Outline

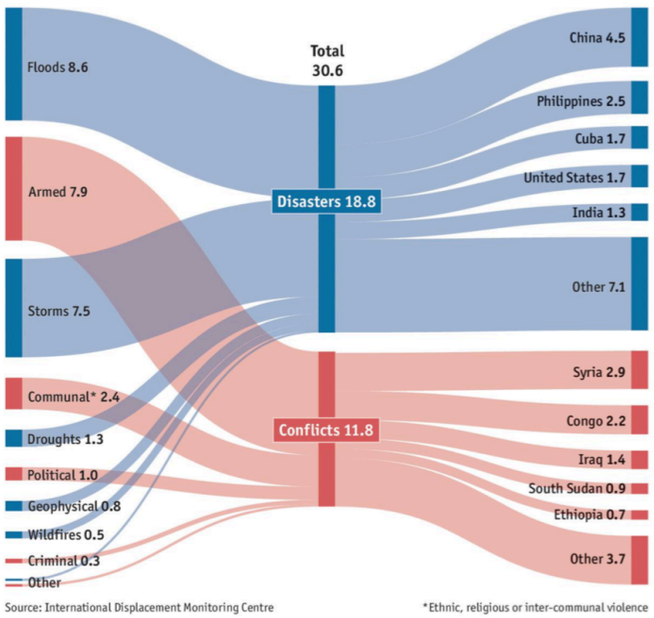
- **Nan:** Fundamentals and State-of-the-art (25-30 minutes)
- **Eugene:** Efficient, Effective and Interactive Visualizations (60-65 minutes)
- **Guoliang:** Recommendation (~60 minutes)
- **Nan:** Uncertainty, collaborative, and immersive data visualizations (~30 minutes)
  - ***uncertainty*** and "***cleaning***" bad data visualizations
  - ***collaborative*** data visualization
  - ***immersive*** data visualization

In (Data) Science, The Only Certainty is

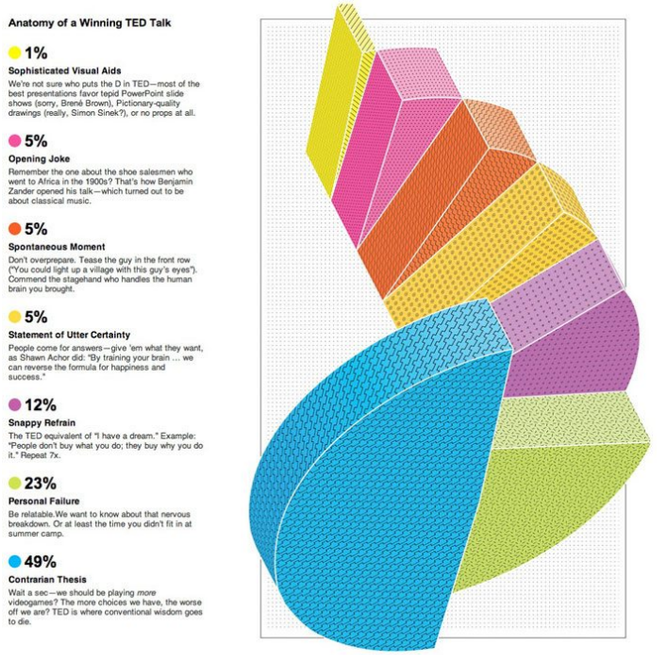
In (Data) Science, The Only Certainty is Uncertainty

In (Data) Science, The Only Certainty is Uncertainty

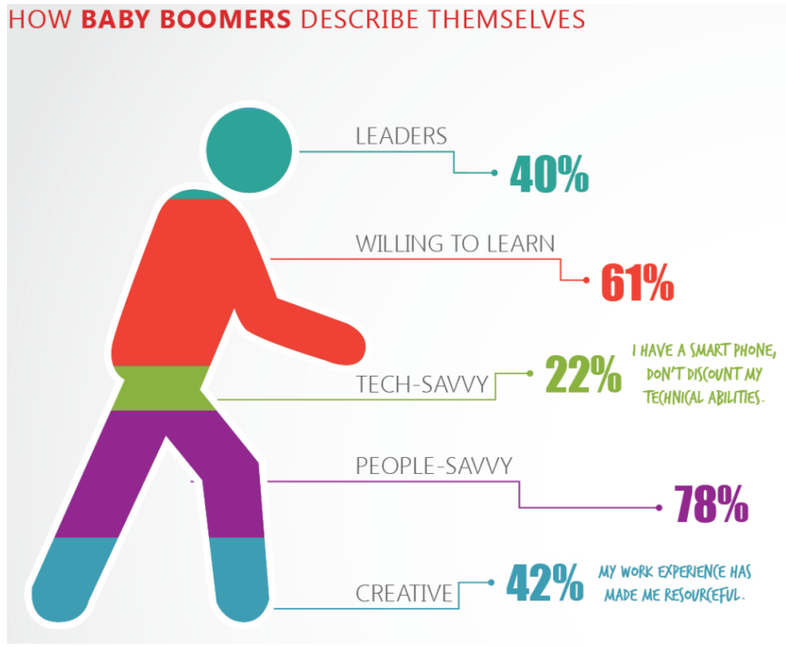
Doubt



Bad



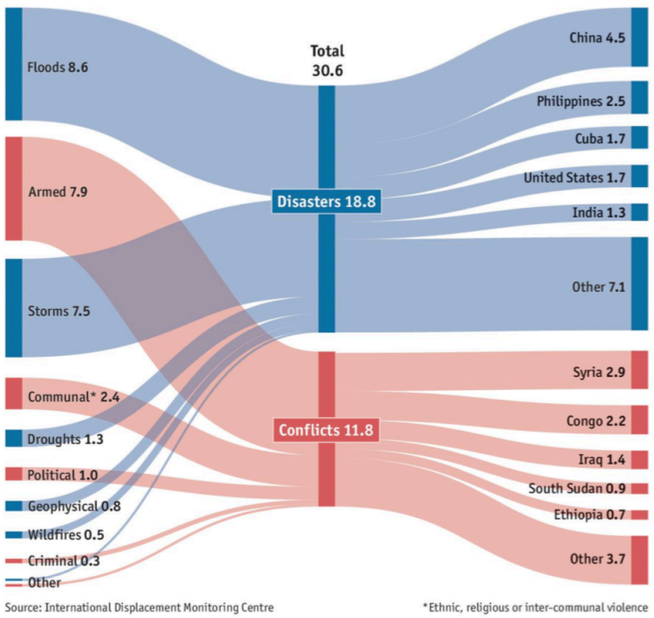
Error



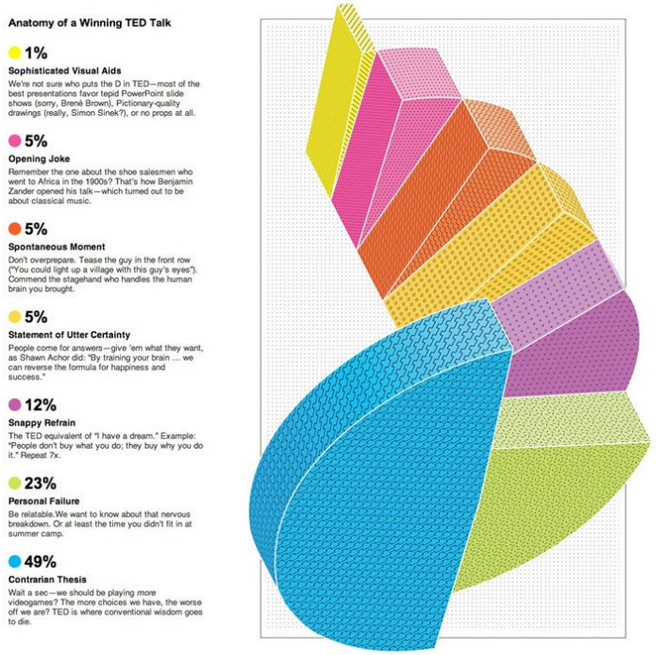
incomplete data, lack of knowledge, variability …

In (Data) Science, The Only Certainty is Uncertainty

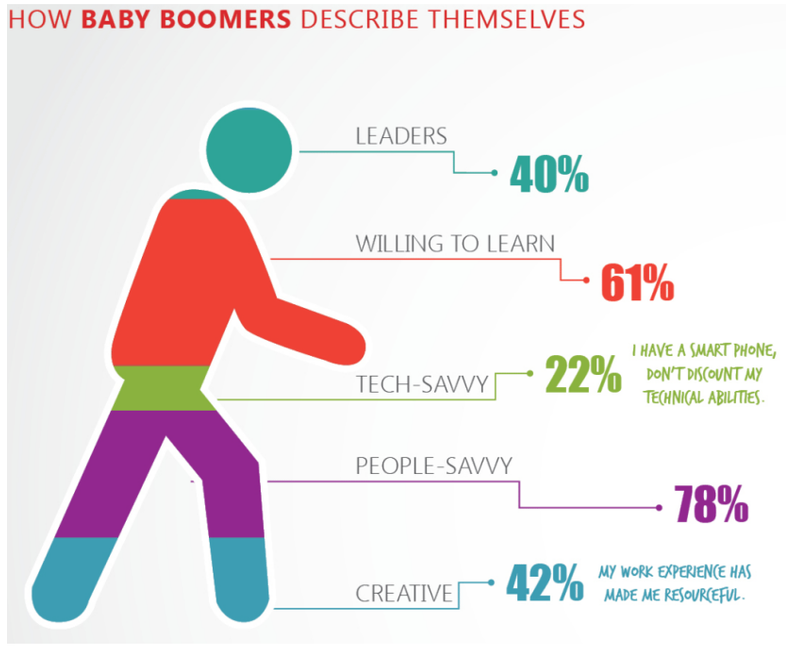
Doubt



Bad



Error



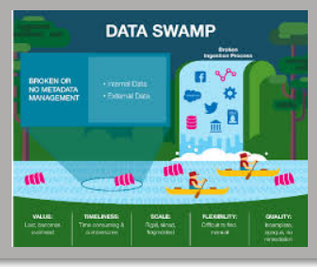
incomplete data, lack of knowledge, variability ...

- What does uncertainty mean? *- Diversified meanings.*
- How should I visualize it? *- It depends.*
- What can go wrong? *- Everything.*

Uncertainty

STEP 1
Discovery

import
discover
collect

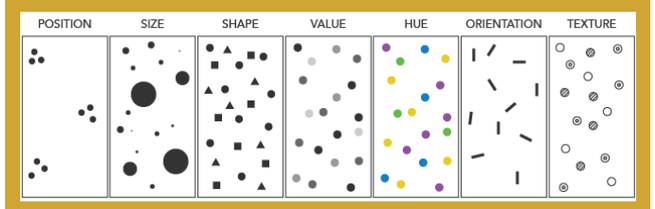


STEP 2
Curation

integration
transformation
cleaning

STEP 3
VISUAL ENCODINGS

map data
to visual variables



STEP 4
Rendering

images

Right ?
Enough ?

Clean?

Good?

Understand?
Trust?

In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation. **Jessica Hullman et al.**, InfoVis 2018.
Uncertainty. CSE 442 - Data Visualization. https://courses.cs.washington.edu/courses/cse442/17au/lectures/CSE442-Uncertainty.pdf

Uncertainty

jupyter user1 (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Run Code

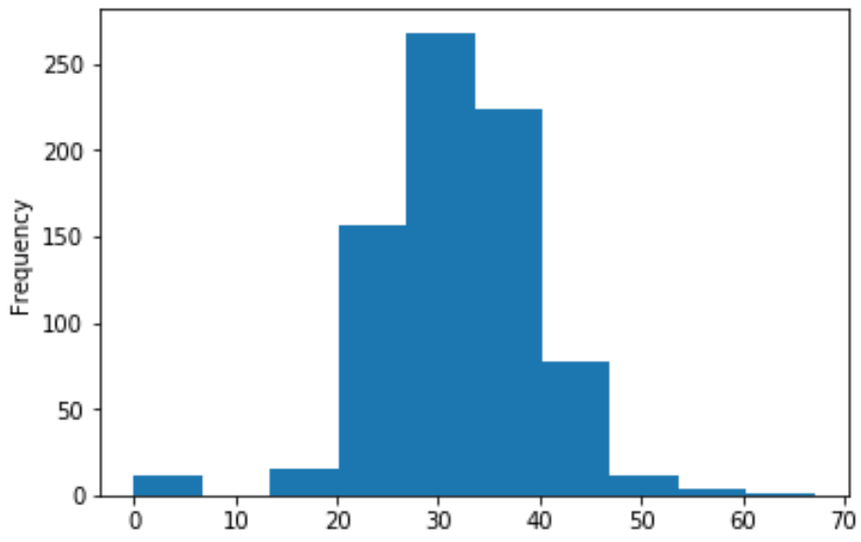
In [2]:
```
import warnings
import sys

warnings.filterwarnings('ignore')
sys.path.append('../')
```

In [ ]:

In [3]:
```
import pandas as pd

df = pd.read_csv('pima-indians-diabetes.csv')
```

In [6]:
```
df['Body mass index'].plot.hist()
```

Out[6]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x113c364e0>
```



Uncertainty

(collaborative) cleaning

(visualization-driven) cleaning

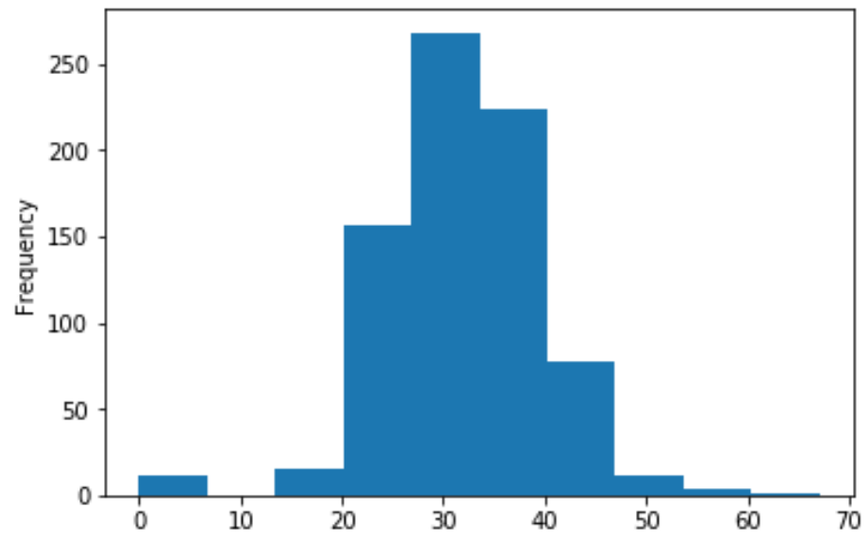
why-analysis

jupyter user1 (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Run Code

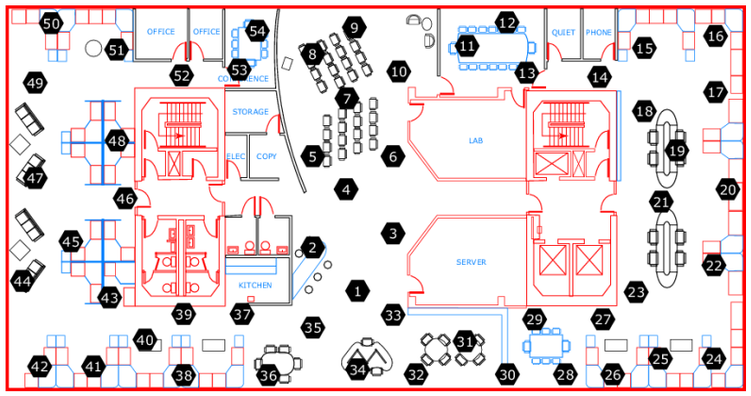
```
In [2]: import warnings
        import sys

        warnings.filterwarnings('ignore')
        sys.path.append('../')

In [ ]:

In [3]: import pandas as pd

        df = pd.read_csv('pima-indians-diabetes.csv')

