Anthropology by Data Science: The EPIC Project with Indicia Consulting as an Exploratory Case Study

By Stephen Paff

Introduction

Thesis: Anthropologists should, when applicable, conduct anthropology *by* science (that is, integrate data science techniques into ethnographic and other anthropological work)

1. I will use Nick Seaver's concept of bastard methodology to ground this theoretically.
2. My practicum with Indicia Consulting is an exploratory case study in such integrative work.

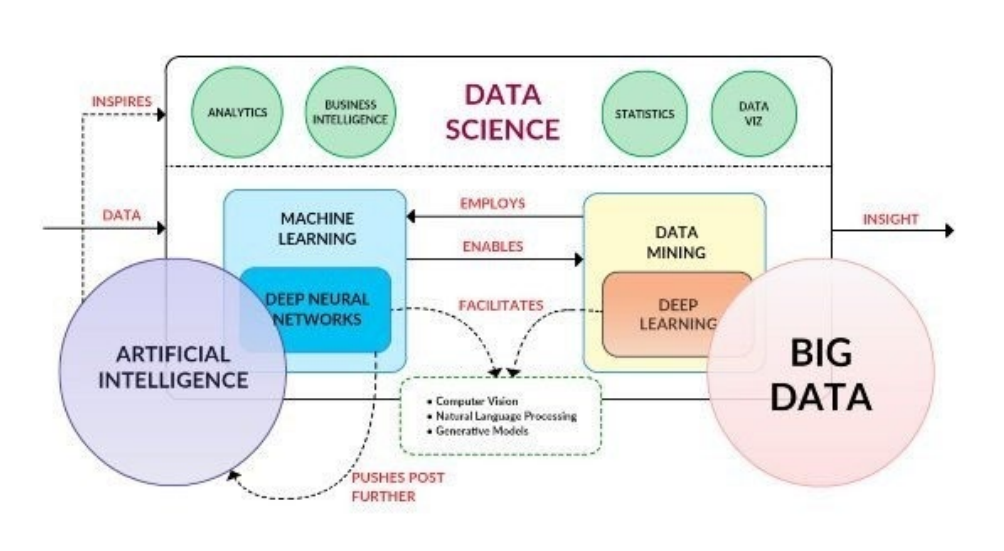
Indicia Consulting "is a mission-driven social enterprise" that seeks to "increase in sustainability and subsequent improvement in the natural environment through engaging behavior through proven social science insights and methods" (EPIC 2018).

Outline:

1. Introduce key concepts
2. Discuss Nick Seaver's concept of bastard methodology
3. Summarize our project
4. Analyze strengths and weaknesses of our project
5. Conclusion

Definitions

- Data: Information (or things known) that form the basis of analysis
- Data science: The science of analyzing data computationally.
- Machine learning algorithms: Algorithms that "learn" by modifying themselves as they iterate through data



https://www.quora.com/What-is-the-difference-between-data-science-data-analysis-data-mining-machine-learning-AI-and-big-data

Interconnections

- Data scientists study data through machine learning algorithms (primarily).

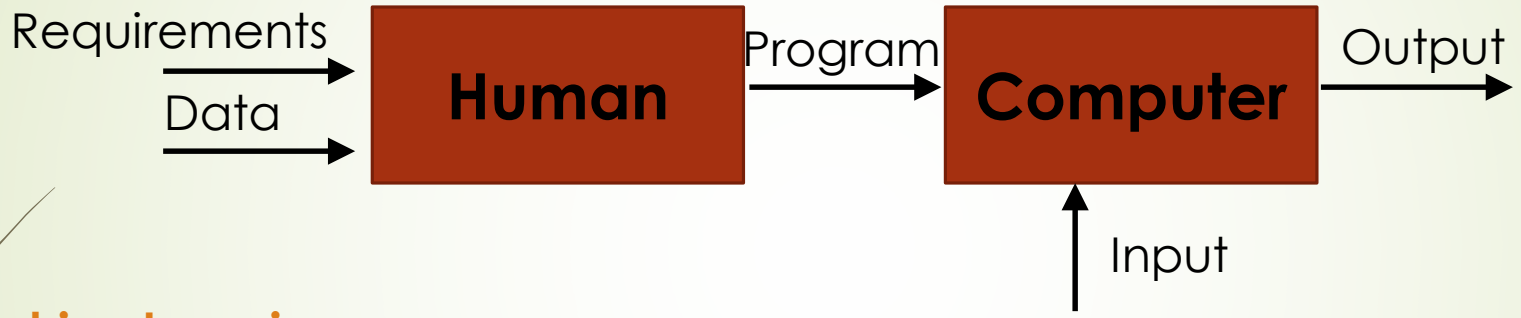
Just like:

- Anthropologists study culture through ethnography (primarily).

