Futzing and Moseying

Interviews with Professional Data Analysts on Exploration Practices

Sara Alspaugh, Nava Zokaei, Andrea Liu, Cindy Lin, Marti Hearst
UC Berkeley

VAST 2018

Supported by the UC Berkeley AMP Lab and a Gift from Tableau Research

The Future of Data Analysis, Tukey 1962

Motivating Question:

Do professional analysts do **exploratory data analysis?**

If so, why? If not, why not?

If so, what kinds of automated tools do they desire?

DEFINITION

Exploration: Open-ended information analysis, which does not require a precisely stated goal.

(Although some EDA begins with preliminary, motivated hypotheses)

EDA: A SPECTRUM

Outline

Motivation / Definitions

Related Work

Method

(Recruitment, Interviews, Coding)

Findings

Ramifications

More Motivations

Many years of
exploratory data analysis student projects

Interest in building tools that
automatically suggest analyses.

More Motivation:

Others motivate their research with claims about EDA:

"Exploratory visual analysis is highly iterative, involving both open ended exploration and targeted question answering."

-- Wongsuphasawat, Moritz, Anand, Mackinlay, Howe, and Heer, *Voyager*, Infoviz 2015.

More Motivation:

Others motivate their research with claims about EDA:

"Data exploration is about efficiently extracting knowledge from data even if we do not know exactly what we are looking for."

—Idreos et al. *Overview of Data Exploration Techniques.* SIGMOD 2015 (Tutorial).

"[Interactive data exploration] is fundamentally a multi-step, non-linear process with imprecise end-goals."

—Cetintemel et al. *Query Steering for Interactive Data Exploration*. CIDR 2013.

Outline

Motivation / Definitions

Related Work

Method

(Recruitment, Interviews, Coding)

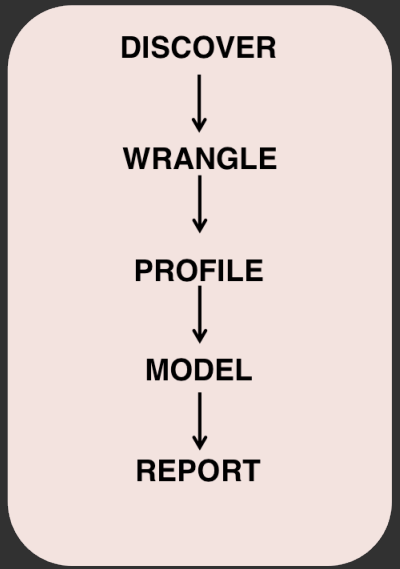
Findings

Ramifications

Related Work

- Data Analyst Interview Studies
  - Kandogen et al. 2014, Kim et al. 2016, Fisher et al. 2012
  - Difference foci (e.g. cloud computing, only one organization)
- Intelligence Analyst Interview Studies / Sensemaking
  - Kang & Stasko 2011, Cowley et al. 2005, Pirolli & Card 2005
  - Intense collaboration, desire for automated tools
- Case Studies
  - Russell 2016, Perer & Shneiderman 2008

Kandel et al.’s 2012 model



Kandel, Paepcke, Hellerstein, Heer, Enterprise Data Analysis and Visualization: An Interview Study, TVCG 2012

Outline

Motivation / Definitions

Related Work

Method

(Recruitment, Interviews, Coding)

