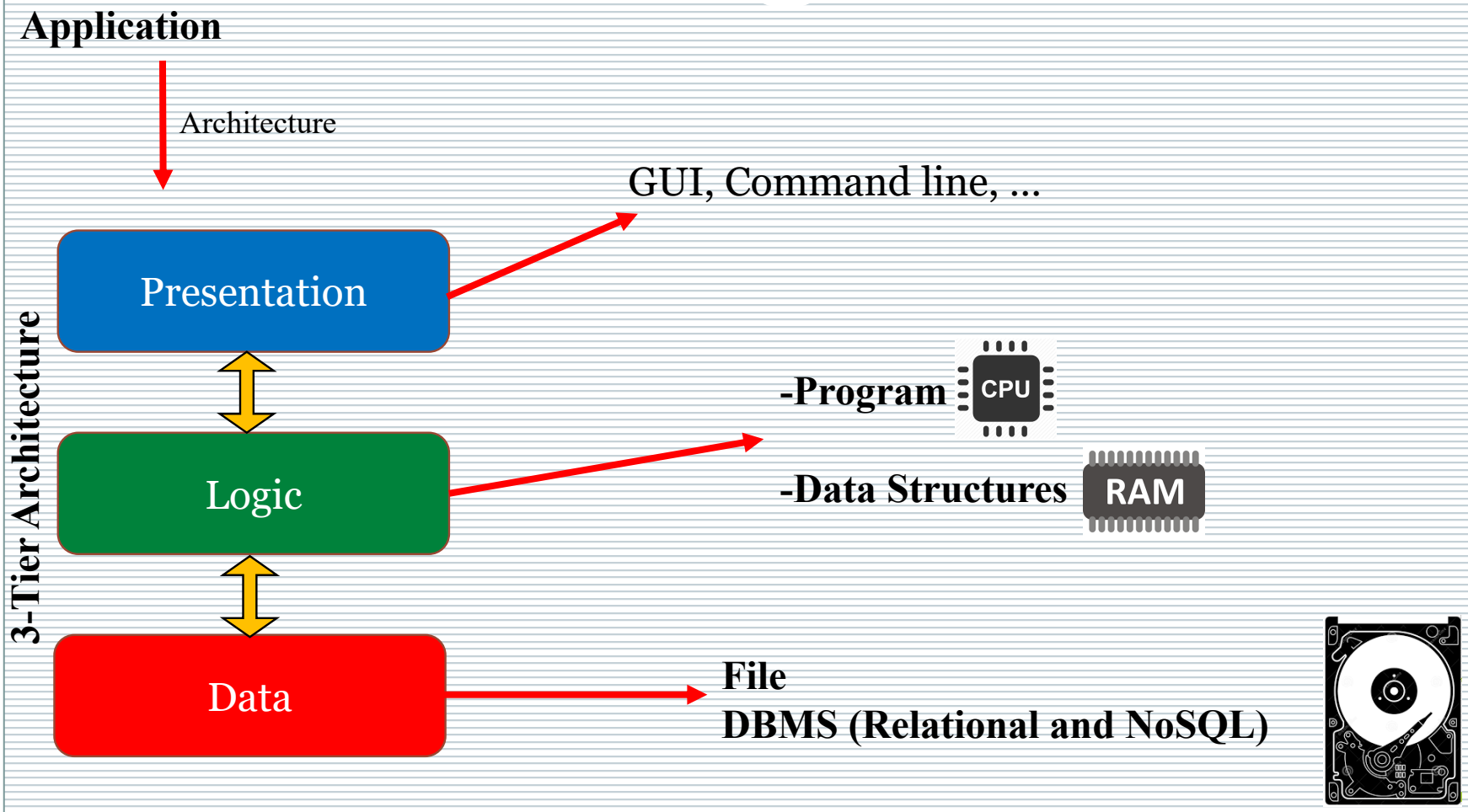
Application

Platform Independent Application
(e.g. Web Applications)

Web Services



Introduction to Big Data



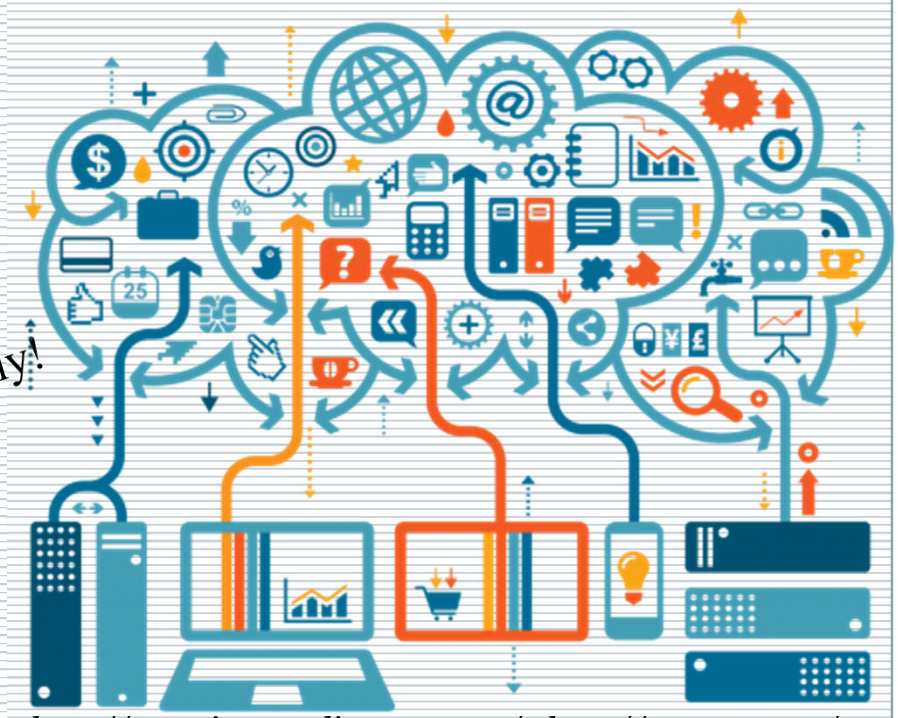
What is Data ?

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
- …

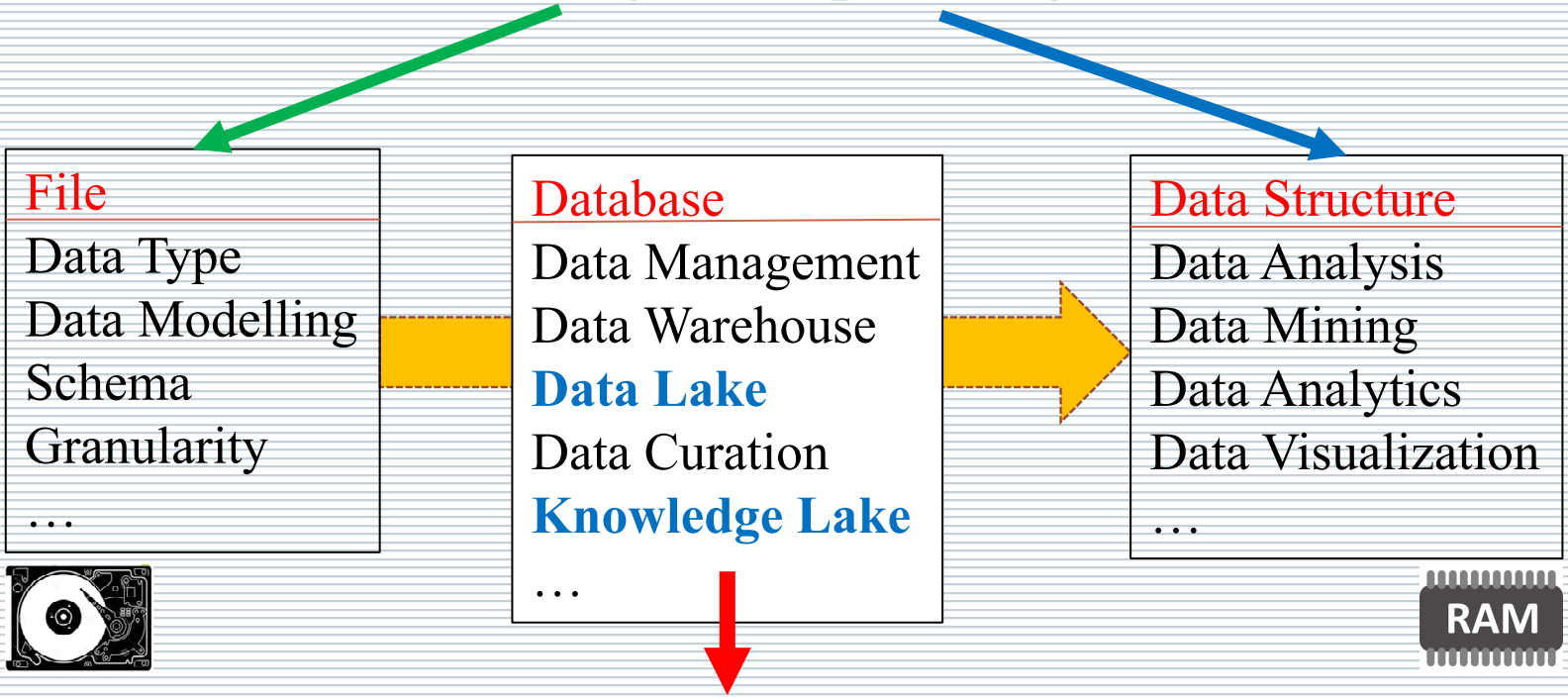
- 500 Million Tweets sent each day!
- 5.75 BILLION Facebook likes every day.
- 3.6 Billion Instagram Likes each day.
- 4.3 BILLION Facebook messages posted daily!
- 6 BILLION daily Google Searches!
- …



https://www-01.ibm.com/; http://www.internetlivestats.com/; http://semeon.com/

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)

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,
  - …

…

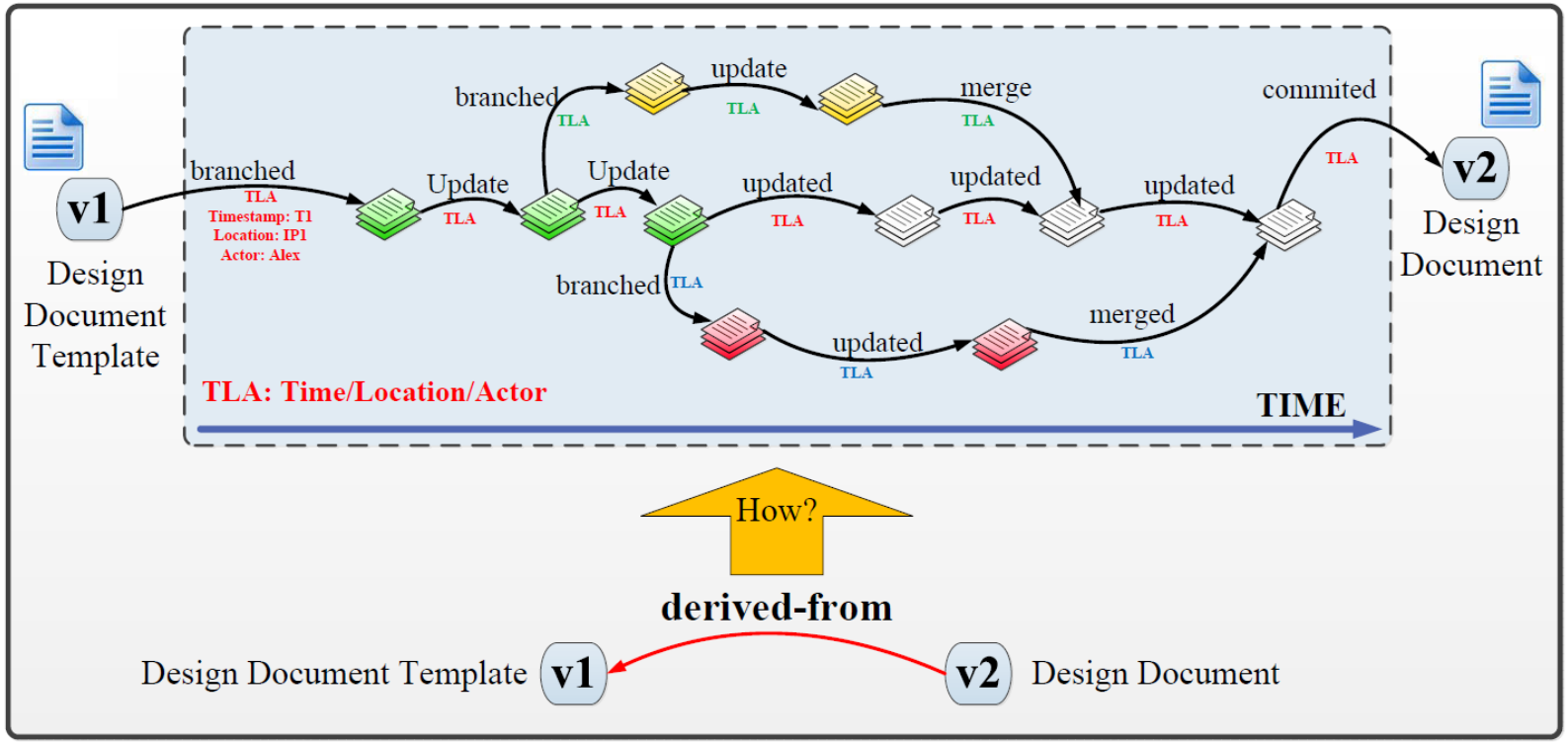
Beheshti et al. "**Enabling the Analysis of Cross-Cutting Aspects in Ad-hoc Processes**", CAiSE Conference (2013)

Example

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.



