Visualization for Data Science

Dr. Yuyu Luo

yuyuluo@hkust-gz.edu.cn

Data Science and Analytics, HKUST(GZ)

Outline

- Background
- Visualization Principles
- Automatic Visualization

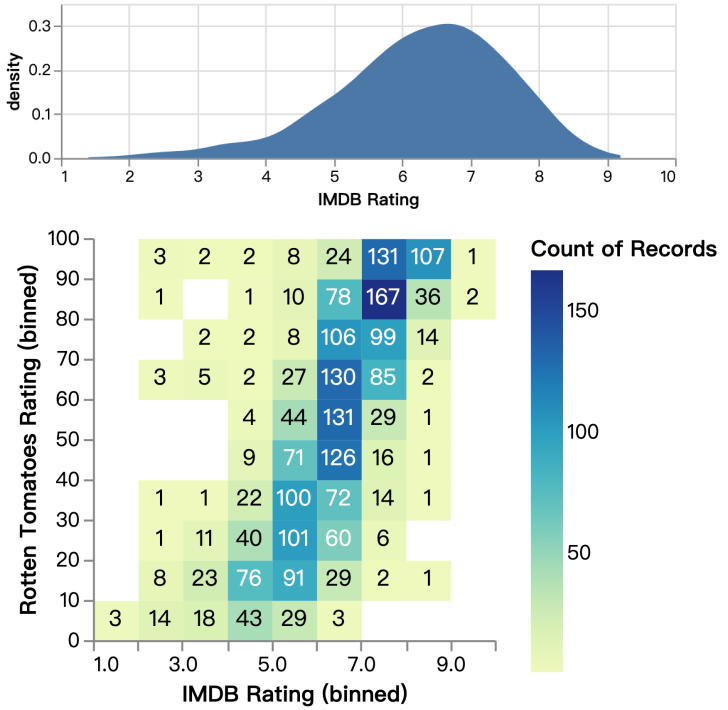
Outline

- **Background**
- Visualization Principles
- Automatic Visualization

What

Data visualization is the graphical representation of data

Worldwide Gross	Production Budget	Release Year	Content Rating	Running Time	Genre	Creative Type	Rotten Tomatoes Rating	IMDB Rating
25728961	20000000	1996	R	107	Horror	Fantasy	63	7.1
148345997	65000000	1996	R	108	Action	Contemporary Fiction	55	5.8
20278055	40000000	1996	R	111	Drama	Contemporary Fiction	55	6.1
38623460	10000000	1996	PG-13	92	Comedy	Contemporary Fiction	58	6.9
51204567	7000000	1996	R	87	Thriller	Contemporary Fiction	94	8.3
55669466	15000000	1996	R	100	Thriller	Fantasy	45	5.9
495900000	88000000	1996	PG-13	117	Action	Contemporary Fiction	57	6
104364680	57000000	1996	PG-13	108	Adventure	Fantasy	50	6.2
17220599	45000000	1996	PG	100	Action	Super Hero	43	4.8
336069511	75000000	1996	R	136	Action	Contemporary Fiction	66	7.2
102825796	47000000	1996	PG-13	95	Comedy	Contemporary Fiction	52	5.8
234400000	100000000	1996	R	115	Action	Contemporary Fiction	34	5.9
325500000	100000000	1996	G	86	Adventure	Historical Fiction	73	6.5
273814019	55000000	1996	PG-13	95	Comedy	Kids Fiction	67	5.6
32773011	50000000	1996	R	115	Comedy	Contemporary Fiction	12	3.9
817400878	75000000	1996	PG-13	145	Adventure	Science Fiction	61	6.5
142836382	32000000	1996	PG	124	Drama	Contemporary Fiction	50	6.3
100833145	46000000	1996	R	115	Drama	Historical Fiction	85	6.6
20133326	45000000	1996	PG-13	117	Comedy	Science Fiction	44	5.7
24000785	3100000	1996	R	94	Drama	Contemporary Fiction	89	8.2
152266007	40000000	1996	R	150	Drama	Contemporary Fiction	68	7.1
32223424	25000000	1996	R	113	Comedy	Contemporary Fiction	51	6.7
36682170	25000000	1996	G	94	Adventure	Fantasy	27	5.3
60209334	55000000	1996	PG-13	106	Action	Contemporary Fiction	13	5.2
25426861	50000000	1996	R	101	Action	Science Fiction	56	5.3
58617334	45000000	1996	PG-13	113	Drama	Contemporary Fiction	17	5.3
75854588	45000000	1996	R	105	Romantic Comedy	Contemporary Fiction	69	6.1

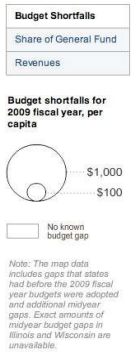


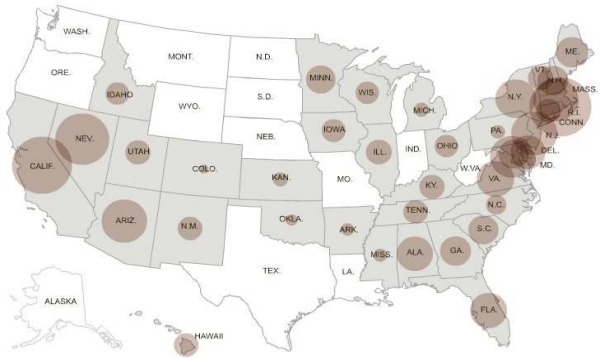
Data

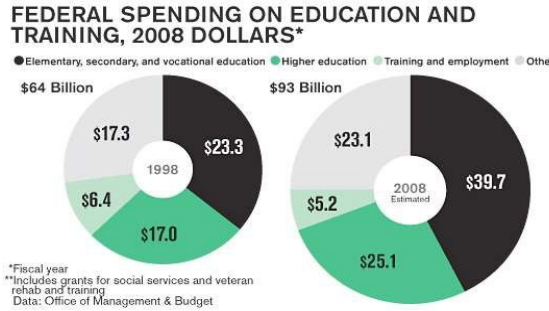
Visualization

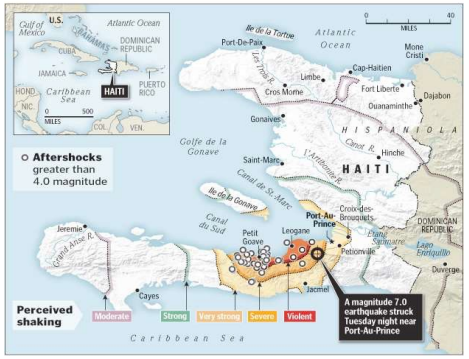
What

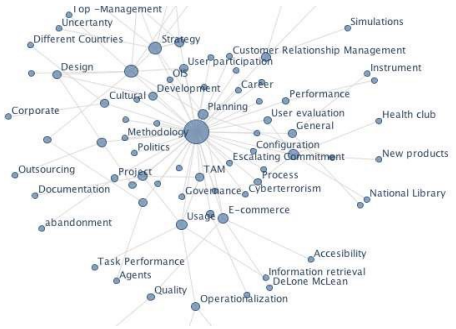
Visualization: To convey information through visual representations