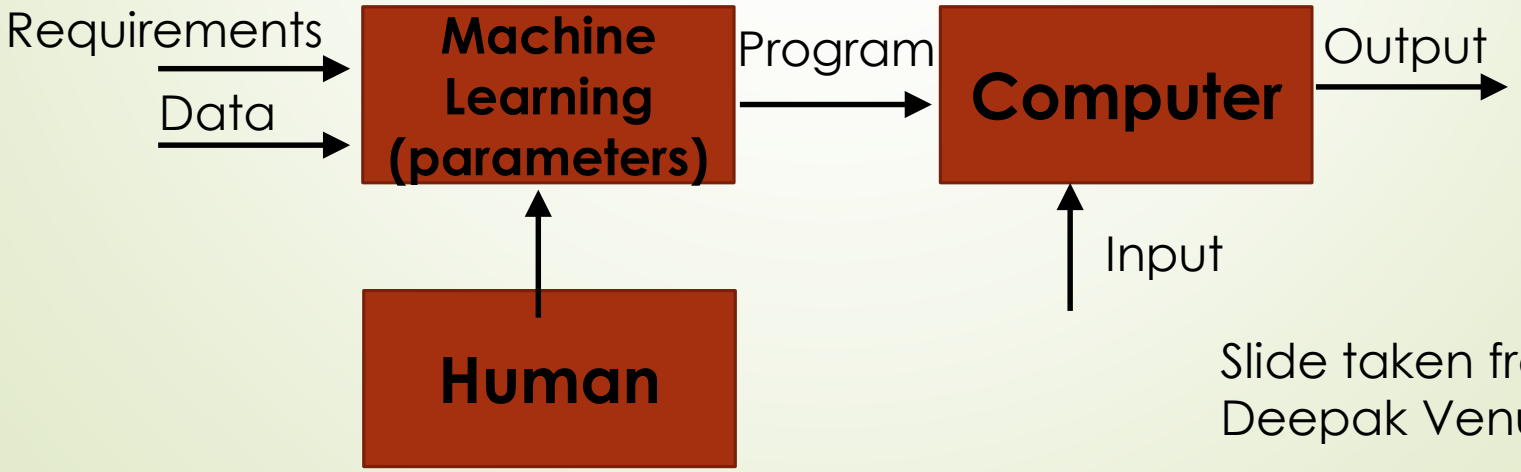
Discipline	Who	What	How/Technique
Data Science	Data scientists	Data	Machine Learning
Anthropology	Anthropologists	Culture	Ethnography

Shift in Relationship in Computer Science

Traditional Programming



Machine Learning



Slide taken from lecture by Dr. Deepak Venugopal

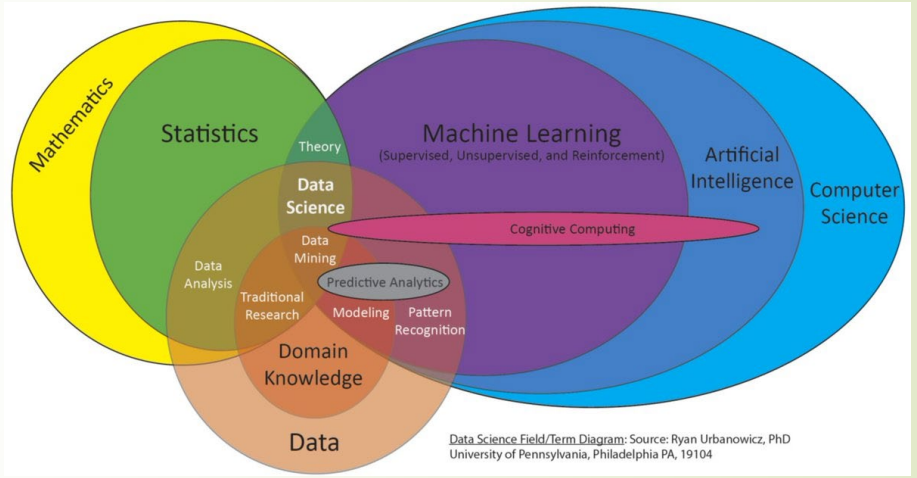
Anthropological Literature about Data Science