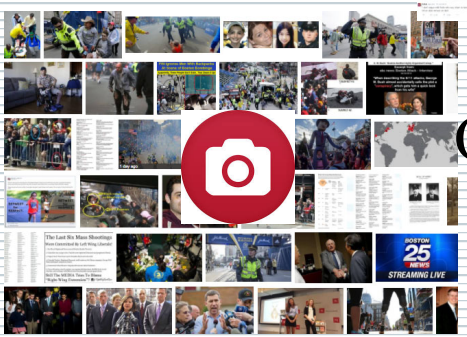
Beheshti et al. "Temporal provenance model (TPM): model and query language, 2012
https://www.w3.org/2005/Incubator/prov/wiki/What_Is_Provenance

What is Big Data ?

Example

What is Big Data ?

Social Data



Open Data

Police Investigation
(e.g. Boston Marathon Bombing)

Private Data
(personal/business)

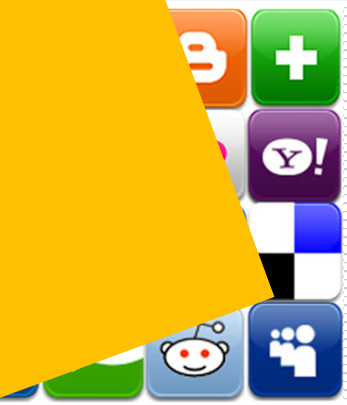
Beheshti et al. “ProcessAtlas: A scalable and extensible platform for business process analytics”, Software: Practice and Experience, 2018

Example

What is Big Data ?

Typical properties of the big data:

- wide physical distribution
- diversity of formats
- non-standard data models
- independently-managed
- heterogeneous semantics



Bombing)

Private Data
(personal/business)

Beheshti et al. “ProcessAtlas: A scalable and extensible platform for business process analytics”, Software: Practice and Experience, 2018

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

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.

STORAGE CAPACITY UNITS

TERM	CAPACITY	ABBREVIATION
Bit	0 or 1 value	b
Byte	8 bits	B
Kilobyte	1024* bytes	KB
Megabyte	1024 KB	MB
Gigabyte	1024 MB	GB
Terabyte	1024 GB	TB
Petabyte	1024 TB	PB
Exabyte	1024 PB	EB
Zettabyte	1024 EB	ZB
Yottabyte	1024 ZB	YB

* 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.

Database Systems, Design, Implementation, & Management, 13th Edition, Carlos Coronel – Steven Morris

Big data - Volume

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!

amazon

Amazon.com

Electronic commerce company

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.

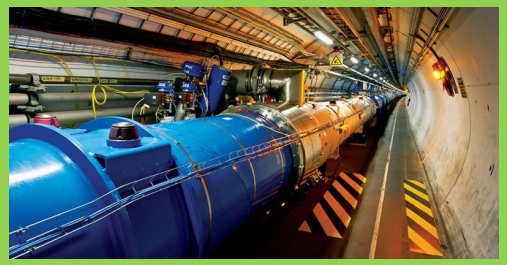
30× increase in the velocity of the data.

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**.

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.

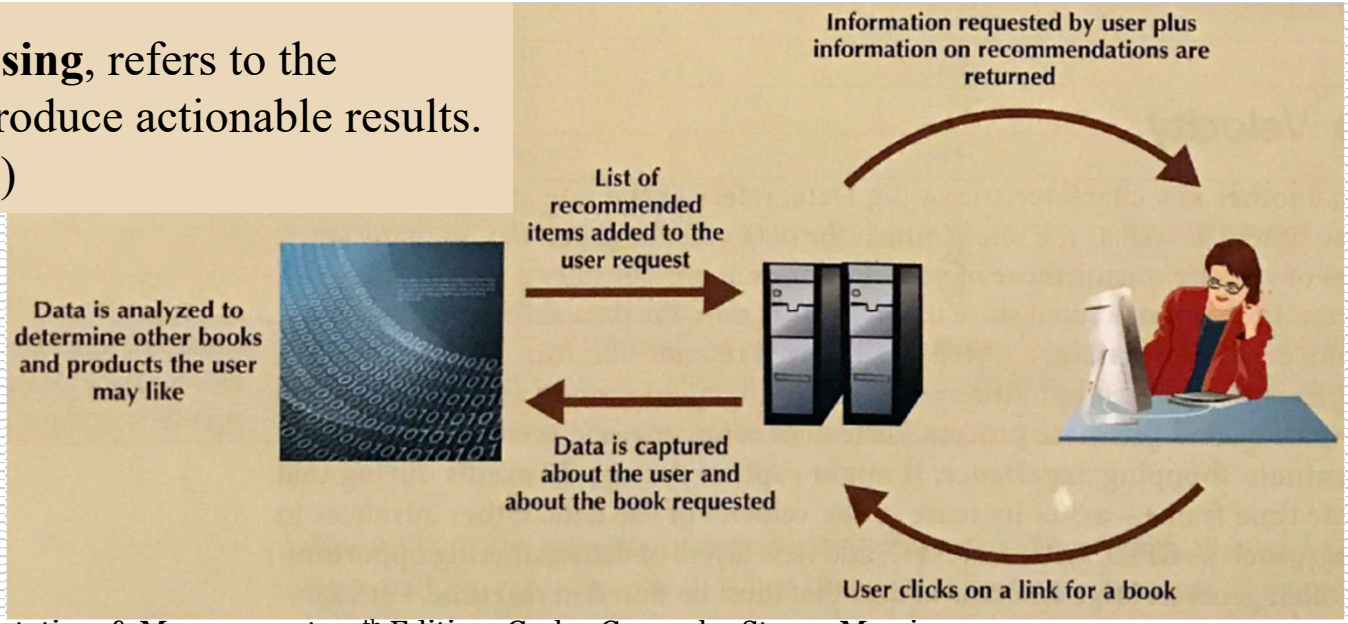
All this data can not be processes, accordingly scientists created algorithms to decide ahead of time which data will be kept; and to **filter the data down** to only about 1 GB per second.

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**.

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)



Database Systems, Design, Implementation, & Management, 13th Edition, Carlos Coronel – Steven Morris

Big data - Variety

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:

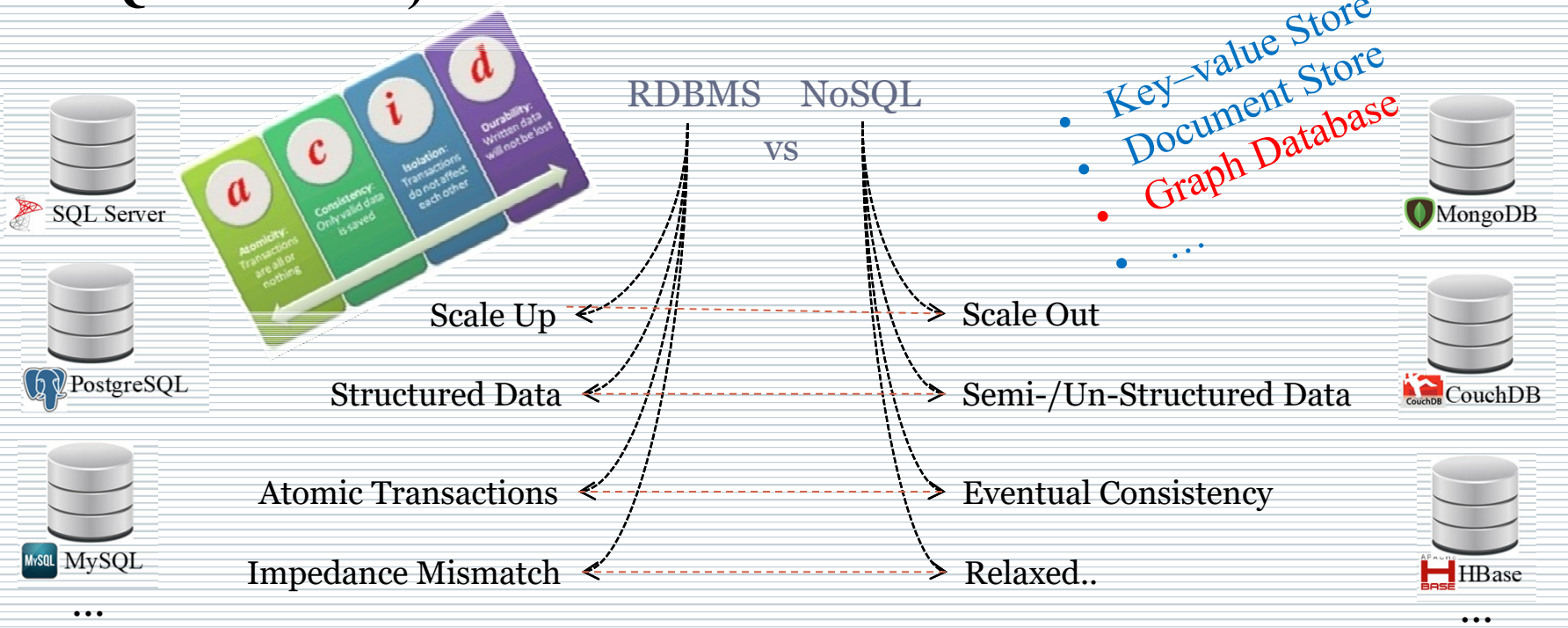
From open to private and social data.

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)?**



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

NoSQL, is a new generation of database management systems that is not based on the traditional Relational Database Model.

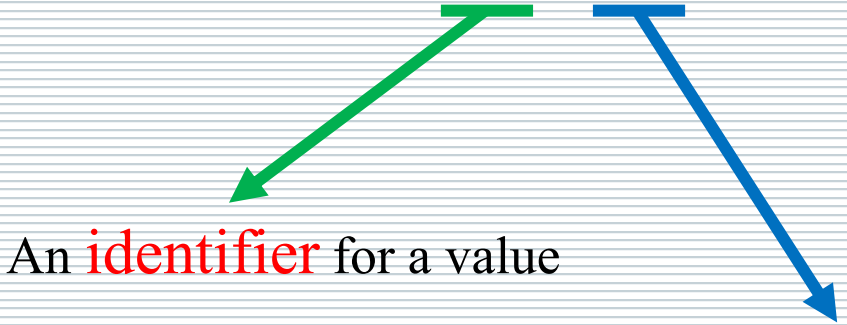
NoSQL DATABASES		
NoSQL CATEGORY	**EXAMPLE DATABASES**	**DEVELOPER**
Key-value database	Dynamo Riak Redis Voldemort	Amazon Basho Redis Labs LinkedIn
Document databases	MongoDB CouchDB OrientDB RavenDB	MongoDB, Inc. Apache OrientDB Ltd. Hibernating Rhinos
Column-oriented databases	HBase Cassandra Hypertable	Apache Apache (originally Facebook) Hypertable, Inc.
Graph databases	Neo4J ArangoDB GraphBase	Neo4j ArangoDB, LLC FactNexus

Database Systems, Design, Implementation, & Management, 13th Edition, Carlos Coronel – Steven Morris

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)

NoSQL – Document Databases

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!

NoSQL – Column-Oriented Databases

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

CUSTOMER relational table

Cus_Code	Cus_LName	Cus_FName	Cus_City	Cus_State
10010	Ramas	Alfred	Nashville	TN
10011	Dunne	Leona	Miami	FL
10012	Smith	Kathy	Boston	MA
10013	Olowski	Paul	Nashville	TN
10014	Orlando	Myron		
10015	O'Brian	Amy	Miami	FL
10016	Brown	James		
10017	Williams	George	Mobile	AL
10018	Farriss	Anne	Opp	AL
10019	Smith	Olette	Nashville	TN

Row-centric storage

Block 1	Block 4
10010,Ramas,Alfred,Nashville,TN 10011,Dunne,Leona,Miami,FL	10016,Brown,James,NULL,NULL 10017,Williams,George,Mobile,AL
Block 2	**Block 5**
10012,Smith,Kathy,Boston,MA 10013,Olowski,Paul,Nashville,TN	10018,Farriss,Anne,OPP,AL 10019,Smith,Olette,Nashville,TN
Block 3	
10014,Orlando,Myron,NULL,NULL 10015,O'Brian,Amy,Miami,FL	

Column-centric storage

Block 1	Block 4
10010,10011,10012,10013,10014 10015,10016,10017,10018,10019	Nashville,Miami,Boston,Nashville,NULL Miami,NULL,Mobile,Opp,Nashville
Block 2	**Block 5**
Ramas,Dunne,Smith,Olowski,Orlando O'Brian,Brown,Williams,Farriss,Smith	TN,FL,MA,TN,NULL, FL,NULL,AL,AL,TN
Block 3	
Alfred,Leona,Kathy,Paul,Myron Amy,James,George,Anne,Olette	

Database Systems, Design, Implementation, & Management, 13th Edition, Carlos Coronel – Steven Morris

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.

NoSQL – Graph Databases

Graph Database, is a NoSQL DB based on Graph theory to store data about relationship-rich environments.

A Mathematical and Computer Science field that 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**.

