A Glossary for Quantitative Research in Social Science

A Working Paper for Uniting Language Across Disciplines

This series of white papers is funded by National Science Foundation awards #1934925 and #1934494 as a collaboration between Georgetown University and University of Michigan. To learn more about The Future of Quantitative Research in Social Science research project, visit www.smrconverge.org.

I. Introduction

Adapted from "Disciplinary Babble: Uniting Language Across Disciplines" by Pam Davis-Kean and Leticia Bode

As a part of our grant, we are hosting a series of meetings about to discuss key issues associated with using social media data in social science research. For each meeting we have invited a fantastic group of experts come from all over the country to join us in thinking about how to converge study design across social science and computer science, especially in light of new data sources like social media data.

We've learned a lot, but one major takeaway was how hard it is to talk to one another.

Every discipline has its own language. We spend years – sometimes even decades – being instructed and socialized into a particular way of thinking about the world, with a set of vocabulary to match. This mild brain-washing, which to be fair has many benefits, turns out to be very difficult to counteract.

At our first meeting, we created a board that was available throughout the two-day workshop, where people added words whose meaning they weren't sure about. We called this our glossary, although that is probably a bit too generous, since many of the words have yet to be fully defined. Throughout this process, we identified two main issues that can lead to language confusion.

First, there are words, phrases, acronyms, and terms that some disciplines use and others do not. Some examples of this include algorithmic bias, and endogeneity. When someone uses such a term, people from other disciplines are confused because they don't understand the term. In this case, however, it is relatively easy to solve the problem. The person who doesn't understand can simply ask for clarification, or point out that they don't know the term, and the original person using the term can offer a definition.

Second, and perhaps more complicated, are words and phrases that are used across disciplines but in different ways. Examples of this second type of confusion include active learning, reliability, model, sample, and certainty. In this case, confusion may persist longer because people may not even realize they are using terms in different ways. This can cause conflict when everyone thinks they are on the same page (that is, using the same terms), but are actually thinking about things very differently.

So how do we overcome these challenges and use a common language? To some extent, language confusion may be inherent in developing interdisciplinary or multidisciplinary collaborations. But we do think some approaches make it easier to navigate than others.

First, speakers (or writers, or presenters) can be explicit about what they mean. Rather than relying on disciplinary terms to do some of the explanation, speakers can take a little extra time to make sure a concept or description is clear.

Second, we think it's really important to facilitate an environment where saying "I don't know what that means" or "I'm not sure we're talking about that idea in the same way" are acceptable statements.

Too often in academia, too much emphasis is placed on being right, and being confident in that right-ness. Fostering an environment where admitting confusion or ignorance is not only acceptable but encouraged as a form of intellectual curiosity, and responded to with respect, can help people to better communicate to one another and overcome these challenges. Interdisciplinary teams need to be open to learning new words and alternative definitions for disciplinary words in order to create a common language for a new team and new ideas. Something as simple as a glossary board can be a great reminder that we don't always speak the same language, and that is ok!

The glossary provided in this document will be continually updated with new words and definitions based off of the discussions at our convergence meetings with the goal of establishing a dictionary that allows for easy translation across social science and computational sciences with respect to concepts connected to social media research. We expect that for many words, there will be multiple definitions written by scholars in the same and across different disciplines. While we are sharing a paper version of this document, we will explain how to add definitions and augment the glossary during the convergence meeting.

This glossary was last updated November 4th, 2020.

II. Glossary

Accurate *Social Science*	This term can be used to refer to different dimensions of data quality. In the Federal Statistical System, accuracy refers to the closeness of estimates to their true values. Generally speaking, accuracy usually refers to the lack of bias (i.e. systematic error) which includes both measurement and representation.
Active learning *Computer Science*	Active learning is a paradigm of machine learning where the algorithm is trained on a small set of labelled examples. Based on this, the algorithm selects which unlabeled examples it is least confident on. Next, it requests user labelling for those to retrain itself.
Algorithm *Computer Science*	An algorithm is a set of steps used to calculate a value or solve a problem.
Algorithmic bias *Computer Science*	Algorithmic bias is when a group of people is unfairly and systematically singled out by a computer algorithm, typically one that is used to make decisions that affect peoples' lives. Algorithmic bias has been a focus in court sentencing guidelines, facial recognition software, mortgage lending, and other areas.
Algorithmic confounding *Computer Science*	One cannot view all behavior online as being naturally occurring. Some of it is a result of system design or engineering goals. These design goals can introduce patterns into the data. These patterns are referred to as algorithmic confounders.
Analytics *Computer Science*	Information that results from a systemic or detailed analysis of data that can be used to predict future behavior.

