COMP41680

# Introduction to Data Science

Derek Greene

UCD School of Computer Science

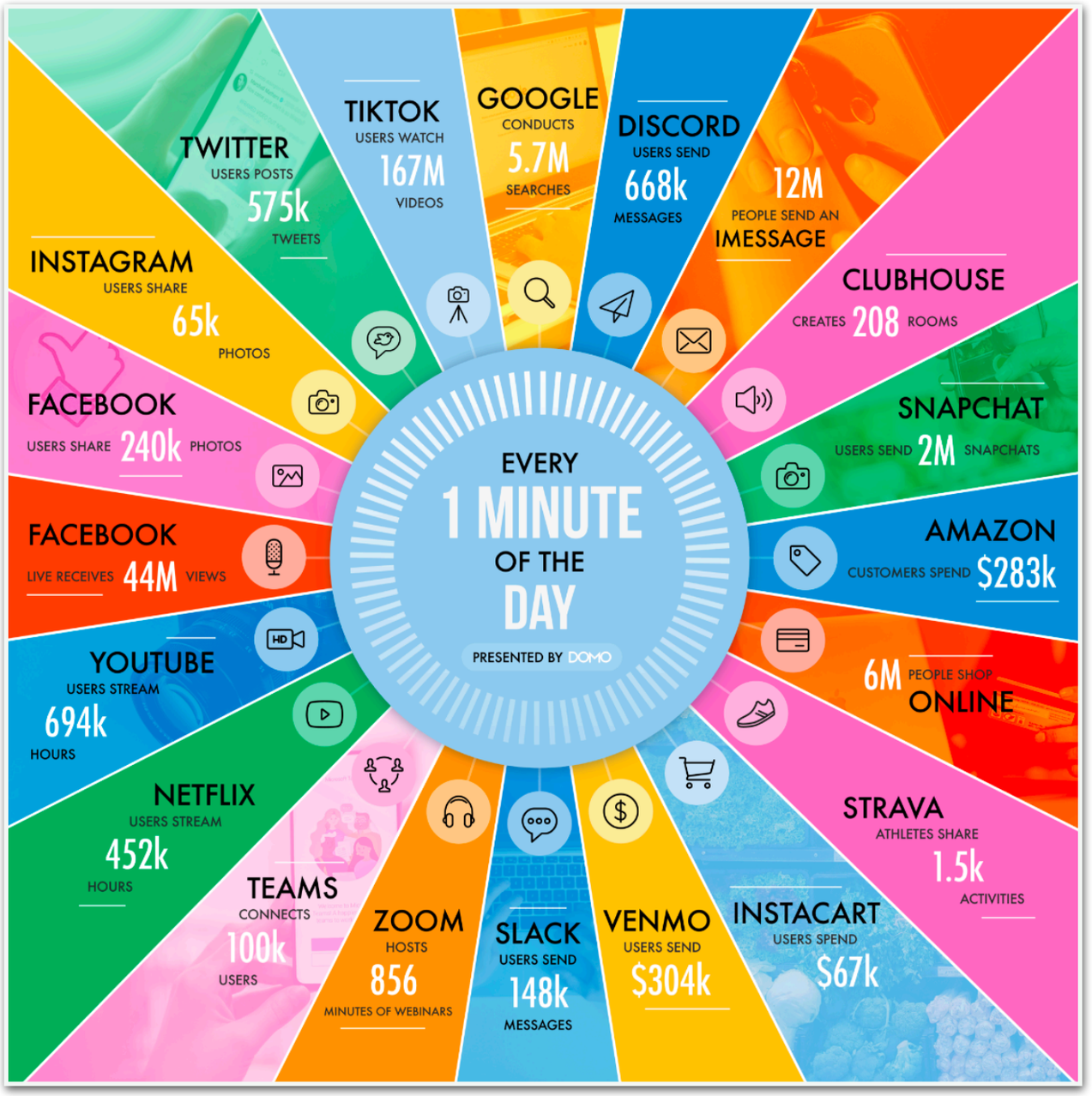
Spring 2023

# The Big Data Era

In the last decade we have witnessed an explosion in the production and availability of digital data.

"Information is the oil of the 21st century, and analytics is the combustion engine." – Peter Sondergaard (Gartner)

How much data is generated every minute?