Beheshti et al., "Galaxy: A Platform for Explorative Analysis of Open Data Sources", **EDBT**, 2016.
http://www.cse.unsw.edu.au/~sbeheshti/EDBT16/

Indexing Big Data

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.

https://www.elastic.co/products/elasticsearch
https://lucene.apache.org/

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.

  https://www.contentful.com/
  https://orchestrate.io

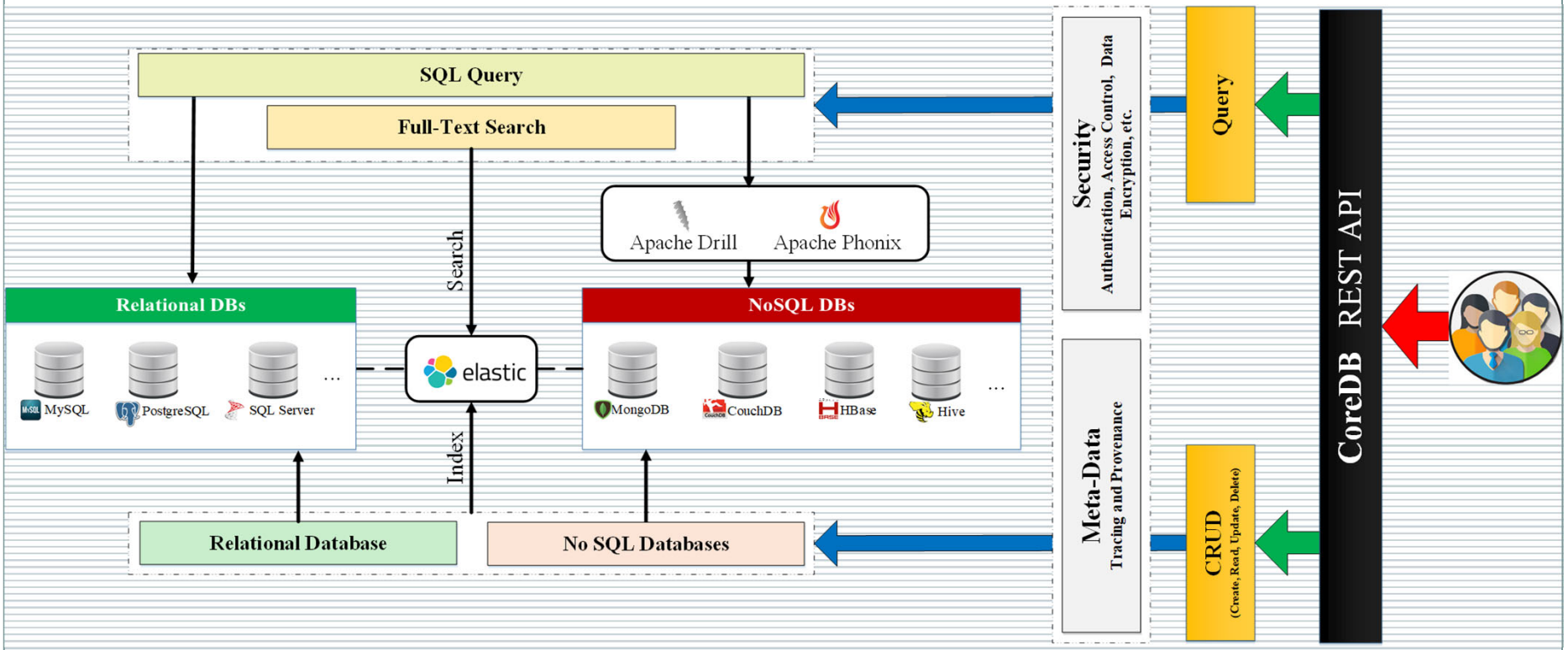
- Dozens of new DBs! how do we choose which DB to use?

- 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

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

Data Lake vs. Data Warehouse

Data Lake		Data Warehouse
Raw Data	Data	Proessed Data
Structured, Semi-Structured, Unstrutured	Variety	Structured
Schema-on-Read	Processing	Schema-on-Write
Designed for Low Cost Storage	Volume	Expensive for Large Data Volumes
Highly Agile, Configured and re-Configure as needed	Agility	Less Agile, Fixed Configuration
Maturing	Security	Mature
Data Scientists and Analysts	Users	Business Analysts

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

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.
Arocena et al., "Benchmarking data curation systems". IEEE Data Eng. Bull., 2016.

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**.

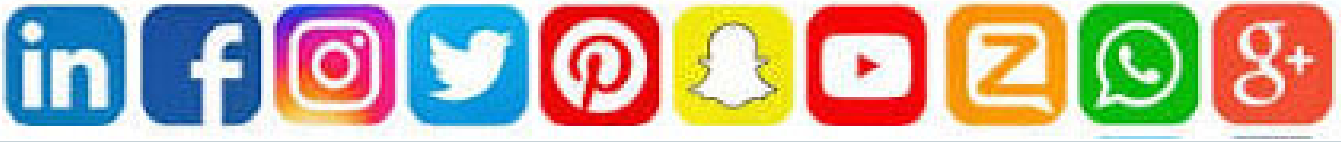
Identifying relevant data sources

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:

linkeddata.org/ ...

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**.

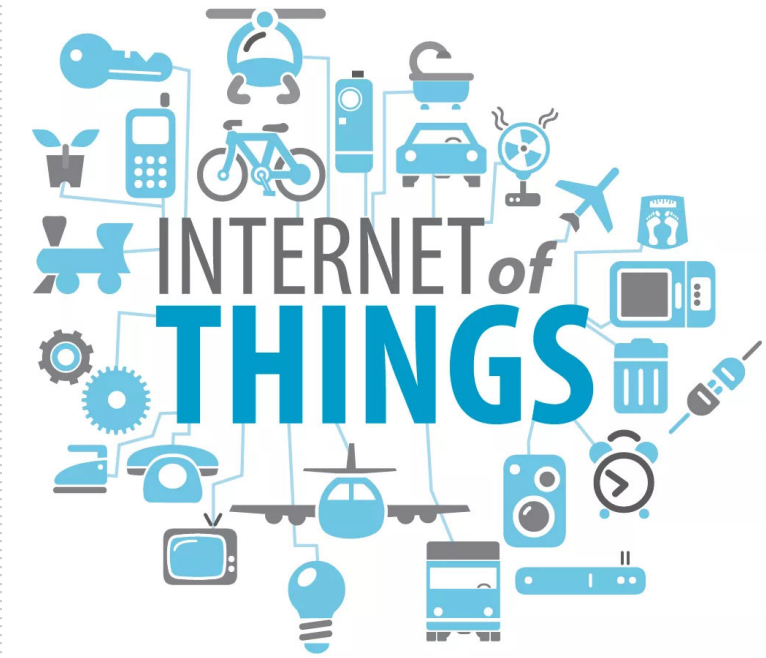
When data is ingested in real time, each data item is imported **as it is** emitted by the source.

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/

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

Big Data **Integration**

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?)

http://www.cs.ccsu.edu/~markov/ccsu_courses/datamining-3.html

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 !)

http://www.cs.ccsu.edu/~markov/ccsu_courses/datamining-3.html

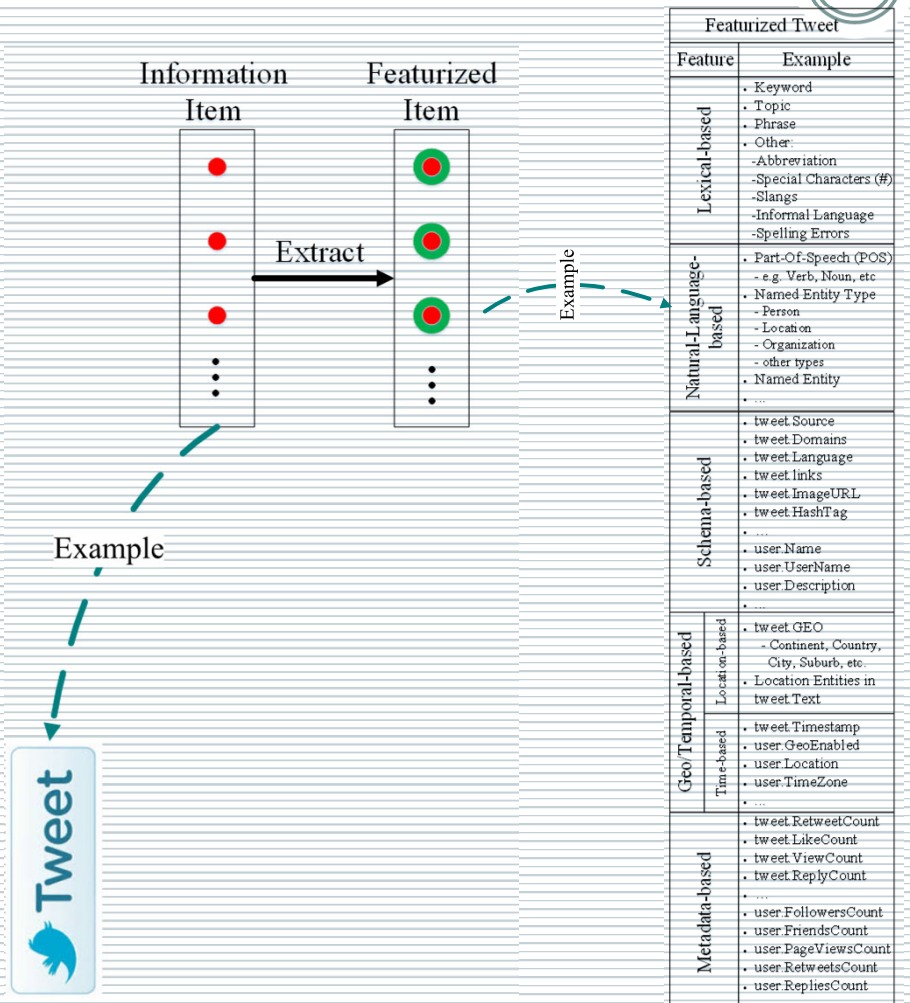
Curation: Tasks for Adding Value

- **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

Extraction – Featurized Item

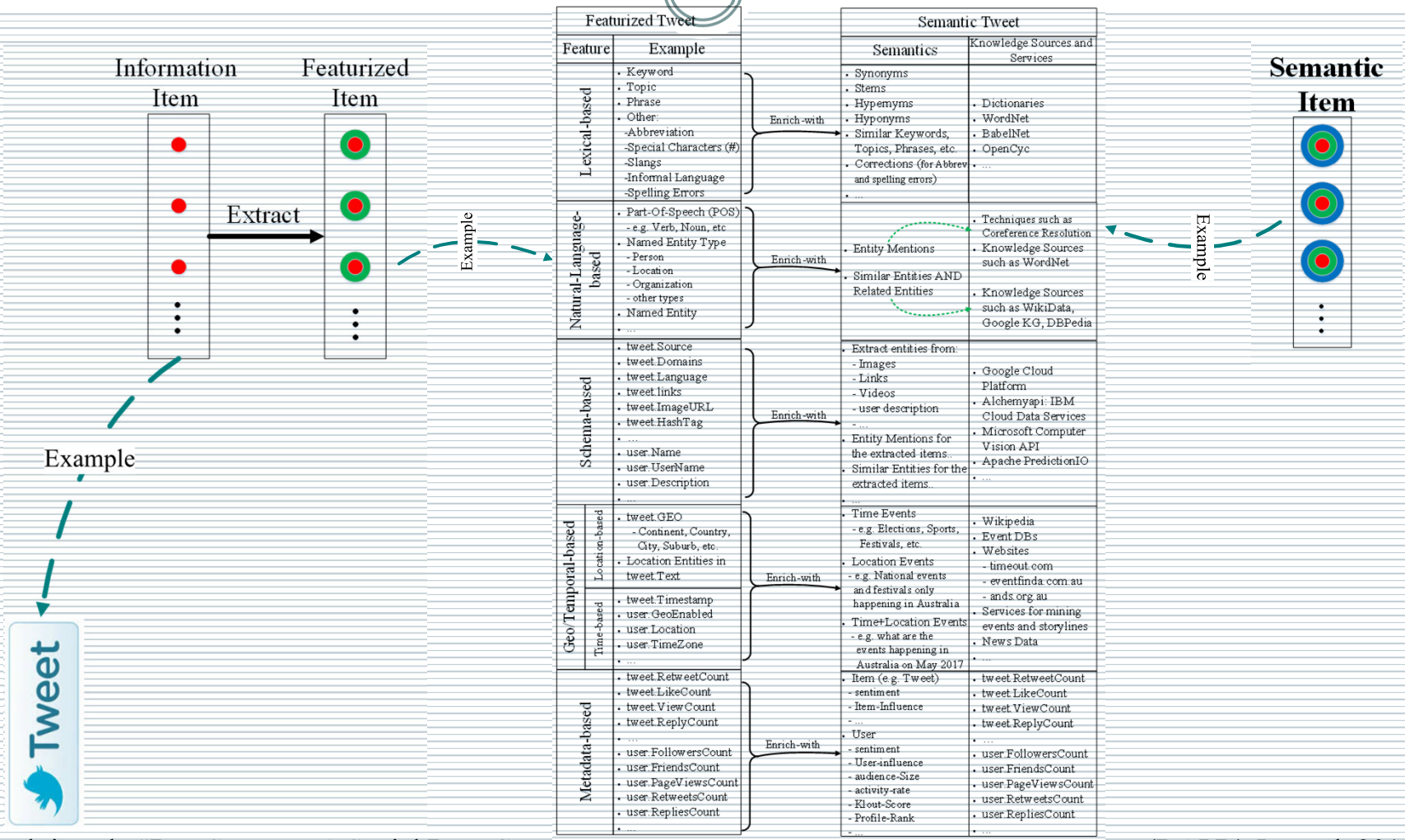


Featurized Tweet

Feature		Example
Lexical-based		• Keyword • Topic • Phrase • Other: -Abbreviation -Special Characters (#) -Slangs -Informal Language -Spelling Errors
Natural-Language-based		• Part-Of-Speech (POS) - e.g. Verb, Noun, etc • Named Entity Type - Person - Location - Organization - other types • Named Entity • ...
Schema-based		• tweet.Source • tweet.Domains • tweet.Language • tweet.links • tweet.ImageURL • tweet.HashTag • ... • user.Name • user.UserName • user.Description • ...
Geo/Temporal-based	Location-based	• tweet.GEO - Continent, Country, City, Suburb, etc. • Location Entities in tweet.Text
	Time-based	• tweet.Timestamp • user.GeoEnabled • user.Location • user.TimeZone • ...
Metadata-based		• tweet.RetweetCount • tweet.LikeCount • tweet.ViewCount • tweet.ReplyCount • ... • user.FollowersCount • user.FriendsCount • user.PageViewsCount • user.RetweetsCount • user.RepliesCount • ...

