Buckaroo

Towards Scalable Visual Data Wrangling via Direct Manipulation

El Kindi Rezig
University of Utah
elkindi.rezig@utah.edu

Mir Mahathir Mohammad
University of Utah
mahathir.mohammad@utah.edu

Nicolas Baret
University of Utah
nicolas.baret@utah.edu

Ricardo Mayerhofer
Hopara, Inc
ricardo@hopara.io

Andrew McNutt
University of Utah
andrew.mcnutt@utah.edu

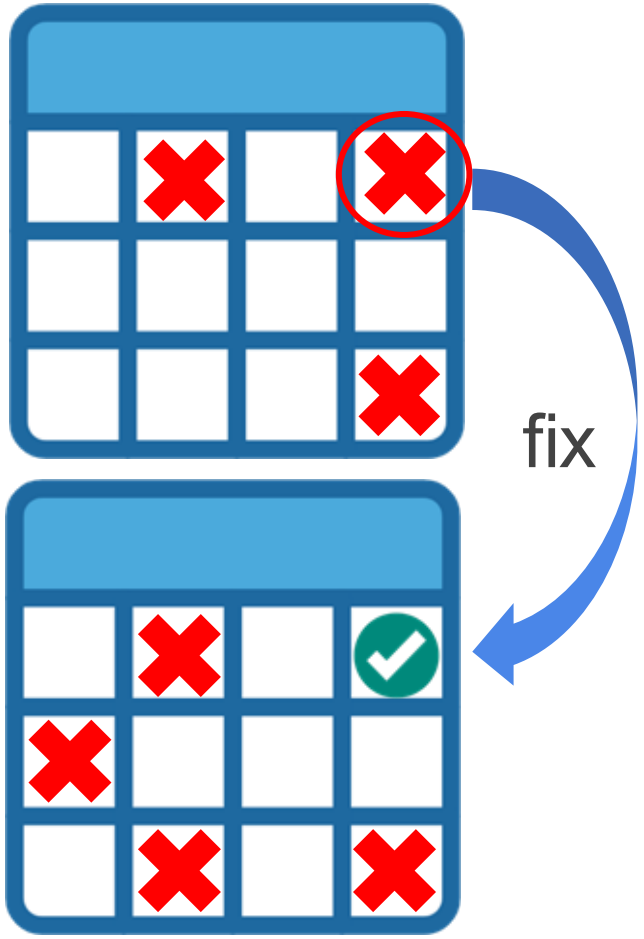
Paul Rosen
University of Utah
paul.rosen@utah.edu

Data preparation is the most time-consuming part of data science development

Data preparation is the most time-consuming part of data science development

Data errors/anomalies are rare →
Hard to spot

Data preparation is the most time-consuming part of data science development



Fixing an error may lead to new errors

But wait, we have lots of data preparation systems out there

But wait, we have lots of data preparation systems out there

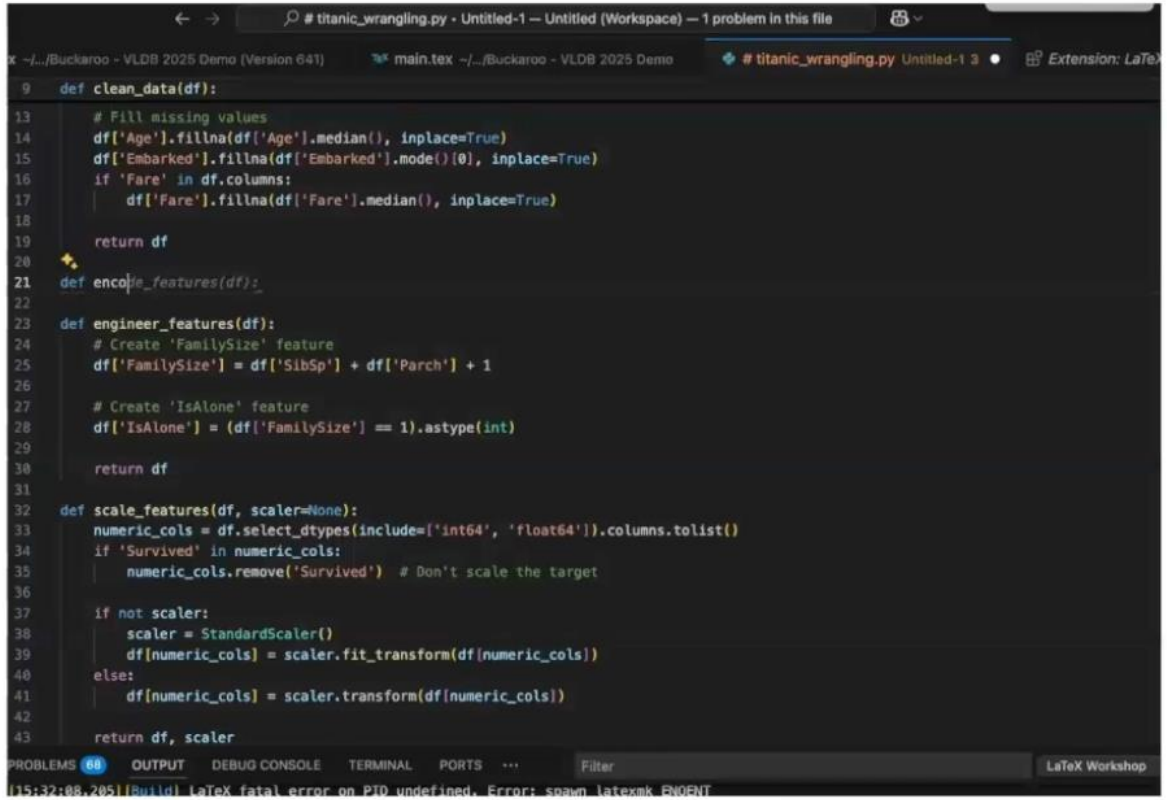
But they often require a steep learning curve, and are specific to certain data errors

