COMP41680

Introduction to Data Science

Derek Greene

UCD School of Computer Science

Spring 2017

Overview

- The Big Data Era
- What is Data Science?
- Basic Data Science Pipeline
- Knowledge Discovery in Databases
- The CRISP-DM Model
  - Business Understanding
  - Data Understanding
  - Feature Engineering

The Big Data Era

In the last decade we have witnessed an explosion in the production and availability of digital data. 90% of the world's data was created over the last 2 years, and by 2020 data will increase by > 4,300%

varonis.com

The Big Data Era

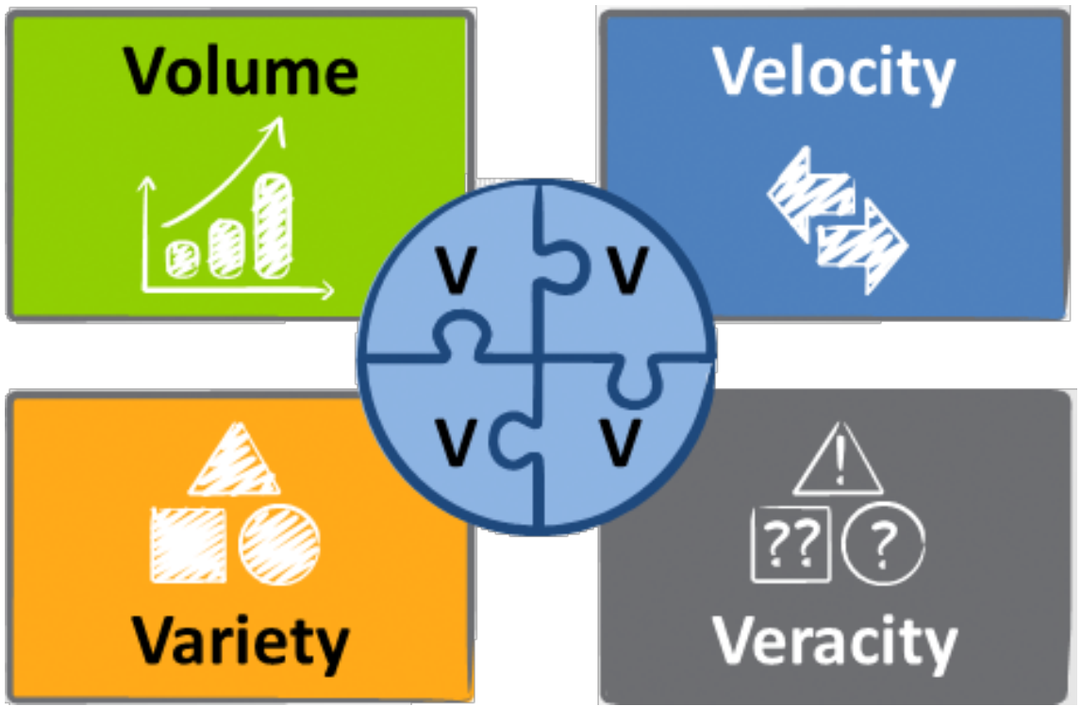
Big Data problems are usually characterised by…

Volume: Huge volumes of data - terabytes or more

Velocity: Continuous streams of data arriving in real time

Variety: Many different types of data - numeric data, structured text, unstructured text, images, video, audio, network data

