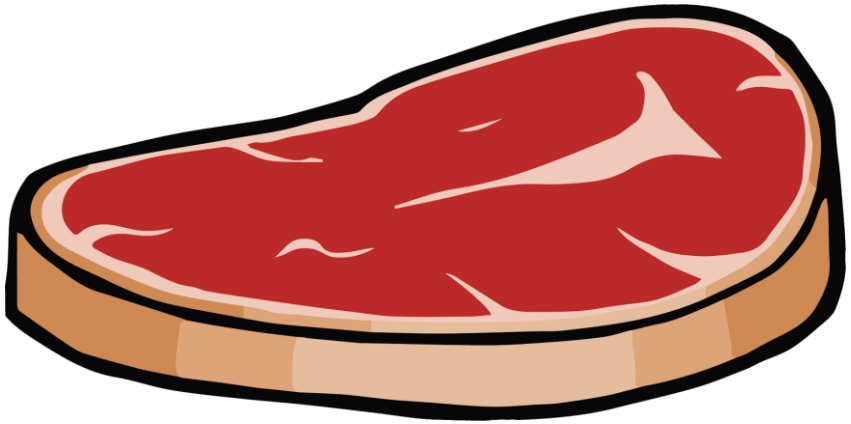
A DAY IN THE LIFE OF A DATA SCIENTIST

- ABHISHEK ARORA



THE MEAT OF IT

- Number of analysis and visualization tools to make the job easy for data analysts
- Little research on how analysis take place within the social and organizational context of companies
- Infrastructure, available data and tools, and administrative & social conventions can have an impact on the analysis process
- In this paper, we try to understand how these factors affect the analysis process

Why you want to do it?

- Understanding how these factors affect can help inform the design of future tools
- An opportunity for visual analytics tools to improve the quality of the process
- This can result in better results and lives for the data scientists

How did we do it?

- Conducted semi-structured interviews with 35 analysts
- From industries like healthcare, finance, retail, to name a few
- Type of information collected:

a) What typical tasks do they perform? Tools they use?

b) Challenges faced by them at the job

c) Studied how organizational features (infrastructure, collaboration etc.) of the company can affect the analysis process

The Participants

- 35 participants from 25 companies were interviewed
- Participants had titles like “Data Scientist”, “Data Analyst”, “Software Engineer”, “CTO”, “Consultant”
- Organizations includes small startups as well as big enterprises with thousands of employees
- Majority of them were located in Northern California

The Interviews

- Interviewed 1 to 4 analysts at a time
- Each interview lasted from 45 minutes to 2 hours
- Mostly in-person (used Skype in case of remote interviews)
- Examples of types of open ended questions asked:

a) What tasks do analyst perform?

b) What kind of data sources do they work with?

c) What tools do they regularly use?

d) Relationship between analysts and other business units?

Etc…

Types of Analysts Found

Hackers	Scripters	App Users
Most proficient programmers	Intermediate proficiency	Rely mainly on spreadsheets or apps
Comfortable in performing complex operations like manipulating data	Able to perform simple manipulations: Filtering, aggregation	Works on smaller data sets than other groups
Spent more time on early stage analytics prior to analytics	Applies models. Generally operates on data pulled by IT staff	Requires someone to pull & prepare their data
Performs less sophisticated models than scripters	Can’t write script that run at scale	Do not write much code
Visualization tools used: Tableau, Excel, Powerpoint, D3	Visualization done using statistical packages during exploration phase	Uses excel or reporting tools for visulizations

Organizational Context

- Enterprise analyst work within the context of a larger organization
- The political and social conventions within the org can affect the analysis process
- Mainly 3 recurring themes

Relationship between Analysts & IT Staff

- Analysts often interacted closely with the IT staff
- Reliance on IT team was found to be high in organizations where data was distributed across many data sources
- Types of regular support provided by IT team:

a) Maintaining the data within a centralized data warehouse

b) Assisting the analysts in acquiring the data. Query data from DW

c) Responsible for operationalizing the workflows

d) IT team serves as a source of documentation

Distributed Data

- 21 analysts reported working with data in at least 3 different formats
- Many analysis tasks involve integrating data from multiple sources
- Some analysts performed this integration themselves while others like scripters and application users relied on the IT team.

