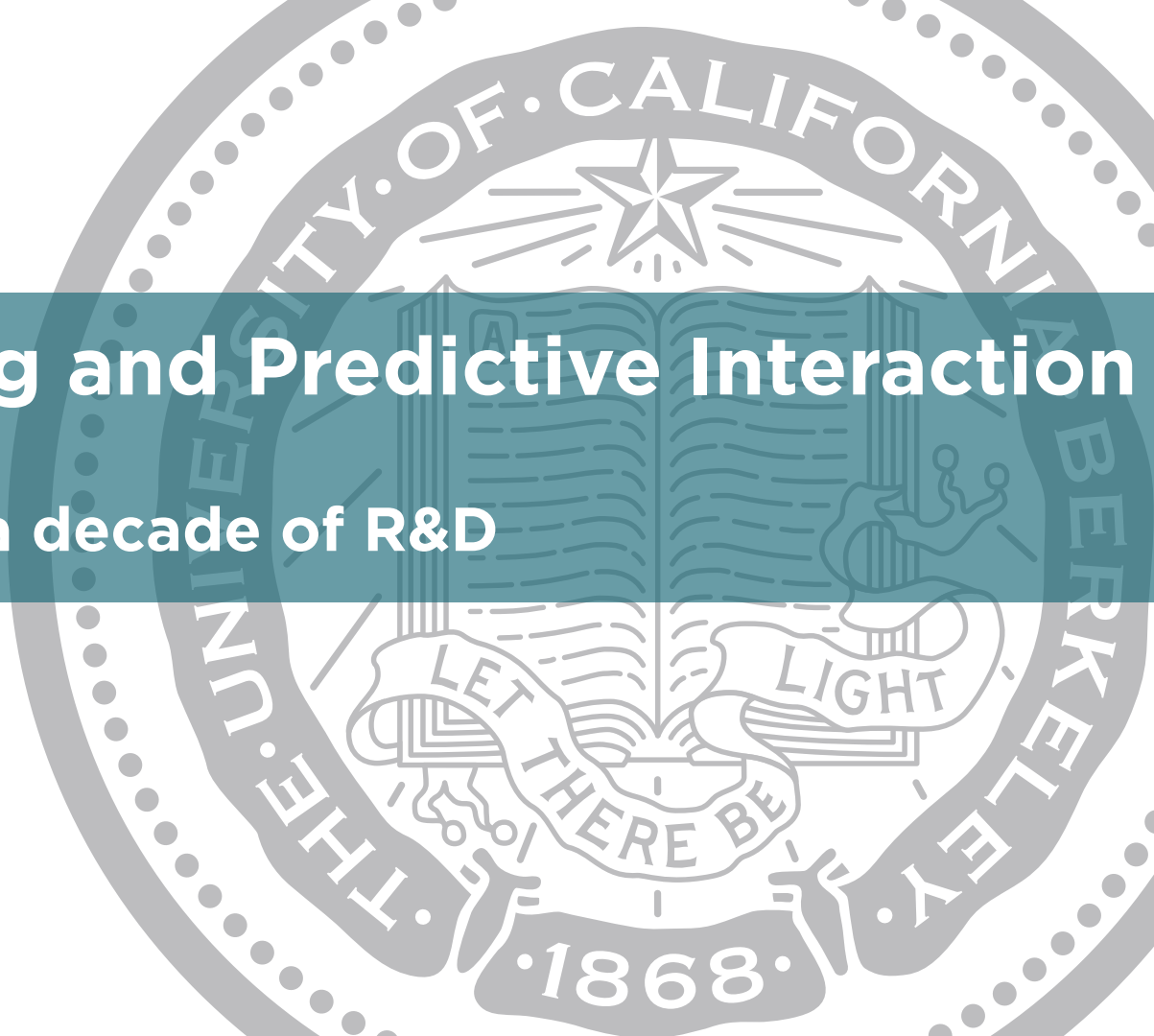
Data Wrangling and Predictive Interaction

Brief lessons from a decade of R&D

Joe Hellerstein



Data Platforms


Databases


Spreadsheets


Log Files


IoT Sensors


Apps

"The hardest part of AI is the **data wrangling.**"
—SWAMI SIVASUBRAMANIAN, VP AWS MACHINE LEARNING

#reInvent2018

80%

"It's impossible to overstress this: **80% of the work** in any data project is in **cleaning the data.**"
— DJ Patil, Data Jujitsu, O'Reilly Media 2012

