

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, Xiaofu Yu



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

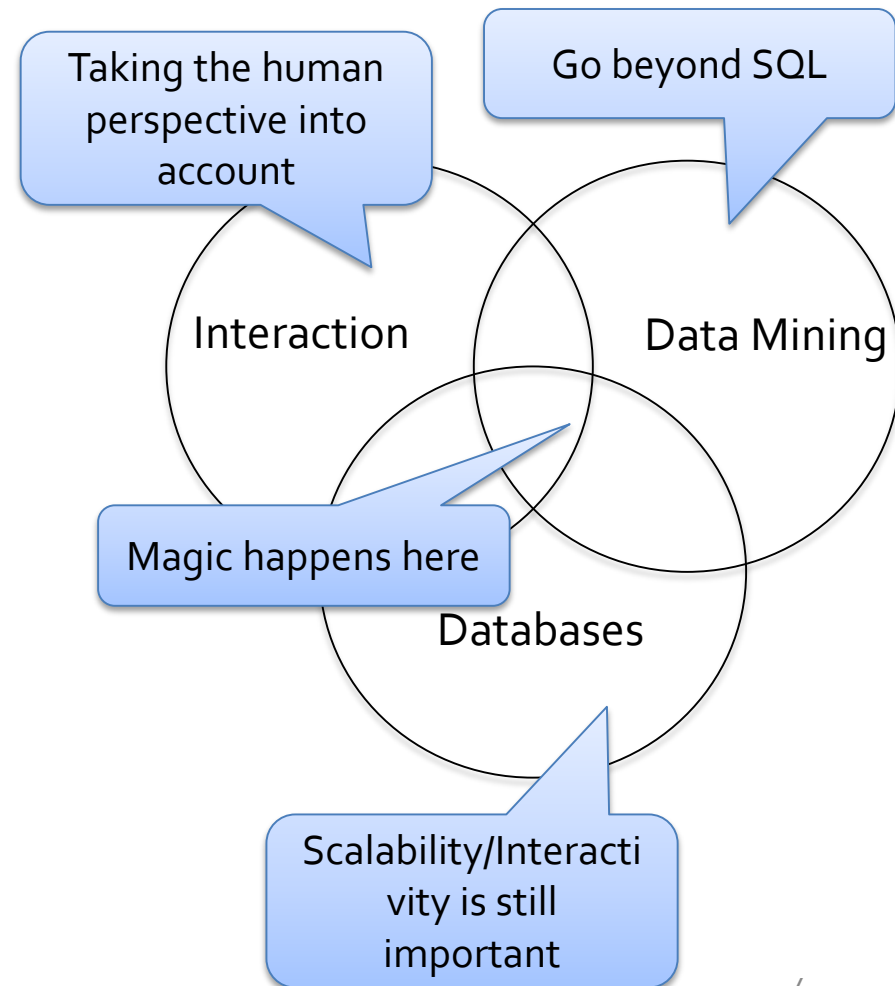
What about the Needs of the 99%?

