title: BAIT507 - Data Management
separator:
verticalSeparator:
theme: solarized
revealOptions:

transition: 'fade'

Scientific Annotation

Using Graphs to Facilitate Interdisciplinary Science

NSF Funded.

New Slide

The degree of interdisciplinary work is increasing significantly.

New Slide

Note: DJ Patil indicated the wealth of data in earth sciences available to data sciences, the challenge is that much of the data requires some level of disciplinary knowledge to analyse.

New Slide

Note: Work by Stephen Richard and myself has shown how mapping variable definitions across open-science databases does indicate a level of apparent alignment between data resources (e.g., Temperature to Temperature), but in practice this alignment may be illusory.

New Slide

Note: From the perspective of equity in academia, and from a project management standpoint this creates two problems: (1) Interdisciplinary research requires a breadth of expertise that may not

be available at all institutions. If this is the case, the development of new tools, and the capacity to answer new questions using data science tool kits gets concentrated in R1 institutions. Whether or not we operate in a meritocracy, the challenge is that the undergraduates who attend HBCUs, Tribal Colleges or regional, undergraduate domainated universities would never have a chance to "prove themselves", since they simply do not have the depth of experience to work with these data types.

New Slide

Note: Within individual projects we also see inequity between early career researchers and established researchers. Because of the structure of the academic reward system, it is the early career researchers who bear the burden of interdisciplinary research, while established researchers gain most of the benefits. Reducing the "time to science" for early career researchers would allow them to gain rewards faster, and compete against researchers working within the more traditional disciplinary silo.

New Slide

Note: Ive been intimately involved in interdisciplinary research for most of my academic career. I manage the Neotoma Paleoecology Database, a community curated data repository that manages records of fossils and environmental data covering the last 2.5 mya, developing tools in R to access this resource, writing scripts and helping build the pipeline to other tools, including the FlyoverCountry App, that puts the power of XXX geological data resurces into the hands of users.

New Slide

Note: My research has focused on building models that relate climate and forest cover over the last 10,000 years, along with historical survey records, modern remote sensing data, real-time sensor data, data assimilation models and General Circulation models to help improve predictions of future climate using models of past vegetation and climate.

New Slide

Note: Through all of this, I have watched students struggle. It's not enough to provide people with data. We need to support people with the kinds of workflows that will help them understand how to use the data.

New Slide

Note: Introducing EarthCube & Throughput

New Slide

Note: Introducting re3data

```
MATCH (:TYPE {type:"schema:DataCatalog"})-[:isType]-(odca:OBJECT) RETURN COUNT(odca
```

2301 Unique online data resources with metadata & defined data schema.

New Slide

Note: Repositories that are linked. Data cleaning that needs work.

```
MATCH (:TYPE {type:"schema:CodeRepository"})-[:isType]-(ocr:OBJECT)
MATCH (:TYPE {type:"schema:DataCatalog"})-[:isType]-(odca:OBJECT) MATCH (:TYPE {typ
WITH ocr, odca, odcb
MATCH p = (odca)-[]-(:ANNOTATION)-[]-(ocr)-[]-(:ANNOTATION)-[]-(odcb)
WITH odca, COLLECT(odcb) AS thing
RETURN odca.name, count(thing)
```

New Slide

Note: Which data resources have the most crosspostings?