For anthropological theories about data science, I am most interested in the relationship they form with data science:

Relationship	Definition
Anthropology of Data Science	Anthropological analysis of data science and/or data scientists
Anthropology over Data Science	Perspectives that anthropology should oversee, regulate, or monitor data science work
Anthropology with Data Science	Collaborative work between the anthropology and data science
Anthropology by Data Science	Using data science tools to conduct anthropological work

Seaver's Bastard Methodology

- Bastard methodology: mishmash approach combining several disciplinary strategies/techniques based on what works to address a specific research question(s) or issue(s)
- Data science techniques and ethnography are both bastard methodologies (2015:44-45).
- I will call the opposite approach a purist disciplinary approach:
  - Separating one's disciplinary methodology to ground its rigor



https://twitter.com/docurbs/status/1007375834347376642

Seaver's Bastard Methodology

- "These relationships between methods are, by now, such disciplinary common sense that we might be surprised how much work goes into rehearsing and reinforcing them. Either ethnography and formal analysis compete for jurisdiction over concepts like 'culture,' or they enter into scripted collaborations with each other, drawing on their apparently complementary strengths. Often, these arguments about how big data and ethnography might get along rehearse claims that have defined ethnography and its relationship to other methods throughout ethnography's entire history." (2015:37)
- "Data scientists work as professional bastard-makers, combining data sets, algorithms, and epistemologies in unauthorized ways to produce illicit offspring" (43).
- "Ethnography is a bastard too, breeding descriptions from illicit encounters, mixing conceptual schemes, and stirring the blood of experience with the ink of theory. As we examine the family situation of bastard algebra [data science], we will have to come to terms with the bastard status of ethnography itself, remaining open to the ubiquity and generative potential of epistemologies that overflow their borders and relate without permission" (44-45).

Anthropology *with* Data Science

- Collaborative
- Separate but complementary
- Combine at the level of people: specialists/people work together
- Action: Seek to work collaboratively data scientists, advancing the value of our "part" of that work

Anthropology *by* Data Science

- Integrative
- Entangles the two through cross-pollination
- Combine at the level of practice: make adjustments to our practice
- Action: Incorporate data science techniques into our ethnographic toolkit as a way to cross-pollinate

Project History

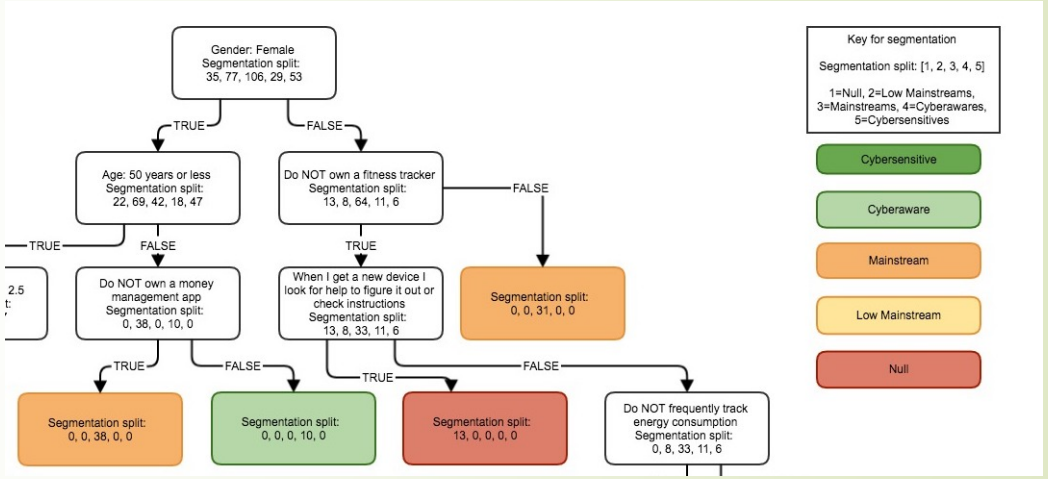
- My practicum focused on Task 6 of the EPIC Project.