The bottleneck is actually the **“humans-in-the-loop”**

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

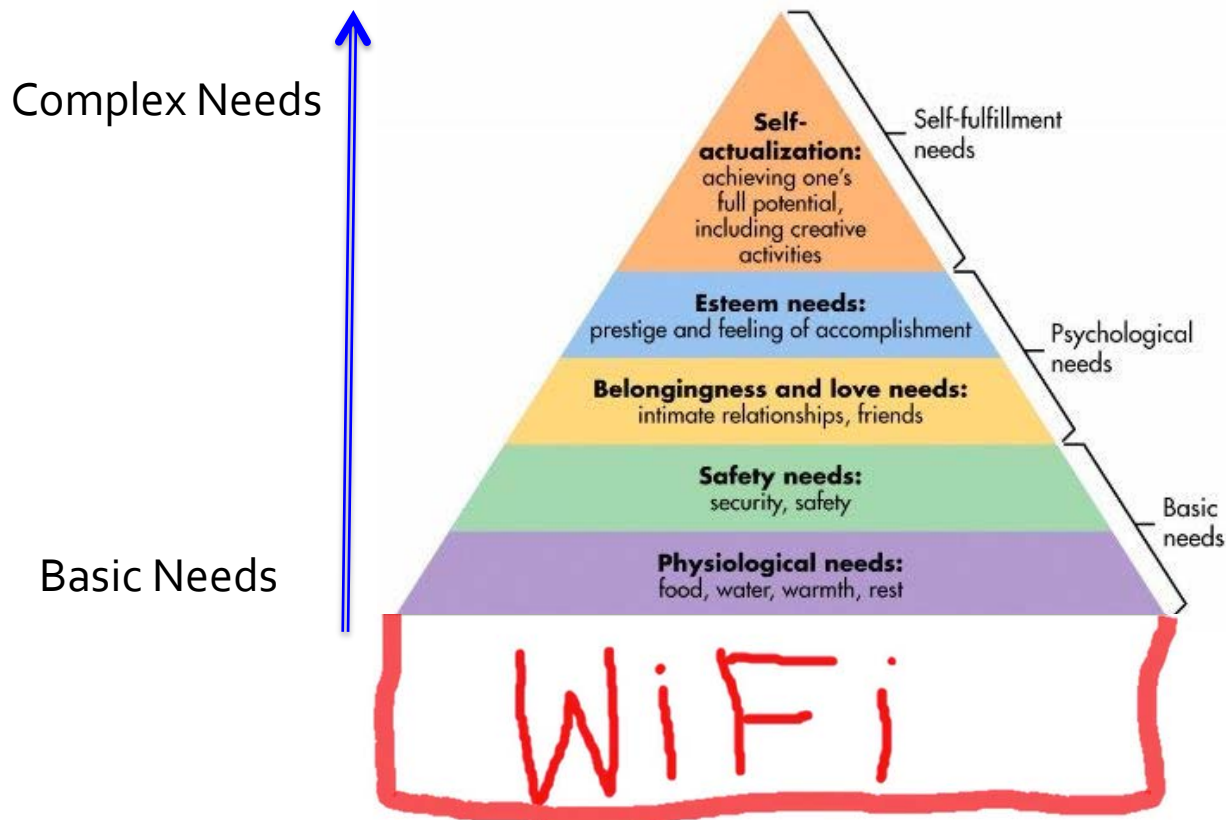
- the **time** for analysis,
- the **cognitive load**,
- the analysis **skills**