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

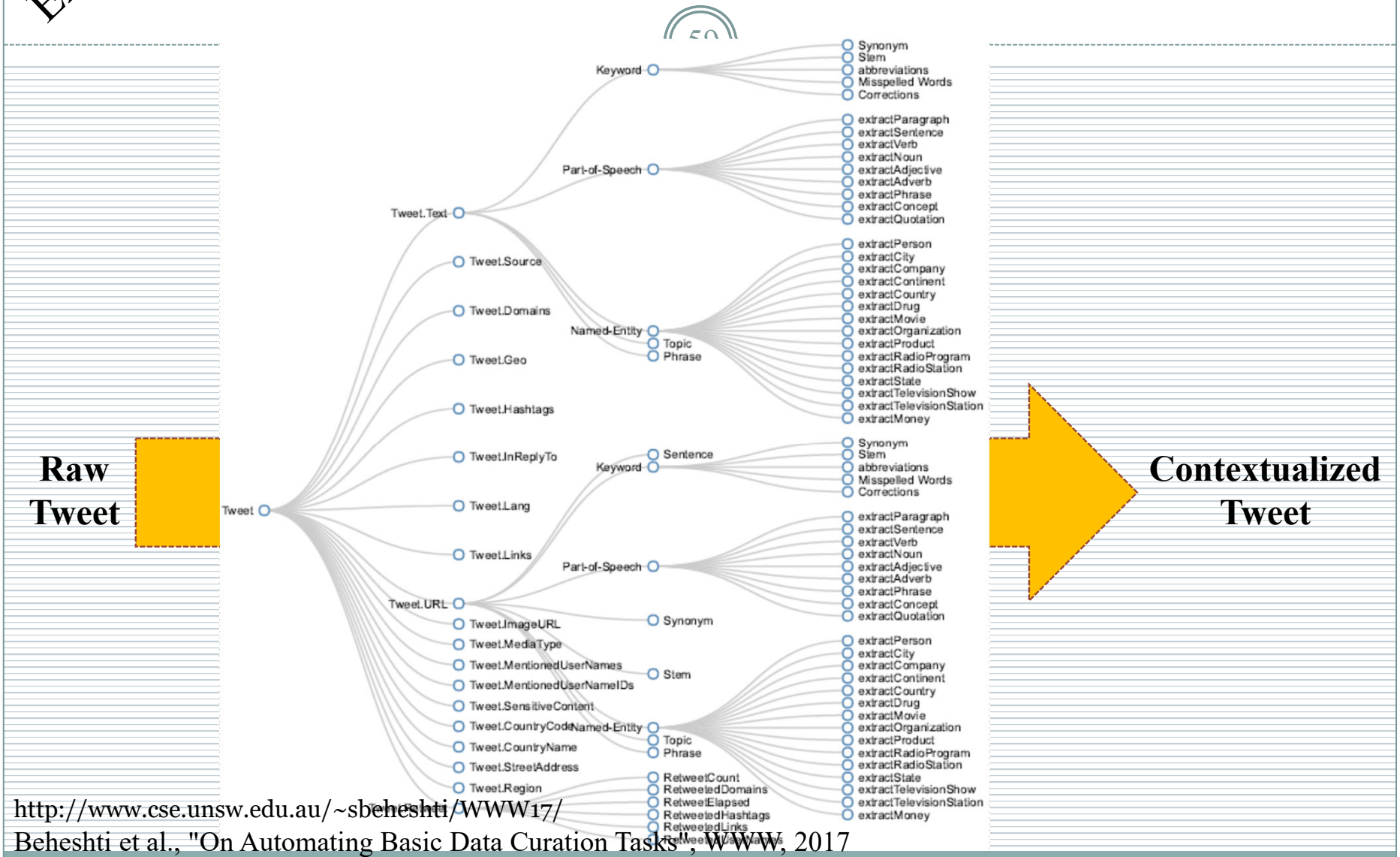
Enrichment – Semantic Item



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

Example

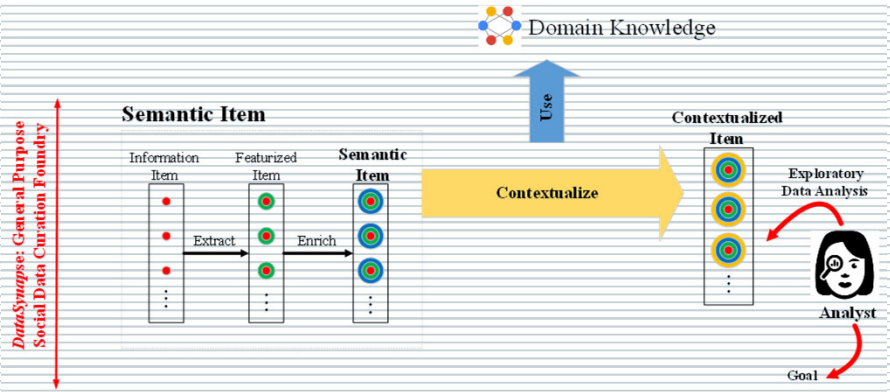
Extraction – Featurized Item

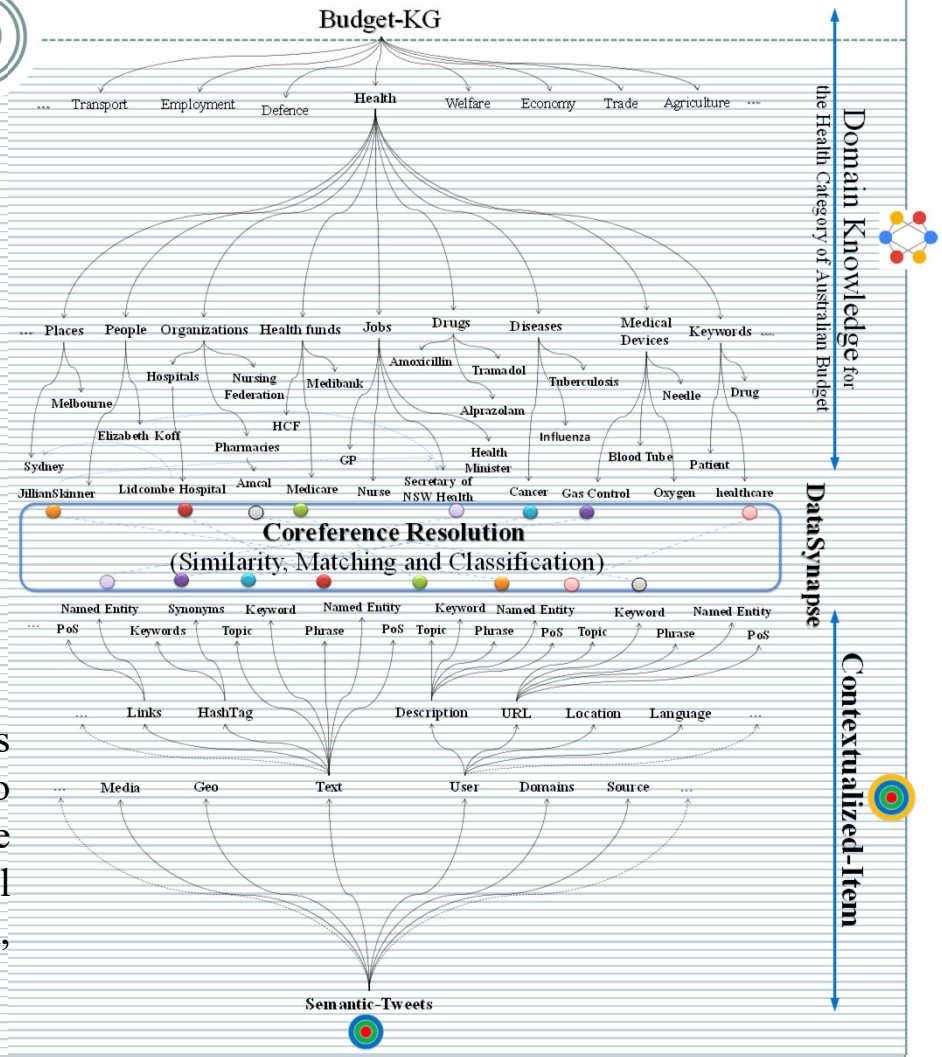


http://www.cse.unsw.edu.au/~sbeheshti/WWW17/

Beheshti et al., "On Automating Basic Data Curation Tasks", WWW, 2017

Linking – Contextualized Item Item





Motivating Example:

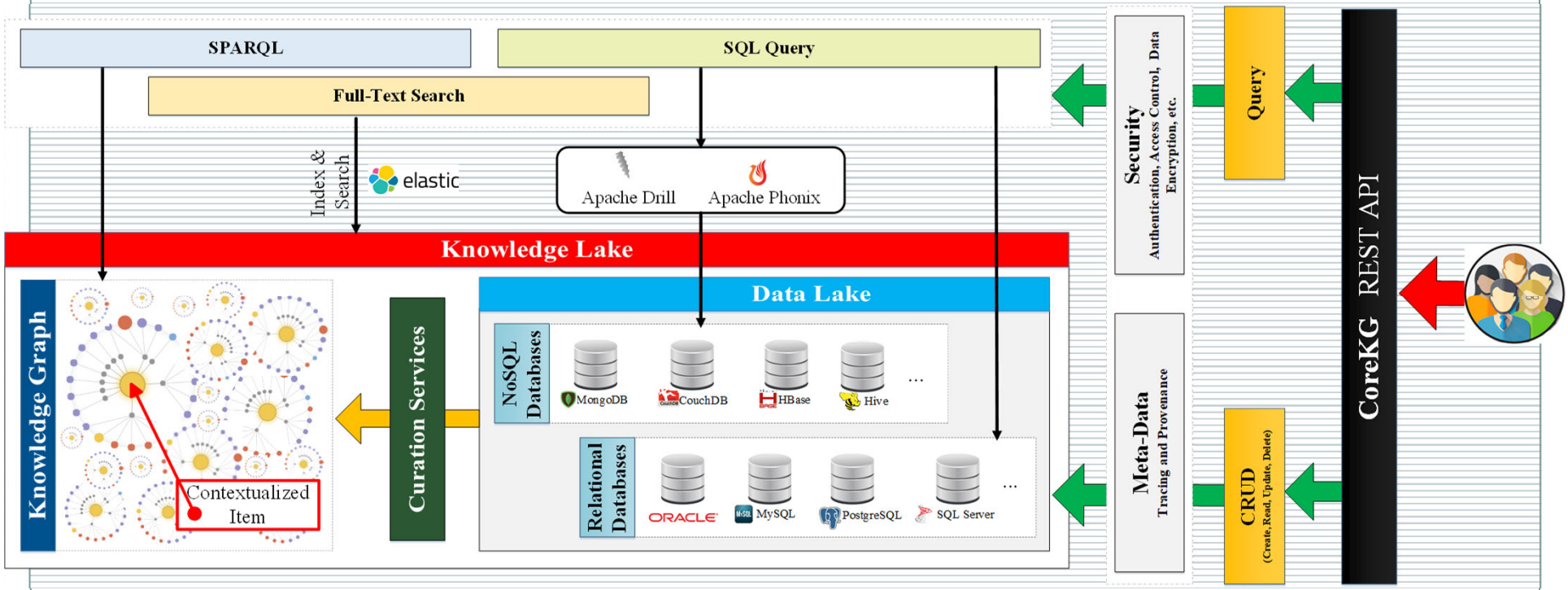
A typical scenario for analyzing Urban Social Issues from Twitter as it relates to the Government Budget, to highlight how DataSynapse significantly improves the quality of extracted knowledge compared to the classical curation pipeline (in the absence of feature extraction, enrichment and domain-linking contextualization).

