Big Data and Open Data

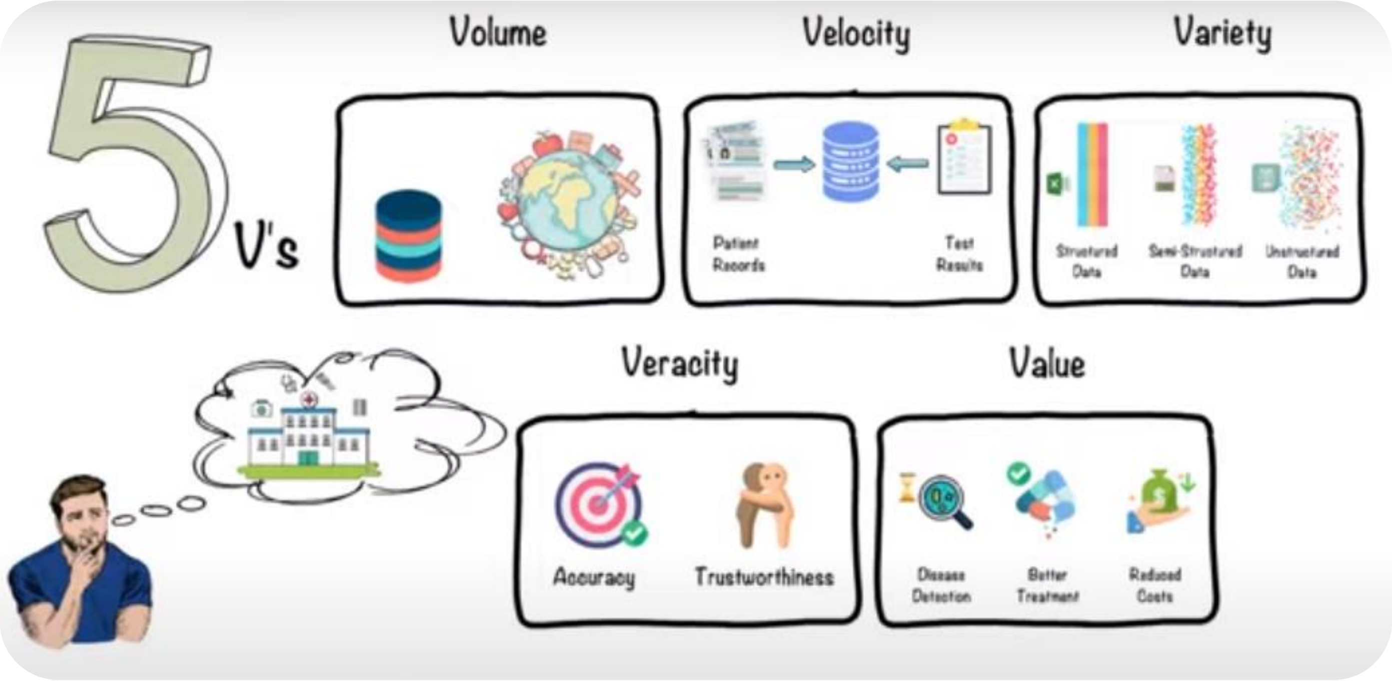
CS5056 - Data Analytics

FRANCISCO J. CANTÚ, HÉCTOR G. CEBALLOS
TECNOLÓGICO DE MONTERREY

MAY 12, 2021

Introduction to Big Data

Big Data
The 5 V's definition



https://youtu.be/bAyrObl7TYE

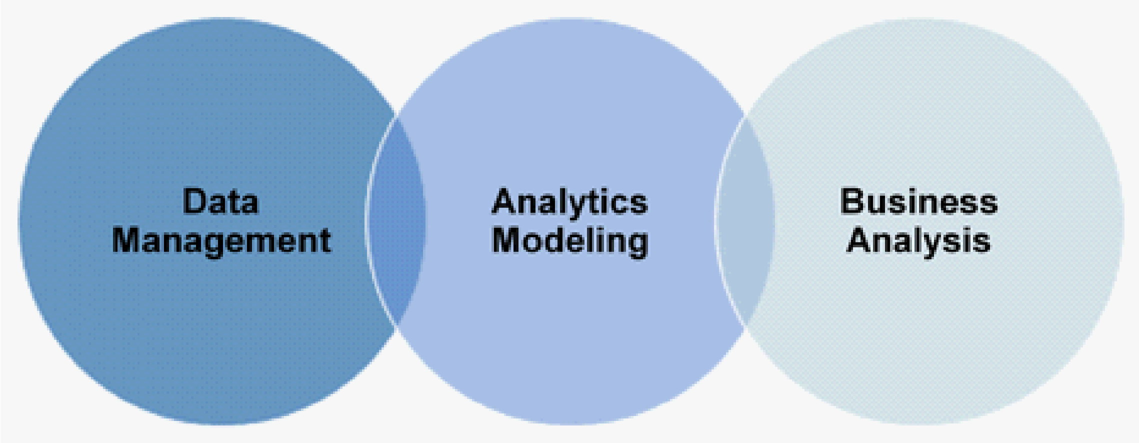
Big Data
Applications

Intro to Big Data: Crash Course Statistics #38

https://youtu.be/vku2Bw7Vkfs

Data Scientist

An individual responsible for modeling complex business problems, discovering business insights and identifying opportunities through the use of statistical, algorithmic, mining and