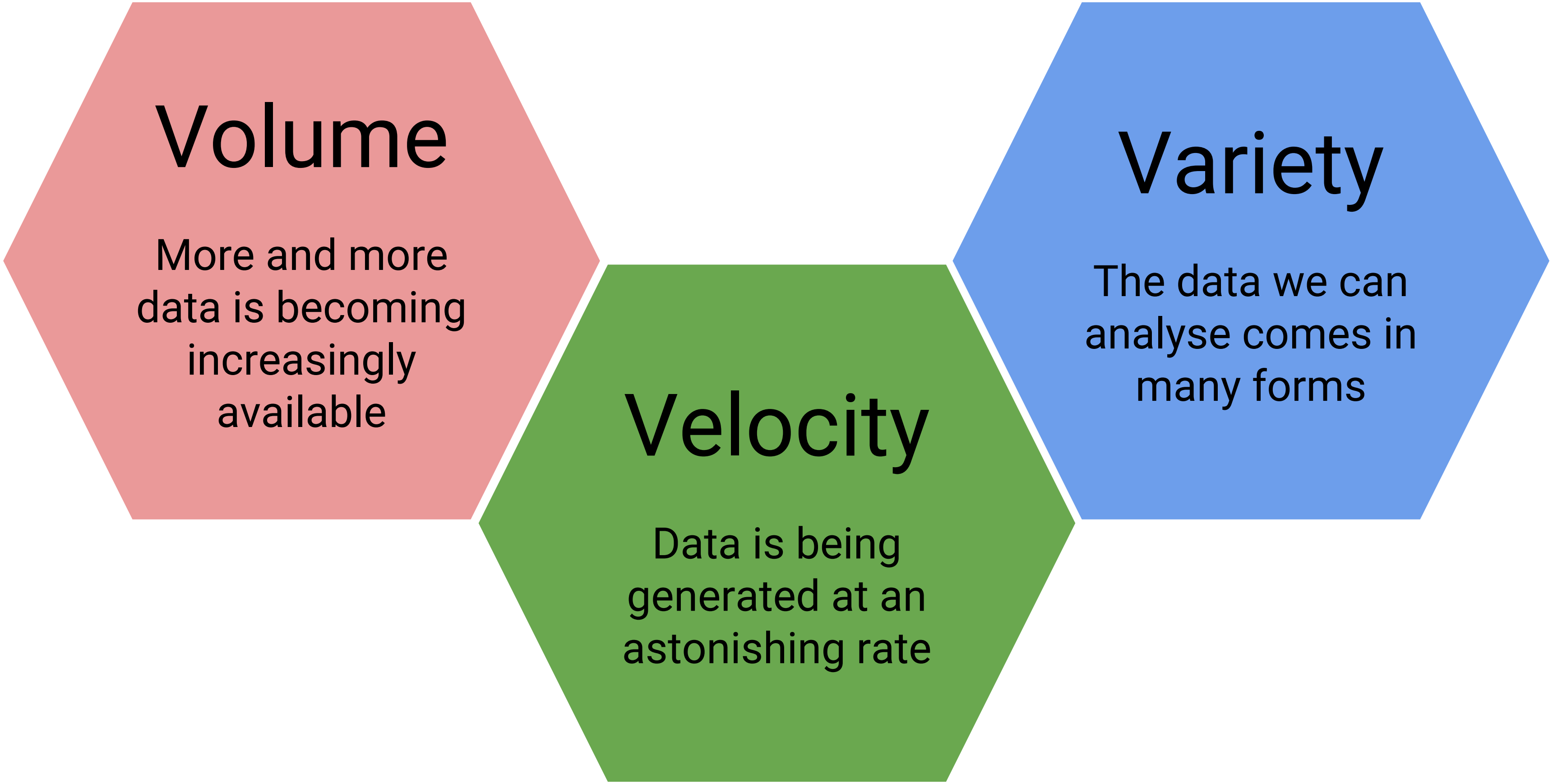
Big Data

The Data Scientist’s Toolbox



Volume

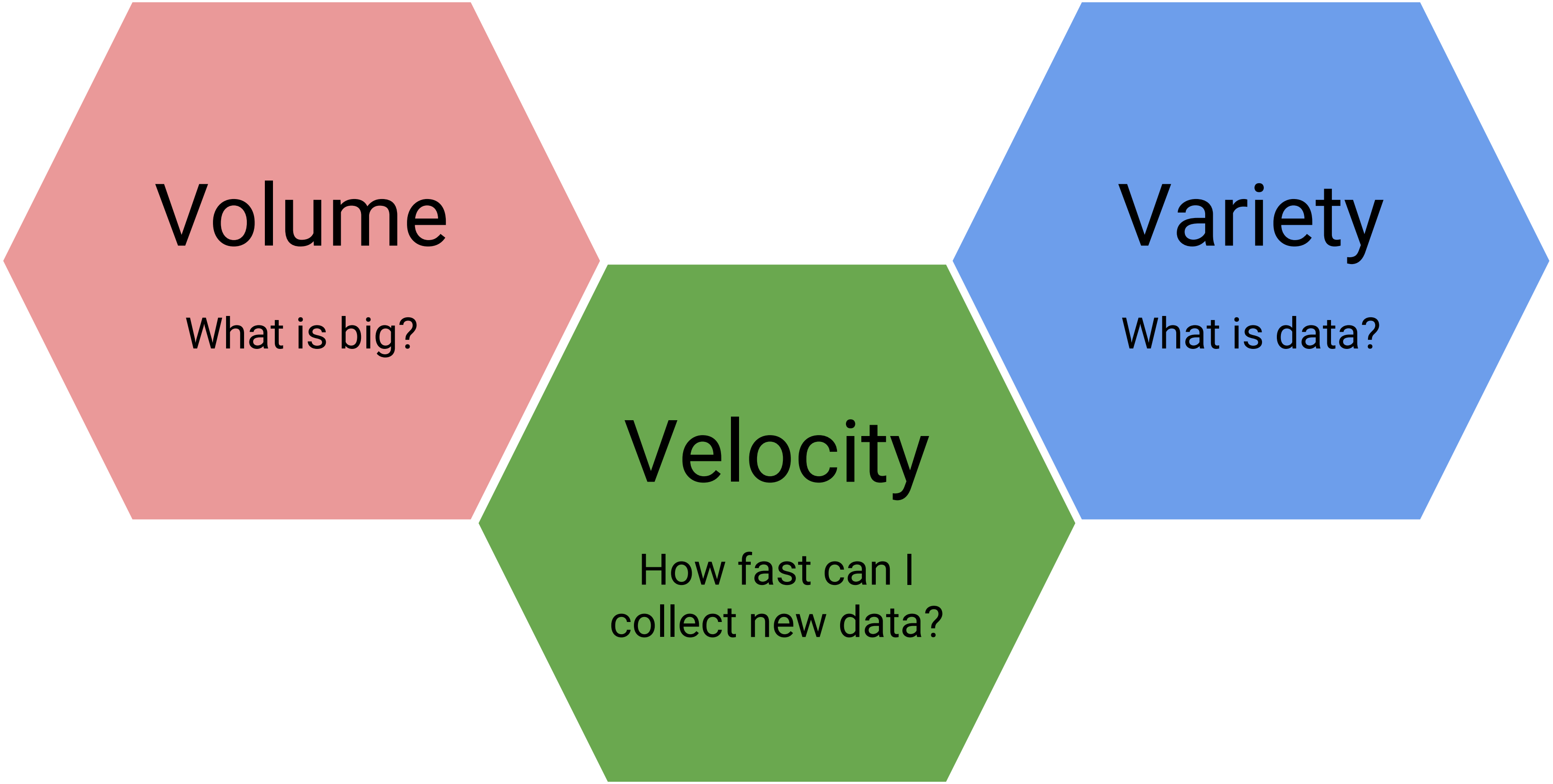
More and more data is becoming increasingly available

Velocity

Data is being generated at an astonishing rate

Variety

The data we can analyse comes in many forms



Volume

What is big?

Velocity

How fast can I collect new data?

Variety

What is data?

