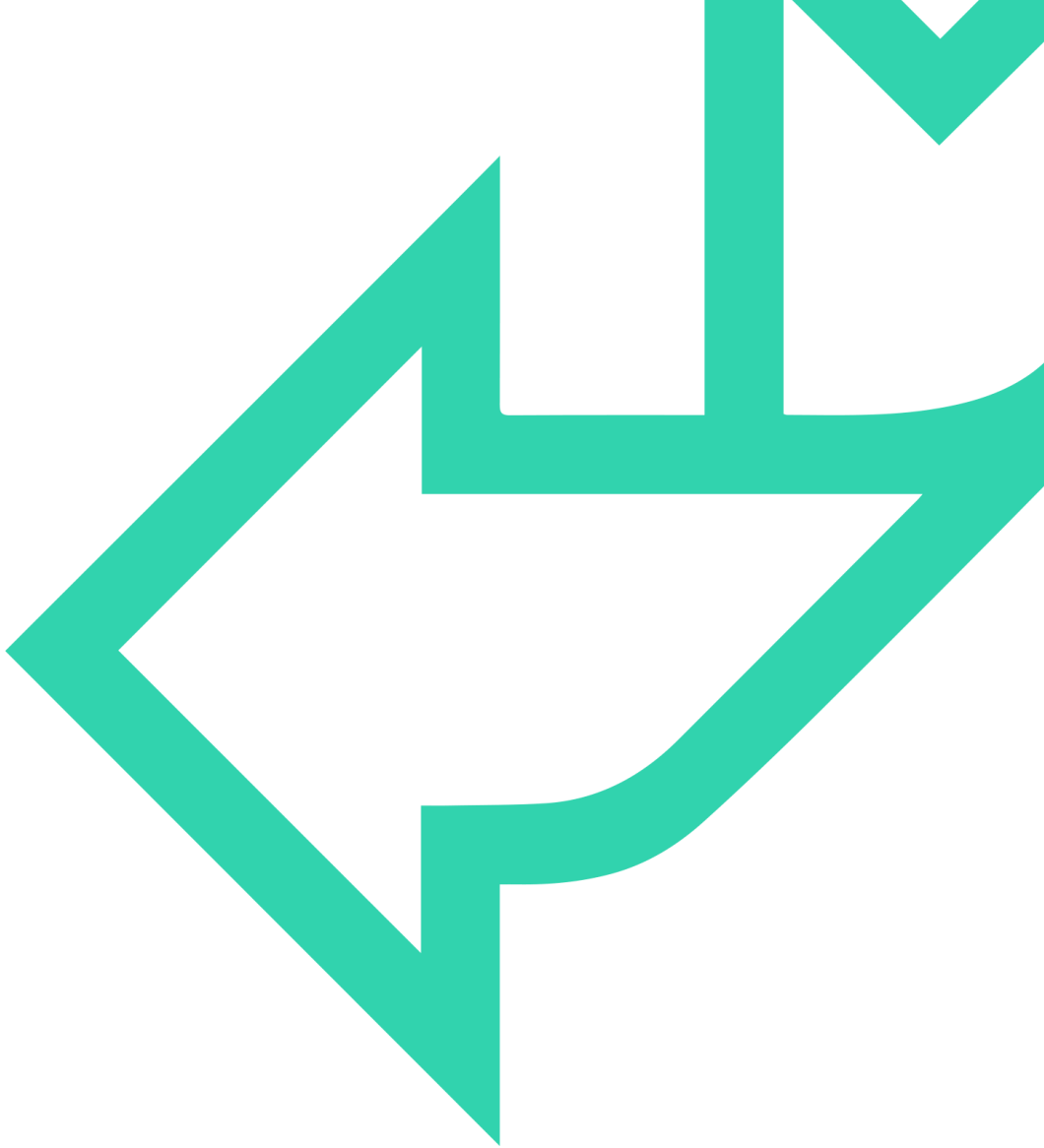
POWERING POTENTIAL



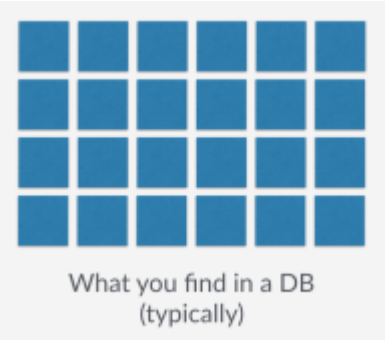
Data Science

the process of extracting knowledge from data automatically

DATA

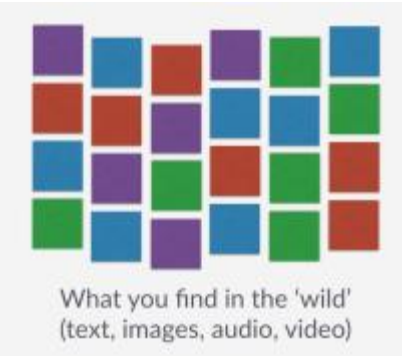
What is data?

- Information
  - Relevant
  - Raw



Data Types? (forms)

