Introduction to Data Science

What is Data Science?

Data science is the study of the computational principles, methods, and systems for extracting knowledge from data.

Data science is the application of the computational principles, methods, and systems in order to extract knowledge from data.

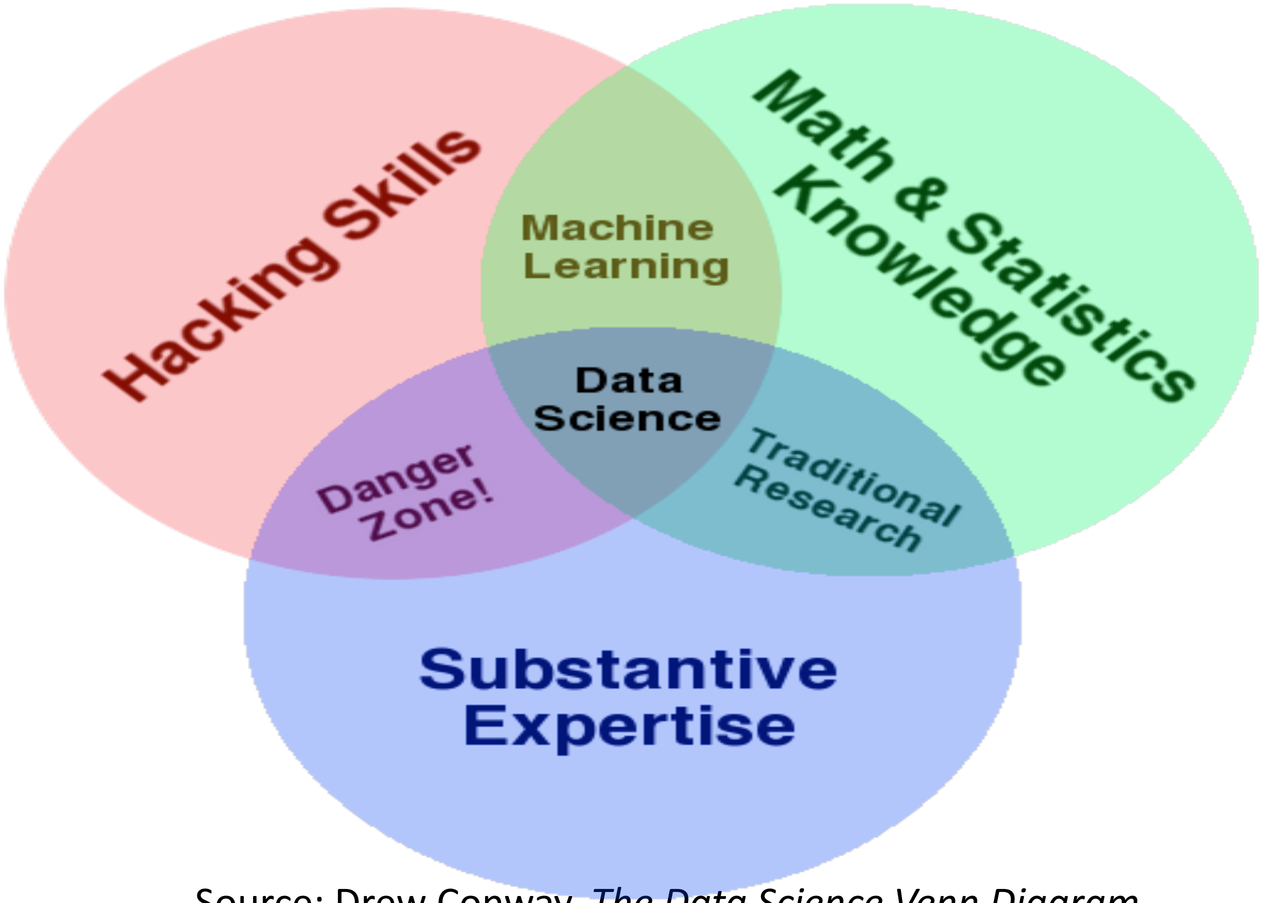
What is Data Science?

“Data Science is the extraction of knowledge from large volumes of data that are structured or unstructured, which is a continuation of the field data mining and predictive analytics, also known as knowledge discovery and data mining (KDD).”