visualization techniques. In addition to advanced analytic skills, this individual is also proficient at **integrating and preparing large, varied datasets, architecting specialized database and computing environments**, and communicating results. A data scientist may or may not have specialized industry knowledge to aid in modeling business problems and with **understanding and preparing data.**



Understand Data

- Integrate
- Manipulate
- QA
- Prep.

Know Analytics

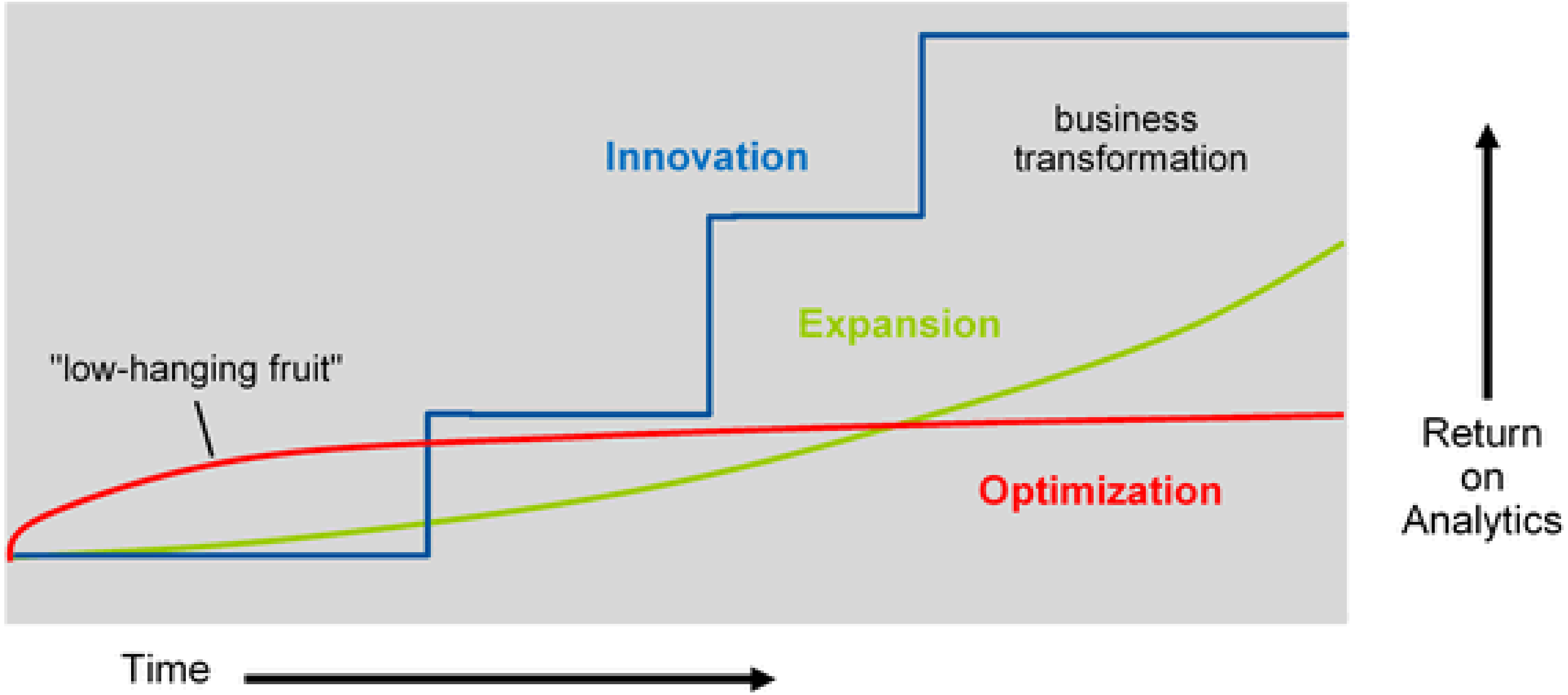
- Appropriate techniques
- Interpret data and diagnose models
- Meet business requirements