In [6]: df['Body mass index'].plot.hist()
Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x113c364e0>
```

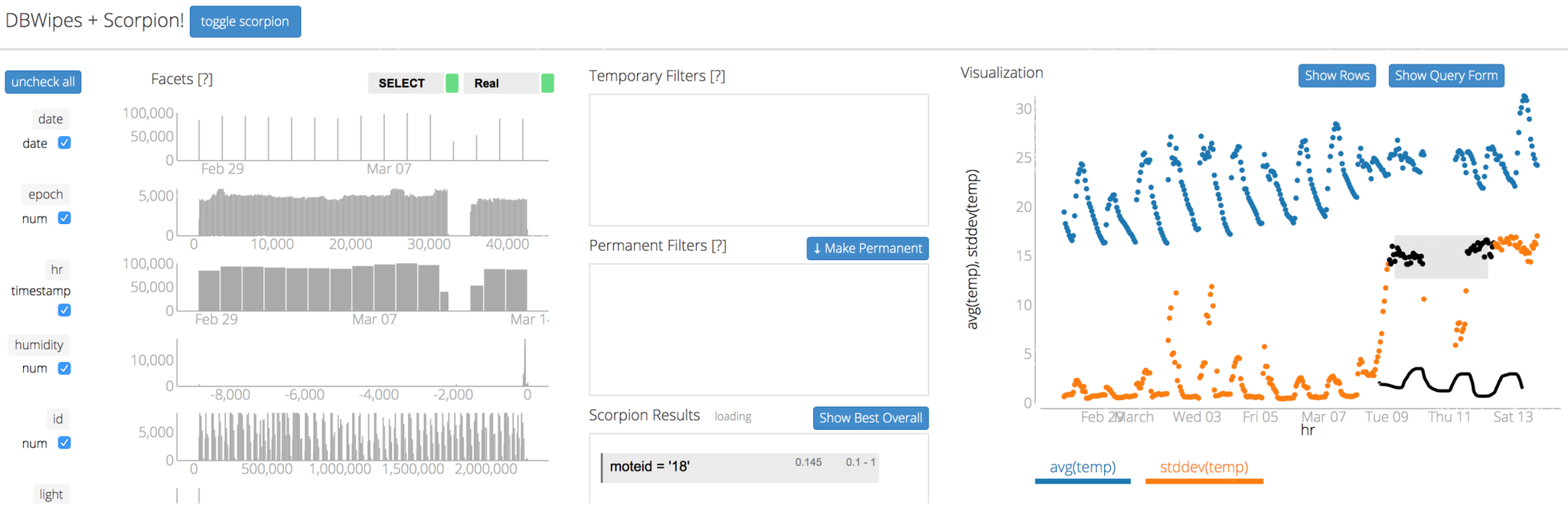


Why-analysis

sensor, light, voltage,
humidity, temperature

54 sensors
3.2k readings/hour



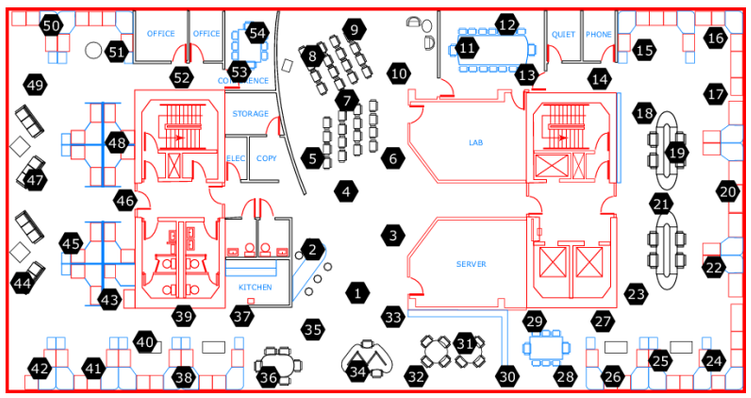


Scorpion: Explaining Away Outliers in Aggregate Queries, **Eugene Wu et al.**, VLDB 2013
A Demonstration of DBWipes: Clean as You Query, **Eugene Wu et al.**, VLDB demo 2012

Why-analysis

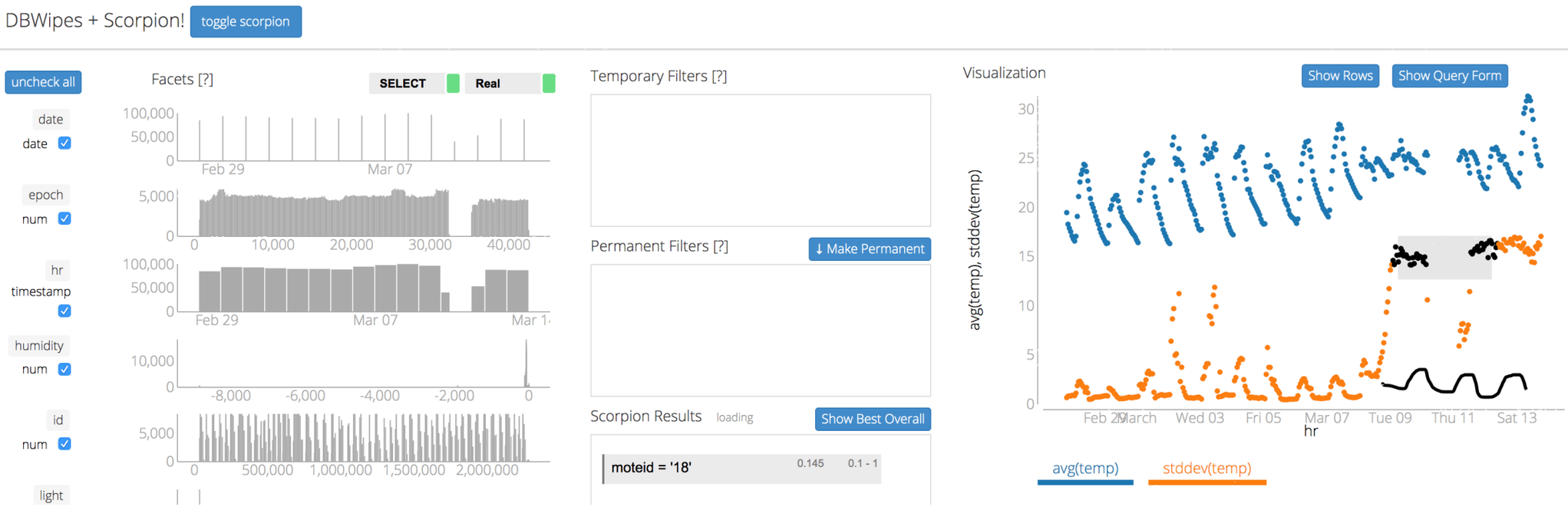
sensor, light, voltage,
humidity, temperature

54 sensors
3.2k readings/hour



Influential Predicates Problem

Given a select-project-group-by query, user inputs ***O***, hold-out results ***H***, it is find the predicate ***p*** that maximizes the influence on ***O***, and minimizes the influence on ***H***.

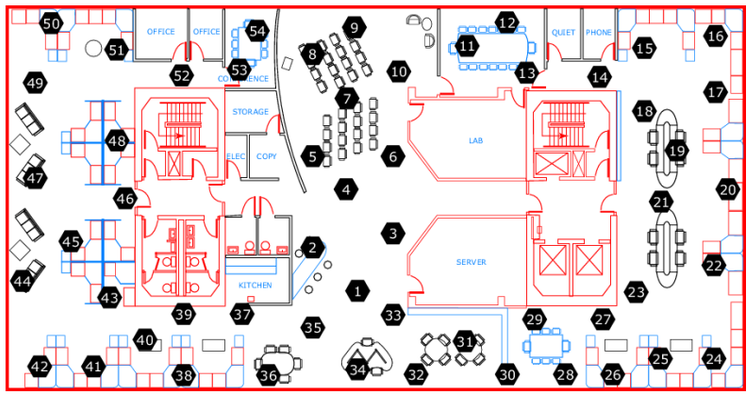


Scorpion: Explaining Away Outliers in Aggregate Queries, **Eugene Wu et al.**, VLDB 2013
A Demonstration of DBWipes: Clean as You Query, **Eugene Wu et al.**, VLDB demo 2012

Why-analysis

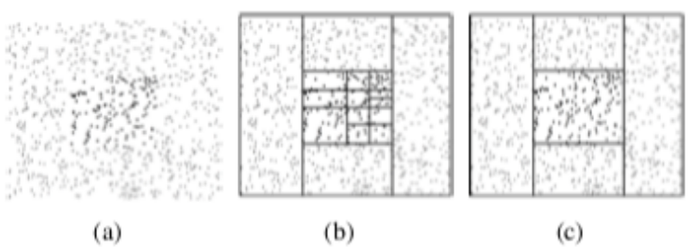
sensor, light, voltage,
humidity, temperature

54 sensors
3.2k readings/hour

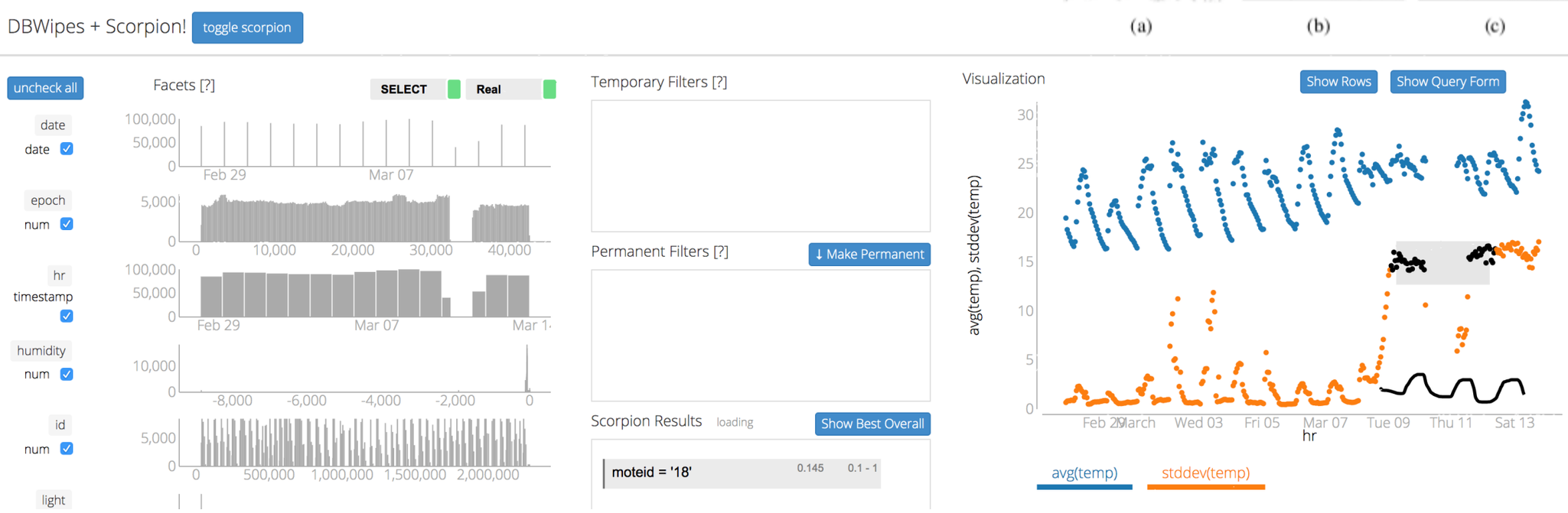


Influential Predicates Problem

Given a select-project-group-by query, user inputs ***O***, hold-out results ***H***, it is find the predicate ***p*** that maximizes the influence on ***O***, and minimizes the influence on ***H***.



Provenance
Partitioner
Merger
Scorer

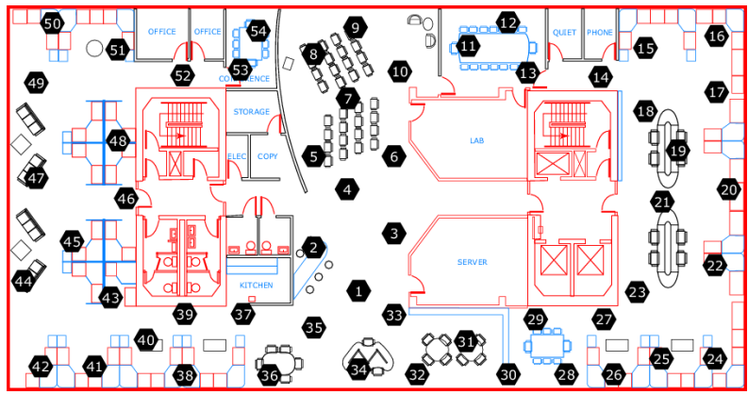


Scorpion: Explaining Away Outliers in Aggregate Queries, **Eugene Wu et al.**, VLDB 2013
A Demonstration of DBWipes: Clean as You Query, **Eugene Wu et al.**, VLDB demo 2012

Why-analysis

sensor, light, voltage, humidity, temperature

54 sensors
3.2k readings/hour



"sensors with low voltage"

Uncertainty

(collaborative) cleaning

(visualization-driven) cleaning

why-analysis

jupyter user1 (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Run Code

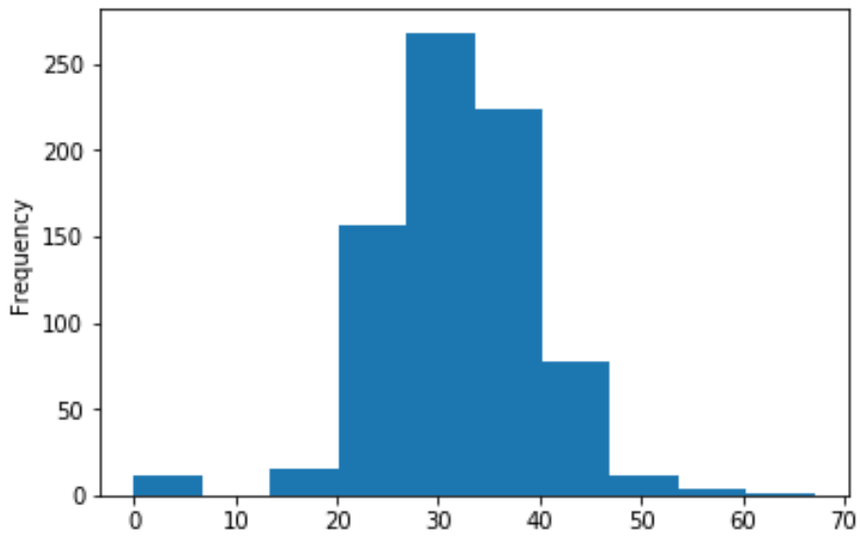
In [2]:
```
import warnings
import sys

warnings.filterwarnings('ignore')
sys.path.append('../')
```

In [ ]:

In [3]:
```
import pandas as pd

df = pd.read_csv('pima-indians-diabetes.csv')
```

In [6]:
```
df['Body mass index'].plot.hist()
```

Out[6]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x113c364e0>
```



Visualization-driven Cleaning

Id	Year	Title (abbr.)	Venue	Affiliation	Citations
t_1	2013	NADEEF	ACM SIGMOD	QCRI	174.0
t_2	2013	NADEEF	SIGMOD Conf.	QCRI, HBKU	1740
t_3	2013	NADEEF	SIGMOD	QCRI HBKU	174.0
t_4	2013	KuaFu	ICDE 2013	Microsoft	15.0
t_5	2013	TsingNUS	SIGMOD'13	Tsinghua	13.0
t_6	2013	TsingNUS	SIGMOD'13	THU	13.0
t_7	2014	SeeDB	VLDB	Stanford Univ.	N.A.
t_8	2014	SeeDB	Very Large Data Bases	Stanford	55.0
t_9	2015	Elaps	ICDE	NUS	42.0
t_{10}	2015	Elaps	IEEE ICDE Conf. 2015	CS@NUS	44.0

TABLE I
AN EXCERPT OF PUBLICATIONS (DIRTY)

Id	Year	Title (abbr.)	Venue	Affiliation	Citations
t_{123}	2013	NADEEF	SIGMOD	QCRI	***174.0***
t_4	2013	KuaFu	ICDE	Microsoft	15.0
t_{56}	2013	TsingNUS	SIGMOD	Tsinghua	13.0
t_{78}	2014	SeeDB	VLDB	Stantford Univ.	***55.0***
t_{910}	2015	Elaps	ICDE	NUS	43.0

TABLE II
AN EXCERPT OF PUBLICATIONS (GROUND TRUTH)

Visualization-driven Cleaning

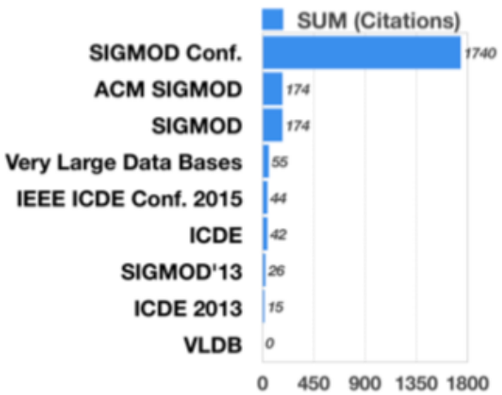
Id	Year	Title (abbr.)	Venue	Affiliation	Citations
t_1	2013	NADEEF	ACM SIGMOD	QCRI	174.0
t_2	2013	NADEEF	SIGMOD Conf.	QCRI, HBKU	1740
t_3	2013	NADEEF	SIGMOD	QCRI HBKU	174.0
t_4	2013	KuaFu	ICDE 2013	Microsoft	15.0
t_5	2013	TsingNUS	SIGMOD'13	Tsinghua	13.0
t_6	2013	TsingNUS	SIGMOD'13	THU	13.0
t_7	2014	SeeDB	VLDB	Stanford Univ.	N.A.
t_8	2014	SeeDB	Very Large Data Bases	Stanford	55.0
t_9	2015	Elaps	ICDE	NUS	42.0
t_{10}	2015	Elaps	IEEE ICDE Conf. 2015	CS@NUS	44.0

TABLE I
AN EXCERPT OF PUBLICATIONS (DIRTY)

Id	Year	Title (abbr.)	Venue	Affiliation	Citations
t_{123}	2013	NADEEF	SIGMOD	QCRI	***174.0***
t_4	2013	KuaFu	ICDE	Microsoft	15.0
t_{56}	2013	TsingNUS	SIGMOD	Tsinghua	13.0
t_{78}	2014	SeeDB	VLDB	Stantford Univ.	***55.0***
t_{910}	2015	Elaps	ICDE	NUS	43.0

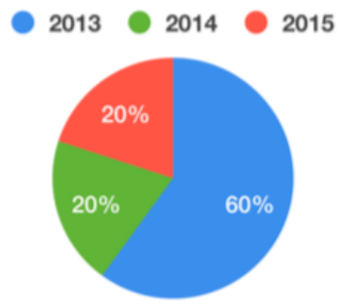
TABLE II
AN EXCERPT OF PUBLICATIONS (GROUND TRUTH)

```
VISUALIZE Bar
SELECT    Venue, SUM(Citations)
FROM      Table I
GROUP BY  Venue
```



(a) An Incorrect Bar Chart

```
VISUALIZE Pie
SELECT    Year, COUNT(Year)
FROM      Table I
GROUP BY  Year
```



(b) A Correct Pie Chart

Visualization-driven Cleaning

Id	Year	Title (abbr.)	Venue	Affiliation	Citations
t_1	2013	NADEEF	ACM SIGMOD	QCRI	174.0
t_2	2013	NADEEF	SIGMOD Conf.	QCRI, HBKU	1740
t_3	2013	NADEEF	SIGMOD	QCRI HBKU	174.0
t_4	2013	KuaFu	ICDE 2013	Microsoft	15.0
t_5	2013	TsingNUS	SIGMOD'13	Tsinghua	13.0
t_6	2013	TsingNUS	SIGMOD'13	THU	13.0
t_7	2014	SeeDB	VLDB	Stanford Univ.	N.A.
t_8	2014	SeeDB	Very Large Data Bases	Stanford	55.0
t_9	2015	Elaps	ICDE	NUS	42.0
t_{10}	2015	Elaps	IEEE ICDE Conf. 2015	CS@NUS	44.0

TABLE I
AN EXCERPT OF PUBLICATIONS (DIRTY)

Id	Year	Title (abbr.)	Venue	Affiliation	Citations
t_{123}	2013	NADEEF	SIGMOD	QCRI	***174.0***
t_4	2013	KuaFu	ICDE	Microsoft	15.0
t_{56}	2013	TsingNUS	SIGMOD	Tsinghua	13.0
t_{78}	2014	SeeDB	VLDB	Stantford Univ.	***55.0***
t_{910}	2015	Elaps	ICDE	NUS	43.0

TABLE II
AN EXCERPT OF PUBLICATIONS (GROUND TRUTH)