“Data Science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, … which is a continuation of some of the data analysis fields such as statistics, data mining, and predictive analytics, similar to Knowledge Discovery in Databases (KDD).”

-- Wikipedia

Data Science as an Intersection of Disciplines



Source: Drew Conway, *The Data Science Venn Diagram*
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Who coined the term?

1962 John W. Tukey, in “The Future of Data Analysis”:

“For a long time I thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and doubt… I have come to feel that my central interest is in data analysis… Data analysis, and the parts of statistics which adhere to it, must…take on the characteristics of science rather than those of mathematics… data analysis is intrinsically an empirical science…”

1974 Peter Naur in *Concise Survey of Computer Methods:*

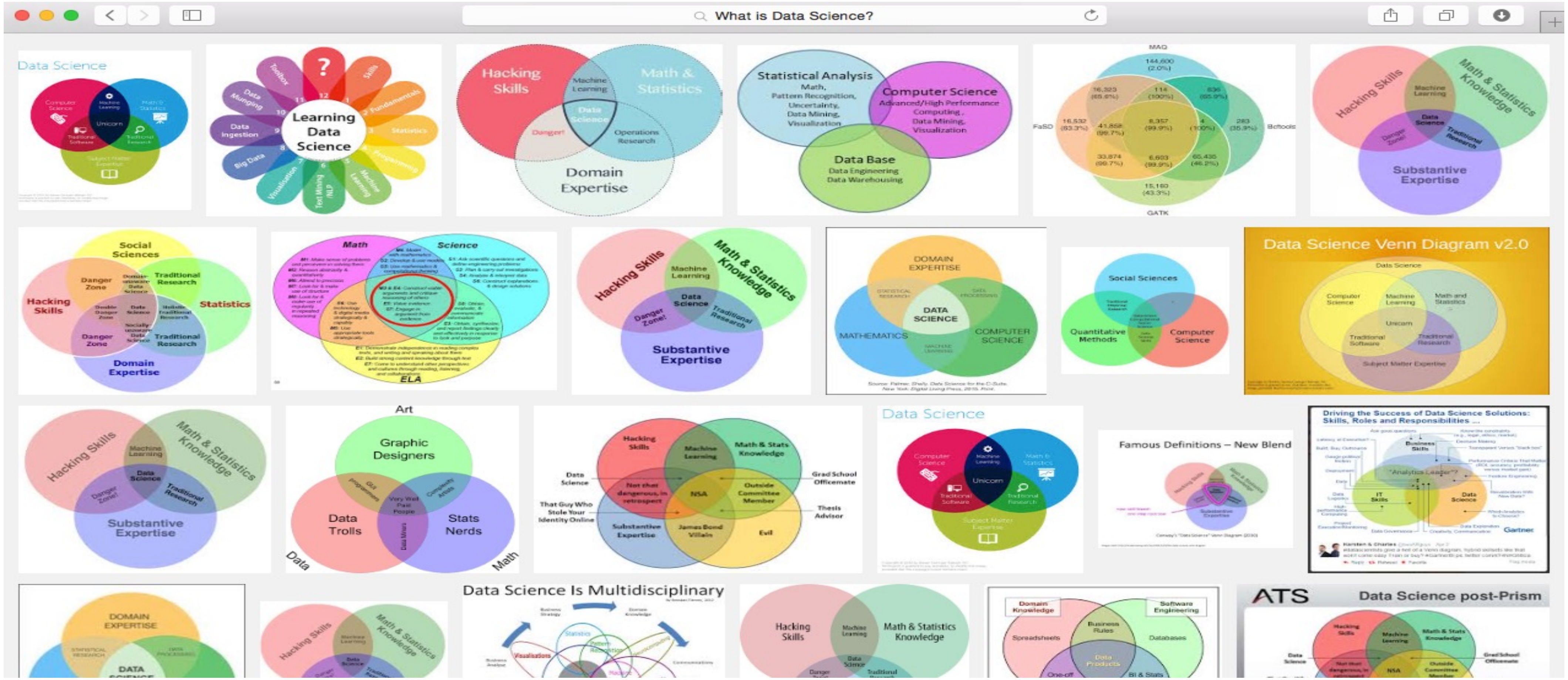
“[in the text of this book], the term ‘data science’ has been used freely.”