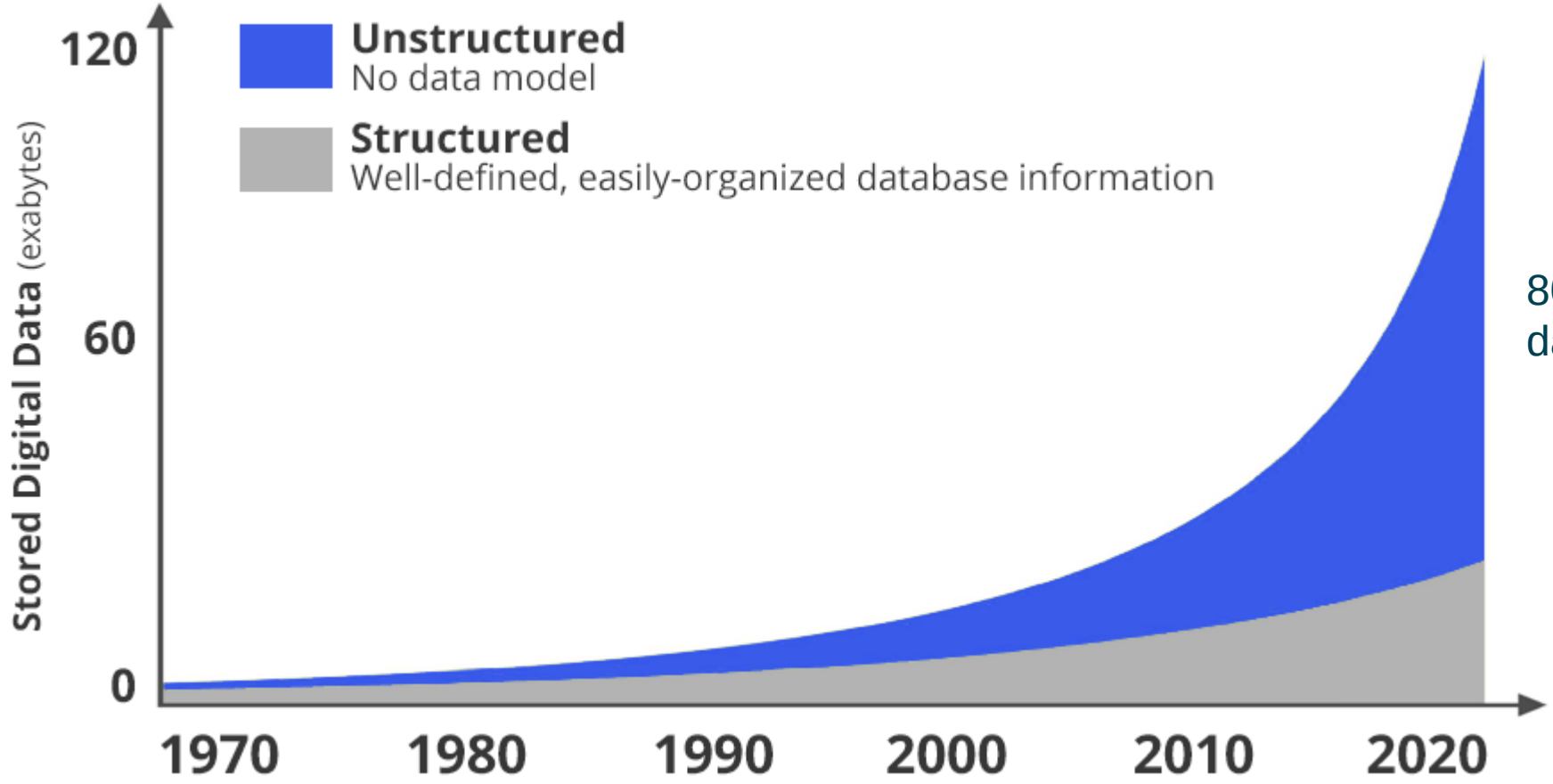
- Structured Data
  - Tabular
  - SQL Databases
- Unstructured Data
  - Photos
  - Sounds
  - Videos



Collected from where?

- Social Media
- Bank Accounts
- ….. everywhere
- Smart devices
- GPS

volume of unstructured data is growing & that growth is accelerating

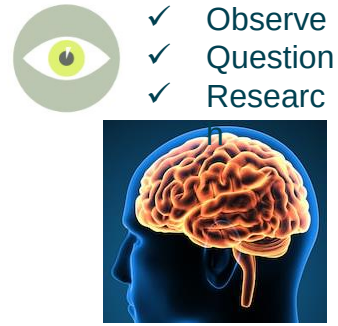


80-90% of
data is unstructured

volume of unstructured data is growing & that growth is accelerating

SCIENCE


Every baby knows the
scientific method!
1 Make an observation.
2 Form a hypothesis.
3 Perform the experiment.
4 Analyze the data.
5 Report your findings.
6 Invite others to reproduce the results.


✓ Observe
✓ Question
✓ Researc

DATA SCIENTIST

DS = evidence-based methodology for solving problems using computational and analytical tools

Data Science = applied branch of statistics

data → valuable asset → insight → decisions → actions

DA

- BA = Business Analysis
- EDA = Exploratory Data Analysis

DS

- ML = Machine Learning Algorithms
- DE = Data Engineering

Explaining

Predicting

automated analysis

extract knowledge

visualisation & communication

DATA SCIENTIST

“A data scientist is someone who is
better at statistics than any software engineer and
better at software engineering than any statistician.”

Wills (2012, Director of Data Science at Cloudera)

“… the sexiest job … will be statisticians….

data … understand it …. extract value from it … visualise it …”

Varian (2008, Chief Economist at Google)

“A data scientist is somebody
who is inquisitive,
who can stare at data and spot trends…
… like a Renaissance individual who really wants to learn and bring change to an organization.”

Bhambhi (2012, VP of Big Data products at IBM)

DATA SCIENTIST

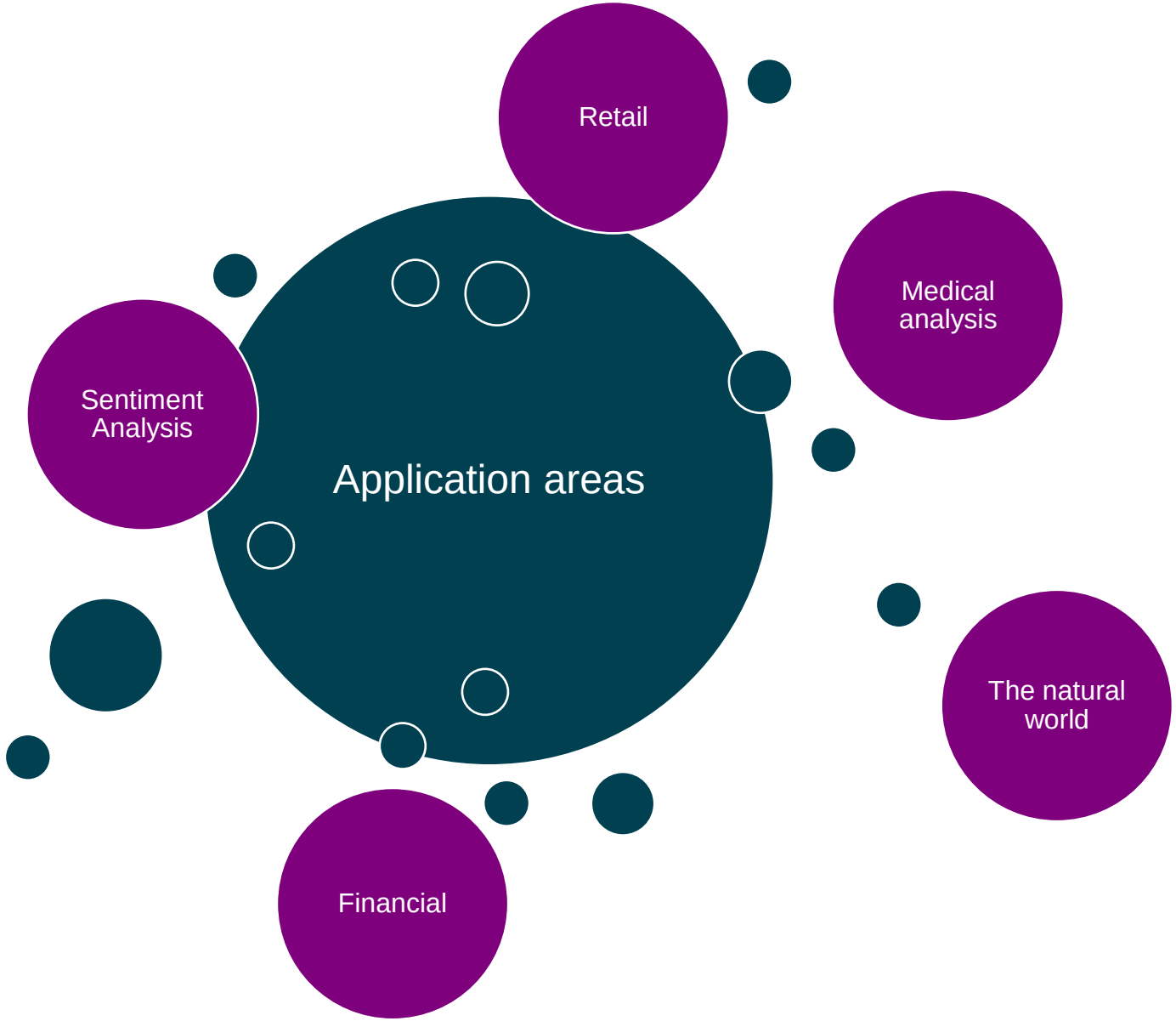
“effective data science should not be treated like just another business process, and can not be operationalized assembly-line style.