Analysis


Enterprise Data Warehouse


AI


Business Intelligence

Data Platforms

Databases

Spreadsheets

Log Files

IoT Sensors

Apps

Discovering → Structuring → Cleaning → Enriching → Validating → Deploying

Data Analyst

Data Engineer

Data Scientist

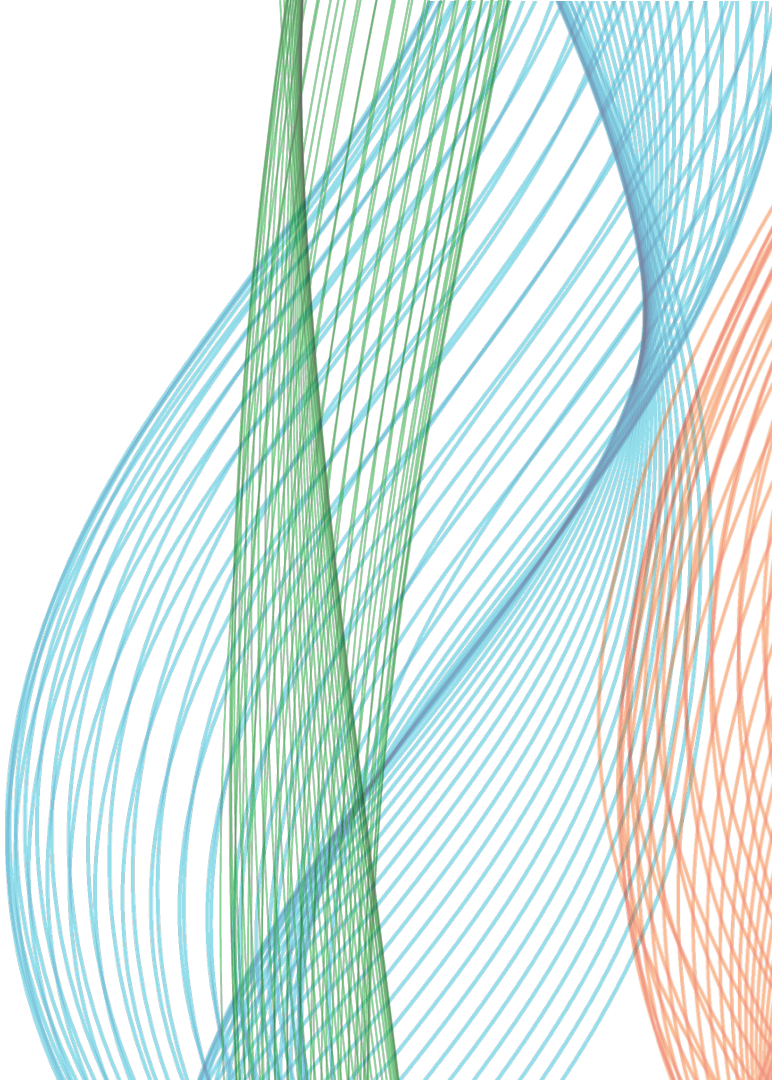
Analysis

Enterprise Data Warehouse

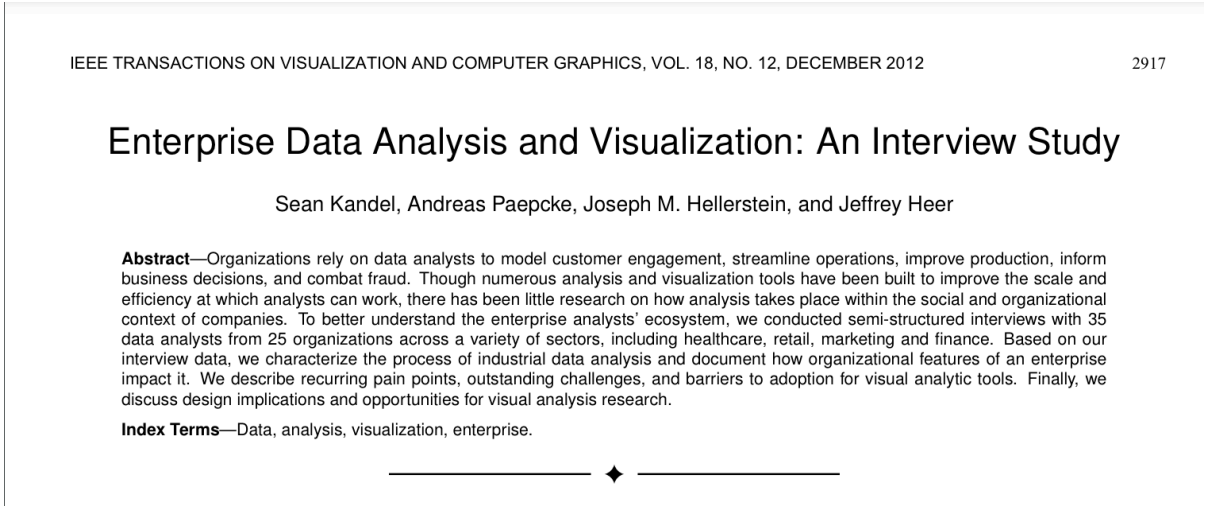
AI

Business Intelligence

Mini Demo



Methodology: Interview Studies!



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.

IEEE Vis Test of Time Award 2022

A Decade of R & D

DataWrangler alpha

TRIFACTA

Wrangler: Interactive Visual Specification of Data Transformation Scripts

Sean Kandel*, Andreas Paepcke*, Joseph Hellerstein† and Jeffrey Heer*
* Stanford University
skandel, paepcke, jheer@cs.stanford.edu
† University of California, Berkeley
hellerstein@cs.berkeley.edu

ABSTRACT
Though data analysis tools continue to improve, analysts still expend an inordinate amount of time and effort manipulating data and assessing data quality issues. Such "data wrangling" regularly involves reformatting data values or layout, correcting erroneous or missing values, and integrating multiple data sources. These transforms are often difficult to specify and difficult to reuse across analysis tasks, teams, and tools. In response, we introduce *Wrangler*, an interactive system for creating data transformations. Wrangler combines direct manipulation of visualized data with

data warehousing projects [4]. Such "data wrangling" often requires writing idiosyncratic scripts in programming languages such as Python and Perl, or extensive manual editing using interactive tools such as Microsoft Excel. Moreover, this hurdle discourages many people from working with data in the first place. Sadly, when it comes to the practice of data analysis, "the tedium is the message."