Veracity: Uncertain if data is reliable, noisy or incorrect



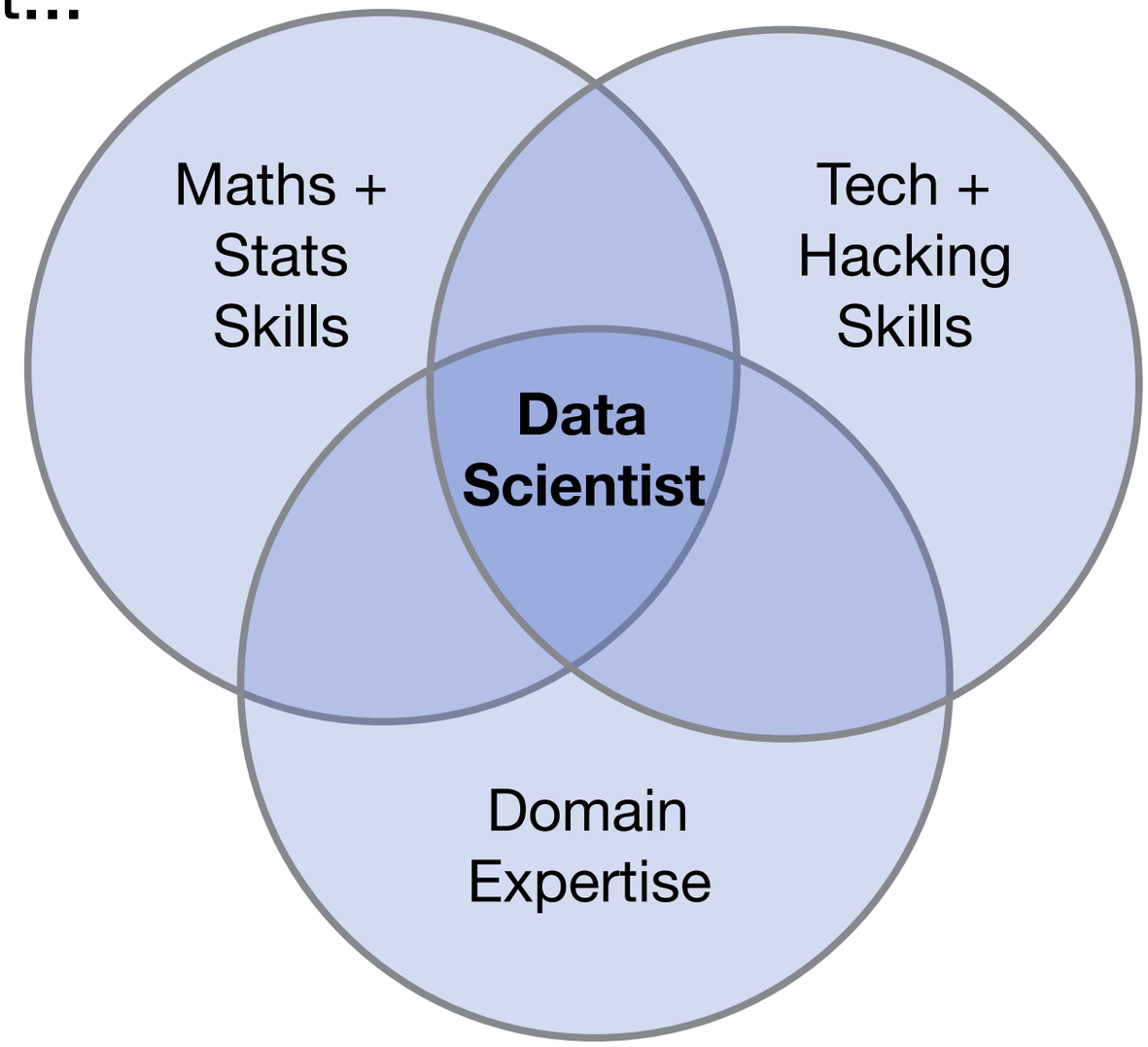
infodiagram.com

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?