Task	Timeline	Task Name	Research Technique	Description
Task 1	June 2015-Sept 2018	General Project Tasks	Administrative (N/A)	Developed project scope and timeline, adjusting as the project unfolds
Task 2	July 2015 – July 2016	Documenting and analyzing emerging attitudes, emotions, experiences, habits, and practices around technology adoption	Survey	Conducted survey research to observe patterns of attitudes and behaviors among cybersensitives/awares.
Task 3	Sept 2016 – Dec 2016	Identifying the attributes and characteristics and psychological drivers of cybersensitives	Interviews and Participant-Observation	Conducted in-depth interviews and observations coding for psych factor, energy consumption attitudes and behaviors, and technological device purchasing/usage.
Task 4*	Sept 2016 – July 2017	Assessing cybersensitives' valence with technology	Statistical Analysis	Tested for statistically significant differences in demographics, behaviors, and beliefs/attitudes between cyber status groups
Task 5	Aug 2017 – Dec 2018	Developing critical insights for supporting residential engagement in energy efficient behaviors	Statistical Analysis	Analyzed utility data patterns of study participants, comparing it with the general population.
Task 6	March 2018 – Aug 2018	Recommending an alternative energy efficiency potential model	Decision Tree Modeling	Constructed decision tree models to classify an individual's cyber status

*I originally joined the project in March 2017 to complete Task 4.

Task 6 Project Goal

Task 6's goal was to use decision tree modeling to represent people's cyber status.



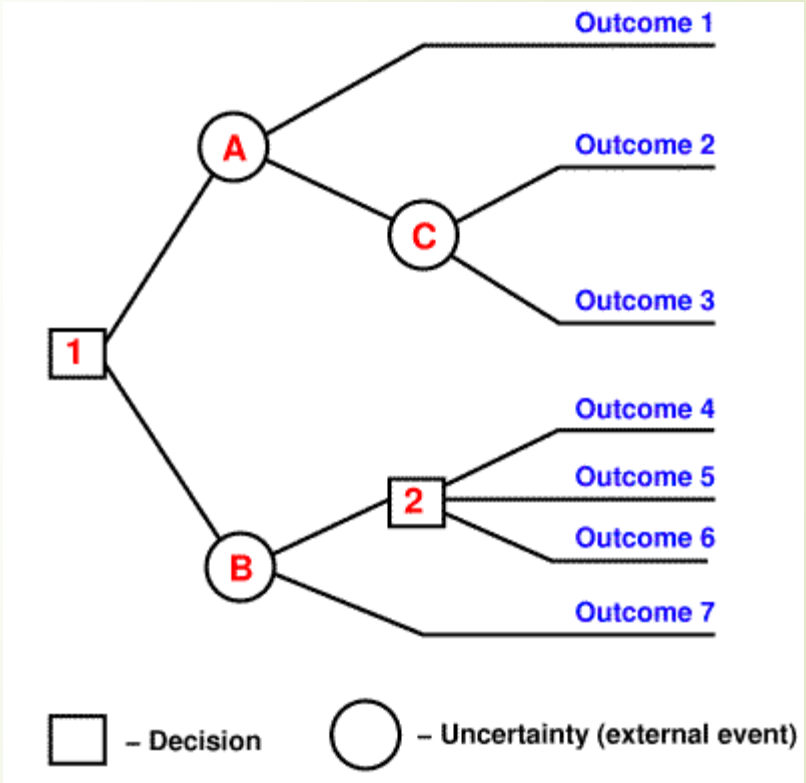
Cyber Status

- Cyber status: psychosocial characteristics Indicia developed in this project to understand an individual's perceived emotional (or lack of) connection with technology and resulting behaviors (Indicia Consulting 2018:5)
- Made up of 5 ranked groups:

Cybersensitive
Cyberaware
Mainstream
Low Mainstream
Null

Decision Tree Modeling

- Decision trees generally model a decision-making flow through a hierarchy of (usually) binary Boolean decisions or criteria, such as true/false or yes/no conditions (Paff 2018:24; Indicia Consulting 2018:6).
- Two types of decision trees:
  1. Ethnographic decision tree modeling (EDTM)
  2. Machine learning decision tree modeling (CART)
- Both are abductive/iterative.