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

Knowledge Lake

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

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:

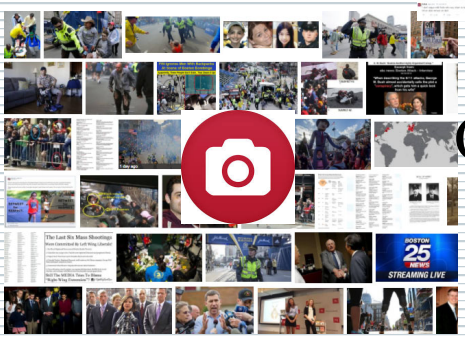
- Fully connected enterprise
- Integrating IoT and Business Data
- Harness real-time intelligence
- Build New Business Models
- Support **Knowledge-Intensive Data-Driven** Processes

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

Social Data



Open Data

Police Investigation

Private Data
(personal/business)

Beheshti et al. "ProcessAtlas: A scalable and extensible platform for business process analytics", Software: Practice and Experience, 2018

Enabling IoT Platforms in Data-Driven Knowledge-Intensive Processes

Motivating Scenario:

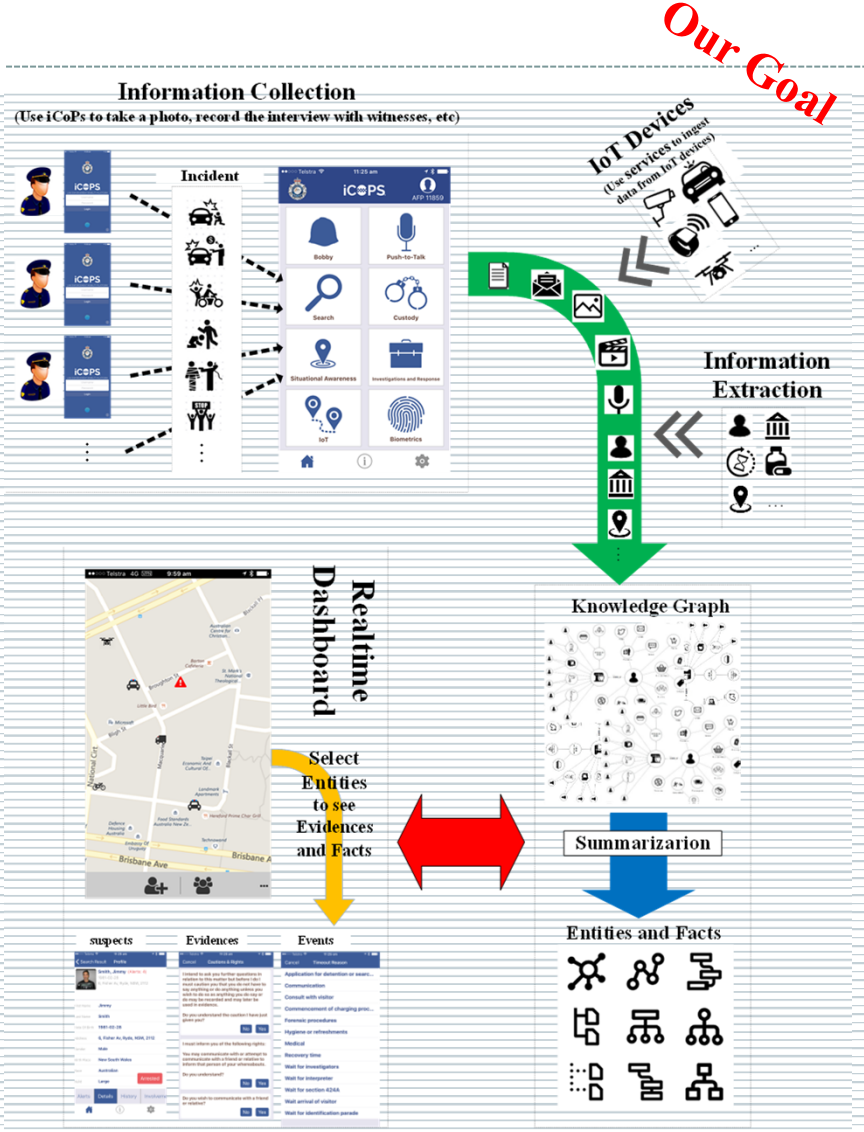
Missing Person !

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

https://missingpersons.gov.au/view-all-profiles

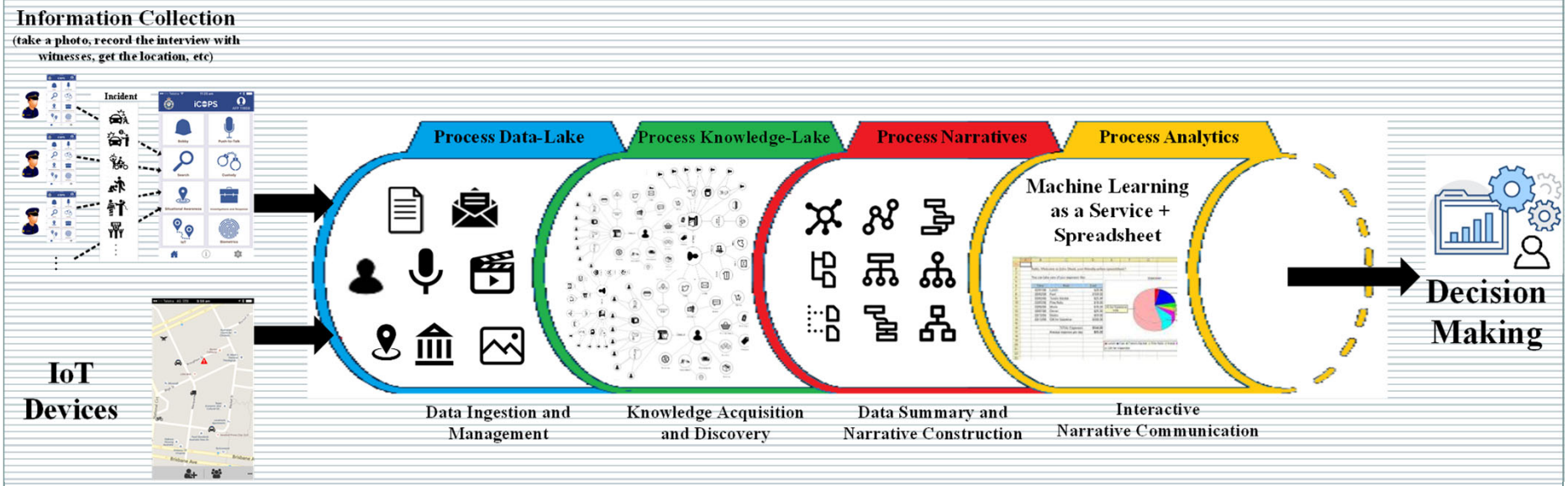
In USA, on any given day, there are as many as 100,000 active missing person's cases.

https://nij.gov



Solution: iProcess

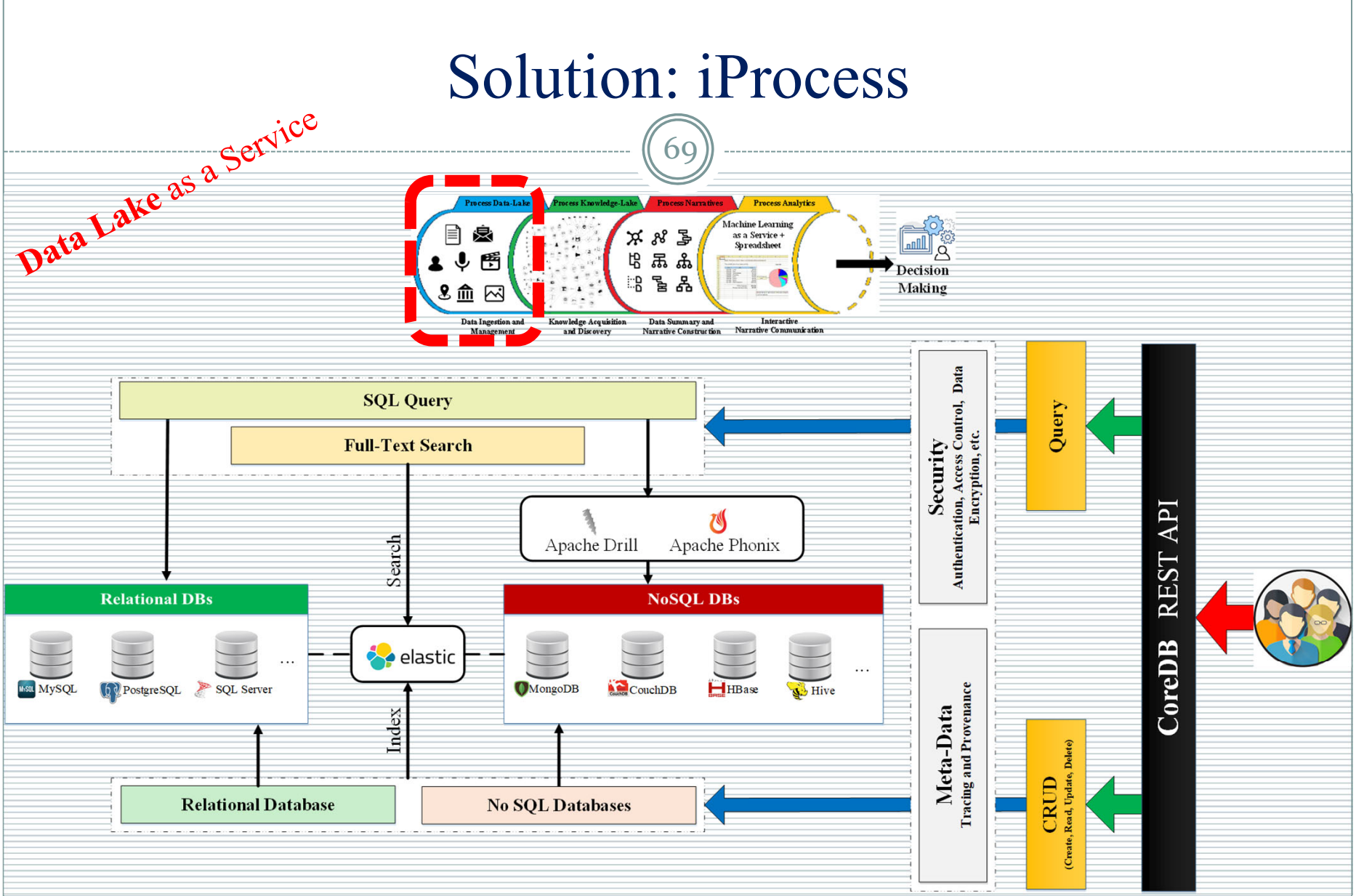
iProcess: is a scalable and extensible IoT-Enabled Process Data Analytics Pipeline to enable analysts ingest data from IoT devices, extract knowledge from this data and link them to process execution data.



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

Solution: iProcess

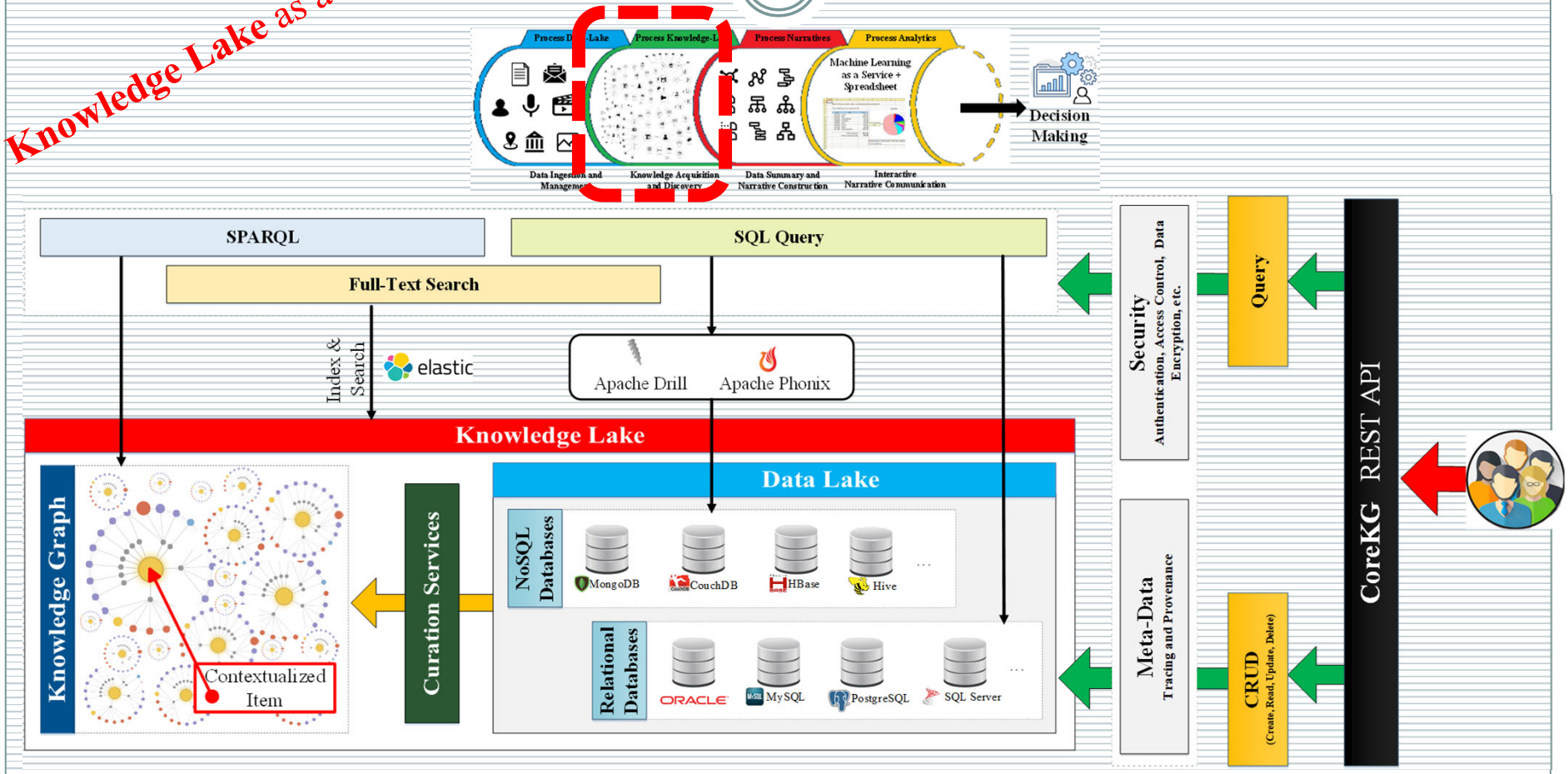
Data Lake as a Service



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

Solution: iProcess

Knowledge Lake as a Service

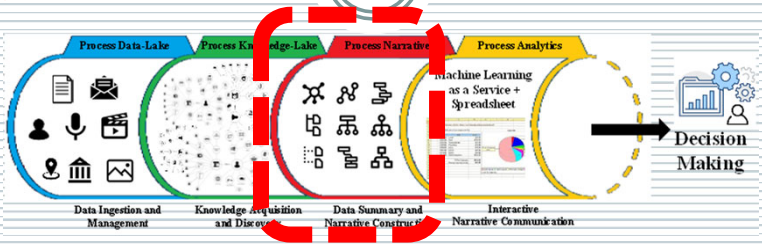


Beheshti, Benatallah, et al., **CoreKG: a Knowledge Lake Service (VLDB'18)** https://github.com/unsw-cse-soc/CoreKG

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

Solution: iProcess

Summarizing the process data and constructing process **narratives**



Process OLAP
Process Cubes
Dimensions
Cells
Measures
Operations

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.