Data errors

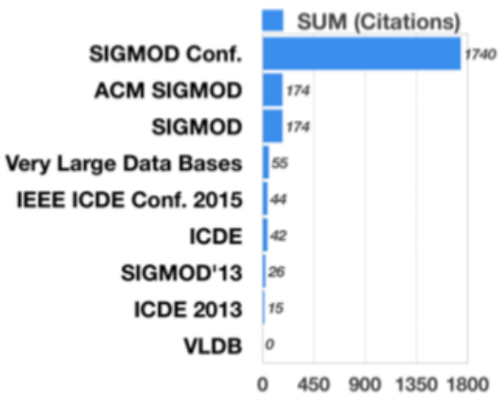
duplicates

synonyms

outliers

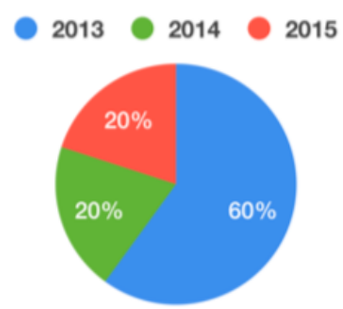
missing values

```
VISUALIZE Bar
SELECT    Venue, SUM(Citations)
FROM      Table I
GROUP BY  Venue
```



(a) An Incorrect Bar Chart

```
VISUALIZE Pie
SELECT    Year, COUNT(Year)
FROM      Table I
GROUP BY  Year
```



(b) A Correct Pie Chart

Visualization-driven Cleaning

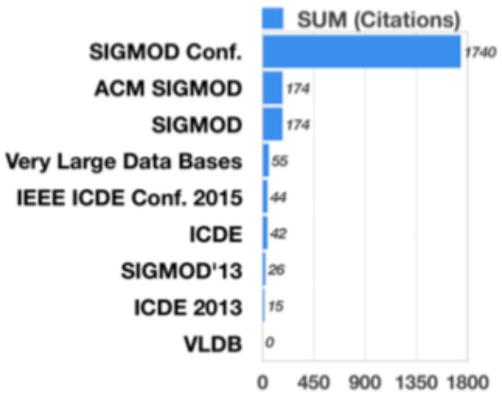
Id	Year	Title (abbr.)	Venue	Affiliation	Citations
t_1	2013	NADEEF	ACM SIGMOD	QCRI	174.0
t_2	2013	NADEEF	SIGMOD Conf.	QCRI, HBKU	1740
t_3	2013	NADEEF	SIGMOD	QCRI HBKU	174.0
t_4	2013	KuaFu	ICDE 2013	Microsoft	15.0
t_5	2013	TsingNUS	SIGMOD'13	Tsinghua	13.0
t_6	2013	TsingNUS	SIGMOD'13	THU	13.0
t_7	2014	SeeDB	VLDB	Stanford Univ.	N.A.
t_8	2014	SeeDB	Very Large Data Bases	Stanford	55.0
t_9	2015	Elaps	ICDE	NUS	42.0
t_{10}	2015	Elaps	IEEE ICDE Conf. 2015	CS@NUS	44.0

TABLE I
AN EXCERPT OF PUBLICATIONS (DIRTY)

Id	Year	Title (abbr.)	Venue	Affiliation	Citations
t_{123}	2013	NADEEF	SIGMOD	QCRI	***174.0***
t_4	2013	KuaFu	ICDE	Microsoft	15.0
t_{56}	2013	TsingNUS	SIGMOD	Tsinghua	13.0
t_{78}	2014	SeeDB	VLDB	Stantford Univ.	***55.0***
t_{910}	2015	Elaps	ICDE	NUS	43.0

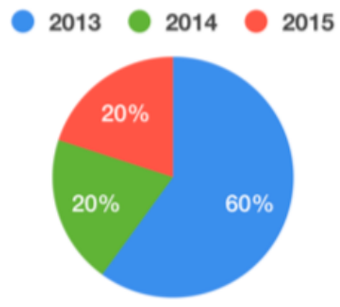
TABLE II
AN EXCERPT OF PUBLICATIONS (GROUND TRUTH)

```
VISUALIZE Bar
SELECT    Venue, SUM(Citations)
FROM      Table I
GROUP BY  Venue
```



(a) An Incorrect Bar Chart

```
VISUALIZE Pie
SELECT    Year, COUNT(Year)
FROM      Table I
GROUP BY  Year
```



(b) A Correct Pie Chart

Data errors

duplicates

synonyms

outliers

missing values

Distance between visualizations

Earth Mover Distance (EMD)

Visualization-driven Cleaning

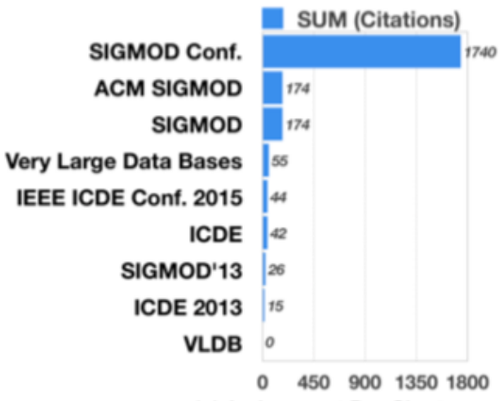
Id	Year	Title (abbr.)	Venue	Affiliation	Citations
t_1	2013	NADEEF	ACM SIGMOD	QCRI	174.0
t_2	2013	NADEEF	SIGMOD Conf.	QCRI, HBKU	1740
t_3	2013	NADEEF	SIGMOD	QCRI HBKU	174.0
t_4	2013	KuaFu	ICDE 2013	Microsoft	15.0
t_5	2013	TsingNUS	SIGMOD'13	Tsinghua	13.0
t_6	2013	TsingNUS	SIGMOD'13	THU	13.0
t_7	2014	SeeDB	VLDB	Stanford Univ.	N.A.
t_8	2014	SeeDB	Very Large Data Bases	Stanford	55.0
t_9	2015	Elaps	ICDE	NUS	42.0
t_{10}	2015	Elaps	IEEE ICDE Conf. 2015	CS@NUS	44.0

TABLE I
AN EXCERPT OF PUBLICATIONS (DIRTY)

Id	Year	Title (abbr.)	Venue	Affiliation	Citations
t_{123}	2013	NADEEF	SIGMOD	QCRI	***174.0***
t_4	2013	KuaFu	ICDE	Microsoft	15.0
t_{56}	2013	TsingNUS	SIGMOD	Tsinghua	13.0
t_{78}	2014	SeeDB	VLDB	Stantford Univ.	***55.0***
t_{910}	2015	Elaps	ICDE	NUS	43.0

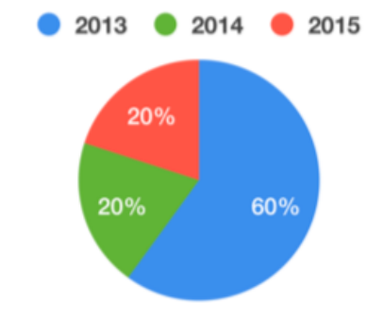
TABLE II
AN EXCERPT OF PUBLICATIONS (GROUND TRUTH)

```
VISUALIZE Bar
SELECT    Venue, SUM(Citations)
FROM      Table I
GROUP BY  Venue
```



(a) An Incorrect Bar Chart

```
VISUALIZE Pie
SELECT    Year, COUNT(Year)
FROM      Table I
GROUP BY  Year
```



(b) A Correct Pie Chart

Data errors

duplicates

synonyms

outliers

missing values

Distance between visualizations

Earth Mover Distance (EMD)

Error repair graph (ERG)

Compile heterogenous errors into one graph

Visualization-driven Cleaning

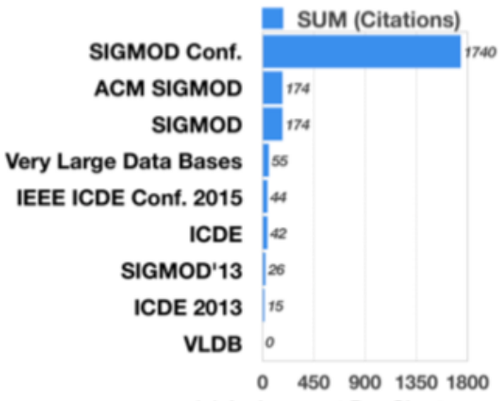
Id	Year	Title (abbr.)	Venue	Affiliation	Citations
t_1	2013	NADEEF	ACM SIGMOD	QCRI	174.0
t_2	2013	NADEEF	SIGMOD Conf.	QCRI, HBKU	1740
t_3	2013	NADEEF	SIGMOD	QCRI HBKU	174.0
t_4	2013	KuaFu	ICDE 2013	Microsoft	15.0
t_5	2013	TsingNUS	SIGMOD'13	Tsinghua	13.0
t_6	2013	TsingNUS	SIGMOD'13	THU	13.0
t_7	2014	SeeDB	VLDB	Stanford Univ.	N.A.
t_8	2014	SeeDB	Very Large Data Bases	Stanford	55.0
t_9	2015	Elaps	ICDE	NUS	42.0
t_{10}	2015	Elaps	IEEE ICDE Conf. 2015	CS@NUS	44.0

TABLE I
AN EXCERPT OF PUBLICATIONS (DIRTY)

Id	Year	Title (abbr.)	Venue	Affiliation	Citations
t_{123}	2013	NADEEF	SIGMOD	QCRI	***174.0***
t_4	2013	KuaFu	ICDE	Microsoft	15.0
t_{56}	2013	TsingNUS	SIGMOD	Tsinghua	13.0
t_{78}	2014	SeeDB	VLDB	Stantford Univ.	***55.0***
t_{910}	2015	Elaps	ICDE	NUS	43.0

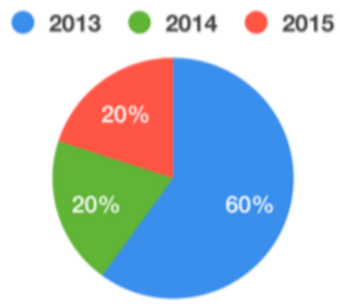
TABLE II
AN EXCERPT OF PUBLICATIONS (GROUND TRUTH)

```
VISUALIZE Bar
SELECT    Venue, SUM(Citations)
FROM      Table I
GROUP BY  Venue
```



(a) An Incorrect Bar Chart

```
VISUALIZE Pie
SELECT    Year, COUNT(Year)
FROM      Table I
GROUP BY  Year
```



(b) A Correct Pie Chart

Data errors

duplicates

synonyms

outliers

missing values

Distance between visualizations

Earth Mover Distance (EMD)

Error repair graph (ERG)

Compile heterogenous errors into one graph

Interactive cleaning for progressive visualization (ICPV)

Given a bad visualization V , it is to obtain a "cleaned" visualization V′ whose distance is far from V , under a given (small) budget w.r.t. human cost.

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Visualization-driven Cleaning

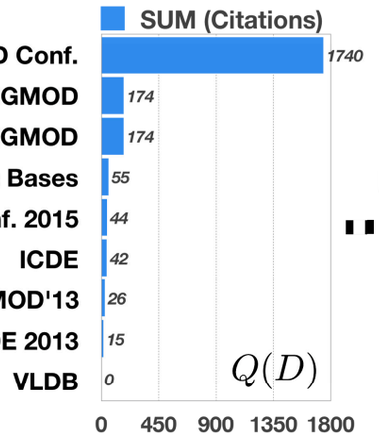
Progressive Visualization

Dirty Table D

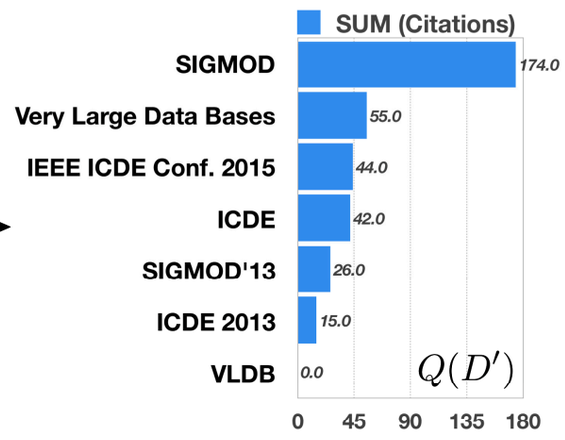
Id	Year	Title (abbr.)	Venue	Affiliation	Citations
t_1	2013	NADEEF	ACM SIGMOD	QCRI	174.0
t_2	2013	NADEEF	SIGMOD Conf.	QCRI, HBKU	1740
t_3	2013	NADEEF	SIGMOD	QCRI HBKU	174.0
t_4	2013	KuaFu	ICDE 2013	Microsoft	15.0
t_5	2013	TsingNUS	SIGMOD'13	Tsinghua	13.0
t_6	2013	TsingNUS	SIGMOD'13	THU	13.0
t_7	2014	SeeDB	VLDB	Stanford Univ.	N.A.
t_8	2014	SeeDB	Very Large Data Bases	Stanford	55.0
t_9	2015	Elaps	ICDE	NUS	42.0
t_{10}	2015	Elaps	IEEE ICDE Conf. 2015	CS@NUS	44.0

② Train

Query Q ①



Update ⑦



⑥ Repair & Update

Label Data

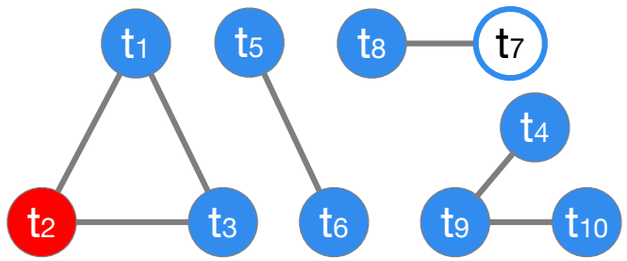
Data Cleaning Models

Errors and Repairs Candidates

Tuple-level Dup.	{t_1=t_2? t_1=t_3? …}
Attr.-level Dup.	{SIGMOD = SIGMOD'13 ? …}
Missing Value	{t_7[Citations] = 55.0? …}
Outliers	{t_2[Citations] = 174.0? …}

③

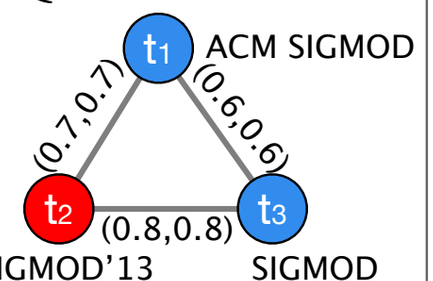
ERG Construction



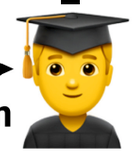
④

CQG Selection

Benefit Model

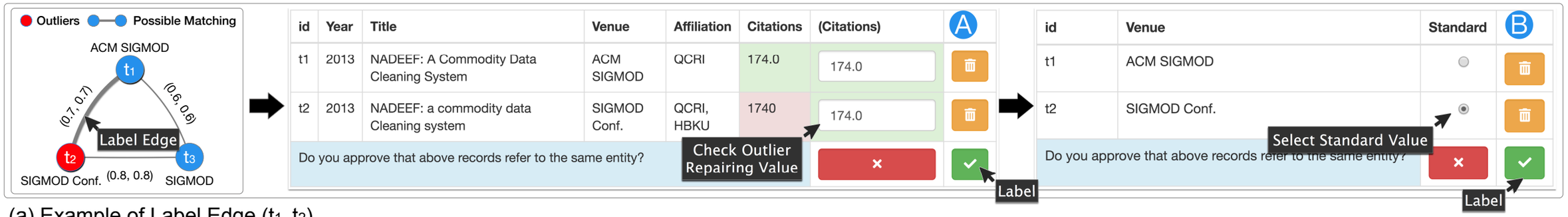


⑤ Interaction

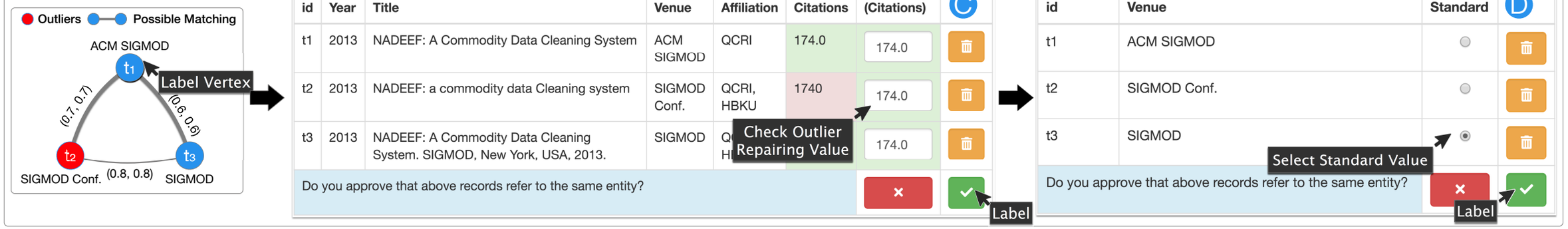


Interactive Cleaning

Visualization-driven Cleaning

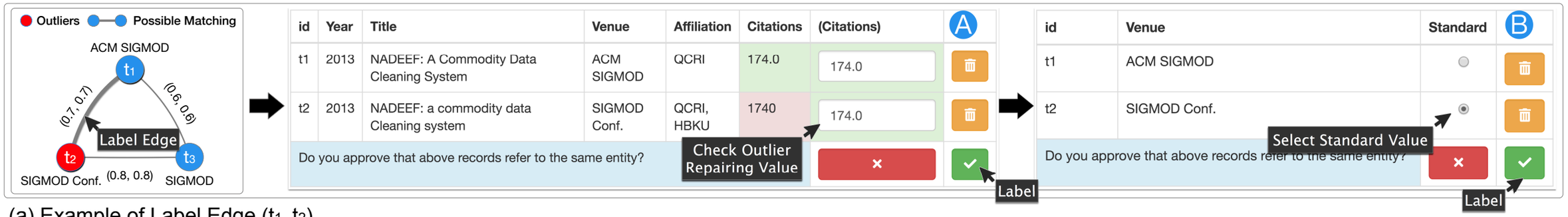


(a) Example of Label Edge (t_1, t_2)

