Three Tools for "Human-in-the-loop" Data Science

Aditya Parameswaran
Assistant Professor
University of Illinois

http://data-people.cs.illinois.edu

Many many contributors!

- **PIs**: Kevin Chang, Karrie Karahalios, Aaron Elmore, Sam Madden, Amol Deshpande (Spanning Illinois, UMD, MIT, Chicago)
- **PhD Students**: Mangesh Bendre, Himel Dev, John Lee, Albert Kim, Manasi Vartak, Liqi Xu, Silu Huang, Sajjadur Rahman, Stephen Macke
- **MS Students**: Vipul Venkataraman, Tarique Siddiqui, Chao Wang, Sili Hui
- **Undergrads**: Paul Zhou, Ding Zhang, Kejia Jiang, Bofan Sun, Ed Xue, Sean Zou, Jialin Liu, Changfeng Liu, Xiaofo Yu

Google Faculty Research Awards
SIEBEL ENERGY INSTITUTE
NIH Big Data to Knowledge (BD2K)

Scale is a Solved Problem

Most work in the database community is **myopically focused on scale:** *the ability to pose SQL queries on larger and larger datasets*.

My claim:
Scale is a solved problem.

Nobody ever got fired for using Hadoop on a cluster

Antony Rowstron Dushyanth Narayanan
Austin Donnelly Greg O'Shea Andrew Douglas
Microsoft Research, Cambridge

Findings:

- Median job size at Microsoft and Yahoo is 16GB;
- >90% of the jobs within Facebook are <100GB

The bottleneck is no longer our ability to pose SQL queries on large datasets!

Of course, exceptions exist: **the "1%"** of data analysis needs

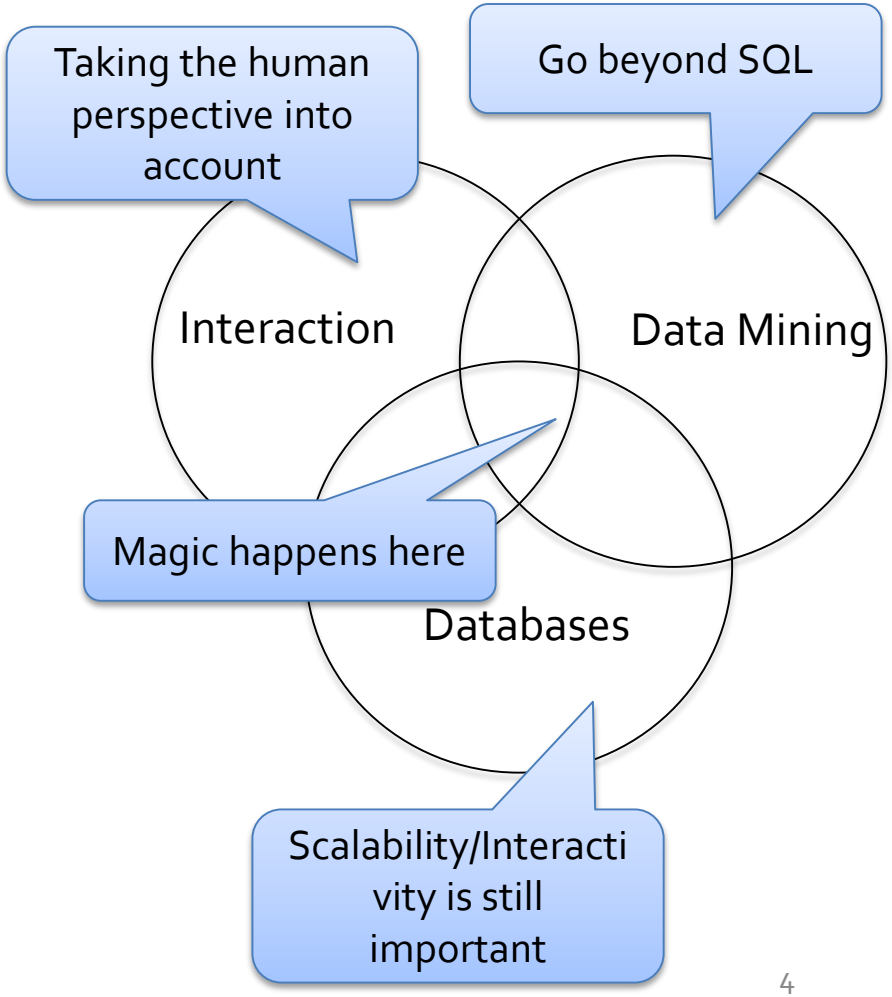
What about the Needs of the 99%?