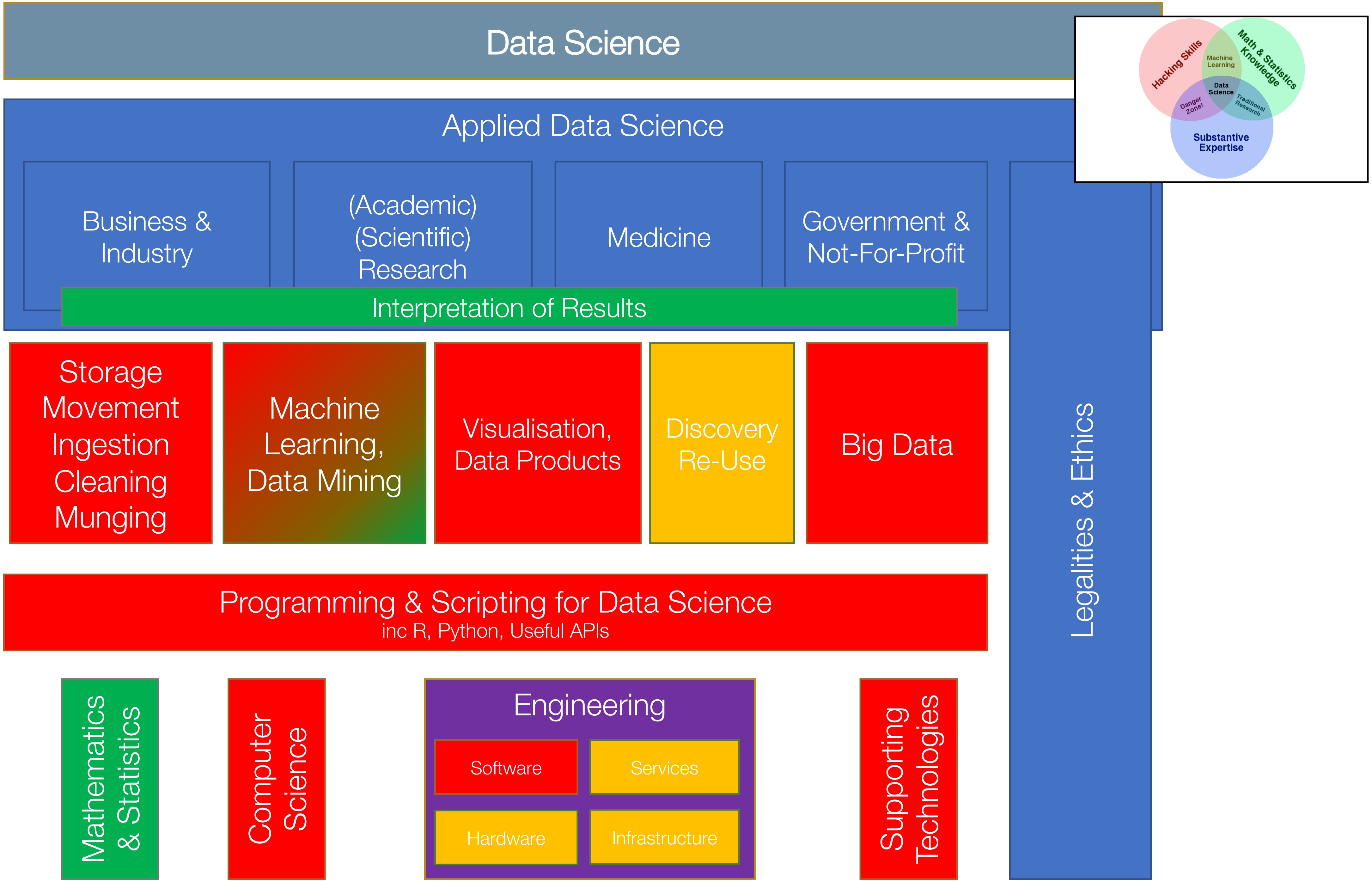
“[data science is] The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences.”

So is Data Science *new*?