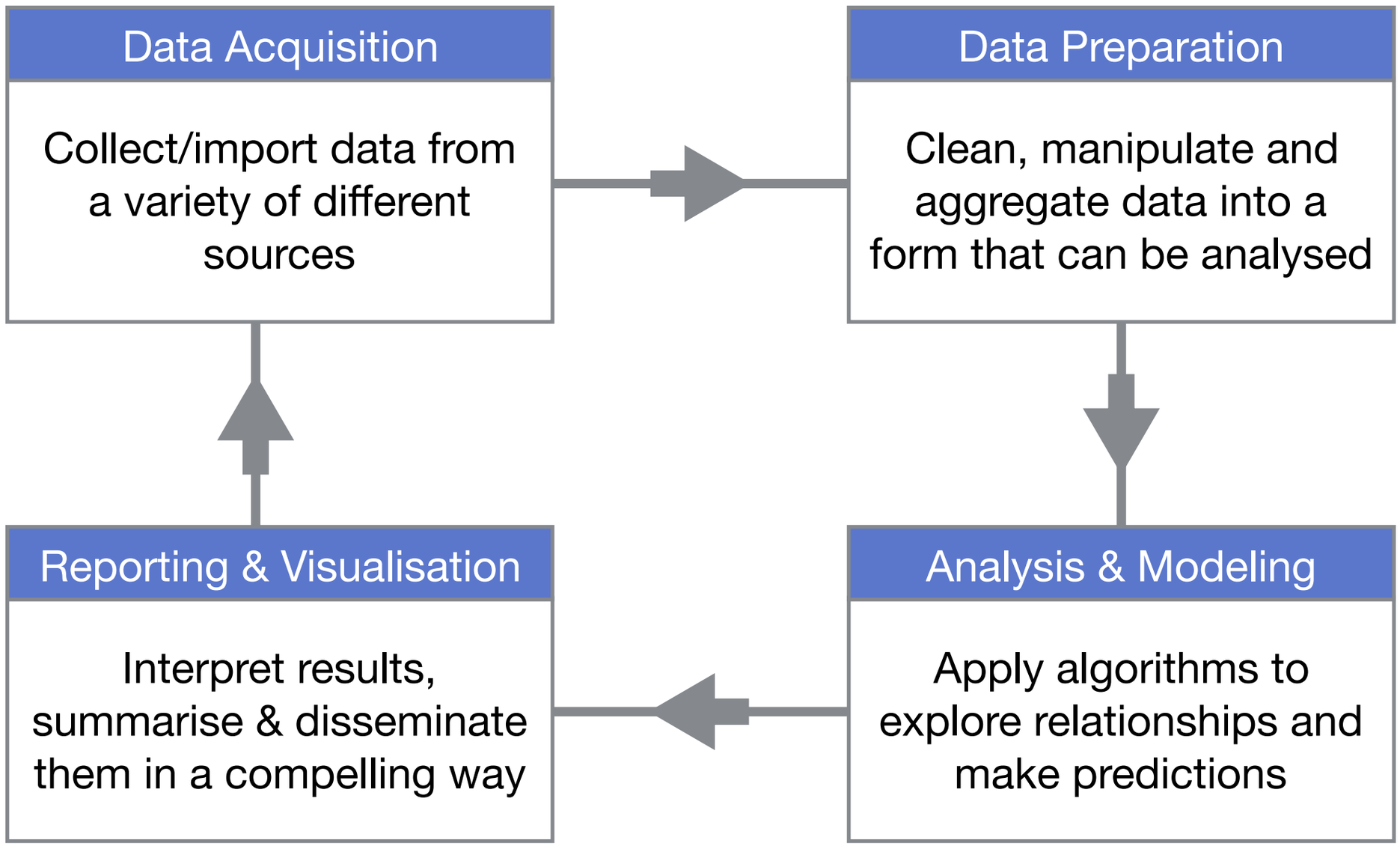
- Data Science: Involves principles, processes, and methods for identifying and understanding phenomena via the automated or semi-automated analysis of data.
- Data Mining: Extraction of knowledge from data, using algorithms that incorporate these principles.

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

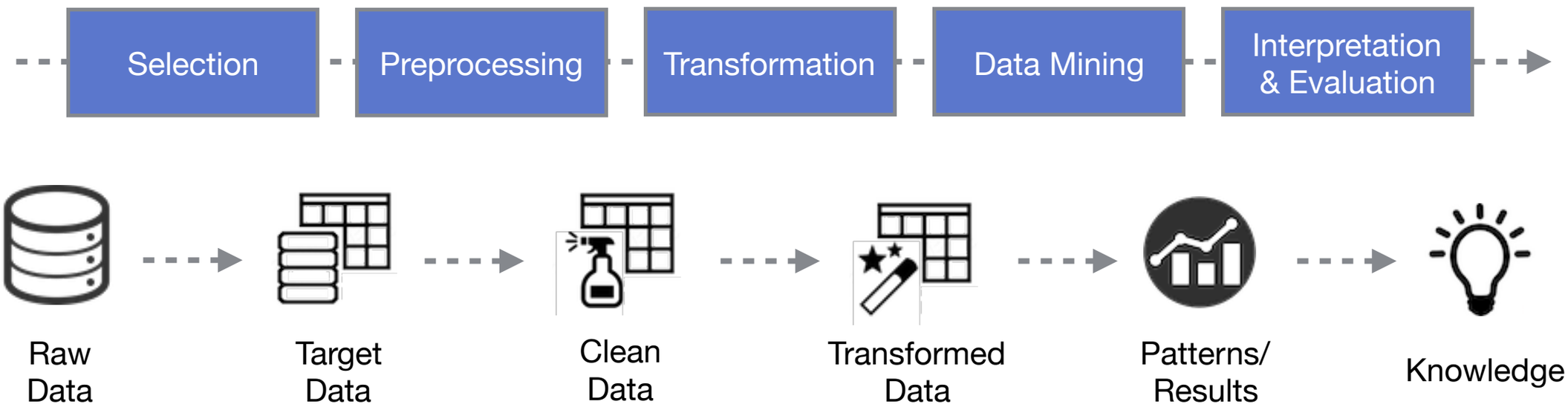
- Data analysis projects typically involve iterating through a pipeline of steps. A simple data science pipeline consists of...



