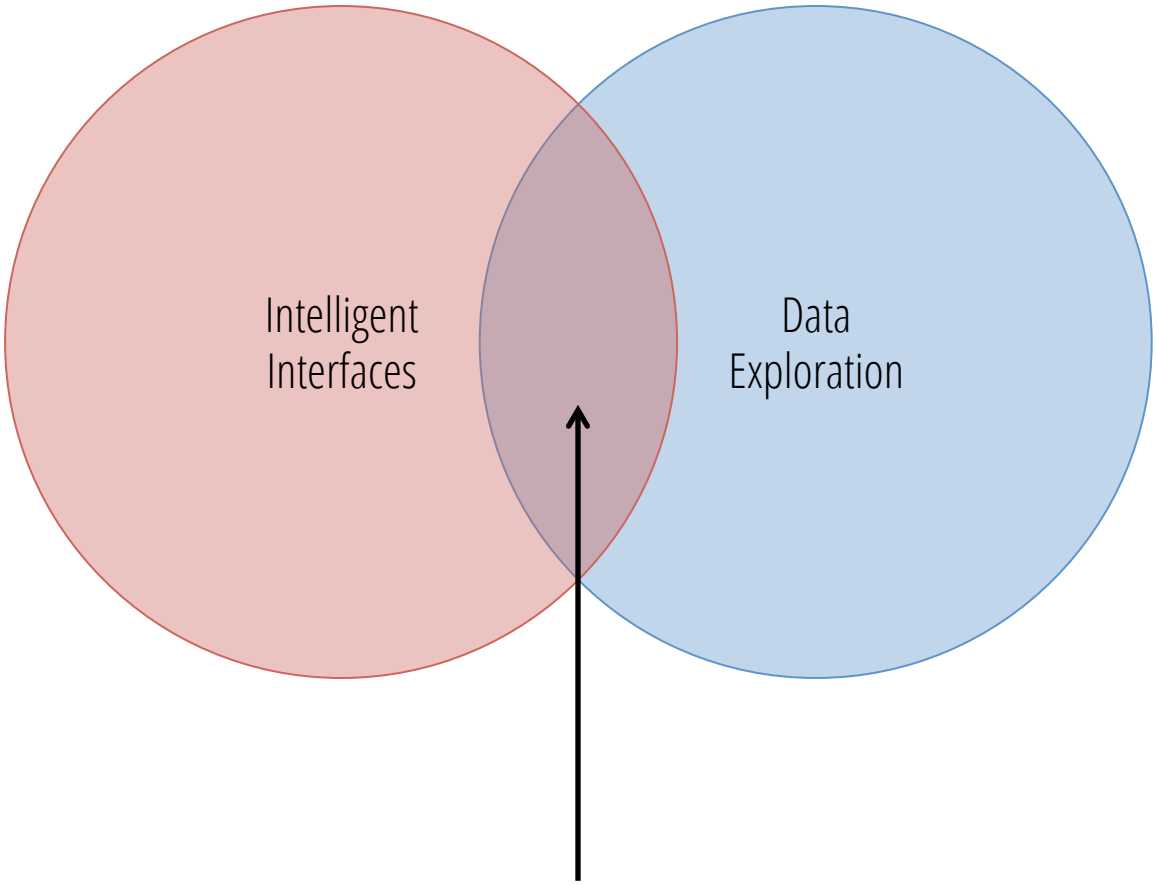
Data Exploration in Practice: Interviews with Expert Analysts and Directions for the Future