But wait, we have lots of data preparation systems out there

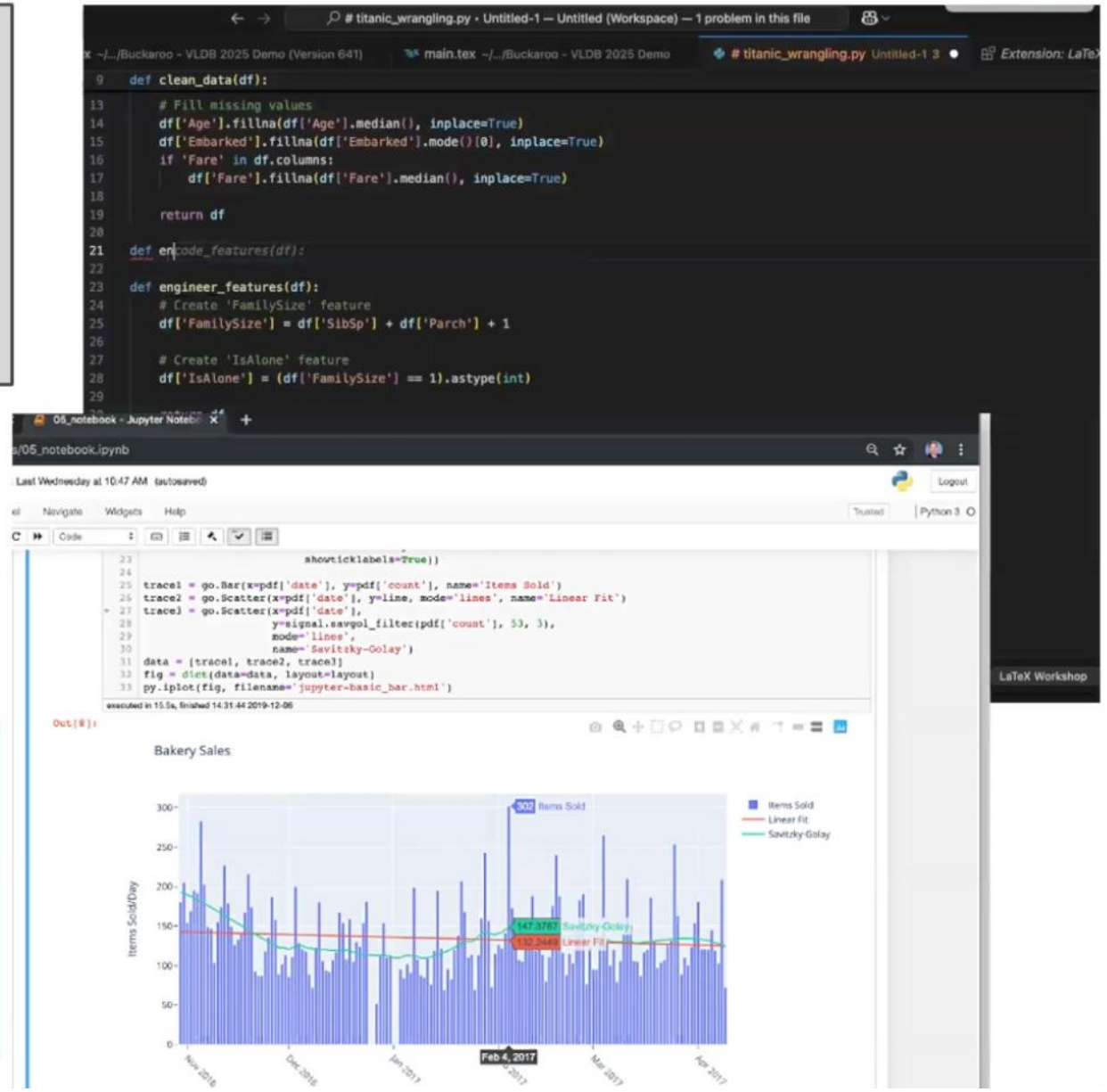
But they often require a steep learning curve, and are specific to certain data errors

There has been little focus on usability of data preparation systems

To prepare datasets, data scientists spend a lot of time refining their scripts to detect and repair data errors



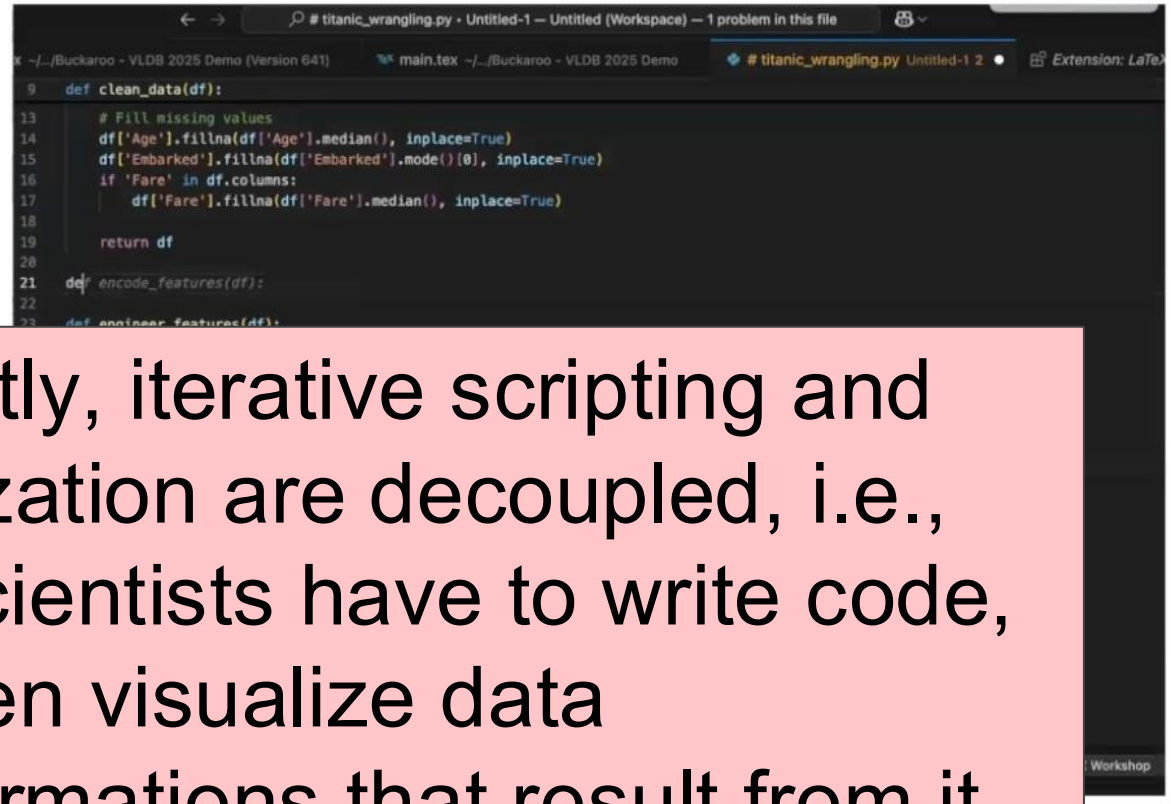
To prepare datasets, data scientists spend a lot of time refining their scripts to detect and repair data errors