domo.com

# The Big Data Era

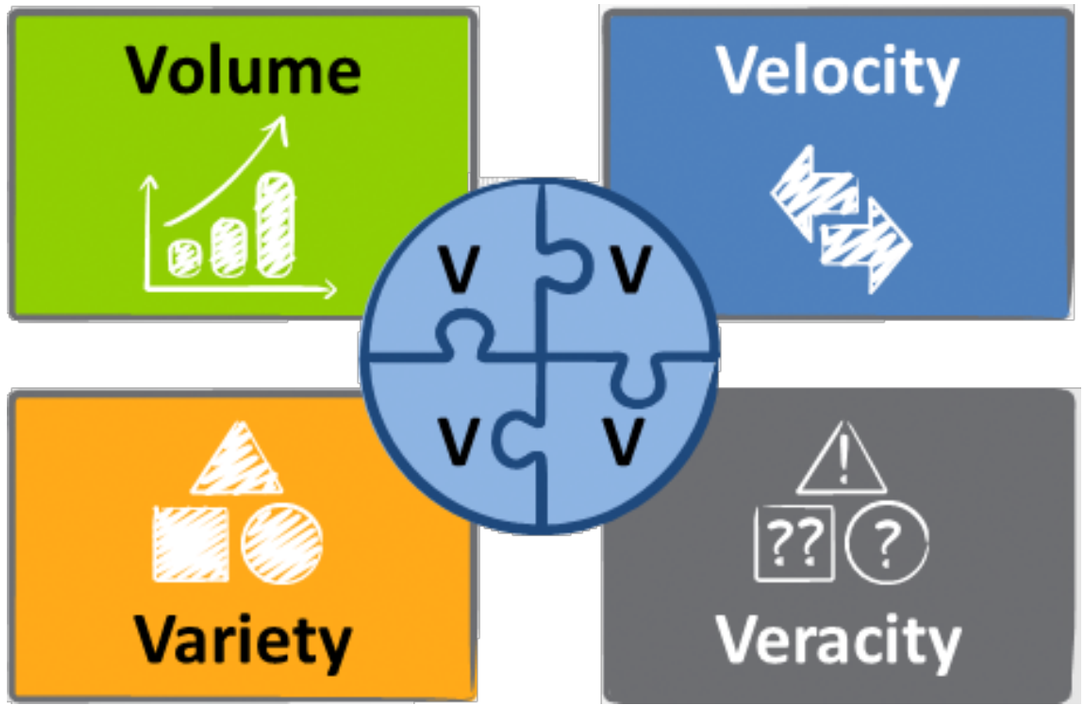
Big Data problems are usually characterised by…

**V**olume: Huge volumes of data - terabytes or more

**V**elocity: Continuous streams of data arriving in real time

**V**ariety: Many different types of data - numeric data, structured text, unstructured text, images, video, audio, time series data