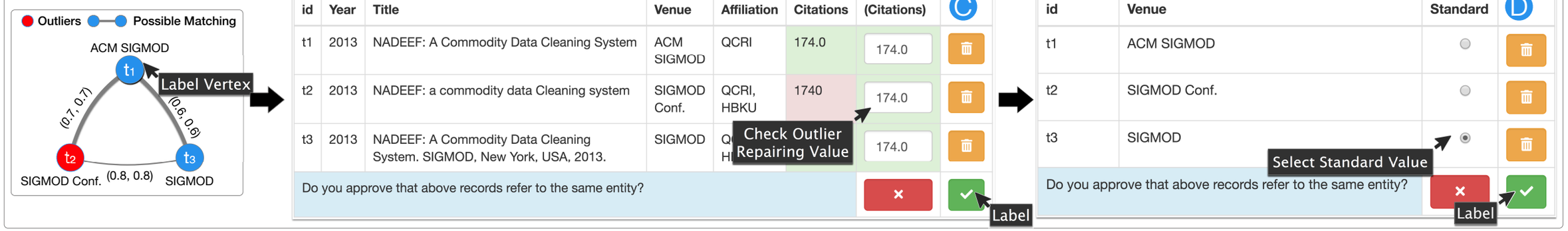


(b) Example of Label Vertex t_1

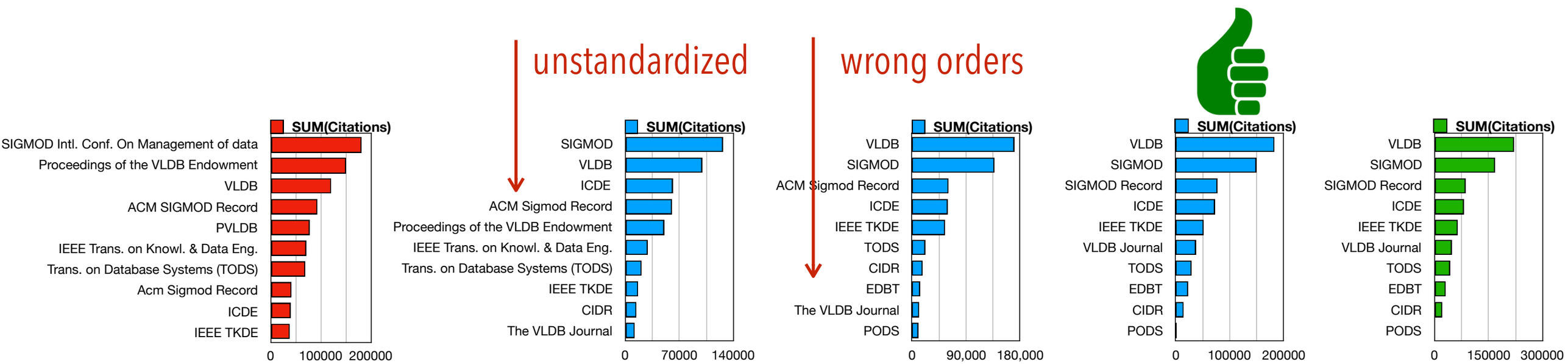
Visualization-driven Cleaning



(a) Example of Label Edge (t_1, t_2)



(b) Example of Label Vertex t_1



(a). Initial Visualization (b). After 5 Questions (c). After 10 Questions (d). After 15 Questions (e). Ground Truth

Uncertainty

(collaborative) cleaning

(visualization-driven) cleaning

why-analysis

jupyter user1 (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

Run Code

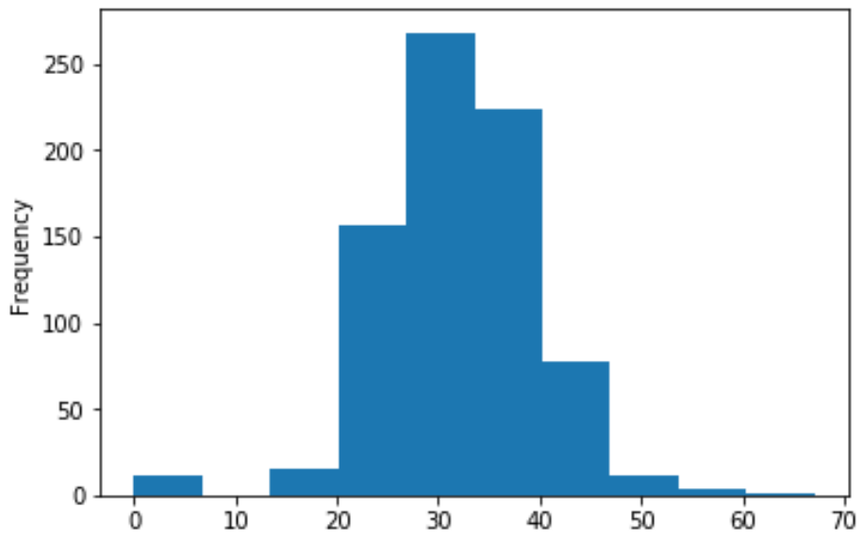
In [2]:
```
import warnings
import sys

warnings.filterwarnings('ignore')
sys.path.append('../')
```

In [ ]:

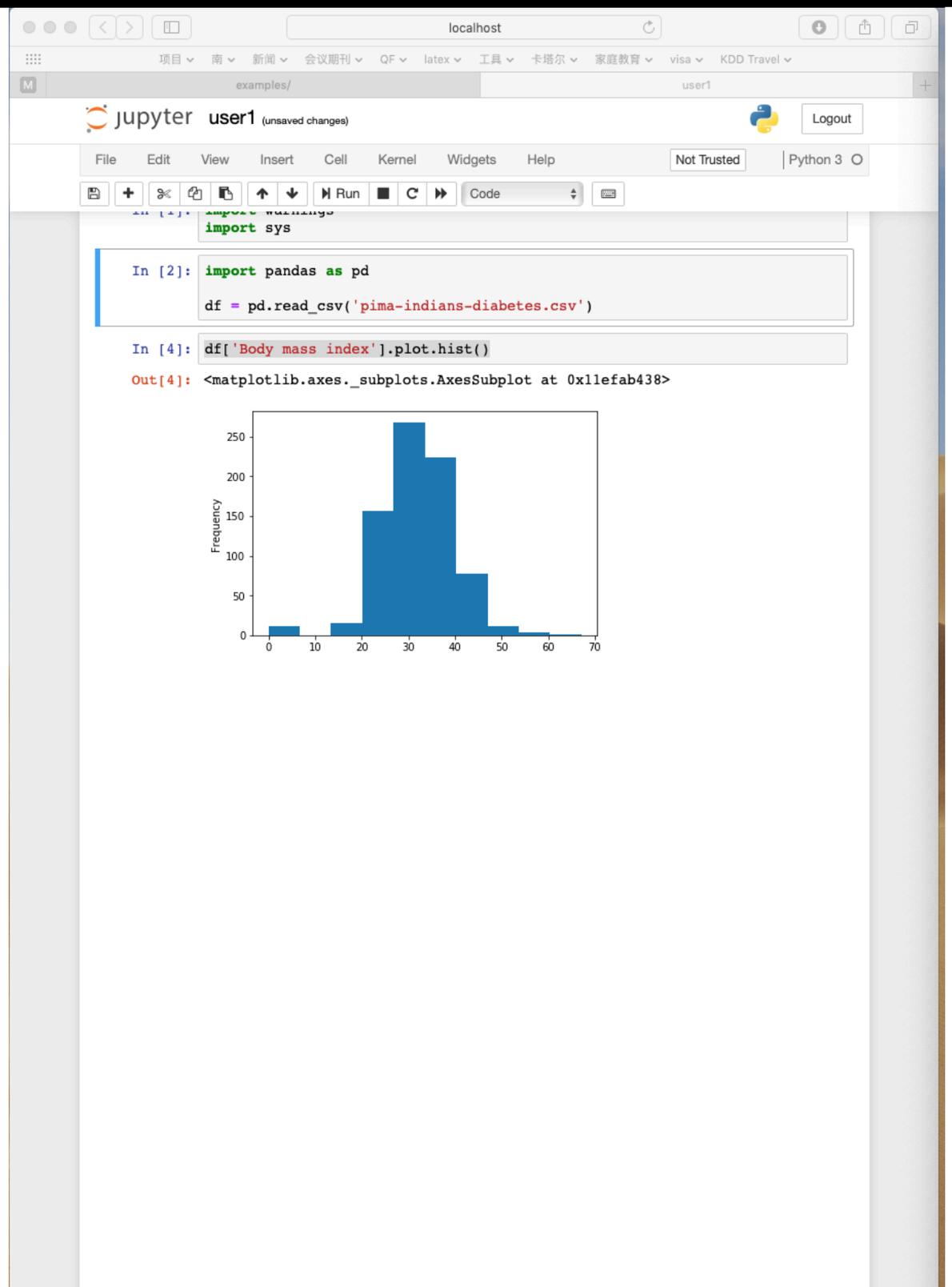
In [3]:
```
import pandas as pd

df = pd.read_csv('pima-indians-diabetes.csv')
```

In [6]:
```
df['Body mass index'].plot.hist()
```

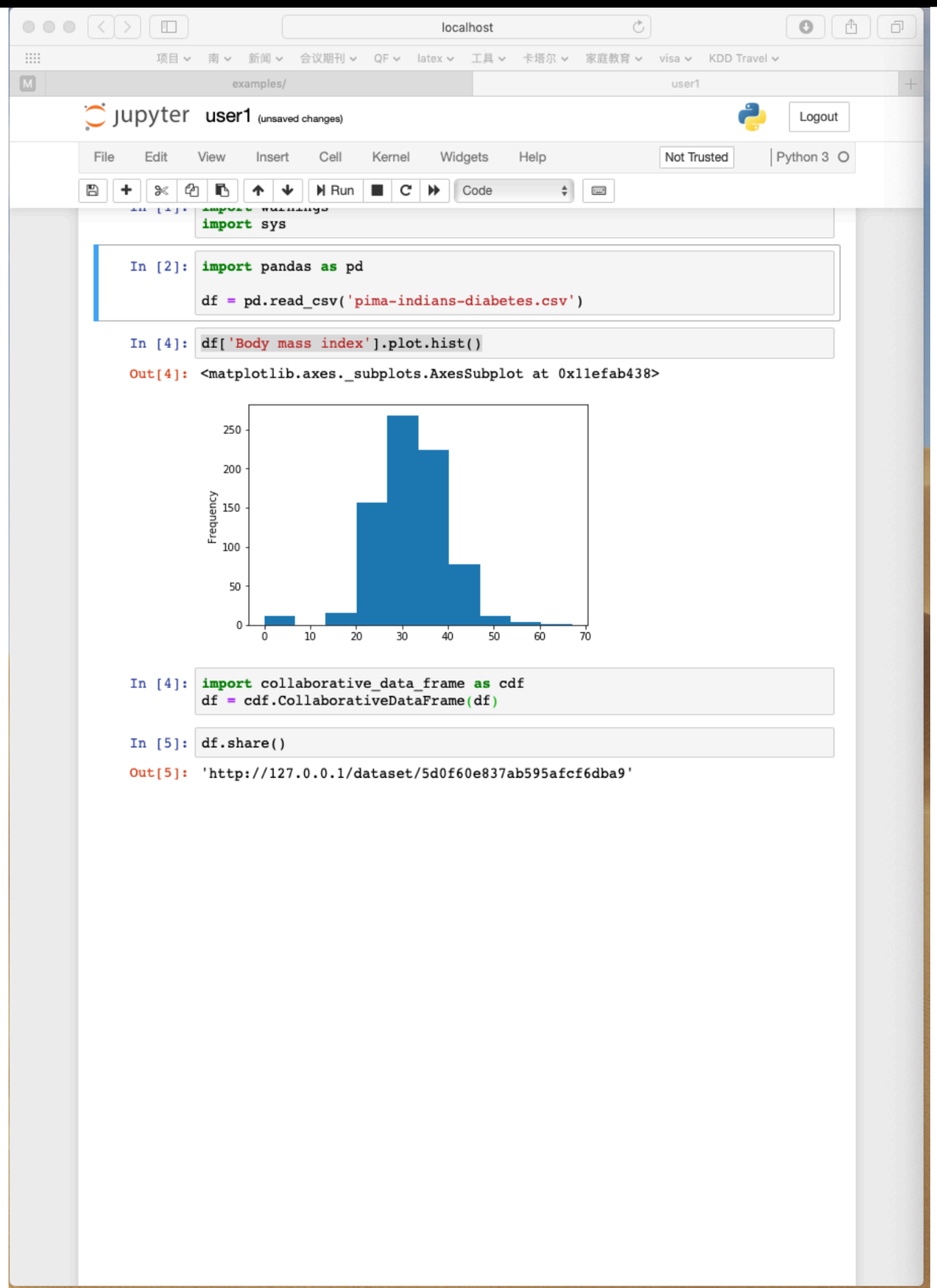
Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x113c364e0>