Findings

Ramifications

METHOD

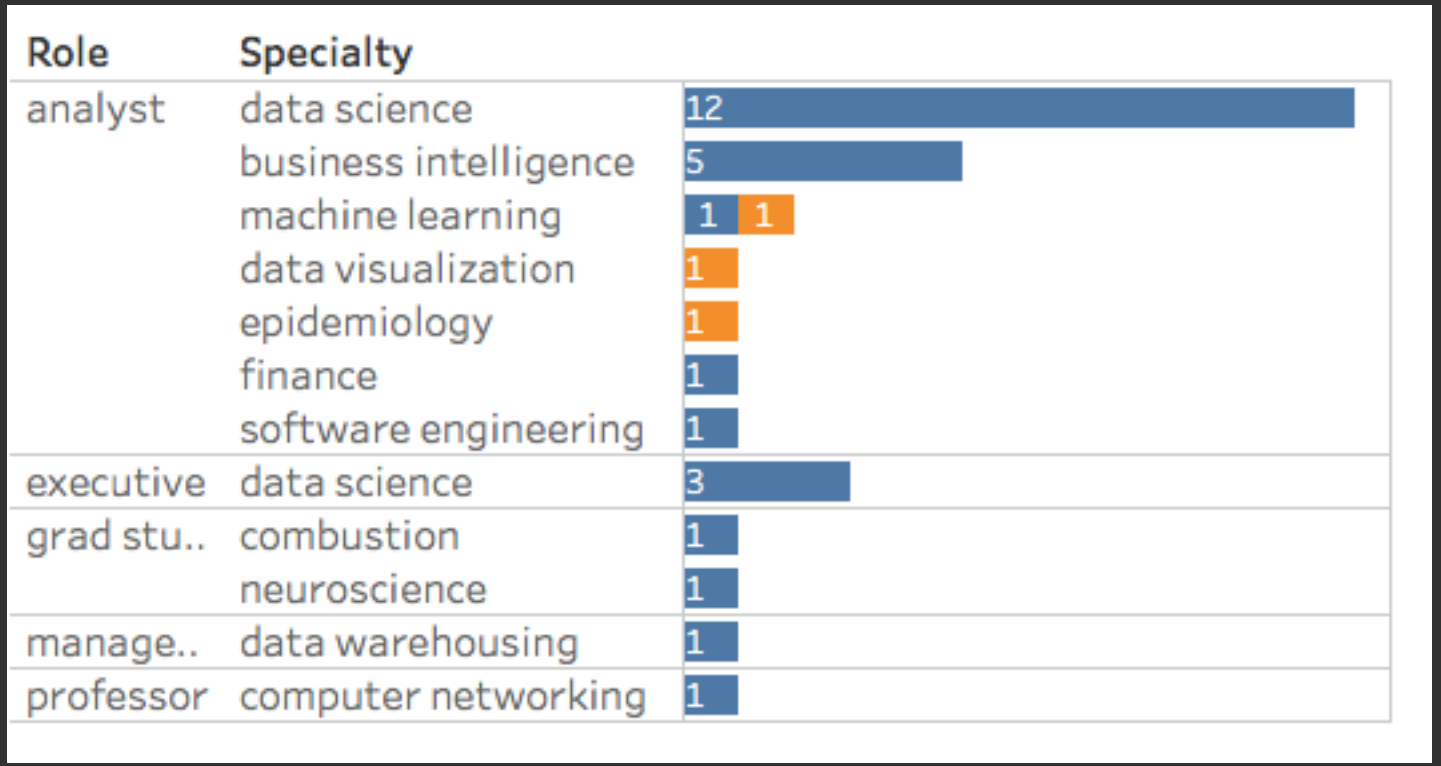
Recruiting, Interviewing, Coding

Recruiting and Interviewing

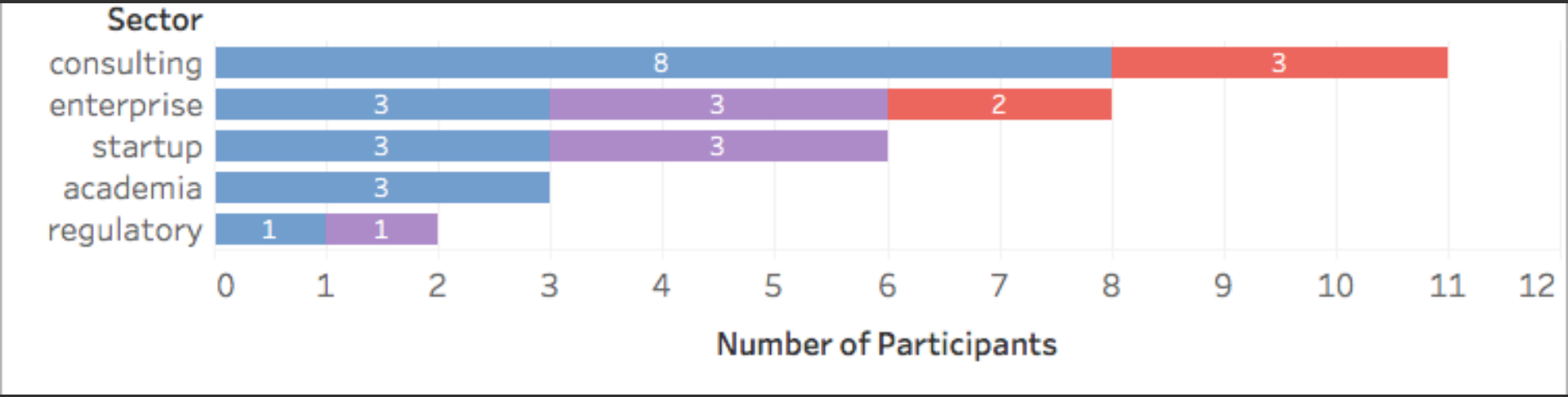
Reached out to professional network

“Professionals who analyze data” daily
Indicated that focus was EDA
30 respondents; 90 minutes on avg

Demographics



Sector by Tool Preference



Developing the Codebook

Grounded theory,
collaborative process

The Codebook

75 codes

Top levels:

- Background
- Workflow stage (exploration goals)
- Tools used
- Desired tools and features
- Homegrown automation

Codes hierarchically organized

Labeling Utterances

Pass 1: using codebook, 2 coders independently labeled each utterance.

