

# Towards Sustainable Insights

Tim Kraska <tim\_kraska@brown.edu>

# A New Study shows: A Glass Of Red Wine Is The Equivalent To An Hour At The Gym [Fox News 02/15 and others]



[http://www.huffingtonpost.co.uk/2016/01/08/a-glass-of-red-wine-is-the-equivalent-to-an-hour-at-the-gym-says-new-study\\_n\\_7317240.html](http://www.huffingtonpost.co.uk/2016/01/08/a-glass-of-red-wine-is-the-equivalent-to-an-hour-at-the-gym-says-new-study_n_7317240.html)



# A new study shows: Secret to winning a nobel prize? Eat More Chocolate [Time 10/12]





# A new study shows: Secret to winning a nobel prize? Eat More Chocolate [Time 10/12]





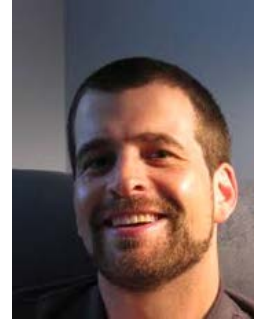
# Scientists find the secret of longer life for men (The bad news: castration is the key) [Daily Mail UK, 09/12]



There has been an **explosion** of (data-driven) discoveries, many of which being **questionable**.



Reasons are manifold, but...  
**the database community**



*... and many others*

**works hard on to be not left out** (again)



# Let me introduce (virtual) Reviewer 2:



The paper's shortcomings are in its motivation, solution, and presentation.

The part of the paper that I did like was the examples given in Sec 2.2.2.

*A note for **Reviewer 2**: We actually liked your comments and it helped us to sharpen our points. If you feel in any way offended by this talk, this was not my intention and I am more than happy to make it up to you with a lot of whisky. Just come to me after the talk and say we need to drink. Knowing this crowd, enough people will do it and I will even never find out your identity if you do not wish so.*



# Outline

## Part I: The problem with:

A. Interactive Data Exploration



B. Visualization Recommendation Systems



C. Hypothesis Generator



A. Part II: Solutions



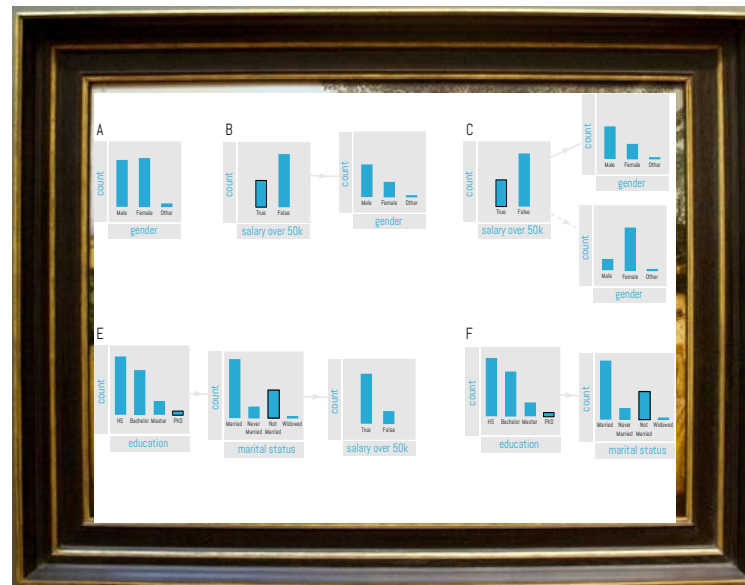
# A) Interactive Data Exploration Tools (Vizdom as an Example)