Name	Country of origin	Sex	Weight (kg)	Height (cm)
A. Bee	Canada	M	75	163
C. Dee	UAE	M	80	180
E. Eff	China	F	72	175
G. Haitch	South Africa	F	68	172
I. Jay	Poland	M	77	168
K. Elle	Japan	N/A	76	173
M. Enn	Chile	M	80	190

Unstructured Data Types

Text files and documents

Websites and applications

Sensor data

Image files

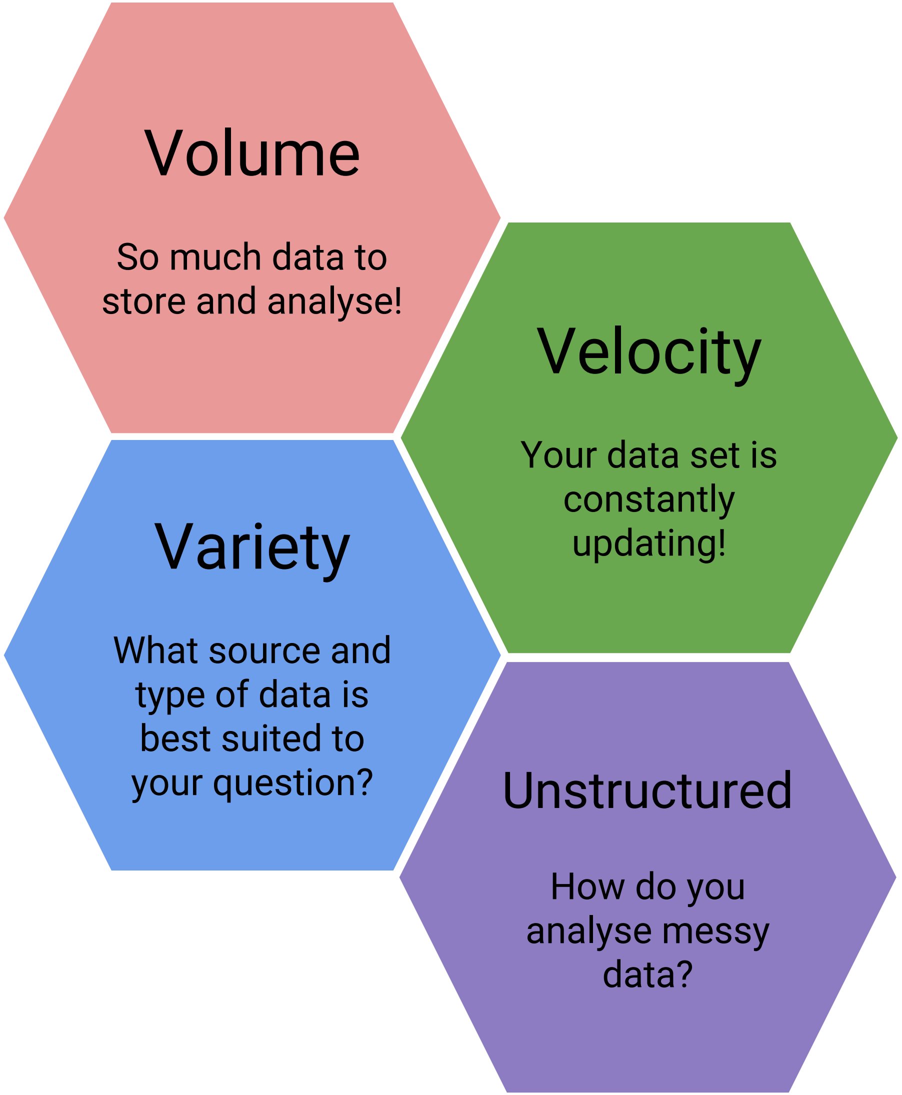
Audio files

Video files

Email data

Social media data

Modified from: https://searchbusinessanalytics.techtarget.com/definition/unstructured-data; using https://icons8.com/



Volume

So much data to store and analyse!