**V**eracity: Uncertain if data is reliable, noisy or incorrect



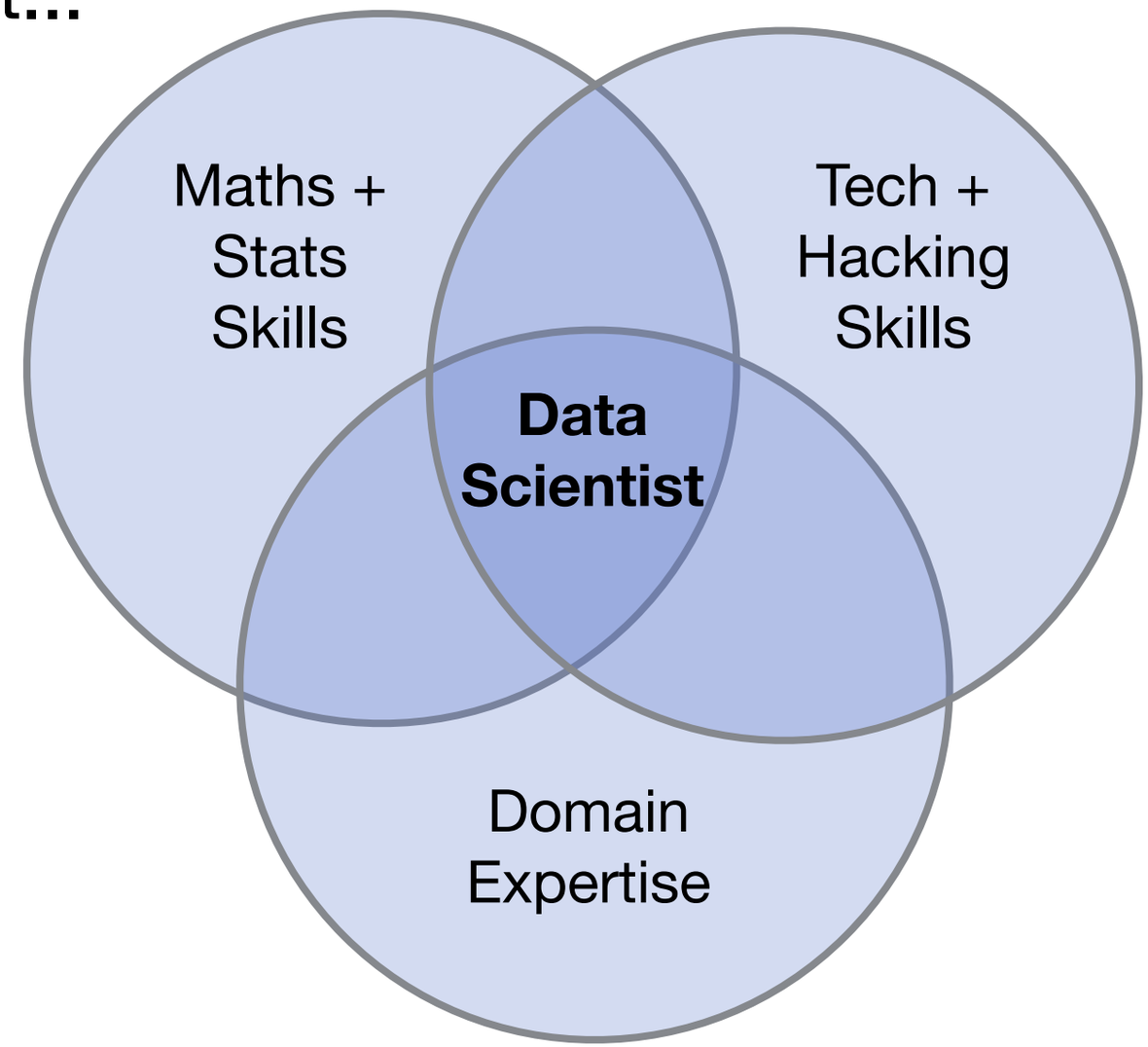

infodiagram.com

# What is Data Science?

Big Data offers great potential, but...

How can we sift through massive amounts of noisy data to capture useful insights?

How can we quantify, interpret, and communicate those insights in a useful way?




*"A Data Scientist's real job is storytelling…*
*Data gives you the what, but humans know the why."*

- Harvard Business Review, March 2013

# What is Data Science?


## USING ARTIFICIAL INTELLIGENCE TO RUN YOUR BEST MARATHON

JOHN × OCTOBER 26, 2017

ARTIFICIAL INTELLIGENCE CUTTING EDGE IRELAND SPORT

Professor Barry Smyth is a data scientist and AI researcher working in the SFI funded Insight Centre for Data Analytics. He also runs marathons on the side. A few years ago, Smyth started to combine these interests by using his data science skills to collect and analyse race data from millions of marathon runners around the world. Now, what started as a hobby, has become part of his day job.

irishtechnews.ie

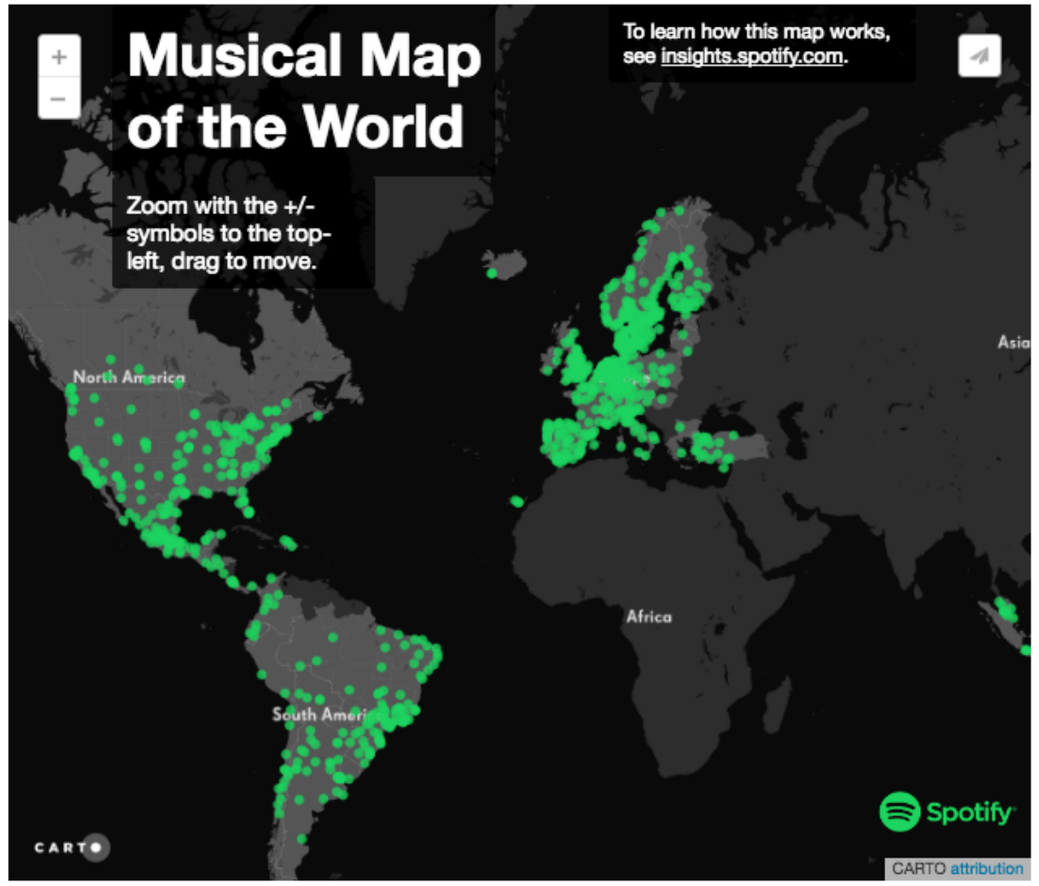
Spotify Insights

## Musical Map of the World 2.0

Posted on December 7, 2016 by Eliot Van Buskirk

Music is a universal language — and maybe one with more variants than any other. To understand these nuances, let's step back and see and hear how people around the world (cities and countries) listen to music differently, what they share the most, and more.