- People have been doing data science for a long time.
- What's new is that it's now important enough for people to be *studying it* as a field in its own right and *doing it*. There's now such a thing as a "Data Scientist".

Google (Images) says…





Data Science

Applied Data Science

Business & Industry

(Academic) (Scientific) Research

Medicine

Government & Not-For-Profit

Interpretation of Results

Storage Movement Ingestion Cleaning Munging

Machine Learning, Data Mining

Visualisation, Data Products

Discovery Re-Use

Big Data

Programming & Scripting for Data Science
inc R, Python, Useful APIs

Mathematics & Statistics

Computer Science

Engineering

Software

Services

Hardware

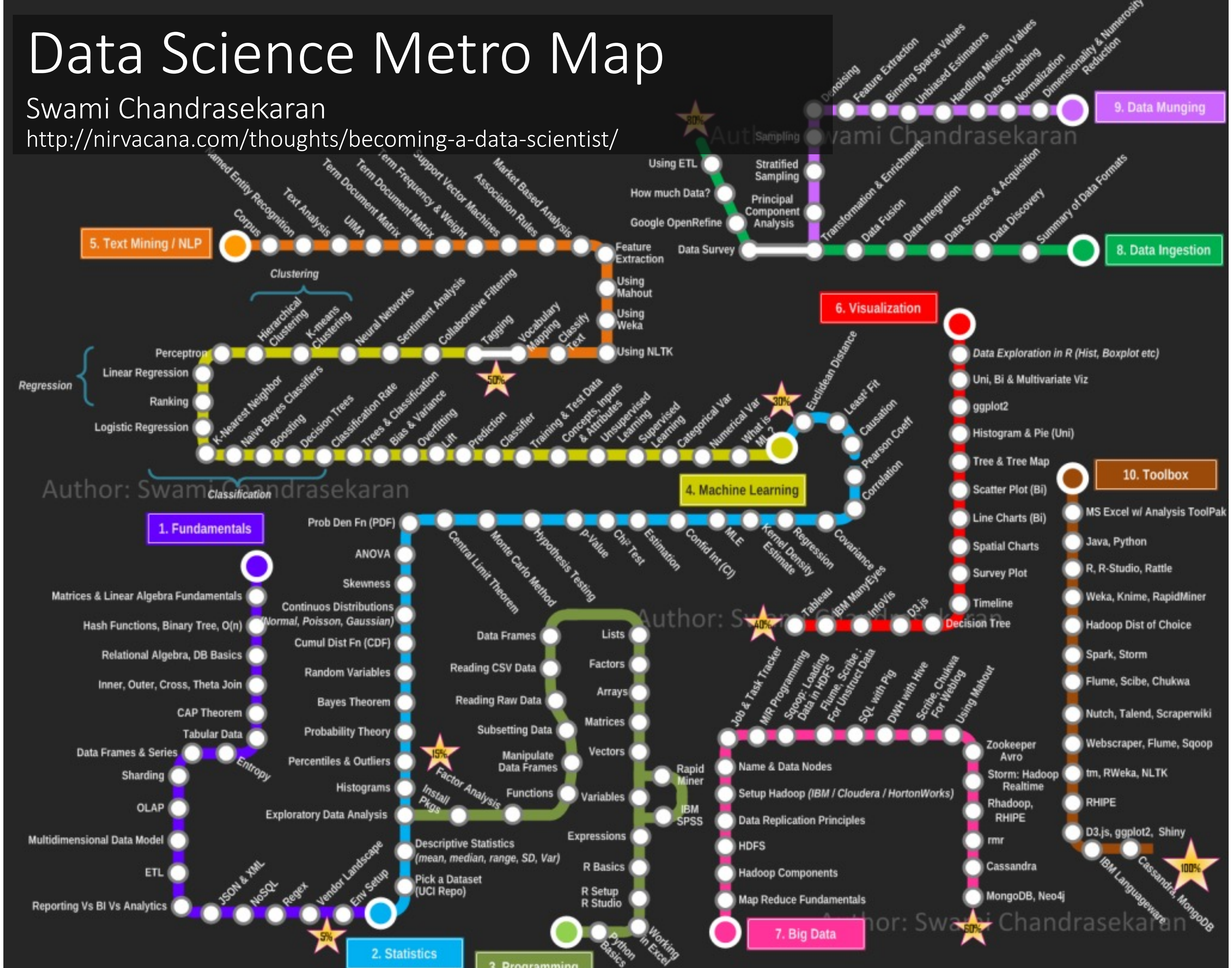