Sara Alspaugh
PhD Candidate
UC Berkeley
advised by Randy Katz and Marti Hearst

Why Data Exploration

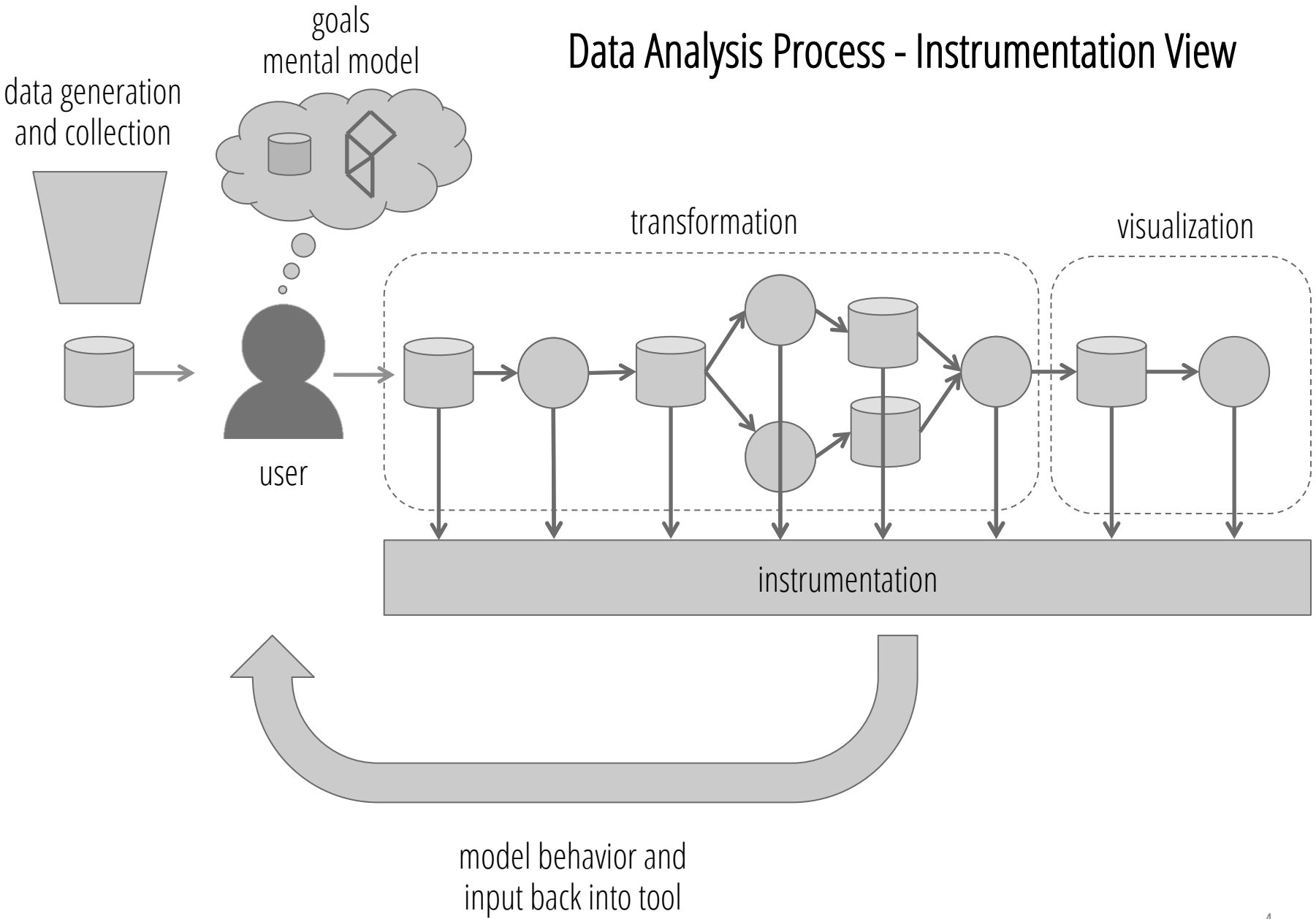


Thesis

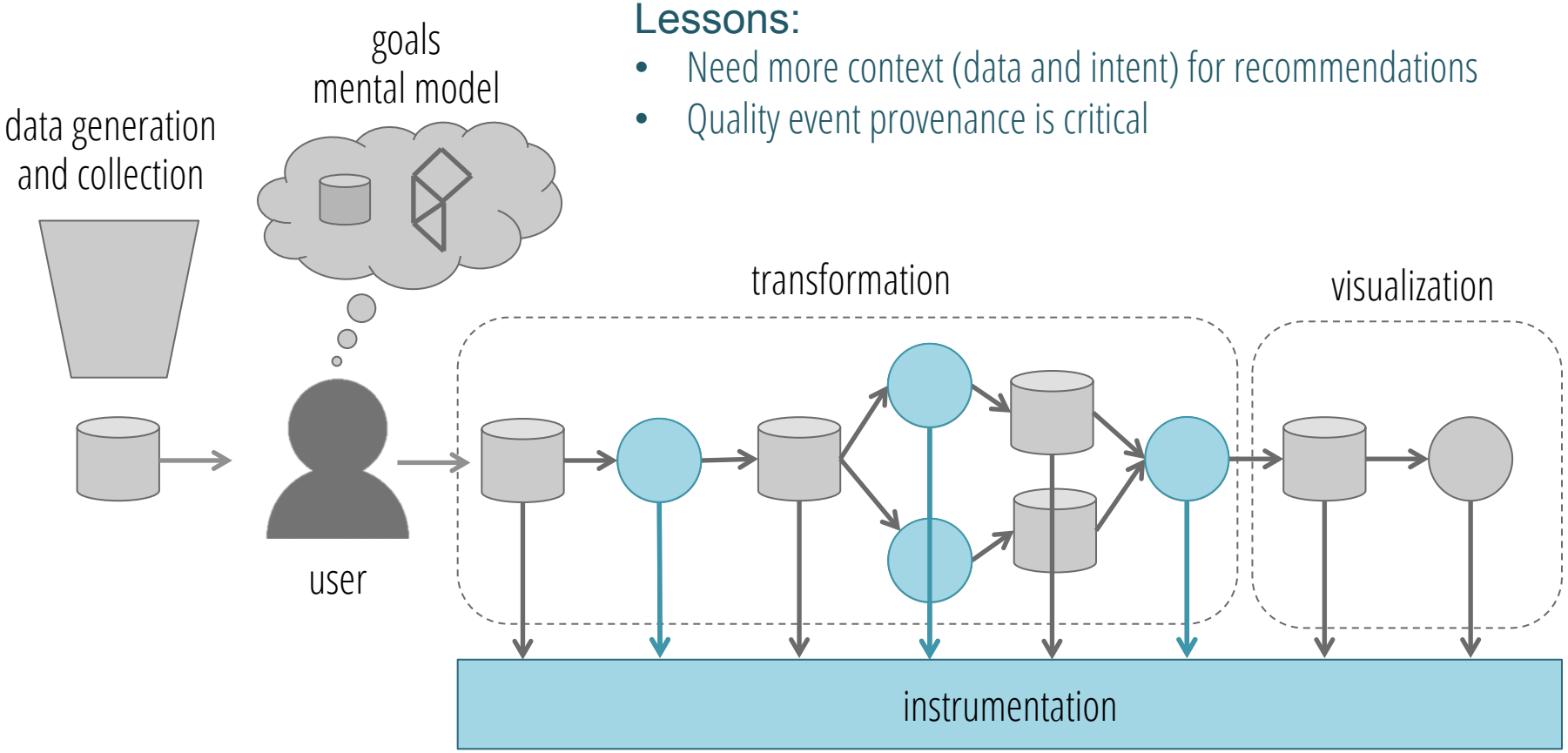
Model usage records logged from data analysis tools:

to characterize tool use
inform product design
feed into intelligent interfaces

Data Analysis Process - Instrumentation View



Usage Data Study #1: Splunk



Lessons:

- Need more context (data and intent) for recommendations
- Quality event provenance is critical

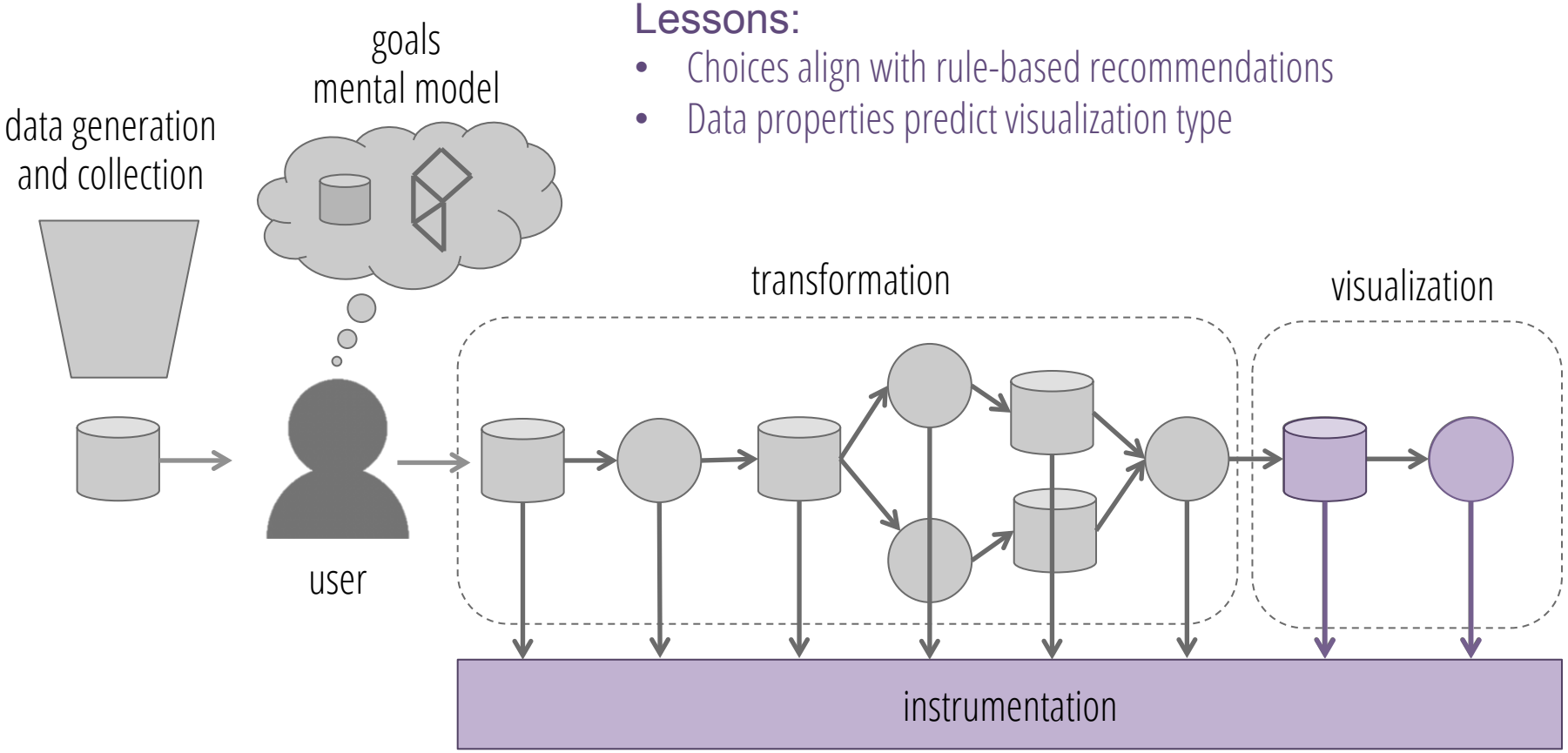
Results:

- Characterization of Splunk usage (still used by PMs at Splunk)
- Lessons for better instrumenting data analysis tools (requested by several startups)

Usage Data Study #2: Vis API

Lessons:

- Choices align with rule-based recommendations
- Data properties predict visualization type