Click a country or city to hear its music (instructions below; full-screen version).




insights.spotify.com

# What is Data Science?

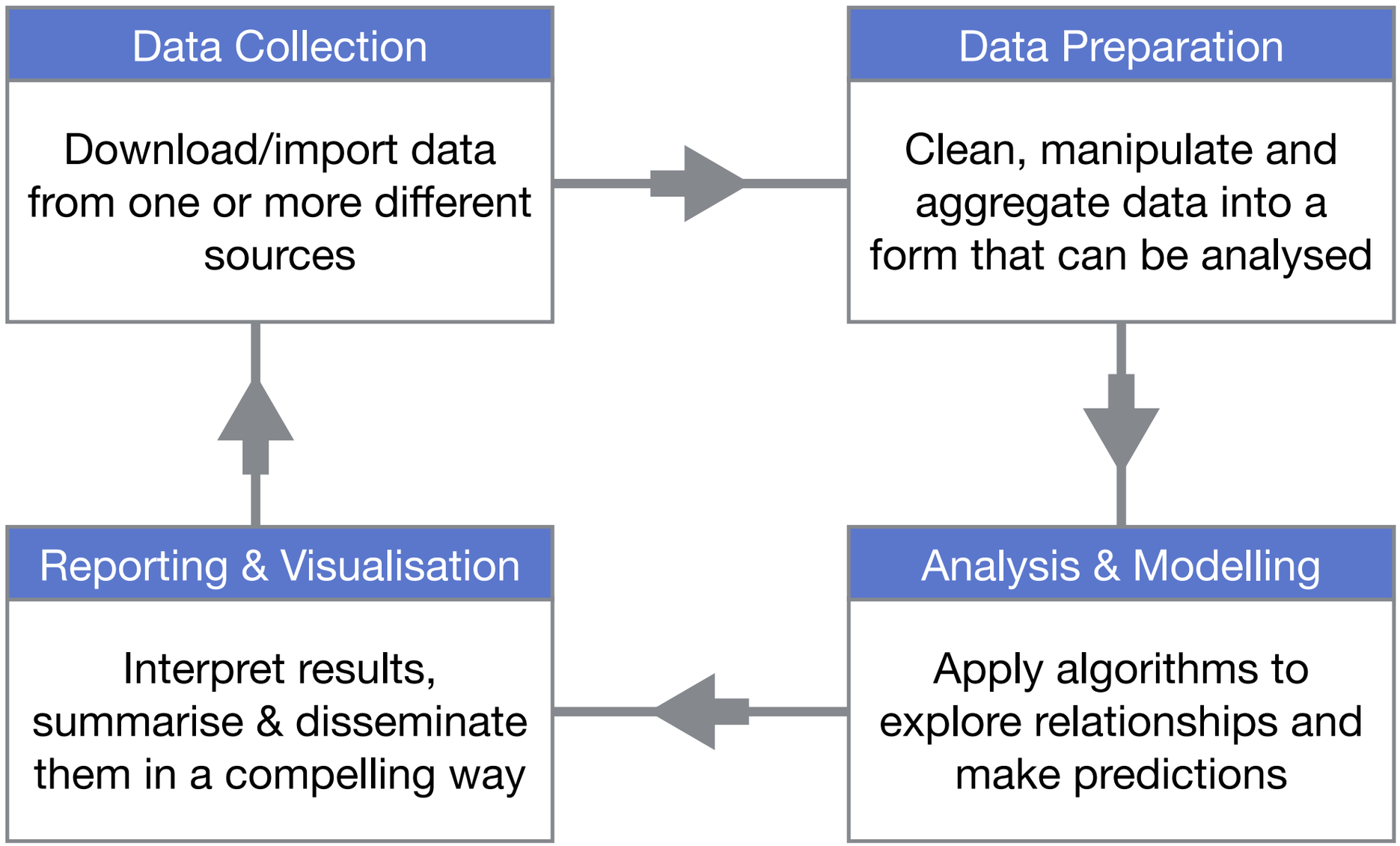
- Data Science: Involves principles, processes, and methods for identifying and understanding phenomena via the automated or semi-automated analysis of data.
- Data Mining: Extraction of knowledge from data, using algorithms that incorporate these principles.