Every utterance gets ≥ 2 coders

Pass 2: each reviewed the other's codes, considered changing if conflict.

Pass 3: any remaining differences were tie-broken by the third coder.

8683 total

Cohen's Kappa=0.91

FINDINGS

The Characteristics of Exploration

EDA: A SPECTRUM

Exploratory or Directed?

exploratory

directed

"A lot of putzing, a lot of trying to parse our text logs to see if I could find anything helpful. … Same as like futzing. … Kind of moseying … I don't know, just poking around with things and see what happens.

Exploratory or Directed?

exploratory

directed

The project I'm doing, I have a very specific question … how does this area connect to this area … I kind of know what I'm going to do.

Exploratory or Directed?

exploratory

directed

Primarily at its root I do video and voltage versus time, basically. And then correlations and all that sort of stuff to try to extract a meaningful result from those.

Exploratory or Directed?

exploratory

directed

directed

Primarily at its root I do video and voltage versus time, basically. And then correlations and all that sort of stuff to try to extract a meaningful result from those.

vague, open-ended

MAIN FINDINGS

Exploratory activities pervade the entire analysis process (for some analysts)

--> Need to extend Kandel et al.'s model with a new phase, EXPLORE:

Find something interesting, check understanding,
generate new questions, demonstrate algorithms or tools

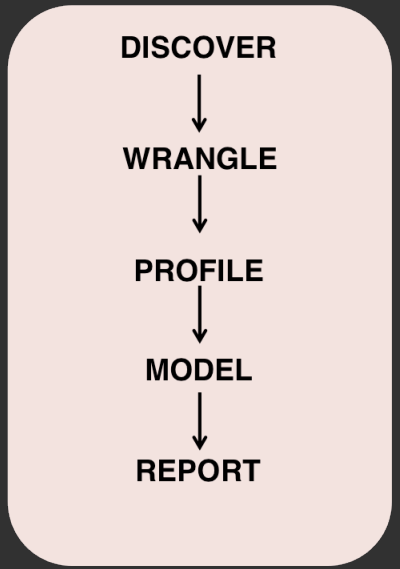
Analysts want a compromise between coding and direct manipulation

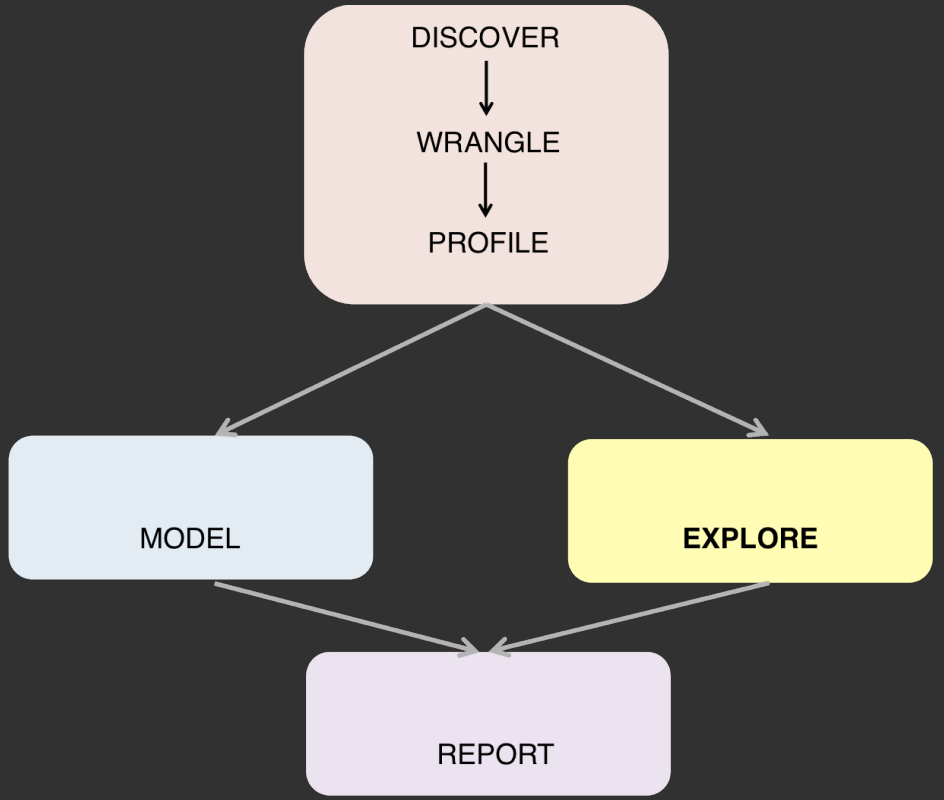
Notebooks with interactive visualizations are promising