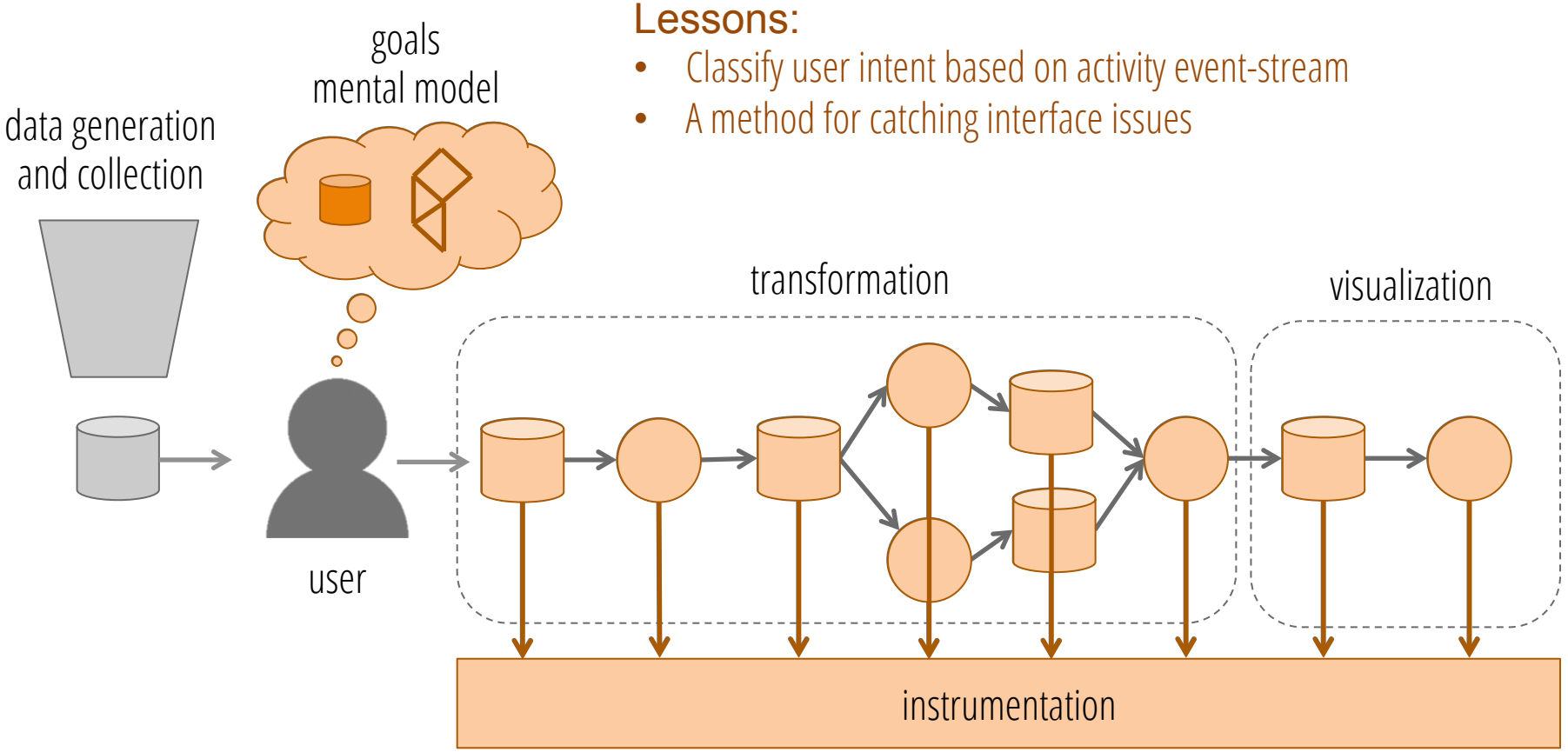


Results: In progress

Usage Data Study #3: Tableau

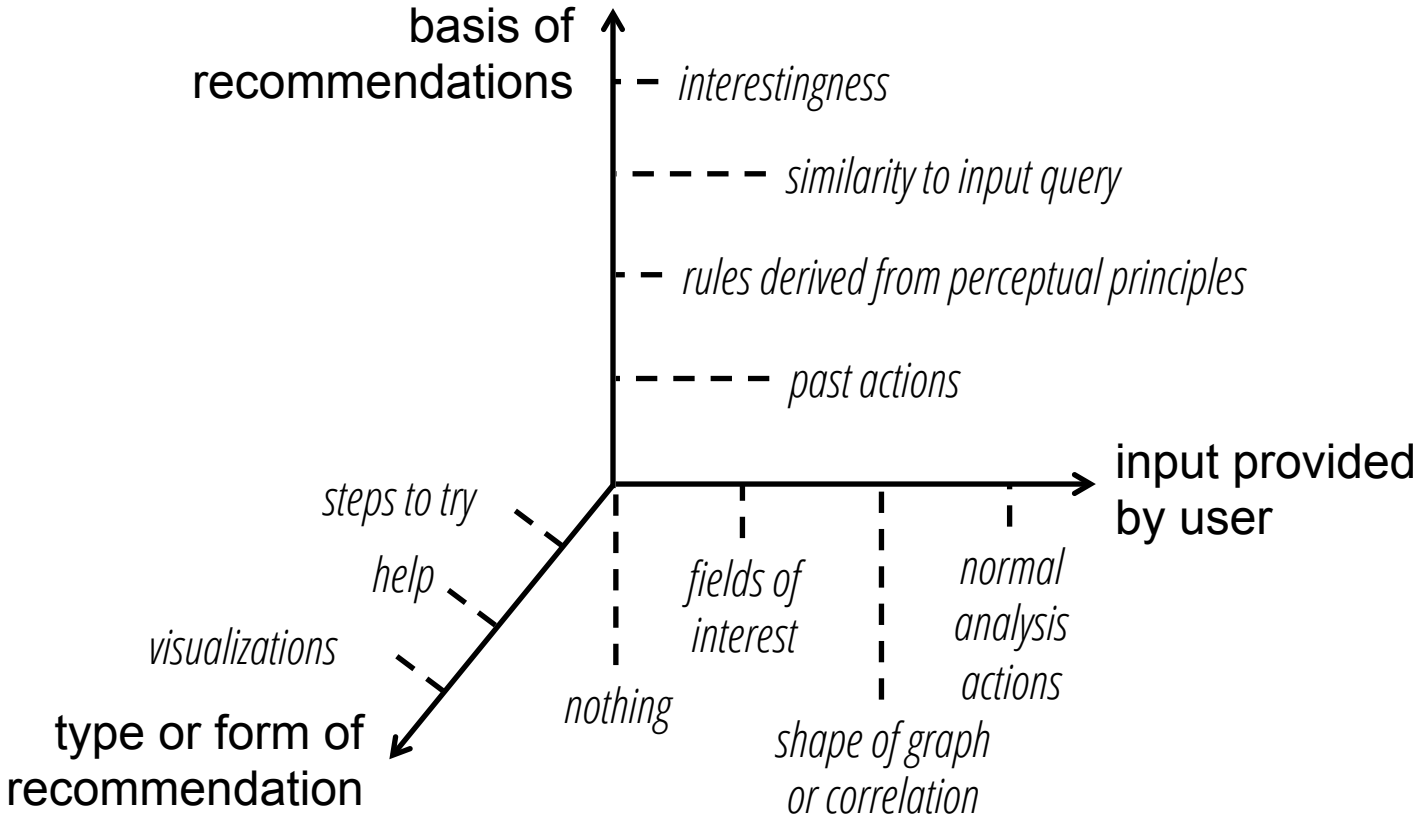
Lessons:

- Classify user intent based on activity event-stream
- A method for catching interface issues



STATUS: In progress

Why Data Exploration



There are many different ideas for intelligent interfaces for data exploration.
To **validate**, need to understand data exploration as it currently exists.

Who We Interviewed

- **Number:** 33 self-identified experts
- **Gender:** 91% male (30/33)
- **Sectors:** industry (26), academia (4), government (3)
- **Roles:** data scientists (15), academics (4), software engineers (3), consultants (3), executives, etc.
- **Industries:** tech (11), consulting (11), finance, healthcare, manufacturing, regulation, marketing. etc.
- **Location:** SF bay area (26), east coast, Australia
- **Education:** PhD or part (15), masters (10), less (8)
- **Experience:** 2 – 52 years (median: 9, average: 13)

Who We Interviewed

- Tools used:
  - Python (17), R (10), Scala (3), command-line tools (1)
  - Tableau (7), Excel (5),
  - relational databases and related tools (6)
  - SAS (2), SPSS (1),
  - Splunk (2), Periscope (1)
- Frequency encountering unfamiliar datasets:
  - Rarely (14)
  - A few times a year (12)
  - A few times a month (4)
  - A few times a week (1)
  - Every day (2)