# Why Visualizations contribute to the problem

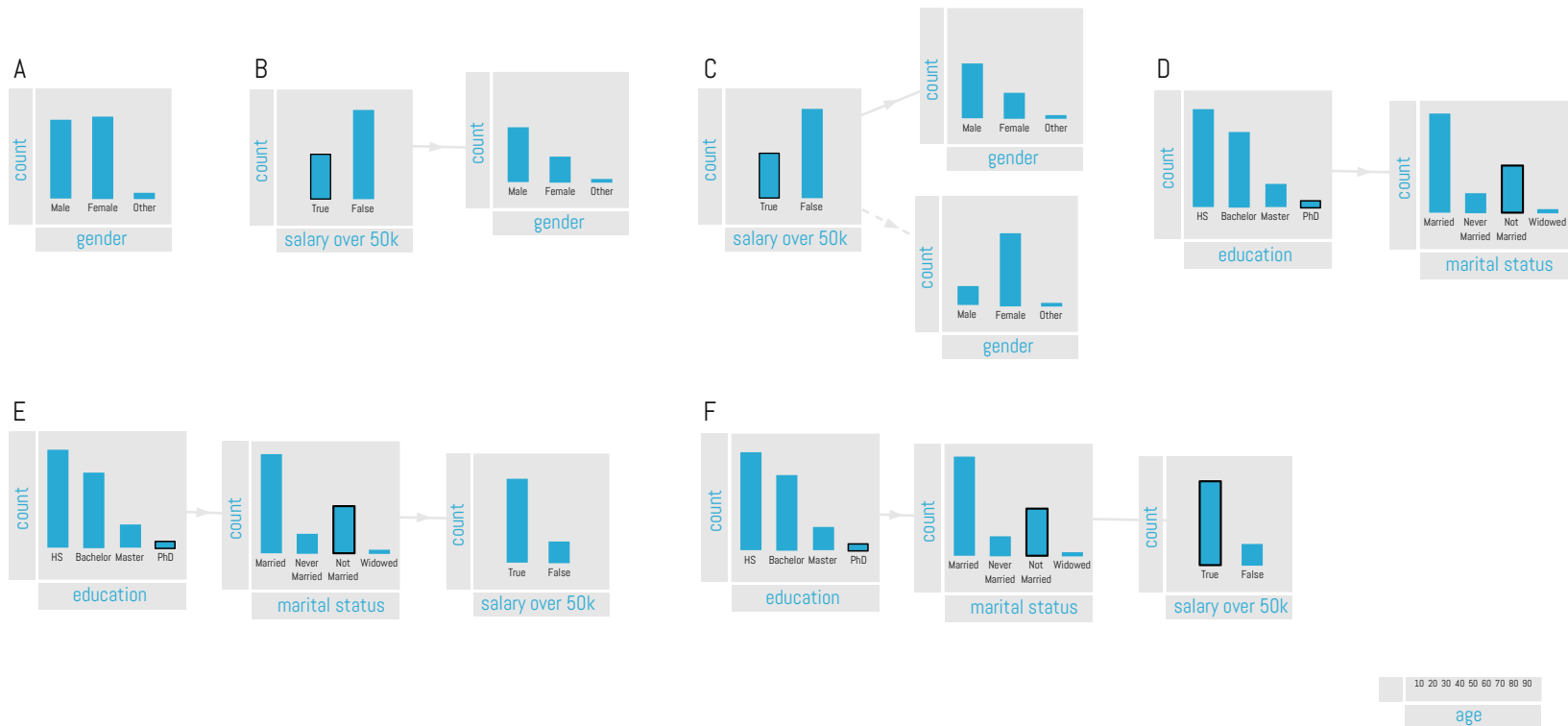
If a visualization provides any insight, it is an **hypothesis test** (just one where you not necessarily know if it is statistical significant)

Otherwise, visualizations have just to be taken as pretty pictures about (potentially) random facts





# If visualizations are used to find something interesting, the user is doing multiple hypothesis testing



# Running Example: Survey on Amazon Mechanical Turk

Project Name:

This name is not displayed to Workers.

Survey about demographics, habits and opinions

Requester: Zheguang Samuel Zhao

Reward: \$2.00 per HIT

HITs available: 0

Duration: 2 Days

Qualifications Required: Masters has been granted

## HIT Preview

**49. Your first guess of "Stonebraker" is?**

- ☐ A Simpsons character
- ☐ A type of stone
- ☐ An antient Egyptian profession
- ☐ A Turing-award winner

**50. Can you jump on one foot for 5 minutes non-stop?**

- ☐ Yes
- ☐ No

**51. Which smartphone operating system do you prefer?**

- ☐ Apple iOS
- ☐ Android



Our goal: To find good indicators (correlations) that somebody knows who Mike Stonebraker is.