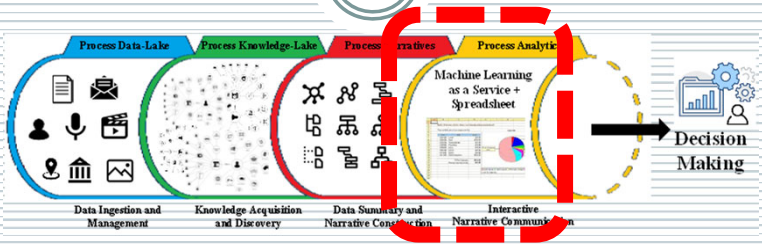
Solution:

- On-Line Analytical Processing on Graphs

Beheshti et al., "**Scalable Graph-based OLAP Analytics over Process Execution Data**", Distributed and Parallel Databases (DAPD) Journal, 34(3), 379-423, 2016

Solution: iProcess

ML as a Service

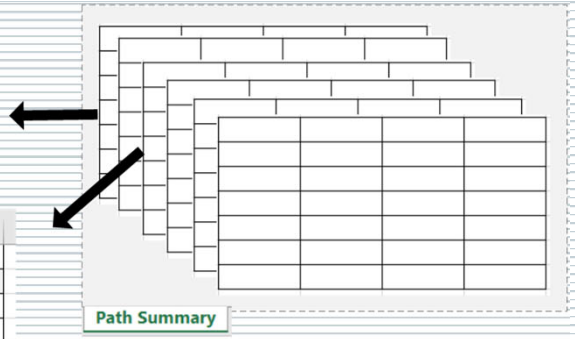


	A	B	C	D	E	F	G	H	I
2		Extracted-from	Data Islands						
3			Twitter	Facebook	eMail	Police Historical Data	news	CCTV1	...
4	D	person (named entity)	summary	summary	summary	summary	summary	summary	...
5	i	Location (named entity)	summary	summary	summary	summary	summary	summary	...
6	m	Organization (namec							
7	e	...							
8	n	Keyword (missing)							
9	s	keyword (Police)							
10	i	keyword (crime)							
11	o	...							
12	n	Phrase (reported mis							
13	s	...							

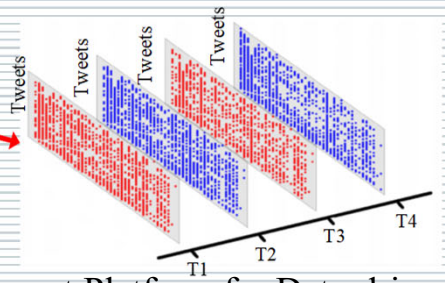
Extracted-from (Relationsh

	A	B	C	D	E	F	G	H	I
1		IMAGE	Data Islands						
2			Twitter	Facebook	eMail	Police Historical Data	CCTV1	CCTV2	...
3	D	Time (captured)	summary	summary	summary	summary	summary	summary	...
4	i	Locaion (continent)	summary	summary	summary	summary	summary	summary	...
5	m	Locaion (Country)	summary	summary	summary	summary	summary	summary	...
6	e	Locaion (City)	summary	summary	summary	summary	summary	summary	...
7	n	Locaion (Suburb)	summary	summary	summary	summary	summary	summary	...
8	s	IP Address (captured)	summary	summary	summary	summary	summary	summary	...
9	i	source (mobile, PC, ...)	summary	summary	summary	summary	summary	summary	...
10	o	size	summary	summary	summary	summary	summary	summary	...
11	n	owner	summary	summary	summary	summary	summary	summary	...
12	s	...	...	...	...	...	...	...	...

Image (Entity Sumamry)



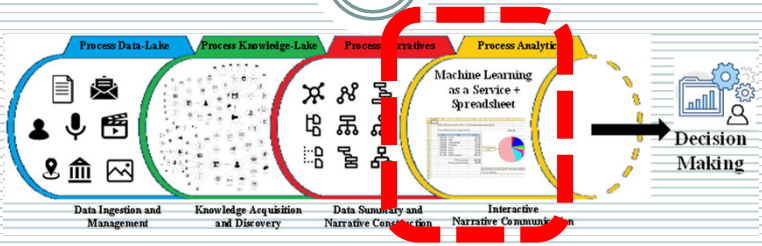
Path Summary

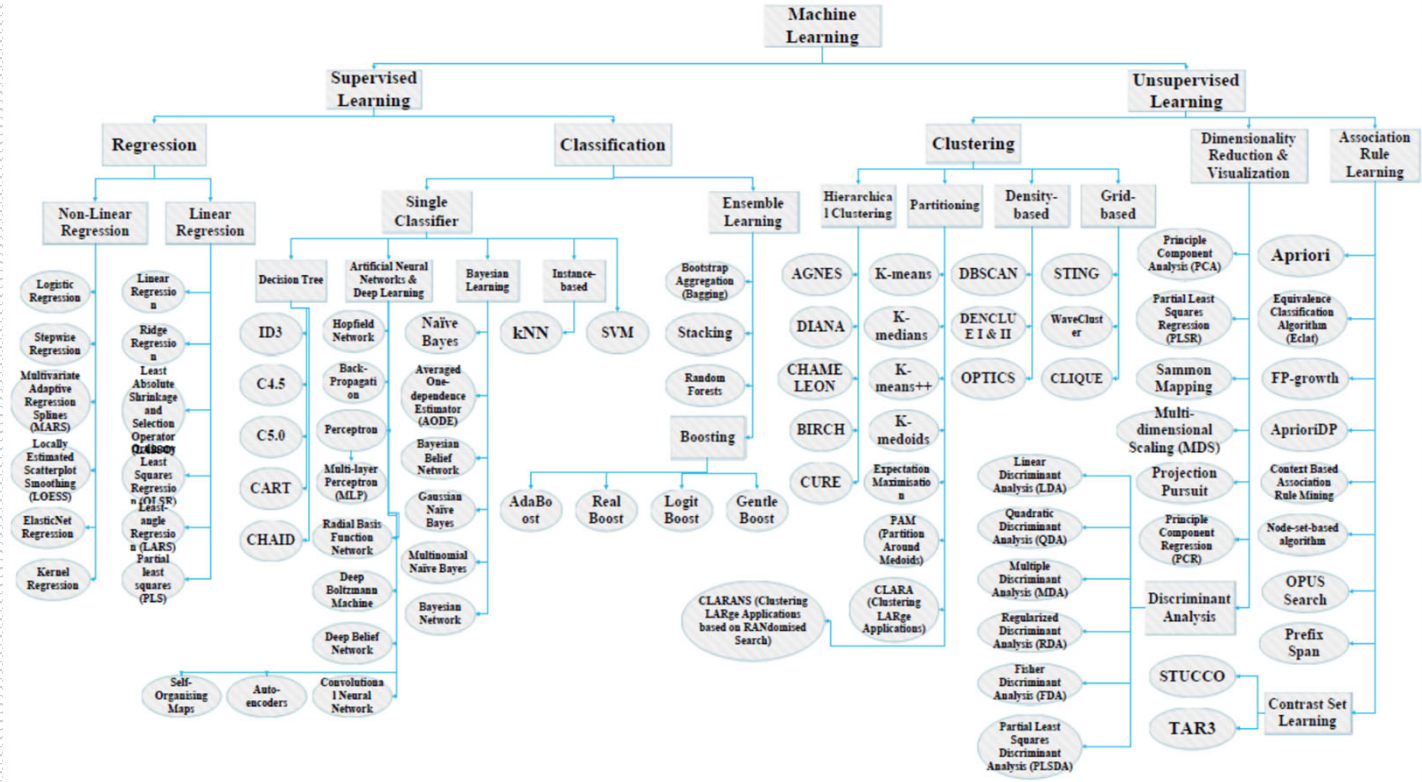


Amouzegar, Beheshti et al., " iSheets: A Spreadsheet-based Machine Learning Development Platform for Data-driven Process Analytics", ICSOC, 2018.

Solution: iProcess

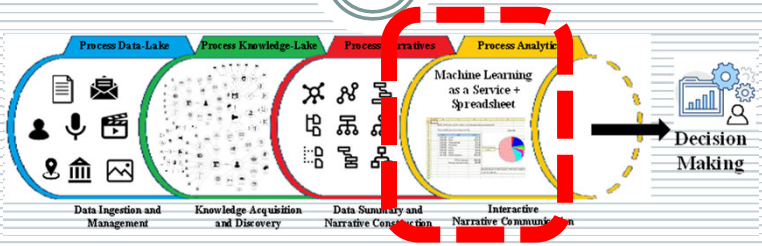
ML as a Service





Solution: iProcess

iSheets



localhost | MLAAS | localhost:8080/

Formula Bar:

Dataset

group("location","time") — Apply

	2016-04-01	2016-04-02	2016-04-03	2016-04-04	2016-04-05	2016-04-06	2016-04-07	2016-04-08	2016-04-09	2016-04-10
Brisbane	N/A	N/A	N/A	Tweets	Tweets	Tweets	N/A	N/A	N/A	N/A
Canberra	N/A	Tweets	N/A	N/A	N/A	N/A	N/A	Tweets	N/A	N/A
Melbourne	Tweets	N/A	N/A	Tweets	N/A	Tweets	N/A	N/A	N/A	Tweets
Perth	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Tweets	N/A
Sydney	Tweets	Tweets	Tweets	N/A	Tweets	Tweets	Tweets	Tweets	Tweets	N/A

classify.bytopic() — Apply

	Sport	Politics	Business	Education	Entertainment
Sydney-2016-04-05	N/A	N/A	Tweets	N/A	N/A
Sydney-2016-04-06	N/A	N/A	Tweets	N/A	N/A
Sydney-2016-04-07	N/A	N/A	Tweets	N/A	Tweets