Methodology

Recruited via email to lists and contacts

One to four hour interviews

Recorded and transcribed

This presentation: first pass to pull out themes

Next step: rigorous coding

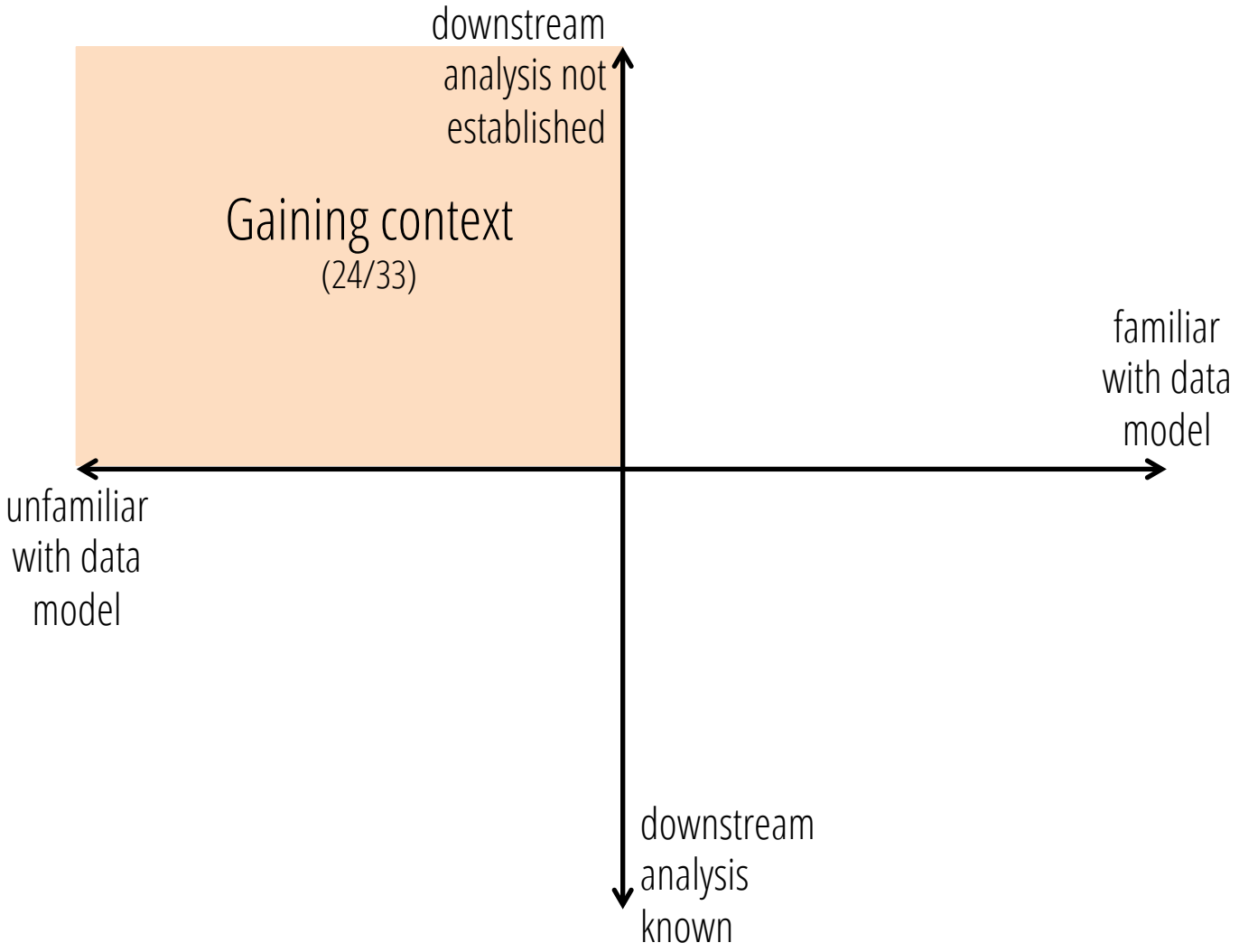
Main Questions

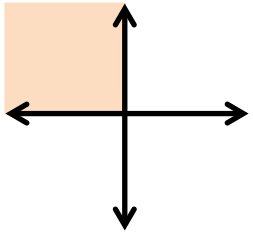
- What characterizes different data exploration **scenarios**?
- What data exploration **practices** do practitioners use?
- In particular, what **automation** do practitioners use?
- How and why do practitioners trade-off between **tools**?

Main Questions

- What characterizes different data exploration **scenarios**?
- What data exploration **practices** do practitioners use?
- In particular, what **automation** do practitioners use?
- How and why do practitioners trade-off between **tools**?

Types of Data Exploration

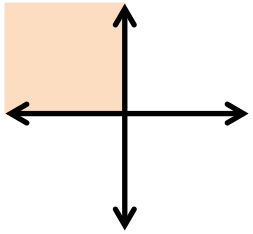




Gaining Context

Example Scenario

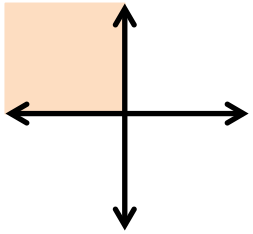
- Position: business intelligence consultant
- Role: help clients in different industries improve business
- Goals:
  - no need to fully characterize data or find interesting phenomenon
  - primarily need to identify subset of data to help client's problems
- Analyses:
  - exact analysis not pre-determined
  - established via data exploration, conversation with client



Gaining Context

Quote

I'm dealing with an application now that has 2700 tables. Not every one of them have I looked at. But I know what tables, just by looking at the names a lot of times, I can start to narrow down what tables do I need to understand about more, and those are the ones I dig into. You know most times, people have an idea -- they're giving you an idea. Well, here's a system, and we'll talk to the DBA or an application programmer and he'll say oh yea, you need to look at these eight tables, that has everything you need.