Data scientists rely on charts as powerful tools to explore and validate intermediate steps in data transformation

To prepare datasets, data scientists spend a lot of time refining their scripts to detect and repair data errors

Data scientists rely on charts as powerful tools to explore and validate intermediate steps in data transformation



Currently, iterative scripting and visualization are decoupled, i.e., data scientists have to write code, and then visualize data transformations that result from it

To prepare datasets, data scientists spend a lot of time refining their scripts to detect and repair data errors

Buckaroo

In Buckaroo, data wrangling is seamlessly integrated with interactive charts, allowing users to directly manipulate visualizations to correct errors and instantly see the results

Data scientists rely on charts as powerful tools to explore and validate intermediate steps in data transformation

Potter's Wheel: An Interactive Data Cleaning System

Vijayshankar Raman and Joseph M. Hellerstein

University of California at Berkeley
{rshankar, jmh}@cs.berkeley.edu

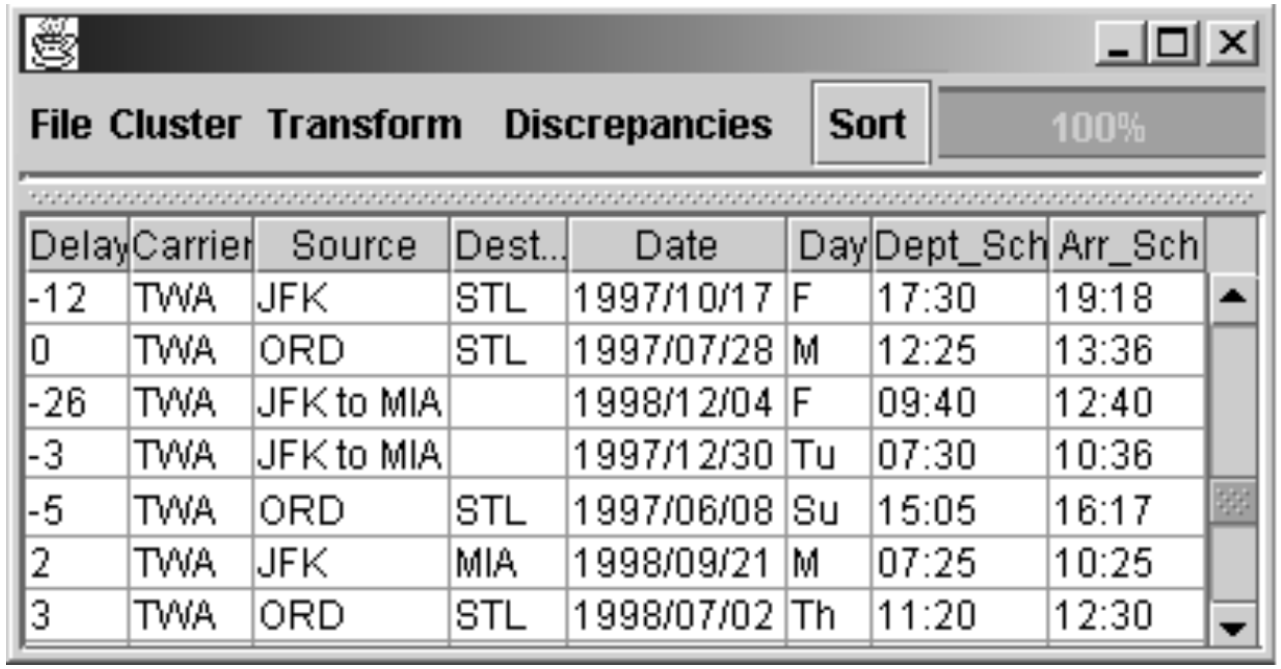


Figure 1: A snapshot of the Potter's Wheel user interface on flight delay data from FEDSTATS (www.fedstats.gov).

Existing work

SEEDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics

Manasi Vartak[1] Sajjadur Rahman[2] Samuel Madden[1] Aditya Parameswaran[2] Neoklis Polyzotis[3]

[1]MIT {mvartak, madden}@csail.mit.edu
[2]University of Illinois (UIUC) {srahman7, adityagp}@illinois.edu
[3]Google npolyzotis@google.com

Potter's Wheel: An Interactive Data Cleaning System

Vijayshankar Raman and Joseph M. Hellerstein

University of California at Berkeley
{rshankar, jmh}@cs.berkeley.edu

Wrangler: Interactive Visual Specification of Data Transformation Scripts

Sean Kandel, Andreas Paepcke, Joseph Hellerstein, Jeffrey Heer
{skandel, paepcke, jheer}@cs.stanford.edu; joeh@cs.berkeley.edu

Visualizing the Scripts of Data Wrangling with SOMNUS

Kai Xiong, Siwei Fu, Guoming Ding, Zhongsu Luo, Rong Yu, Wei Chen, Hujun Bao, Yingcai Wu

Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment

Sean Kandel*, Ravi Parikh*, Andreas Paepcke*, Joseph M. Hellerstein†, Jeffrey Heer*
*Stanford University †University of California, Berkeley
{skandel, rparikh, paepcke, jheer}@cs.stanford.edu, hellerstein@cs.berkeley.edu

Lux: Always-on Visualization Recommendations for Exploratory Dataframe Workflows

Doris Jung-Lin Lee, Dixin Tang, Kunal Agarwal, Thyne Boonmark, Caitlyn Chen, Jake Kang, Ujjaini Mukhopadhyay, Jerry Song, Micah Yong, Marti A. Hearst, Aditya G. Parameswaran
UC Berkeley
{dorislee,totemtang,kagarwal2,thyneboonmark,caitlynachen,cjache,ujjaini,jerrysong,micahtyong,hearst,adityagp}@berkeley.edu

Existing work

SEEDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics

Manasi Vartak[1] Sajjadur Rahman[2] Samuel Madden[1] Aditya Parameswaran[2] Neoklis Polyzotis[3]

[1]MIT {mvartak, madden}@csail.mit.edu
[2]University of Illinois (UIUC) {srahman7, adityagp}@illinois.edu
[3]Google npolyzotis@google.com

Potter's Wheel: An Interactive Data Cleaning System

Vijayshankar Raman and Joseph M. Hellerstein

University of California at Berkeley
{rshankar, jmh}@cs.berkeley.edu

Wrangler: Interactive Visual Specification of Data Transformation Scripts

Sean Kandel, Andreas Paepcke, Joseph Hellerstein, Jeffrey Heer
{skandel, paepcke, jheer}@cs.stanford.edu; joeh@cs.berkeley.edu

Visualizing the Scripts of Data Wrangling with SOMNUS

Kai Xiong, Siwei Fu, Guoming Ding, Zhongsu Luo, Rong Yu, Wei Chen, Hujun Bao, Yingcai Wu

Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment

Sean Kandel*, Ravi Pa
*Sta
{skandel, rparikh

Lux: Always-on Visualization Recommendations for Exploratory Dataframe Workflows

ark, Caitlyn Chen, Jake Kang,
rst, Aditya G. Parameswaran

htyong,hearst,adityagp}@berkeley.edu

Buckaroo adds to this line of work by bringing the wrangling process into the visualization for subgroup anomaly detection and wrangling

Group-level anomalies

Country	Income	Degree
Bhutan	100,000	MS
Bhutan	12k	MS
Bhutan	NaN	PhD
Bhutan	0	BS
Bhutan	0	BS
Lesotho	NaN	BS
Lesotho	NaN	BS
Lesotho	13,000	BS

G1

G2

(Data sample)

Built in error types

Missing values	Type mismatch
Outliers	No anomalies

User defined types

We have two groups:

G1 = Country = “Bhutan”
G2 = Education = “BS”

Both groups have anomalies such as outliers, missing values and type mismatch

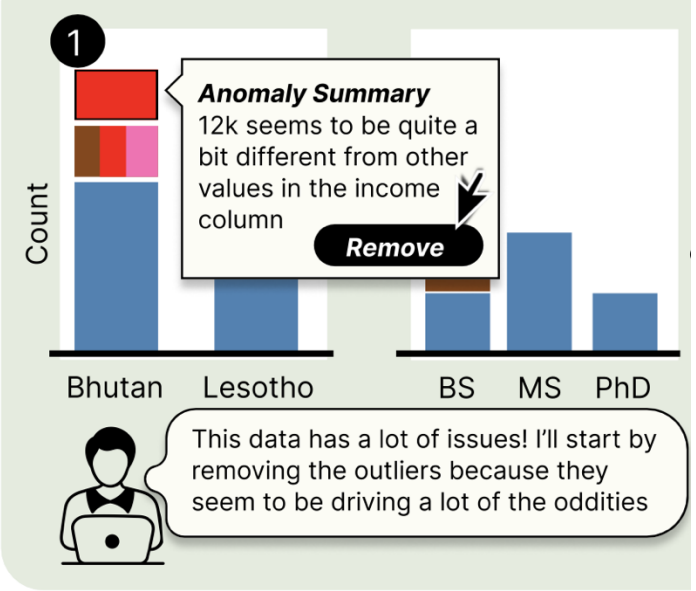
Group-level anomalies

Country	Income	Degree
Bhutan	100,000	MS
Bhutan	12k	MS
Bhutan	NaN	PhD
Bhutan	0	BS
Bhutan	0	BS
Lesotho	NaN	BS
Lesotho	NaN	BS
Lesotho	13,000	BS

G1

G2

(Data sample)