Consumer of Analysis

- The result of analysis served many different departments in the org like marketing, sales, operations, and business development.
- The results were generally generated in the form of summary statistics, charts or recommendations
- Static results were usually delivered in the form of ppt documents
- Dynamic results were delivered in the form of interactive dashboards
- In other cases, the reports were in the form of recommendations of actions to take
- The results were shared via email, shared file system, or during a group meeting/presentation

Collaboration

- Analysts often collaborated with people in their own team and department
- Analysts reported meeting regularly to discuss long term projects and immediate next steps
- Most analysts reported that they rarely interacted with other analysts to complete a given task
- Shared four types of resources: data, scripts, results and documentation
- The least commonly shared resource was the data processing scripts
- Analysts rarely stored their analytics code in source control
- Shared their final results in the form of reports or charts

Impediments to Collaboration

3 Common Impediments observed:

a) The diversity of tool and programming made it difficult to share intermediate code to other analysts

b) analysts reported that finding a script or intermediate data product someone else produced was often more time consuming than writing the script from the scratch

c) Many analysts (25/35) also expressed a general attitude that intermediate products such as code and data were “ad hoc”, “experimental” or “throw-away.”

Challenges in the Analysis Process

- 5 high level tasks were identified based on the interviews:

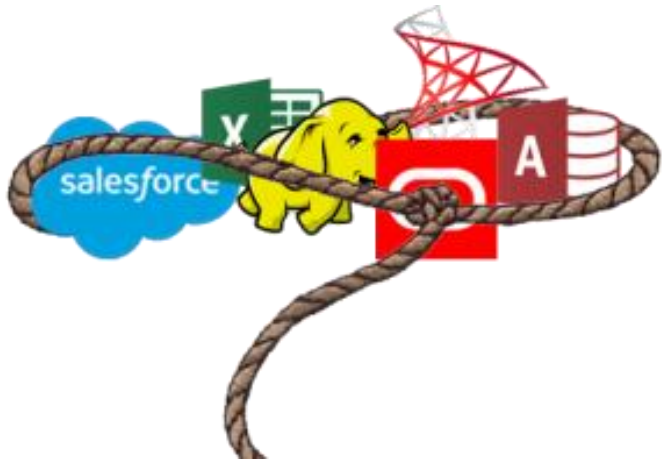
a) Discover
b) Wrangle
c) Profile
d) Model
e) Report

1) Challenges in Discovery

- Within large organizations, finding and understanding relevant data was often a significant bottleneck.
- For 17 analysts, finding relevant data distributed across multiple databases was very time consuming
- Organizations often lack sufficient documentation to identify data
- Analysts relied on their colleagues: like often asking the database admins or other for help
- This difficulty is often compounded by the difficulty of interpreting certain fields in a database.

2) Challenges in Wrangling

- Ingesting semi-structured data. For example, ingesting log files (requires writing complex regular expressions).
- Another difficulty reported by data analysts was integrating data from multiple sources
- Identifiers useful for joining records across data sets were often missing in one or more data sets
- Identifiers used slight variations in spelling or formatting that make direct matches difficult. For example
  First Name: 'John'
  First Name: 'Jonathan'
- Some identifiers used different encoding. For example,
  State: 'British Columbia'
  State: 'BC'



3) Challenges in Profiling

- Data sets may contain a number of quality issues that affect the validity of results, such as missing, erroneous or extreme values
- Many analysts (22/35) reported issues dealing with missing data
- In other cases, entire observations were missing from a data set which were much more difficult to detect.
- A column with an expected type may contain values of another type.

“” in one data set there were 4 males [between the ages] 0 to 9 who were pregnant. If I make an assumption about what that means and filter the data, then I am destroying data.””