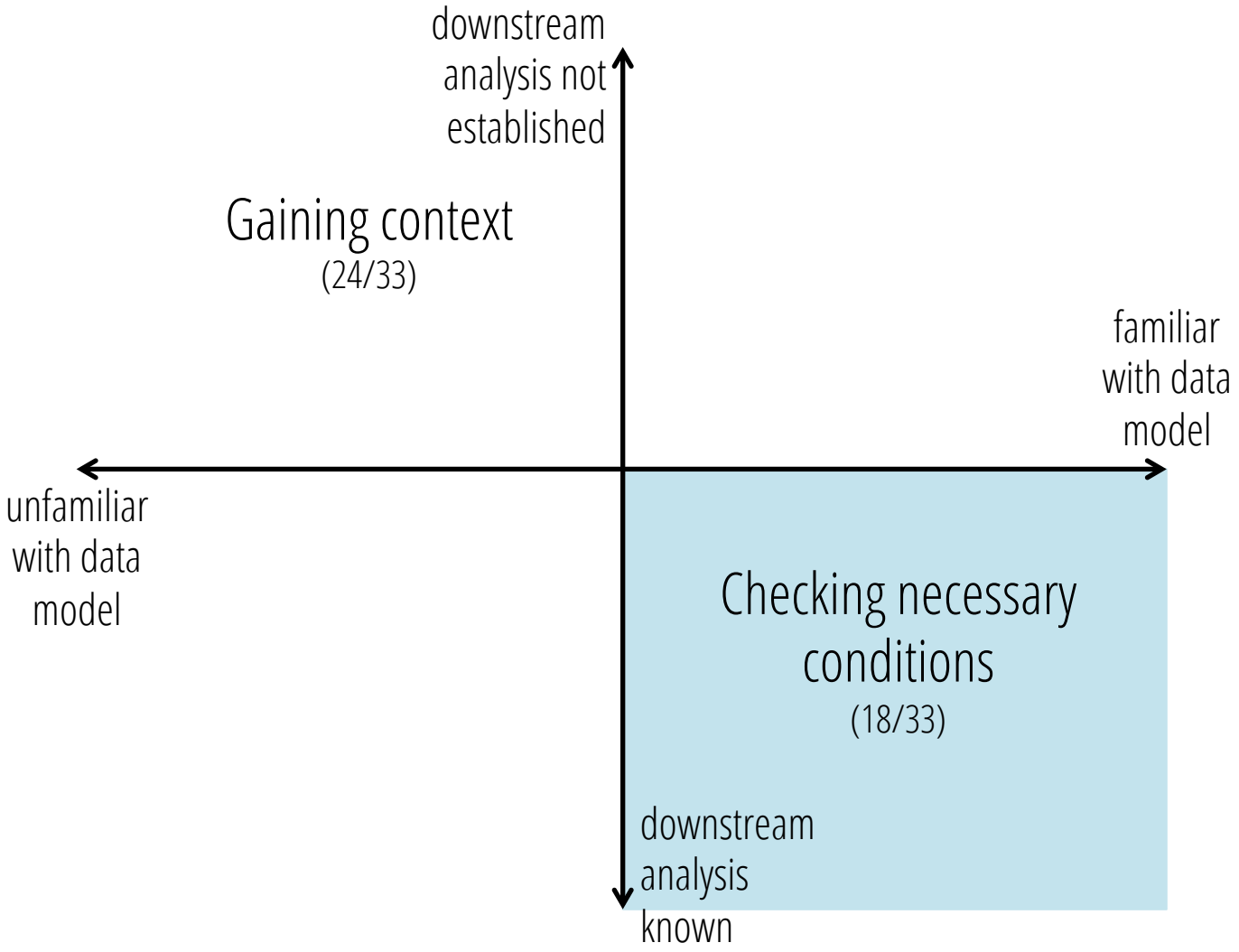


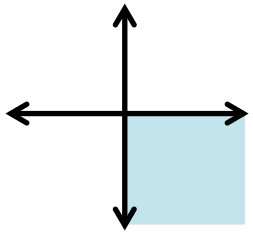
Gaining Context

Tools Needed

- Data discovery
- Detecting and describing data models
  - fields, meanings, structure, data types, distributions, etc.
  - exists for well-organized, relational environments
  - important in disorganized environments with complex data

Types of Data Exploration

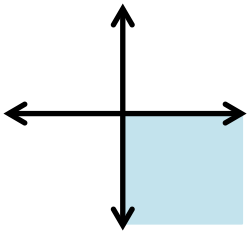




Checking Necessary Conditions

Example Scenario

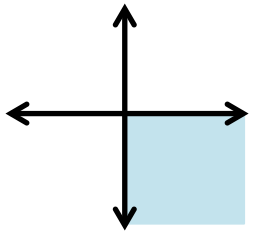
- Position: analyst at public health department
- Role: run quarterly reports
- Goals:
  - check data for changes or unexpected issues
- Analyses:
  - data domain and model is understood
  - code already exists or analysis problem otherwise well defined



Checking Necessary Conditions

Quote

It was just a lot of time spent trying to understand what changed from the last time you received it…Sometimes it would be issues like looking at why suddenly the length of one variable changed and sometimes you would find out later it was because one of the test assays expanded the possible response values, and so it was legitimate. But if you had SAS code that didn't account for that length you could be truncating values. So that was where it was interesting. Even though you had SAS code for things, you still had to double check that it was appropriate to keep using those assumptions of things as simple as variable length.

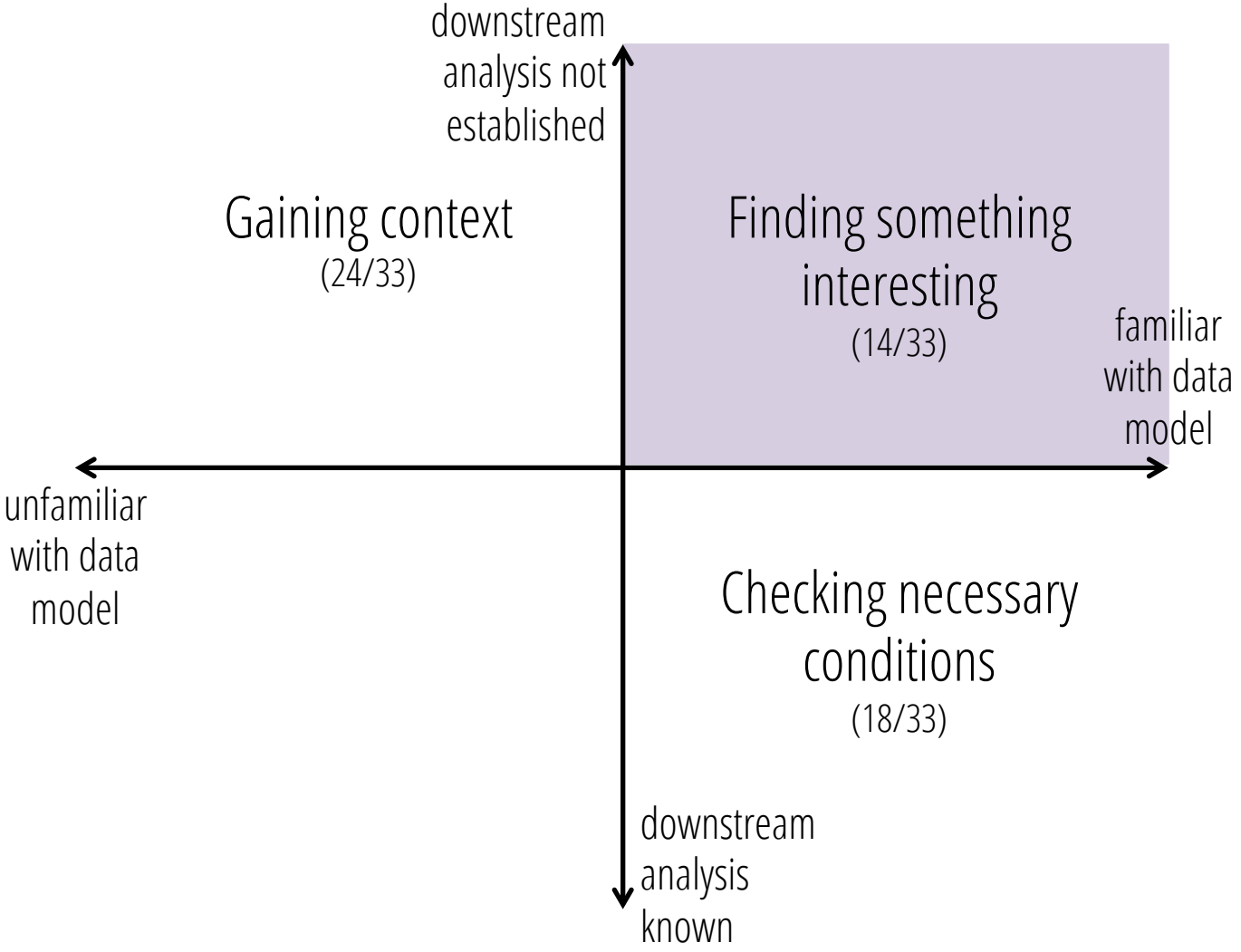


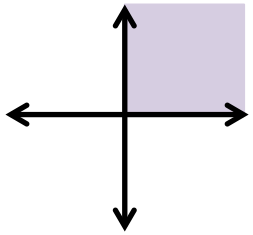
Checking Necessary Conditions

Tools Needed

- Expressing assumptions of downstream code
- Automating checking of input conditions
  - not always feasible
- Making better data generation process documentation
  - provenance
  - metadata

Types of Data Exploration

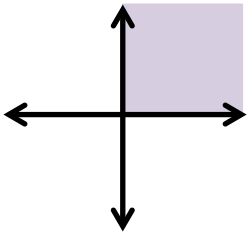




Finding Something Interesting

Example Scenario

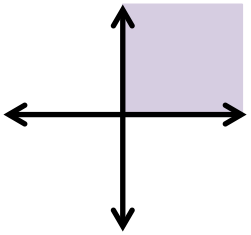
- Position: executive at data science consulting firm
- Role: help clients with data science problems
- Goals:
  - find something interesting (relevant, reasonable, unexpected)
- Analyses:
  - data domain understood through gaining context phase
  - exact analysis for finding interesting things not known



Finding Something Interesting

Quote

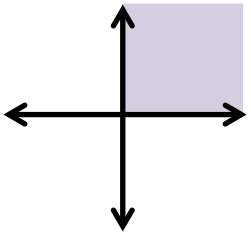
You're trying to understand this thing that isn't well described. A lot of times, this still happens with surprising regularity, we get clients who are just like, can you tell me what's interesting in this data and how I can make a bazillion dollars in it because I've read this article in Forbes and it says that there's gold in these there hills, and all I need to do is take my data and exhaust it, I can turn it into money...I would say that it's a reasonable hypothesis that you can do many useful things for people even without understanding [their] specific intent.