Velocity

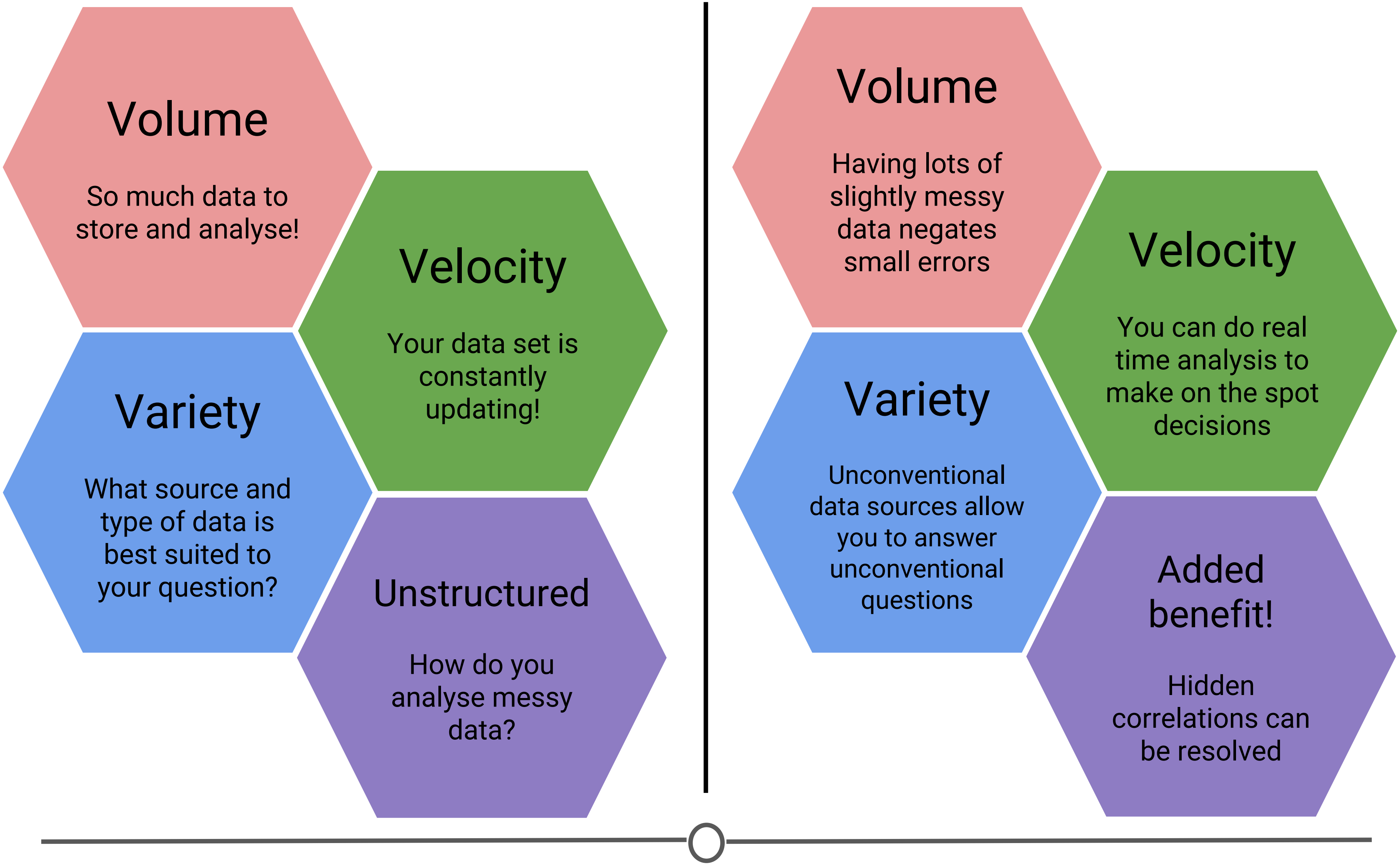
Your data set is constantly updating!

Variety

What source and type of data is best suited to your question?

Unstructured

How do you analyse messy data?



Volume

So much data to store and analyse!

Velocity

Your data set is constantly updating!

Variety

What source and type of data is best suited to your question?

Unstructured

How do you analyse messy data?