Continued…

- Analysts reported using visualization and statistical routines to detect errors in their data.
- Most analysts reported using a combination of visualization and statistics to inspect data
- During inspection, they were also able to gain an understanding of what assumptions they could make
- Common assumptions made:

a) How values were distributed within an attribute

b) How different attributes relate to each other

”””Once you play with the data you realize you made an assumption that is completely wrong. It’s really useful, it’s not just a waste of time, even though you may be banging your head.””

4) Challenges in Modeling

- Biggest challenge as per the analysts was to understand which fields were most relevant
- Issues with scaling:

a) Most existing analytics packages & tools did not scale well with the size of their data sets

b) Scripters were typically limited by the memory requirements of their machine.

c) Hackers were often limited by the types of analysis they could run as some models do not have parallelized implementation

Continued…

- Other analysts used sampling but cited that sampling was hard to do correctly without introducing bias into the analysis.
- Scaling issues were more prominent with visualization tools
- Existing visualization tools simply could not load enough data into the tool
- Interviewees also believed that visualization does not scale to high dimensional data

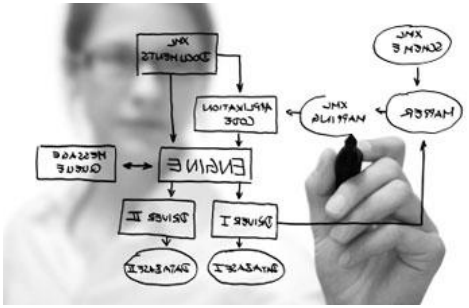
As one analyst described,

””Graphical representation is at best two or three dimensional. Three dimensions won’t tell me very much about how 300 variables interact””

5) Challenges in Reporting

- One complaint about distributing and consuming reports (made by 17 analysts) is poor documentation of assumptions made during analysis
- A number of analysts (17/35) also complained that reports were too inflexible and did not allow interactive analysis.

DESIGN IMPLICATIONS



- Visualization is often typically applied to isolated late stages of the workflow
- Little visualization research addresses discovery, wrangling or profiling challenges
- Visual analytic tools that enable efficient application of these data mining routines could significantly speed up the analysis process
- Tools that extend their data query facilities to operate over partially structured data will enable analysts to immediately apply visualization and analytics to a much wider set of data
- For example, Splunk enables to write queries directly against log files without first structuring data, but has limited support for visualization
- Tools sometimes require the data in a specific format. Should provide more flexibility
- Tools lacked common transformation tasks such as integrating new data, filters or aggregations.

Design Implications Continued…

- Support scalable visual analytics:

a) Visual analytic tools must consider using density or aggregation based plots such as histograms and binned scatter plots for large data sets.

b) Improve scalability by leveraging existing data processing engines for manipulating data. For instance, Tableau can translate statements in its internal representation into queries that run on distributed databases.

c) Tools could benefit from server-side pre-processing and aggregation to enable interactive exploration in the client

d) Tool builders should also consider how approximation approaches might be applied to scale visual analytic tools

e) Sampling data can speed up querying but may introduce bias

Design Implications Continued…

- Bridge the Gap in Programming Proficiency:

a) Tools should try to bridge the gap between capabilities of hackers, scripters, and application users.

b) Provide direct manipulation interfaces for data acquisition and wrangling

c) To empower hackers, direct manipulation interfaces might expose the underlying logic of the tool

Design Implications Continued…

- Capture Metadata at natural annotation points:

a) Tools should augment intermediate products such as scripts and data with additional metadata

b) Metadata might include the script's author, the rationale for an analysis procedure or assumptions made about the input data

c) metadata can enable more efficient search over products and simplify the interpretation of results by others

Conclusion

- An opportunity for visual analytic tools to improve the quality of analysis and the speed at which it takes place
- Tools that simplify tasks across the analytic pipeline could empower nonprogrammers to apply their statistical training and domain expertise to large, diverse data sets.
- Can make lives easier for analysts, which can result in better results and process execution

Thank You!

Q&A

A possible response may include (but not limited to) the following:

a) I don’t know

b) I know, but I won’t tell you

c) Let’s discuss it over a beer for deeper insights