Skepticism toward automated analysis tools

Augmenting a Model

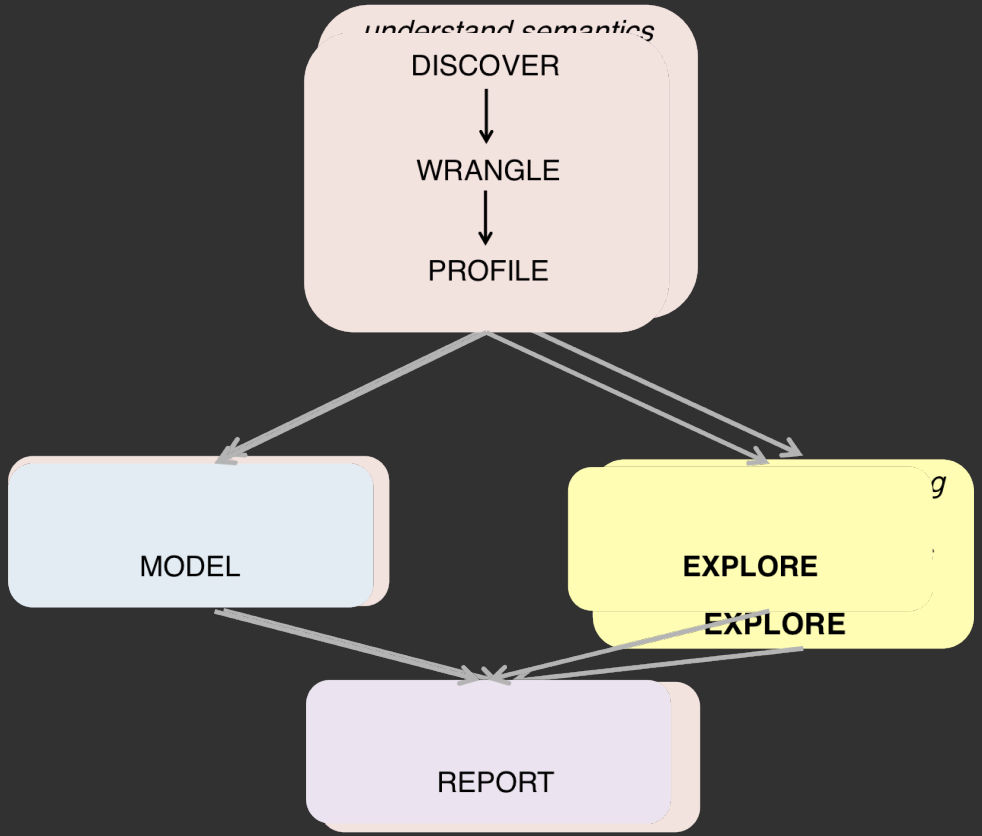
Kandel et al. 2012





Augmenting a Model

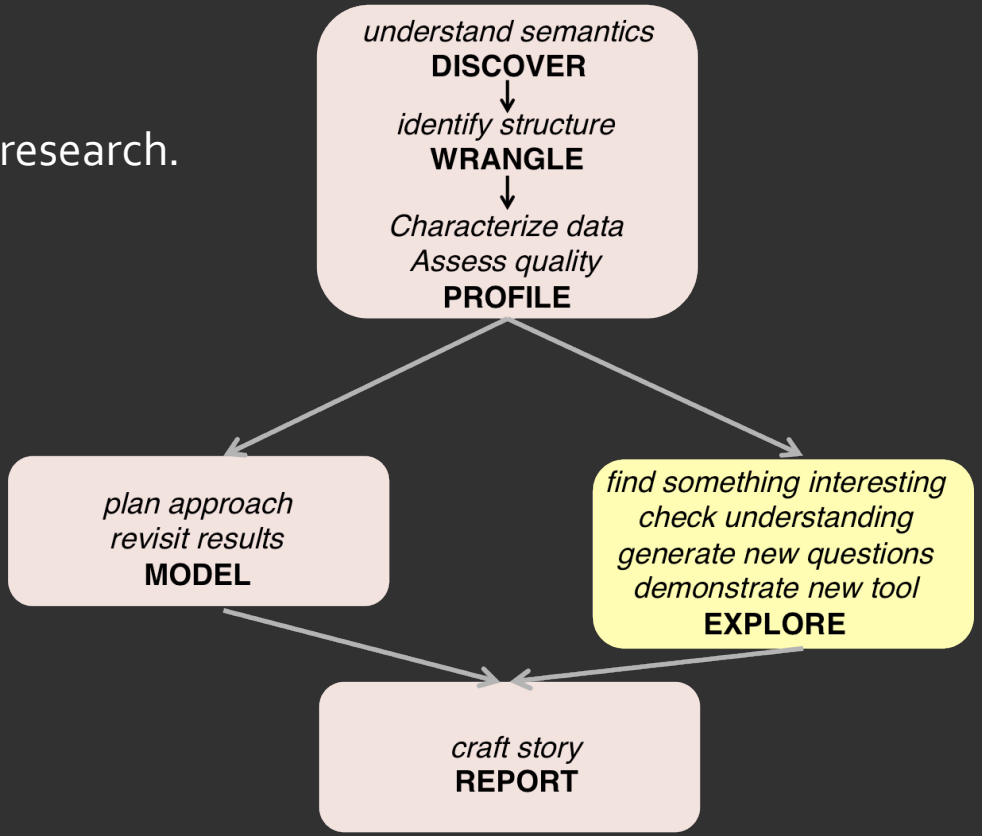
Kandel et al. 2012



Augmenting a Model

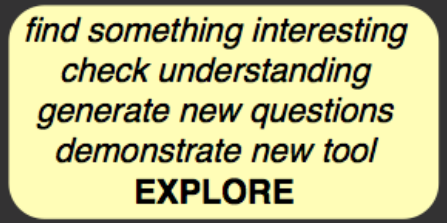
Kandel et al. 2012

Italicized parts are contributions of this research.
They show where exploration occurs.



Core Exploratory: Find Something Interesting (12/30)

"I have all this time series data, I've got terabytes and terabytes of data, I know what tasks they are, I know what people they were, I know everything about it, and now I can just put it through whatever I want, and figure out what's there, what's interesting."



Core Exploratory: Find Something Interesting (12/30)

"So most of the jobs I apply for they'll give you a take-home exercise where they give you a couple hundred Megs of data in whatever format. And they're like, "Tell us something interesting with our data."...It's general data analytics and finding the needle in the haystack. Developing interesting understandings, finding trends."

find something interesting
check understanding
generate new questions
demonstrate new tool
EXPLORE

Core Exploratory: Find Something Interesting (12/30)

"Correlating … data and using this prediction to be able to reduce the time and make a better … customer experience. That's one use case. I'd start with that. Then we can get correlations … with other … sources … [and] see different patterns … It's just finding out the different ideas, and that's what makes the whole process interesting and not often repeatable, unfortunately."

find something interesting
check understanding
generate new questions
demonstrate new tool
EXPLORE

Core Exploratory: Check Understanding (9/30)

Compare data with the current understanding of an underlying phenomenon in a highly open-ended fashion.

"What you are looking for when you look for when you say, ``Look, Tableau, just show me these seven fields, just show me a a chart that lets me understand.'' If you look at it, most of the time, you are going to go, ``Yeah, that is what I figured. … What you are looking for is those outliers. … I need to understand that one. Is there fraud involved? "

find something interesting
check understanding
generate new questions
demonstrate new tool
EXPLORE

Core Exploratory: Check Understanding (9/30)

Compare data with the current understanding of an underlying phenomenon in a highly open-ended fashion.

