Facilitator: Corey Harper
Notetaker: Corey Harper

Metadata as Ethnography

APIs are a moving target. Google's not exposing anything about that dataset,

Amazon, Google, IBM. Nothing is documentable in some of these cases.

Systemic problems in the science of how this stuff even works. What's happening with it.

* Example of AllState data.
* Uses are commercial. Biases that come out with it affect people's relationships to car insurance company
* Deeper, more general problem is if you don't know how data was collected, there's a reproducibility problem
* Provenance, implied bias, but the bias is a product of the set, how you process, your expectations / interpretations
* How the learning is happening, how taht works.
* Anecdote about MIT Facial recognition gender thing
* Story last week, dataset floating around made available by NIST: https://slate.com/technology/2019/03/facial-recognition-nist-verification-testing-data-sets-children-immigrants-consent.html
* Gendering question Oz Keys on gendering of non-binary folk
* Word embeddings, the bias therein, that paper.
* Is there a role for libraries
  * Needing to understand what's happening here in order to use it
  * Libraries don't have the context here
  * Using the bias in the embeddings, and then trying to understand what's going on
* Chillian that was designed to detect aboriginal terrorism. But was trained on US data.
* Question about health data, iot stuff, fitbit stuff, etc
  * Does this mean we'll build medicine that is just for rich white folk?
  * That's a bigger issue
* Library workflow collection
  * Docuemnting, cataloging, capturing datasets developed by it's own instutions
  * Dataset in figshare or mendeley or on a paper
  * Proposal that every dataset needs to have a contextual paper on it
* Example from Norway. All the stuff they did is in a single temporal slot.
  * All the inferences they made are coherent due to that problem
  * Being able to define what things make up this context
* How do we bring this knowledge to the community
  * What roles do Orgs like metro have in this?
  * Finding teaching plans that fit current time

  * We're still trying to teach semantics and ontologies
* Algorithms will be re-learning from things that were already processed by Algorithms
  * You'll end up with biased and untracable datasets...
  * How do we step up here. Publishing data that tracks that Provenance
* A lot of this is already part of data management policy and approach
  * Document provenance, use, history, etc
  * Data curation is a nightmare
  * What's novel about this...
  * An interest in doing this for big public datasets that have broad use
  * Challenging the ethics of using datasets in a certain way
* dataset citations. In context of google for example.
  * Reproducbility -- what tools are getting used here? What versions?
  * How do we capture versions of models, versions of algorithms, too?
* Can we borrow from Genetics here.
  * They have tools to try to docuemnt some of hits stuff.
* We got off into reproducibility from data Ethnography
* What does it take to make something reproducible?
* What are the practices around software publication, data publication, etc?
  * Reproducible environment with docker containers
  * Is the data or the software in version control?
  * DOIs for software should go at least to a commit, if not to environment
  * Reproducibility vs replicability
  * Carrying out the experiment again, vs reproducing the analyis.
* Ocean plastics example. Meta-analysis. Attitudes from researchers on this stuff are varied
  * Researchers in "Big Data" area. Data intensive science.
  * N = all, sampling methodology doesn't matter.
  * Don't they teach bayes theorum in data science any more?
  * If you're close enough to your sample being the population