Knowledge Discovery in Databases (KDD)

Raw data ≠ valuable knowledge or actionable insights

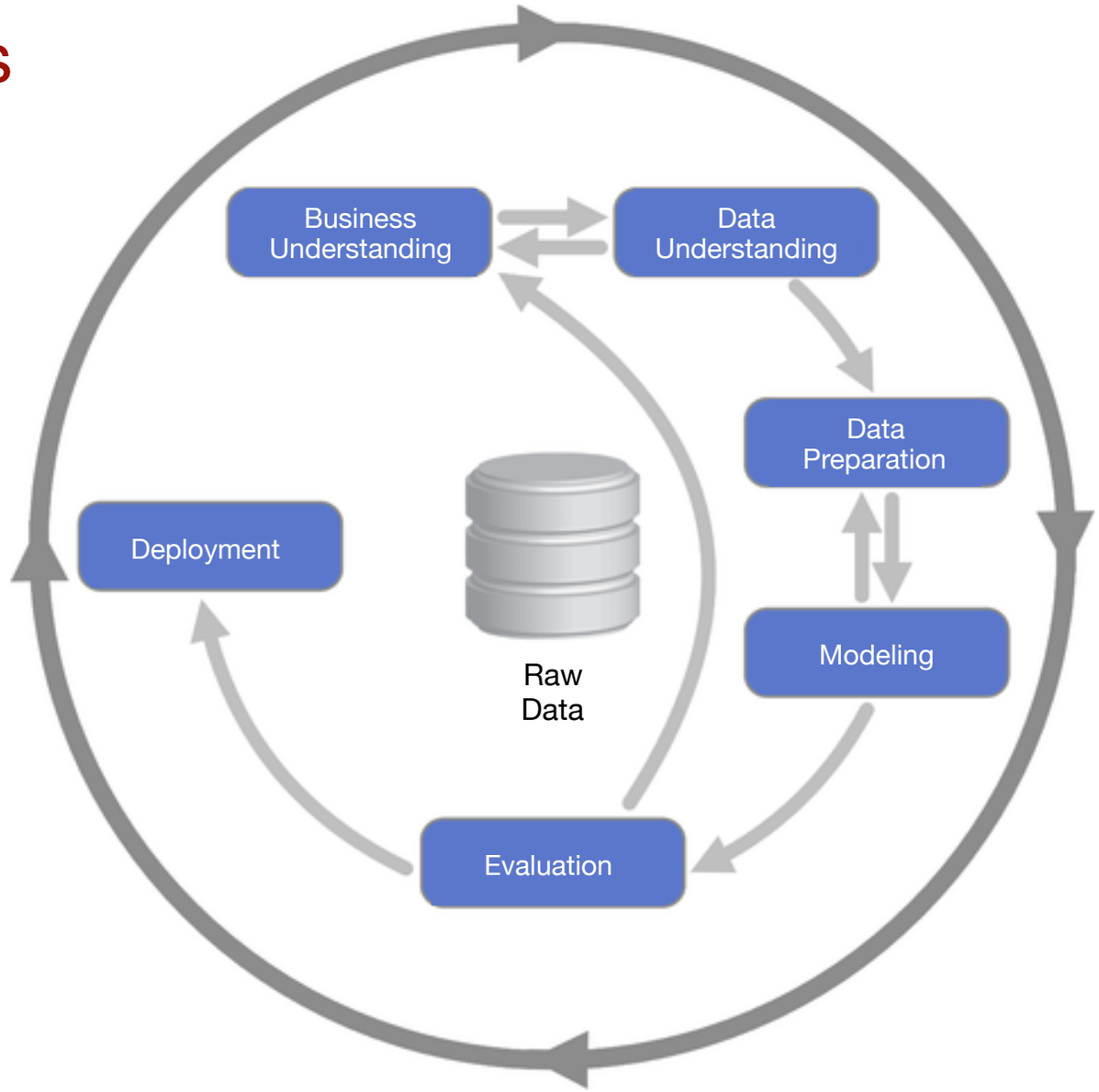
- KDD Process: Goal is to uncover useful knowledge hidden in large lower-level databases. This is achieved by applying data mining algorithms to identify and extract knowledge.



- 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?