Data science -- as the name suggests -- is a mode of inquiry and exploration similar to “real” science.

Just as a physicist uses math to reason about the natural world, data scientists harness mathematical and computational tools to reason about the business world.”

Peter Wang (2019, CEO Anaconda)

DATA SCIENCE APPLICATIONS



DATA SCIENCE APPLICATIONS

Marketing

Optimizing ads

Identifying the best ad to run online

Forecasting Churn

Predict which type of customers are most likely to leave your service

Segmenting Customer Base

Deeply understand customer base for better tailored products and services

Finance

Stock Price Prediction

Forecasting share indices and specific stock prices

Improved Customer Service

Using text data for chatbots and call center routing

Fraud Detection

Identify fraudulent transactions based on key patterns

DATA SCIENCE APPLICATIONS

Sentiment Analysis - Social media – can we gain intuition/vibe?

Medical analysis - Are you likely to get a certain illness based on your lifestyle (or other factors)?

The natural world - What is going on with the weather? snow?

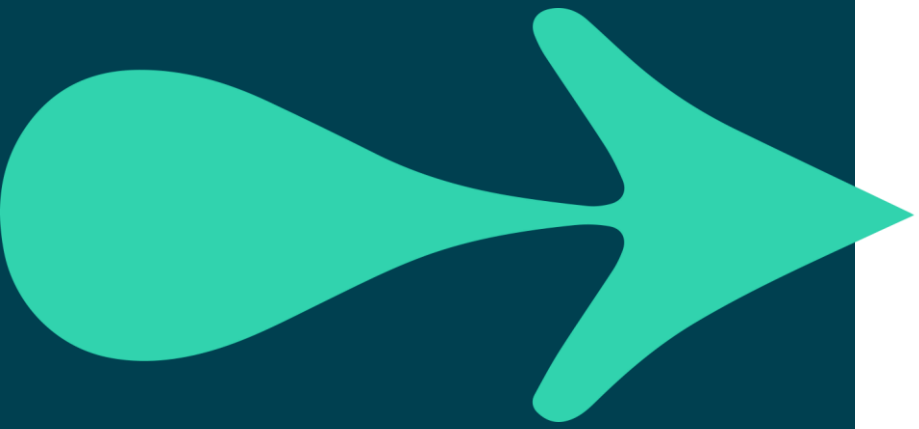
Financial - Is the market going to go up or down?

Retail - Who is most likely to buy something?

Supermarkets - What stock to put in the front of the shops?
What discounts to give?

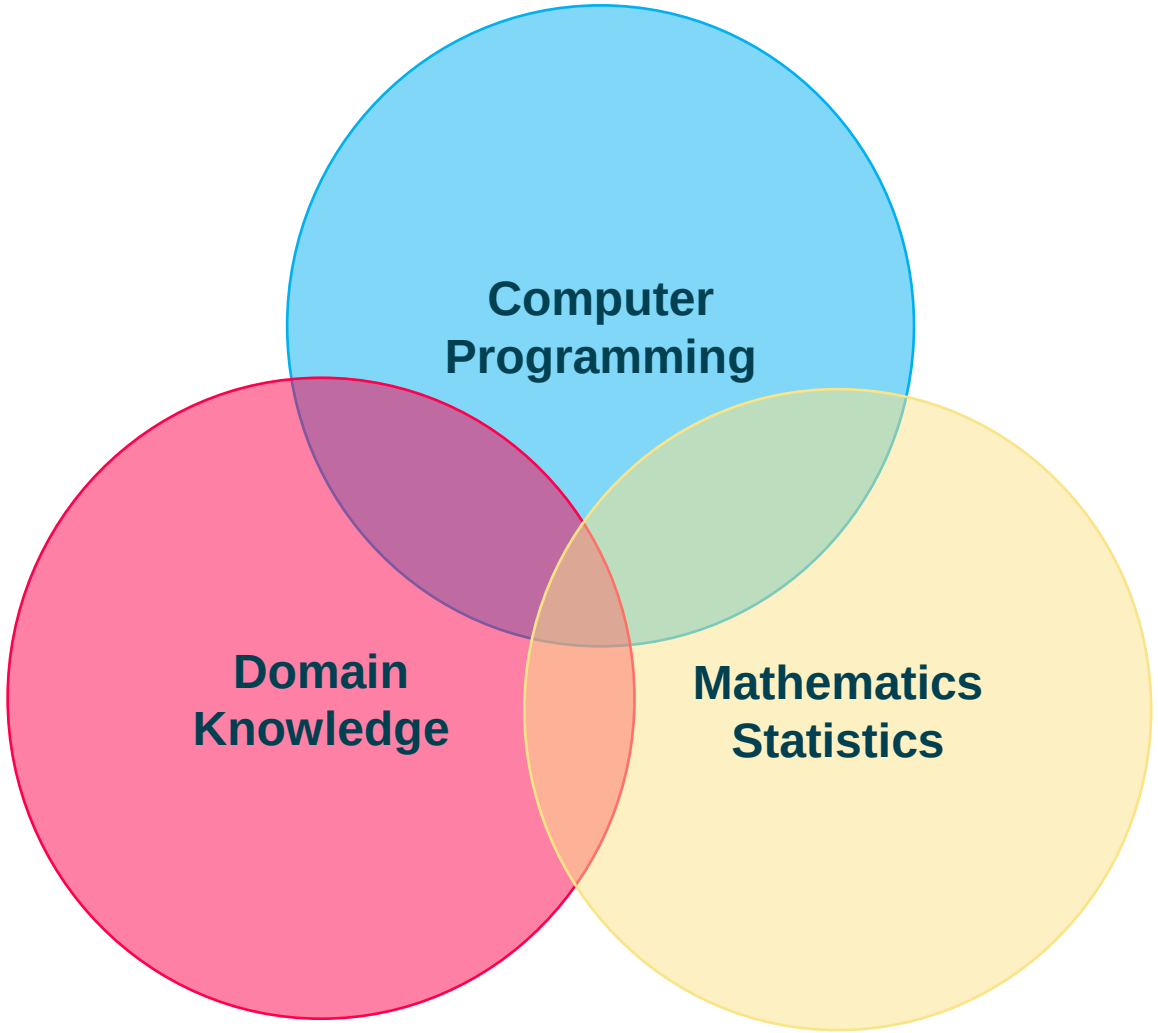
Personal - Why can't I hit 10,000 steps a day?