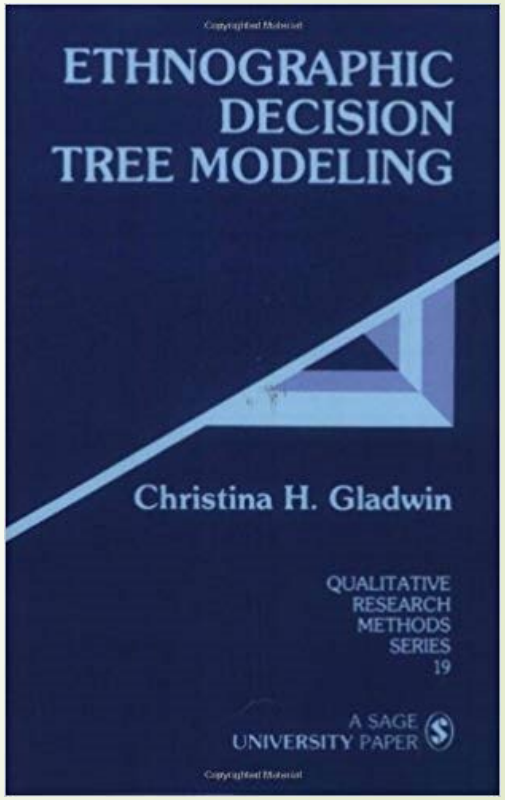


https://medium.com/@SeattleDataGuy/what-is-a-decision-tree-algorithm-4531749d2a17