Examples of common data science tasks…

- Prediction: Identify customers likely to move to a competitor
- Regression: Forecast revenue based on historic data
- Clustering: Segment customer base into meaningful groups
- Anomaly Detection: Identify fraudulent customer behaviour
- Visualisation: Support and explain the above tasks

# Basic Data Science Pipeline

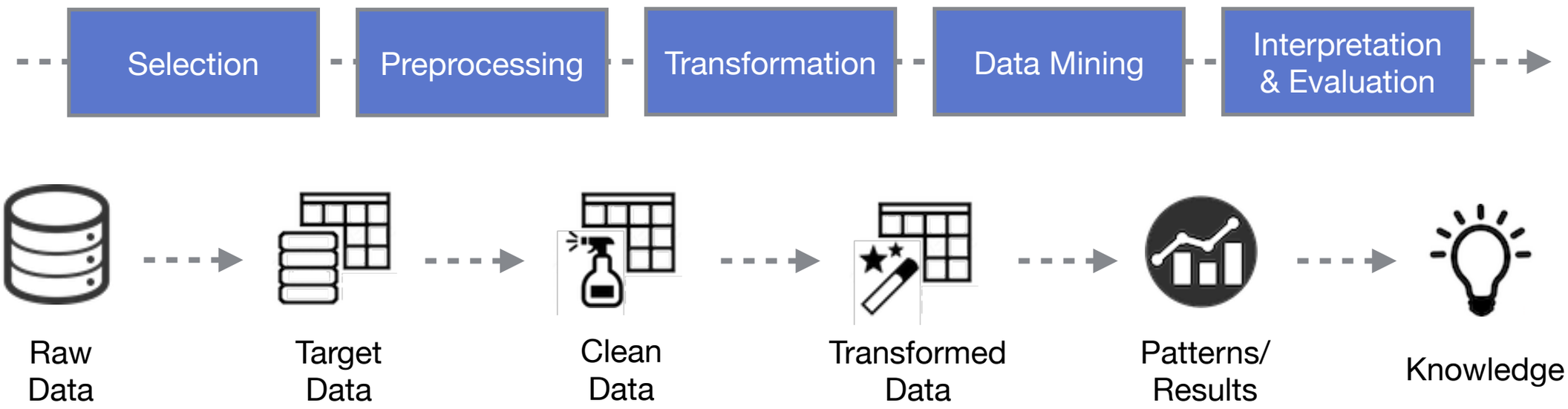
- Data analysis projects typically involve iterating through a pipeline of steps. A simple data science pipeline consists of...



# Knowledge Discovery in Databases (KDD)

Raw data ≠ valuable knowledge or actionable insights

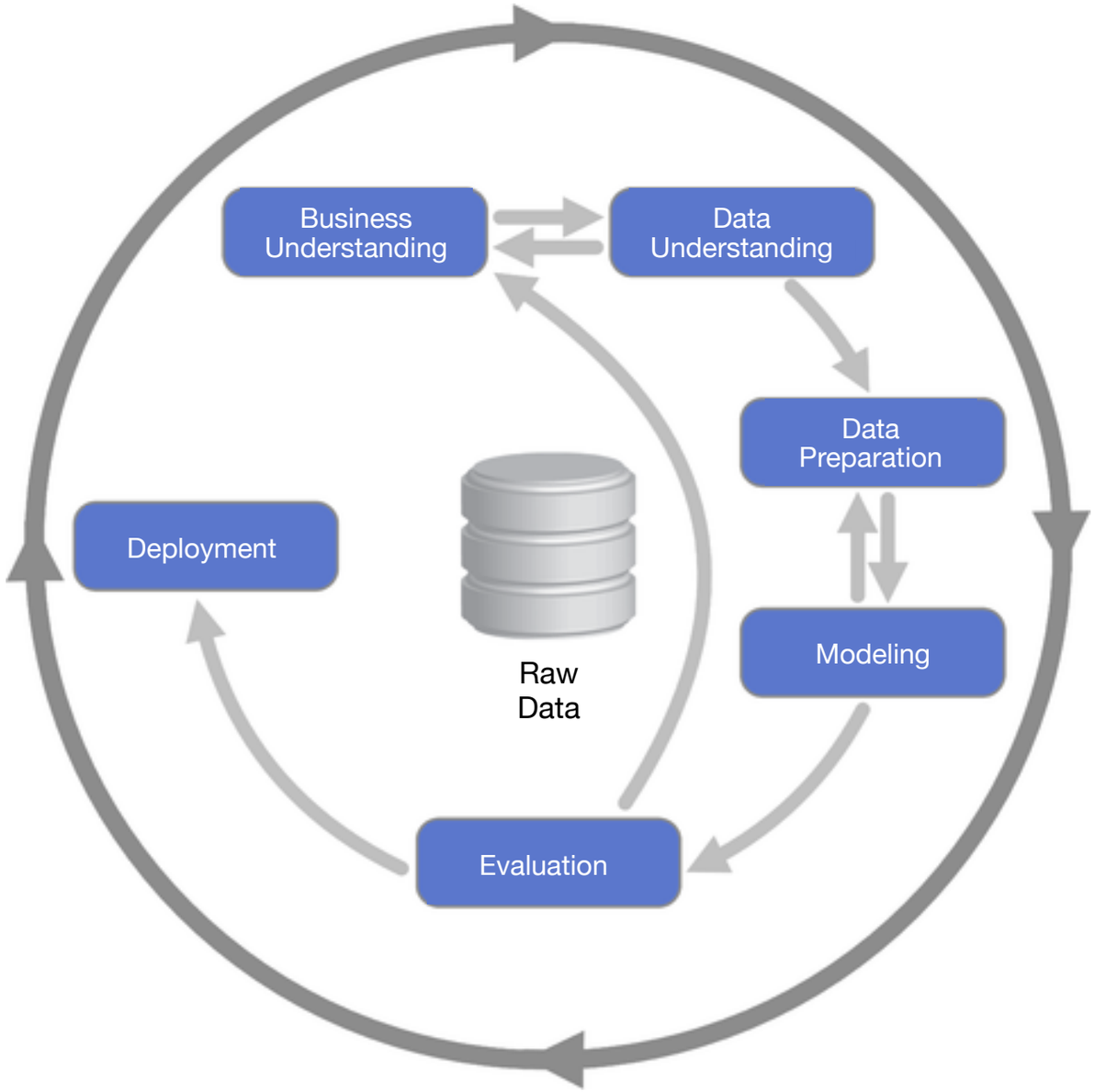
- KDD Process: Goal is to uncover useful knowledge hidden in large lower-level databases. This is achieved by applying data mining algorithms to identify and extract knowledge.