NOMENCLATURE

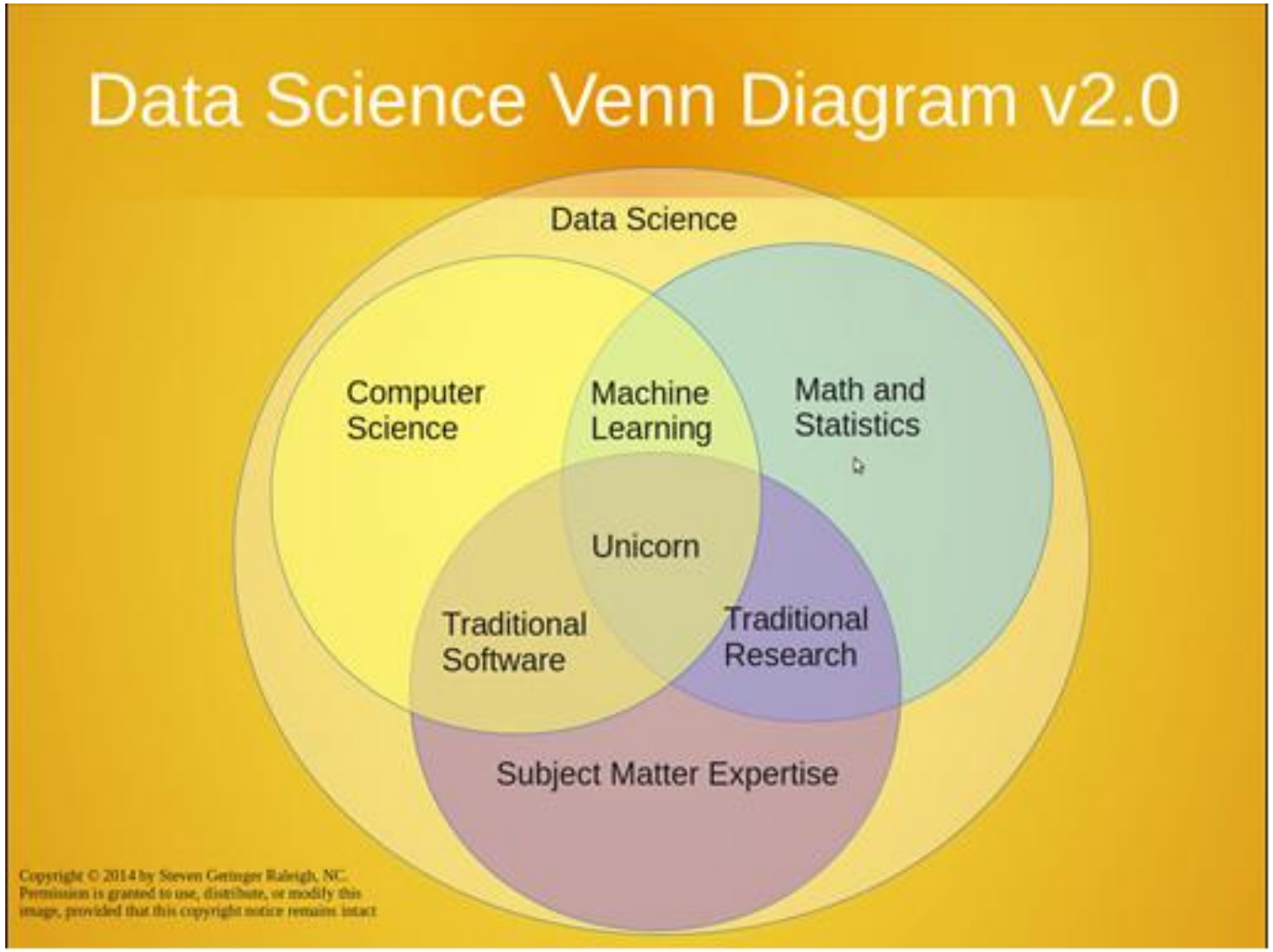


Data scientists	Statistician
“example”	“observation”
“feature”	“predictor” “independent variable”
“label”	“response” “dependent variable”

DATA SCIENCE SKILLSET



DATA SCIENCE SKILLSET



© Steve Geringer

DATA SCIENCE SKILLSET

- ✓ **Domain Knowledge**
- ✓ **Programming**
- ✓ **Statistics**
- ❑ Curiosity
- ❑ Visualisation
- ❑ Communications
- ❑ Storytelling
- ❑ Project Management/DevOps
- ❑ Databases
- ❑ Data Mining
- ❑ Machine Learning