Infrastructure

Supporting Technologies

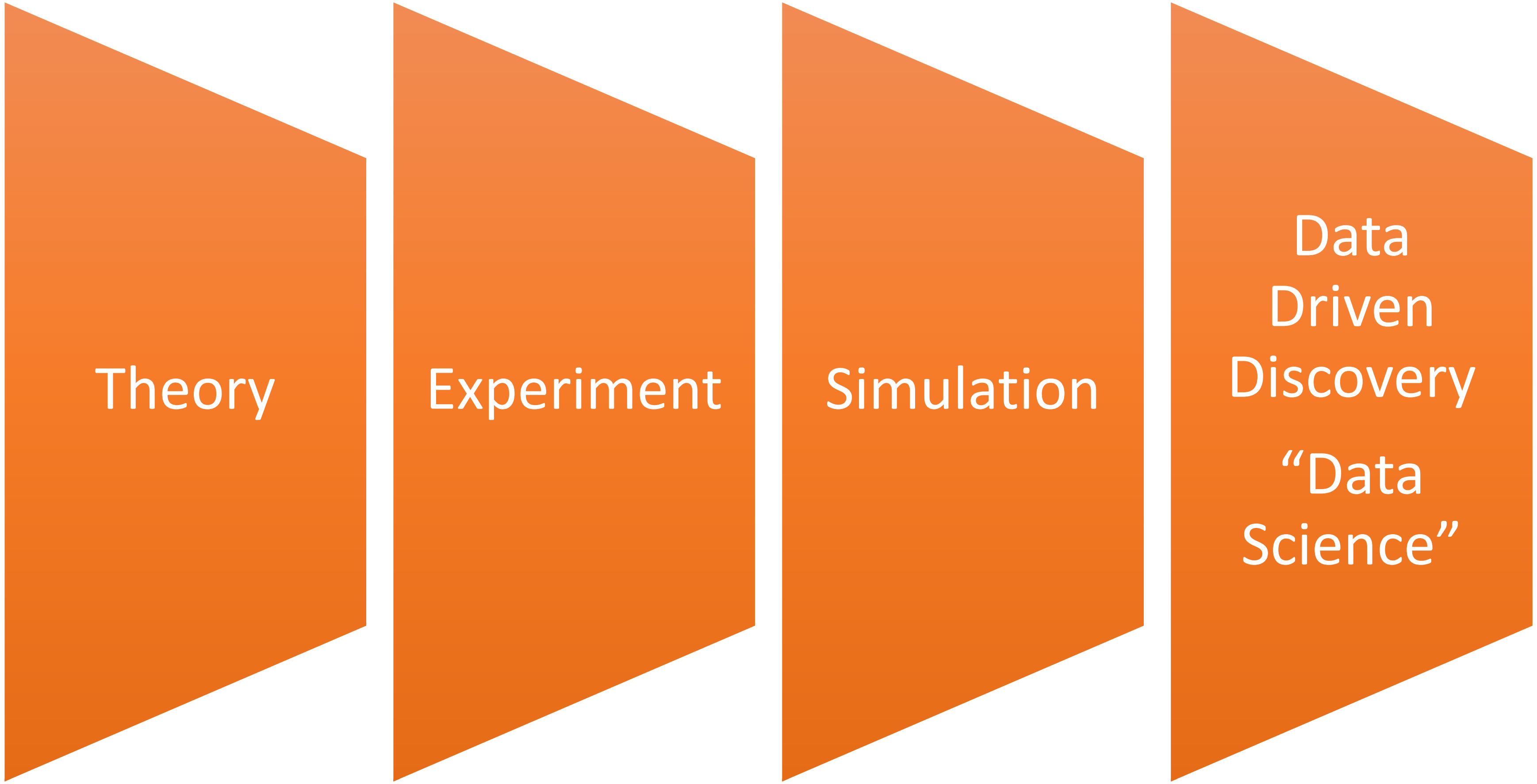
Legalities & Ethics

Data Science Metro Map

Swami Chandrasekaran
http://nirvacana.com/thoughts/becoming-a-data-scientist/



Data Science as the Fourth Paradigm



Theory

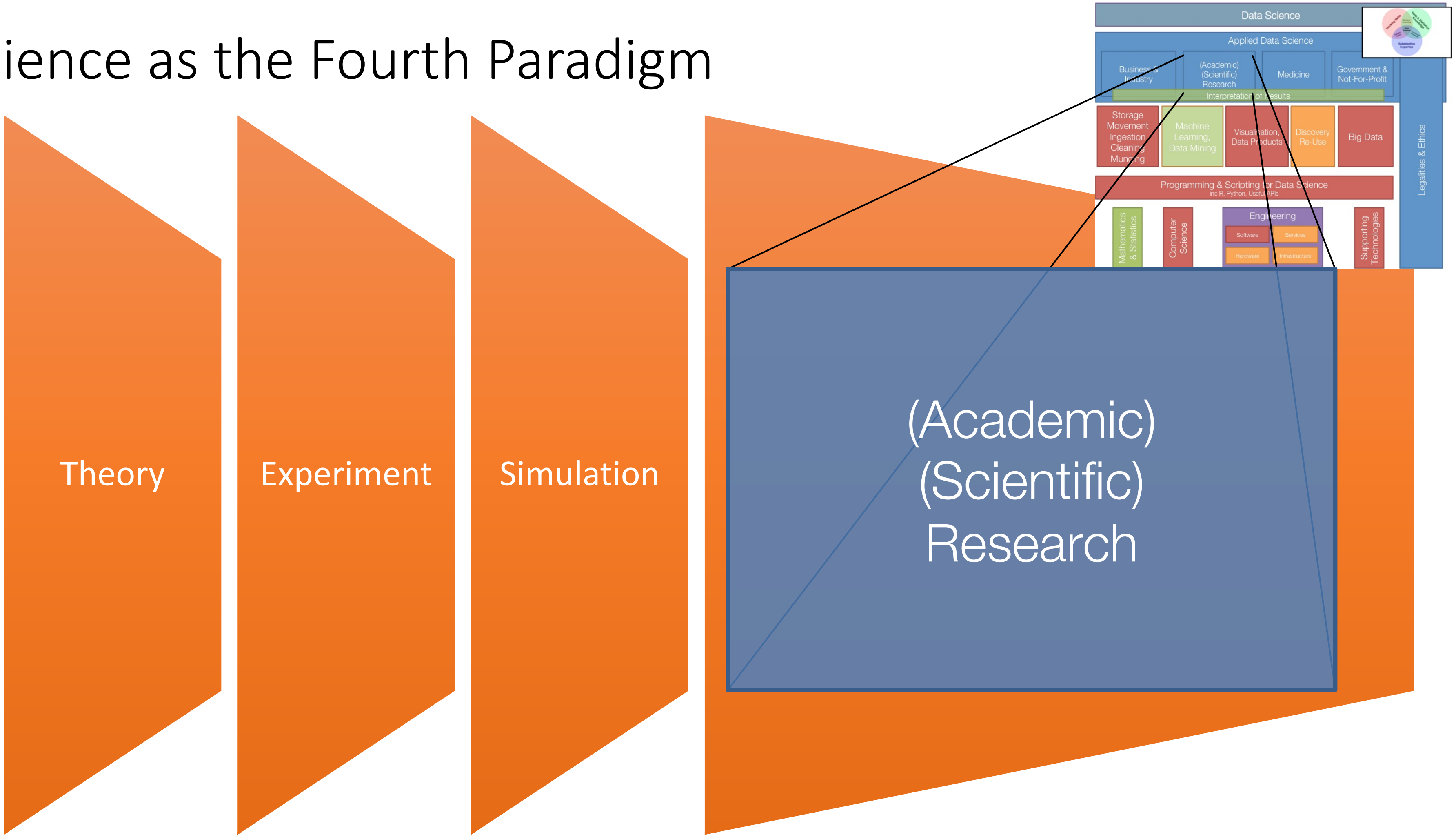
Experiment

Simulation

Data Driven Discovery

“Data Science”