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

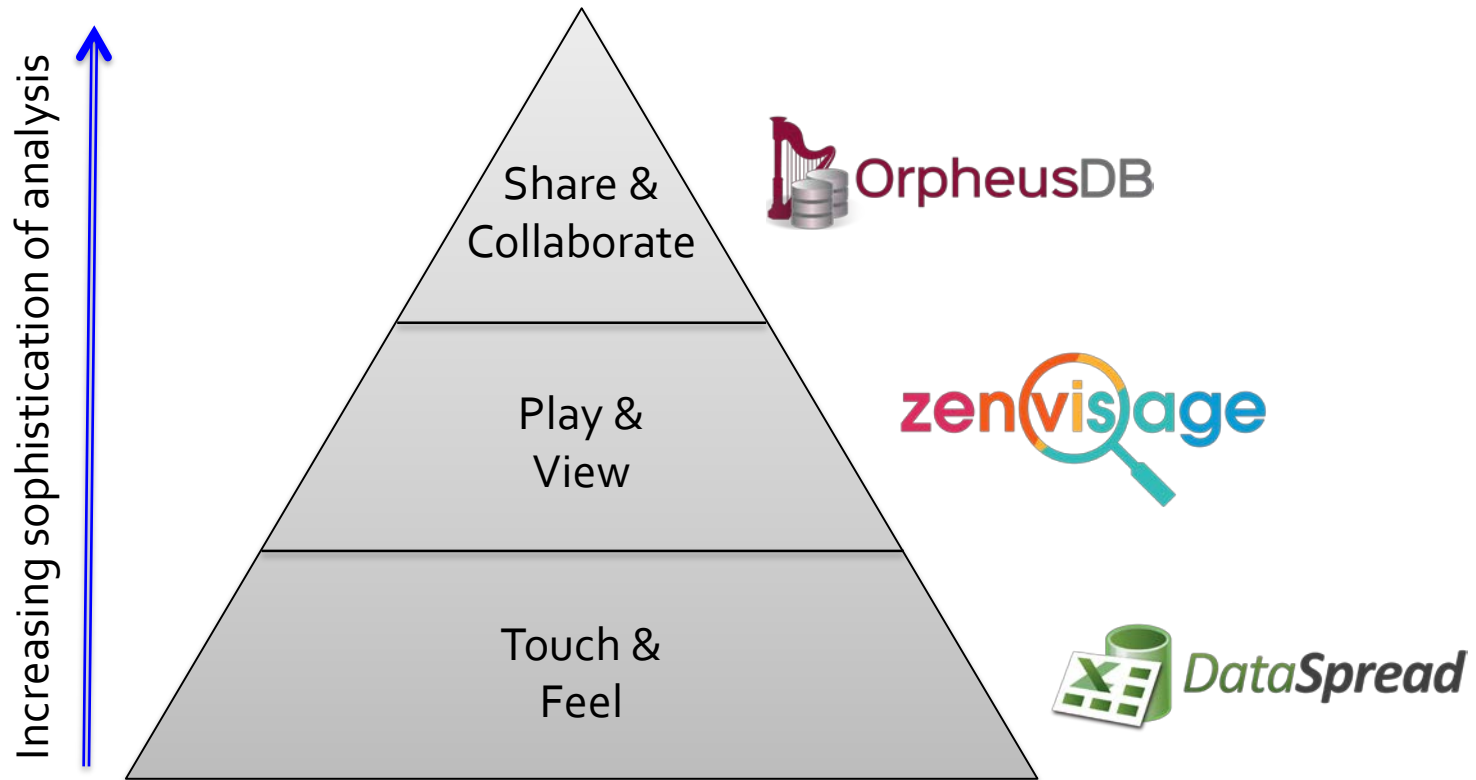


A Maslow's Hierarchy for HILDA

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



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  : a collab. analysis system w/ MIT & UMD)
datahub

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



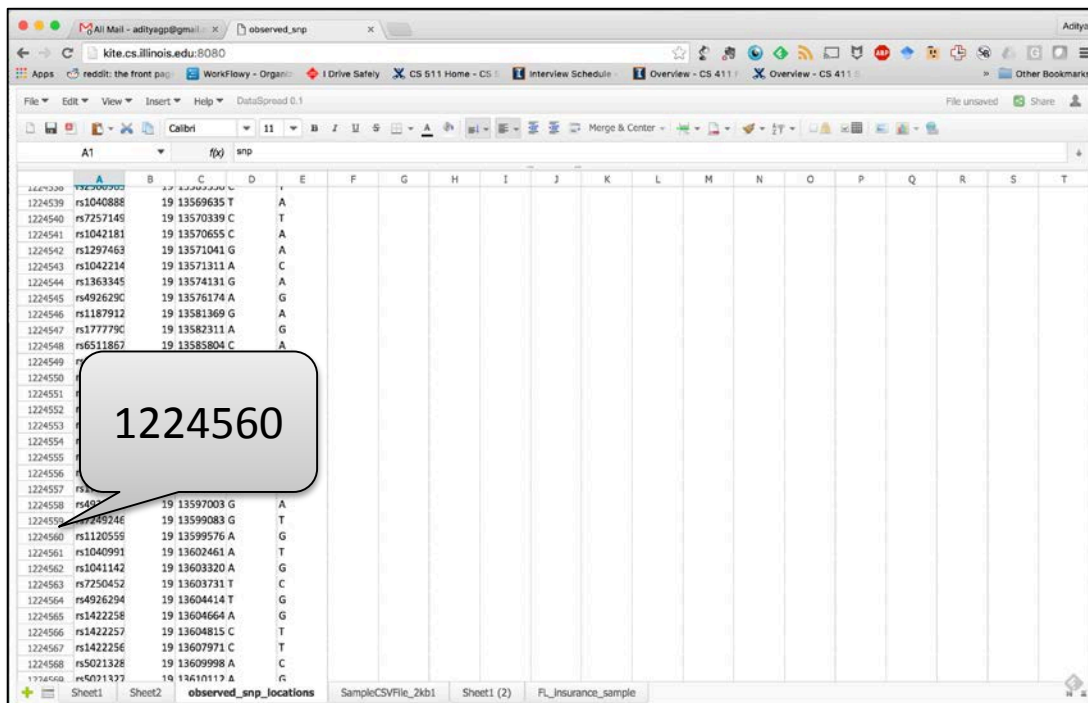
Postgres backend

ZK spreadsheet

- open-source web frontend

Comfortably scales to arbitrarily many rows + handle SQL queries

Hopefully bring spreadsheets to the big data age!



The screenshot shows a web browser window displaying a spreadsheet application. The spreadsheet has a grid with columns labeled A through T and rows numbered 1224530 through 1224569. A speech bubble points to row 1224560, which contains the value "1224560" in column A. The spreadsheet is titled "observed_snp" and is part of a larger application named "DataSpread 0.1".

AT	fn	snp
1224530	rs1040888	19 13569635 T
1224539	rs1040888	19 13569635 T
1224540	rs7257148	19 13570339 C
1224541	rs1042181	19 13570655 C
1224542	rs1297463	19 13571041 G
1224543	rs1042214	19 13571311 A
1224544	rs1363345	19 13574131 G
1224545	rs4926290	19 13576174 A
1224546	rs1187912	19 13581369 G
1224547	rs1777790	19 13582311 A
1224548	rs6511867	19 13585804 C
1224549		
1224550		
1224551		
1224552		
1224553		
1224554		
1224555		
1224556		
1224557		
1224558	rs4926290	19 13597003 G
1224559	rs7257148	19 13599083 G
1224560	rs1120556	19 13599576 A
1224561	rs1040991	19 13602461 A
1224562	rs1041142	19 13603320 A
1224563	rs7250452	19 13603731 T
1224564	rs4926290	19 13604414 T
1224565	rs1422258	19 13604664 A
1224566	rs1422257	19 13604815 C
1224567	rs1422256	19 13607971 C
1224568	rs5021328	19 13609998 A
1224569	rs6071327	19 13610117 A

