# LISTEN UP. MOVE UP. STEP UP.

Joe Hellerstein

Berkeley
UNIVERSITY OF CALIFORNIA

TRIFACTA

# BIG DATA HAPPENED.  NOW WHAT?

Listen up:

- Know your users.  They are changing.
- Research starts with interviews.

Move up:

- The future of the field is broader, and higher up.
- Trees grow beyond their roots.

Step up:

- We are paying for our own lack of initiative.
- We are lucky: opportunity to lead has come to us.

# BACKGROUND

Fraction of well-structured data shrinks exponentially

- Relative to the deluge of ungoverned Big Data

Data infrastructure is a race to the bottom.

- Good enough: $0.00

The user base is changing.

- Shrinking users: IT and developers.
- Growing users: analysts and subject-matter experts.

What do users say?

# LISTEN UP

## Enterprise Data Analysis and Visualization: An Interview Study

Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer

**Abstract**—Organizations rely on data analysts to model customer engagement, streamline operations, improve production, inform business decisions, and combat fraud. Though numerous analysis and visualization tools have been built to improve the scale and efficiency at which analysts can work, there has been little research on how analysis takes place within the social and organizational context of companies. To better understand the enterprise analysts' ecosystem, we conducted semi-structured interviews with 35 data analysts from 25 organizations across a variety of sectors, including healthcare, retail, marketing and finance. Based on our interview data, we characterize the process of industrial data analysis and document how organizational features of an enterprise impact it. We describe recurring pain points, outstanding challenges, and barriers to adoption for visual analytic tools. Finally, we discuss design implications and opportunities for visual analysis research.

**Index Terms**—Data, analysis, visualization, enterprise.

### 1 Introduction

Organizations gather increasingly large and complex data sets each year. These organizations rely on data analysis to model customer engagement, streamline operations, improve production, inform sales and business decisions, and combat fraud. Within organizations, an increasing number of individuals — with varied titles such as "business analyst", "data analyst" and "data scientist" — perform such analyses. These analysts constitute an important and rapidly growing user population for analysis and visualization tools.

Enterprise analysts perform their work within the context of a larger organization. Analysts often work as a part of an analysis team or business unit. Little research has observed how existing infrastructure, available data and tools, and administrative and social conventions [illegible] enterprise. Understanding how these issues shape analytic workflows can [illegible]

ery and wrangling, often the most tedious and time-consuming aspects of an analysis, are underserved by existing visualization and analysis tools. We discuss recurring pain points within each task as well as difficulties in managing workflows across these tasks. Example pain points include integrating data from distributed data sources, visualizing data at scale and operationalizing workflows. These challenges are typically more acute within large organizations with a diverse and distributed set of data sources.

We conclude with a discussion of future trends and the implications of our interviews for future visualization and analysis tools. We argue that future visual analysis tools should leverage existing infrastructures for data processing to enable scale and limit data migration. [illegible] systems that specify analysis as data processing operations in a high-level lan- [illegible] note