Data Science as the Fourth Paradigm



Theory

Experiment

Simulation

(Academic)
(Scientific)
Research

Data Science

Applied Data Science

Business & Industry

(Academic) (Scientific) Research

Medicine

Government & Not-For-Profit

Interpretation of Results

Storage Movement Ingestion Cleaning Munging

Machine Learning, Data Mining

Visualisation, Data Products

Discovery Re-Use

Big Data

Programming & Scripting for Data Science
inc R, Python, Useful APIs

Mathematics & Statistics

Computer Science

Engineering

Software

Services

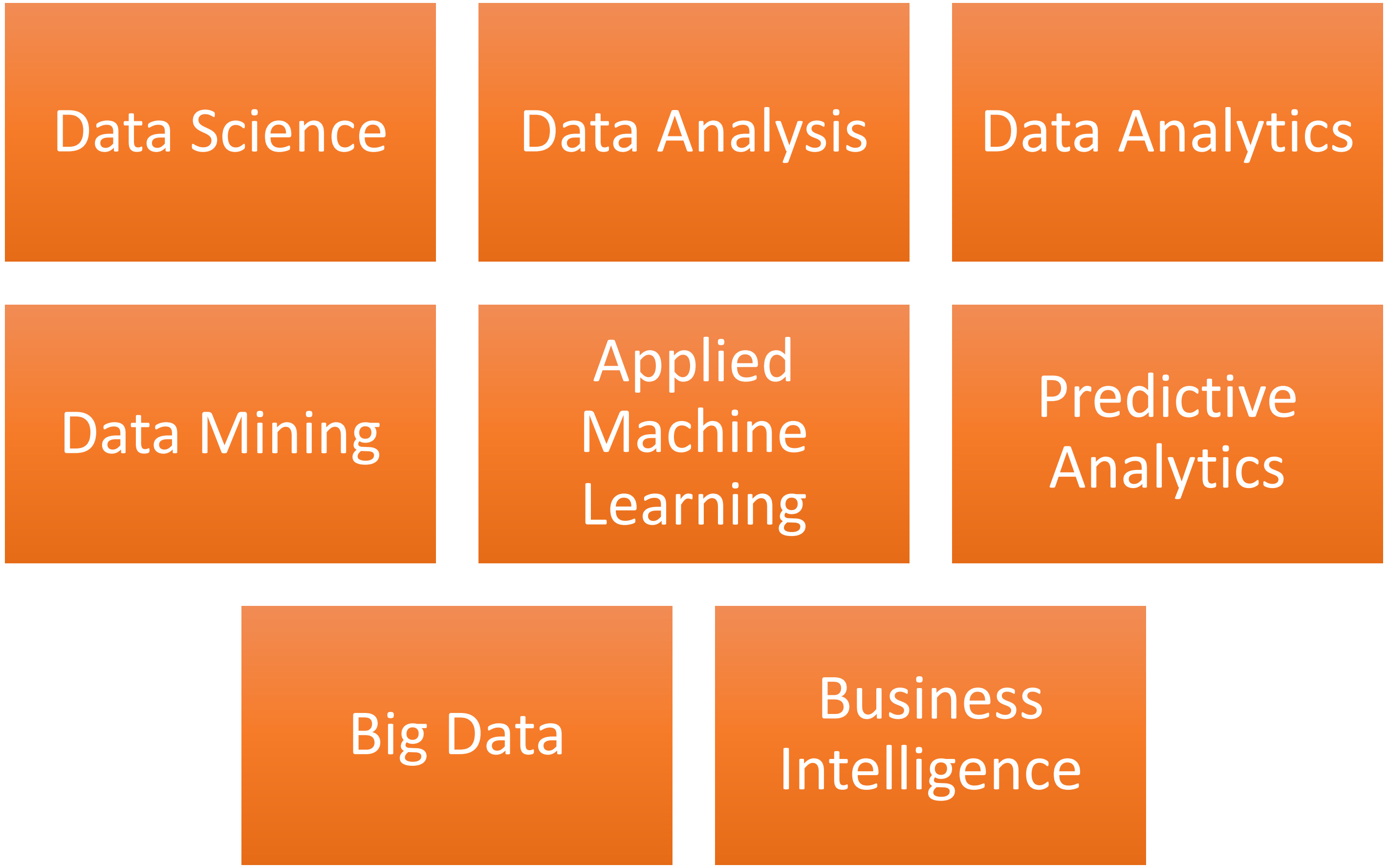
Hardware

Infrastructure

Supporting Technologies

Legalities & Ethics

Related “fields”