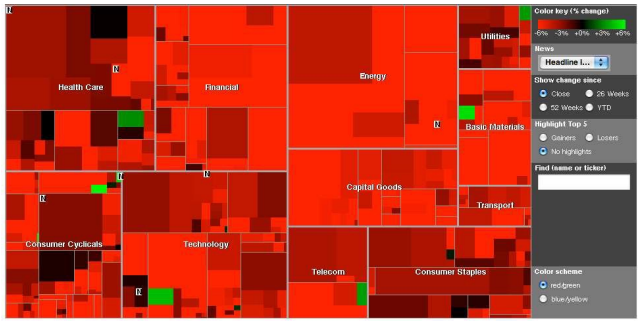


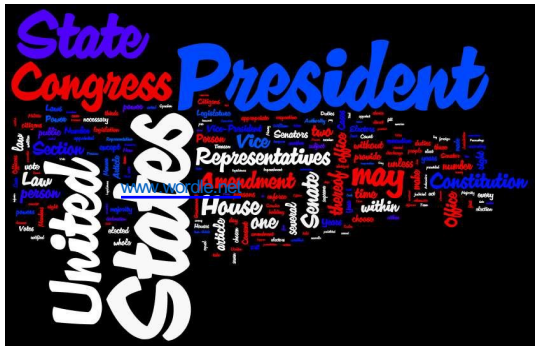






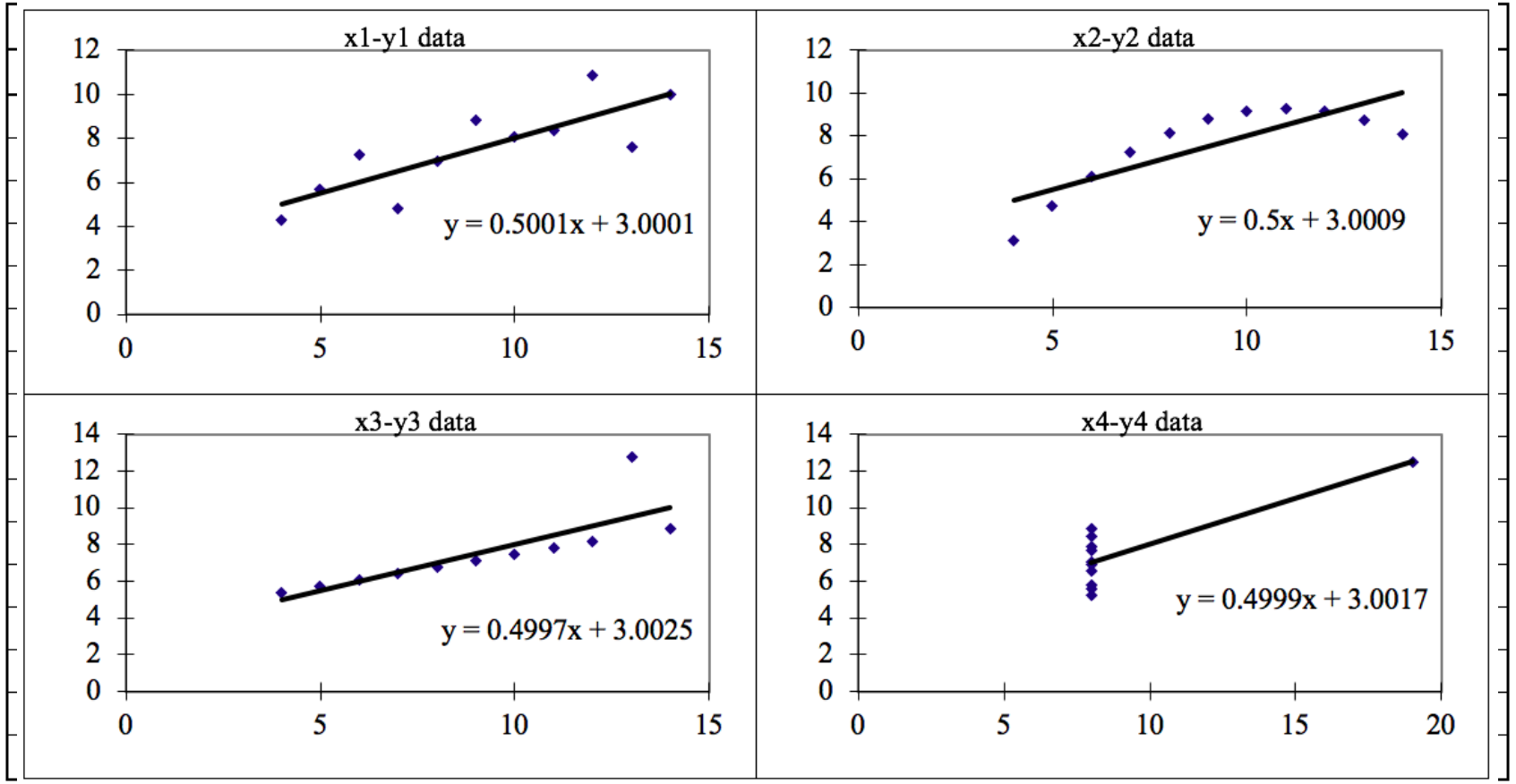






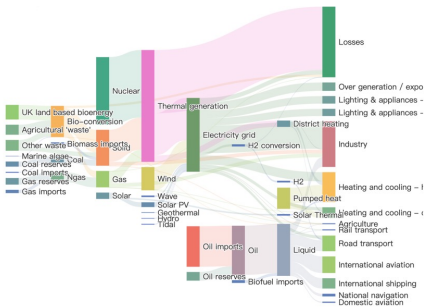
Why

Example: Anscombe's Quartet – the importance of Visualization

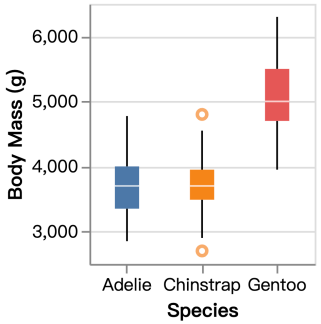


Making data more accessible and understandable by human

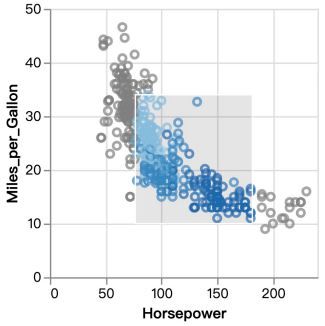
Why



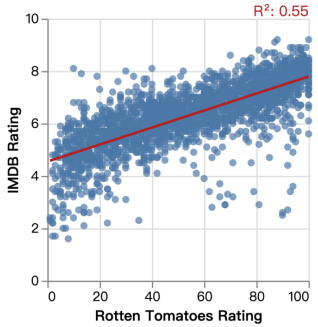
Communication



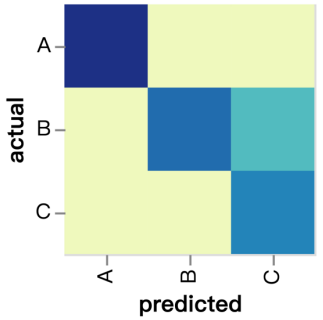
Understanding



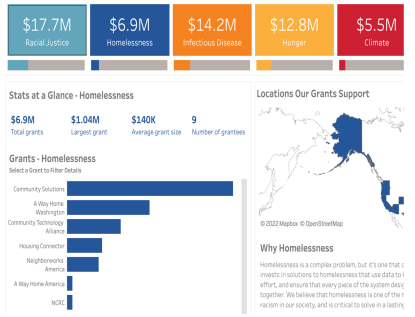
Exploration



Modelling



Evaluation



Monitoring & Communication

Problem Understanding → **Data Preparation** → **Data Exploration** → **Modelling** → **Evaluation** → **Development**

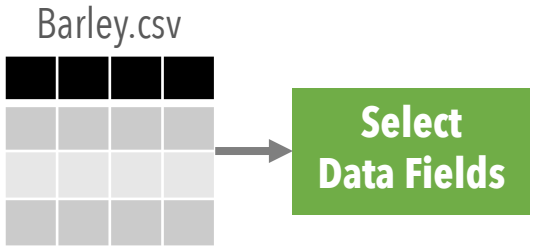
Data Science Pipeline

Making *Data Science and Analytics* More Effective !

How to effectively visualize data?

yield	variety	year	site
27	Manchuria	1931	University Farm
48.86667	Manchuria	1931	Waseca
27.43334	Manchuria	1931	Morris
39.93333	Manchuria	1931	Crookston
32.96667	Manchuria	1931	Grand Rapids
28.96667	Manchuria	1931	Duluth