"I spend probably 25% of my time looking at user stats and metrics and how people are using the site in various ways … exploring how users are using our product and identifying our user base."

find something interesting
check understanding
generate new questions
demonstrate new tool
EXPLORE

Core Exploratory: Check Understanding (9/30)

Compare data with the current understanding of an underlying phenomenon in a highly open-ended fashion.

"… we were trying to understand user behavior more generally and so in that case we were like, "Okay what are some different dimensions that we can approach this problem?" We can approach it from an activity standpoint. We can approach it from an interest standpoint."

"We spend many, many, many hours, calibrating often… shorthand for data quality issues and understanding both the nature of the data, its flaws, and then of the underlying phenomena, both."

find something interesting
check understanding
generate new questions
demonstrate new tool
EXPLORE

Core Exploratory: Generating New Questions (5/30)

Coming up with new analysis questions or hypotheses by looking through data in a free-form fashion.

> “There's learnings that I'm having in the data that are leading more questions but aren't necessarily germane to the task at hand. … So there's just kind of tagging it for later. ”

find something interesting
check understanding
generate new questions
demonstrate new tool
EXPLORE

Core Exploratory: Generating New Questions (5/30)

Coming up with new analysis questions or hypotheses by looking through data in a free-form fashion.

"That's part of the cycle is, figuring out what you don't want to look at. It's this exploratory. Go everywhere as fast as you can, and look at all sorts of weird things, in order to figure out really what do you care about."

find something interesting
check understanding
generate new questions
demonstrate new tool
EXPLORE

Core Exploratory: Demo New Tool/Method (5/30)

"Whenever I come to a new database and one of the things that I do is, when a user first lands with [our tool], we want it to be interesting…so what we'll do is we'll go into their data for them and we'll just try to make some somewhat interesting charts that really showcase the product. They don't necessarily have to actually be useful but they should be a little bit interesting and make sense."

find something interesting
check understanding
generate new questions
demonstrate new tool
EXPLORE

Core Exploratory: Demo New Tool/Method (5/30)