Built in error types

Missing values	Type mismatch
Outliers	No anomalies
User defined types	

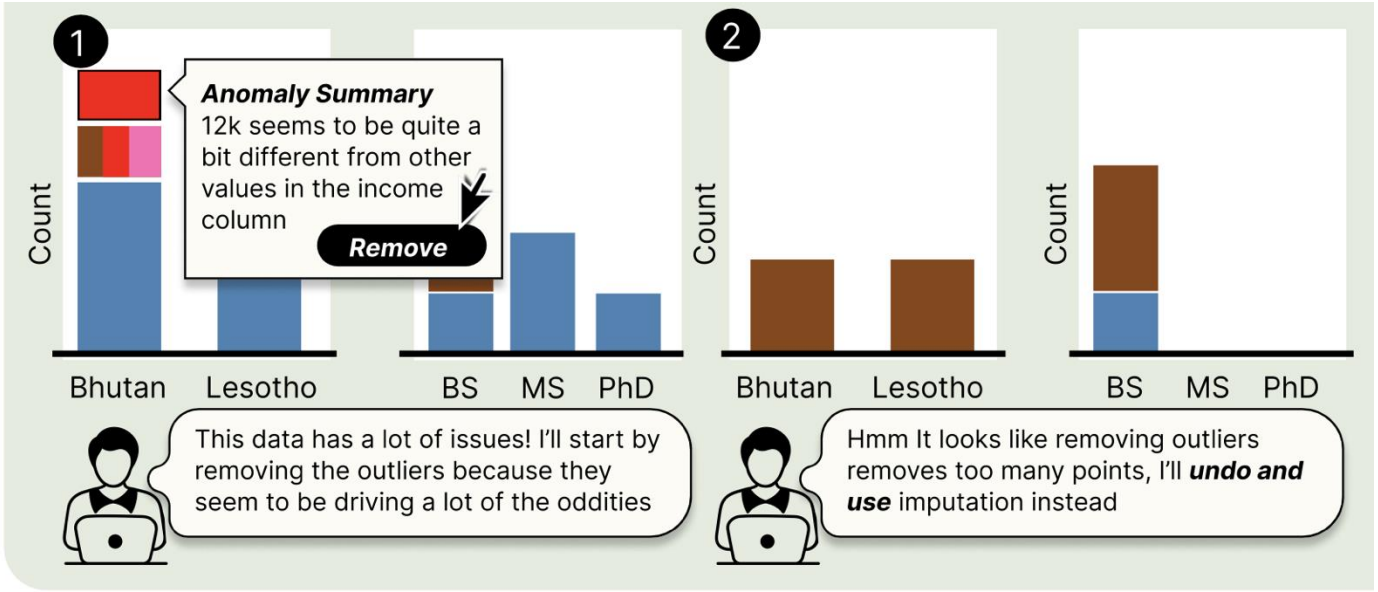
Group-level anomalies

Country	Income	Degree
Bhutan	100,000	MS
Bhutan	12k	MS
Bhutan	NaN	PhD
Bhutan	0	BS
Bhutan	0	BS
Lesotho	NaN	BS
Lesotho	NaN	BS
Lesotho	13,000	BS

G1

G2

(Data sample)



Built in error types

Missing values	Type mismatch
Outliers	No anomalies
User defined types	

Group-level anomalies

Country	Income	Degree
Bhutan	100,000	MS
Bhutan	12k	MS
Bhutan	NaN	PhD
Bhutan	0	BS
Bhutan	0	BS
Lesotho	NaN	BS
Lesotho	NaN	BS
Lesotho	13,000	BS

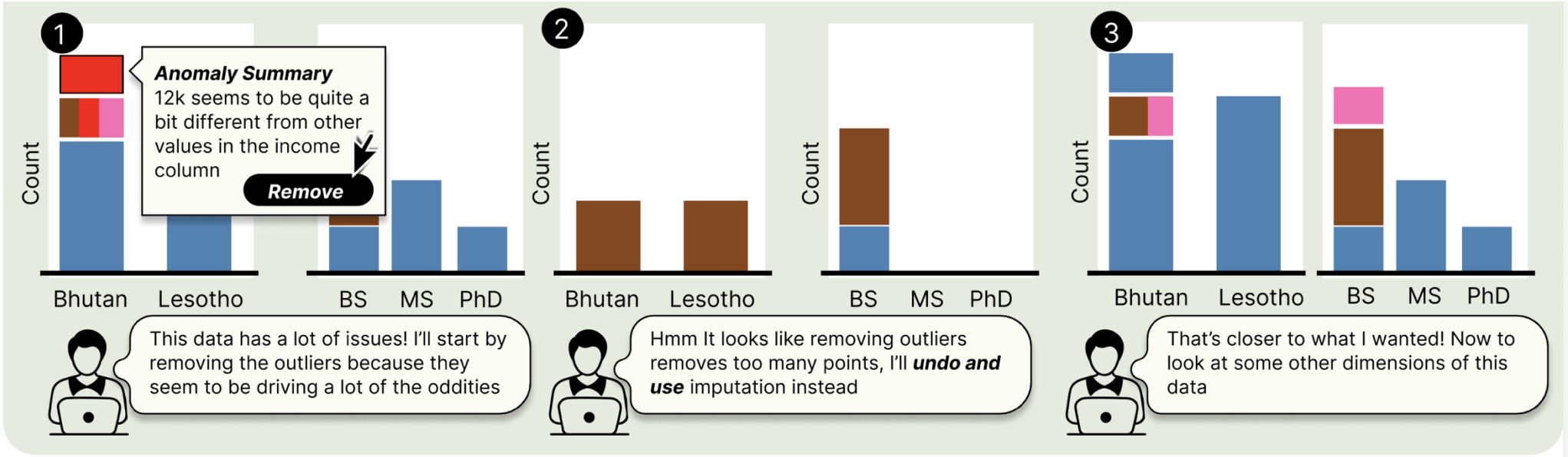
G1

G2

(Data sample)

Built in error types

Missing values	Type mismatch
Outliers	No anomalies
User defined types	



What if visualizations were the interface for repair?

- Treat charts as **active manipulation surfaces.** Let users:
  - See errors
  - Interact with them
  - Fix them directly

What if visualizations were the interface for repair?

- Treat charts as **active manipulation surfaces.** Let users:
  - See errors
  - Interact with them
  - Fix them directly