Focus on the Business

- Goals
- Constraints
- Decisions
- Communication of results

Gartner (2013). Emerging Role of the Data Scientist and the Art of Data Science

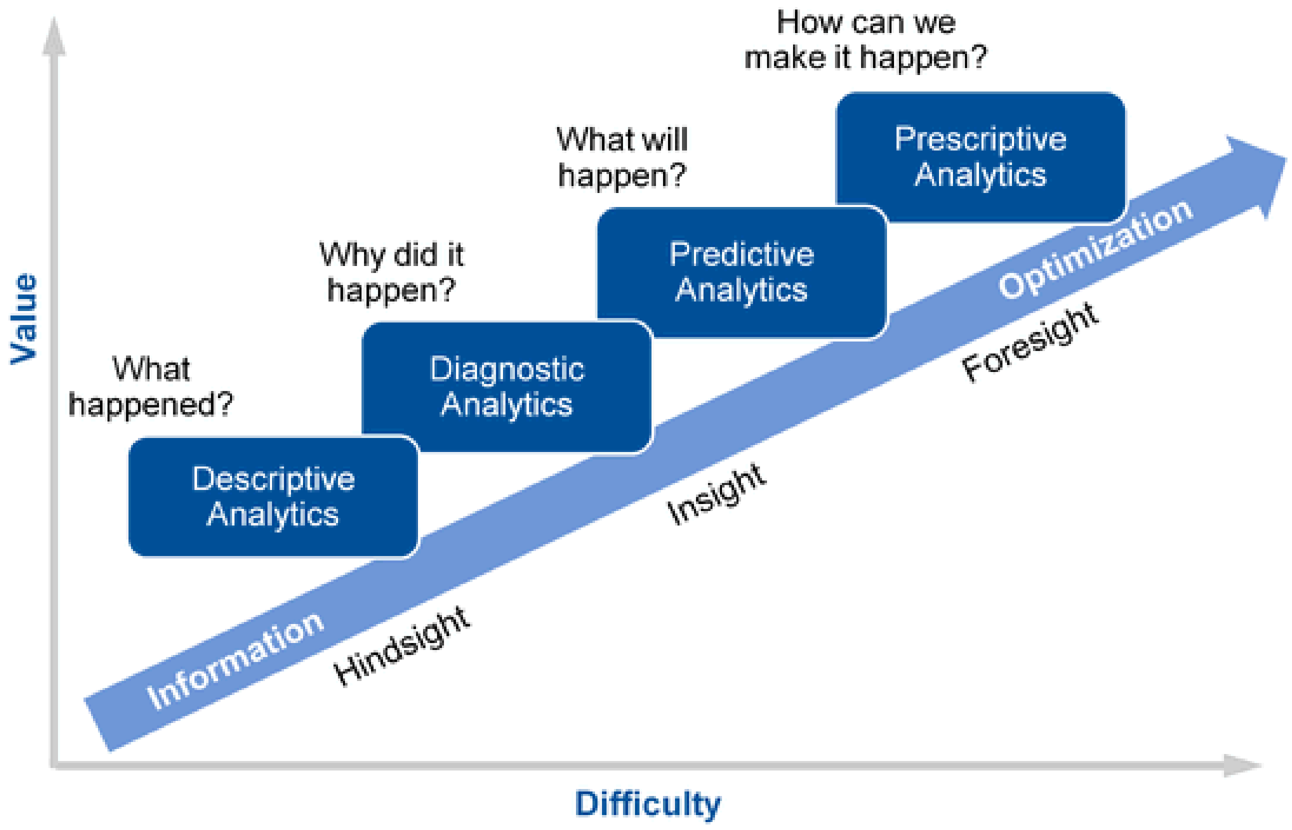
Relative return on Analytics



Gartner (2013). Emerging Role of the Data Scientist and the Art of Data Science

Gartner Analytic Ascendancy Model

Which kind of model requires each type?



Gartner (2014). Big Data Strategy Components: IT Essentials

Big Data Keys

"Big data" is not about size alone. This year's big data is next year's normal-sized data. Generally, volume quickly gives way to the more defining requirements of **variety, velocity and complexity**.

Business benefits are frequently higher when addressing the **variety of the data** as opposed to addressing volume.

Specific **low-cost solutions** for processing approaches are permitting many organizations to **deploy** big data at a much faster rate than previous technological advances.

Many business organizations continue to struggle with applying big data processing and datasets to **business outcomes**.