- Human interaction is involved throughout the process. Final step is the interpretation and documentation of results, translating knowledge into a form understandable by the end user.

# CRISP-DM Model

- Cross Industry Standard Process for Data Mining introduced a standard model for the lifecycle of commercial data mining projects.
- Designed to convert real-world business problems in data mining solutions.
- Provides a structured approach to planning a project.
- Emphasises the iterative nature of the data mining process.



# CRISP-DM Model

- Different stages of the data analytics project lifecycle:
  1. Business Understanding: What is the business problem?
  2. Data Understanding: What is the data required to solve the business problem?
  3. Data Preparation: Where is the data, how should it be collected, transformed, and stored?
  4. Modeling: What data mining algorithms should be used to solve the business problem, given the data available?
  5. Evaluation: How well do the algorithms work?
  6. Deployment: How can the analytics results/model be integrated into the current workflow for the organisation?
- In this module we will primarily focus on Steps 3-5: Data Preparation, Modeling, and Evaluation.

# Business Understanding

- Firstly, a number of fundamental tasks need to be completed to convert a business problem into an analytics solution:
  - Q. What is the business problem?
  - Q. What are the goals that the customer really wants to achieve?
  - Q. How does the business currently work?
  - Q. In what ways could a data analytics solution help to solve the business problem?
- Next, we need to confirm that an analytics solution is feasible in this scenario:
  - Q. Is the data required by the solution available to us? Could it be made available?
  - Q. What is the capacity of the business to actually use the results and insights that the analytics solution will provide?

# Data Understanding

- Data understanding: Start with initial data collection. Proceed with activities that enable us to become familiar with the data, identify data quality problems, discover first insights into the data, detect interesting subsets.
- In CRISP-DM, the basic structure used to represent datasets is the analytics base table (ABT). Each row represents a case, and is composed of a set of descriptive features and a target feature.
- A key task prior to modelling is building the ABT.

| | Descriptive Features | | | | | | | | | Target Feature |
|---|---|---|---|---|---|---|---|---|---|---|
| Cases | ----- | ----- | ----- | ----- | ----- | ----- | ----- | ----- | ----- | ----- |
| | ----- | ----- | ----- | ----- | ----- | ----- | ----- | ----- | ----- | ----- |
| | ----- | ----- | ----- | ----- | ----- | ----- | ----- | ----- | ----- | ----- |
| | ----- | ----- | ----- | ----- | ----- | ----- | ----- | ----- | ----- | ----- |

# Data Understanding: Example

**Car insurance fraud prediction:** Table of sample descriptive features with numeric, ordinal, date, categorical, and text types.

| ID | Name | Date of Birth | Country | Salary | Credit Rating |
|---|---|---|---|---|---|
| 8586 | Dermot Curren | 22/05/1988 | Belgium | 25,000 | C |
| 5265 | Joan Flanagan | 04/06/1974 | Ireland | 65,000 | AA |
| 4309 | Mila Weber | 04/06/1983 | Germany | 56,750 | A |
| 6194 | Sheila Shannon | 11/11/1997 | Ireland | 37,000 | B |
| 2383 | Emer McDermott | 01/12/2000 | Ireland | 42,500 | AA |
| 8194 | Lucas Durand | 17/09/1993 | France | 45,000 | B |

Numeric Textual Date Categorical Numeric Ordinal

# Creating Features

Good input features are essential for analytics. Creating appropriate descriptive features can be difficult and is often the most time-consuming part of developing an analytics solution.

- Three basic practical considerations are important when designing features:
  1. Data availability: Do we have access to the data required to create the feature?
  2. Timing: When will the data required for the feature be available?
  3. Longevity: How long will the data used in a feature stay relevant? Will it quickly become "stale" or inaccurate?
- Two further considerations have become increasingly important:
  1. What is the financial cost of producing data for new features?
  2. What are the privacy or ethical considerations?

# Creating Features

- Cost-sensitive features: Sometimes we need to factor in the financial cost of producing new features and acquiring new data.
- Can the expense be justified? Is there a fixed budget available?
- Example: In clinical, pharmaceutical or manufacturing applications, acquiring new data may require considerable time and resources to set up and run experiments.