DATA SCIENCE SKILLSET

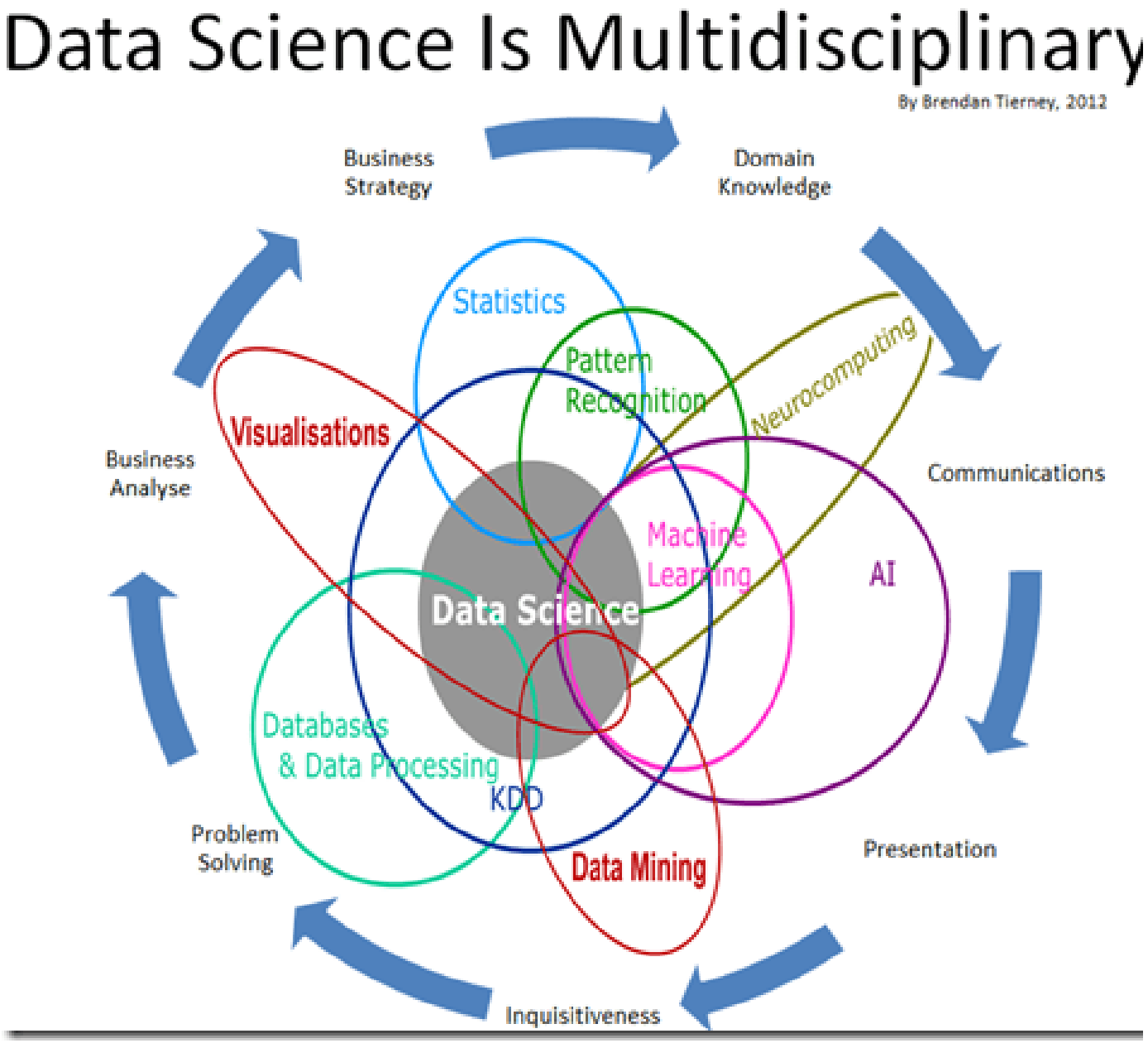
- ✓ **Domain Knowledge**
- ✓ **Programming**
- ✓ **Statistics**
- ❑ Curiosity
- ❑ Visualisation
- ❑ Communications
- ❑ Storytelling
- ❑ Project Management/DevOps
- ❑ Databases
- ❑ Data Mining
- ❑ Machine Learning

In many walks of life evolution ***selects against*** the kind of person who decides to find out what happens “if I push that button”.

In Data Science it ***selects for*** it.

2012, Ross
(Director Data Science at Teradata)

DATA SCIENCE SKILLSET



ANALYTICS

Descriptive Analytics

Making data visible and getting it in the right hands

Predictive Analytics

Predict the future with data

Prescriptive Analytics

Making data-driven decisions

MATHEMATICS & STATISTICS

Probability

- Belief of event occurrence
- Bayes rule

Statistics

- Random Variables (Features)
- Samples vs Populations
- Distributions
- Inference

Linear Algebra

- Transformations on data sets
- Weighted Sums
- Models (Tensor, Matrix, Vector)

Calculus

- Areas and Rates
- Rates of error functions
- Parameter Optimisation

PYTHON LIBRARIES

NumPy

- fast numerical arrays
- optimized fortran and C extensions

Pandas

- numpy wrapper
- provides "data frames"
- tabular model over numpy arrays

matplotlib

- visualization and plotting

seaborn

- convenience matplotlib wrapper

Bokeh

- alternative graphing library (for the web)
- especially useful for geoplots and other complex plots

SciKit Learn

- comprehensive machine learning library
- provides good-enough implementations of most key algorithms

Tensorflow

- fast (concurrent, distributed, gpu) numerical computing library
- describes computations as optimizable graphs