And after searching for a bit,  
one of my favorites



**Pearson correlation significance-level  $p < 0.05$**

But Why Does the DB community make the situation worse?





# So What Did Reviewer 2 say?



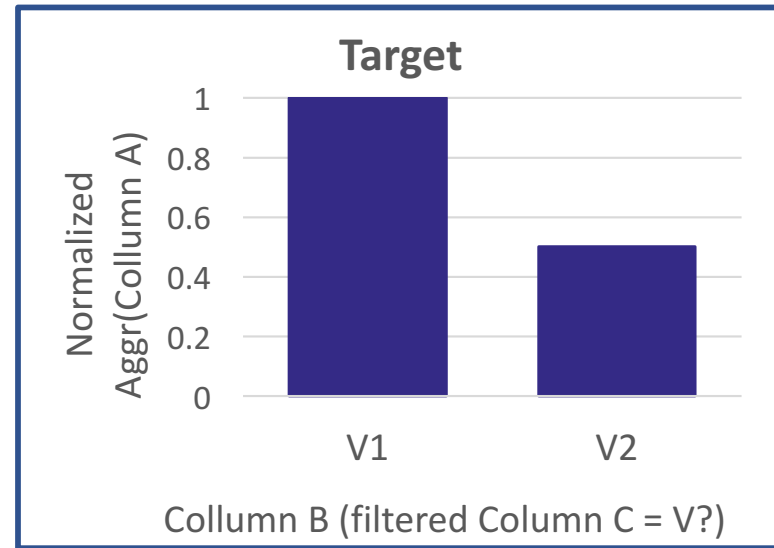
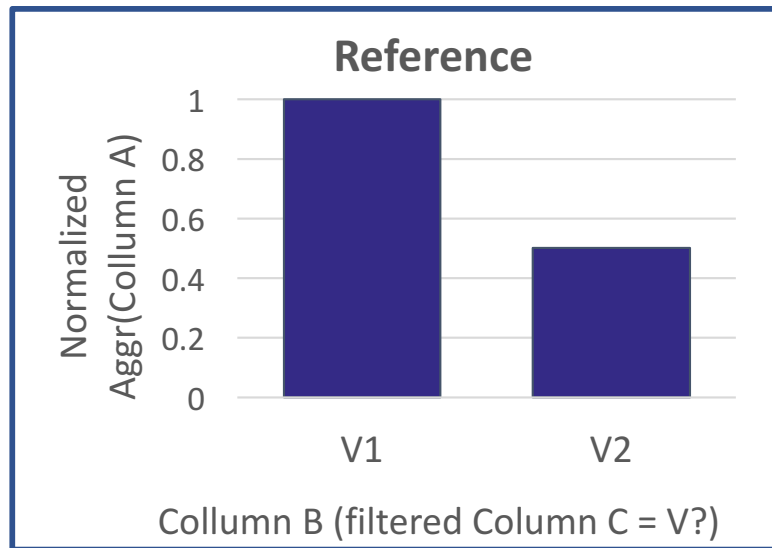
Blaming the multiple-comparison problem on fast visualization-generation is like blaming fast cars for child driver casualties due to car accidents...

But...