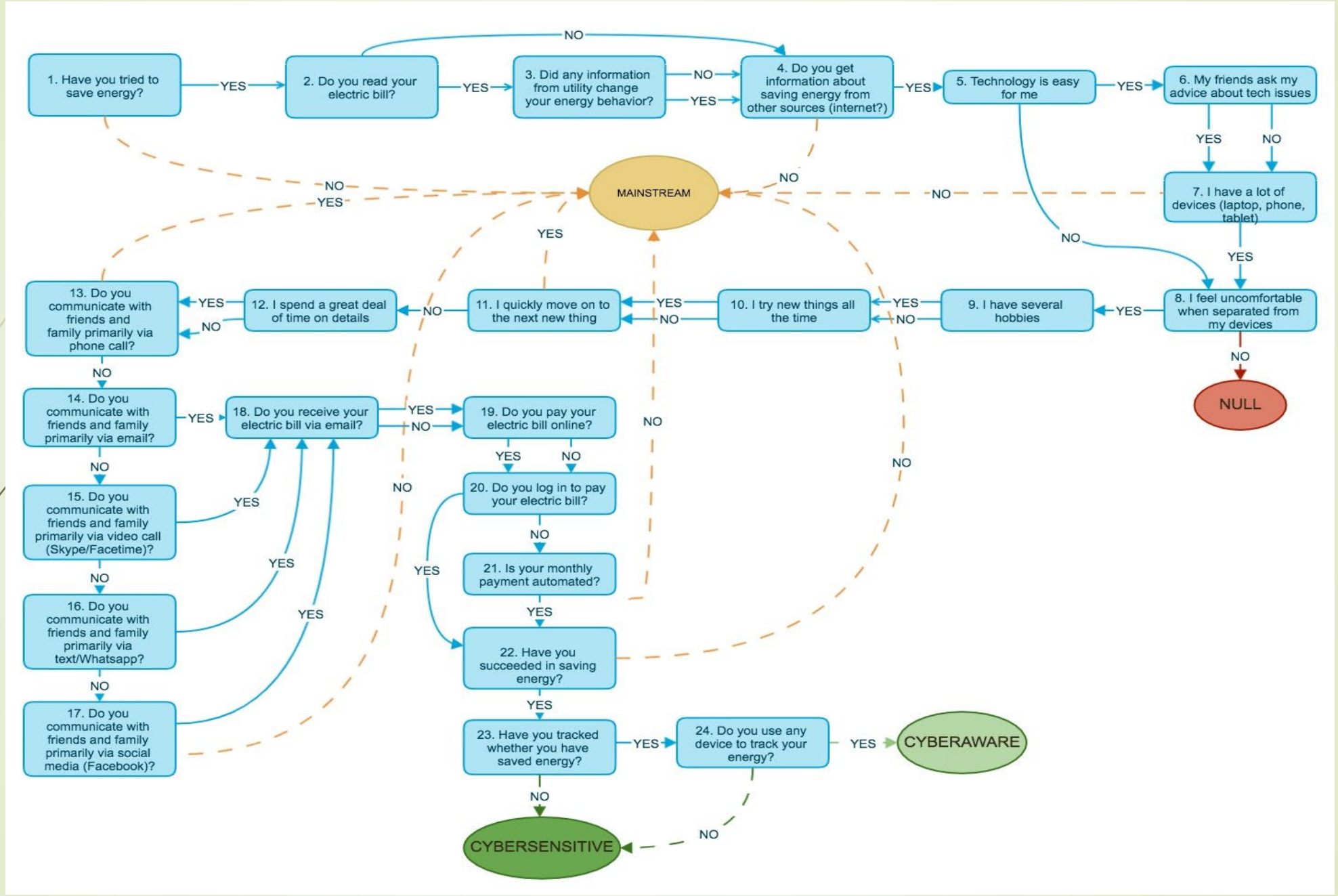
Stages of EDTM

1. Interview: Conduct ethnographic interviews of one group about their decision-making process (Task 3)
2. Analyze: Review responses and organize into decision flow
3. Iterate: Run these set of choices past another similar group and modify accordingly (Gladwin 1997; Indicia Consulting 2018:11-12)

I was not the primary lead on the EDTM for this project.



https://www.amazon.com/Ethnographic-Decision-Modeling-Qualitative-Research/dp/0803934874


1. Have you tried to save energy?
2. Do you read your electric bill?
3. Did any information from utility change your energy behavior?
4. Do you get information about saving energy from other sources (internet?)
5. Technology is easy for me
6. My friends ask my advice about tech issues
7. I have a lot of devices (laptop, phone, tablet)
8. I feel uncomfortable when separated from my devices
9. I have several hobbies
10. I try new things all the time
11. I quickly move on to the next new thing
12. I spend a great deal of time on details
13. Do you communicate with friends and family primarily via phone call?
14. Do you communicate with friends and family primarily via email?
15. Do you communicate with friends and family primarily via video call (Skype/Facetime)?
16. Do you communicate with friends and family primarily via text/Whatsapp?
17. Do you communicate with friends and family primarily via social media (Facebook)?
18. Do you receive your electric bill via email?
19. Do you pay your electric bill online?
20. Do you log in to pay your electric bill?
21. Is your monthly payment automated?
22. Have you succeeded in saving energy?
23. Have you tracked whether you have saved energy?
24. Do you use any device to track your energy?
MAINSTREAM
NULL
CYBERAWARE
CYBERSENSITIVE
YES
NO

Stages of CART

1. Pre-Development: Cleaned and prepared data
   a) Resampling
2. Development: Developed the decision tree model
   a) Gini Index
3. Pruning: Tested the model's accuracy and further refined by adjusting the (hyper)parameters and ensemble methods
   a) Leave-One-Out Cross-Validation
   b) Random Forests

