Towards Reliable Interactive Data Cleaning:

A User Survey and Recommendations

Sanjay Krishnan, **Daniel Haas**, Eugene Wu, Michael Franklin
HILDA 2016

Data Scientist:

The Sexiest Job of the 21st Century

coax treasure out of messy, unstructured data

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

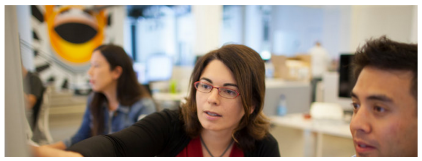
For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

Email
Share
Tweet
Save

Technology revolutions come in measured, sometimes foot-dragging steps. The lab science and marketing enthusiasm tend to underestimate the bottlenecks to progress that must be overcome with hard work and practical engineering.

The field known as "big data" offers a contemporary case study. The catchphrase



Dirty data costs the U.S. economy $3 trillion+ per year

JANUARY 30, 2013 by Joe Fusaro in CRM

Estimates show dirty data is a big problem for the U.S. economy. How big? **About 2x the national deficit**.

Software expert **Hollis Tibbets**, the Global Director of Marketing at **Dell**, estimates that duplicate data and bad data combined cost the U.S. economy over $3 trillion every year - which is just about two times the national deficit.

In his post "*$3 Trillion Problem: Three Best Practices for Today's Dirty Data Pandemic*," Hollis points to a few key facts and figures to back up his estimate.

204 papers on data cleaning* since 2012 in VLDB, ICDE, SIGMOD

(papers mentioning data cleaning in title or abstract, possibly **dirty data**)

The tutorial you missed

- How can statistical techniques improve efficiency or reliability of data cleaning? (**Data Cleaning with Statistics**)
- How how can we improve the reliability of statistical analytics with data cleaning? (**Data Cleaning For Statistics**)

The tutorial you missed

Data cleaning *with* statistical techniques

ERACER 2010
Guided Data Repair 2011
Corleone 2014
Wisteria 2015
Deep Dive 2014
Katara 2014
Trifacta 2015
Data Tamer 2013
….

Data cleaning *for* statistical analysis

Sensor Net/Stream+ 2000s
Scorpion 2013
SampleClean+ 2014
Unknown Unknowns 2016
…

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.

35 People from 25 organizations

[Sean Kandel et. al, VAST, 2012]

DeepDive KeystoneML TensorFlow™ ML

In practice, **how is data cleaned** before analysis?

What are the **limitations** of **existing processes**?

How can **database researchers contribute**?

9

Data Cleaning Survey

Introduction

We're conducting research on industrial perspectives and practices on managing, analyzing, and serving "dirty" data. Dirtiness is broadly defined as any type of corruption that can negatively affect subsequent analysis. We hope to publish these results to help the research community bridge the gap between data cleaning theory and practice. These results will also inform the design of a new open-source data cleaning platform (sampleclean.org). The survey should take about 15 minutes, and your responses are completely anonymous. If you are willing to allow us to follow-up with you through email, please provide your email address. We will not reveal or publish your email address.

If you have any questions about the survey, please email us: sanjay@eecs.berkeley.edu dhaas@cs.berkeley.edu eugenewu@mit.edu

We really appreciate your input!

1. Describe your company/organization and your role there.

2. Describe the types of data you use (i.e., source, input format, size, etc.).

3. Are any of the data that you work with "dirty"? Dirtiness is any corruption that can negatively affect subsequent analysis (e.g., missing, incorrect or duplicate records).

Yes

No

10

How do you determine whether the data is sufficiently clean to trust the analysis?

I am comfortable writing a program that reads a large textual log file on HDFS and computes the number of errors from each IP using Apache Spark or Hadoop.

Describe your company/organization and your role there.

Which of the following tools/interfaces/languages do data scientists at your organization prefer for manipulating data, including extraction, schema transformations, and outlier removal, to make analysis easier or more reliable. Please Rank

Which of the following most closely describes your job?

Describe your organization's data analysis pipeline, including data acquisition, pre-processing, and applications that use the processed data. If possible, list it as a sequence of steps that each describe the intended goal and the tools used.

Are any of these steps, or the downstream applications that use the data, affected by dirty data (i.e., inconsistent, missing, or duplicated records)? If so, please describe how you identify dirty records, repair dirty records, and maintain the processing pipeline.

How do you validate the correctness of the processing pipeline?