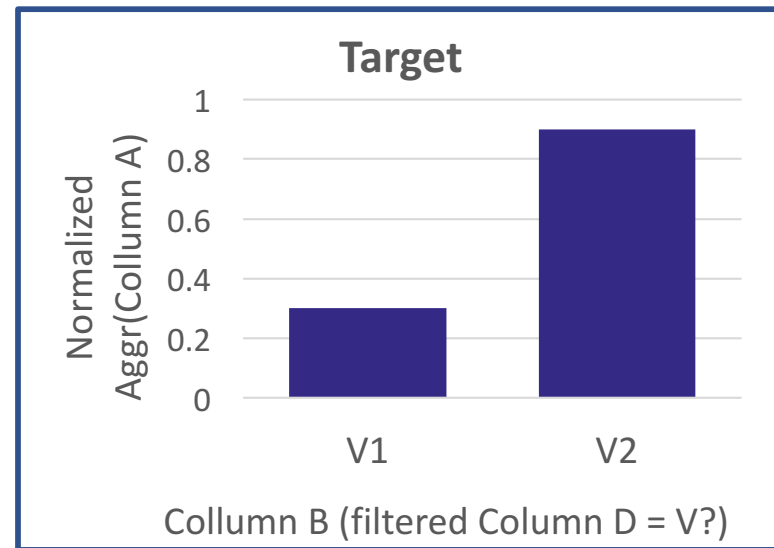




## 2) Visual Recommendation Systems (SeeDB as an Example)