classify.bySentiment() — Apply

	Positive	Negative
Sydney-2016-04-07-Se	Tweets	Tweets

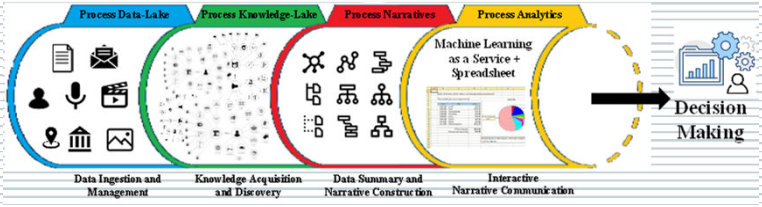
Twitter | Group

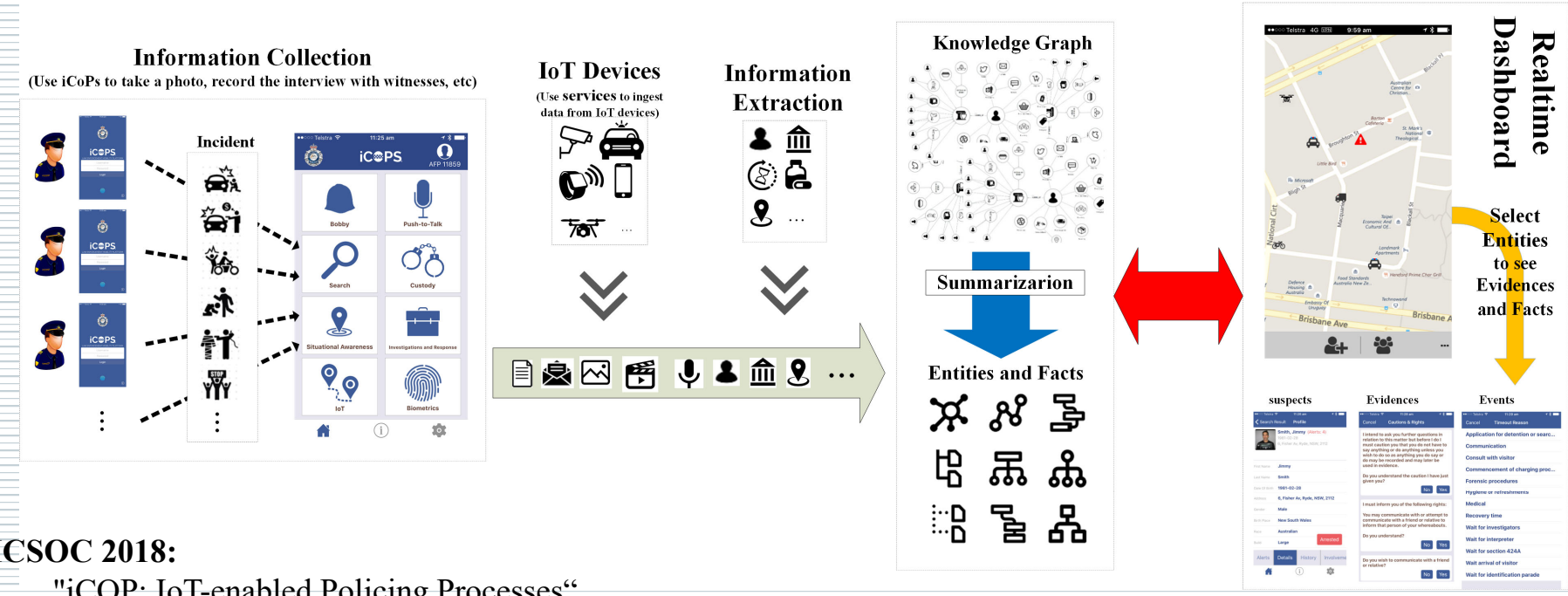
ML as a Service

- Classification
  - SVM
  - kNN
  - Logistic Regression
  - C5.0
  - MLP
  - Trained:Topic Classifier
  - Trained:Sentiment Classifier
- Clustering
- Association Learning
- Operations
  - Select
  - Group
  - Sort
  - Partition
  - Addition

Solution: iProcess

Motivating Scenario
Missing People!



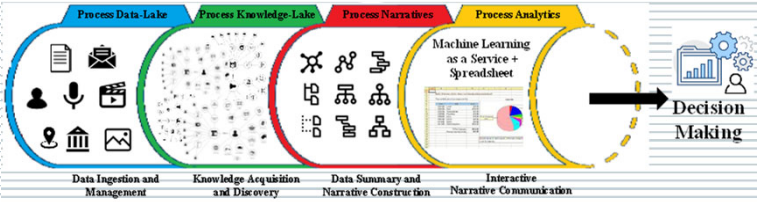


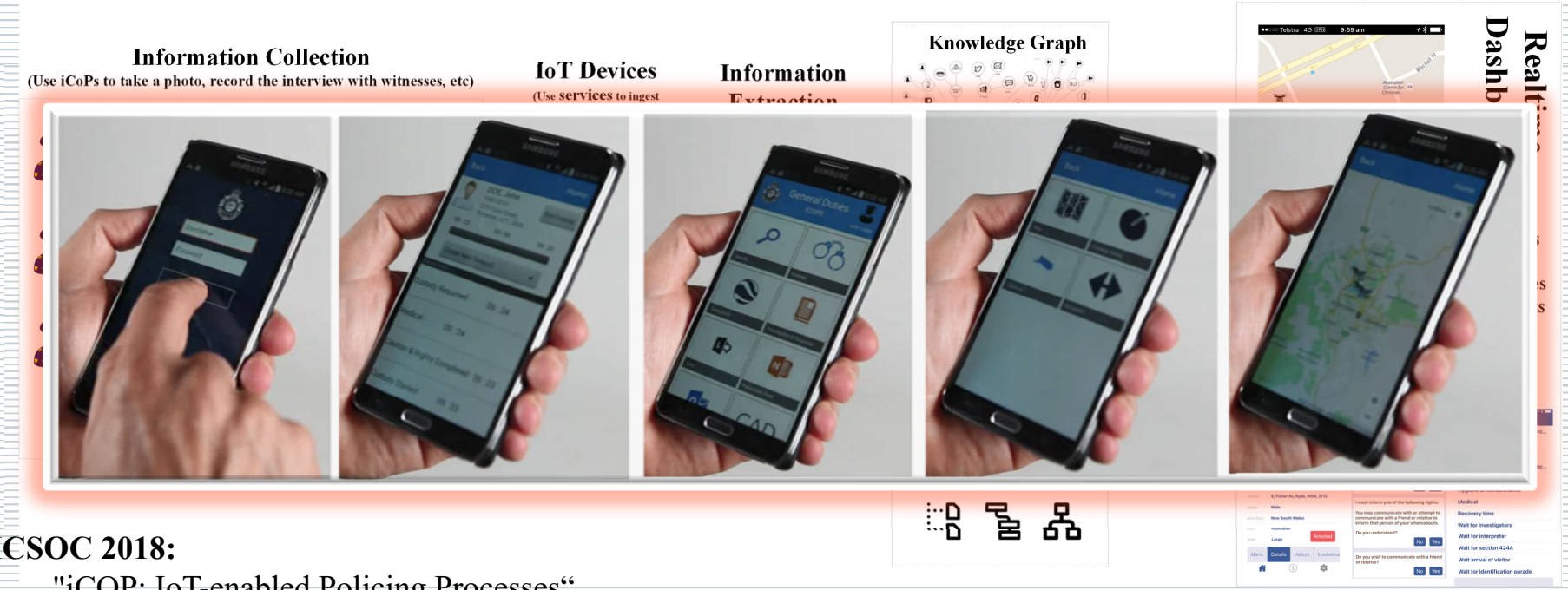
ICSOC 2018:

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

Solution: iProcess

Motivating Scenario
Missing People!

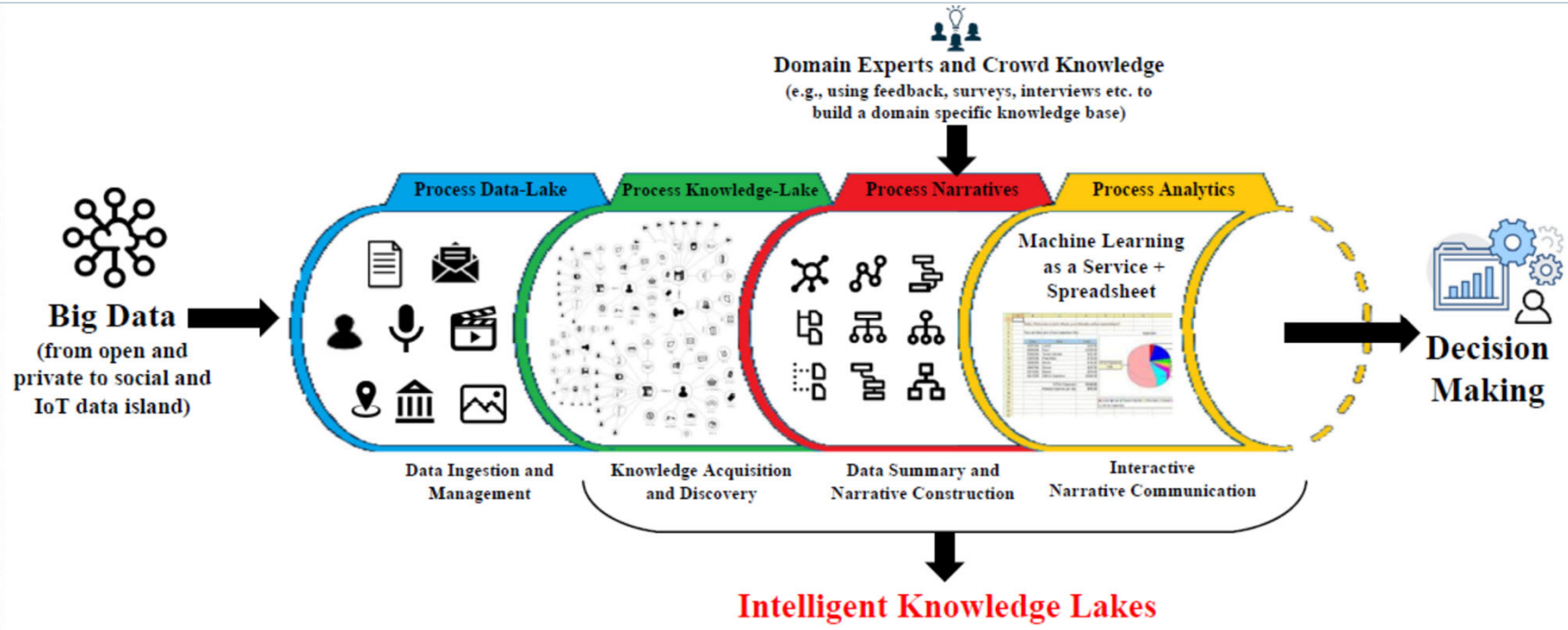




ICSOC 2018:

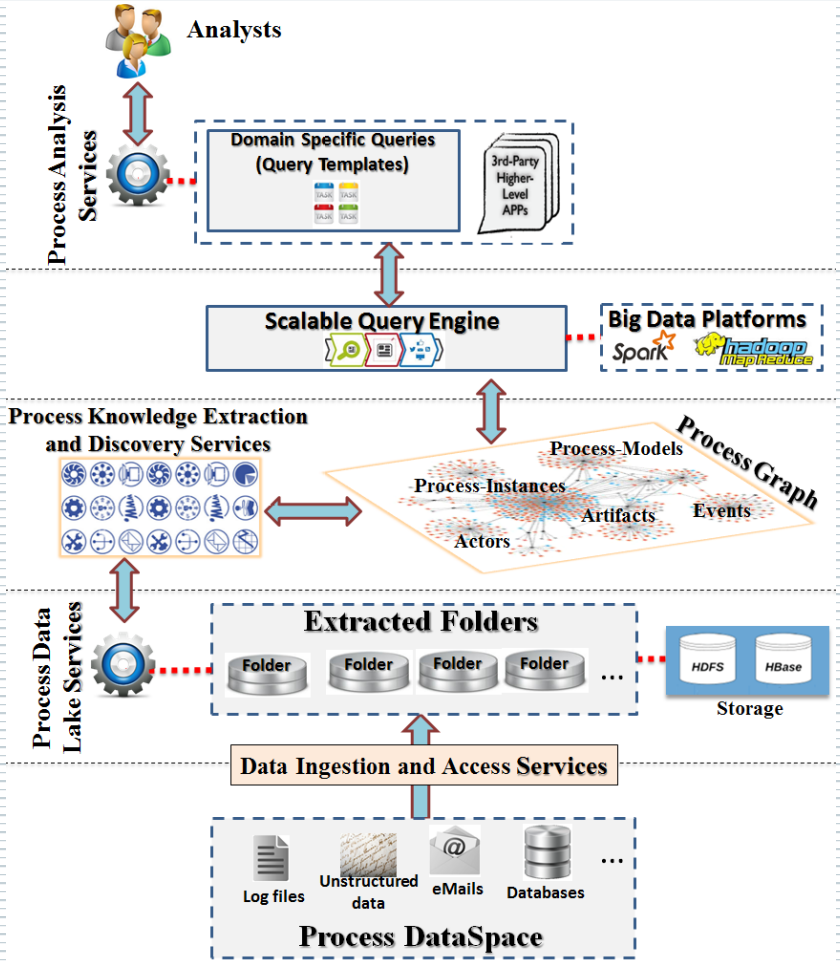
- "iCOP: IoT-enabled Policing Processes“
- "iSheets: A Spreadsheet-based Machine Learning Development Platform for Data-driven Process Analytics“.

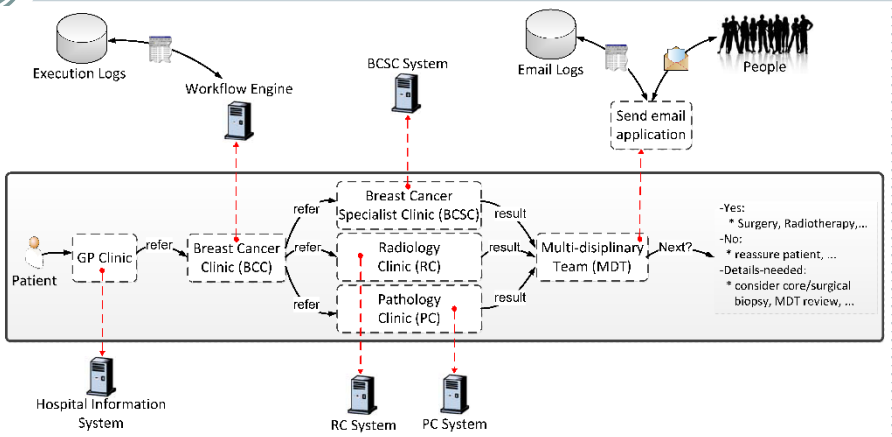
Intelligent Knowledge Lakes !