Keras

- tensorflow (et al.) wrapper providing neural network abstractions

TOOLS

Programming Languages



Databases


Command Line Tools



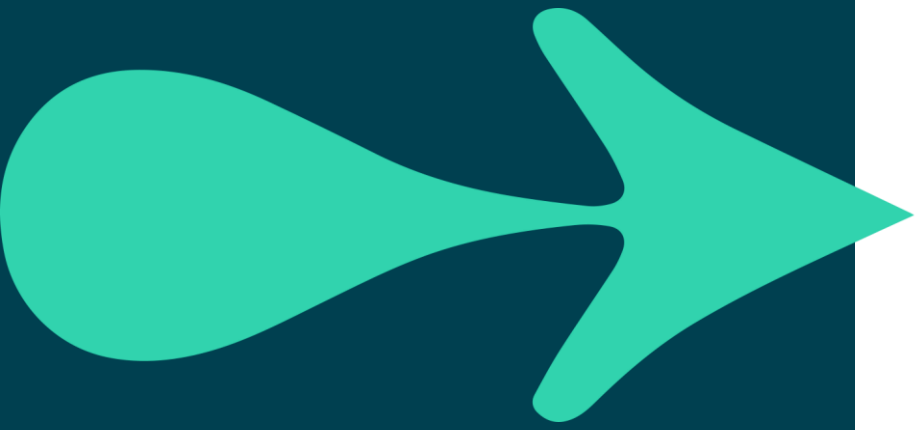
Spreadsheets

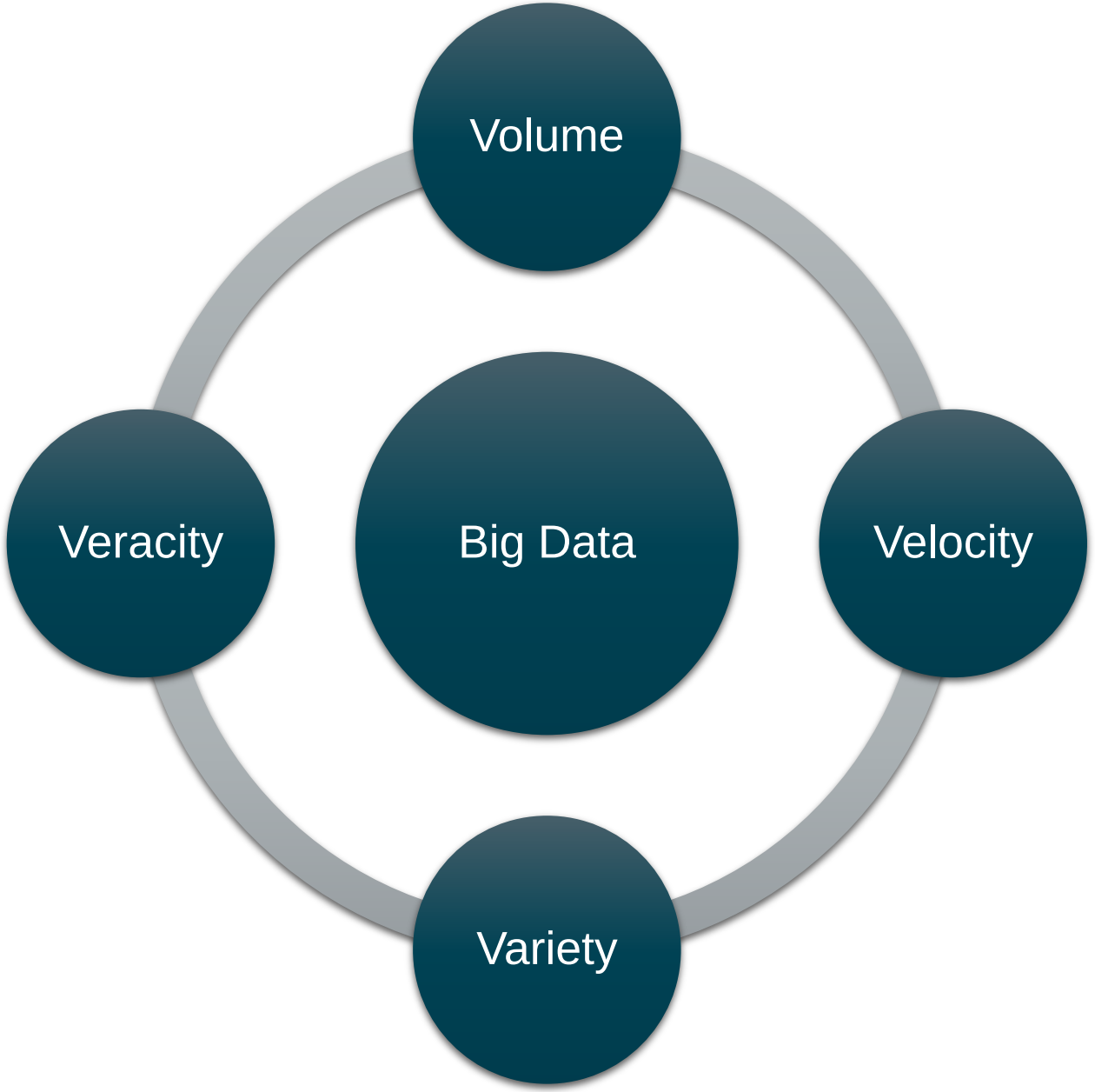
Business Intelligence Tools


Big Data

collection, ingestion, processing and analysis
of datasets **too large**, and generated **too quickly**,
to be analyzed effectively
by traditional analytical tools and methodologies

BIG DATA





VOLUME

Huge amount of data

VERACITY

Inconsistencies and uncertainty in data

VARIETY

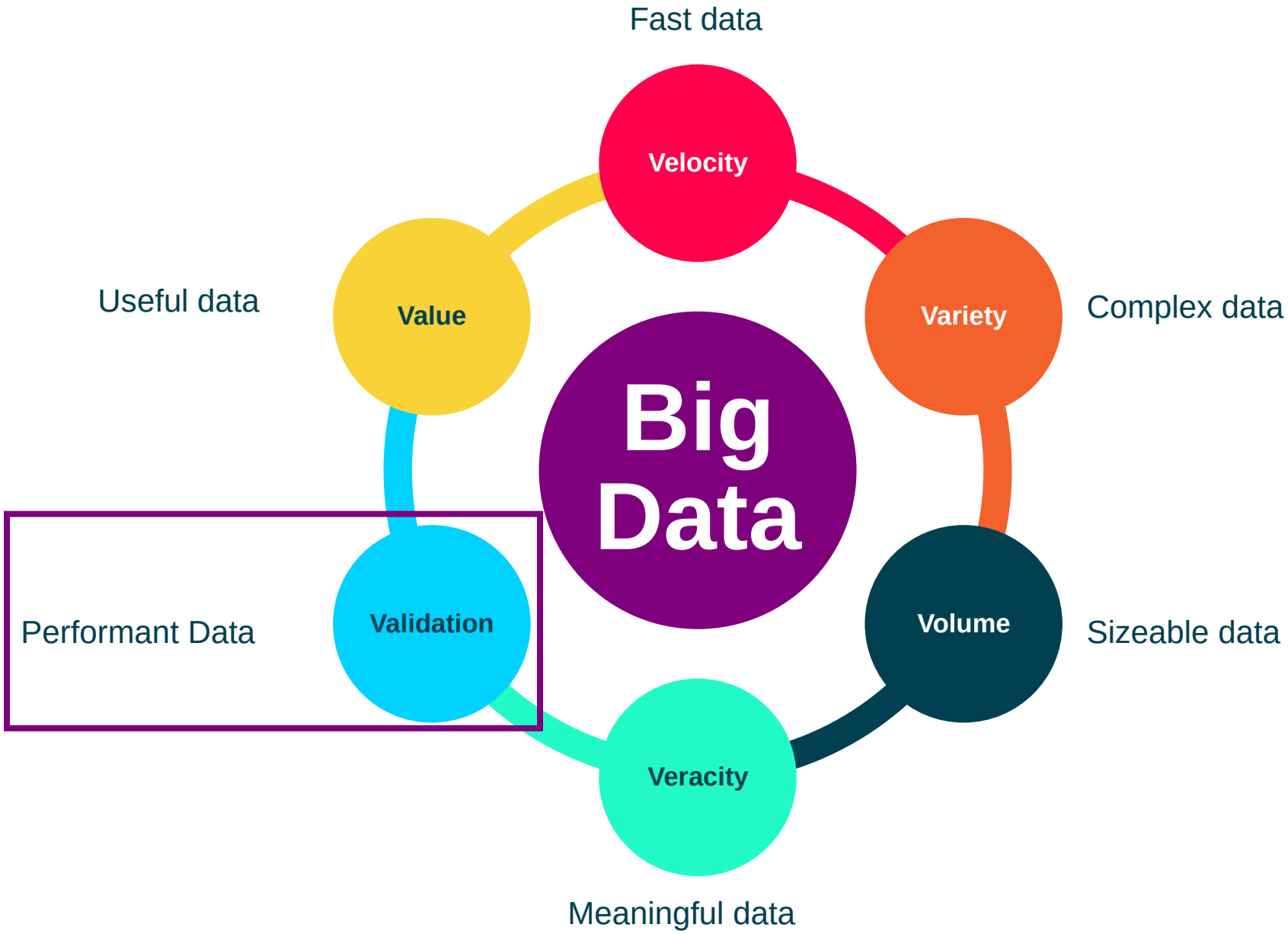
Different formats of data from various sources