The bottleneck is actually the **"humans-in-the-loop"**

As our data size has grown, what has stayed constant is

- the **time** for analysis,
- the **cognitive load**,
- the analysis **skills**

➔ **Human-in-the-loop Data Analytics** (HILDA) tools

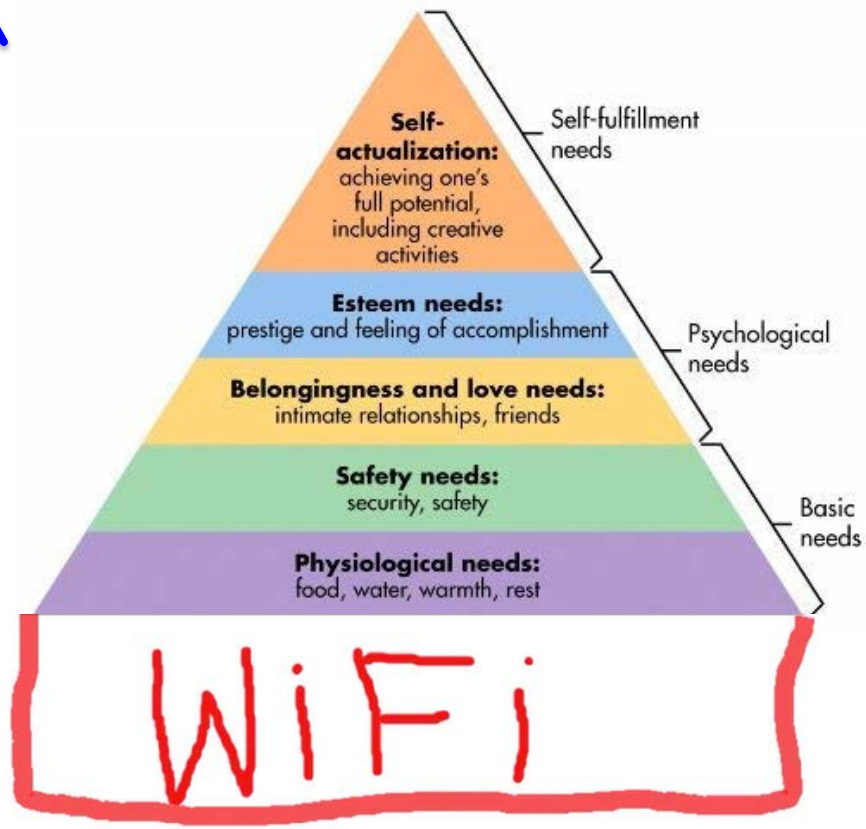


A Maslow's Hierarchy for HILDA

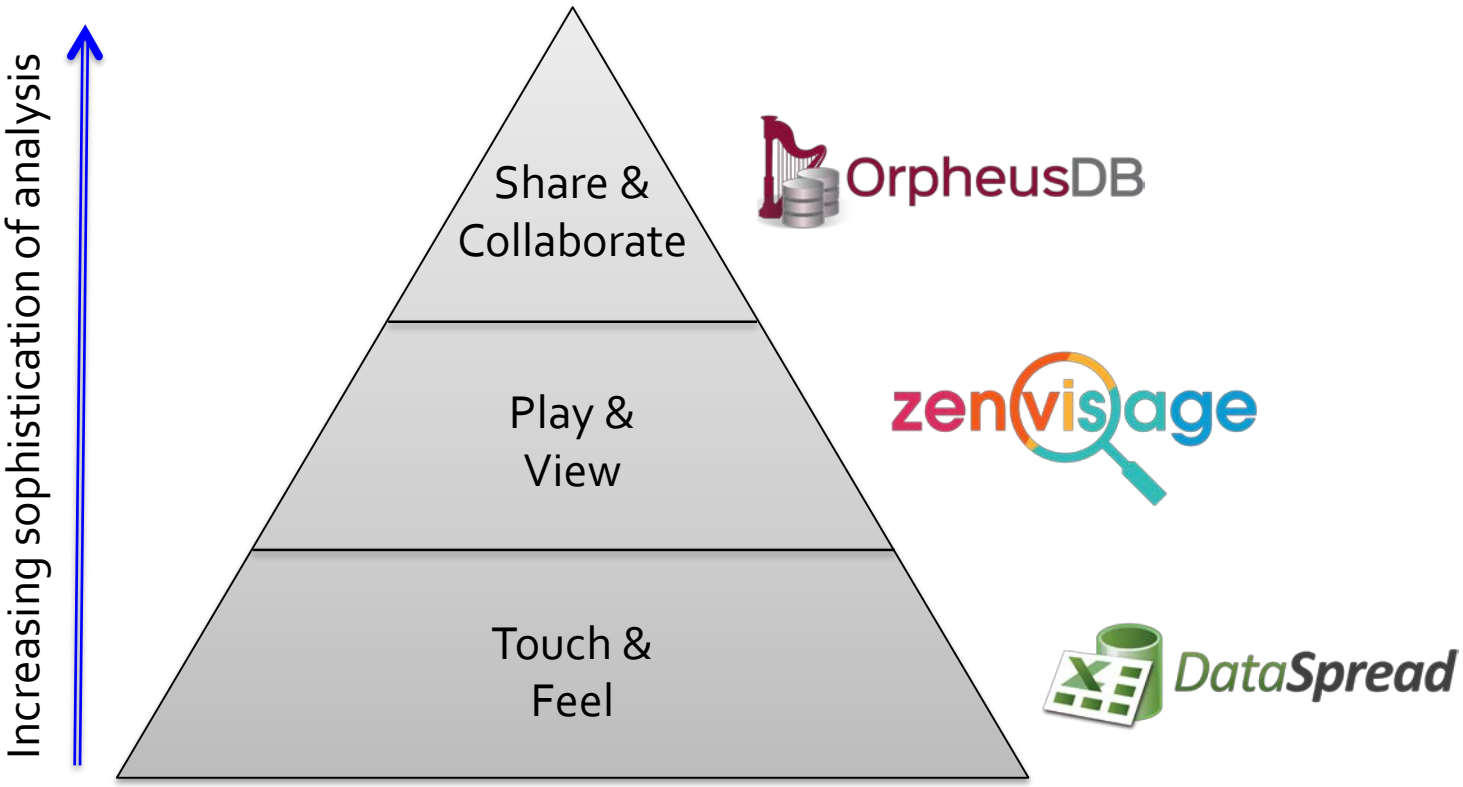
Background: Maslow developed a theory for what motivates individuals in 1943; highly influential

Complex Needs

Basic Needs



A Maslow's Hierarchy for HILDA



Touch and Feel: DataSpread

DataSpread is a **spreadsheet-database hybrid**:

Goal: Marrying the flexibility and ease of use of spreadsheets with the scalability and power of databases

Enables the "99%" with large datasets but limited prog. skills to open, touch, and examine their datasets

http://dataspread.github.io

[VLDB'15,VLDB'15,ICDE'16]

Play and View:

Zenvisage is **effortless visual exploration tool.**

Goal: "fast-forward" to visual patterns, trends, without having analyst step through each one individually

Enables individuals to play with, and extract insights from large datasets at a fraction of the time.

http://zenvisage.github.io

[TR'16,VLDB'16,VLDB'15,DSIA'15,VLDB'14,VLDB'14]

Collaborate and Share: OrpheusDB

OrpheusDB is a tool for **managing dataset versions** with a database

Goal: building a versioned database system to reduce the burden of recording datasets in various stages of analysis

Enables individuals to collaborate on data analysis, and share, keep track of, and retrieve dataset versions.

http://orpheus-db.github.io

[VLDB'16,VLDB'15,VLDB'15,TAPP'15,CIDR'15]

(also part of datahub : a collab. analysis system w/ MIT & UMD)

Combining the benefits of spreadsheets and databases

Spreadsheet as a frontend interface

Databases as a backend engine

Result: retain the benefits of both!