Volume

Having lots of slightly messy data negates small errors

Velocity

You can do real time analysis to make on the spot decisions

Variety

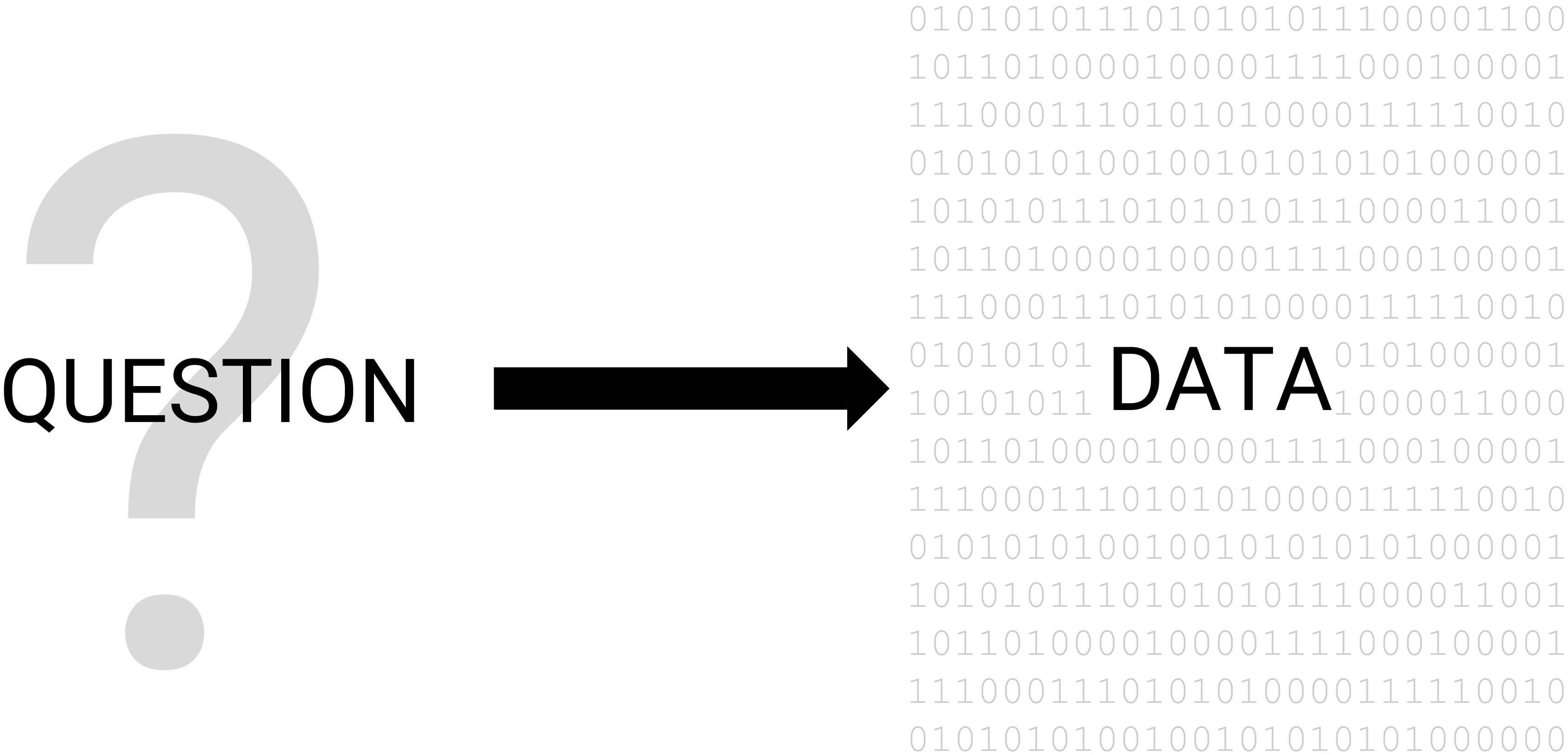
Unconventional data sources allow you to answer unconventional questions

Added benefit!

Hidden correlations can be resolved

The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.

- John Tukey, 1986



QUESTION

DATA

Summarizing: Big Data

The Data Scientist’s Toolbox