Part of the problem is that reformatting and validating data requires transforms that can be difficult to specify and evalu-

CHI 2011

Google Cloud Dataprep by TRIFACTA

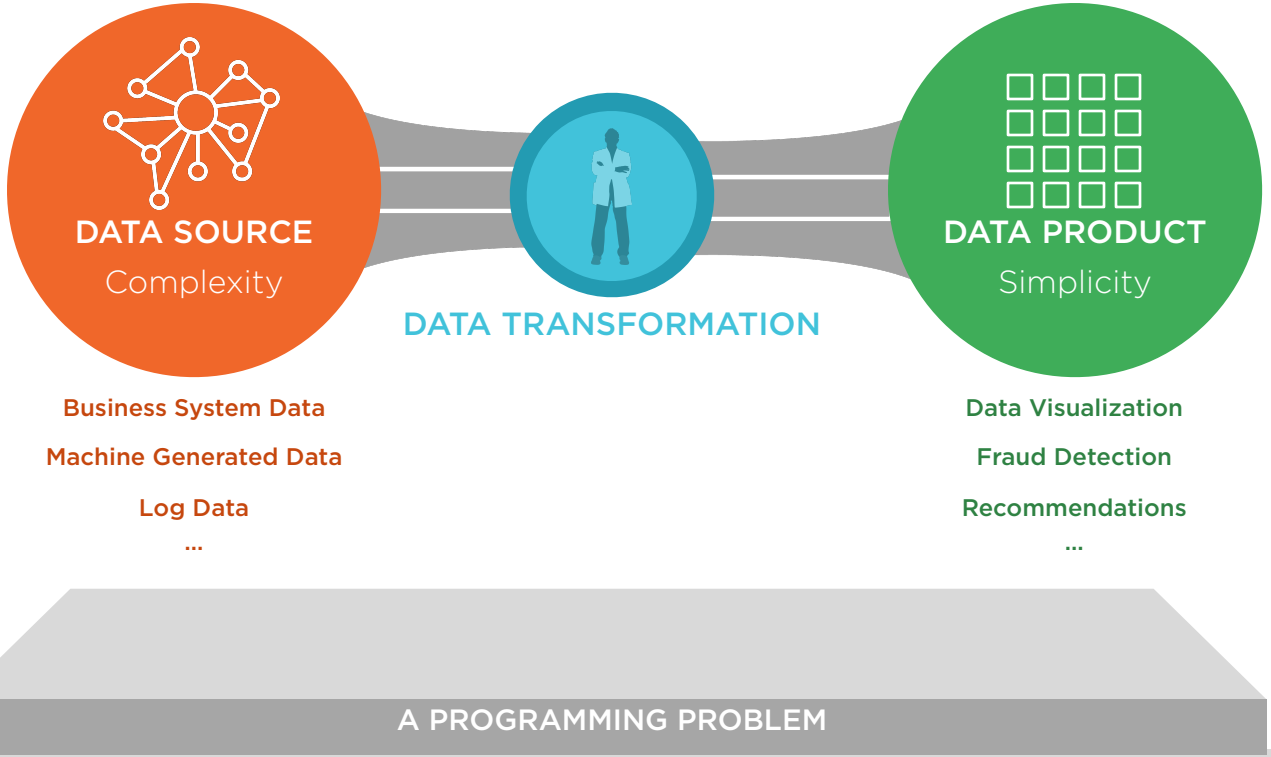
See also:

- OpenRefine
- MS PowerQuery
- Tableau Dataprep
- AWS Databrew
- Zoho Dataprep
- ...