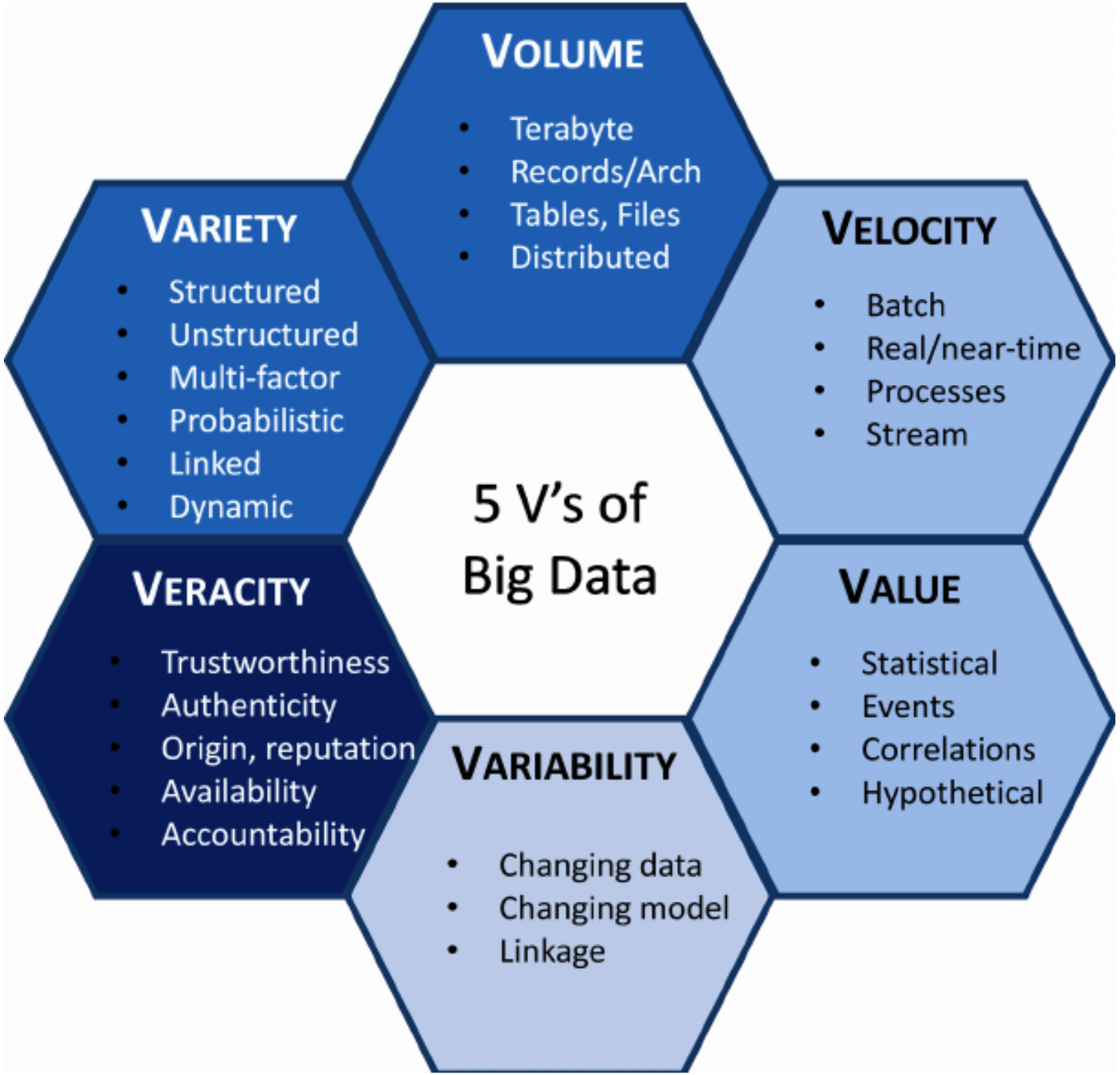
VELOCITY

High speed of accumulation of data

VALUE

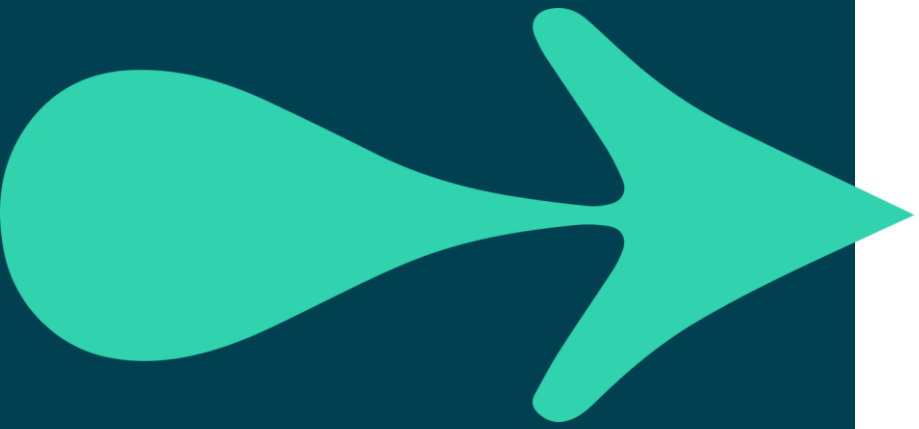
Extract useful data


Fast data
Velocity
Useful data
Value
Variety
Complex data
Big Data
Performant Data
Validation
Volume
Sizeable data
Veracity
Meaningful data


VOLUME
Terabyte
Records/Arch
Tables, Files
Distributed
VARIETY
Structured
Unstructured
Multi-factor
Probabilistic
Linked
Dynamic
VELOCITY
Batch
Real/near-time
Processes
Stream
5 V's of
Big Data
VERACITY
Trustworthiness
Authenticity
Origin, reputation
Availability
Accountability
VALUE
Statistical
Events
Correlations
Hypothetical
VARIABILITY
Changing data
Changing model
Linkage


BIG DATA
with 8 V's
01
VOLUME
Can you find the information you are looking for?
02
VALUE
Can you find it when you most need it?
03
VERACITY
Are you dealing with information or disinformation?
04
VISUALISATION
Can you make sense at a glance? Does it trigger a decision?
05
VARIETY
Is a picture worth a thousand words in 70 languages? Is your information balanced?
06
VELOCITY
Information gains momentum and crises & opportunities evolve in real time. How is outlook for today?
07
VISCOSITY
Does it stick with you? Does it call for action?
08
VIRALITY
AHA to-go? Does it convey a message that can be pasted into a presentation or Instgrammed?