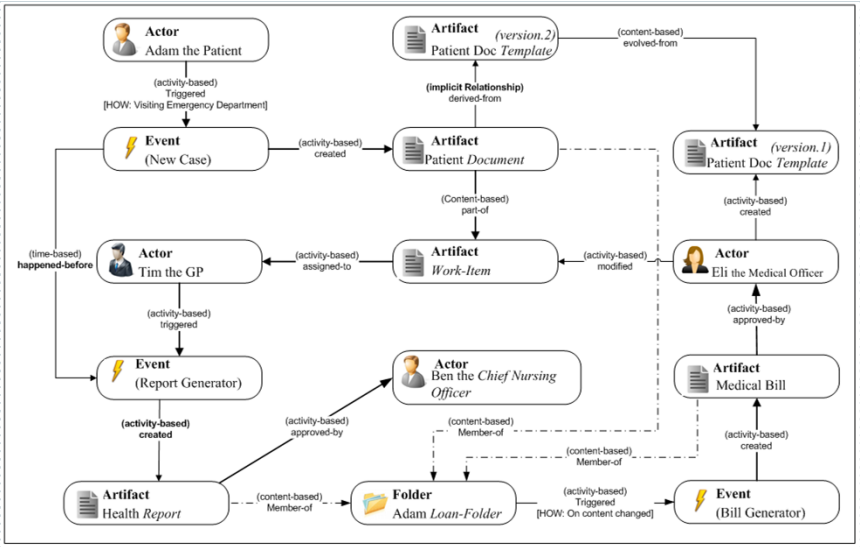


Other Scenarios

Health

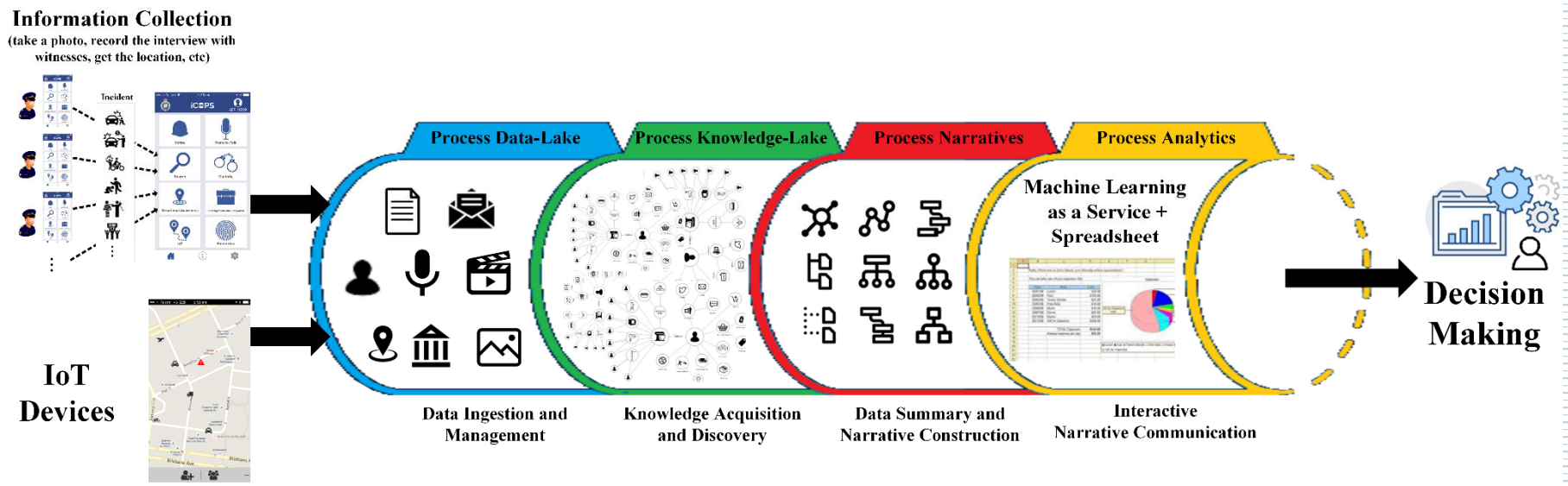


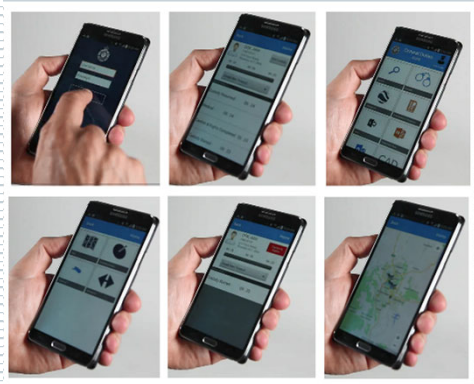


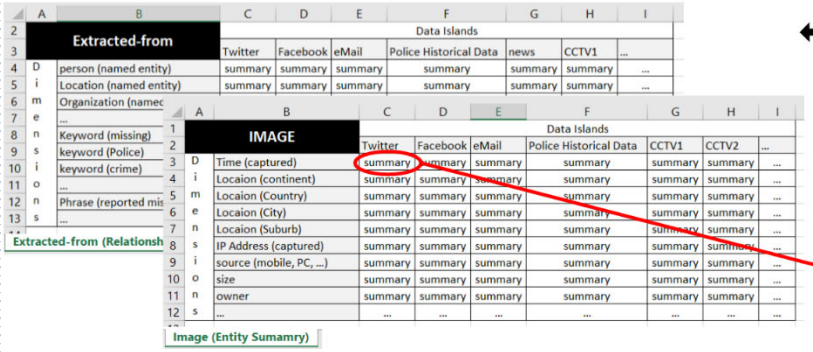


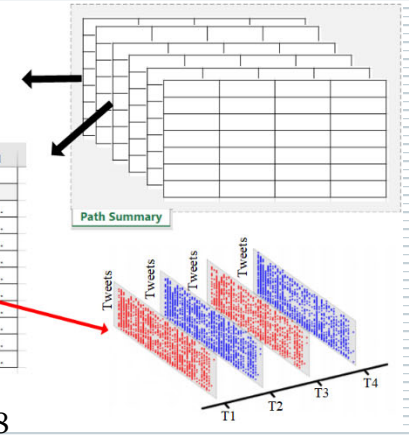
Beheshti et al., "**Galaxy: A Platform for Explorative Analysis of Open Data Sources**", EDBT, 2016.

IoT





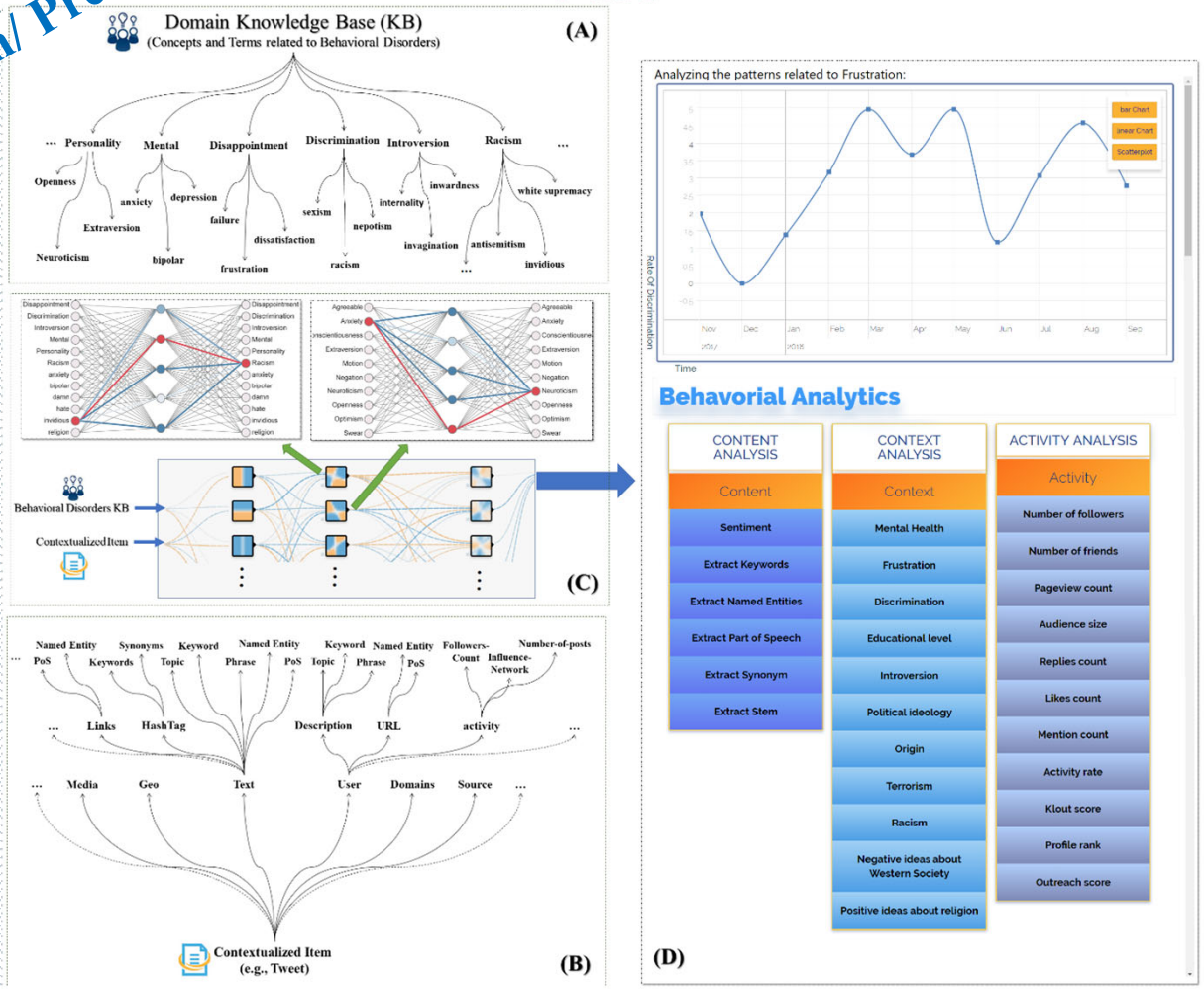




Beheshti et al., "iProcess: Enabling IoT Platforms in Data-Driven Knowledge-Intensive Processes", BPM, 2018

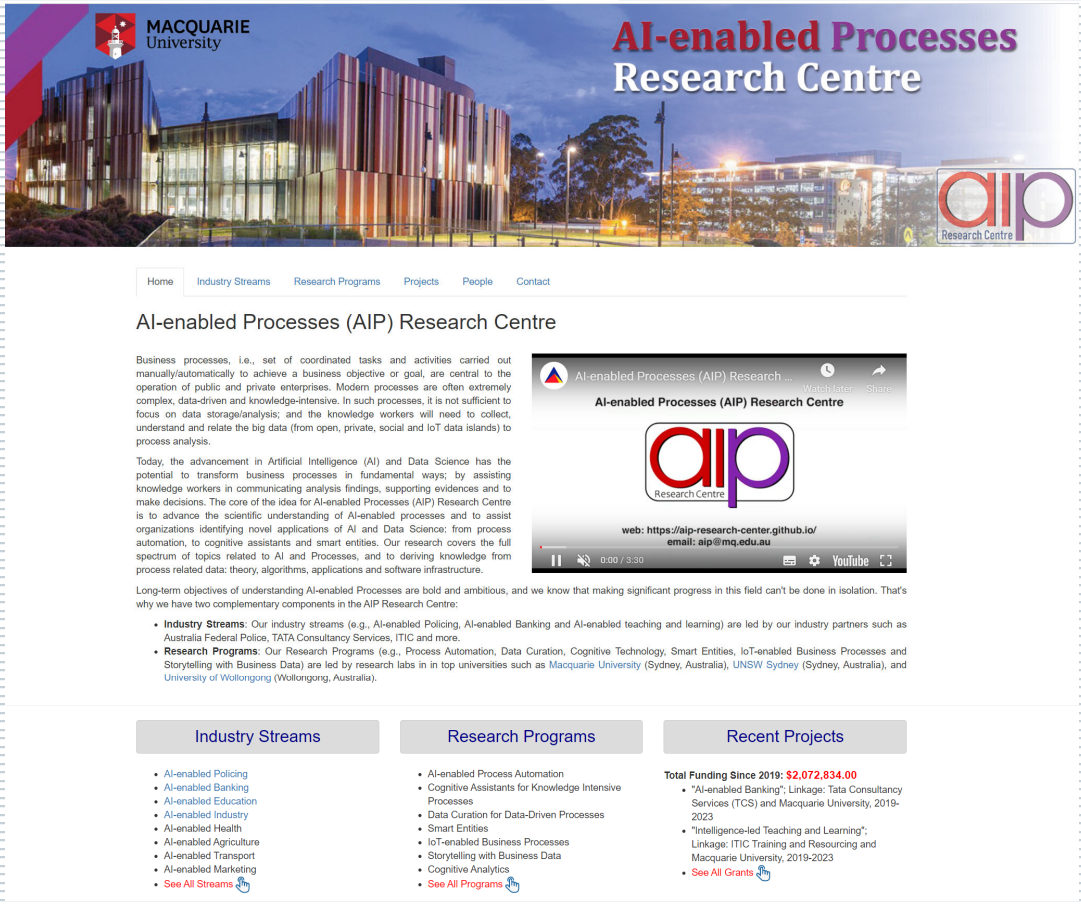
AI-Enabled Policing (Applications)

Crime Detection/ Prevention



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*)

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.

Thank You !