43.06666	Glabron	1931	University Farm
55.2	Glabron	1931	Waseca
28.76667	Glabron	1931	Morris
38.13333	Glabron	1931	Crookston
29.13333	Glabron	1931	Grand Rapids

How to effectively visualize data?



Data

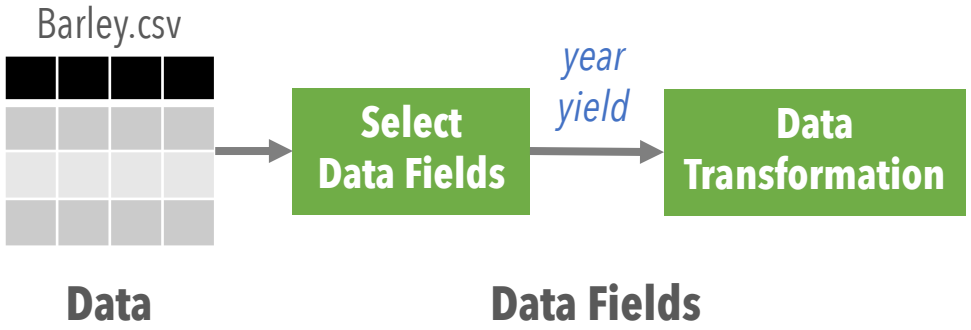
yield	variety	year	site
27	Manchuria	1931	University Farm
48.86667	Manchuria	1931	Waseca
27.43334	Manchuria	1931	Morris
39.93333	Manchuria	1931	Crookston
32.96667	Manchuria	1931	Grand Rapids
28.96667	Manchuria	1931	Duluth
43.06666	Glabron	1931	University Farm
55.2	Glabron	1931	Waseca
28.76667	Glabron	1931	Morris
38.13333	Glabron	1931	Crookston
29.13333	Glabron	1931	Grand Rapids

How to effectively visualize data?

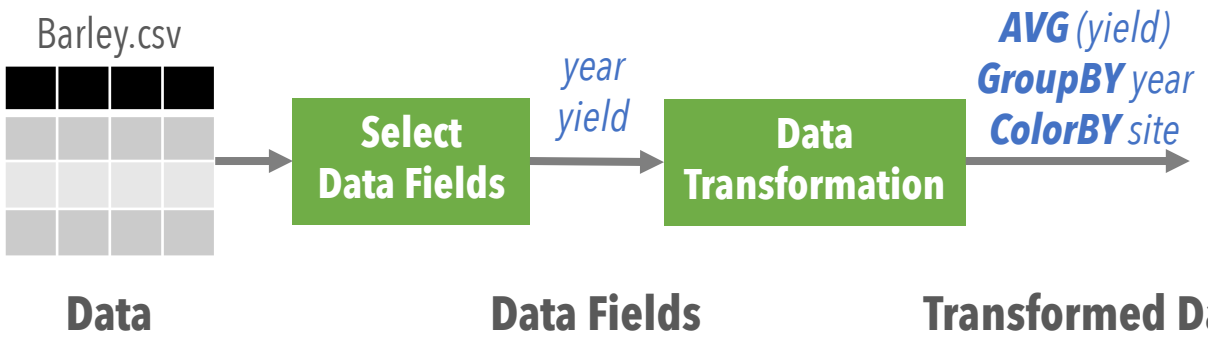
Data **Data Fields**

yield	variety	year	site
27	Manchuria	1931	University Farm
48.86667	Manchuria	1931	Waseca
27.43334	Manchuria	1931	Morris
39.93333	Manchuria	1931	Crookston
32.96667	Manchuria	1931	Grand Rapids
28.96667	Manchuria	1931	Duluth
43.06666	Glabron	1931	University Farm
55.2	Glabron	1931	Waseca
28.76667	Glabron	1931	Morris
38.13333	Glabron	1931	Crookston
29.13333	Glabron	1931	Grand Rapids

How to effectively visualize data?