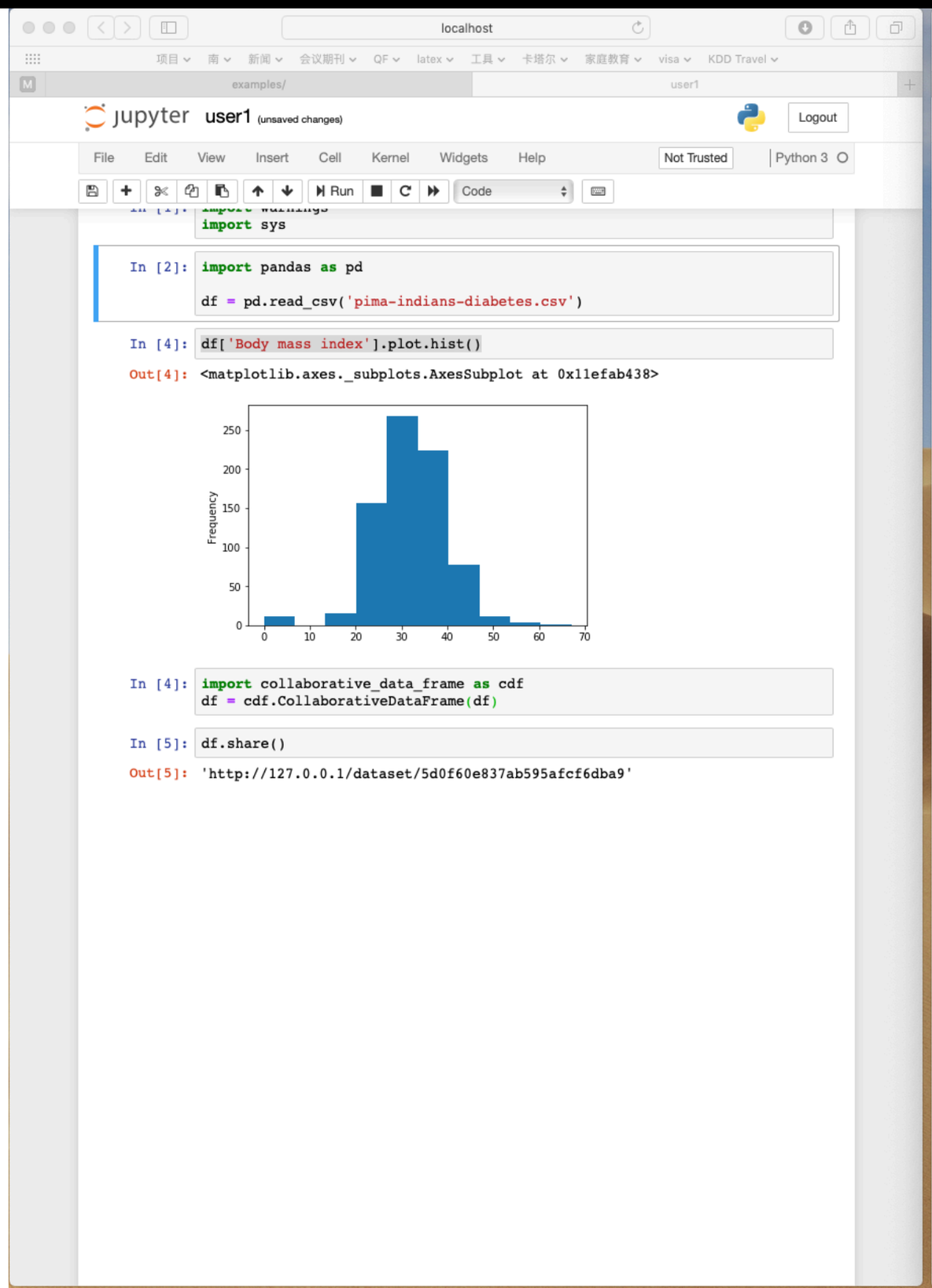
Collaborative Data Cleaning

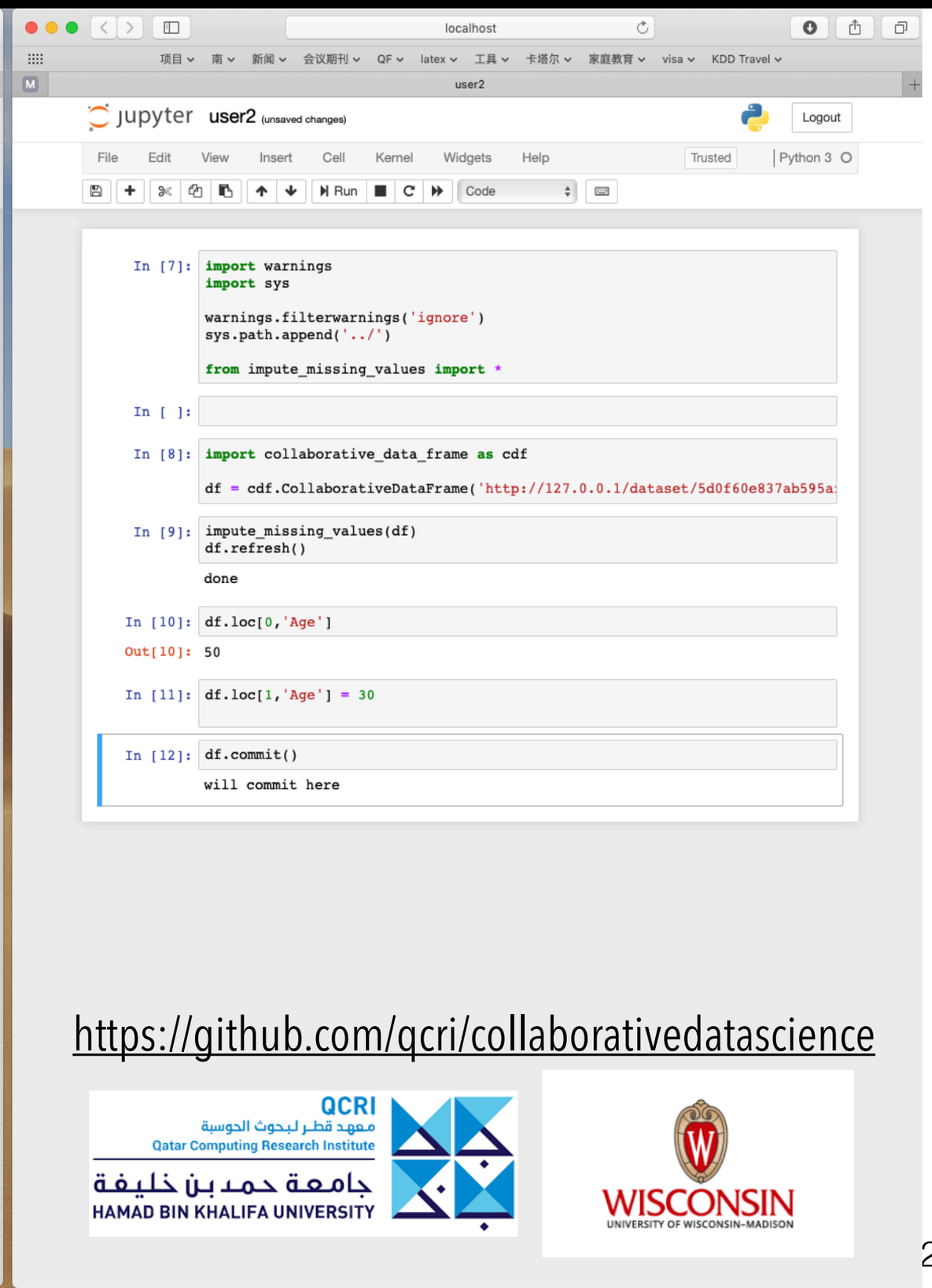


Collaborative Data Cleaning



Collaborative Data Cleaning



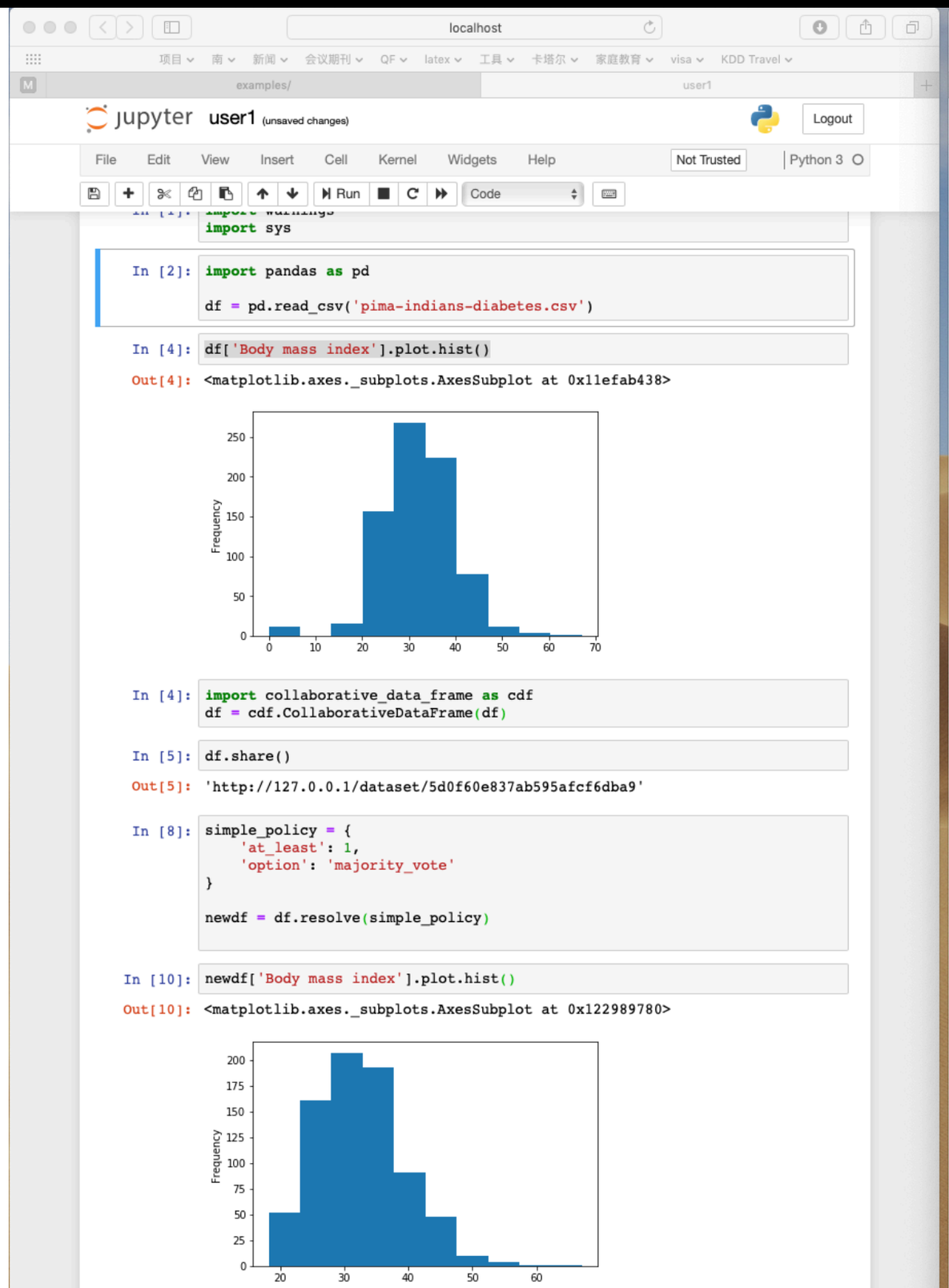


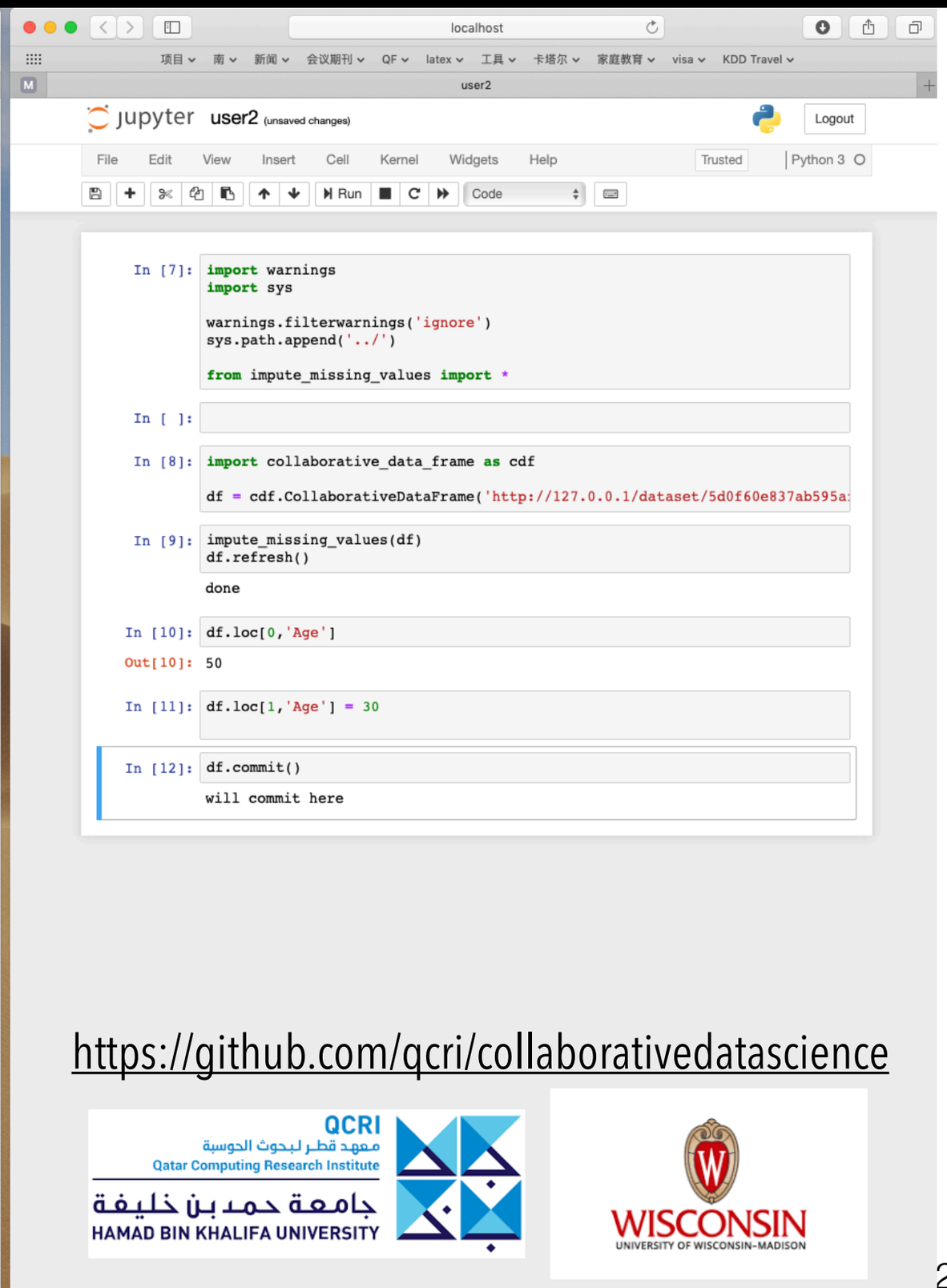
https://github.com/qcri/collaborativedatascience

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

Collaborative Data Cleaning



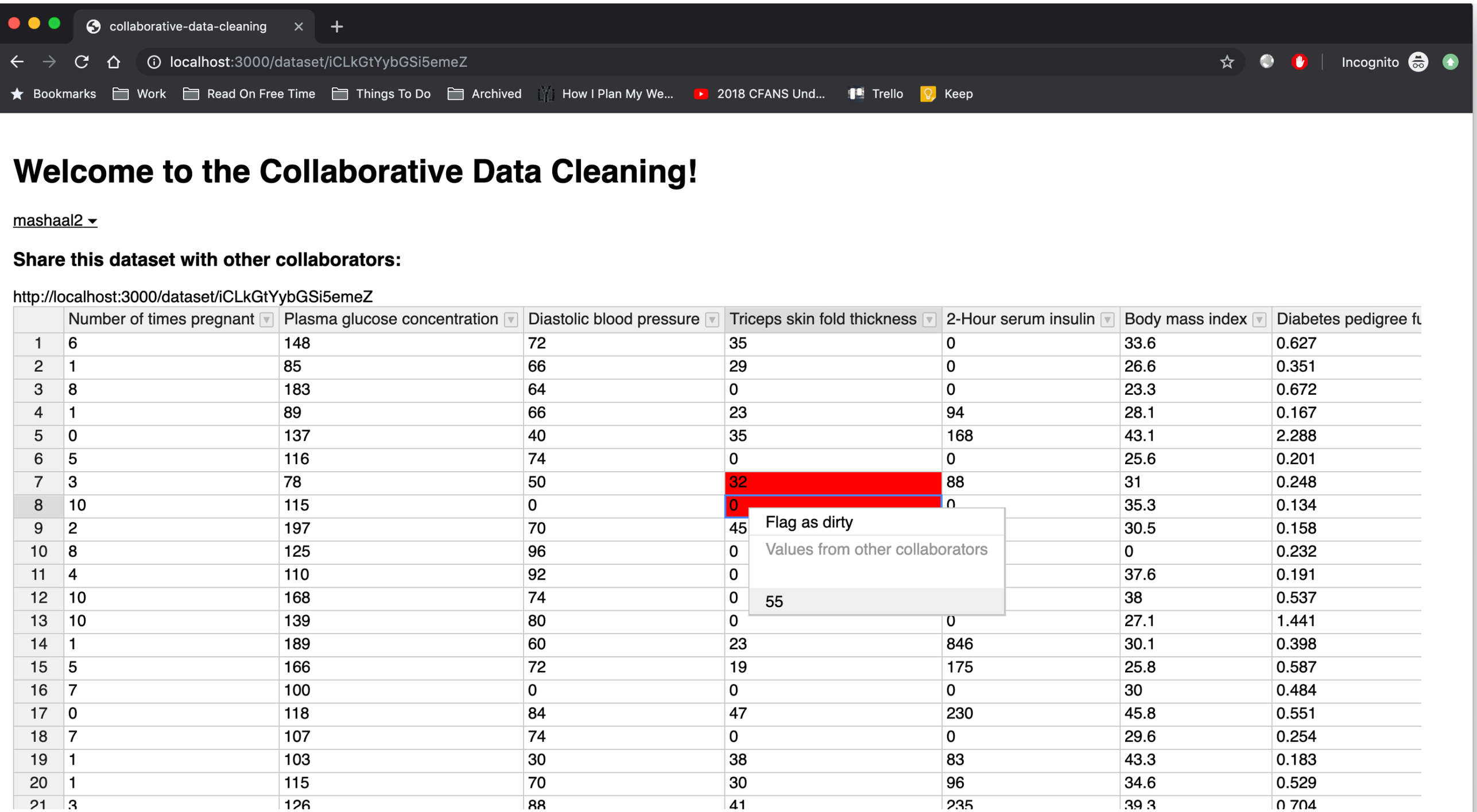


https://github.com/qcri/collaborativedatascience

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

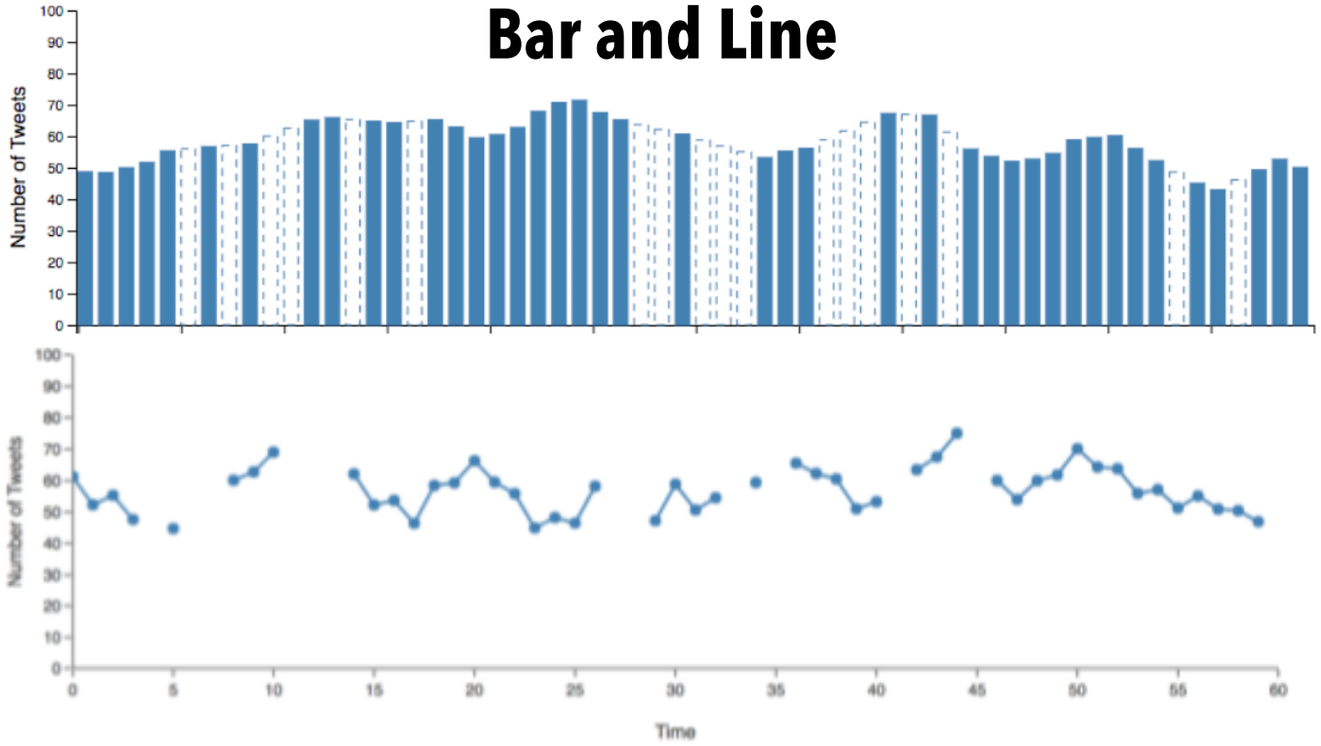
Collaborative Data Cleaning (GUI)



	Number of times pregnant	Plasma glucose concentration	Diastolic blood pressure	Triceps skin fold thickness	2-Hour serum insulin	Body mass index	Diabetes pedigree fu
1	6	148	72	35	0	33.6	0.627
2	1	85	66	29	0	26.6	0.351
3	8	183	64	0	0	23.3	0.672
4	1	89	66	23	94	28.1	0.167
5	0	137	40	35	168	43.1	2.288
6	5	116	74	0	0	25.6	0.201
7	3	78	50	32	88	31	0.248
8	10	115	0	0	0	35.3	0.134
9	2	197	70	45		30.5	0.158
10	8	125	96	0		0	0.232
11	4	110	92	0		37.6	0.191
12	10	168	74	0		38	0.537
13	10	139	80	0	0	27.1	1.441
14	1	189	60	23	846	30.1	0.398
15	5	166	72	19	175	25.8	0.587
16	7	100	0	0	0	30	0.484
17	0	118	84	47	230	45.8	0.551
18	7	107	74	0	0	29.6	0.254
19	1	103	30	38	83	43.3	0.183
20	1	115	70	30	96	34.6	0.529
21	3	126	88	41	235	39.3	0.704

Missing Data

How the methods used to impute and visualize missing data may influence analysts' perceptions of data quality and their confidence in their conclusions?