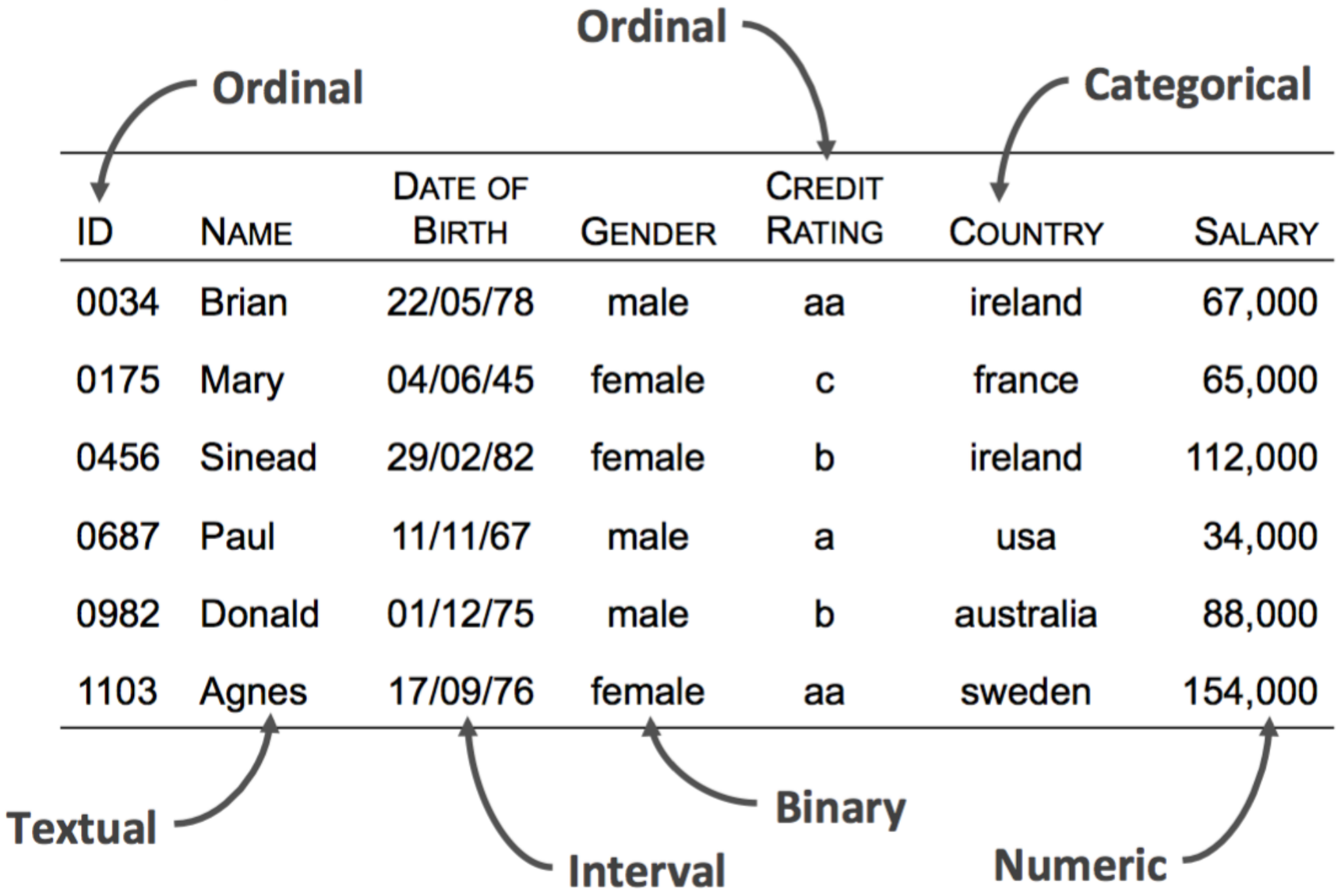
Data Understanding

- 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 modeling is building the ABT.

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

Data Understanding: Example

Car insurance fraud prediction: Table of sample descriptive features illustrating numeric, binary, ordinal, interval, categorical, and text feature types:



Kelleher et al.

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 data considerations are particularly 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?
- Also need to factor in the cost of producing new features.
  e.g. in clinical, pharma or manufacturing scenarios, obtaining data may require considerable time and resources to run experiments.

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: Process of transforming raw features into new features that better represent the underlying problem, and lead to better models.
- Common feature engineering tasks often involve:
  - Remove unnecessary and/or redundant variables
  - Modify variable 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

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.