**Enterprise Data Analysis and Visualization: An Interview Study**

Sean Kandel, Andreas Paepcke, Joseph Hellerstein, Jeffrey Heer

*IEEE Visual Analytics Science & Technology (VAST)*, 2012

# INTERVIEWEES × CHALLENGES/TOOLS

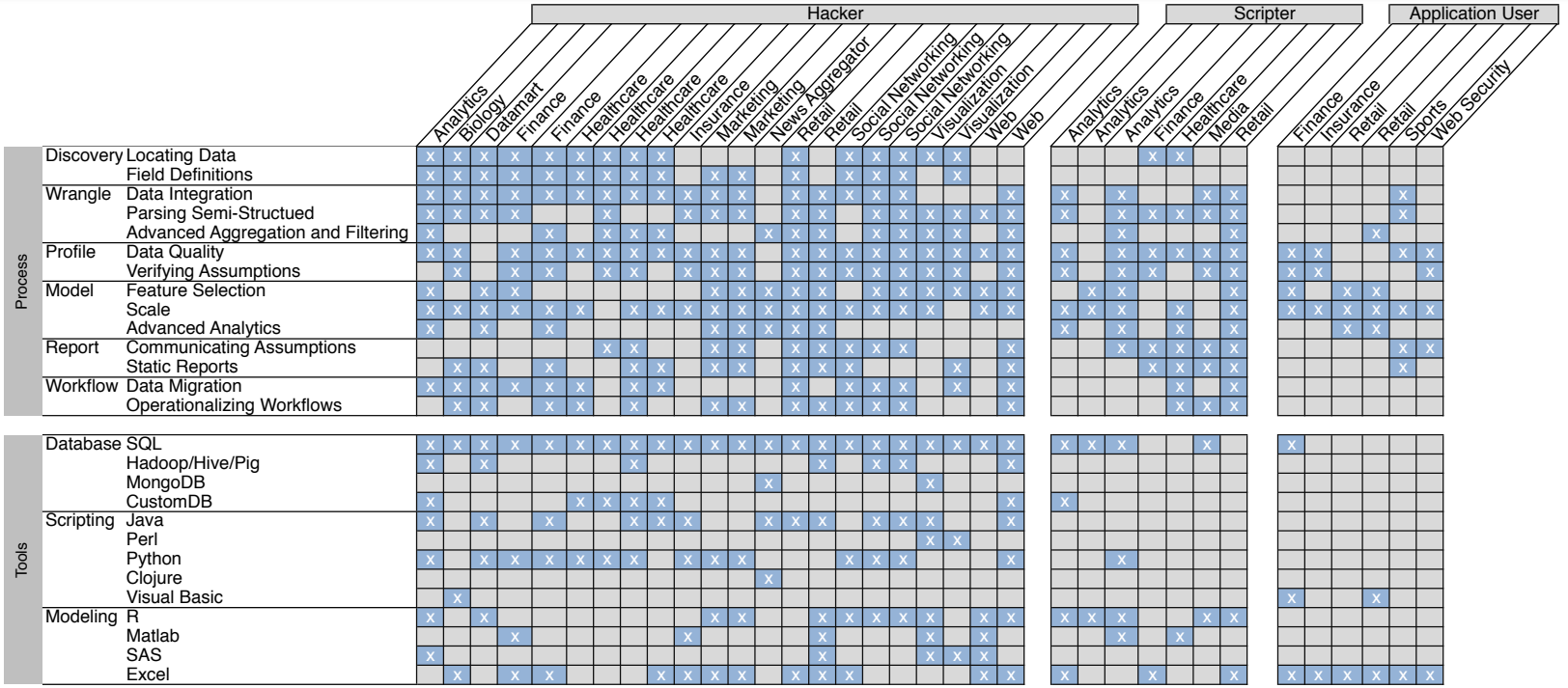


Fig. 1. Respondents, Challenges and Tools. The matrix displays interviewees (grouped by archetype and sector) and their corresponding challenges and tools. *Hackers* faced the most diverse set of challenges, corresponding to the diversity of their workflows and toolset. *Application users* and *scripters* typically relied on the IT team to perform certain tasks and therefore did not perceive them as challenges.

Enterprise Data Analysis and Visualization: An Interview Study
Sean Kandel, Andreas Paepcke, Joseph Hellerstein, Jeffrey Heer
*IEEE Visual Analytics Science & Technology (VAST)*, 2012

"I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any 'analysis' at all…

"I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any 'analysis' at all...

Most of the time once you transform the data ... the insights can be scarily obvious."

“It is really hard to know where the data is. We have all the data, but there is no huge schema where we can say this data is here and this variable is there.”

“In practice right now the biggest differentiator is feature selection: knowing what columns to pay attention to and how to sensibly transform them. Do you take the log of these, do you combine these two? A lot of work is just finding what the units of the columns should be.”

“The only code that doesn’t go in github is our analytics code...

svn is more like backup than version control.”

"You go down a lot of dead ends, and you come up with a bunch of hypotheses. 8 out of 10 are dead ends…

Especially your dead ends…there's no remnant of that."

“Analysts that can’t program are disenfranchised here.”

# MOVE UP

Areas:

- Interaction/Visualization
- Data Analysis
- Many Application Verticals
  - *Go native!*

Bring our signal skills with us:

- Declarativity and semantics
- Appreciation of both algorithms and systems
- Strong connections to industry

The “complete” computer scientist.

# STEP UP

Computer Science has finally moved our way.

We have yet to take leadership:

- Reframe CS curriculum from Day 1.
- Send reps we're proud of to Washington.
- Assail the popular press.
- Embrace open source.
- Found bold companies.
- Forge strategic alliances.

> Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018.
>
> —McKinsey Global Institute, 2011