Gartner (2013). The Importance of 'Big Data': A Definition

No-SQL data storage

Wide-column: it uses tables, rows, and columns, but unlike a relational database, the names and format of the columns can vary from row to row in the same table.

- E.g. Apache Accumulo, Amazon DynamoDB

Document-oriented: storing, retrieving and managing document-oriented information. XML databases are a subclass.

- E.g. Apache CouchDB, ArangoDB, BaseX, IBM Domino, MarkLogic, MongoDB.

Key-value: storing, retrieving, and managing *associative arrays*, and dictionaries or hash tables.

- E.g. Apache Ignite, ArangoDB, Berkeley DB, MemcacheDB, MUMPS, Oracle NoSQL Database.

Graph: uses graph structures for semantic queries with nodes, edges, and properties to represent and store data.

- E.g. AllegroGraph, ArangoDB, InfiniteGraph, Apache Giraph, MarkLogic, Neo4J, OrientDB, Virtuoso

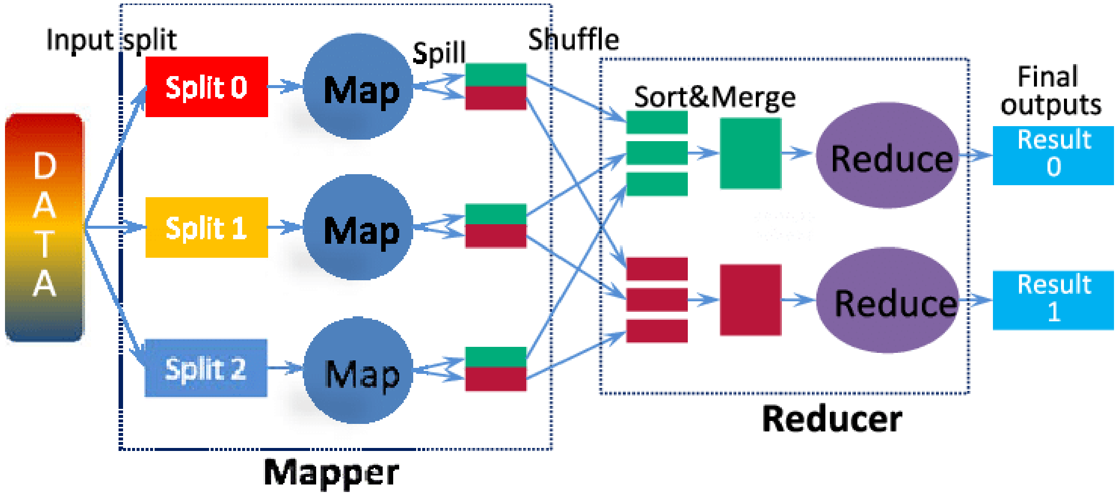
Data Warehouse vs Data Lake

Characteristics	Data Warehouse	Data Lake
Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
Schema	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price / Performance	Fastest query results using higher cost storage	Query results getting faster using low-cost storage
Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (ie. raw data)
Users	Business analysts	Data scientists, Data developers, and Business analysts (using curated data)
Analytics	Batch reporting, BI and visualizations	Machine Learning, Predictive analytics, data discovery and profiling

https://aws.amazon.com/es/big-data/datalakes-and-analytics/what-is-a-data-lake/?nc=sn&loc=2

Massive processing

The MapReduce programming model



Fuente: https://www.researchgate.net/publication/305716828_In-Storage_Computing_for_Hadoop_MapReduce_Framework_Challenges_and_Possibilities/figures?lo=1