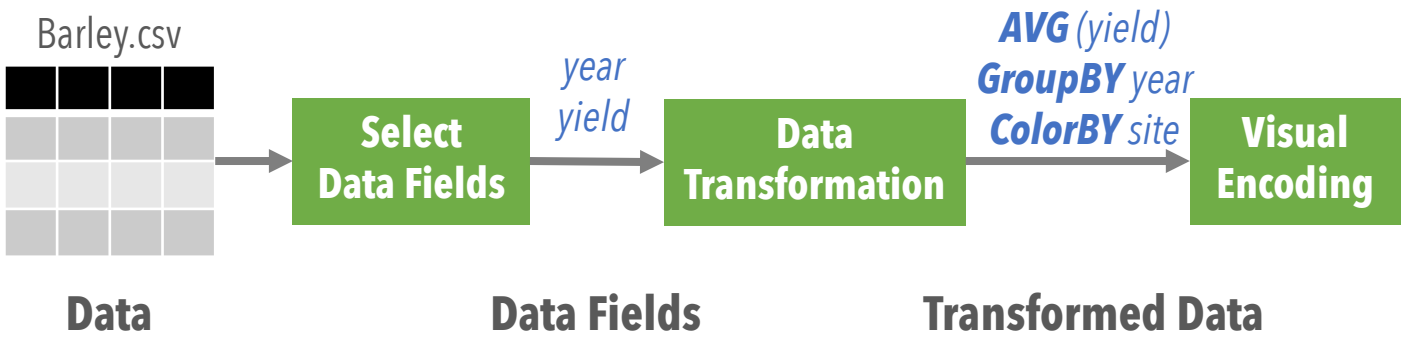


How to effectively visualize data?



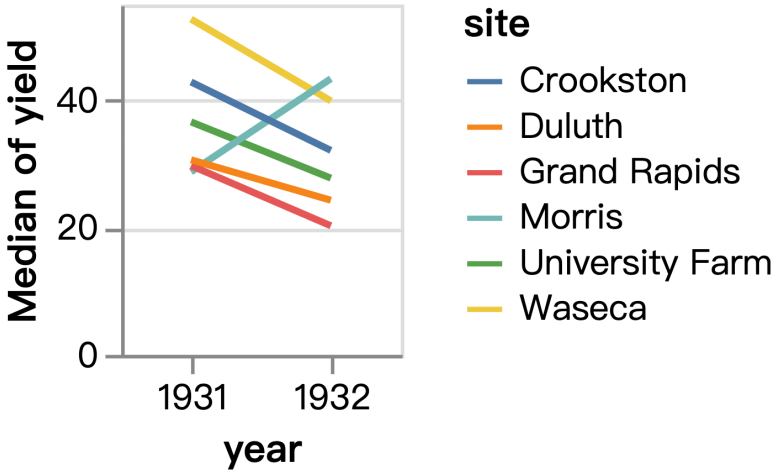
year	site	AVG(yield)
1931	"University Farm"	36.583330000000004
1931	"Waseca"	52.716665000000006
1931	"Morris"	28.733335
1931	"Crookston"	42.85
1931	"Grand Rapids"	29.71667
1931	"Duluth"	30.633335000000002
1932	"University Farm"	27.750005
1932	"Waseca"	39.883335
1932	"Morris"	43.36667
1932	"Crookston"	32.099995

How to effectively visualize data?

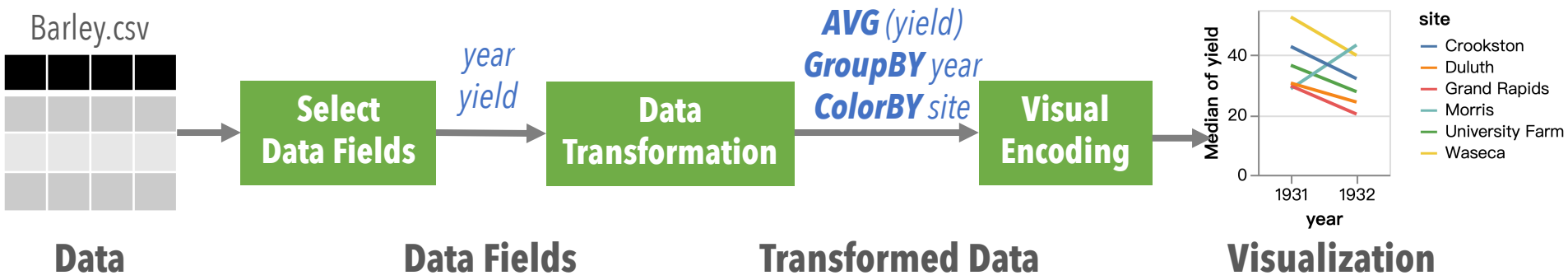


year	site	AVG(yield)
1931	"University Farm"	36.583330000000004
1931	"Waseca"	52.716665000000006
1931	"Morris"	28.733335
1931	"Crookston"	42.85
1931	"Grand Rapids"	29.71667
1931	"Duluth"	30.633335000000002
1932	"University Farm"	27.750005
1932	"Waseca"	39.883335
1932	"Morris"	43.36667
1932	"Crookston"	32.099995

```
"mark": "line"
"x": "year"
"y": "AVG(yield)"
"color":  "site"
```

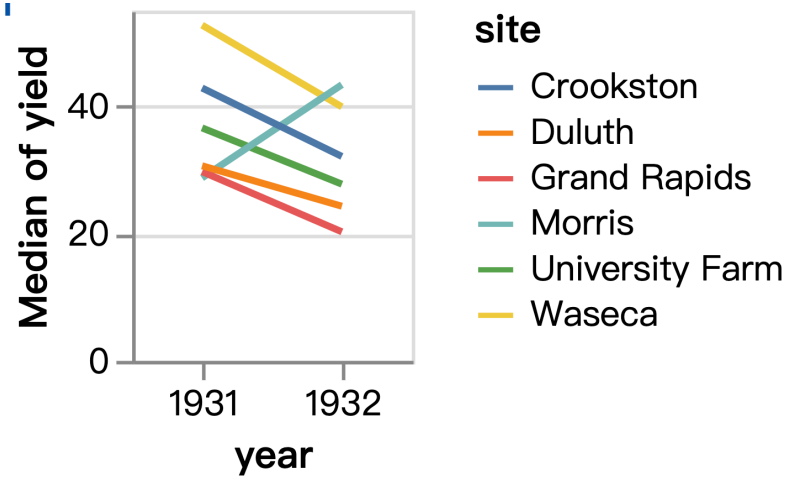


How to effectively visualize data?