“If [the potential customer is] not sure if [our tool] is going to be useful for them, or they're not quite sure how to make use of it, … there's a team … who will use [our tool] to generate and annotate a visualization. …They'll do some analysis and come up with some interesting findings, they'll be like, “Hey, we found this vulnerability that just came out in tools that you use. This is the kind of thing that you could discover using [our tool].''. ”

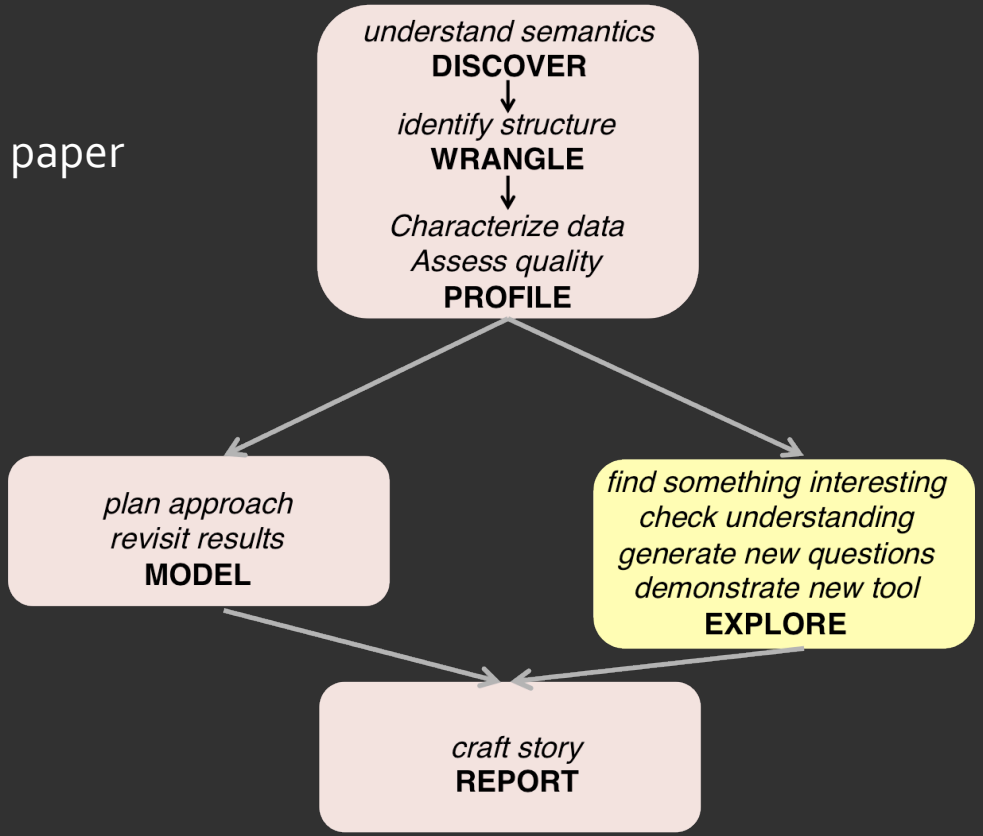
find something interesting
check understanding
generate new questions
demonstrate new tool
EXPLORE

Opposition to Exploratory Analysis

- Concern about **"fishing expeditions"** and spurious correlations (2/30)
- **Open-endedness** is likely to result in wasted time (3/30)
- **Data is expensive** to pull together, so it does not make sense to gather it without a well-defined goal (3/30)

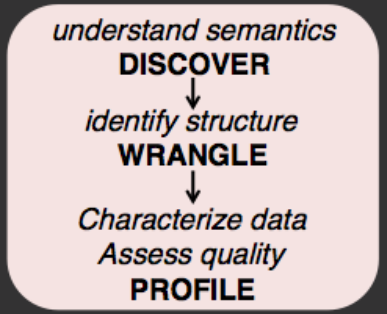
Augmented Model

Italicized parts are contributions of this paper
They show where exploration occurs.



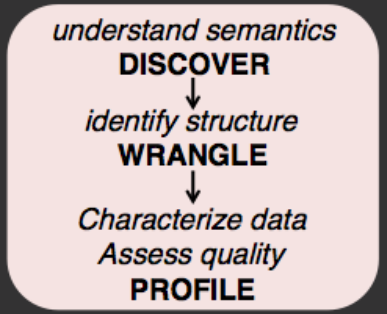
Discover Phase: Understand Semantics of the Data (21/30)

"I am dealing with an application now that has 2700 tables… Just by looking at the name a lot of the times you can figure out 'I need to see invoicing, so I going to look at invoicing tables…I am going to find how they relate. In doing that, I can start to narrow down what tables do I really need to understand about more and those are the ones I dig into."