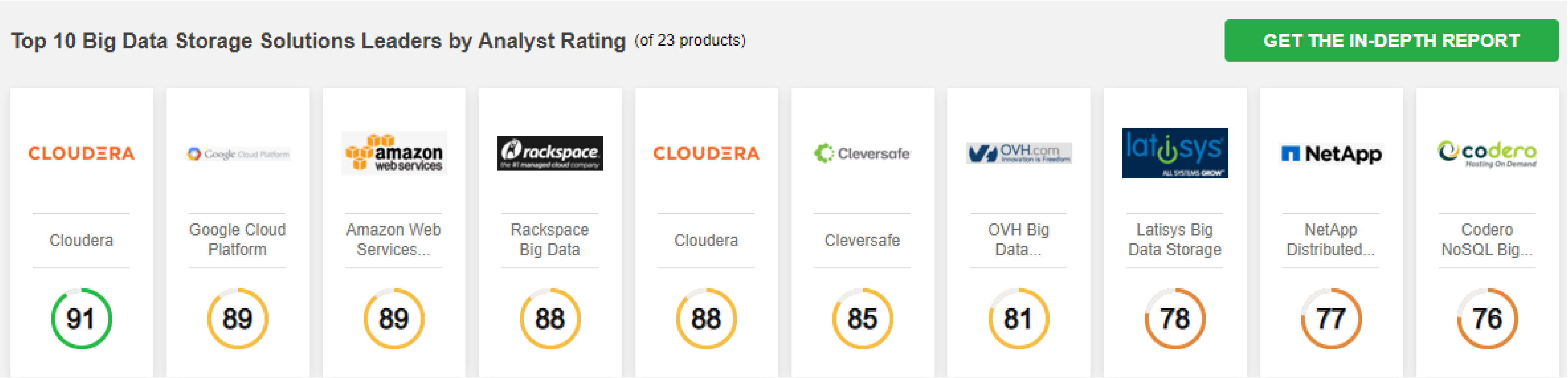
Apache Hadoop: Processing of big data using the MapReduce programming model.

Apache Sparkl: it adds the ability to set up many operations (not just map followed by reducing).

https://spark.apache.org/docs/2.2.0/ml-classification-regression.html

Big Data Storage Solutions

Current Cloud solutions provide: data integration, access to IoT devices, automation of machine learning and artificial intelligence processes, analysis of data streaming, and business intelligence tools.



Top 10 Big Data Storage Solutions Leaders by Analyst Rating (of 23 products)

Cloudera	Google Cloud Platform	Amazon Web Services...	Rackspace Big Data	Cloudera	Cleversafe	OVH Big Data...	Latisys Big Data Storage	NetApp Distributed...	Codero NoSQL Big...
91	89	89	88	88	85	81	78	77	76

https://www.selecthub.com/big-data-storage-software/

Big Data and CRISP-DM

What is Deployment?

- How do you expect to provide a business solution?
- Will it be cost-effective?
- How much information will you need for it?
- Are you dealing with a Big Data scenario?

Activity: Big Data Scenario

1. Teams of 3 persons will be formed automatically in Zoom.
2. You have 15 minutes to describe a Big Data scenario using the 5 Vs.
3. Send your scenario in a Word or Powerpoint document (ceballos@tec.mx)
4. Return to the main room in Zoom.
5. Two teams will present their scenario in 3 minutes.

Activity: Big Data Scenario

1. Elaborate an example of a Big Data scenario and describe its five characteristics

Scenario: ________________________________

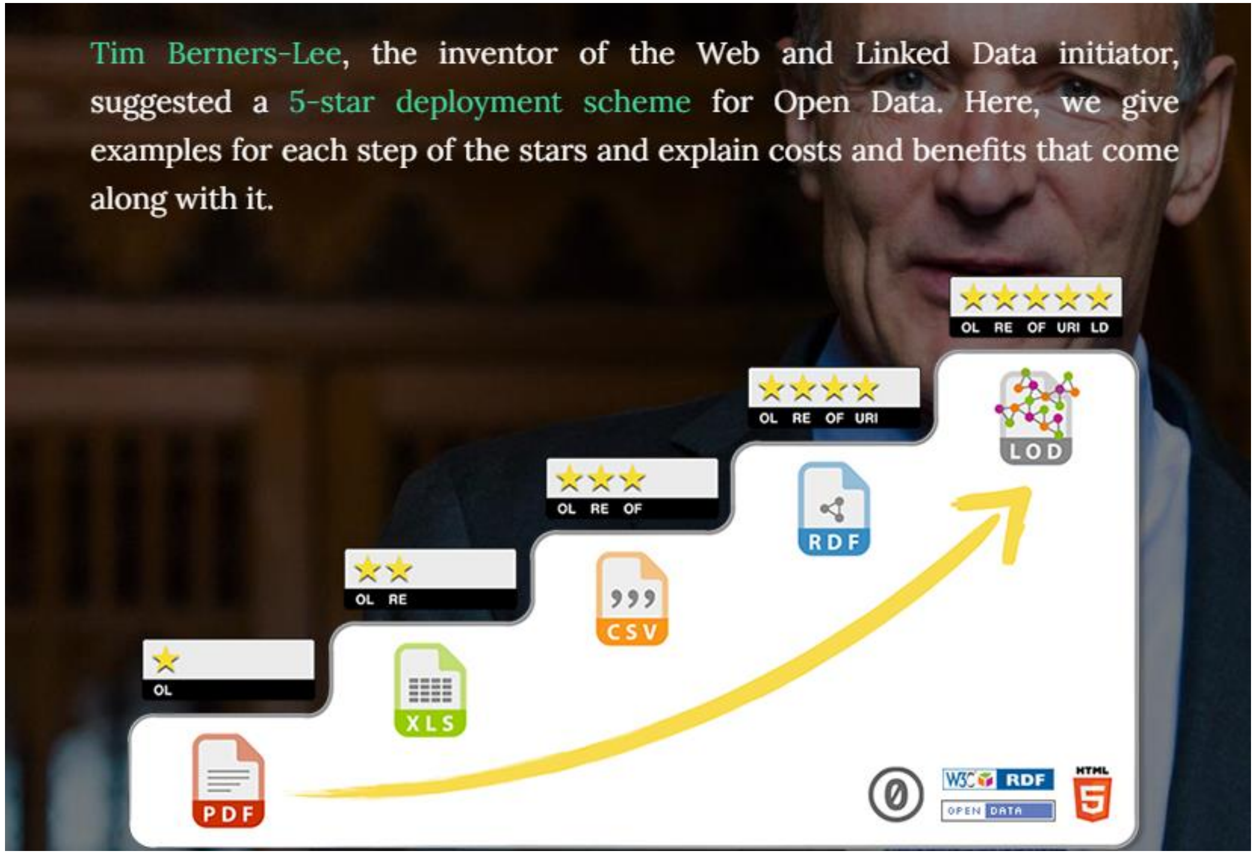
Characteristic	Description
Volume	How much data is generated every year?
Velocity	How fast is it generated?
Variety	How is this data structured? Metadata? How is it stored?
Veracity	How trustworthy is the information generated? Is it noise? How it can be cleaned?
Value	Stakeholders? Business questions? Potential earnings/savings?