BIG DATA: SOURCES



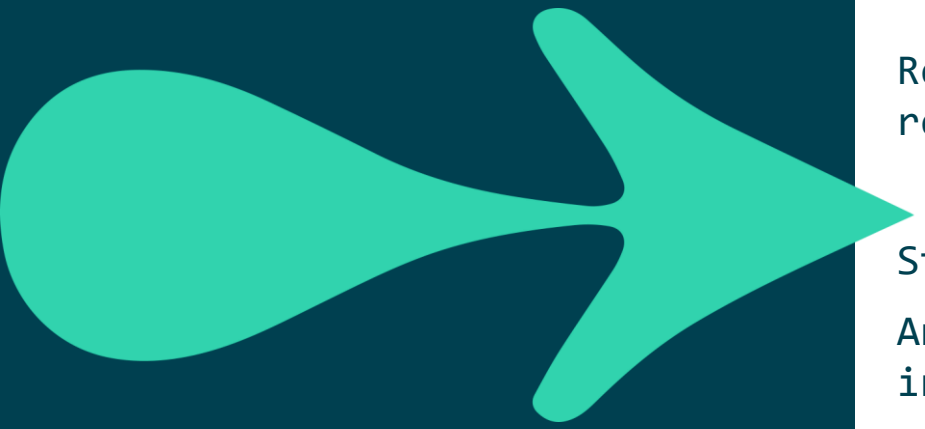
Human Generated

- Photos, videos
- Text
- Click Like
- Web search
- Emails & SMS
- Online Purchases

Machine Generated

- Cell Phones , GPS
- Industrial Process Monitoring
- Climate monitoring
- Medical Devices
- IoT

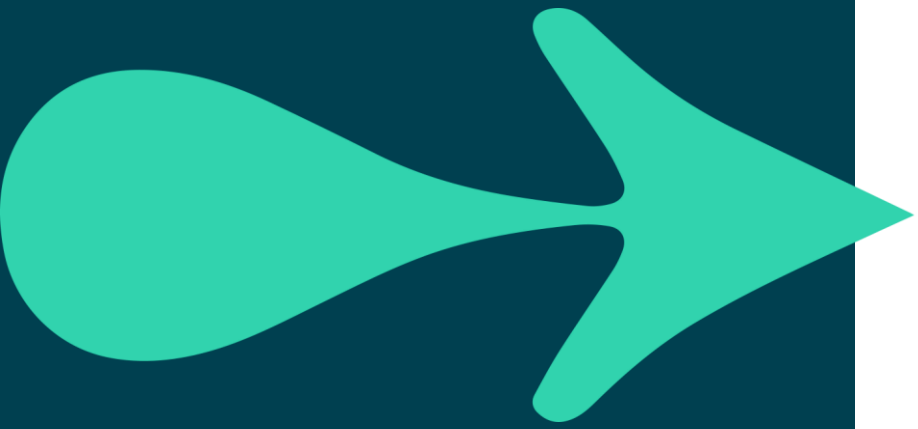
BIG DATA VS SMALL DATA



Small Data	Big Data
Goals: Specific	Goals: Evolving
Location: In one place	Location: distributed
Structure: structured	Structure: unstructured & multi formats
Preparation: by end user for their own purpose	Preparation: by many, used by many
Longevity: kept for specific time	Longevity: kept for long period of time
Reproducibility: can be reproduced	Reproducibility: may not be possible
Stakes: data loss cost is limited	Stakes: data loss cost is high
Analysis: can be analysed at once in one PC	Analysis: Multiple steps as files can be in different places and formats

APPENDIX: Jargon

ARTIFICIAL INTELLIGENCE



Neural Networks

- A particular pattern-finding algorithm

Deep Neural Networks

- A particular generalization of the neural network algorithm

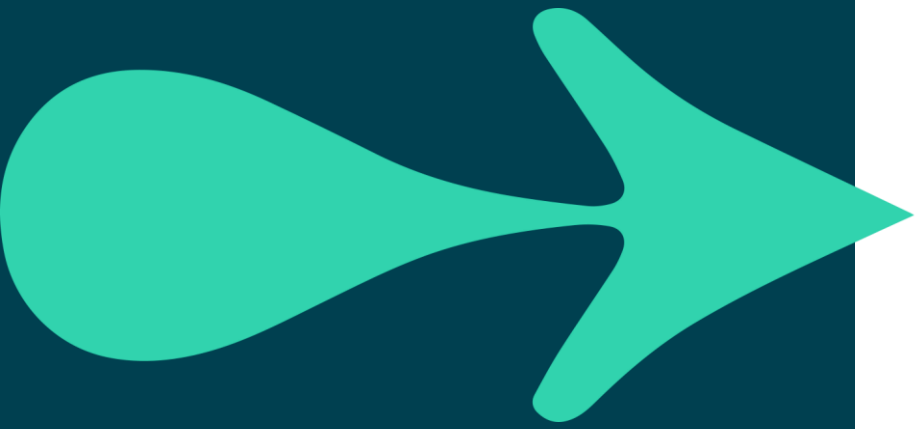
Artificial Intelligence

- (boad def.) the project of simulating animal intelligence

Artificial Intelligence

- whatever artifical system is the best at decision making
- 50s: computers running simple programs
- 80s: computes running expert programs
- 00s: computers running machine learning programs
- 10s: computers running machine learning with neural networks

DATA SCIENCE



Statistics

- Describing and finding patterns in data

Patterns

- Correlations between variables

Statistical Inference

- Finding novel patterns / making predictions / generalizing from observation

Machine Learning

- Computational Statistical Inference = Statistical Inference with Computers

Learning

- Finding the “best” patterns

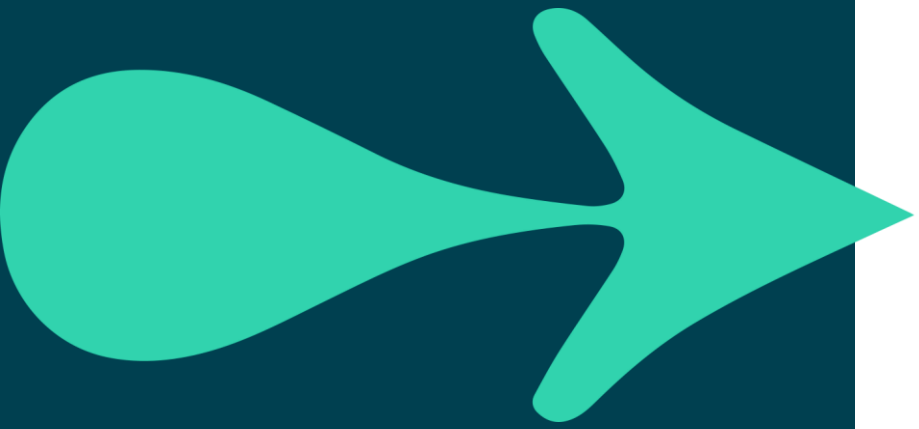
ROLES

Data Analysis

- Historical Trends, Data Fact Finding

Data Science

- Future Trends, Inference & Prediction
- Domain Understanding
- Data Understanding
- Data Exploration, Preparation
- Statistical Modelling
- Evaluation of Models
- Deploying solutions



DATA INSIGHT

Data Scientist

- Machine Learning Developer
- Researcher

Data Engineer

- Big Data Engineer
- DataOps
- Data Science with DevOps

Data Leader

- Project Manager