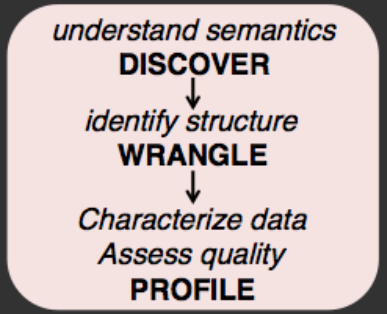
Discover Phase: Understand Semantics of the Data (21/30)

"… [continued] Maybe there is a ninth one that they haven't thought of but you will find it. You will come back and say 'Here is a code in one table I don't know what it means, but it is just a number.' 'Oh, well that just links to another table that has a description.' And you do this incremental process. It takes a little bit of time."



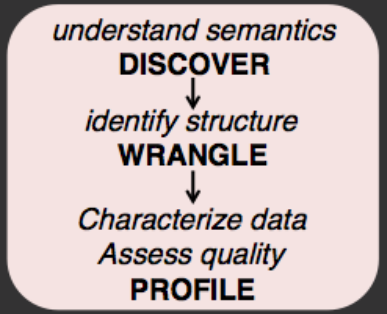
Discover Phase: Understand Semantics of the Data (21/30)

"I guess step one is describe the table that I'm looking at…Describing the table so that it's not too impossible. Sometime and then I'll check the wiki that we keep… If there isn't, I'll do an aggregate to see what kind of values it can possibly take and then usually I'll save the results of that query to a little note that I keep to myself.



Profile Phase: Characterize Data (25/30)

"if I have this data and I don't really know anything about it. I don't know which ones are correlated. So probably the first thing I'll do is I'll get a correlation matrix. You know, what are -- how did the scores on one question correlate with the ones on another. Or you know, what do the histograms look like? Does everyone answer this one question one or five and no one answers in the middle."



Model Phase: Plan Analysis Approach (16/30)

Often borderline between exploratory and directed
Includes:

- Scoping the problem in the context of the data
- Then deciding upon the analysis plan
- Also, defining metrics for later monitoring and analysis

"The open-ended question really came up first three months ago in first defining what these metrics should be. It's based on how I define things like script creation, or what sort of content creation metric that we're tracking.."

plan approach
revisit results
MODEL

Report Phase: Craft Story (3/30)

"…working through the data and trying to see what would be the best way to present the data to them in a way that they go, oh, I get it."

FINDINGS

Visualization and Automation

Most Use Visualization

28 use standard charts

4 use “advanced” visualizations (Tableau experts)

9 use interactive visualizations

7 said they do not

- 1 frequently re-plots instead
- 2 used tools with the capability but lacked support
- 1 wishes RStudio had it

Many Use Visual Analytics Tools

14 use Tableau, SAS, Splunk, Stata, Alteryx, Periscope, and others (all of these confirmed using them for exploring data)

12 said it was their primary tool

2 Do not use the tools interactively (i.e., Tableau for reports only)

Many Use Notebooks

13

Mentioned Jupyter notebooks or similar (e.g., R Markdown)

(primarily programmers)

Benefit: easier to document the analysis process

Drawbacks: hard to create rich interactive visualizations, bad engineering analysis pipelines, hard to extend.

Homegrown Automation

19 Described tools they had built themselves for repetitive tasks (including wrapping common visualization commands)

3 Wrote code to profile all columns of a data set

1 Wrote their own visualization library

Homegrown Automation

(continued)

12 Copy-paste reuse: many scripts with minor variations, hard to manage

6 Barrier to home-grown automation: difficulty of generalizing solutions so others could use them.

Many other frustrations

Computer Automation?

