COMP 499 Introduction to Data Analytics

Lecture 1 — Data Analytics

Greg Butler

Data Science Research Centre

and

Centre for Structural and Functional Genomics

and

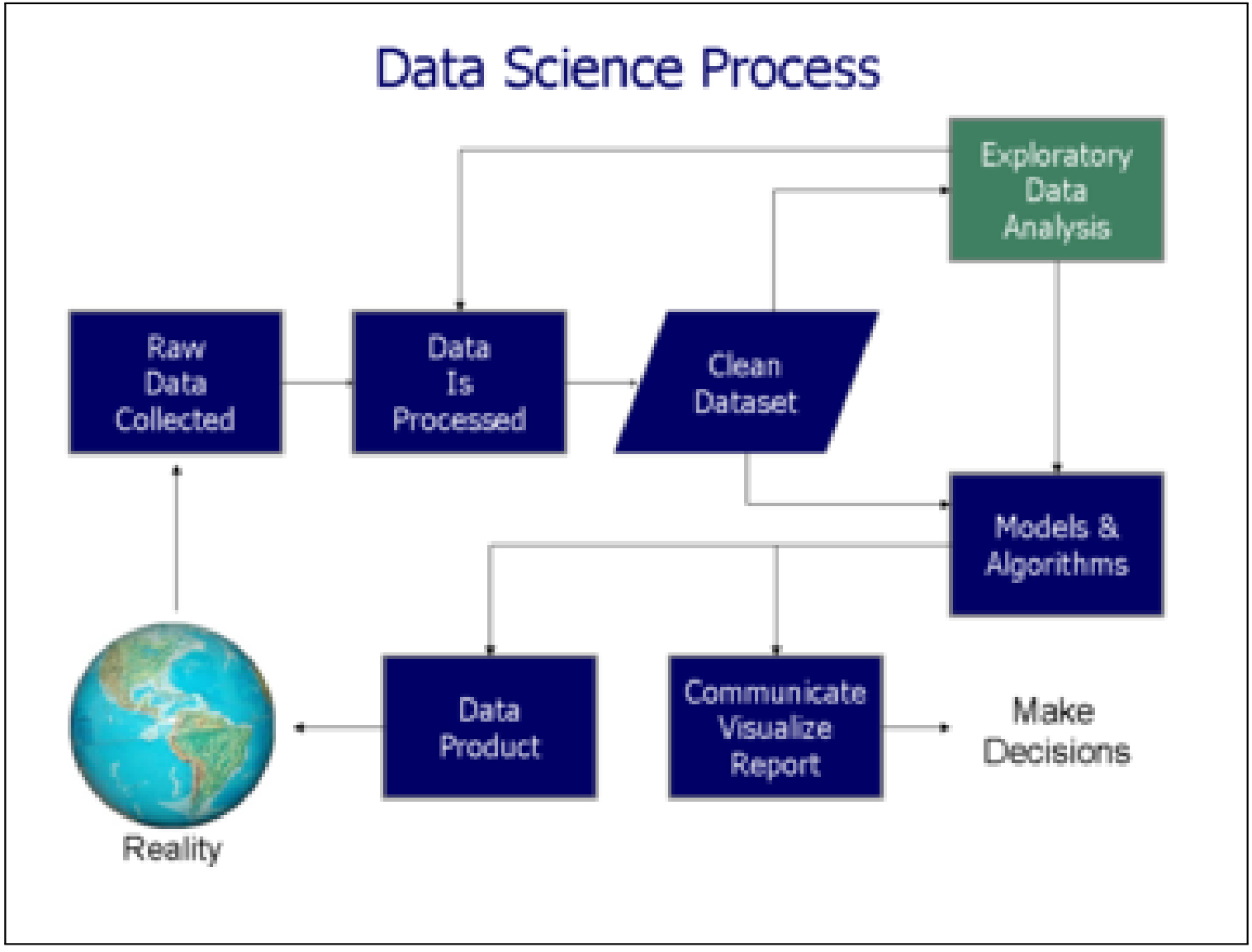
Computer Science and Software Engineering
Concordia University, Montreal, Canada

gregb@cs.concordia.ca

Overview of Lecture

1. Data Analytics
   - Data Wrangling
   - Exploratory Data Analysis
   - Modeling
   - Story telling

Data Analytics



wikipedia

Data Analytics — Example from wikipedia

Find the variables which best predict the tip given to the waiter.

The variables available in the data collected:

- the tip amount,
- total bill,
- payer gender,
- smoking/non-smoking section,
- time of day,
- day of the week, and
- size of the party

The approach is to fit a regression model to predict the tip rate. The fitted model is

- $tip\_rate = 0.18 - 0.01 \times party\_size$

if size of the dining party increases by one (leading to a higher bill), the tip rate will decrease by 1%.

Data Analytics

Steps in Data Analytics

- Setting Questions
- Data Wrangling
- Exploratory Data Analysis
- Modeling
- Story telling

Iterative process

The process is rarely linear.
Each step can push data scientist to revisit methods, techniques ... or reconsider whether the original question was the right one?
And the final answer simply sparks more questions!

Data Analytics

Hypothesis-driven Experimental Design and Analysis

Not exploratory.

You have a single, specific hypothesis to accept or reject.

Steps

- Set null hypothesis H_0 and alternative hypothesis H_1
- Design experiment to collect data, and
- Design analysis of experimental data to accept/reject hypothesis
- Determine *statistical power* of experiment
  Do you have enough data points?
- Do experiment, do analysis, accept/reject hypothesis

Data Analytics: Setting Questions

Ask an Interesting Question

Steps

- Is there a business goal to achieve?
- Some object of scientific interest that would be helpful to discover?
- What parameters would the ideal answer fulfill?

Data Analytics: Data Wrangling

Design a Data Collection Program

- Establish whether or not the data exists in the real world and is relevant to the question
- Devise a collection scheme to acquire it
  Logistical considerations? Cost? Privacy issues?
- Coordinate with departments or agencies needed for collection program liaison

Collect and Review the Data

- Store the incoming data to allow modeling and reporting
- Join data from multiple sources in relevant & logical manner
- Check for anomalies or unusual patterns
  - Caused by the collection process?
  - Inherent to topic of investigation?
  - Correct them, or develop new collection scheme?

Data Analytics: Data Wrangling

Data Wrangling or Data Munging

Bring skills and intuition to bear ...
to take messy, incoherent information ...
and shuffle it into clean, accessible sets

"Munging" the Data

- Select your tools to comb through raw
- Store the munged data as a fresh data set, **or**
- use programmatic pre-processing for each subsequent query

Data Analytics: Exploratory Data Analysis

Exploratory Data Analysis

Learn about the properties of the data

Steps

- Descriptive statistics: mean/median and variance, quantiles, outliers
- Correlation
- Fitting curves and distributions
- Dimension reduction
- Clustering

Data Analytics: Modeling

Modeling

the fun stuff of getting *"meaning"* from a clean data set

Steps

- Build a data model to fit the question
- Validate the model against the actual collected data
- Perform the necessary statistical analyses
- Machine-learning or recursive analysis
- Regression testing and other classical statistical analysis techniques
- Compare results against other techniques or sources

Data Analytics: Story telling

Visualize and Communicate the Results

The most challenging part of the data scientist's job is taking the results of the investigation and presenting them to the public or internal consumers of information in a way that makes sense and can be easily communicated.

Steps

- Graph or chart the information
- Tell a story to fit the results: Interpret the data to describe the real-world sources in a plausible manner
- Assist decision-makers in using the results to drive their decisions