alteryx

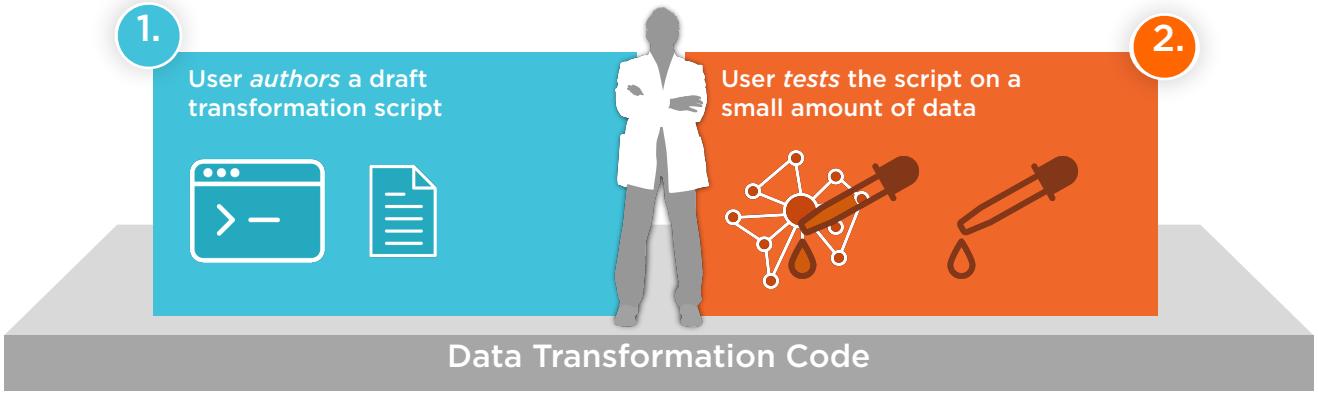
DESIGNER CLOUD
Powered by TRIFACTA

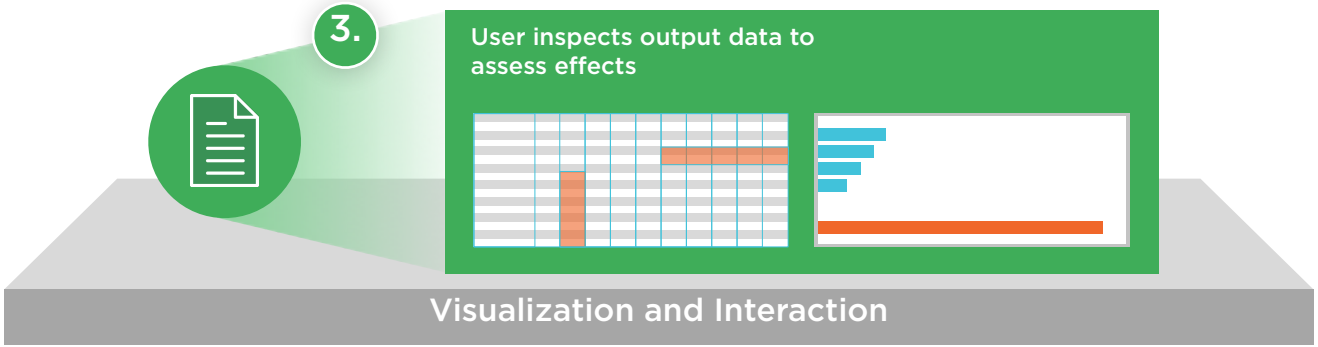
THE DATA TRANSFORMATION PROBLEM



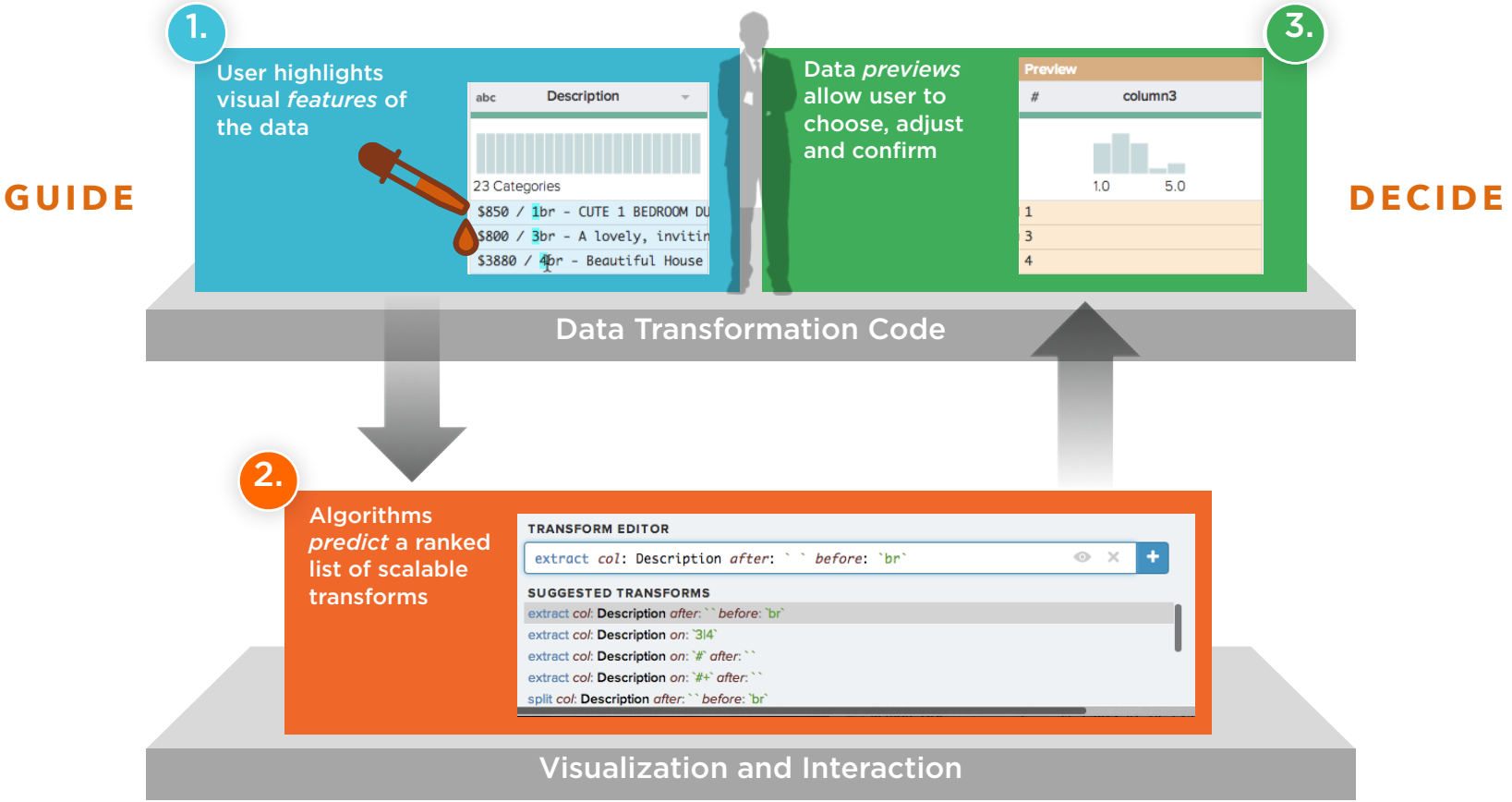
TRADITIONAL ROLE OF VISUALIZATION

Python pandas,
R dplyr,
SQL, etc...

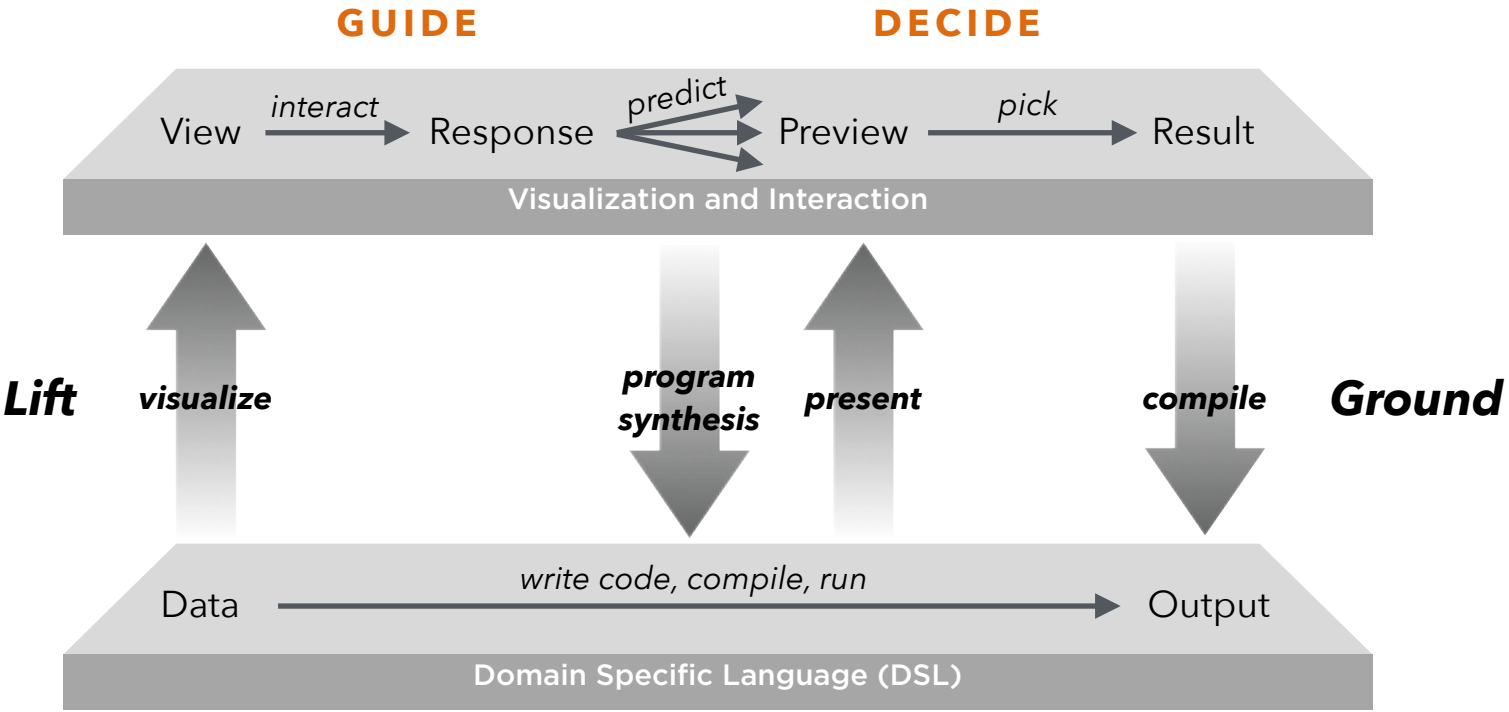




PREDICTIVE INTERACTION



PREDICTIVE INTERACTION



[Heer, Hellerstein, Kandel, CIDR15]

Predictive Interaction for Data Transformation

Jeffrey Heer
U. Washington & Trifacta Inc.
jheer@uw.edu

Joseph M. Hellerstein
UC Berkeley & Trifacta Inc.
hellerstein@berkeley.edu

Sean Kandel
Trifacta Inc.
skandel@trifacta.com

ABSTRACT

The human work involved in data transformation represents a major bottleneck for today's data-driven organizations. In response, we present Predictive Interaction, a framework for interactive systems that shifts the burden of technical specification from users to algorithms, while preserving human guidance and expressive power.