I am comfortable explaining when to use regularization for Support Vector Machines

Does your organization employ teams of people or crowdsourcing for any of the steps described above?

Has the scale of your dataset ever made it challenging to clean?

I analyze my organization's customer data for modeling, forecasting, prediction, or recommendation.

Describe your data analysis, ideally as a sequence of steps that each describe the intended goal and the tools you use to achieve the goal. Include descriptions of where the data comes from (including the number and variety of sources), properties of the data (e.g., the format, amount), each preprocessing step, and the final result.

Is your analysis affected by dirty data (i.e., inconsistent, missing, or duplicated records)? If so, please provide examples of how the data is dirty, what the cleaned versions look like, and how it affects the final result.

Describe your data cleaning process including how you identify errors, steps to mitigate errors in the future, and how you validate data cleaning with your analysis.

...

11

How do you determine whether the data is sufficiently clean to trust the analysis?

I am comfortable writing a program that reads a large textual log file on HDFS and computes the number of errors from each IP using Apache Spark or Hadoop.

Describe your company/organization and your role there.

Which of the following tools/interfaces/languages do data scientists at your organization prefer for manipulating data, including extraction, schema transformations, and outlier removal, to make analysis easier or more reliable. Please Rank

Which of the following most closely describes your job?

Describe your organization's data analysis pipeline, including data acquisition, pre-processing, and applications that use the processed data. If possible, list it as a sequence of steps that each describe the intended goal and the tools used.

Are any of these steps, or the downstream applications that use the data, affected by dirty data (i.e., inconsistent, missing, or duplicated records)? If so, please describe how you identify dirty records, repair dirty records, and maintain the processing pipeline.

How do you validate the correctness of the processing pipeline?

I am comfortable explaining when to use regularization for Support Vector Machines

Does your organization employ teams of people or crowdsourcing for any of the steps described above?

Has the scale of your dataset ever made it challenging to clean?

I analyze my organization's customer data for modeling, forecasting, prediction, or recommendation.

Describe your data analysis, ideally as a sequence of steps that each describe the intended goal and the tools you use to achieve the goal. Include descriptions of where the data comes from (including the number and variety of sources), properties of the data (e.g., the format, amount), each preprocessing step, and the final result.

Is your analysis affected by dirty data (i.e., inconsistent, missing, or duplicated records)? If so, please provide examples of how the data is dirty, what the cleaned versions look like, and how it affects the final result.

Describe your data cleaning process including how you identify errors, steps to mitigate errors in the future, and how you validate data cleaning with your analysis.

...

12

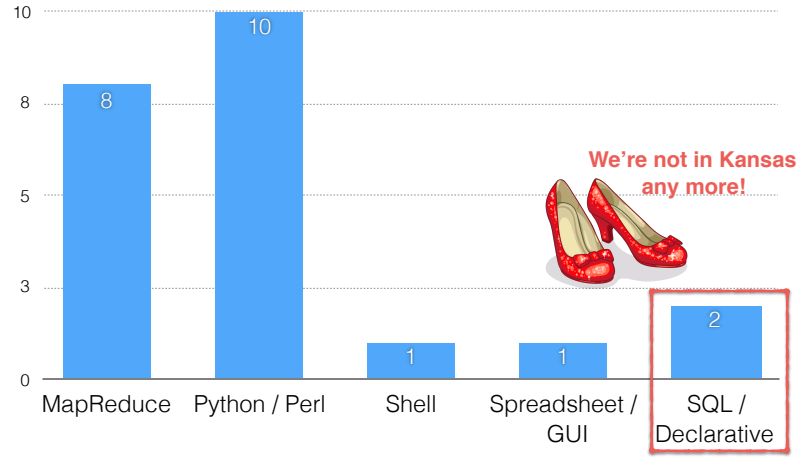
Our Survey: Participants

Initial results from N = 29
Largely Technology Sector

Organization Size	#
Small	7
Large	17
N/A	5

Job Desc.	#
Infrastructure	10
Analysis	12
Both	7

Our Survey: Tools



Our Misconceptions

~~The **end-goal** of data cleaning is **clean data**~~

"We typically clean our data **until the desired analytics works** without error."

Icons created by Clara Joy from Noun Project

Our Misconceptions

~~Data cleaning is a **sequential operation**~~

"[It's an] **iterative process**, where I assess biggest problem, devise a fix, re-evaluate. It is dirty work."