https://en.wikipedia.org/wiki/Big_data#Applications

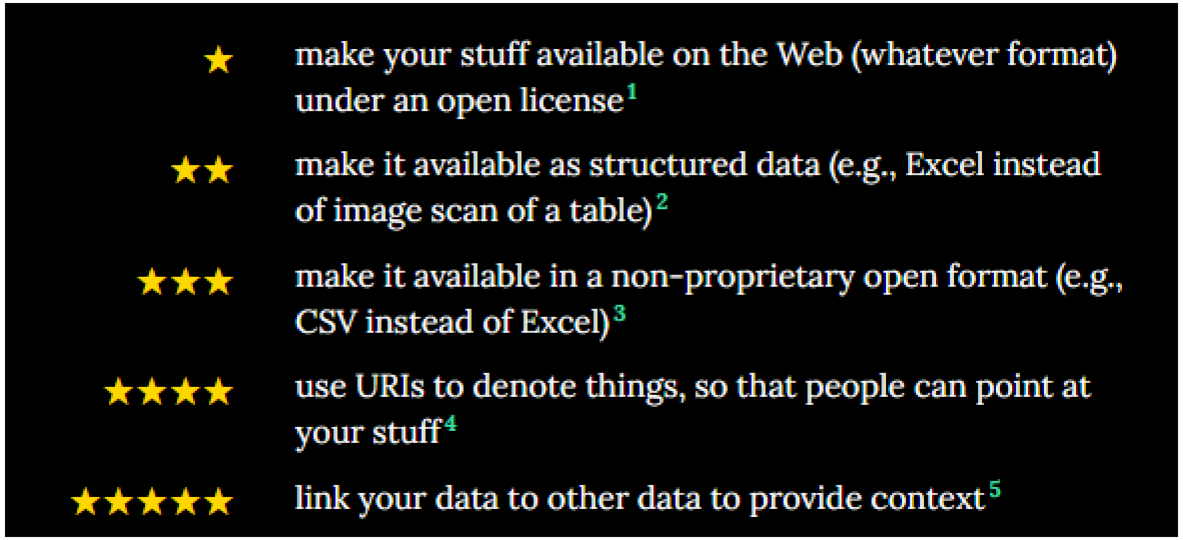
Teams and topics

Open Data

5 stars Open Data



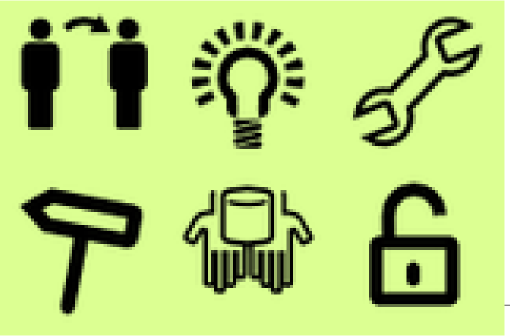
Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a 5-star deployment scheme for Open Data. Here, we give examples for each step of the stars and explain costs and benefits that come along with it.