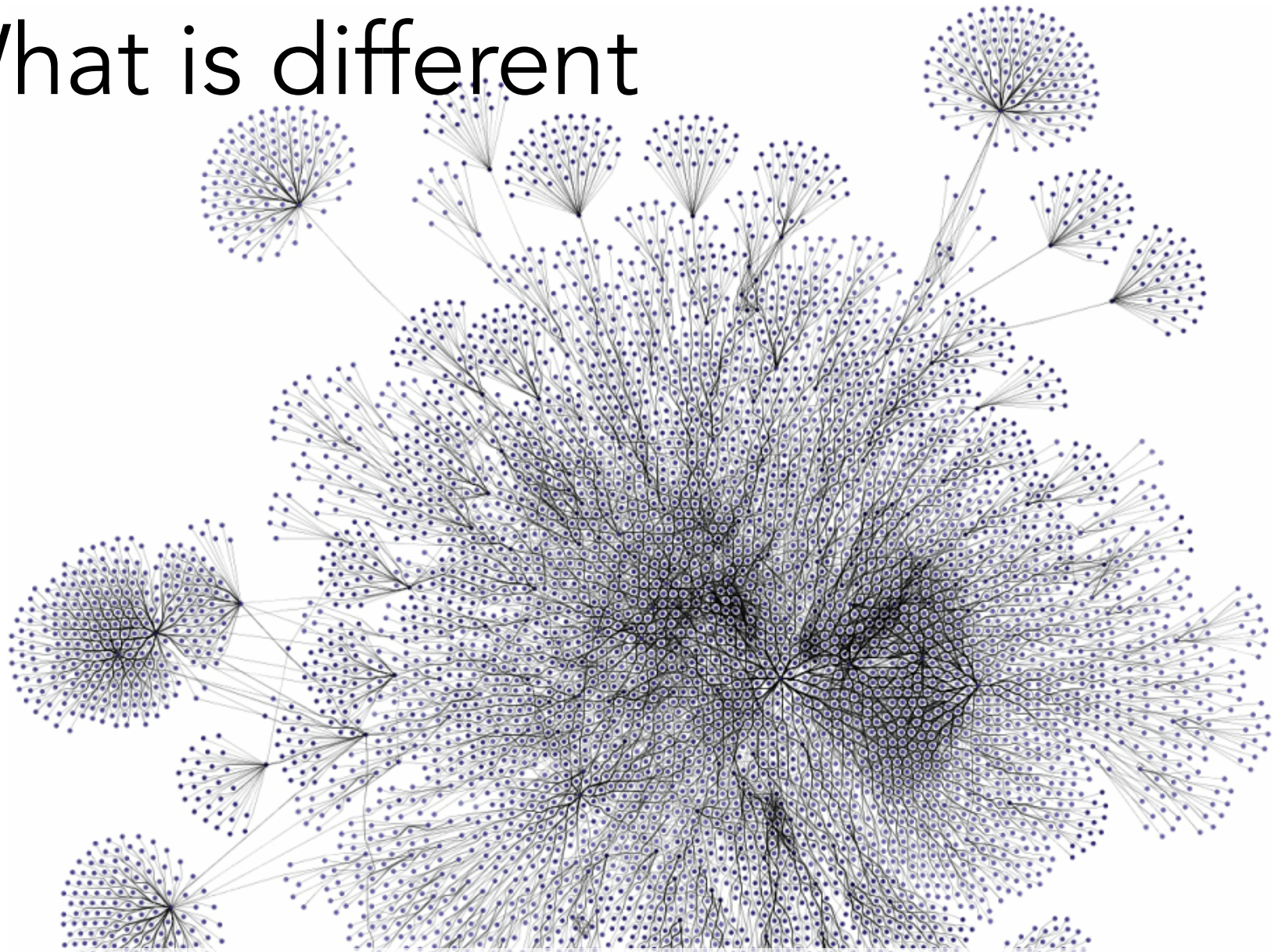


Uninteresting



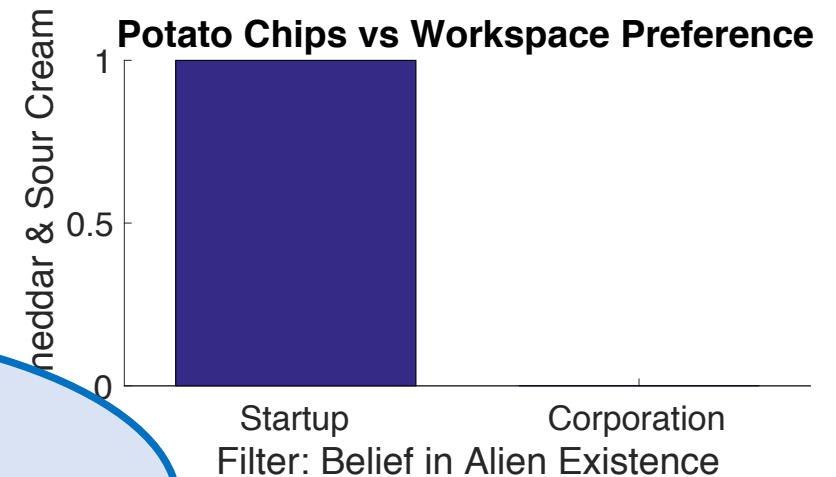
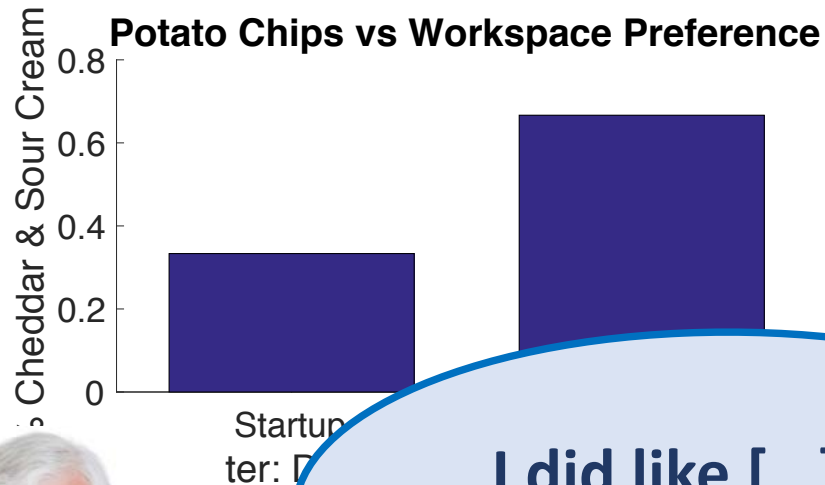