1. INTRODUCTION

Many of today's data management challenges stem from the increasing variety of scenarios where data is being exploited. This

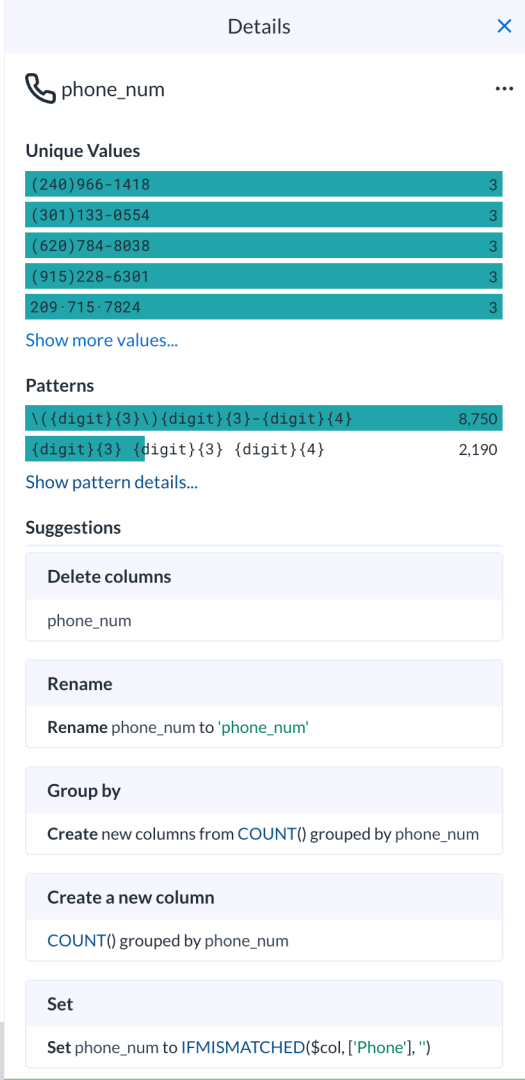
data-to-metadata transformations like Pivot/Unpivot), cleaning (e.g., standardization, entity resolution, dictionary management), enrichment (e.g., joins and references), and distillation (e.g., sampling, filtering, aggregation, windowing).

There are a number of recurring themes in previous work on data transformation. One is to develop new *user interfaces* for graphical specification of queries and transforms [11, 23, 26, 32], and visualization of outputs [3, 19]. A second theme is to innovate at the *data management* layer, with domain-specific languages that are well-suited to data transformation and can be executed in a scalable, high-performance manner [8, 18, 28]. And despite the