Big Data *Computer Science*	"Big Data" refers to data that is larger than you can manage on your own computer. It is typically generated in an automated or computer-aided fashion and provides information on a massive scale. Examples of big data include information on online usage and behavior (e.g. Google searches, profile clicks), mobile device data, data collected by health or home devices, satellite imagery, surveillance data, reports by citizen journalists, and social media data.
Causal inference *Computer Science*	Causal inference is the process of analyzing and answering the question: whether the given factor(s) is a cause or not for the observed?
Coding *Computer Science & Social Science*	*Computer Science:* Coding is the process of programming software. There are different coding/programming languages. Two popular languages are Python and Java. Programmers code algorithms, i.e. a set of instructions for completing a computational task. *Social Sciences:* Coding is the process of labeling and organizing your qualitative data to identify different themes and the relationships between them.
Computational Social Science *Social Science*	Within computational social science, researchers are analyzing large data sets to answer social science questions. They use both data science and computer science methods to model and analyze the data.
Confidence interval *Social Science*	A confidence interval is a range of values we are fairly sure our true value lies.
Controlled observations *Social Science*	A type of observational study where the conditions are contrived by the researcher. This type of observation may be carried out in a laboratory type situation and because variables are manipulated is said to be high in control.
Data *Computer Science*	Information that is generally collected by observation and (in computer science) stored on a computer.

Data lake *Computer Science*	A "data lake" is a single store of all enterprise data. It includes raw copies of the original data from one or more sources and transformed versions of the data.
Data leakage *Computer Science*	Generally speaking, data leakage occurs when sensitive data is exposed. In machine learning, data leakage occurs when information that is not part of the training data set is used to create the model.
Data shadows *Social Science*	A data shadow (to follow) is the collective body of data that is automatically generated and recorded as we go about our lives rather than intentionally created.
Deliberation Measures *Political Science*	Deliberation refers to the process of thoughtfully weighing options with the intent of making a decision, such as a vote. Social media has become an increasingly popular way to measure political deliberation, for example toxicity.
Descriptive *Psychology*	Analyses of data that do not yield causal inferences about an association but a descriptive portrait of a sample or population on constructs of interest.
Dimensional Reduction *Computer Science*	Dimensional reduction refers to a mathematical approach for reducing the dimensionality (number of variables) of a data set to a smaller number. Standard dimensionality reduction techniques include Principal Component Analysis and Singular Value Decomposition.
Domain *Computer Science*	It is the discipline or subdiscipline the research question is connected to. When we think about this from a computer science perspective, it helps us determine the background knowledge or subdiscipline of knowledge that can be useful for algorithms to understand. For example, a question about election dynamics would have a domain of politics or political science.

Elites *Political Science*	Elites are often studied in political science. They refer to individuals such as journalists and politicians who are able to make their opinions known to a wide audience, or the mass public.
Emotions *Psychology*	Emotions are defined by the American Psychology Association as a complex reaction pattern, involving experiential, behavioral, and physiological elements, by which an individual attempts to deal with a personally significant matter or event. Example emotions of interest with regards to social media include happy and sad.
Endogeneity *Social Science*	The potential that a relation observed between two or more variables may be a function of something outside of those variables. This often includes a common cause (x and y are related because z causes them both).
Event Detection *Computer Science*	Event detection refers to using automated techniques for identifying events from text - as opposed to how events impact behaviors, attitudes, etc.
Exploratory vs. Hypothesis testing *Social Science*	Hypothesis-driven research is based on scientific theories, while exploration is based on a search for discovery backed by few theories or none at all.
Feature vs Variable vs Parameter *Computer Science*	In computer science, a feature is a variable that can be used to train a machine learning model. A variable may be a feature or it may be the outcome the machine learning algorithm is attempting to predict. Parameters have a number of meanings. In statistics, one estimates the parameters of the model. Sometimes that same idea is used in computer science. Sometimes, parameters are the conditions specified at the start of an algorithm. For example, when running the k-means clustering algorithm, k must be defined. k is an example of a parameter that needs to be set or input for the algorithm to run.

Feature Engineering *Computer Science*	Feature engineering focuses on how we construct different variables (e.g. topics or sentiment) from text. For example, we can use opinion mining to determine a stance or position of a user about a topic, e.g. opinion on breastfeeding.
Generalizability *Social Science*	Generalizability refers to the extent to which the results of a study can be applied to a broader population. In the context of social media research, a common critique is that the results should only be interpreted to reflect users on the platform.
Generative Model *Computer Science*	A generative model is one that learns the distribution of each class or category, while a discriminative model models the decision boundary of each class. Many topic models, including LDA, are examples of a generative model.
Granger Causality *Statistics*	A way to measure causality between two time series variables. The Granger causality test is a statistical test that can be used to determine if one time series can be used to predict or forecast another time series.
Graphical model *Computer Science*	Generally, probabilistic graphical models use a graph-based representation as the foundation for encoding a distribution over a multi-dimensional space to expresses the conditional dependence structure between random variables.
Interpretivist *Social Science*	An approach to social science that opposes the positivism of natural science and allows for human interest and interpretation (qualitative).
Latent Dirichlet Allocation (LDA) *Computer Science*	LDA is probabilistic, generative model for identifying topics from documents.