Icons created by Clara Joy from Noun Project

Our Misconceptions

~~Data cleaning is performed by **one person**~~

“There are often long back and forths with senior **data scientists**, **devs**, and the **business units** that provided the data on data quality.”

Icons created by Clara Joy from Noun Project

17

Our Misconceptions

~~Data quality is **easy to evaluate**~~

“I wish there were a more rigorous way to do this but we **look at the models and guess** if the data are correct.”

Icons created by Clara Joy from Noun Project

18

Our Misconceptions

~~Data quality is **easy to evaluate**~~

“Other than **common sense** we do not have a procedure to do this.”

Icons created by Clara Joy from Noun Project

19

Our Misconceptions

~~Data quality is **easy to evaluate**~~

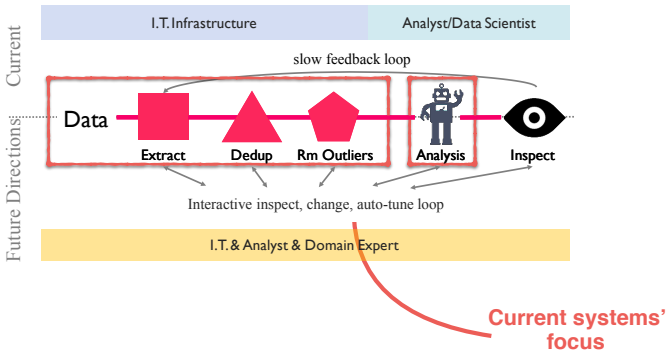
“Usually [a data error] is only caught **weeks later** after someone notices.”

Icons created by Clara Joy from Noun Project

20

How can **database researchers contribute**?

1. A New Architecture



2. New Challenges

Data cleaning **evaluation**

Debugging workflows

Usable **collaborative** interactions

…

Evaluation: Metrics

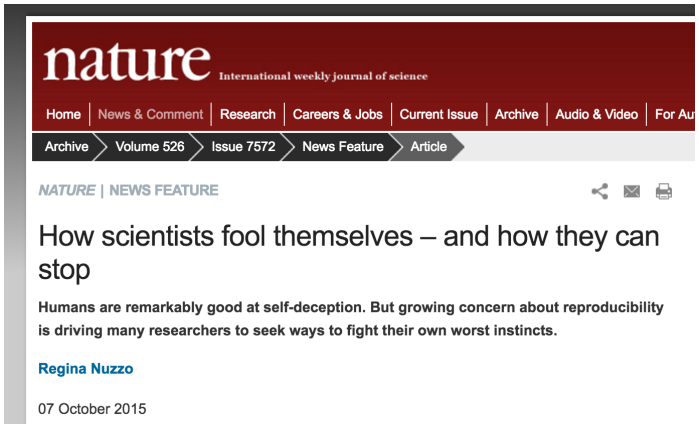
Goal: **Quantitative** understanding of how well cleaning has worked

Techniques: **gold standard** data, **benchmark** datasets, your idea here?

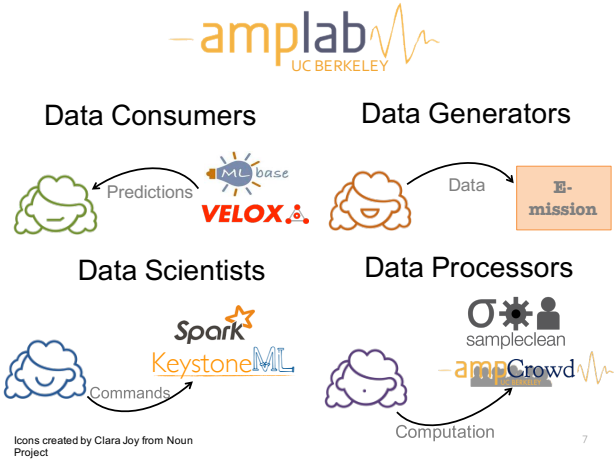
Feedback: design systems that use data quality evaluations to **optimize the pipeline**

[Patricia Arocena et. al, VLDB 2016]

Evaluation: Overfitting and Confirmation Bias



Final thoughts: People are Everywhere



Thank you!

- **Now**: Questions at the panel
- **Later**: Check out our poster
- **Whenever**:
  {sanjay, dhaas, franklin}@cs.berkeley.edu,
  ewu@cs.columbia.edu