- Privacy-sensitive features: Some features might be useful from a purely algorithmic perspective, but it might not be appropriate to include them due to legal or ethical concerns.
- Sensitive features can range from gender or race, to medical history or conditions, to unique identifiers (e.g. passport ID).
- Example: In certain settings, gender information cannot be used in insurance pricing models.

# Features: Example

- Here we are using six features to represent each customer.
  - Are these relevant to the task we are trying to perform?
  - Do any more features need to be added from other sources?
  - Should any misleading or "noisy" features be removed?
  - Do we need to alter, normalise or "engineer" some or all of these features in any way?

| ID | Name | Date of Birth | Country | Salary | Credit Rating |
|---|---|---|---|---|---|
| 8586 | Dermot Curren | 22/05/1988 | Belgium | 25,000 | C |
| 5265 | Joan Flanagan | 04/06/1974 | Ireland | 65,000 | AA |
| 4309 | Mila Weber | 04/06/1983 | Germany | 56,750 | A |
| 6194 | Sheila Shannon | 11/11/1997 | Ireland | 37,000 | B |
| 2383 | Emer McDermott | 01/12/2000 | Ireland | 42,500 | AA |
| 8194 | Lucas Durand | 17/09/1993 | France | 45,000 | B |

# Feature Engineering

- Features in the ABT can be of two types:
  - Raw features: These come directly from the original data.
  - Derived features: Do not exist in the original data, but are constructed in some way from the raw data.
- Feature engineering: Using domain knowledge to transform raw features into new features that better represent the underlying problem, and lead to better results in downstream tasks.
- Common feature engineering tasks often involve:
  - Remove unnecessary and/or redundant features
  - Modify feature data types - e.g. from categorical to numeric
  - Combine two or more existing features
  - Transform existing features
  - Create new features

# Feature Engineering

- Feature engineering itself often involves its own iterative process which feeds back into the data analytics pipeline...

1. Brainstorm features: Understand the business problem, explore the data, study previous solutions from other tasks/domains. Identify availability, timing, and longevity constraints.
2. Devise features: Manually and/or automatically create raw and/or derived representative features from the data.
3. Select features: Identify subset of all possible features, which provide one or more “views” for our models to operate upon.
4. Evaluate models: Estimate how effective the features were for the analytics problem that we are trying to solve.

- While some automated methods try to avoid manual feature engineering (e.g. deep learning), in most cases we still require considerable manual effort, working closely with domain experts.

# Example: Feature Engineering

- **Analysing student performance**: Is student activity in an online course management system predictive of good grades?

| *Activity Timestamp* | *IP Address* |
|---|---|
| 14/01/2016 10:12 | 96.37.123.145 |
| 29/01/2016 21:05 | 137.43.230.43 |
| 03/02/2016 20:34 | 92.44.542.331 |
| 06/02/2016 14:01 | 137.43.145.53 |

Raw data (System activity log)

- Can we create better features from the activity timestamp?
  - Was activity daytime or nighttime?
  - Was activity weekday or weekend?
  - Was activity during or outside semester?
- Can we create better features from IP?
  - Was activity on or off campus?

| *is_daytime* | *is_weekday* | *is_insemester* | *is_oncampus* |
|---|---|---|---|
| True | True | False | False |
| False | True | True | True |
| False | True | True | False |
| True | False | True | True |

New data representation with 4 derived features

# Feature Vectors

- Feature vectors: Many machine learning algorithms require the input data to be represented as real-numbered vectors of equal length.
- The number of features or dimensions in each vector is often referred to as its dimensionality.
- Integer and real-valued data do not need a special encoding because they can be represented directly by a numeric value.

| ID | Name | Date of Birth | Gender | Credit Rating | Country | Salary |
|---|---|---|---|---|---|---|
| 0034 | Brian | 22/05/78 | male | aa | ireland | 67,000 |
| 0175 | Mary | 04/06/45 | female | c | france | 65,000 |
| 0456 | Sinead | 29/02/82 | female | b | ireland | 112,000 |
| 0687 | Paul | 11/11/67 | male | a | usa | 34,000 |
| 0982 | Donald | 01/12/75 | male | b | australia | 88,000 |
| 1103 | Agnes | 17/09/76 | female | aa | sweden | 154,000 |

Raw Data

| feature1 | feature2 | ... |
|---|---|---|
| 67000.0 | ... | ... |
| 65000.0 | | |
| 112000.0 | | |
| 34000.0 | | |
| 88000.0 | | |
| 154000.0 | | |

Feature Vectors

# Categorical Features

- Categorical features take a discrete set of possible values, typically represented as strings. Not suitable for algorithm input.
- One hot encoding: process by which categorical variables are converted into a numeric form that can be passed to an algorithm.
- The values in a feature are split into multiple binary columns, one for each possible category.

| ID | Credit Rating |
|---|---|
| 0034 | aa |
| 0175 | c |
| 0456 | b |
| 0687 | a |
| 0982 | b |
| 1103 | aa |

Raw Data

| aa | a | b | c |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 |
| 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 |