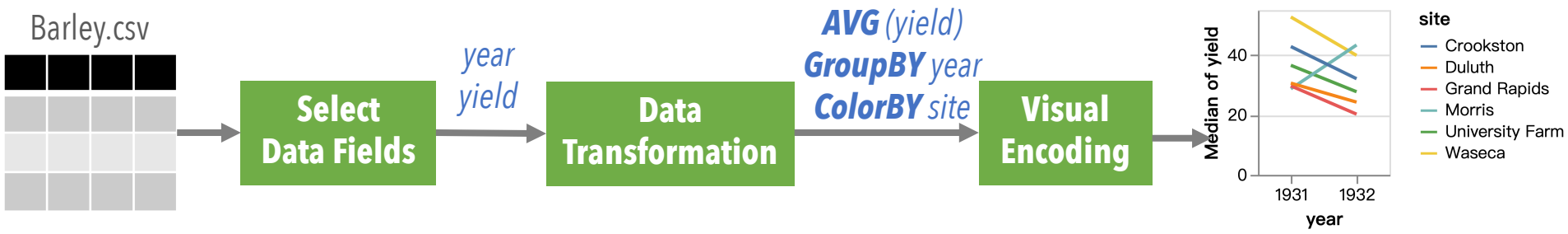


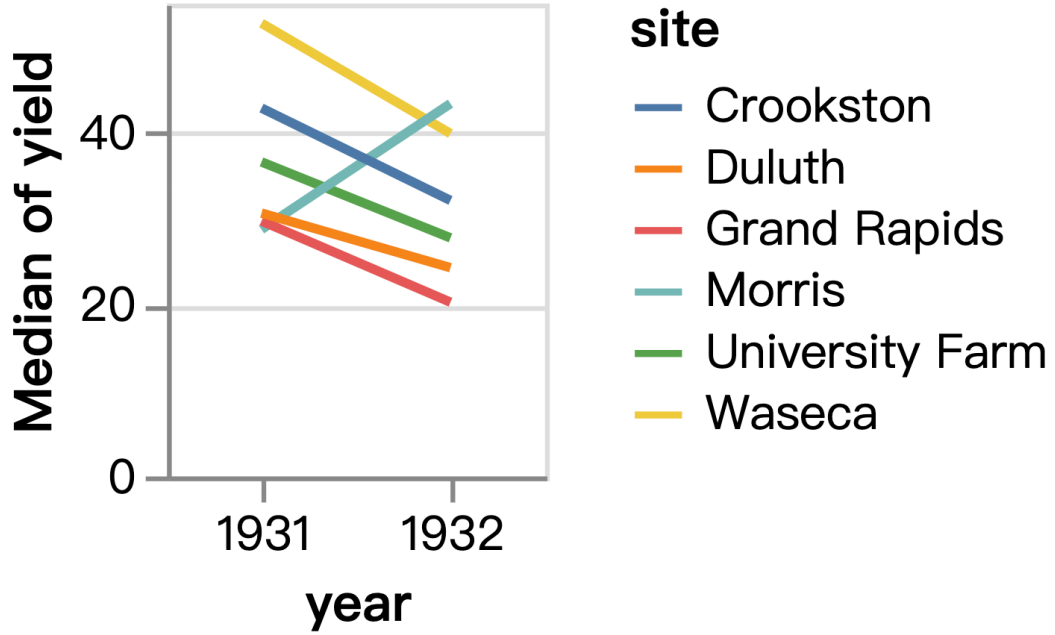
year	site	AVG(yield)
1931	"University Farm"	36.583330000000004
1931	"Waseca"	52.716665000000006
1931	"Morris"	28.733335
1931	"Crookston"	42.85
1931	"Grand Rapids"	29.71667
1931	"Duluth"	30.633335000000002
1932	"University Farm"	27.750005
1932	"Waseca"	39.883335
1932	"Morris"	43.36667
1932	"Crookston"	32.099995

```
"mark": "line"
"x": "year"
"y": "AVG(yield)"
"color":  "site"
```



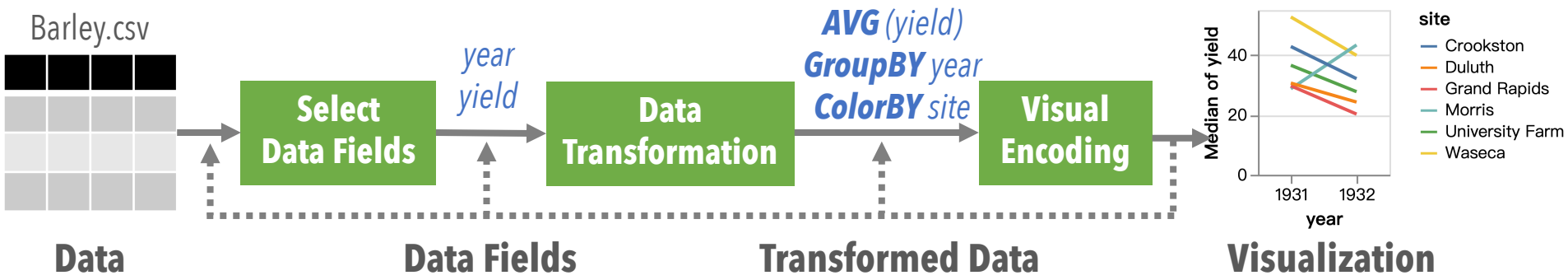
How to effectively visualize data?





Creating effective visualizations is hard!

How to effectively visualize data?