★	make your stuff available on the Web (whatever format) under an open license[1]
★★	make it available as structured data (e.g., Excel instead of image scan of a table)[2]
★★★	make it available in a non-proprietary open format (e.g., CSV instead of Excel)[3]
★★★★	use URIs to denote things, so that people can point at your stuff[4]
★★★★★	link your data to other data to provide context[5]

https://5stardata.info/en/

Open Data License



“You are free: To **Share'**, ***To*** *Create,*
To *Adapt* as long as you:
Attribute, **Share-Alike**, **Keep open**”

Open Data Commons Open Database License (ODbL)

- Open Commons
- https://opendatacommons.org/licenses/odbl/

Licenses

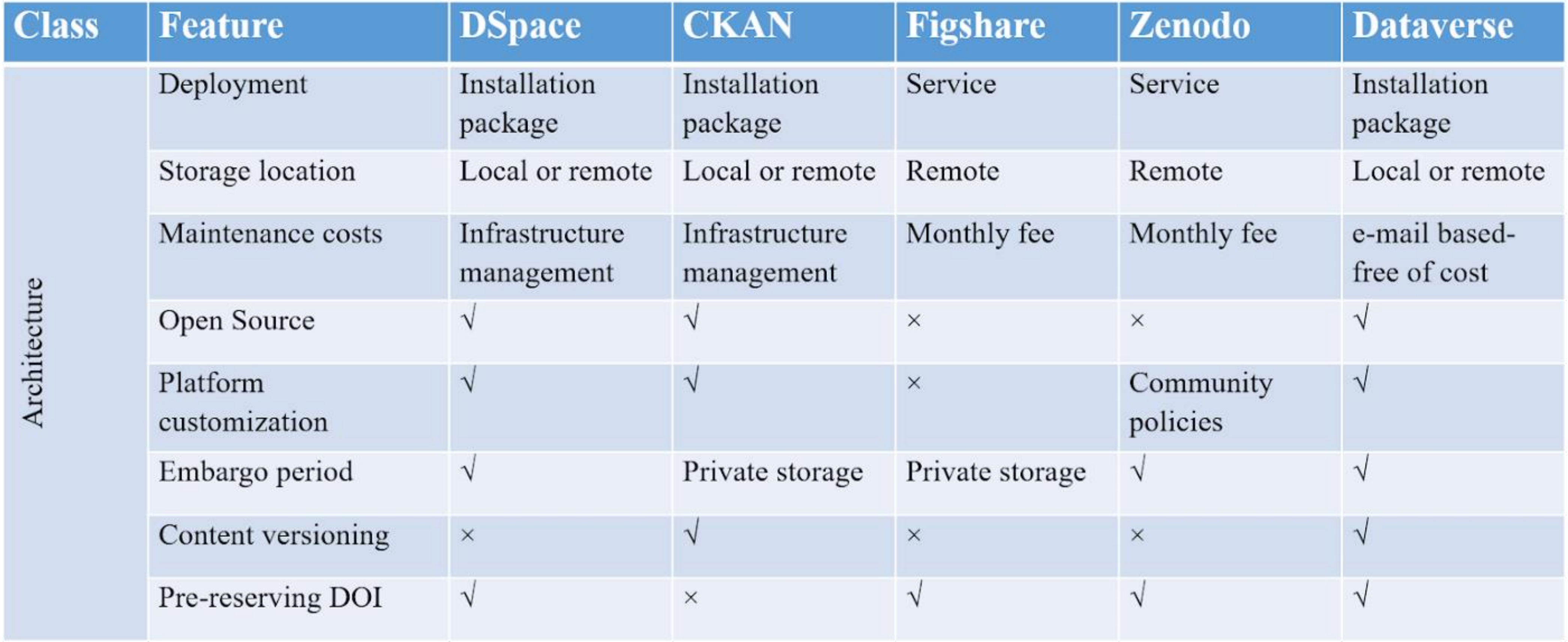
Open Data Commons Open Database License (ODbL) — “Attribution Share-Alike for data/databases”

Open Data Commons Attribution License — “Attribution for data/databases”

Open Data Commons Public Domain Dedication and License (PDDL) — “Public Domain for data/databases”

Legend: This {DATA(BASE)-NAME} is made available under the Open Database License: http://opendatacommons.org/licenses/odbl/1.0/. Any rights in individual contents of the database are licensed under the Database Contents License: http://opendatacommons.org/licenses/dbcl/1.0/