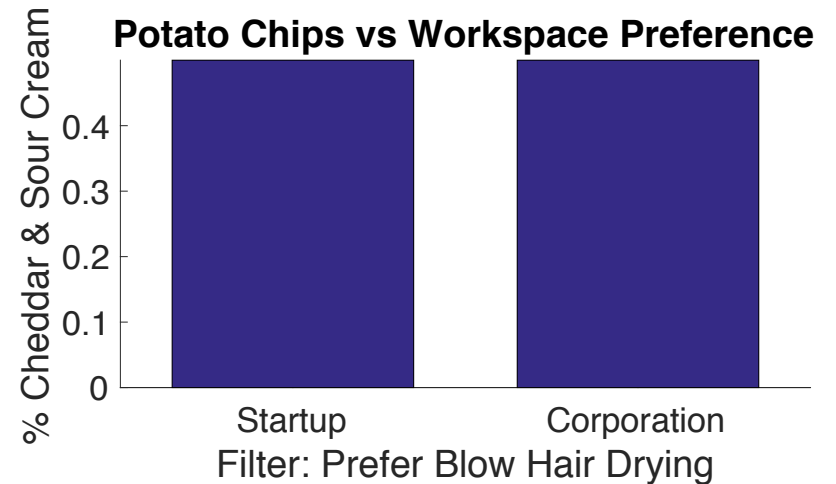
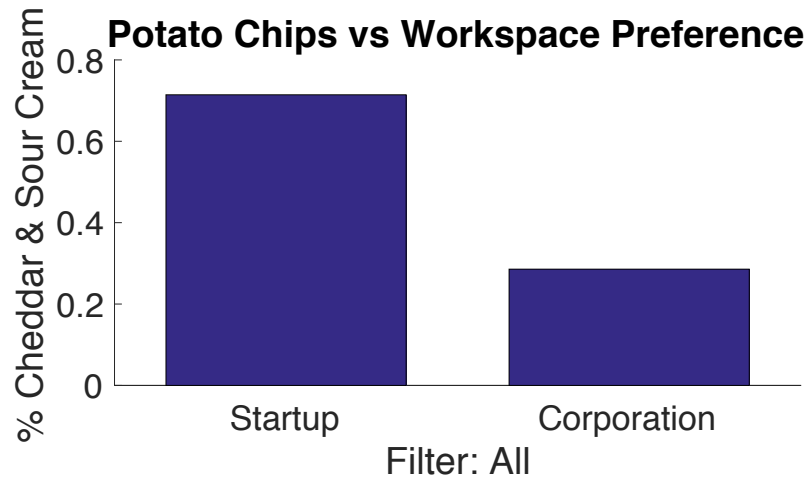
Interesting

