









University of  
South Australia

## Big Data search and mining

1

### The six Vs of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: *volume, variety and velocity*. Over time, other Vs have been added to descriptions of big data:

| VOLUME  | VARIETY   | VELOCITY  | VERACITY  | VALUE  | VARIABILITY   |
|---|---|---|---|--|---|
| The amount of data from myriad sources.   | The types of data: structured, semi-structured, unstructured.                       | The speed at which big data is generated.   | The degree to which big data can be trusted.  | The business value of the data collected.  | The ways in which the big data can be used and formatted.                             |
|  |  |  |  |  |  |

ICONS: ALEXANDRO/ADORE STOCK

©2018 TECHTARGET. ALL RIGHTS RESERVED. TechTarget

2

## Big Data mining



- Exploring and analysing large amounts of data to discover and extract **patterns** in Big Data.
- Methods at the intersection of **machine learning**, **statistics** and **database systems**.

## Common tasks in data mining

- Anomaly detection
- Association rule mining
- **Clustering**
- Classification
- Regression
- Summarisation



## Challenges with Big Data mining

- Search for the relevant data first, then apply appropriate techniques to find patterns.
- Distributed storage means new kinds of **index structures** and associated **search technologies**.
  - A high indexing rate initially but also during additions, deletions and updates.



## Search engines for Big Data (open source)

- Scalable, high-performance indexing for performing information retrieval functions on Big Data.



## Apache Solr

---

- Search engine server that uses Lucene.
- Applications communicate with Solr using XML and HTTP to index documents or execute searches.
- Supports a rich schema specification that allows for a wide range of flexibility in dealing with different document fields.
- Also has an extensive search plugin API for developing custom search behaviour.

## elasticsearch

---

- Elasticsearch is built on top of Lucene.
- Flexible and powerful distributed RESTful search engine and analytics engine for the cloud.
- No need for upfront schema definition.
- API driven, almost any action can be performed using a simple RESTful API using JSON over HTTP.

## Speaking of search engines...

*'Google is actually a mountain of data and a set of [Big Data] tools for working with it.'* – Bernard Marr

Deep learning  
Semantic indexing  
Natural language processing

### WARNING

This material has been reproduced and communicated to you by or on behalf of the **University of South Australia** in accordance with section 113P of the *Copyright Act 1968 (Act)*.

The material in this communication may be subject to copyright under the Act. Any further reproduction or communication of this material by you may be the subject of copyright protection under the Act.

**Do not remove this notice**