Machine learning *Computer Science*	Machine learning is a subfield of artificial intelligence that aims to teach computers to learn and improve from experience. Machine learning algorithms identify patterns in existing data, which are then used to make predictions. For example, by learning distinctive patterns of spam emails from existing data, machine learning algorithms can automatically detect spam email.
Mechanical Turk *Computer Science*	An Amazon-owned crowdsourcing platform. Crowdsourcing, in layperson's terms, is to ask a large number of people to complete a task together, such as asking 1000 people to label 1 million dog pictures into different dog breeds. The idea is that large groups of people can do something way faster and even better in certain situations. So, Amazon's Mechanical Turk (MTurk) is where workers and requesters come together. The workers on MTurk are called MTurkers. They are freelancers that get paid for doing crowdsourced work provided by various requesters such as businesses, researchers, etc.
Missing data *Statistics*	Missing data occurs when there is no observed value for a certain variable. This occurs when a respondent on a survey does report an answer to a question such as their income, reports don't know, or refuses to report a value, or when a coder of a text cannot determine the content of the text.
Mixed Methods *Methodology*	Mixed Methods is a type of data collection method utilized in a research project to explore or investigate a topic. This design usually utilized more than one method of data collection such as a qualitative method (observation, focus group discussion) and a quantitate method (survey, experiment). Results from both type of methods are compared and assimilated to draw conclusions about the research questions.
Observation *Psychology*	A type of data collection in which participant behavior is observed and recorded or evaluated by a neutral third party, not reported on by the participant him/herself.

Observation vs. Exploratory *Social Science*	Studies that help us gather this information are considered observational because data are collected as they naturally exist, rather than through manipulation of variables as in experiments. Observational studies may be considered descriptive or exploratory.
Observational research *Social Science*	Technique that involves the direct observation of phenomena in their natural setting.
Optimal *Computer Science*	An optimal solution is a solution to an optimization problem that is feasible and has been mathematically proven to be the best solution given all feasible solutions.
Participant observation *Social Science*	Researcher inserts himself/herself as a member of a group, aimed at observing behavior in a naturalistic setting. Taking notes of what is observed.
Personally identifiable information *Computer Science*	Personally identifiable information is data that could potentially be used to identify a specific person. Personally identifiable information could directly identify a person, such as a driver's license number or email address. It also includes information that could be indirectly distinguish individuals even without such direct identifiers; for example, there may be only one specific person who fits a particular combination of demographics, place of employment, and job title.
Population *Survey Methodology*	A set of units (individuals, schools, students, organizations, etc..) to whom inference is to be made to.
Population frame *Survey Methodology*	A list of units in the population from which a sample will be drawn.
Positivist *Social Science*	Knowledge is exclusively derived from experience of natural phenomena and their properties and relation.

Precision/Recall *Computer Science*	Precision is a measure of quality, it is the percentage of predictions made by the information retrieval or classification system that are correct. Recall is a measure of completeness, it is the percentage of the total items that have been correctly identified by the system.
Prospective vs. Retrospective *Social Science*	In reference to studies, prospective refers to a study that will collect data multiple times in the future. Retrospective refers to studies that ask individuals to report information that happened in the past.
Reliable *Psychology*	There are many types of reliability, but in measurement generally a measure is reliable if it yields the same value for the same target on repeated observations or measurements.
Rigorous *Social Science*	Extremely thorough, exhaustive, or accurate.
Sample *Social Science*	A groups of units selected from the target population to generate estimates about the target population. The sample is usually selected from a sampling frame. Definitions of the "target population" and "sampling frame" are provided elsewhere.
Secondary use *Social Science*	Secondary use of data refers to using research data to study a problem that was not the focus of the original data collection.
Study Design *Social Science*	A framework, or the set of methods and procedures used to collect and analyze data on variables specified in a particular research problem.
Training *Computer Science*	Training is the process of tuning machine learning models using ground truth examples.
Valid *Psychology*	There are many ways to establish validity, but in measurement generally a measure is valid if it accurately measures the construct of interest. The best analogy for the difference between reliability and validity is that a scale is reliable if it gives you the same weight every time the same object is put on a scale, and a scale is valid if that weight is actually correct.