# What is different



The system automatically generates thousands of visualizations and ranks them somehow (e.g., based effect size)

# SeeDB on Our Survey Data



...I did like [...] the example ...





# What is the Problem?



The user is in the dark what the system did.  
The system might have "tested" thousands of potential  
visualization, just to find something interesting.

# What did Reviewer 2 say?



These systems are **not designed for an average person** to run and get insights that they can publish medical articles on! The end users are still analysts. The only difference is that **they automate hypotheses generation and NOT hypotheses testing,...**

My suggestions, papers should include in the future a a warning like

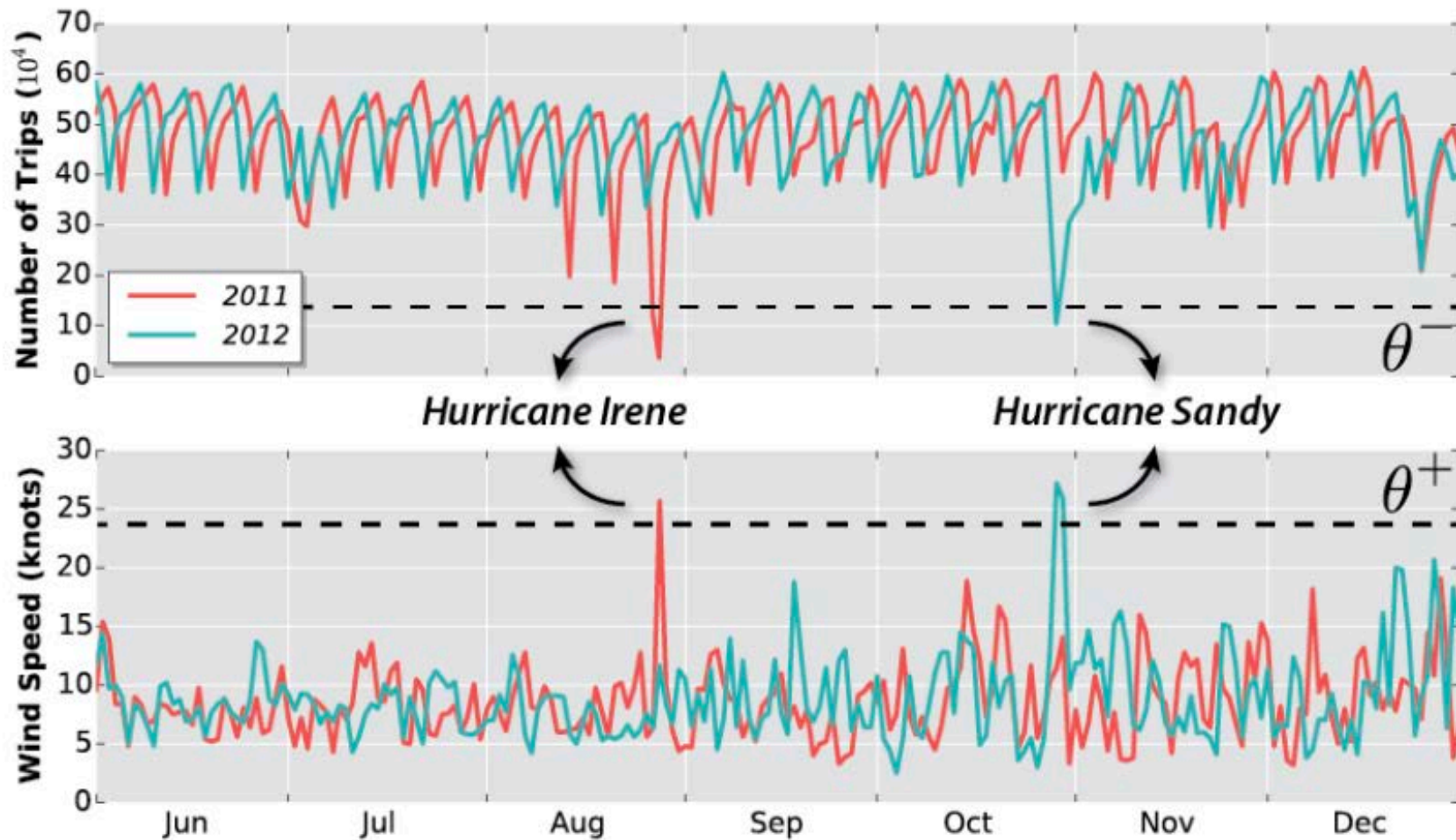
**WARNING**

After using the tool,  
**throw away the data.**

It is not safe!<sup>1</sup>

<sup>1</sup>To be more precise: you do not have to throw it all away, but you can not use the same data anymore for significance testing

### 3) Real Hypothesis Generators (Data Polygamy as an Example)





(Data) Polygamy is bad, especially if *you* do not know what is going on.

# Outline

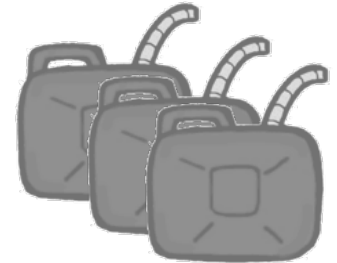
Part I: The problem with:

A. Interactive Data Exploration

B. Visualization Recommendation Systems

C. Hypothesis Generator

**A. Part II: Solutions**



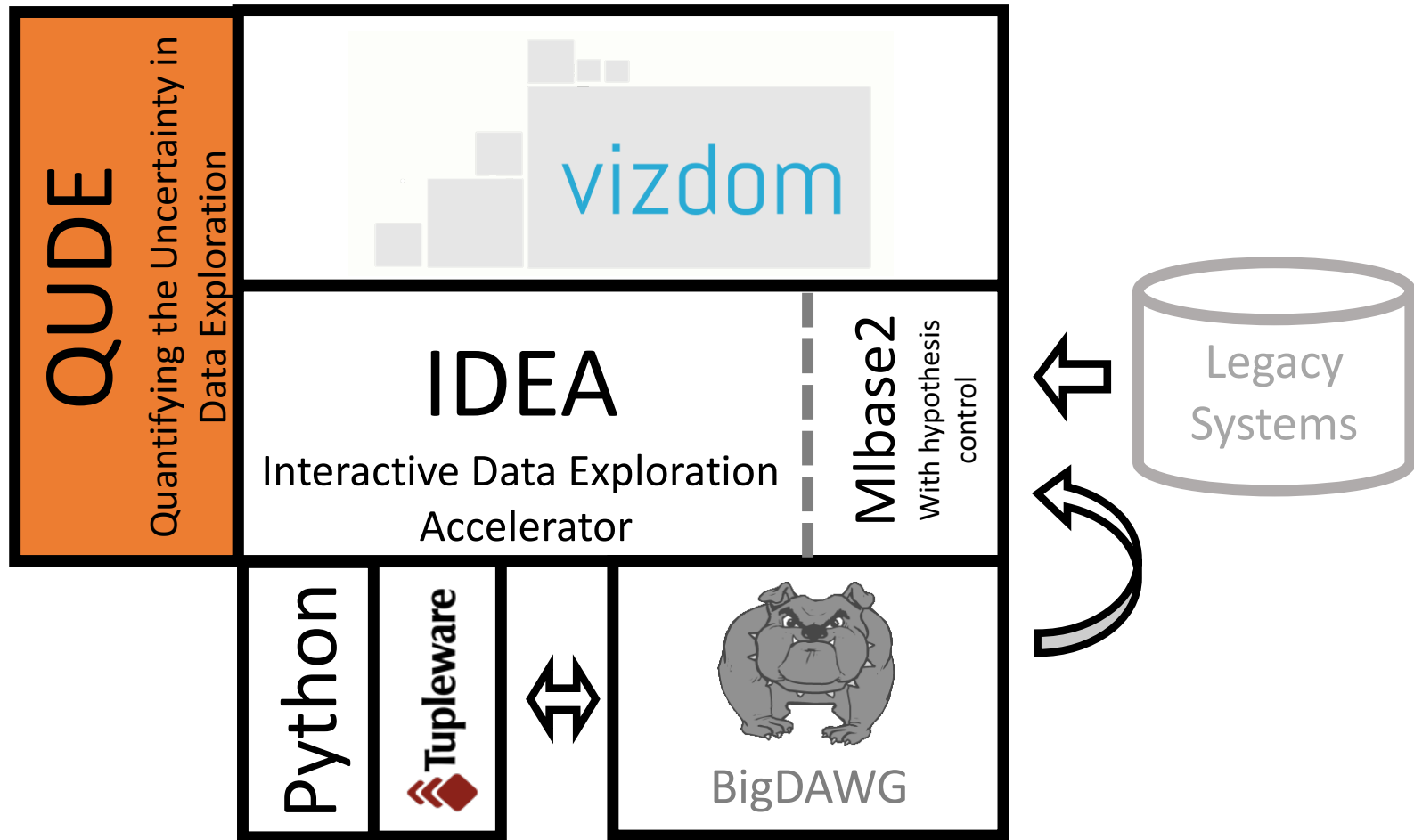
# Should we stop working on IDE, Recommenders, etc?

# NO

- Actively inform the user about the risk factors
- Try your techniques over random data with different data sizes
- *If possible*, split data into **a exploration and a validation set**.
  - Be aware, **significantly lowers the power**
  - Everything on the validation data set has to be carefully handled (i.e., use multi-hypothesis control)
- *If possible*, use **additional experiments** (e.g., A/B testing)
  - Requires a small number of hypothesis and careful design
  - Might not always be possible or is very expensive

**Better: control the multi-hypothesis problem from the start**

# Our Interactive Data Exploration Stack (BIDES)





# Many Interesting Open Problems

We are just at the beginning

- **Transparent hypothesis testing**  
how to automatically derive what the hypothesis is the user is testing
- **How to convey the meaning to the user**  
(e.g., FDR vs family-wise error)
- **Safe recommender techniques**  
(we are currently exploring new techniques based VC-dimensions to control the error)
- **Incremental multiple-hypothesis control techniques**  
(for example, see "Controlling False Discoveries During Interactive Data Exploration" [CoRR abs/1612.01040](#) how we use new alpha-investing policies to do that)
- **Dependencies between hypothesis**  
(this can safe "hypothesis budget")
- ...

# A Final Note from Reviewer 2 on Is the Situation really so Bad?



.., the systems that are criticized by this paper are essentially three tools [4,6,28] ... So the problem is not really as serious as it might seem as **none of these systems are used by anyone in practice**

Tim Kraska  
<tim\_kraska@brown.edu>

Special thanks to:



A last note to Reviewer 2:

- 1<sup>st</sup> I sincerely hope you are not one of my letter writers for my tenure case :)
- 2<sup>nd</sup> Your comments actually helped us to improve the paper and helped with the talk. So thank you!
- 3<sup>rd</sup> I am happy to pay for your drinks tonight to make it up to you.