CIDR 2015

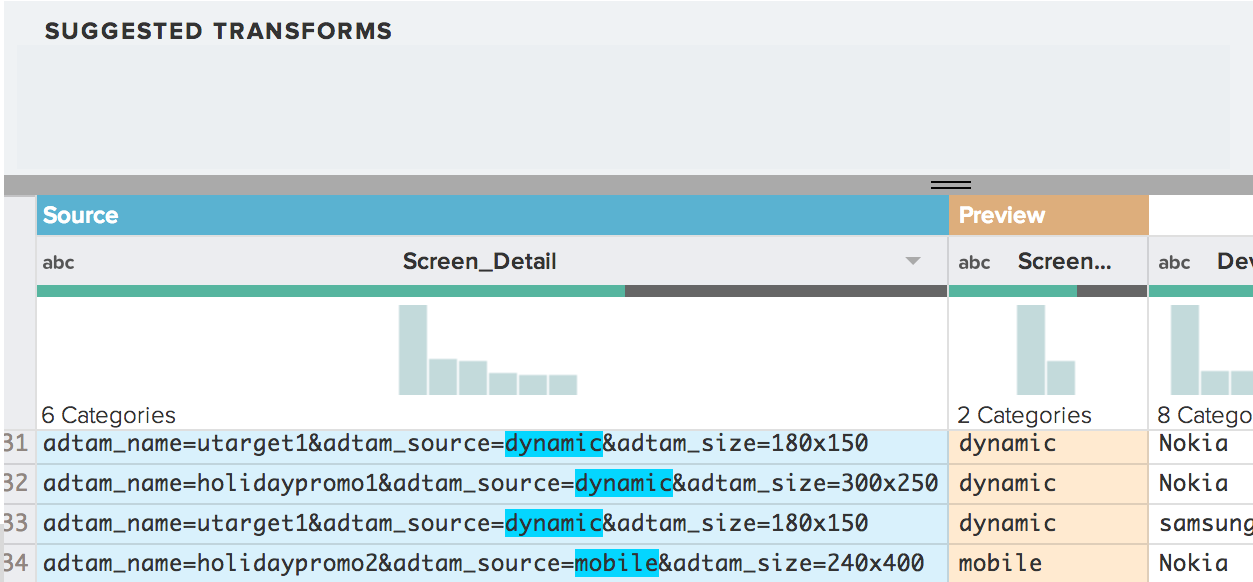
Surprising Lessons

➔ **UX matters far more than prediction quality**



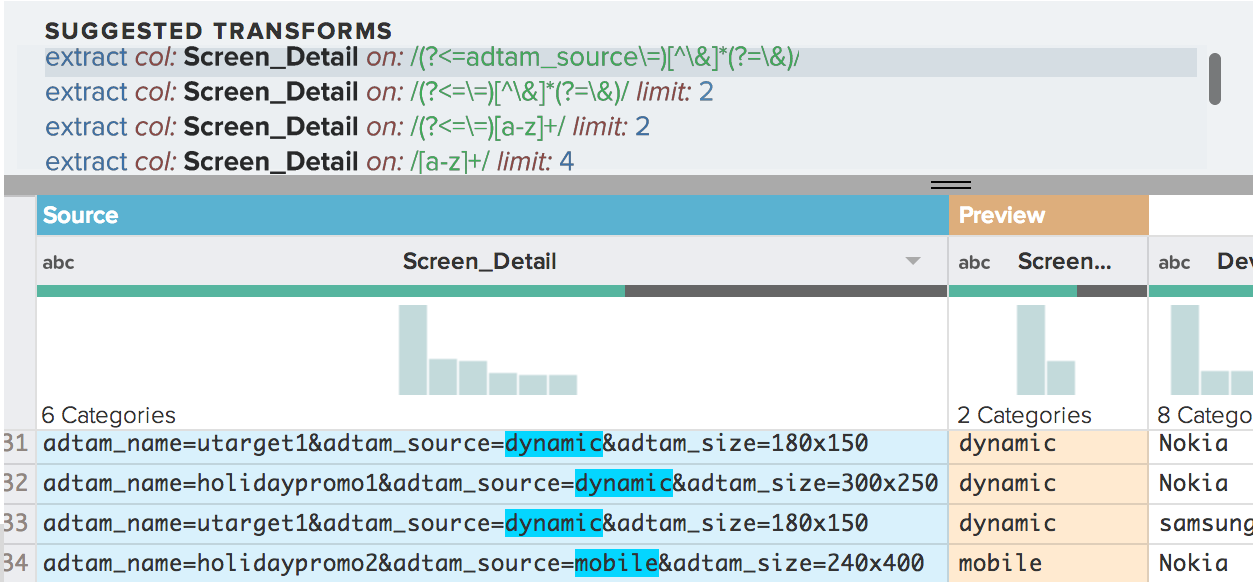
Surprising Lessons

- **UX matters far more than prediction quality**
- **Embrace “read-centric” rendering of code**



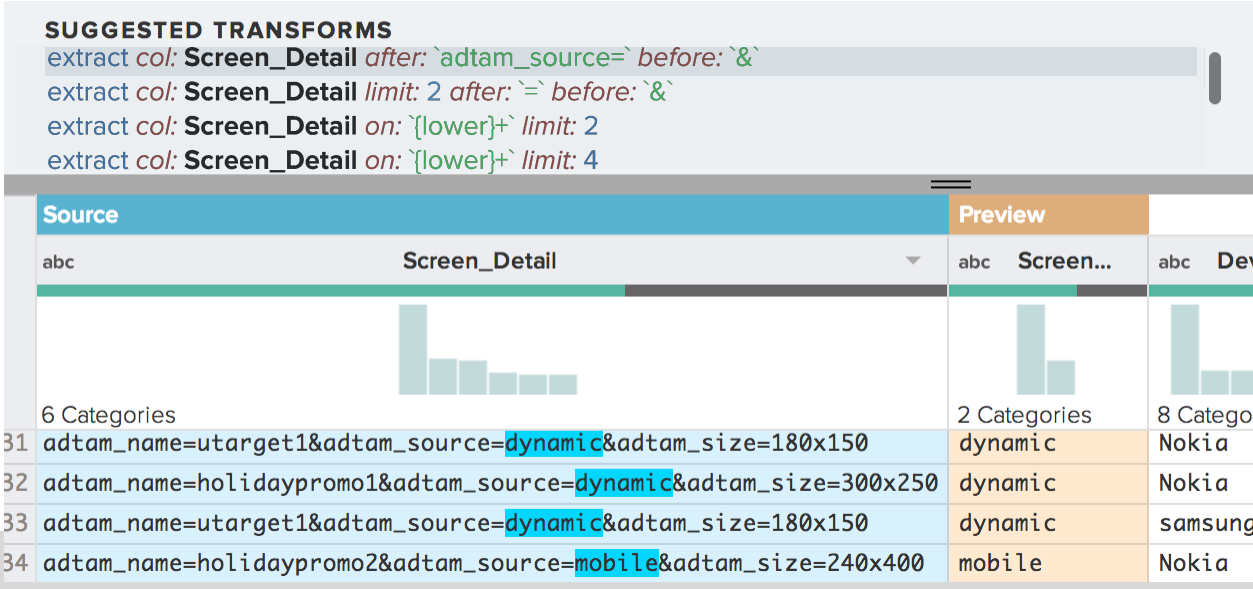
Surprising Lessons

- **UX matters far more than prediction quality**
- **Embrace “read-centric” rendering of code**