- Preserve:
  - Undo/redo
  - Provenance
  - Script generation

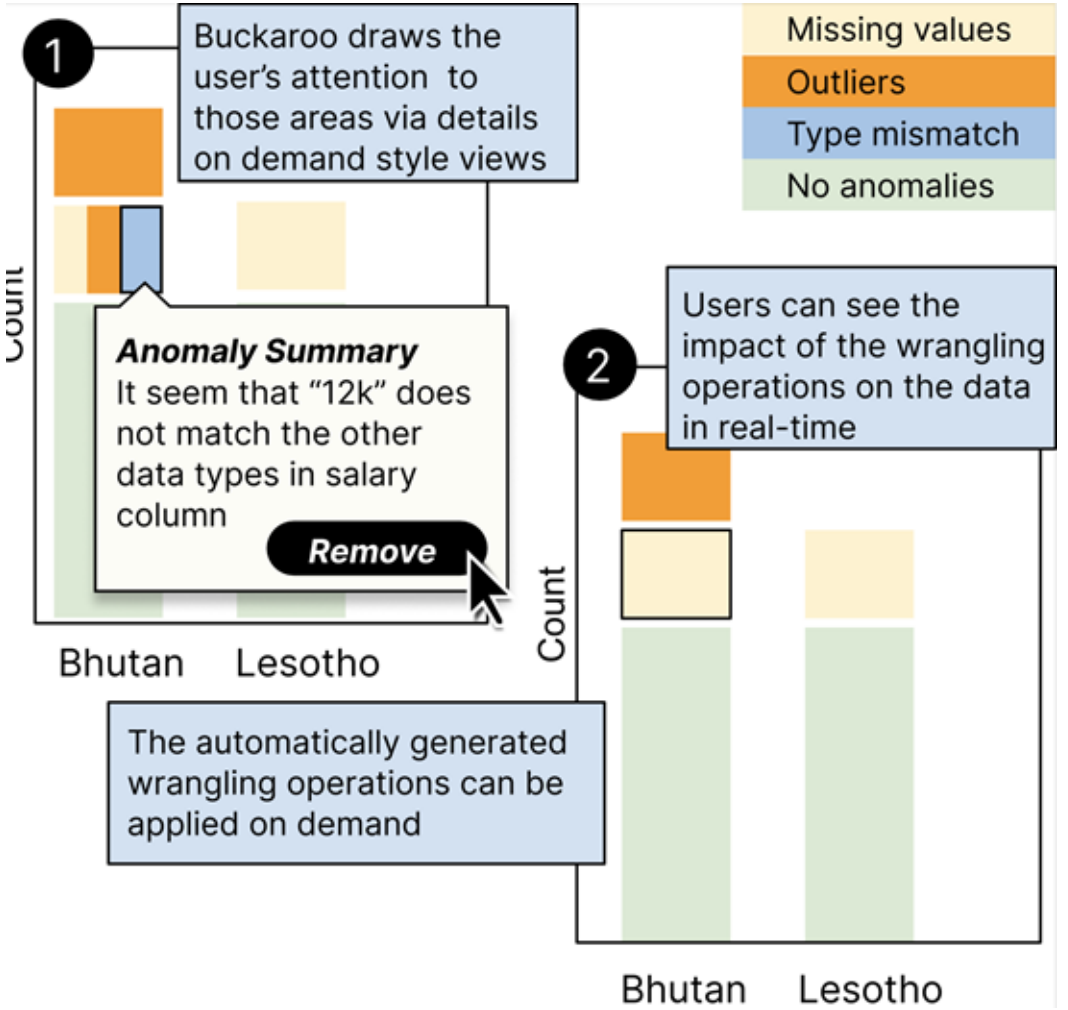
Buckaroo reframes wrangling as visual direct manipulation

Buckaroo is a system that satisfies three requirements

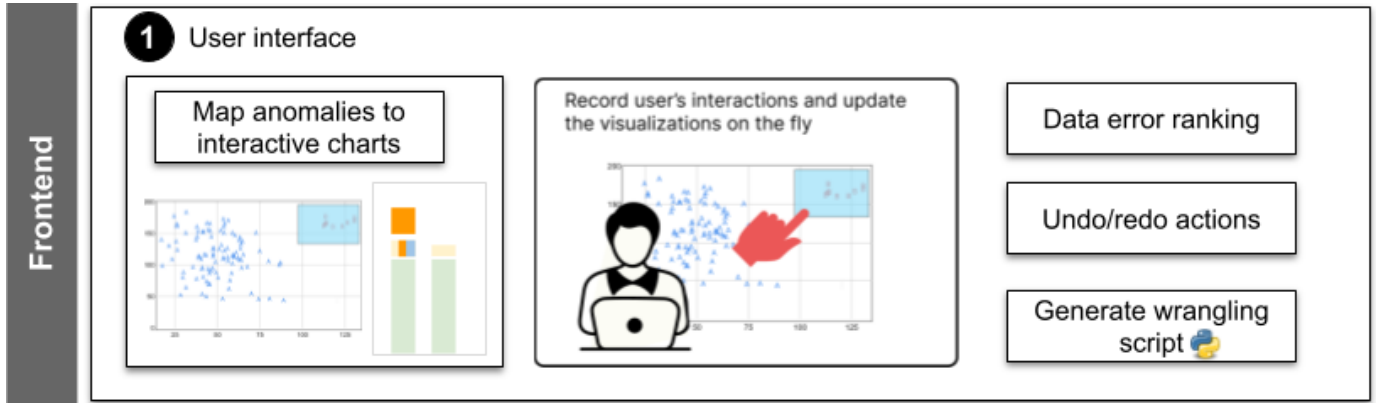
(1) It generates interactive charts to highlight data errors

(2) It enables users to make on-the-spot error corrections by directly manipulating the visual objects