Challenge 1. Creating effective visualizations is **iterative and error-prone**

Try right combination of data fields
Try right data transformation
Try right visual encoding

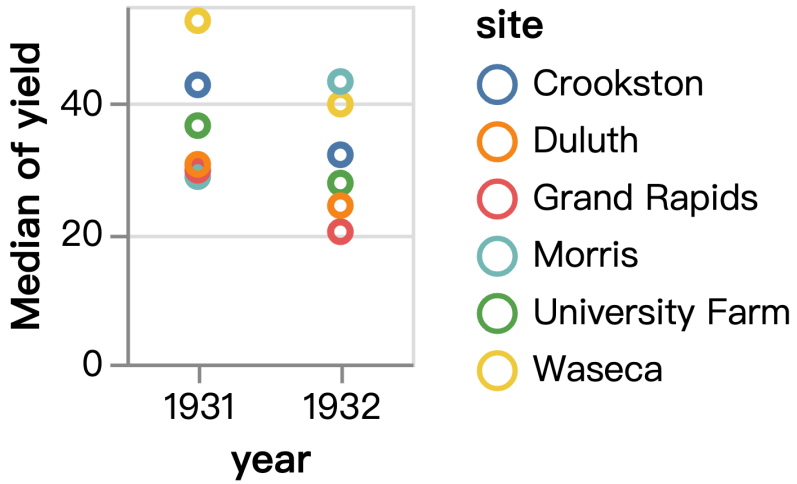
→ ***Huge search space*** for visualizations
Time-consuming user interaction

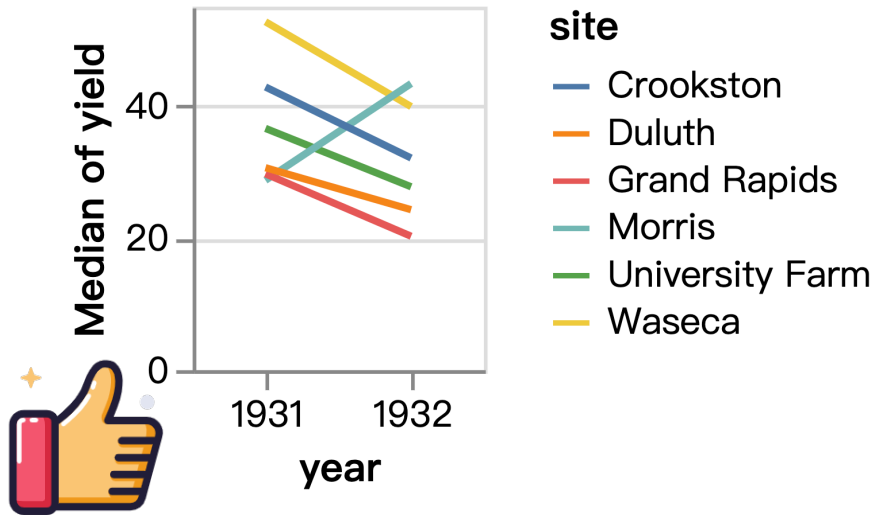
Challenge 1. Creating effective visualizations is **iterative and error-prone**

Try right combination of data fields
Try different data transformation
Try different visual encoding

→ ***Huge search space*** for visualizations
Time-consuming user interaction

Challenge 2. It **requires** *domain- and data-specific* **expertise**





Better interpreting the trend of the data.

Challenge 2. It **requires** *domain- and data-specific* **expertise**

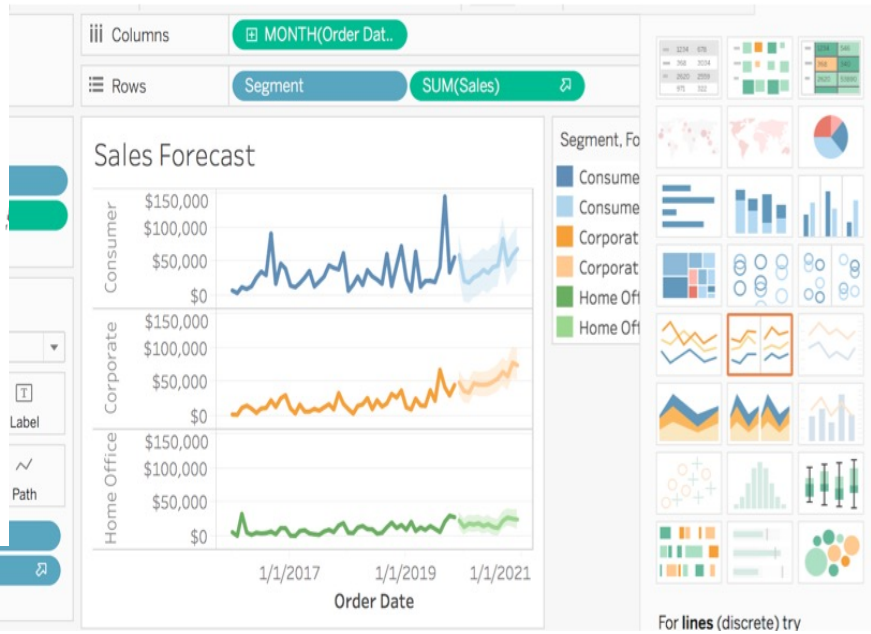
```
{
 "background": "white",
 "height": 150,
 "data": [
    {
      "name": "source_0",
      "url": "data/cars.json",
      "format": {"type": "json"},
      "transform": [
        {
          "type": "aggregate",
          "groupby": ["Origin"],
          "ops": ["count"],
          "fields": [null],
          "as": ["__count"]
        }
      ]
    }
  ],

 "marks": [
     {
       "name": "marks",
       "type": "rect",
       "style": ["bar"],
       "from": {"data": "source_0"},
       "encode": {
         "update": {
           "fill": {"value": "#4c78a8"},
           "ariaRoleDescription": {"value": "bar"},
           "x": {"scale": "x", "field": "Origin"},
           "width": {"scale": "x", "band": 1},
           "y": {"scale": "y", "field": "__count"},
           "y2": {"scale": "y", "value": 0}
         }
       }
     }
   ],
    ...
 }
```