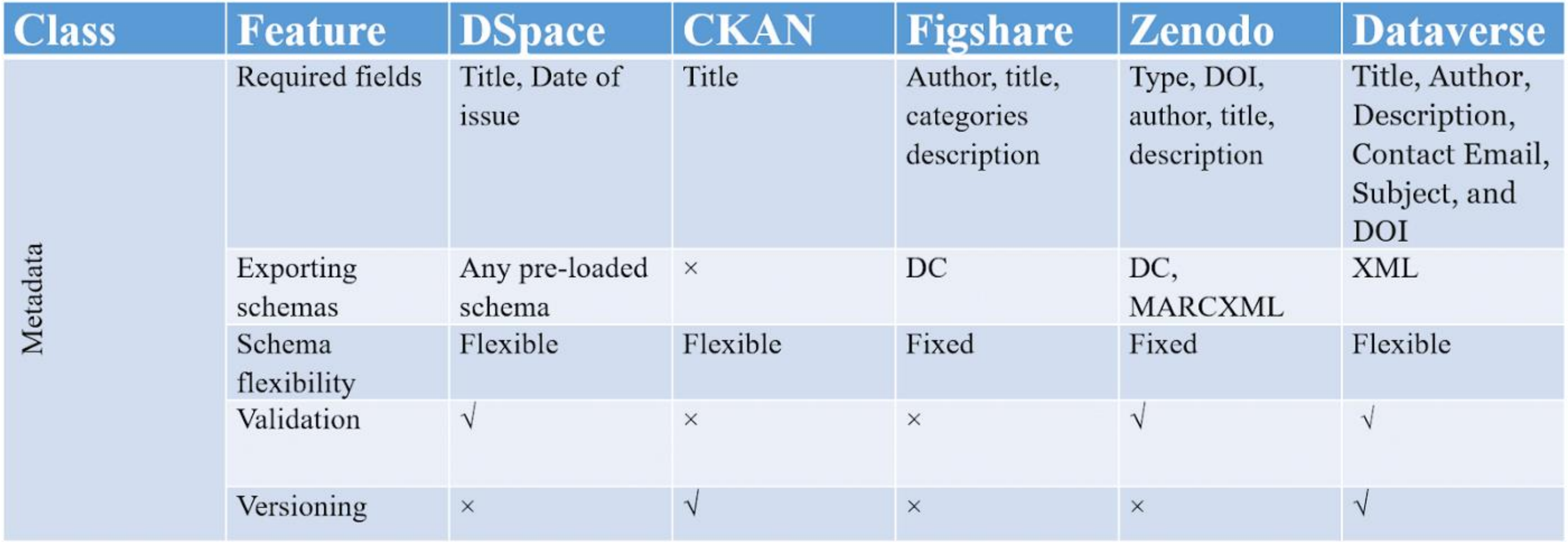
Research Data Management Platforms



Class	Feature	DSpace	CKAN	Figshare	Zenodo	Dataverse
Architecture	Deployment	Installation package	Installation package	Service	Service	Installation package
	Storage location	Local or remote	Local or remote	Remote	Remote	Local or remote
	Maintenance costs	Infrastructure management	Infrastructure management	Monthly fee	Monthly fee	e-mail based-free of cost
	Open Source	√	√	×	×	√
	Platform customization	√	√	×	Community policies	√
	Embargo period	√	Private storage	Private storage	√	√
	Content versioning	×	√	×	×	√
	Pre-reserving DOI	√	×	√	√	√

https://medium.com/analytics-vidhya/comprehensive-study-of-open-data-platforms-a63d702ef0d5

Research Data Management Platforms



Class	Feature	DSpace	CKAN	Figshare	Zenodo	Dataverse
Metadata	Required fields	Title, Date of issue	Title	Author, title, categories description	Type, DOI, author, title, description	Title, Author, Description, Contact Email, Subject, and DOI
	Exporting schemas	Any pre-loaded schema	×	DC	DC, MARCXML	XML
	Schema flexibility	Flexible	Flexible	Fixed	Fixed	Flexible
	Validation	√	×	×	√	√
	Versioning	×	√	×	×	√

https://medium.com/analytics-vidhya/comprehensive-study-of-open-data-platforms-a63d702ef0d5

Open Data Repositories

Data World (742): https://data.world/datasets/open-data

Google Data Set Search: https://datasetsearch.research.google.com/

Kaggle: https://www.kaggle.com/

R datasets: https://www.reddit.com/r/datasets/

UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php

United States Government: https://www.data.gov/

Mexico Government: https://datos.gob.mx/

- INEGI: http://en.www.inegi.org.mx/servicios/datosabiertos.html