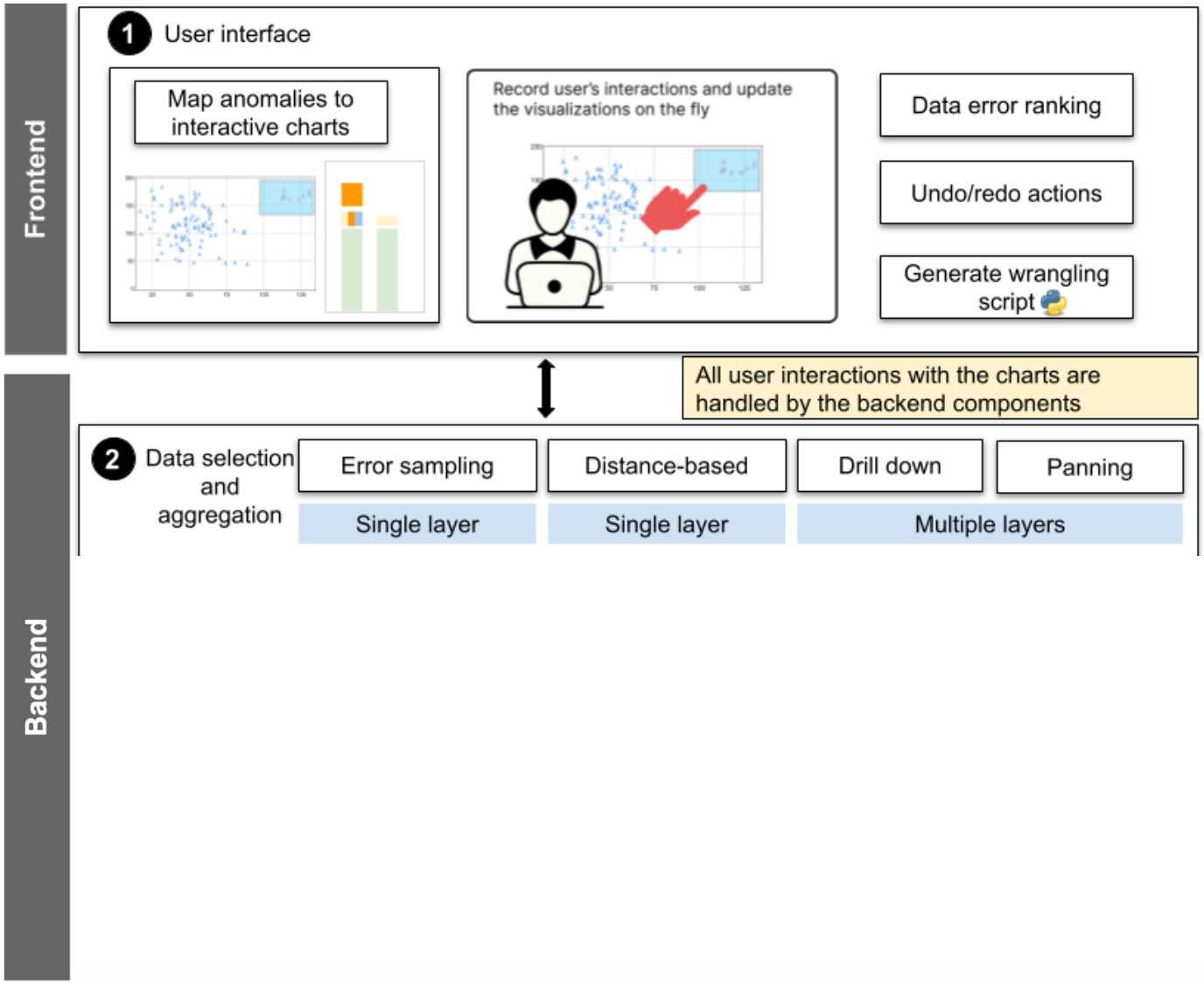
(3) It is flexible to allow user-defined error detection and correction through an API