Expertise on

ming-based visualization script

visualization tools



pandas plotly ggplot2 +ableau Qlik Power BI Excel

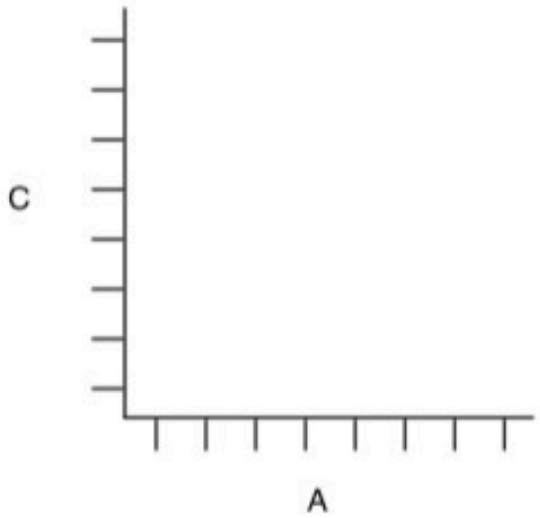
Outline

- Background
- **Visualization Principles**
- Automatic Visualization

Elements of a Plot

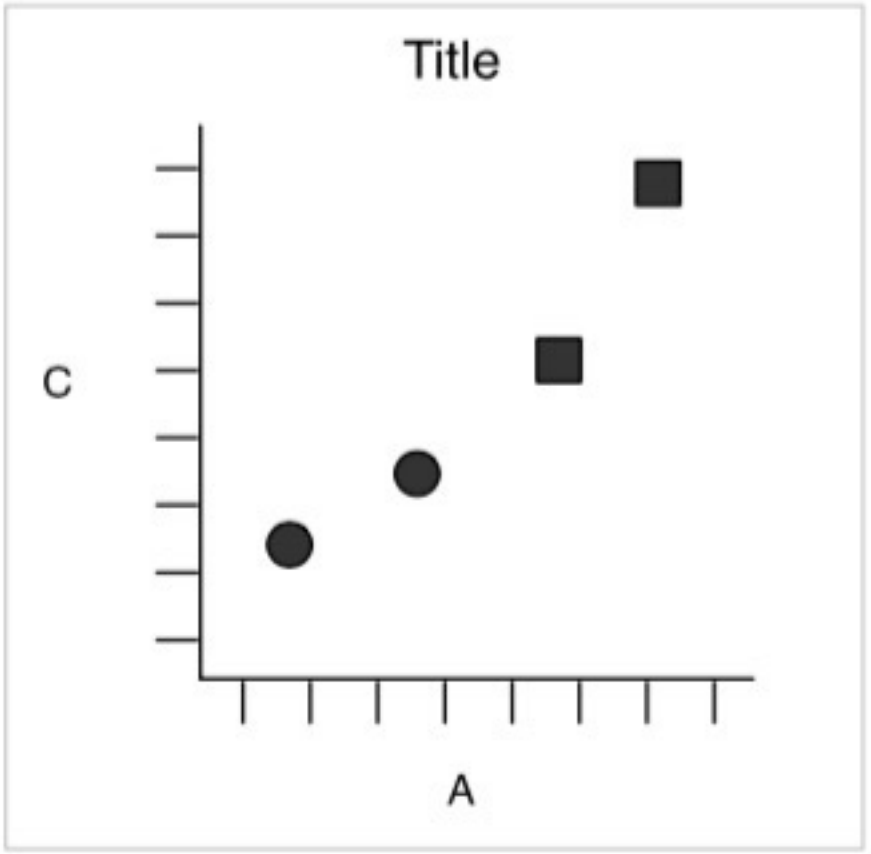
