

Data

Brian Keegan, Ph.D.

Northeastern University

HGSE T553, March 25, 2014

@bkeegan

Dimensions

Quantitative ... Qualitative

Big ... Small

Structured ... Unstructured

Fast ... Slow

Found ... Made

Precise ... Representative

Open ... Closed

Quantitative ... Qualitative

Tribal distinction of little use

Quantitative data demands interpretation

Qualitative data must be abstracted

Stands in for more meaningful distinctions

Big ... Small

Volume : “Does it fit in Excel?”

Different strategies for analysis and management

Preponderance of small datasets

New insights from joining big & small together

Structured ... Unstructured

Variety : numbers, text, images, documents

Cleaned versus raw data

Always takes more time than you expect

Uncritical assumptions abound

Fast ... Slow

Velocity : Analyzing fast and slow

Newest is not best → Past was not so different

Fast : Exploratory tests of a proto-hypothesis

Slow : Making sense of anomalies

Found ... Made*

Observe and create new facts about the world

Reflection on research questions and methods

Critical of assumptions and generating processes

Others will want to use your data and credit you

* [Sean Taylor](#): “Real scientists make their own data”

Precise ... Representative

TINSTAAFL

Detailed records expensive and difficult

Large-scale records noisy and superficial

If this were solved ... everyone would be doing it

Open ... Closed

Replicability fundamental to scientific method

Open access to data, methods, and software

Wonderful ecosystems of people and tools

Tough issues around privacy and anonymization

Thank you!

www.brianckeegan.com

[@bkeegan](https://twitter.com/bkeegan)

github.com/brianckeegan