Data Science

Data Analysis

Data Analytics

Data Mining

Applied Machine Learning

Predictive Analytics

Big Data

Business Intelligence

Conclusions

- Data Science means different things to different people
    - …although it’s normally a matter of emphasis

- Important aspects of data science:
    - Extraction of knowledge or insight from data
    - Interdisciplinary
    - Applications in multiple fields
    - The combination of skills is new but most are drawn from well-established fields

What is Big Data?

A few slides based on presentation originally created for the Edinburgh Data Science event “Demystifying Big Data”

What is…

BIG DATA

?

What is Big Data?

- “Big Data” is a buzzword. It means different things to different people.
- Data is[*] everywhere. We’re all using it.
- One person’s big is another person’s normal
- Today, I’ll try to pick out some useful stuff from the hype!

[*] http://nxg.me.uk/note/2005/singular-data/

The 3 Vs of Big Data

Data is BIG DATA if it’s too big in _________ to work with in the ways that we’ve been used to

VOLUME

VELOCITY

VARIETY

The Vs o 3 ig Data

VOLUME

VELOCITY

VARIETY

VALUE

The Vs o 4 ig Data

VOLUME

VELOCITY

VARIETY

VERACITY

VALUE

The Vs o 5 g Data

VOLUME

VELOCITY

VARIETY

VALUE

VALIDITY

VERACITY

The Vs o 6 ig Data

VOLUME

VELOCITY

VARIETY

VALUE

VALIDITY

VERACITY

VISUALISATION

The Vs o 7 ig Data

VOLUME

VELOCITY

VARIETY

VALUE

VERACITY

VALIDITY

VISUALISATION

VISIBILITY

The Vs o 8 ig Data

VOLUME

VELOCITY

VARIETY

VALUE

VALIDITY

VERACITY

VISUALISATION

VISIBILITY

VARIABILITY

VOLUME

Too many bytes

Too many files

Too many records

…to fit in cache

…to fit in memory

…to fit on disk

…to store at one site

VELOCITY

Data arrives too quickly

…to be processed

…to be analysed

Data needs to be available very quickly

VARIETY

Too unstructured

…to store in a traditional database

…to process with traditional tools

Multiple Content Types

Images Sensor Data

Video Sound

Sound

Tweets

Multiple Formats

VALUE

This is the reason that people are “doing” Big Data!

The large amount of data that we have available means that *we can do things that we couldn’t do before*

VERACITY

The more that you use data that others have collected, the harder it is to be sure of where it came from, and *whether it’s actually correct*

VALIDITY

The data might be “correct” but is it actually suitable to answer your question?

Of course, you could always take the data, and see what questions it could answer…

VISUALISATION

At some point, humans have to consume the data. As it becomes larger, how do we do this effectively?

VISIBILITY

How do we make sure that the data is **visible** to everyone who’s meant to see it?

…and invisible to everyone who’s not meant to see it?

(and who decides who’s *meant* to see it anyway?)

VARIABILITY

All of the above can vary with time…

One (partial) solution: Cloud Computing

Big Data = Hadoop,

No!

Hadoop is just *one* solution to *some of* the problems of Big Data.

It’s an implementation of MapReduce, which is a re-usable parallelisation strategy, applicable to large data sets.

What about NoSQL?

…again, just one more solution to *some* of the problems

NoSQL databases are databases with their constraints relaxed so that they can work more efficiently at scale

In Summary…

- Big Data is about having to think differently about how you work with your data
    - It possibly means **scaling out** and not just **scaling up**
    - It possibly means using **new tools**
- …but it really can provide **new insight** and offer more **Value**

Acknowledgements and Re-Use

This talk was adapted by Adam Carter, EPCC for the ARCHER2 course Introduction to Data Science & Machine Learning. It is based on a talk originally created for an Edinburgh Data Science event.