UNIVERSITY OF EXETER

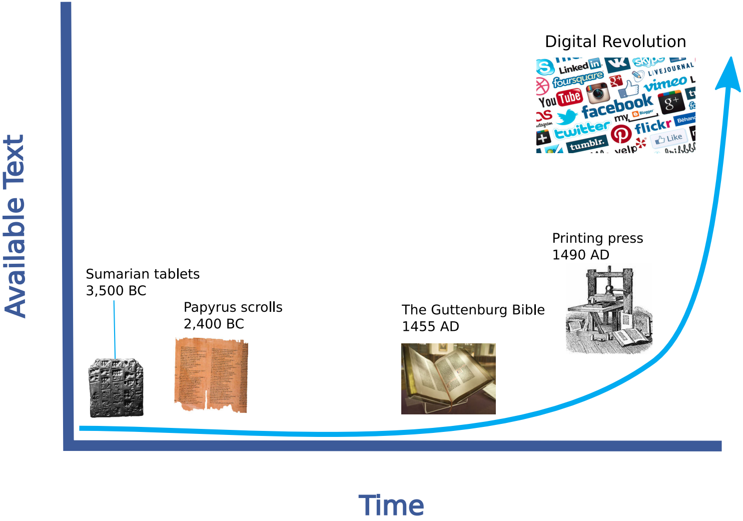
Q-Step

Text as Data

A guided tour

Dr. Travis G. Coan
T.Coan@exeter.ac.uk

Text as data



Digital Revolution

Printing press
1490 AD

The Guttenburg Bible
1455 AD

Sumarian tablets
3,500 BC

Papyrus scrolls
2,400 BC

Available Text

Time

Language is the medium for politics and policy

> *Language is the medium for politics and political conflict. Candidates debate and state policy positions during a campaign. Once elected, representatives write and debate legislation. After laws are passed, bureaucrats solicit comments before they issue regulations. Nations regularly negotiate and then sign peace treaties, with language that signals the motivations and relative power of the countries involved. News reports document the day-to-day affairs of international relations that provide a detailed picture of conflict and cooperation. Individual candidates and political parties articulate their views through party platforms and manifestos. Terrorist groups even reveal their preferences and goals through recruiting materials, magazines, and public statements.*
>
> *(Grimmer and Stewart 2013)*

The **problem**: Too much text to read!

© 2012 Ted Goff

"Here's a list of 100,000 warehouses full of data. I'd like you to condense them down to one meaningful warehouse."

Topics for the morning session

We will cover the following topics:

1. Day 1: Processing text & building lexicons
2. Day 2: Supervised learning for text classification
3. Day 3: Topic modelling
4. Day 4: Word embeddings, language models, and semantic similarity between texts

Computer-**assisted** content analysis

> *We emphasize that the complexity of language implies that automated content analysis methods will never replace careful and close reading of texts. Rather, the methods that we profile here are best thought of as* **amplifying** *and* **augmenting** *careful reading and thoughtful analysis. Further, automated content methods are* **incorrect** *models of language. This means that the performance of any one method on a new data set cannot be guaranteed, and therefore validation is essential when applying automated content methods.*
>
> *(Grimmer and Stewart 2013, emphasis in original)*

Four **principles** of text as data

Grimmer and Stewart (2013) outline four general principles that are absolutely essential to keep in mind when using text as data:

Principle 1: All Quantitative Models of Language Are Wrong—But Some Are Useful.

Principle 2: Quantitative methods for text amplify resources and augment humans.

Principle 3: There is no globally best method for automated text analysis.

Principle 4: Validate, Validate, Validate.

Disclaimer: programming required!

Off-the-shelf software: There are a number of “off-the-self” software solutions for text analysis (e.g., WordStat). However, this software is expensive and lacks flexibility.

Open-source solutions: There are a number of excellent open-source alternatives for text mining.

1. **R**: There are a number of different options to choose from, but I highly recommend taking a look at quanteda.
2. **Python**: The remainder of this class!

Why Python?

Why Python?

1. Any language named after the Monty Python is worth giving a try!
2. It's easy
3. It's fast enough (and pretty easy to speed up)
4. Large user community
5. It is becoming the industry standard for scientific computing.