Finding Something Interesting

Example Scenario

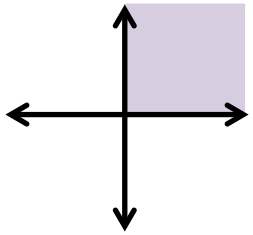
- Position: computer science networking researcher
- Role: publishes papers about Internet phenomenon
- Goals:
  - find something interesting (relevant, reasonable, unexpected)
- Analyses:
  - data domain understood through gaining context phase
  - exact analysis for finding interesting things not known



Finding Something Interesting

Quote

What I would like, but don't have, is the ability to say, just crunch every single distribution and identify ones that are pretty. Now "pretty" takes some thought, how do you define that. But ones that have a certain sort of succinctness or surprise to them, one or the other. So succinctness meaning, once you do this plot, you can describe the gist of it, really quite simply -- it's linear, it's Gaussian, whatever. Surprising meaning you do this plot, and there's two groups and boy they're different. That's often what the manual process strives towards.



Finding Something Interesting

Tools Needed

- Identifying anomalous or unexpected patterns
  - challenge to avoid emphasizing coincidental correlations
- Explaining remaining phenomenon coherently

Main Questions

- What characterizes different data exploration **scenarios**?
- What data exploration **practices** do practitioners use?
- In particular, what **automation** do practitioners use?
- How and why do practitioners trade-off between **tools**?

Main Questions

- What characterizes different data exploration **scenarios**?

  Familiarity with dataset and whether downstream analysis is known.

- What data exploration **practices** do practitioners use?

- In particular, what **automation** do practitioners use?

- How and why do practitioners trade-off between **tools**?

Main Questions

- What characterizes different data exploration **scenarios**?

  Familiarity with dataset and whether downstream analysis is known.

- What data exploration **practices** do practitioners use?

- In particular, what **automation** do practitioners use?

- How and why do practitioners trade-off between **tools**?

Main Questions

- What characterizes different data exploration **scenarios**?

  Familiarity with dataset and whether downstream analysis is known.

- What data exploration **practices** do practitioners use?

- In particular, what **automation** do practitioners use?

- How and why do practitioners trade-off between **tools**?

Opportunistic Behavior

Hoped to identify exploration checklist to plug into tools

Many analysts could not be precise about their process

Define goal as something that the achievement of which can be effectively measured

- run one mile
- find out if sales increased or decreased

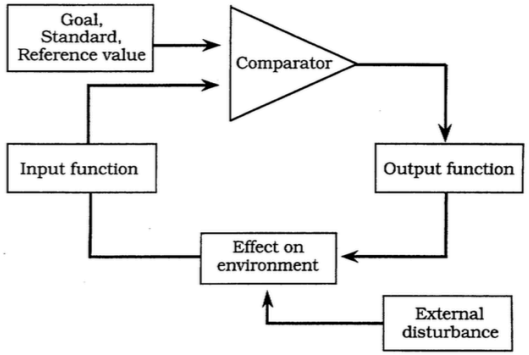
Opportunistic Behavior

Imprecise Goals

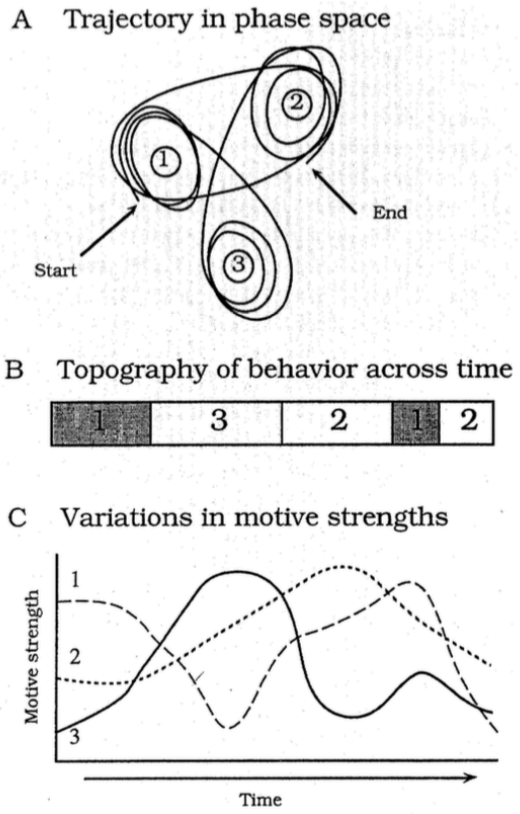
[There's] a lot of putzing, a lot of trying to parse our text logs to see if I could find anything helpful…[Putzing is the] same as like futzing…Kind of moseying…I don't know, just poking around with things and see what happens.

I start, you know, just start, you know…summarizing it. Start playing around with looking at -- because I'm very much, I come from the world of time-series data. I come from the world of the FFT. I come from, you know -- so those are the things that come naturally to me to look for. I look for their recurrences. I look for ways to sum over time, look for things like that.

Opportunistic Behavior



behavior: top-down goal-driven
evaluation: identify goal and measure progress to it



behavior: opportunistic data-driven
evaluation: ?

Workflow Tracking

- Many analysts did not track work and thought process
  - Partial exception: Jupyter notebook users
    - Unclear they would track if not for built-in aspect
  - Partial exception: command-line or IDE history
    - Unclear this is ever really referenced again
  - True exception: one data scientist using OSX notes
- Good idea: automatic tracking **and linking** of provenance

Workflow Tracking

Tracking Thought Process

I use just the Mac OSX notes app. I do a nice little job of formatting it to have my thought process up top, then the results of the query and then the conclusions that I draw from it. Usually I'll have one folder per project and anytime something comes up weird about that project I'll say here's the query I used to demonstrate that something is weird. Like if I think that a field isn't being populated anymore, I'll have a query that filters by date and then aggregates all the values in it and I'll say like hey look, overall there's this nice distribution between the four categories you list yourself as but ever since June it's all been nothing.

Workflow Tracking

Analysis Provenance

Easy annotation and then slicing of the data related to the annotation would be great. So [often] there are artifacts in [a] plot. And so, the first thing I want to do is annotate all the artifacts that catch my eye. And then often the answer for exploring one of them is now slice the data on that. And that can be a point -- that's easy -- or it can be a region, which is sometimes not as easy because you are going to mix in some things of one type and some things of another…I would love to be able to say, tell me things that differ between these, and not the fact that they're different places on this plot, I already know that. But these points are not just pairs, (x, y) pairs, they're linked to entire records.

Workflow Tracking

Analysis Provenance (continued)

An environment that would preserve those linkages so that I can point to things on this plot and say that essentially, that point's provenance, it started from here, here, and here in the raw data, and it went through this, and so forth, these are all the attributes I have associated with that. That would be great. Because then I can look at that and say well how do you differ from this. Or for just an individual point, I can say, tell me about yourself, what is the raw data that went into you. Because sometimes for an individual point I'll say, oh that's bizarre! That thing is zero, that should never be zero. Or whatever. You start getting ideas. Today, that's really clunky.

Reverse Engineering Data Generation

- Data exploration textbooks use well-documented datasets
- IRL documentation is often non-existent, patchy, or incorrect
- Most spent time figuring out how given dataset was generated
  - Resembles reverse engineering or detective work

- Challenges analysts encountered:
  - data generation process changes mid-collection
  - information at best locked-up in other people's heads

Reverse Engineering Data Generation

Lack of Documentation*

A lot of the data isn't really documented very well and each specific component of different data sources is owned by different teams...You may know what one or two fields mean, but you don't know what the other twenty do. Either you have to find who knows how to do it, whoever's working with it, or you have to do some exploratory analysis to find out what's there. I've been in many situations where I've been like, I think this field is something I need because the column name makes sense to me, so I'm just going to pull it out and see what's there, and it turns out it's full of null values because it's something some team decided they wanted but then they didn't end up recording any of that information. Then the list of columns is exhaustive, so if you really wanted to try to understand every single column and figure out each one you wouldn't do anything because you'd spend so much time just trying to understand this encyclopedia of this information.

Reverse Engineering Data Generation

Reverse Engineering

I'll go to a user and I'll say, do me a favor. After hours tonight, I want you to add one new invoice. I want you to bill someone. And when you're done, I want you to pretend like they paid it. Use a dummy invoice. Then I'll go look at the table catalogues and say alright, what tables have just been updated in the last twenty minutes? That tells me, from that application, every single table that's influenced when they do a billing, when they get money in. Every table, and maybe there are fifteen of them. But out of hundreds of tables, I know which fifteen I need to deal with…That's a very common way of analyzing data when you don't have good documentation, or you don't have … a database programmer or an application programmer.