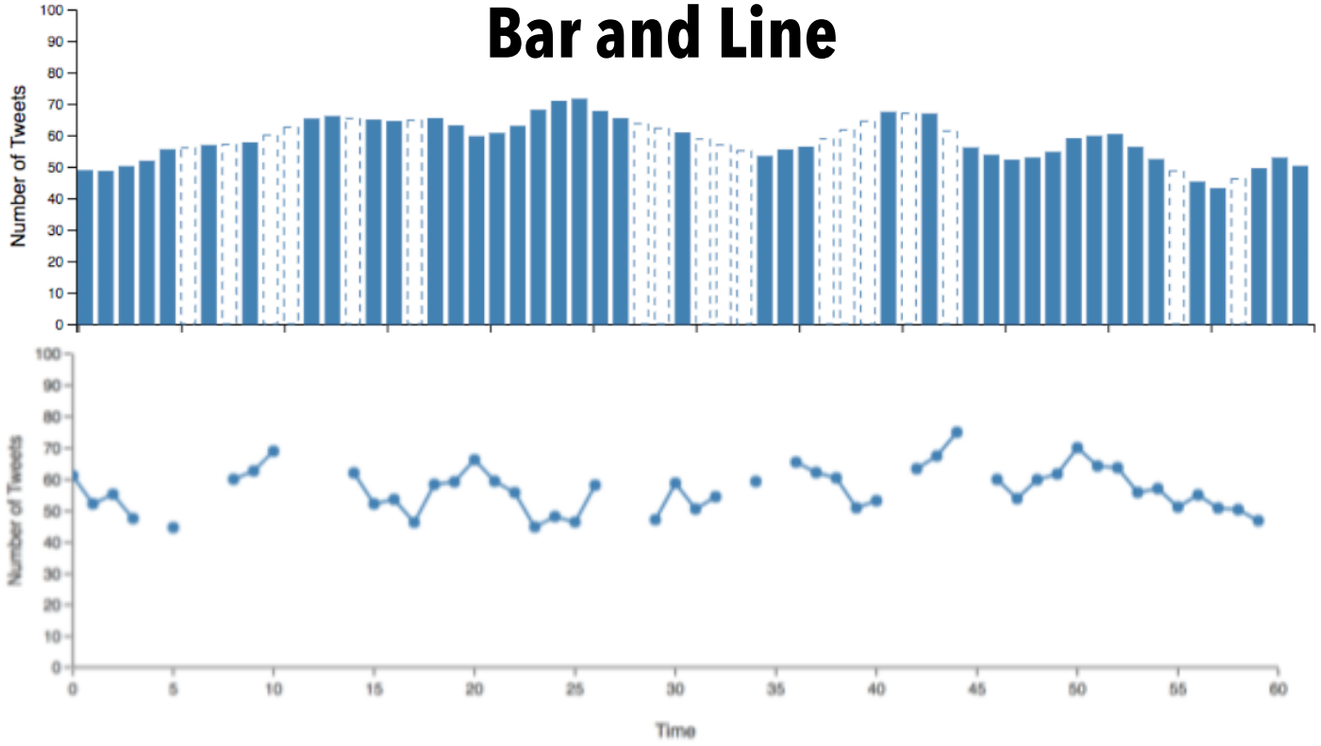


Where is My Data? Evaluating Visualizations with Missing Data. **Hayeong Song et al.**, InfoVis 2018.

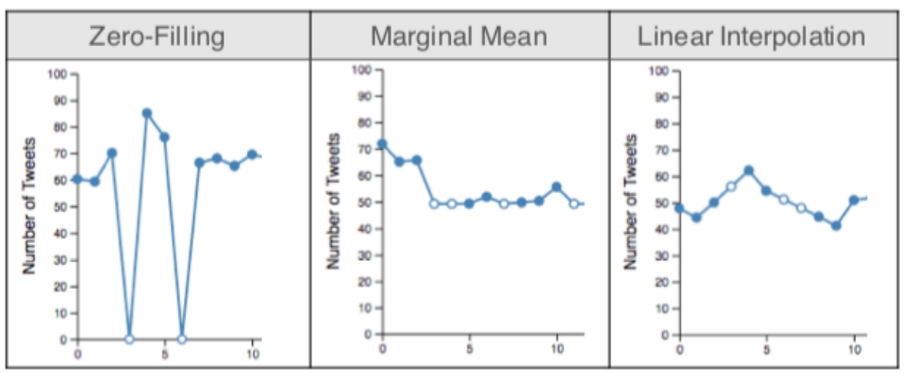
Missing Data

How the methods used to impute and visualize missing data may influence analysts' perceptions of data quality and their confidence in their conclusions?

Bar and Line



Imputation Methods

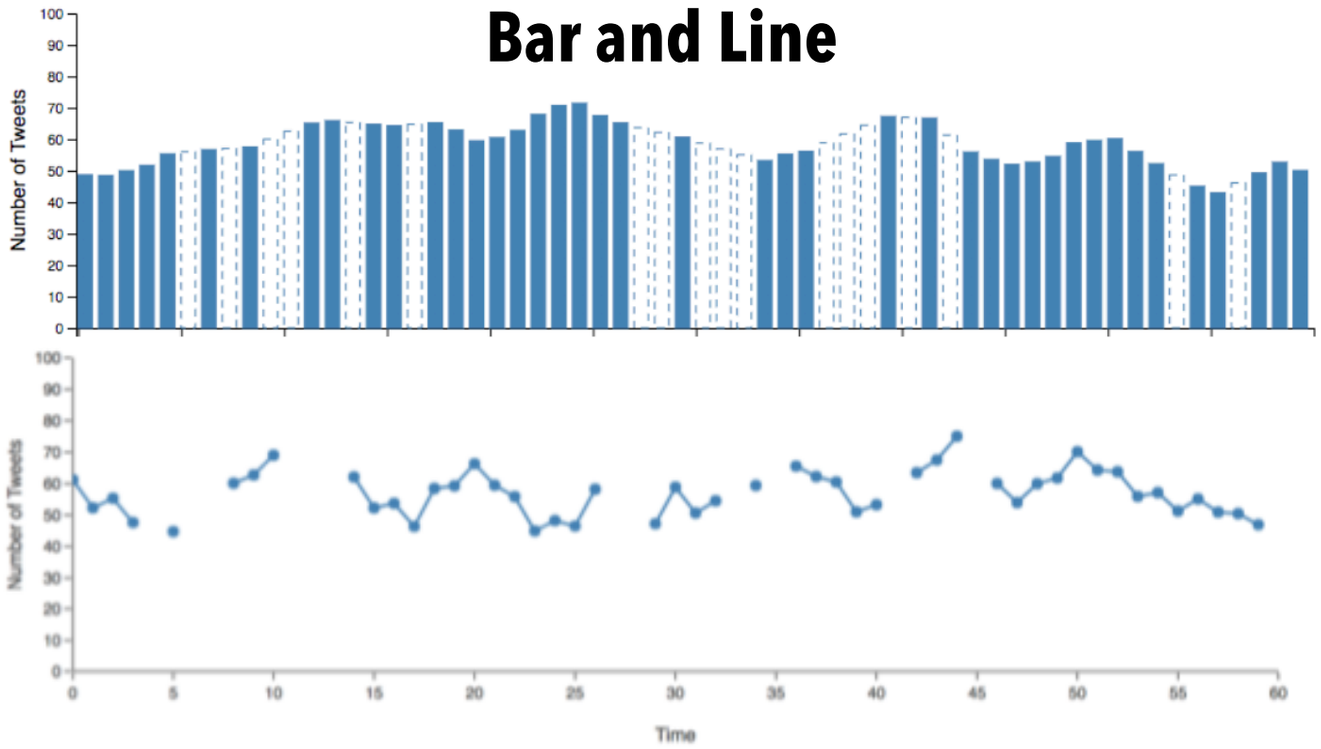


Where is My Data? Evaluating Visualizations with Missing Data. **Hayeong Song et al.**, InfoVis 2018.

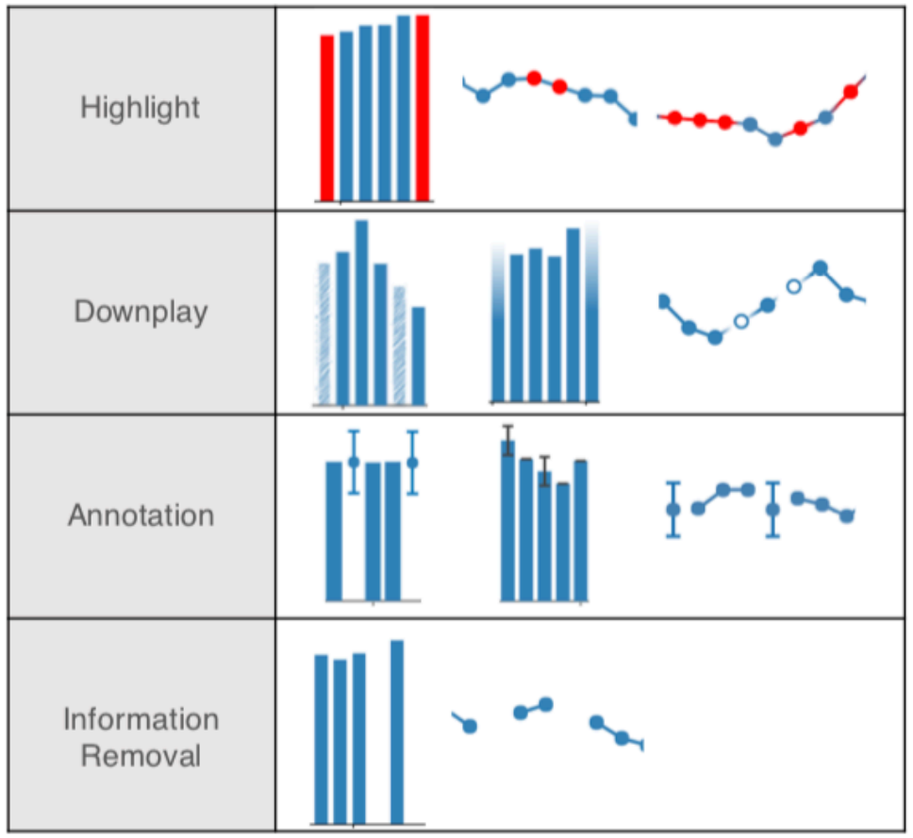
Missing Data

How the methods used to impute and visualize missing data may influence analysts' perceptions of data quality and their confidence in their conclusions?

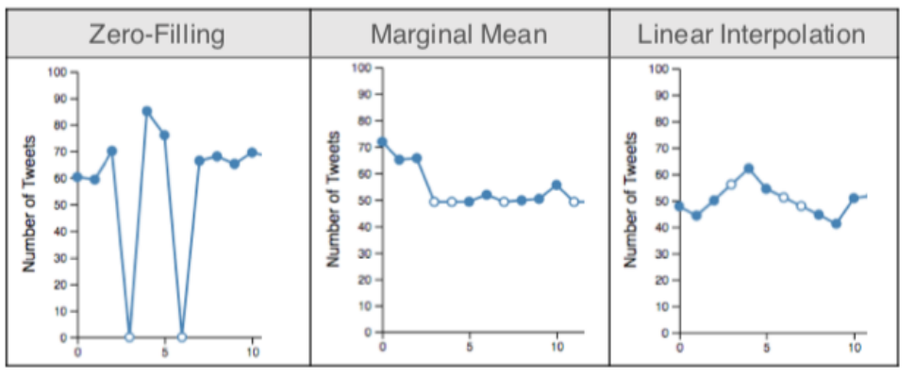
Bar and Line



Encode Imputed Values



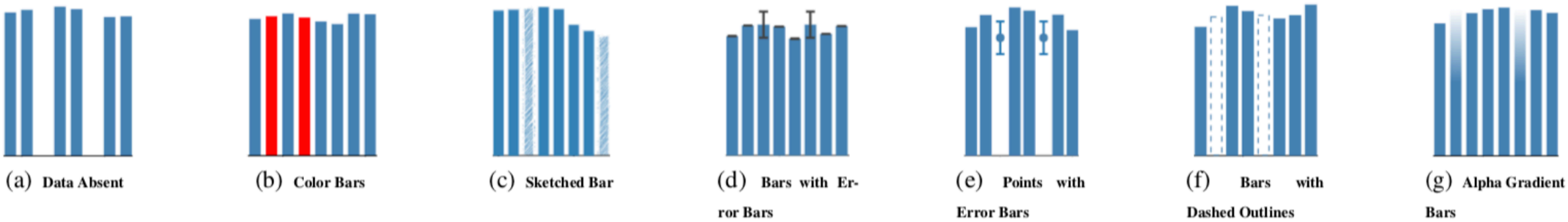
Imputation Methods



Where is My Data? Evaluating Visualizations with Missing Data. **Hayeong Song et al.**, InfoVis 2018.

Missing Data

(a) Data Absent (b) Color Points (c) Color Points & Line Gradients (d) Connected Error Bars (e) Disconnected Error Bars (f) Unfilled Points (g) Unfilled Points & Line Gradients



(a) Data Absent (b) Color Bars (c) Sketched Bar (d) Bars with Error Bars (e) Points with Error Bars (f) Bars with Dashed Outlines (g) Alpha Gradient Bars

Questions

- Is the overall rate of change larger in the first or second half-hour?
- How confident are you in your response?
- How complete is this data?
- How reliable is this data?

H1– Perceived data quality will degrade as the amount of missing data increases.

H2– Highlighting will achieve higher perceived data quality than downplaying and information removal

H3– Linear interpolation will lead to higher perceived confidence and data quality than marginal means or zero-filling

H4– Imputed values will lead to higher perceived data quality than removed values.

Implicit Error

A Framework for Externalizing Implicit Error Using Visualization, **Nina McCurday et al.**, InfoVis 2018.

Implicit Error

Measurement error: the difference between a measured value and the true value as it exists

Implicit error: measure error that is inherent to a dataset, assumed to be present and prevalent, but not explicitly defined or accounted for

- exists as tacit knowledge in the minds of experts
- is rarely quantifiable
- is accounted for subjectively during expert interpretation of the data

Implicit Error

Measurement error: the difference between a measured value and the true value as it exists

Implicit error: measure error that is inherent to a dataset, assumed to be present and prevalent, but not explicitly defined or accounted for

- exists as tacit knowledge in the minds of experts
- is rarely quantifiable
- is accounted for subjectively during expert interpretation of the data

detection
recording
collection
processing
reporting

distributed
heterogeneous
data generation
pipelines

Implicit Error

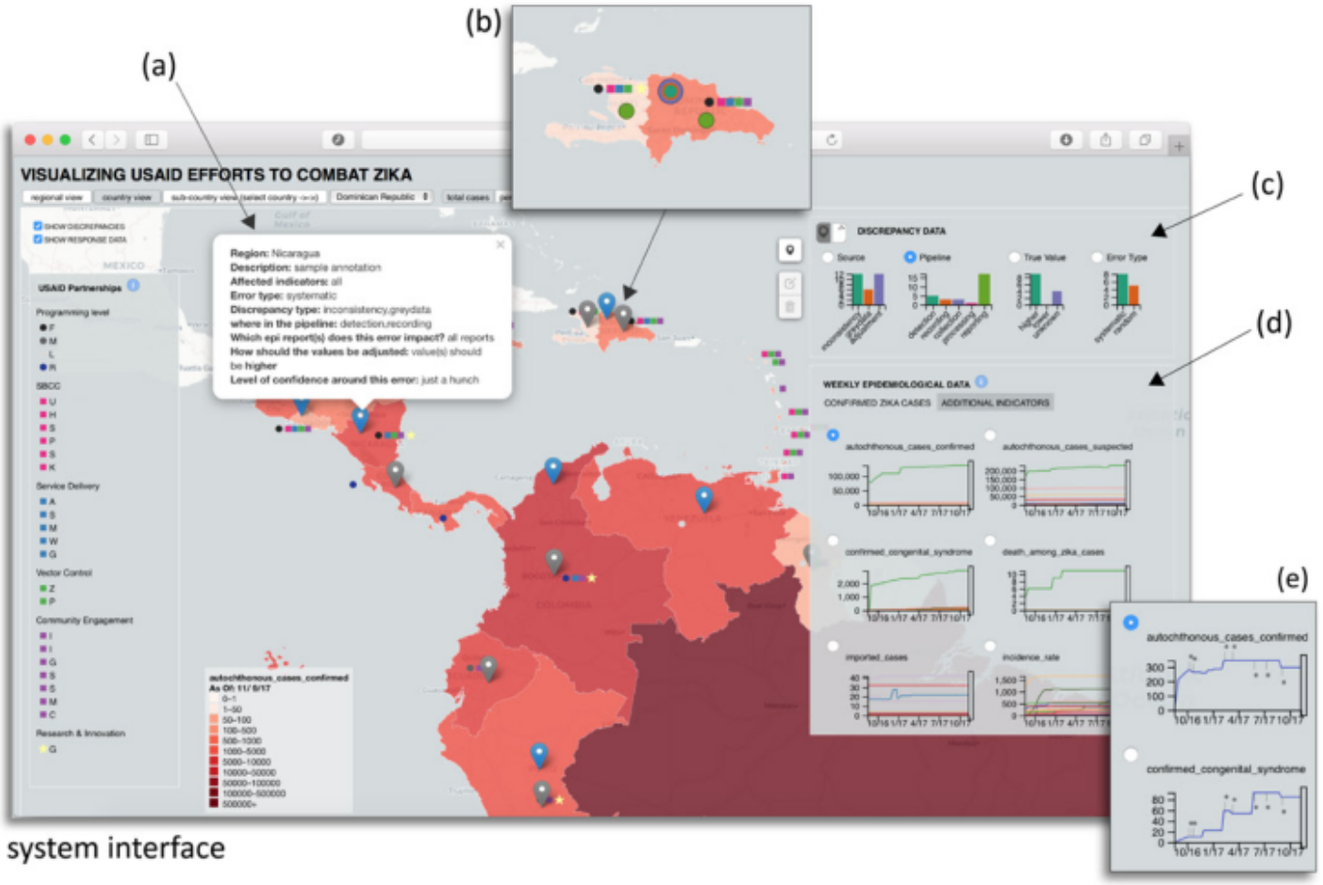
Measurement error: the difference between a measured value and the true value as it exists

Implicit error: measure error that is inherent to a dataset, assumed to be present and prevalent, but not explicitly defined or accounted for