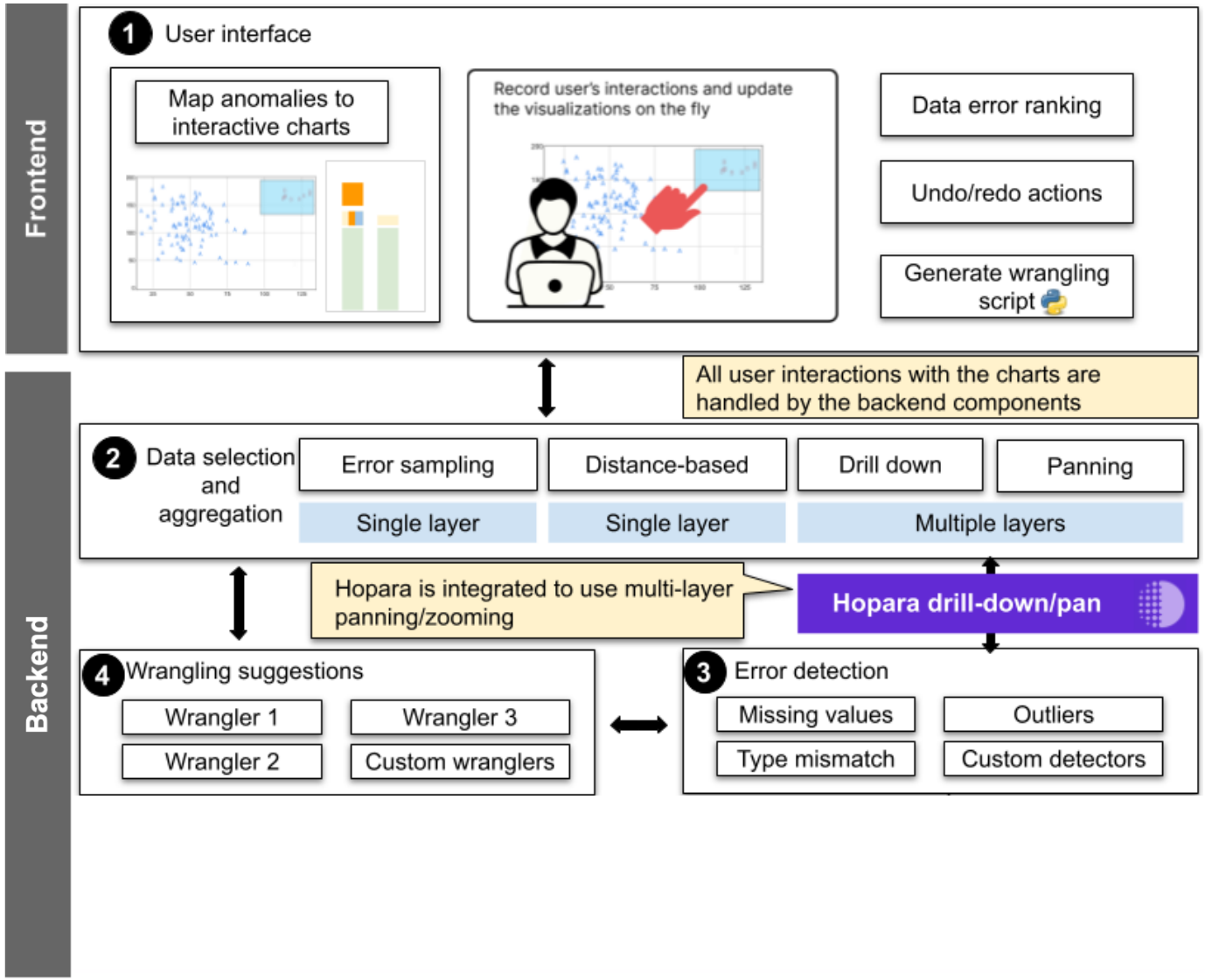
Buckaroo Architecture



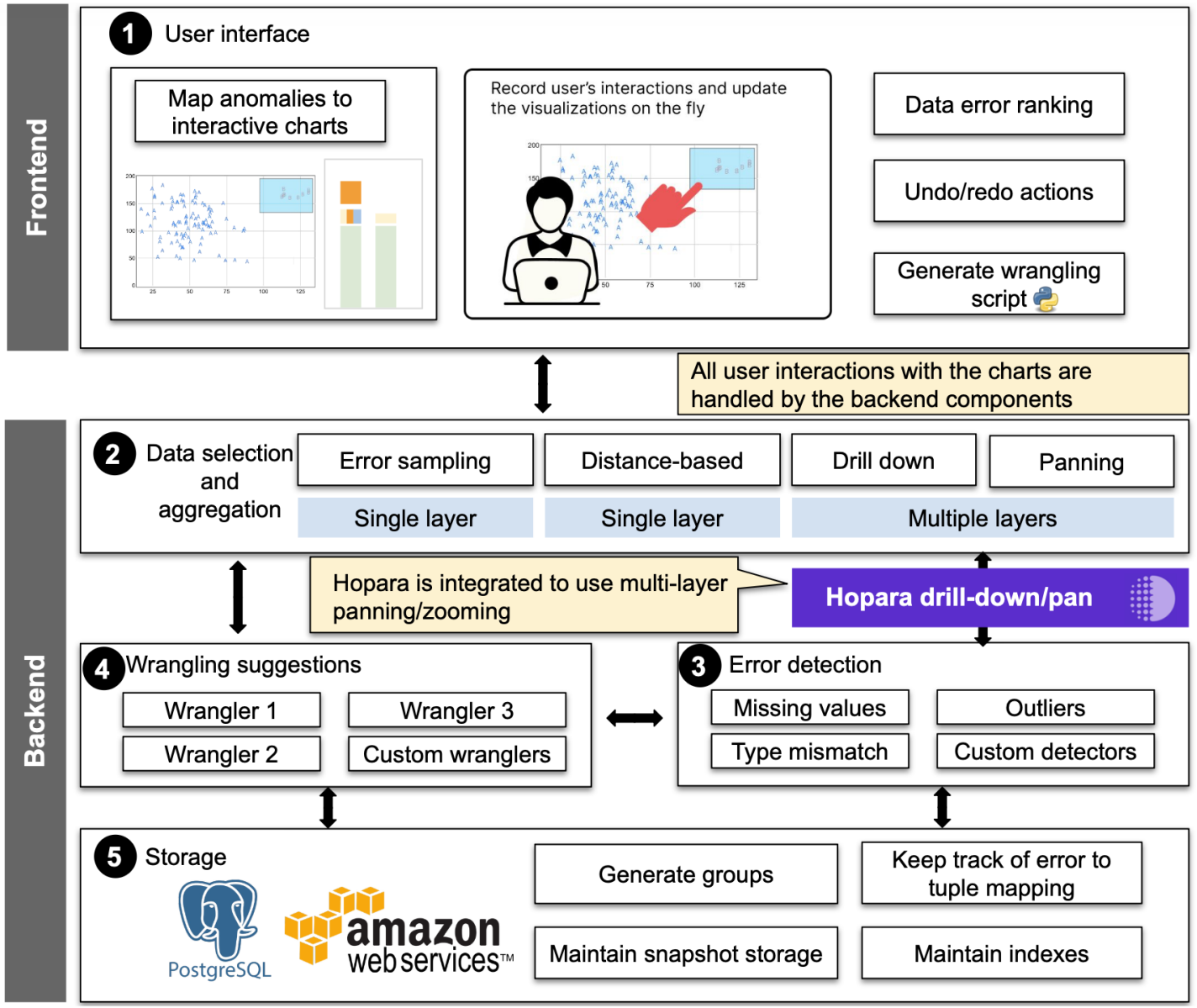
Buckaroo Architecture



Buckaroo Architecture



Buckaroo Architecture



Error Detection

Built-in detectors

- Missing values
- Outliers
- Type mismatches
- Small/incomplete groups

Error Detection

Built-in detectors

- Missing values
- Outliers
- Type mismatches
- Small/incomplete groups

Custom detectors

- User-defined logic
- SQL or Pandas
- Group-level semantics

How to do direct manipulation data wrangling at scale?

Two navigation modes

- **Single-layer**
  - Error-first sampling
  - Distance-based sampling

- **Multi-layer**
  - Drill-down
  - Pan & zoom (Hopara)

Errors remain visible even when rare

Hopara Integration

- Multi-layer navigation engine
- SQL-backed pan/zoom
- Only visible data is loaded
- Enables interactivity on large datasets

Preliminary Results

Dataset	Postgres (removal)	Postgres (impute)	Pandas (removal)	Pandas (impute)
StackOverflow	0.18 sec	0.16 sec	1.69 sec	1.27 sec
Adult Income	0.15 sec	0.13 sec	1.40 sec	1.17 sec
Chicago Crime	0.71 sec	0.68 sec	5.87 sec	5.29 sec

Parting thoughts

- Buckaroo is actively being developed

Try Buckaroo now!

Parting thoughts

- Buckaroo is actively being developed
- We are conducting a large user study to assess how its usability compares to other systems like Trifacta, ChatGPT, and Jupyter notebooks

Try Buckaroo now!

Parting thoughts

- Buckaroo is actively being developed
- We are conducting a large user study to assess how its usability compares to other systems like Trifacta, ChatGPT, and Jupyter notebooks
- Usability is a big problem for data prep today, which limits adoption of our systems by practitioners

Try Buckaroo now!