Reverse Engineering Data Generation

Inaccessibility of Information*

We're in a situation where all of our knowledge isn't even like regular old talk-to-someone accessible. You can't even get people to respond to emails asking like, hey remember that thing that you built for us, we paid you like many thousands of dollars to build, could you tell us how it works?

Reverse Engineering Data Generation

Tools that help: data integration, wikis, digitization, CIM

Some analysts skeptical that this could be automated now

To improve: focus on data generators

Main Questions

- What characterizes different data exploration **scenarios**?

  Familiarity with dataset and whether downstream analysis is known.

- What data exploration **practices** do practitioners use?

  Standard, domain-specific techniques opportunistically, plus detective work.

- In particular, what **automation** do practitioners use?

- How and why do practitioners trade-off between **tools**?

Main Questions

- What characterizes different data exploration **scenarios**?

  Familiarity with dataset and whether downstream analysis is known.

- What data exploration **practices** do practitioners use?

  Standard, domain-specific techniques opportunistically, plus detective work.

- In particular, what **automation** do practitioners use?

- How and why do practitioners trade-off between **tools**?

Main Questions

- What characterizes different data exploration **scenarios**?

  Familiarity with dataset and whether downstream analysis is known.

- What data exploration **practices** do practitioners use?

  Standard, domain-specific techniques opportunistically, plus detective work.

- In particular, what **automation** do practitioners use?

- How and why do practitioners trade-off between **tools**?

Spectrum of Re-Use

- Where there is re-use there is opportunity for intelligent interfaces
- Removes friction of having to think about low-level steps to take
- Analysts fell along spectrum of re-use
  - Most sophisticated: R package that automatically graphs of all fields
  - Least sophisticated: copy-paste re-use or none
- Re-use was more common among programmers
  - programming more expressive at expense of being more verbose
  - re-use not well-supported in some GUI environments
  - re-use less important with efficiency of expression

Spectrum of Re-Use

Not Having to Think Low-Level

Some people use cut, and I used to use awk... one thing I realized is that generating these tiny little things that do something I can do with more key strokes, it's worth it. My mind just starts working more quickly, knowing I can just grab that. I need column five, give it to me. As opposed to "awk quote" ... or whatever, "cut". I have a bunch of those very low level ones, and then I've got another that ... I swear everybody who does command line analysis has written this script which I call "unique count" and then I alias it to "uc", so that's all I ever type. All it is, is it's: sort the data, count, and now sort .... on the count. I want the top N things. That's it. I've seen other people type that longer version of it over and over. I'm sure I did, too, and eventually I just scripted it.

Spectrum of Re-Use

Automatic Profiling Packages*

I've been doing data profiling stuff for a long time. If you go to my Web site, there's an archive presentation I did 10 years ago on data profiling. The idea is basically, if somebody gives you a data set, and the first thing I do is, I do a profile. I do solution plots variable by variable, and looking for funny stuff.

Spectrum of Re-Use

Copy Paste Re-use

I ended up having the same exact Python script with very minor variations. But instead of having it with like, beautiful git commit history or anything like that, I just replicated the file approximately 25 times with time stamps embedded in the file name so that I would know when I'd made it, because I was able to match that to what I was doing.

I do have commands built for part of this but I still have to kind of struggle.

Spectrum of Re-Use

Re-use in Tableau

The template doesn't support reusability. We can kind of hack it to get it so that we can reuse something. I created a workbook the other day for court diagrams, or you created one for process control that was kind of reusable...It's not well supported...We can try to do it, but it's still not what Tableau was designed to do.

Opinions on Intelligent Interfaces

- Some drawn to appeal of fewer steps to do something
  - lower level steps are done automatically
- Spectrum from excitement to indifference or skepticism
- Reasons for skepticism:
  - distrust in tool output, NLP-skepticism
  - inability to clearly define work and decision conditions
  - thinking it will cause shift in work rather than reduction
  - fear of distance from data, need to interact with it
  - belief that creativity and human interaction aren't automatable
  - worries about spurious correlations, p-hacking, fishing expeditions

Opinions on Intelligent Interfaces

For: Not Having to Think Low-Level

A lot of the time, regardless of the data source, there some sort of complex transformation that's needed. And so rather than running twenty lines of code it would be great if there was something that says I want to do this this, this, and this, in this particular order. Go and do it. So I think some sort of short cut for what is effectively another series of operations on the same data set, would be useful. So there's a package in R that kind of is handy in this regard. You still end up running a lot of code but you can kind of pipe one result into another into another and so it's called magrittr...that's kind of a step in that direction because it condenses the amount of work you do down to one line of code.

Opinions on Intelligent Interfaces

For: Automate Generation of Visualizations

That would be awesome. I'm just not pro enough to pull a generalized solution for that sort of thing.

Opinions on Intelligent Interfaces

Against: Automation Merely Shifts Work

I think I really don't like tools that do any sort of code generation where I could do it myself because I feel like then I start using the tools so that it generates the code that I would have written had I not had the tool, right? It's kind of similar to an automatic transmission in a car. If I want the car to shift, I have to push the pedal down harder so that it knows I want the car to go faster so that it downshifts or whatever. It's the same thing.

I think I'd generally much rather have a high-level language that lets me precisely express exactly what I want to do than to have some tool that generates something and then I have to look at that.

Opinions on Intelligent Interfaces

Against: Can’t Automate Understanding

Where Spotfire says, "Spotfire will automatically examine your data and recommend the best visualizations for it." That it will be the future of analytics…The response to that is: that the notion that the software can automatically enable people without analytical skills to make sense of data is ludicrous.

Opinions on Intelligent Interfaces

Against: Can’t Automate Creativity

Interviewer: *If you had to write an if statement that described ... I really like the guided exploration idea, I kind of think of you as like a roomba. (laughter)*

Interviewee: *I see it as like this binary tree, that's just branching out, branching out, branching out. And that there's lots of x's. X, x, x, x, x, x, and there's one check mark. And that's how I know when I'm done.*

Interviewer: *Could you tell me what makes a thing an x?*

Interviewee: *If it doesn't work, if it's not effective, if it's not ...*

Interviewer: *What does doesn't work mean, though? Could you be more precise? If I were trying to write code, I wouldn't be able to say, if doesn't work. What is the actual condition?*

Interviewee: *Is it interesting? Is it illuminating? Does it drive knowledge that is important to make a decision in the business?*

Interviewer: *How do you decide if it's interesting?*

Interviewee: *I'm a human being. (laughs)*

Opinions on Intelligent Interfaces

Against: Can’t Automate Creativity

I think there's a misconception about data analysts and data analysis because we work on a computer that we are algorithmically driven in some way. What I have noticed of the people that are really good at that, at analyzing data, at presenting data visualizations, and certainly everyone at this table is that we've had very wide experiences in our lives, and we've been artists and scientists, we have that sort of Leonardo da Vinci sort of mix of the two and bounce between extremes. I think the idea of a computer doing data analysis or data prep, it's like you take a block of marble and you see what's inside before you start carving because you have a computer tell you what it is.

Opinions on Intelligent Interfaces

Against: Can't Automate Human Interaction

I would say like 90% of my work is working with somebody else.

Opinions on Intelligent Interfaces

Against: Need for Interaction with Data

I've seen the demos of some of these tools. They're supposed to go in and they'll put up network graphs and show you all of these correlations out of your data automatically and everything. For me, I look at the demos and there's this bit of sterility there for me around it. That bit of mucking around in the data and looking at it gives me this much more visceral feel for the data. I feel it, I know it, and when I see this, it's different. There's all these associations around it that when I'm presented with a whole bunch of stuff on a screen all at once, it's hard to know what's important and what's not, but when I've been mucking around with it, there's that whole physical piece.