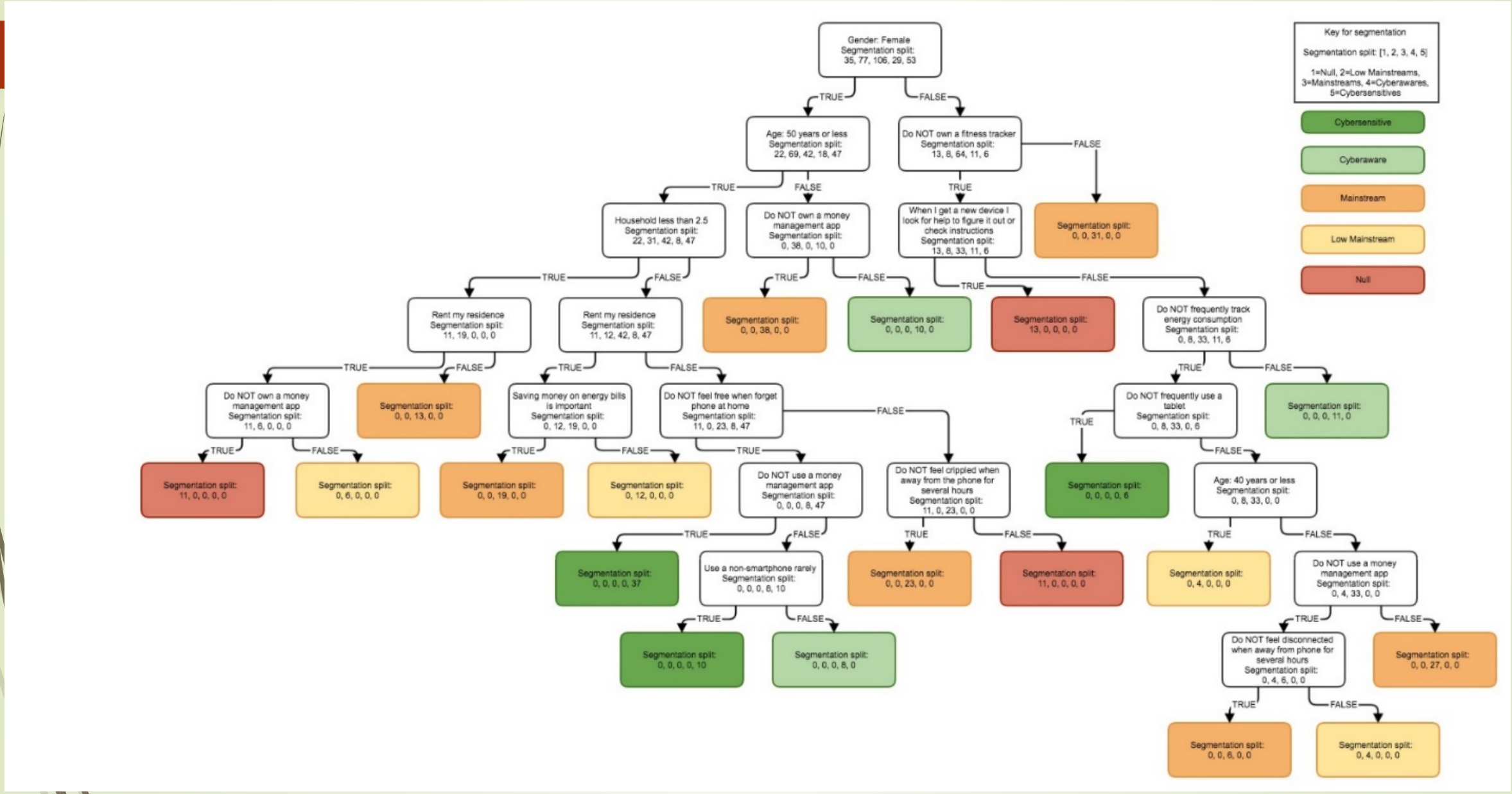
Our random forest had 100% accuracy on our sample.

Thus we will use the gini index with a maximum depth of 7.

In [10]:
```
loocv = model_selection.LeaveOneOut()
accuracy = pd.DataFrame( columns = ['Maximum Depth', 'Gini Index Accuracy', 'Entropy Accuracy'])

for depth in range(1, 34):
    score = [depth]

    clf_gini = tree.DecisionTreeClassifier(criterion = "gini", random_state = 100, max_depth=depth, min_samples_leaf=1)
    results = model_selection.cross_val_score(clf_gini, X_resample, y_resample_qual, cv=loocv)
    score.append(results.mean())

    clf_entropy = tree.DecisionTreeClassifier(criterion = "entropy", random_state = 100, max_depth=depth, min_samples_leaf=1)
    results = model_selection.cross_val_score(clf_entropy, X_resample, y_resample_qual, cv=loocv)
    score.append(results.mean())

    accuracy.loc[len(accuracy)] = score

accuracy
```

Out[10]:

	Maximum Depth	Gini Index Accuracy	Entropy Accuracy
0	1.0	0.253333	0.333333


Gender: Female
Segmentation split:
35, 77, 106, 29, 53
TRUE
FALSE
Age: 50 years or less
Segmentation split:
22, 69, 42, 18, 47
Do NOT own a fitness tracker
Segmentation split:
13, 8, 64, 11, 6
FALSE
Household less than 2.5
Segmentation split:
22, 31, 42, 8, 47
Do NOT own a money management app
Segmentation split:
0, 38, 0, 10, 0
When I get a new device I look for help to figure it out or check instructions
Segmentation split:
13, 8, 33, 11, 6
Segmentation split:
0, 0, 31, 0, 0
Rent my residence
Segmentation split:
11, 19, 0, 0, 0
Rent my residence
Segmentation split:
11, 12, 42, 8, 47
Segmentation split:
0, 0, 38, 0, 0
Segmentation split:
0, 0, 0, 10, 0
Segmentation split:
13, 0, 0, 0, 0
Do NOT frequently track energy consumption
Segmentation split:
0, 8, 33, 11, 6
Do NOT own a money management app
Segmentation split:
11, 6, 0, 0, 0
Segmentation split:
0, 0, 13, 0, 0
Saving money on energy bills is important
Segmentation split:
0, 12, 19, 0, 0
Do NOT feel free when forget phone at home
Segmentation split:
11, 0, 23, 8, 47
Do NOT frequently use a tablet
Segmentation split:
0, 8, 33, 0, 6
Segmentation split:
0, 0, 0, 11, 0
Segmentation split:
11, 0, 0, 0, 0
Segmentation split:
0, 6, 0, 0, 0
Segmentation split:
0, 0, 19, 0, 0
Segmentation split:
0, 12, 0, 0, 0
Do NOT use a money management app
Segmentation split:
0, 0, 0, 8, 47
Do NOT feel crippled when away from the phone for several hours
Segmentation split:
11, 0, 23, 0, 0
Segmentation split:
0, 0, 0, 0, 6
Age: 40 years or less
Segmentation split:
0, 8, 33, 0, 0
Segmentation split:
0, 0, 0, 0, 37
Use a non-smartphone rarely
Segmentation split:
0, 0, 0, 8, 10
Segmentation split:
0, 0, 23, 0, 0
Segmentation split:
11, 0, 0, 0, 0
Segmentation split:
0, 4, 0, 0, 0
Do NOT use a money management app
Segmentation split:
0, 4, 33, 0, 0
Segmentation split:
0, 0, 0, 0, 10
Segmentation split:
0, 0, 0, 8, 0
Do NOT feel disconnected when away from phone for several hours
Segmentation split:
0, 4, 6, 0, 0
Segmentation split:
0, 0, 27, 0, 0
Segmentation split:
0, 0, 6, 0, 0
Segmentation split:
0, 4, 0, 0, 0
Key for segmentation
Segmentation split: [1, 2, 3, 4, 5]
1=Null, 2=Low Mainstreams, 3=Mainstreams, 4=Cyberawares, 5=Cybersensitives
Cybersensitive
Cyberaware
Mainstream
Low Mainstream
Null

Project Strengths

Our decision tree models had the following strengths:

1. Intelligible narratives for both humans and computers
2. Intermixable parts, allowing for unique connections
3. Translatable:
    a. Between disciplines
    b. Into actionable insights for policy-makers

http://www.invistaperforms.org/strengths-key-higher-productivity-effectiveness-employee-engagement-joy/

Project Weaknesses

1. Unable to test against an external dataset
2. Potential for overfitting (representing features specific to the group studied)
3. Models developed separately to be integrated in testing

https://www.potentash.com/2018/02/20/important-embrace-your-weaknesses-lifestyle/

Strategy to Address Weaknesses

We obtained a second, external dataset of Prince Williams County for testing from Network Dynamics Simulation Science Laboratory at Virginia Tech University.

<u>Ultimate Goals:</u>

- To create a general criteria to classify the cyber status of individuals in a whole population
- To segment market specific energy-saving campaigns towards each group

https://www.marketingtechnews.net/news/2015/aug/21/big-data-and-personalisation-truly-comes-age-next-steps-cmo/

Conclusion

- Nick Seaver's concept of bastard methodology/disciplines helps ground a cross-pollinating approach between the anthropology and data science,
  - Within anthropology, this manifests as anthropology *by* data science work: using data science techniques in ethnographies when applicable.
- Decision tree modeling is helpful for this.
- The project is a exploration in anthropology *by* data science.

Work Cited

- EPIC. Indicia Consulting. 2018. <https://www.epicpeople.org/business-directory/4430/indicia-consulting/>.
- Gladwin, Christina H. *Ethnographic Decision Tree Modeling*. Newbury, CA: Sage, 1997.
- Indicia Consulting. "Engaging Cybersensitives and Cyberawares in Energy Efficiency Part 1." Epic Project Task 6. 2018.
- Paff, Stephen. *Anthropology by Data Science: The EPIC Project with Indicia Consulting as an Exploratory Case Study*. Practicum Report. Memphis: University of Memphis, 2018.
- Seaver, Nick. "Bastard Algebra." Boellstorff, Tom and Bill Maurer. *Data, Now Bigger and Better*. Chicago: Prickly Paradigm Press, 2015. 27-46.