5 Expressed interest in automated wrangling tools

3 Pointed out that manual wrangling yields valuable insight

7 Suggested tools for automatically profiling data

9 Expressed skepticism

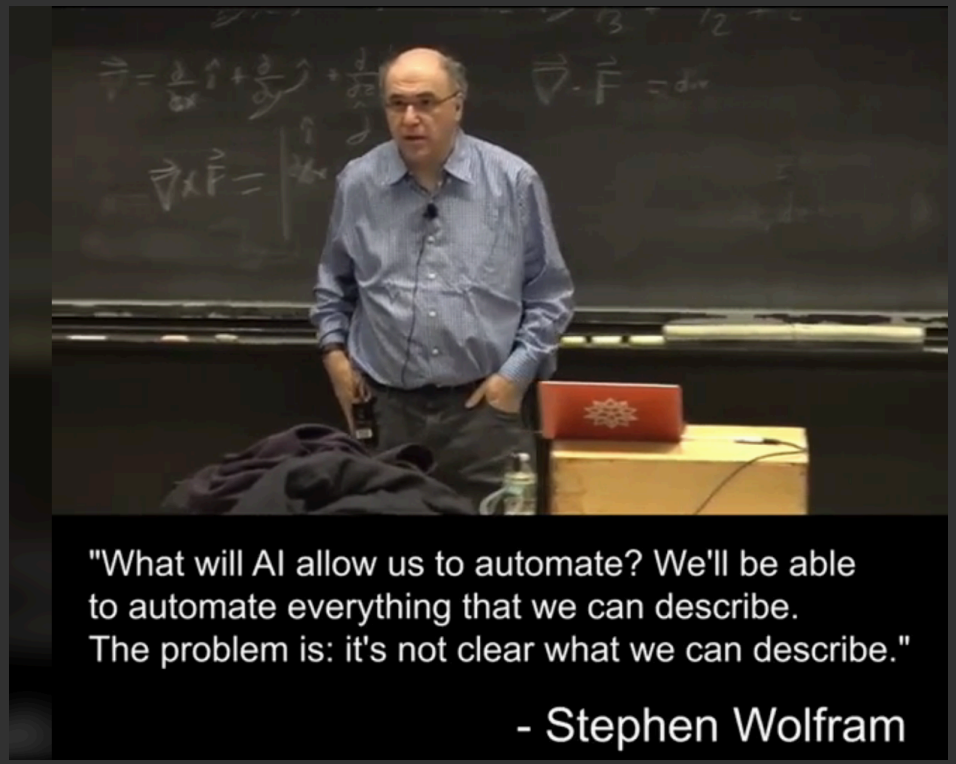
"the parts that are easy are easy; the hard parts are difficult to automate"

Automation?
(continued)

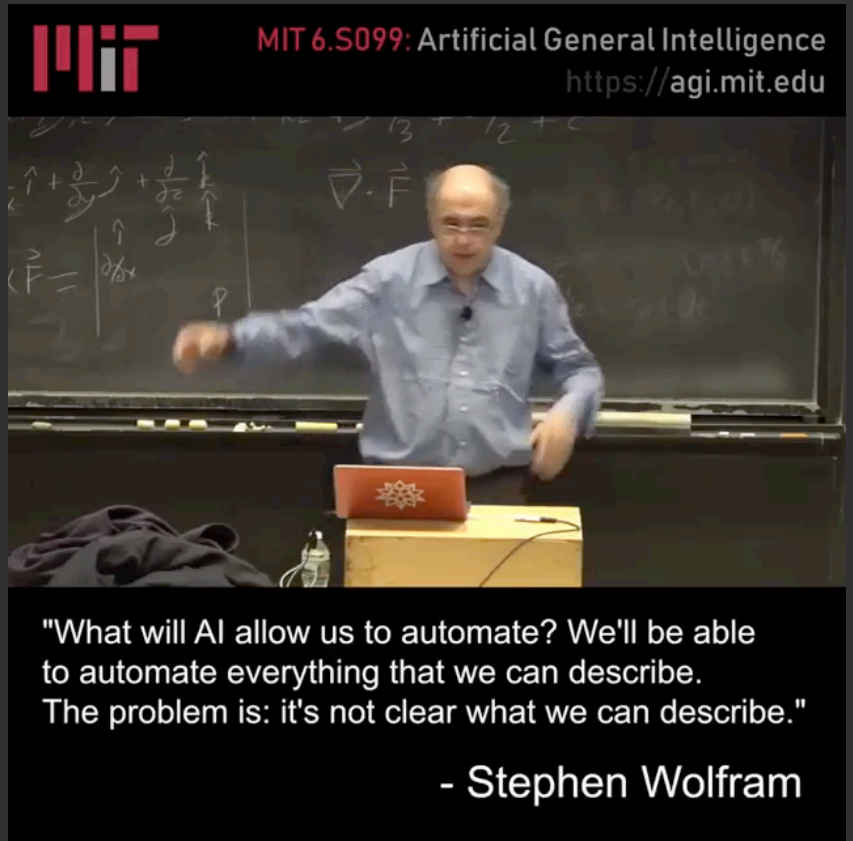
3 Expressed interest in automated suggestions of interesting relationships

12 Thought that for recommenders to be useful, they must navigate between being a black box and making the user do tedious work.

The Hard Part of Automation



The Hard Part of Automation



Outline

Motivation / Definitions

Related Work

Method

(Recruitment, Interviews, Coding)

Findings

Ramifications

RAMIFICATIONS

Improves an existing data analysis workflow framework by including open-ended exploration activities

Empirical justification for tool builders' intuitions

Suggests for aligning research and practitioner interests

Identifies promising research directions for continuing focus

reusability, provenance,
combining direct manipulation and programmatic interfaces

LIMITATIONS

Unbalanced gender split

Not a random sample

Data was collected in 2015

Questions?

Interviews with Professional Data Analysts on Exploration Practices

Sara Alspaugh, Nava Zokaei, Andrea Liu, Cindy Lin, Marti Hearst
UC Berkeley

VAST 2018

Supported by the UC Berkeley AMP Lab and a Gift from Tableau Research