Opinions on Intelligent Interfaces

Against: Garden of Forking Paths

I put myself very firmly in the pre-establish the goal, figure out how to walk towards to the goal [camp]. Like the reason that I say that I don't do EDA is for exactly that kind of…like I don't really want to nebulously float around and play with the data. That sounds like a recipe for a multiple comparison to end all multiple comparisons. Anything that you find through this is: okay you learned something along the way, you make some pretty neat graphs. But like as a statistician I can never guarantee that…it will generalize well to the larger population…"Hey look this is neat" has always been too risky for my blood. I've read so many things where like at the end of it, I'm like, I didn't think that was neat…My solution is to come up with a problem where you need a measurable answer, if you don't have one of those then I'm not sure what to tell you.

Opinions on Intelligent Interfaces

Against: Wants Less Computer

I think, for me, one of my things is, I want my software to disappear. I want the computer to get out of my way and let me do what I want to do. I don't want to have to fight with the software and figure out some weird combination of setting that it needs to be just right for me to get the answer I want...I want to pay attention to my data...[Software] takes me out of the flow. Now I've got to sit here and figure out, all right, now I've got to think like the engineer of this software...I think what we have is a lot of good stuff, but it's not invisible enough. I don't need new logic. What I need is better software that allows me to be more me instead of me trying to emulate a computer. That's why I'm saying, what Bret Victor says for the future, that's what I want. I want that direct manipulation. I don't want to have to run a whole bunch of code. I want to say, take this, do this, now take this and apply it to that, and do this with it, and then let it do it. Let the computer do what the computer's good at.

Main Questions

- What characterizes different data exploration **scenarios**?

  Familiarity with dataset and whether downstream analysis is known.

- What data exploration **practices** do practitioners use?

  Standard, domain-specific techniques **opportunistically**, plus detective work.

- In particular, what **automation** do practitioners use?

  Varies. Re-use challenging, varying in importance. Some suspicion of automation.

- How and why do practitioners trade-off between **tools**?

Main Questions

- What characterizes different data exploration **scenarios**?

  Familiarity with dataset and whether downstream analysis is known.

- What data exploration **practices** do practitioners use?

  Standard, domain-specific techniques **opportunistically**, plus detective work.

- In particular, what **automation** do practitioners use?

  Varies. Re-use challenging, varying in importance. Some suspicion of automation.

- How and why do practitioners trade-off between **tools**?

Main Questions

- What characterizes different data exploration **scenarios**?

  Familiarity with dataset and whether downstream analysis is known.

- What data exploration **practices** do practitioners use?

  Standard, domain-specific techniques **opportunistically**, plus detective work.

- In particular, what **automation** do practitioners use?

  Varies. Re-use challenging, varying in importance. Some suspicion of automation.

- How and why do practitioners trade-off between **tools**?

Command-line versus Visual-first

- Friction of switching environments is pain for some
  - incur cost of switching tools, *or*
  - incur cost of doing something tool wasn't designed to do well

- Some want programmability and direct manipulation
  - Programmability for extensibility and control
  - Direct manipulation for efficiency in supported scenarios
  - Combination also helps cater to wider user base

Command-line versus Visual-first

Friction

Tableau is not doing what I want it to do, go to Google and look it up. As soon as I do that, I'm out of that flow to get back into the flow, as soon as I've switched into Google, oh, well why not check e-mail, somebody calls, something like that. It's easily taking me out of that space, and it really sucks up time.

Command-line versus Visual-first

Tool Limitations: Capabilities

Tableau also doesn't have iterative functions necessarily...whereas [in] Alteryx you can do those things.

Tableau is essentially a single query deep. If you cannot do it in a single query, you probably can't do it inside Tableau.

I use R for some stuff, I use Python for some stuff. If there was something that managed to combine the best of both world,s that would be great.

Tool Limitations: Control

Tableau does not expose control over those hundred steps, that how we have influence over those hundred steps, is indirectly and through a combination of hidden and undocumented ways.

Command-line versus Visual-first

More Programming in Direct Manipulation Interfaces

There were a set of human beings that developed that interface so they've made a lot of assumptions and decisions about how people work and what they do. And so there's some boundaries there that it creates... Programming languages have those built in as well. But the programming languages are at a lower level so...ultimately you can be more expressive. ...I think that creates some of the distinction there, some of the program people would never, their definition of what they need for control is big enough to be outside the bounds of what Tableau could ever do sort of thing. Because Tableau is limited by these constraints. For what a lot of the people in this room a huge amount of the time we can live within the constraints of Tableau and enjoy the things that we get from Tableau by being faster to get these results. ...The fact that Tableau's base choices work really well for so much of what I want to do is great. I really love that, that I don't have to re-enter which color [like] I would if I was working D3 or something like that. At the same time, I want to be able to open up the lid of that and change that and I want that tool to be responsive to me and learn from me. Which you could talk about it as, in one sense as templates and macros, to enable that, but I want it more open-ended than that.

Command-line versus Visual-first

More Direct Manipulation in Programming Interfaces

One problem I have with IPython notebook is that it's very hard to automate visualization...it's not something I can give the dataframe, and then have a Tableau interface in the notebook and then I can just manipulate it. I have to write a new, like, something every time, either in seaborn or matplotlib or something, right? I can't just visually do that. It's hard to make because the environment itself on the front end is, for me, completely opaque. There's no way--my environment itself does not lend itself to being automated...because of the way the Python server model works for me...I have never been able to understand the JavaScript side of things. ... That for me is a major shortcoming, which would easily address a lot of the issues, where it's easy for us to automate the code type of things, because that's just Python. It's hard to automate the visual side of things because then I have to mess around with this other environment that's not really easy to automate in....I want every cell to be like a GUI in itself where I don't necessarily have to write a lot of Python code to do what I want.... Every visual is interact-able, kind of like Bokeh, but in a way that I don't have to decide in advance what I want that to be. I just decide the data...Sort of like for every column it would have, could make histogram of this one, you know, a bar plot of this one.

Command-line versus Visual-first

Combined Direct Manipulation and Programming

A lot of the data team here serves business users or our nurse team or our press service team, people who just don't have, necessarily, backgrounds are uncomfortable for multiple reasons, which is fine. They don't have to. Splunk, we've built a fun, little power user structure that lets folks … It helps us be more data self-sufficient across a broader span of the company. Even though Splunk wasn't meant for that, it's nice to be able to say, "Hey, here's the dashboard full of the 10 charts you need. If you need more, push this button." We do a lot of tokenization. You push these radio buttons to get or you push these check boxes to get in the fields you want and we'll slice it for you. These wasn't the things Splunk was meant to do, but it's been really, really helpful in having minimal data resources but being able to just spread it out across the company to visualize their own data.

Command-line versus Visual-first

Programming for Control

The more fine grain control you want over what you're doing with your data the less room there is for just coming up with a slightly different GUI.

Main Questions

- What characterizes different data exploration **scenarios**?

  Familiarity with dataset and whether downstream analysis is known.

- What data exploration **practices** do practitioners use?

  Standard, domain-specific techniques **opportunistically**, plus detective work.

- In particular, what **automation** do practitioners use?

  Varies. Re-use challenging, varying in importance. Some suspicion of automation.

- How and why do practitioners trade-off between **tools**?

  Poorly supported use cases cause friction. Want programmability and direct manipulation especially for visualization.

Main Questions

- What characterizes different data exploration **scenarios**?

  Familiarity with dataset and whether downstream analysis is known.

- What data exploration **practices** do practitioners use?

  Standard, domain-specific techniques **opportunistically**, plus detective work.

- In particular, what **automation** do practitioners use?

  Varies. Re-use challenging, varying in importance. Some suspicion of automation.

- How and why do practitioners trade-off between **tools**?

  Poorly supported use cases cause friction. Want programmability and direct manipulation especially for visualization.

Main Questions

- What characterizes different data exploration **scenarios**?

  Familiarity with dataset and whether downstream analysis is known.

- What data exploration **practices** do practitioners use?

  Standard, domain-specific techniques **opportunistically**, plus detective work.

- In particular, what **automation** do practitioners use?

  Varies. Re-use challenging, varying in importance. Some suspicion of automation.

- How and why do practitioners trade-off between **tools**?

  Poorly supported use cases cause friction. Want programmability and direct manipulation especially for visualization.

Future Directions

- Dissertation: inform log analysis about workflow and tasks
- Post-dissertation: data analysis IDE
  - thorough instrumentation and provenance
  - aide for workflow tracking and re-use
  - programmability plus direct manipulation
  - recommendation for goal-directed and opportunistic behavior
  - test analyst ideas raised in interview