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



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

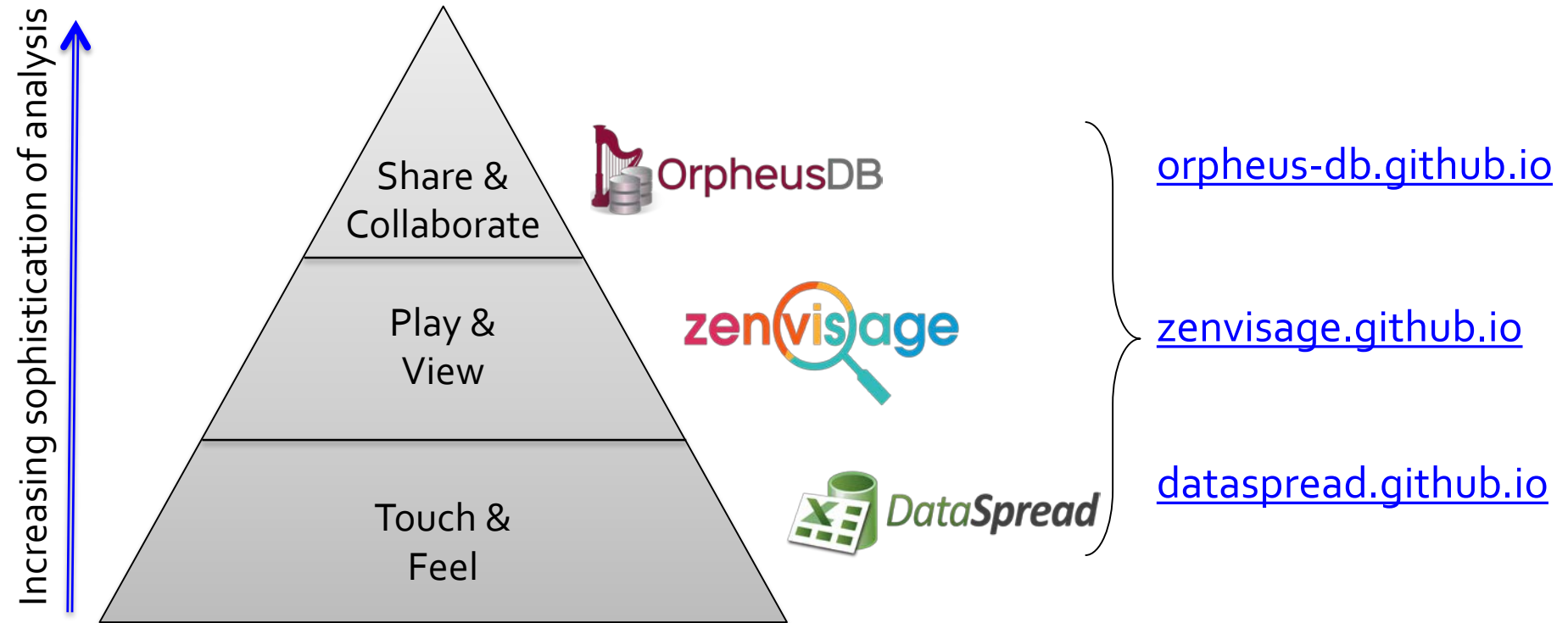
The screenshot displays the ZenVisage interface with several key components highlighted by blue callout boxes:

- Attribute Selection:** A sidebar on the left lists various attributes under categories like Metro, X-axis, Y-axis, and Quarter. The 'Real Estate' dataset is selected.
- Sketching Canvas:** A large central area with a drawing toolbar (Draw, Modify, Clear, Line) and a grid. A green curve is plotted on the canvas.
- ZQL: Advanced Exploration Interface:** A control bar with fields for X, Y, Z, Constraints, and Process, and a lightning bolt icon.
- Matches:** A 'Results' section showing six small line charts for different locations: 1: Naples, 2: Key West, 3: Sacramento, 4: Hilo, 5: Santa Cruz, and 6: Salinas.
- Typical Trends and Outliers:** Two sections: 'Representative patterns' showing 1: Panama City and 2: San Jose, and 'Outliers' showing 1: Pittsburgh and 2: Peoria, and 3: Cedar Rapids.

Attribute Selection

Summary:

Make Data Analytics Great Again!



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

Twitter: @adityagp