But it's not that simple…

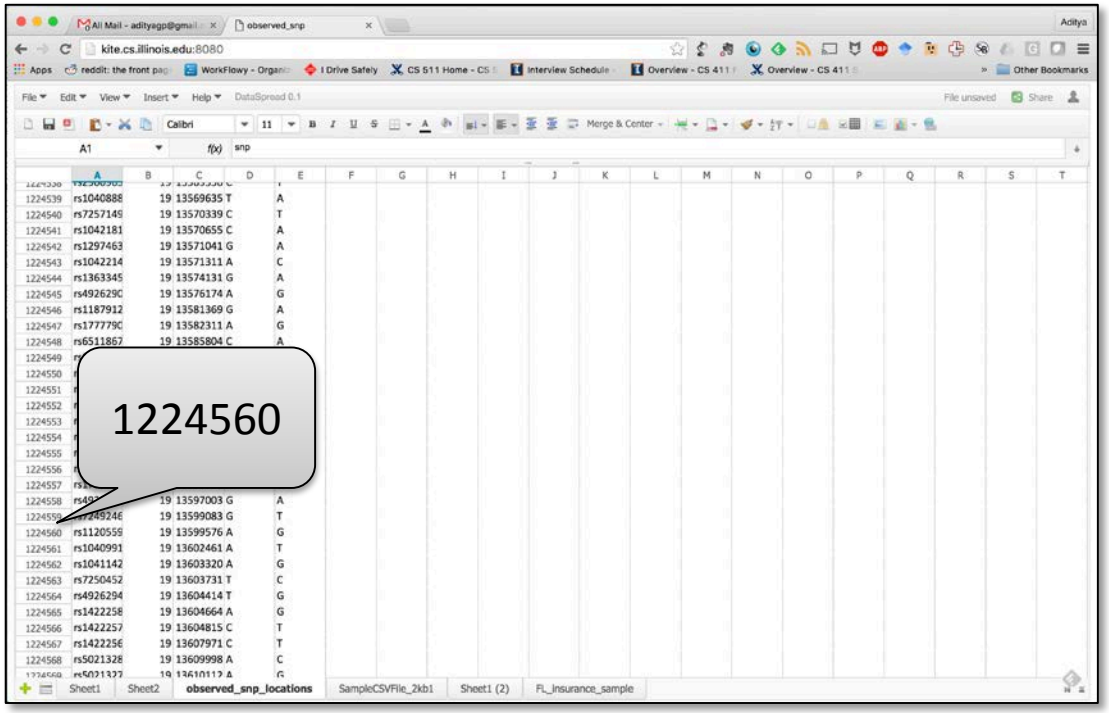
Different Ideologies

Databases and spreadsheets have different ideologies that need to be reconciled…

Feature	Databases	Spreadsheets
Data Model	Schema-first	Dynamic/No Schema
Addressing	Tuples with PK	Cells, using Row/Col
Presentation	Set-oriented, no such notion	Notion of current window, order
Modifications	Must correspond to queries	Can be done at any granularity
Computation	Query at a time	Value at a time

Due to this, the integration is not trivial…

Initial Progress and Architecture



Postgres backend

ZK spreadsheet

- open-source web frontend

Comfortably scales to arbitrarily many rows

+ handle SQL queries

Hopefully bring spreadsheets to the big data age!

Standard Visual Data Analysis Recipe:

1. Load dataset into viz tool
2. Select viz to be generated
3. See if it matches desired visual pattern
4. Repeat until you find a match

➔ *Tedious and time-consuming*

Effortless Visual Exploration of Large Datasets with zenvisage

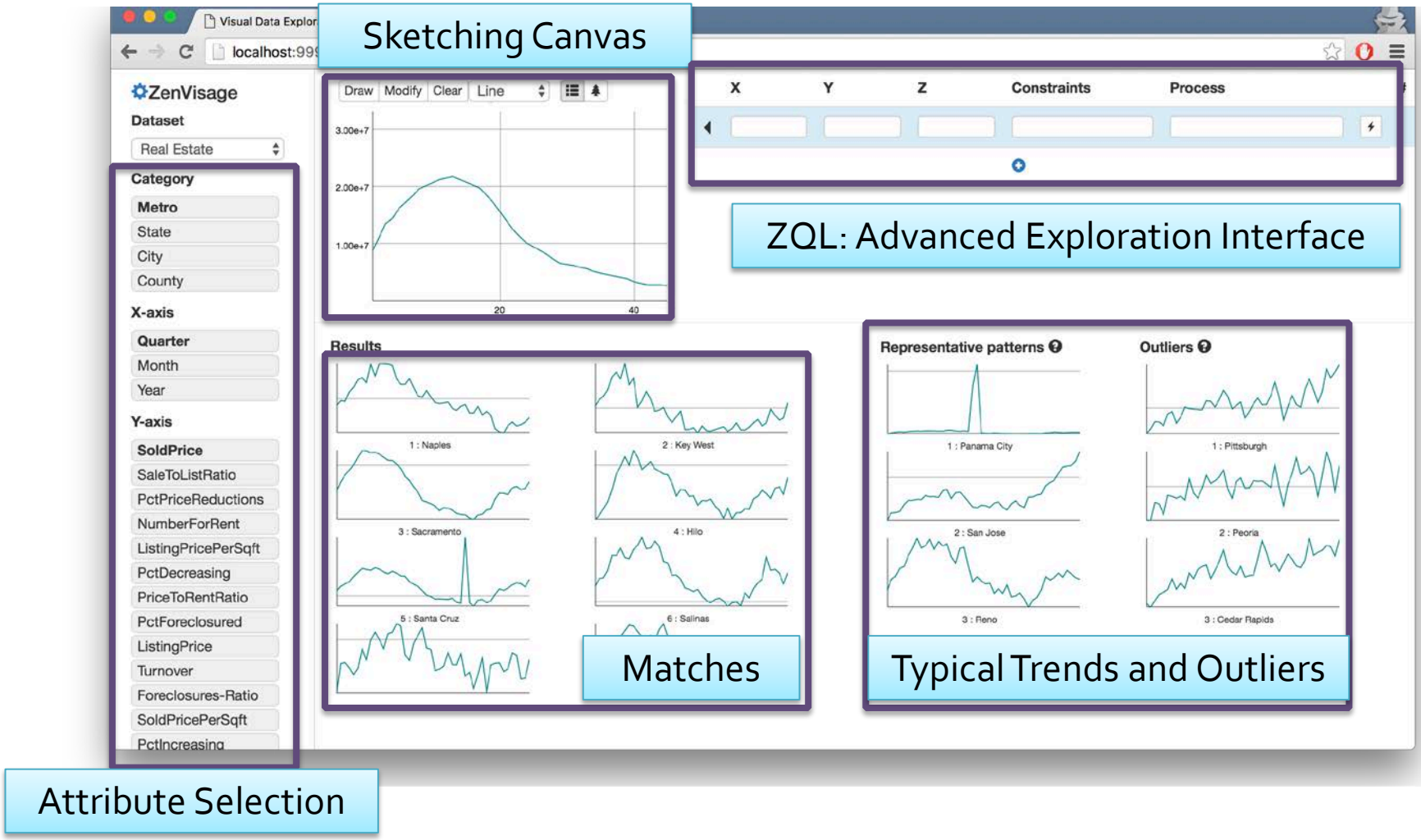
We can automate that!