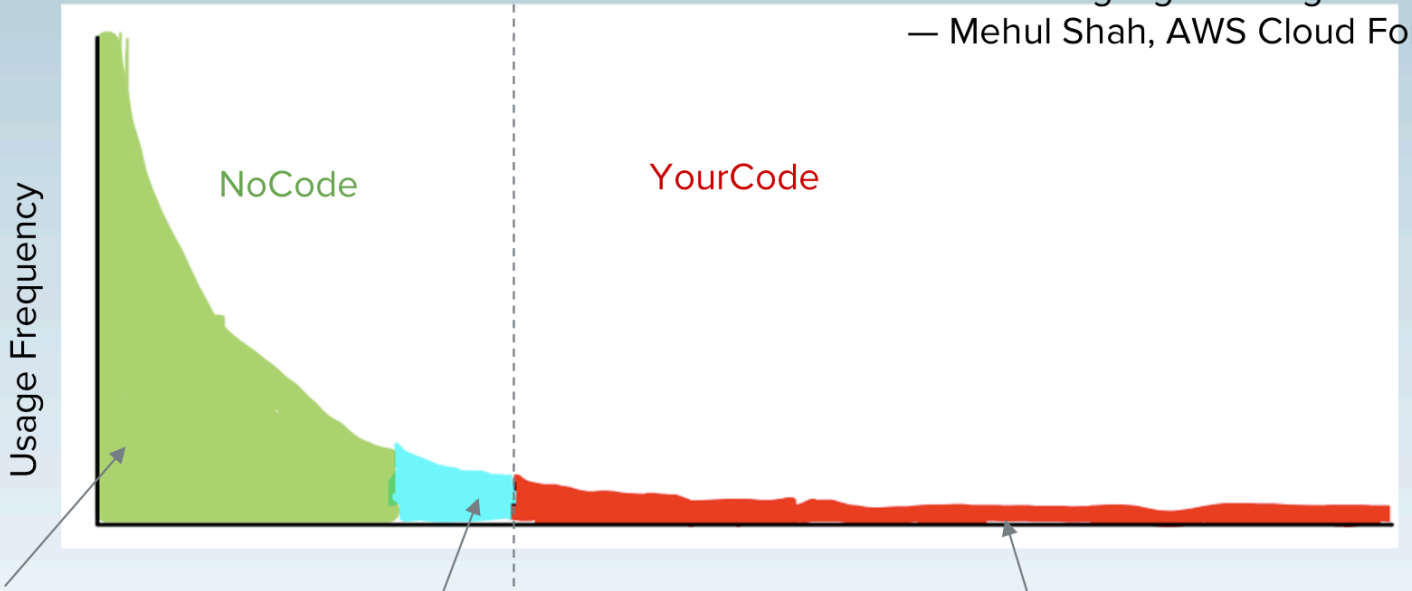
Surprising Lessons

- **UX matters far more than prediction quality**
- **Embrace “read-centric” rendering of code**



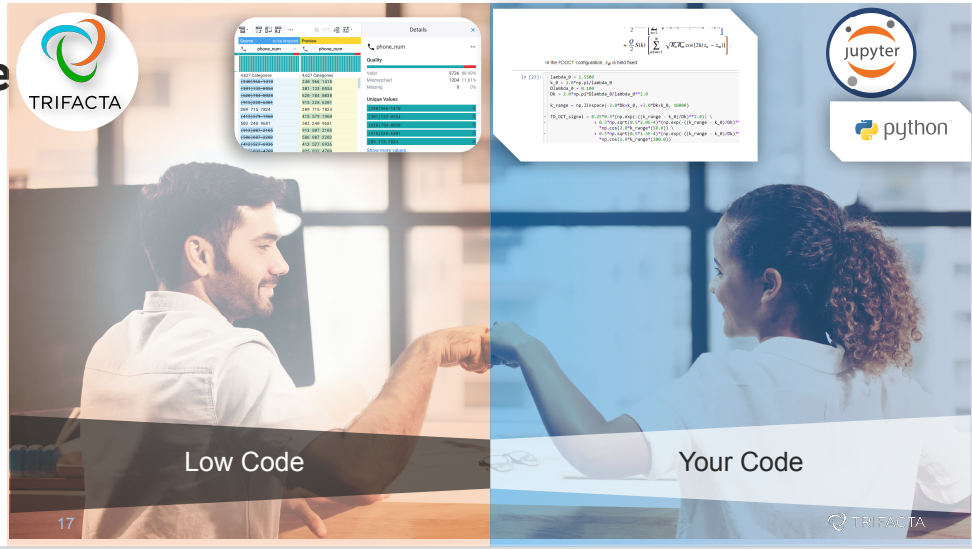
Surprising Lessons

"Data wrangling is a long-tail business."
— Mehul Shah, AWS Cloud Formation & Glue



Surprising Lessons

- **UX matters far more than prediction quality**
- **Embrace read-only rendering of code**
- **Solve for Low Code + Your Code**



Selected Academic Papers

Raman, Vijayshankar, and Joseph M. Hellerstein. "Potter's wheel: An interactive data cleaning system." In *VLDB*, 2001.

Kandel, Sean, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. "Wrangler: Interactive visual specification of data transformation scripts." In *SIGCHI,* 2011.

Kandel, Sean, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. "Enterprise data analysis and visualization: An interview study." *IEEE Trans. on Vis. and Comp. Graphics* 18, no. 12 (2012).

Heer, Jeffrey, Joseph M. Hellerstein, and Sean Kandel. "Predictive Interaction for Data Transformation." In *CIDR,* 2015.

Hellerstein, Joseph M., Jeffrey Heer, and Sean Kandel. "Self-Service Data Preparation: Research to Practice." *IEEE Data Eng. Bull.* 41, no. 2 (2018).

@joe_hellerstein
hellerstein@berkeley.edu