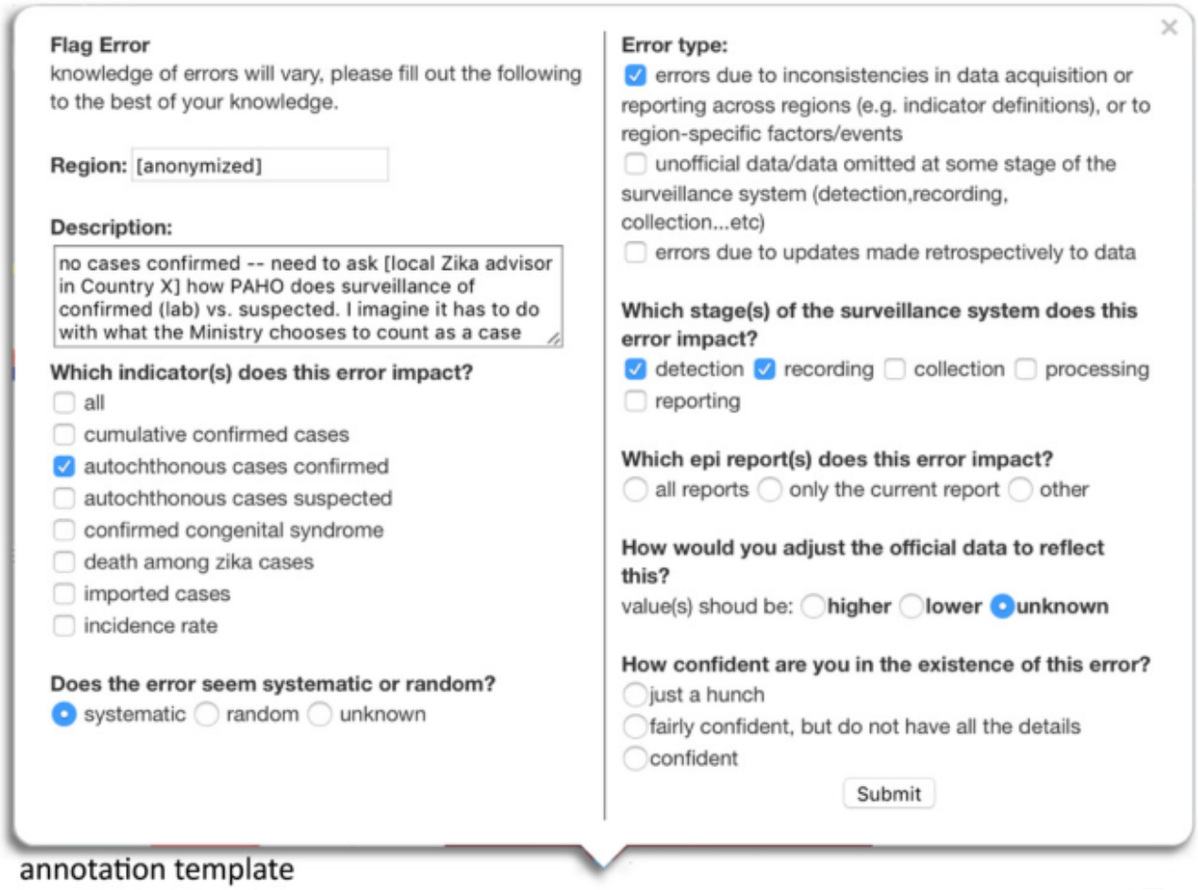
- exists as tacit knowledge in the minds of experts
- is rarely quantifiable
- is accounted for subjectively during expert interpretation of the data

detection
recording
collection
processing
reporting

distributed
heterogeneous
data generation
pipelines



system interface



annotation template

False Discoveries

Amazon Mechanical Turk

People who go to saunas are more likely to know that Mike Stonebraker is not a character in "The Simpsons"

Towards Sustainable Insights, **Carsten Binning et al.**, CIDR 2017.

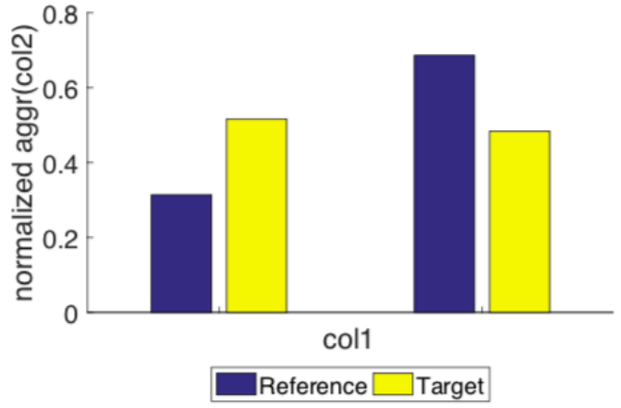
False Discoveries

Amazon Mechanical Turk →

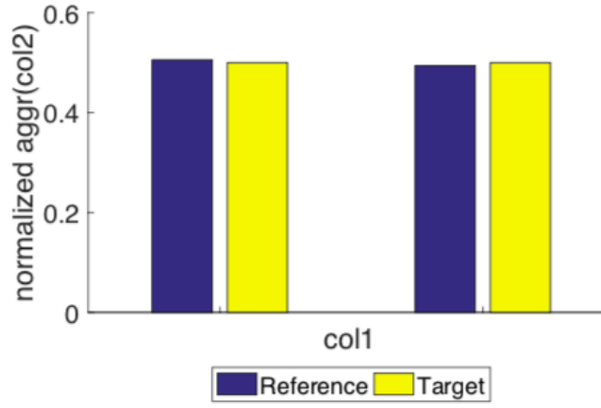
People who go to saunas are more likely to know that Mike Stonebraker is not a character in "The Simpsons"

SeeDB

Discovery rate: three times higher than for manual exploration tools such as Tableau or Vizdom



(a) Interesting visualization



(b) Uninteresting visualization

Towards Sustainable Insights, **Carsten Binning et al.**, CIDR 2017.

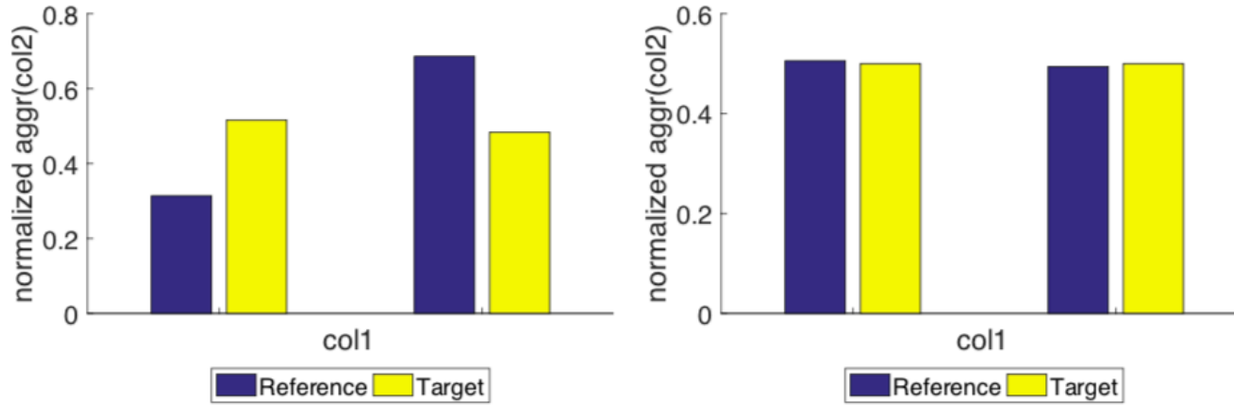
False Discoveries

Amazon Mechanical Turk →

People who go to saunas are more likely to know that Mike Stonebraker is not a character in "The Simpsons"

SeeDB

Discovery rate: three times higher than for manual exploration tools such as Tableau or Vizdom



(a) Interesting visualization (b) Uninteresting visualization

Survey Data

Reference views: 9,996

Target views: 2,078,608

Recommended views: 708,109

Towards Sustainable Insights, **Carsten Binning et al.**, CIDR 2017.

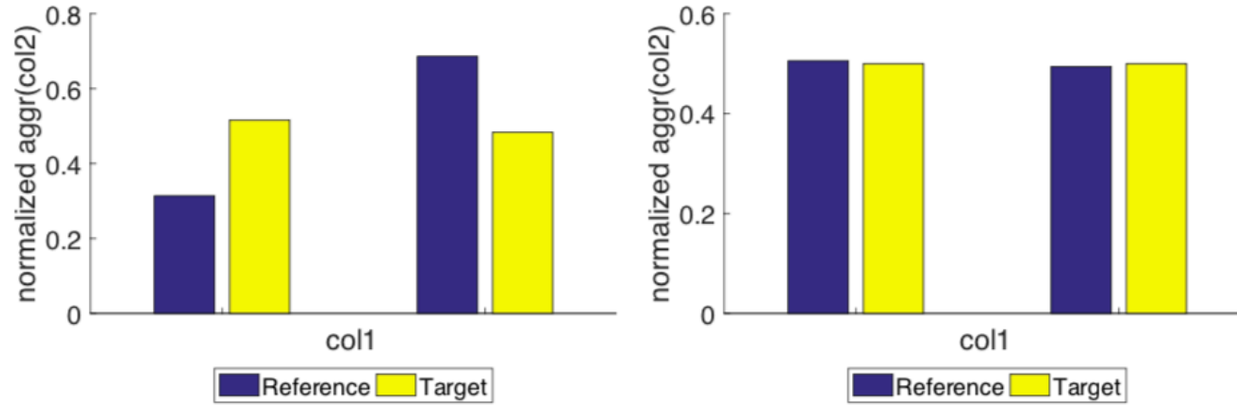
False Discoveries

Amazon Mechanical Turk →

People who go to saunas are more likely to know that Mike Stonebraker is not a character in "The Simpsons"

SeeDB

Discovery rate: three times higher than for manual exploration tools such as Tableau or Vizdom



(a) Interesting visualization (b) Uninteresting visualization

Survey Data

Reference views: 9,996
Target views: 2,078,608
Recommended views: 708,109

QUDE

Quantify the **U**ncertainty in **D**ata **E**xploration

– False Discovery Rate –

Towards Sustainable Insights, **Carsten Binning et al.**, CIDR 2017.

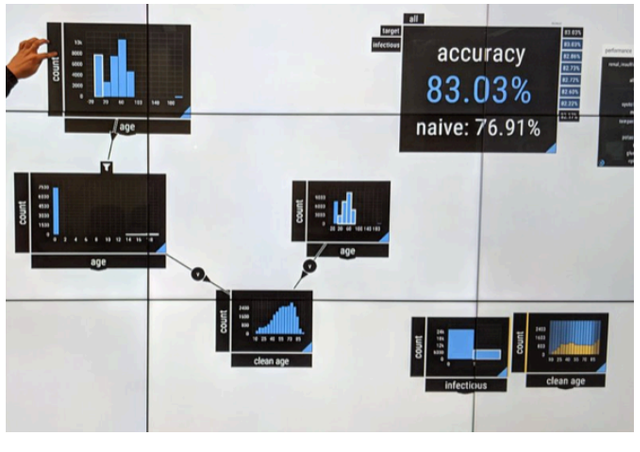
Collaborative Data Visualization

https://vimeo.com/yalongyang

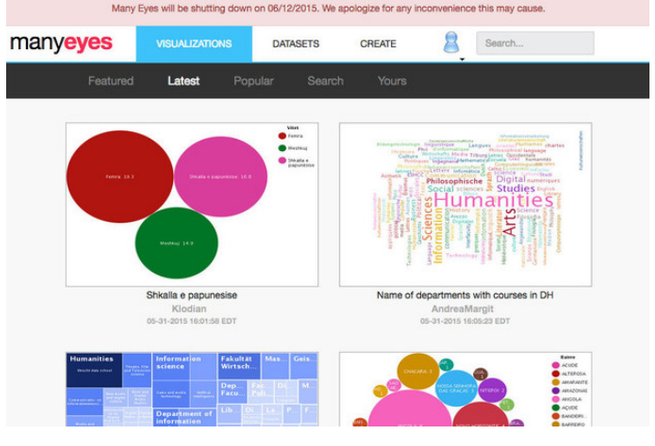
Enhances the traditional visualization by bringing together **many experts** so that each can contribute toward the common goal of the **understanding** of the object, phenomenon, or data

Synchronous

Northstar



Asynchronous



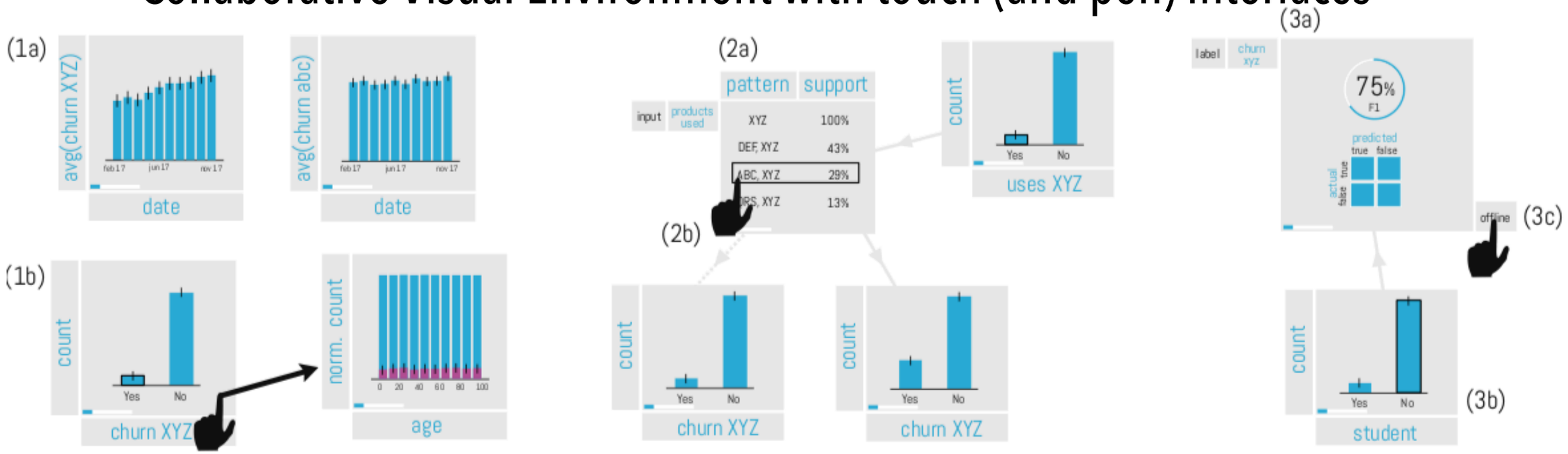
manyeyes

Northstar: An Interactive Data Science System

Writing scripts and SQL-like languages

↓ 10 years from now

Collaborative Visual Environment with touch (and pen) interfaces



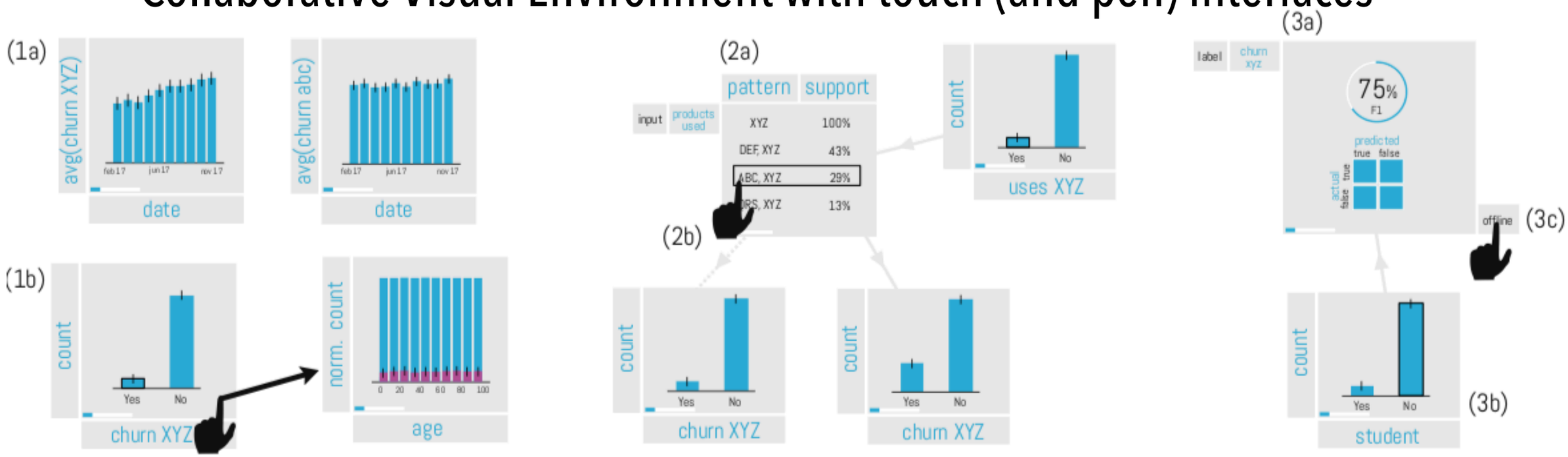
Northstar: An Interactive Data Science System, **Tim Kraska**. VLDB 2018.

Northstar: An Interactive Data Science System

Writing scripts and SQL-like languages

↓ 10 years from now

Collaborative Visual Environment with touch (and pen) interfaces



- **Vizdom**: a visual data exploration environment pen and touch interface
- **IDEA**: an intelligent cache and string approximation engine
- **QUDE**: warn about common mistakes and problems
- **Alphine Meadow**: automatically orchestrate a ML pipeline (i.e., plan)

Northstar: An Interactive Data Science System, **Tim Kraska**. VLDB 2018.