- instead of combing through visualizations manually
- tell us what you want, and we can "fast-forward"

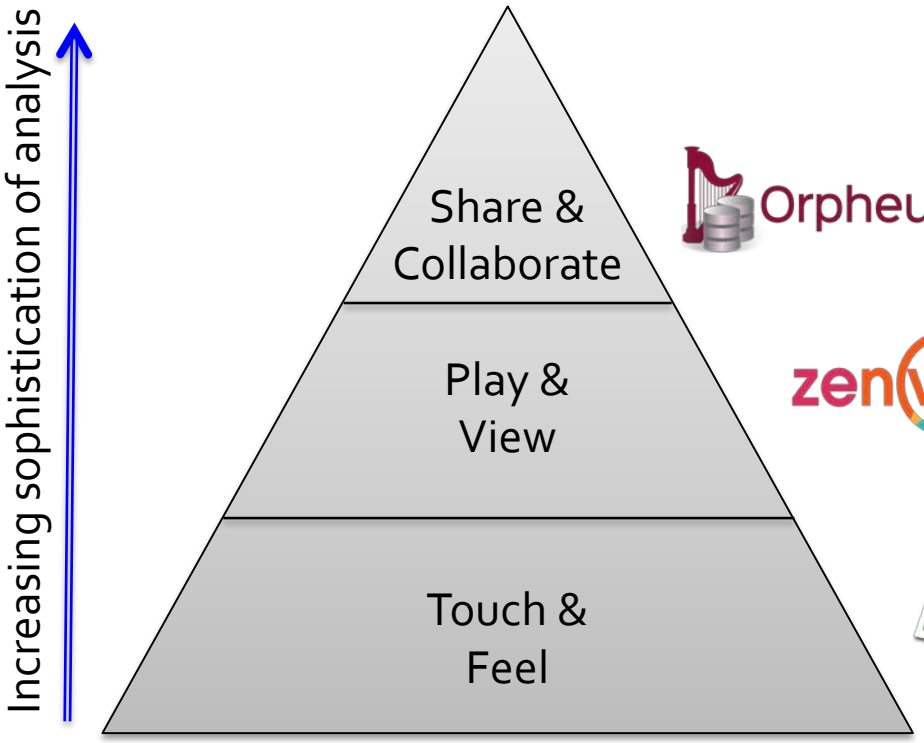
Ingredients:

- *Drag-and-drop and sketch based interactions*
  - to find specific patterns
- *Sophisticated visual exploration language, ZQL*
  - to ask more elaborate questions
- *Scalable visualization generation engine*
  - preprocess, batch and parallel eval. for interactive results
- *Rapid pattern matching algorithms*
  - sampling-based techniques

Screenshots



Summary:
Make Data Analytics Great Again!



OrpheusDB — orpheus-db.github.io

zenvisage — zenvisage.github.io

DataSpread — dataspread.github.io

My website: http://data-people.cs.illinois.edu

Twitter: @adityagp