Encoded Data

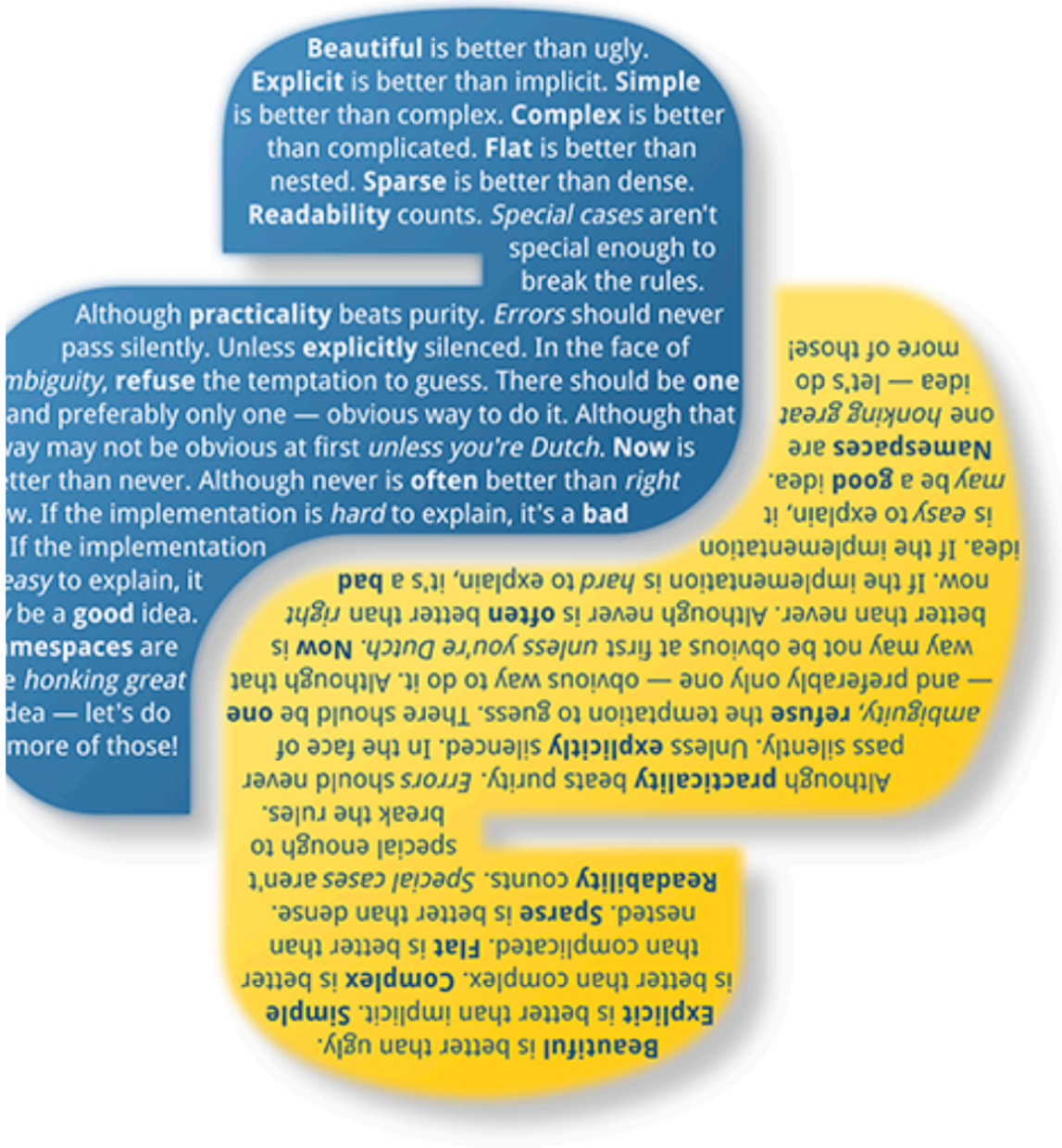
# Common Problems in Data Science

- Problems can arise at many different levels in a practical data science project...
  - Initial assumptions about the data may be incorrect.
  - Inherent noise and unreliability in the data source.
  - Process of collecting the data yields a dataset that is biased, incomplete or non-representative.
  - Inappropriate representation, algorithms or parameters used during the modelling process.
  - Evaluation of the modelling process is flawed or incomplete.
  - Interpretation of the results flawed or biased - e.g. pareidolia or "patterns in the clouds".
  - Failure to communicate the results to stakeholders.
  - ...

# Data Science in Python

pandas

matplotlib




scikit

Python

# References


- J. D. Kelleher, B. Mac Namee, A. D'Arcy. "Fundamentals of Machine Learning for Predictive Data Analytics", 2015.
- F. Provost, T. Fawcett. "Data Science for Business", 2013.
- U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From data mining to knowledge discovery in databases." AI magazine, 1996.
- P. Guo. "Software Tools to Facilitate Research Programming", PhD Thesis, Stanford, 2012.
- W. Yan "Feature Engineering for PHM Applications", PHM 2015.
- J. Brownlee "Discover Feature Engineering, How to Engineer Features and How to Get Good at It", 2014.