Democratizing Data Visualization

Immersion

- exploration
- inside-feeling

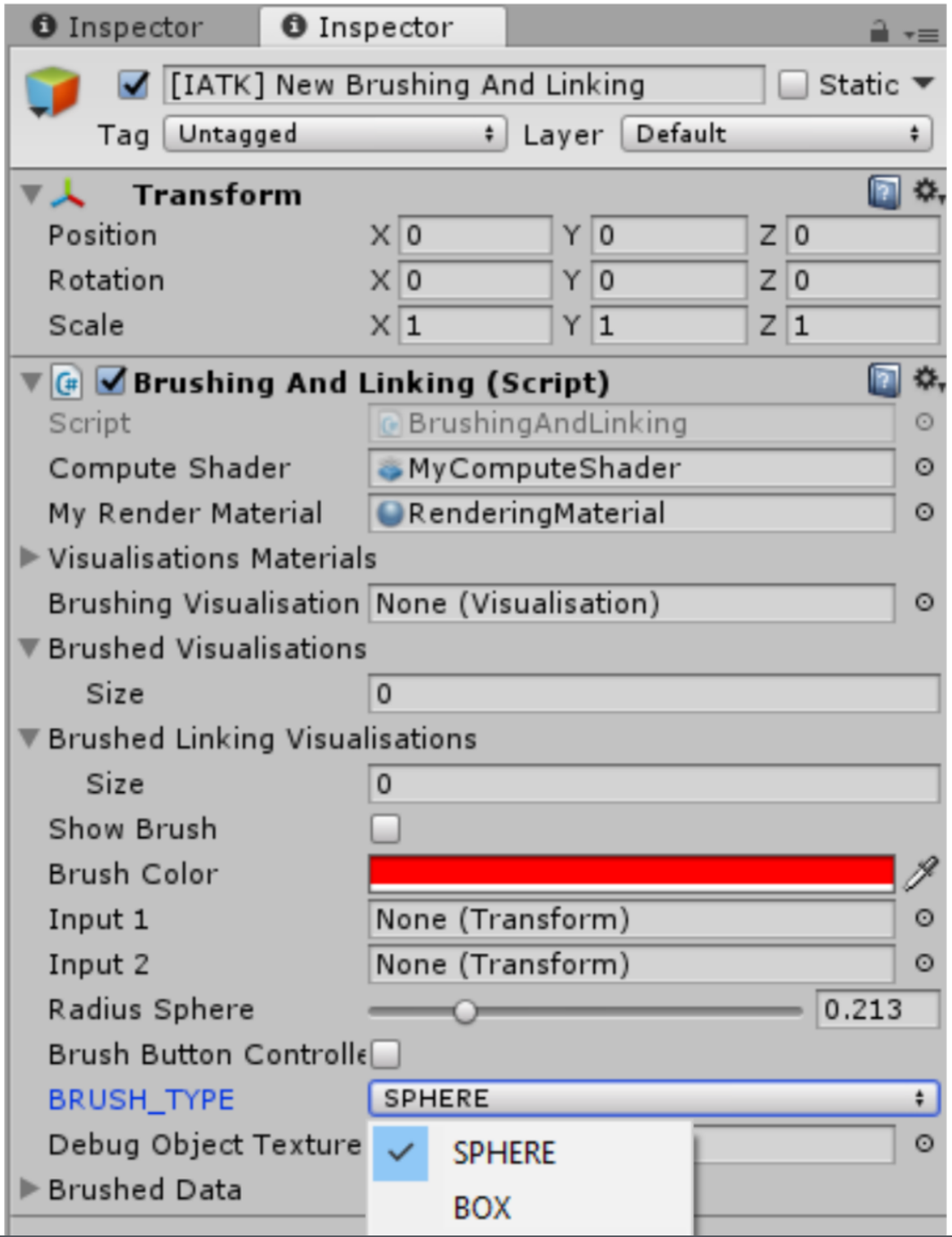
Interaction

- gesture + voice
- real-time

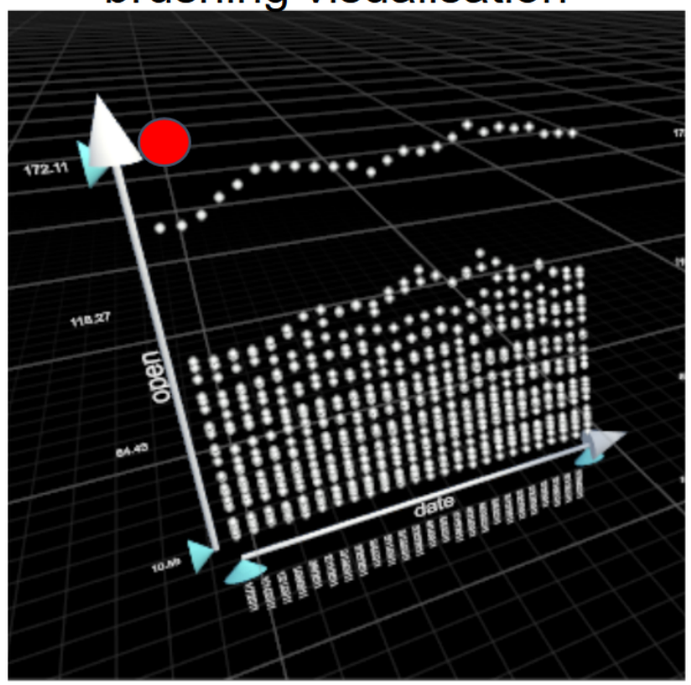
Intelligence

- reactive
- proactive

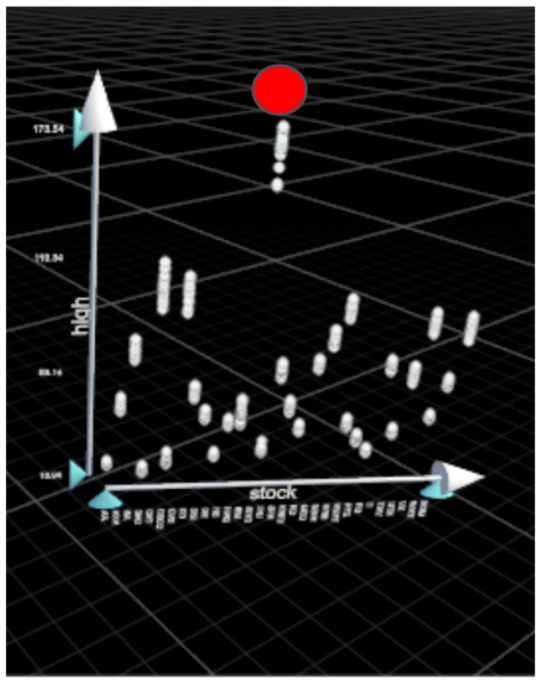
IATK: Tableau for Virtual Reality



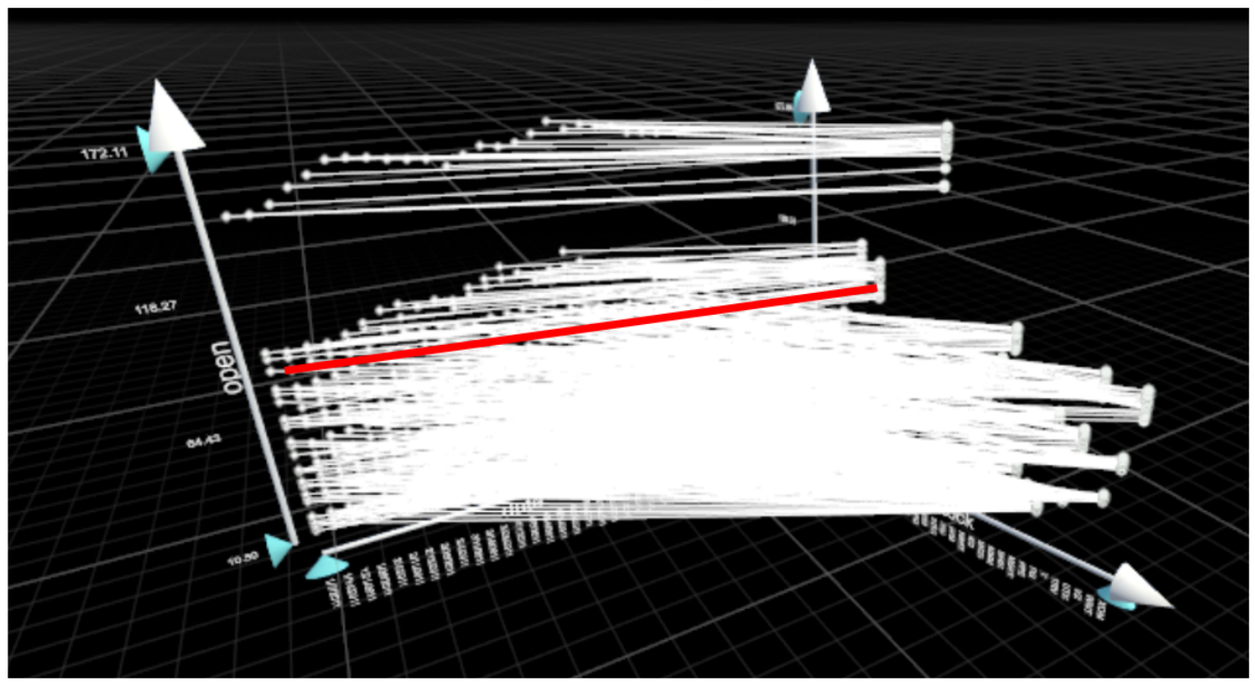
brushing visualisation



brushed visualisation

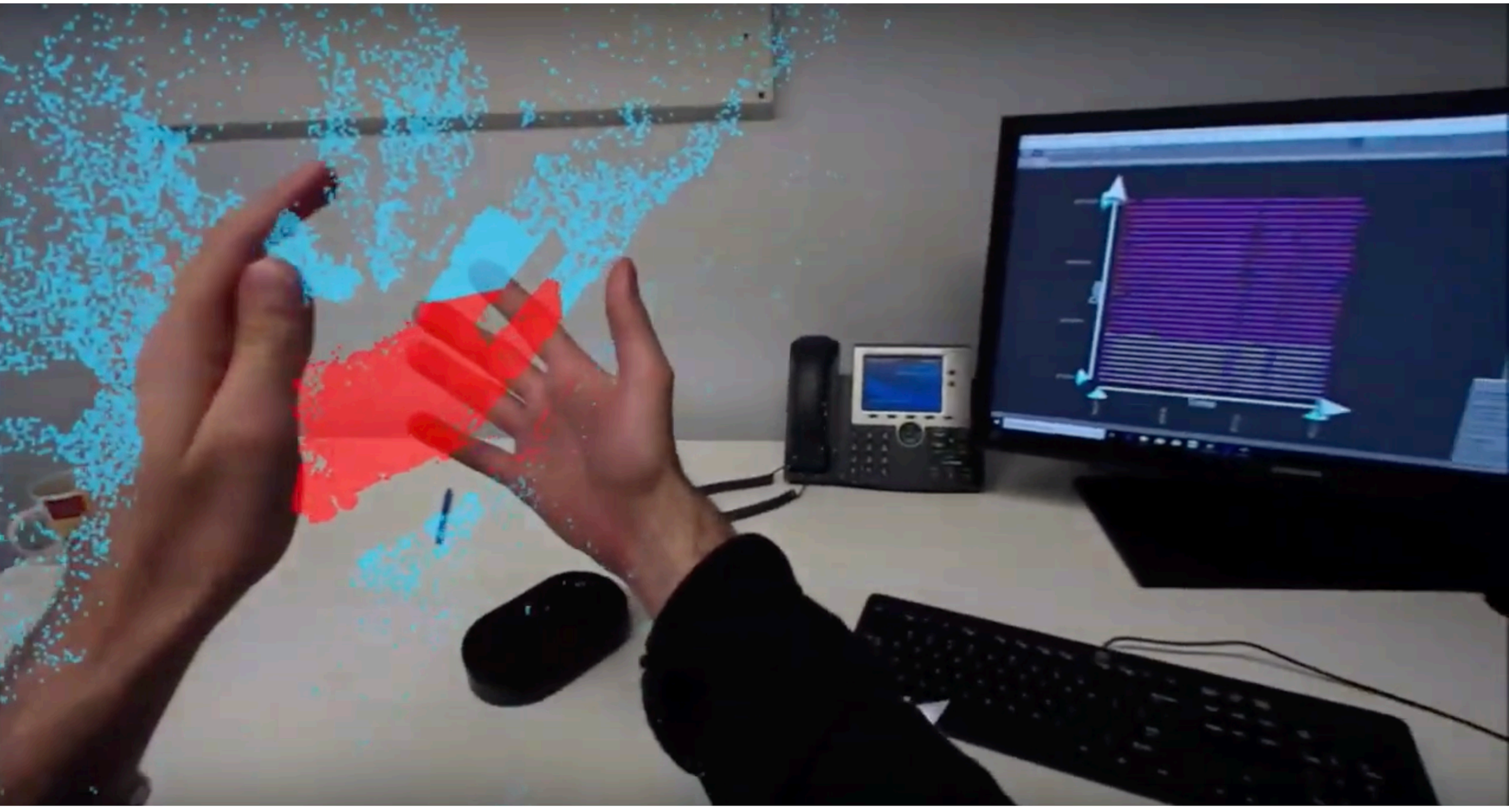


Brushed Linking Visualisation

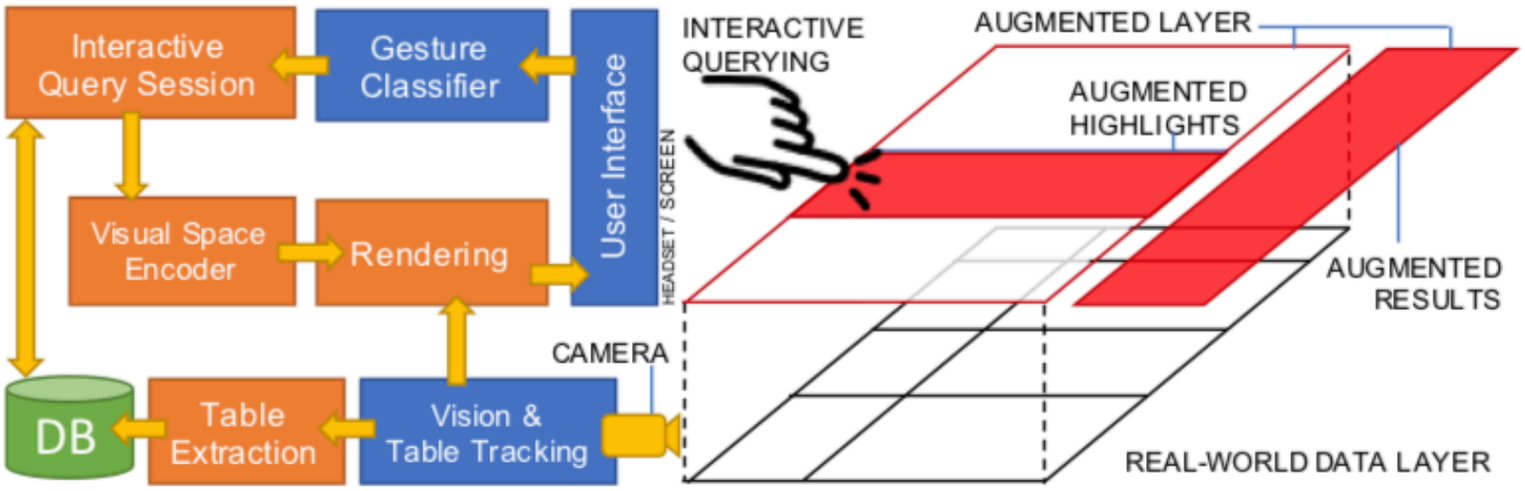


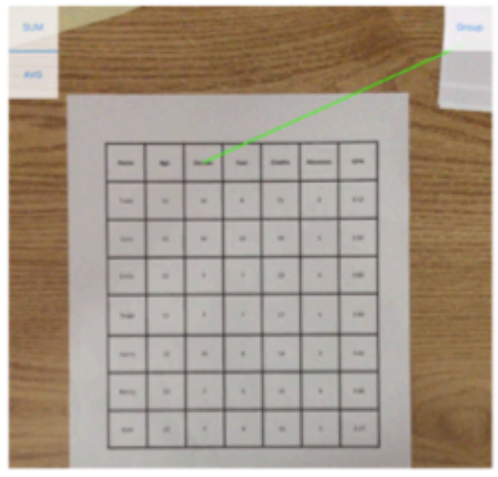
IATK: An Immersive Analytics Toolkit. **Maxime Cordeil et al.**, IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 2019.

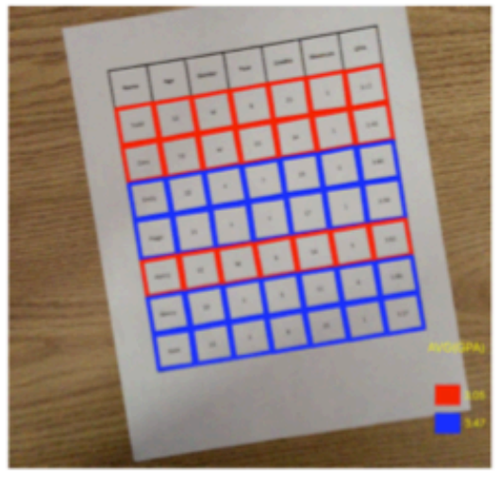
IATK: Tableau for Virtual Reality



Querying Real-World Data using Augmented Reality







ARQuery: Hallucinating Analytics over Real-World Data using Augmented Reality, **Codi Burley et al.**, CIDR 2019.

Future Work

1. Visualization-driven data preparation/curation
2. A generic system to support efficient/approximate/progressive data visualization
3. (Asynchronous) Collaborative data visualization
4. Data Visualization Benchmarks
5. Immersive data analytics for abstract data
6. Deep learning for data visualization

Further Readings

- Stephen Macke, "***Adaptive Sampling for Rapidly Matching Histograms***", VLDB 2018
- Doris Jung-Lin Lee et al., "***The Case for a Visual Discovery Assistant: A Holistic Solution for Accelerating Visual Data Exploration***", IEEE Data Bulletin 2018
- Kelly Mack et al., "***Characterizing Scalability Issues in Spreadsheet Software using Online Forums***", CHI 2018
- Leilani Battle et al., "***Beagle: Automated Extraction and Interpretation of Visualizations from the Web***", CHI 2018
- Wenbo Tao et al., "***Kyrix: Interactive Visual Data Exploration at Scale***", CIDR 2019
- Kevin Zeng Hu et al., "***VizNet: Towards A Large-Scale Visualization Learning and Benchmarking Repository***", CHI 2019
- Codi j. Burley et al., "***ARQuery: Hallucinating Analytics over Real-World Data using Augmented Reality***", CIDR 2019
- Meraj Ahmed Khan et al., "***Data Tweening: Incremental Visualization of Data Transforms***", PVLDB 2017
- Ciro Donalek et al., "***Immersive and Collaborative Data Visualization Using Virtual Reality Platforms***", IEEE International Conference on Big Data, 2014.
- Fernanda B. Viegas et al., "***ManyEyes: A Site for Visualization at Internet Scale***", IEEE Transactions on Visualization and Computer Graphics, 2007.