Geometric Objects

Scales & Coordinates

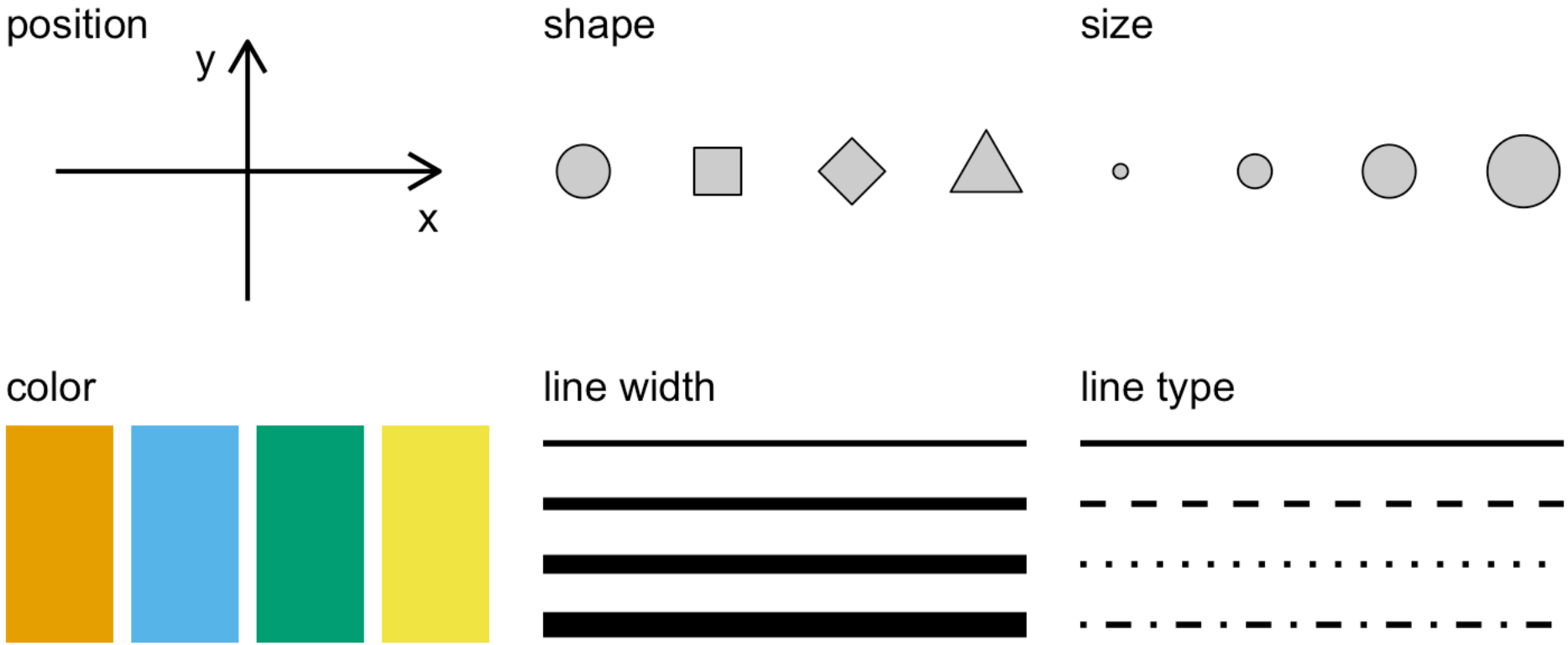


Annotations

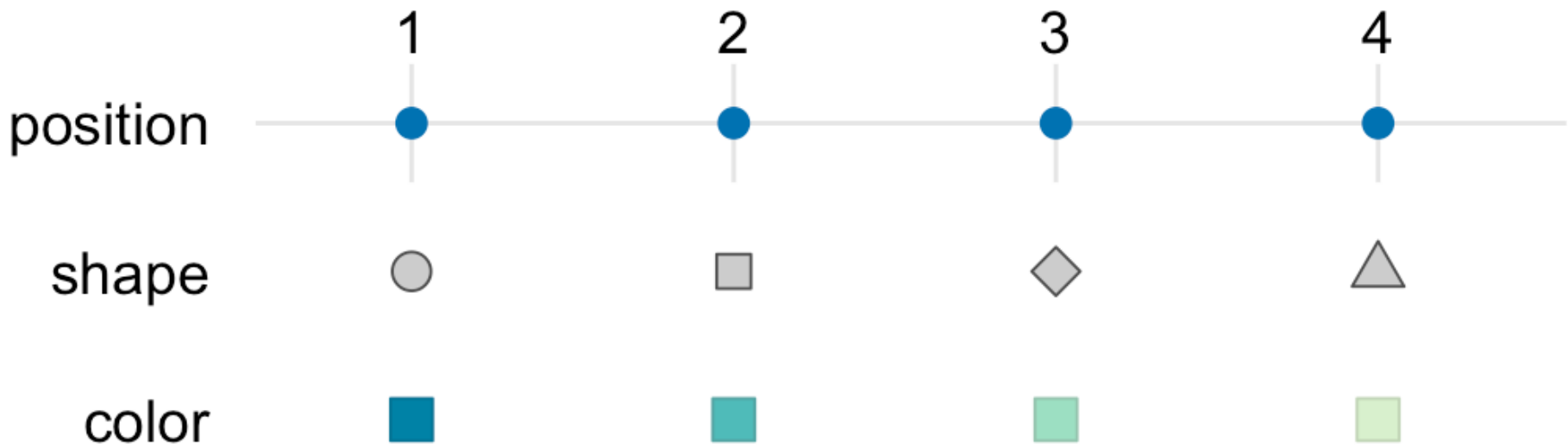
Elements of a Plot



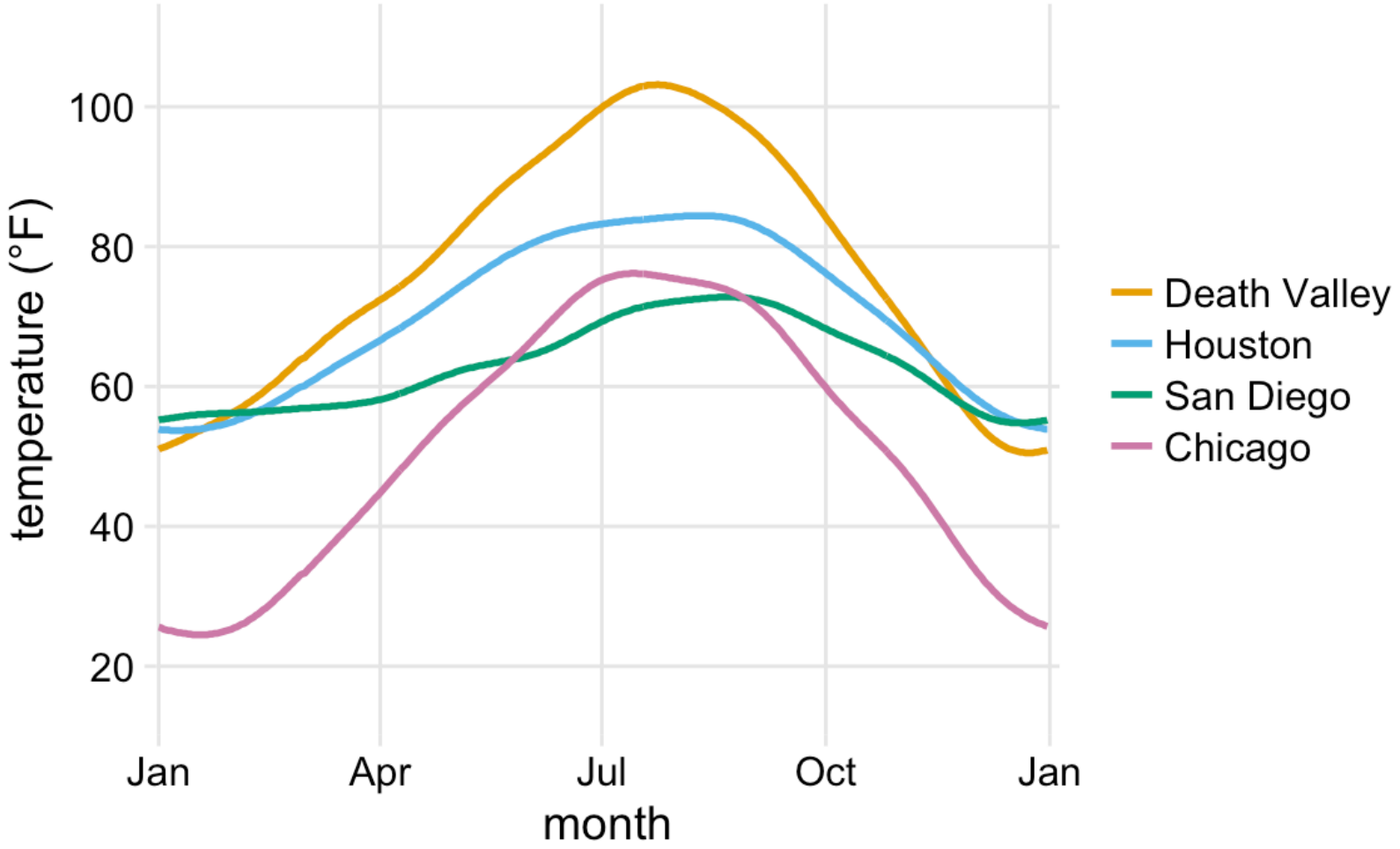
Aesthetics of a Plot



Aesthetics Map Data to Visual Representation

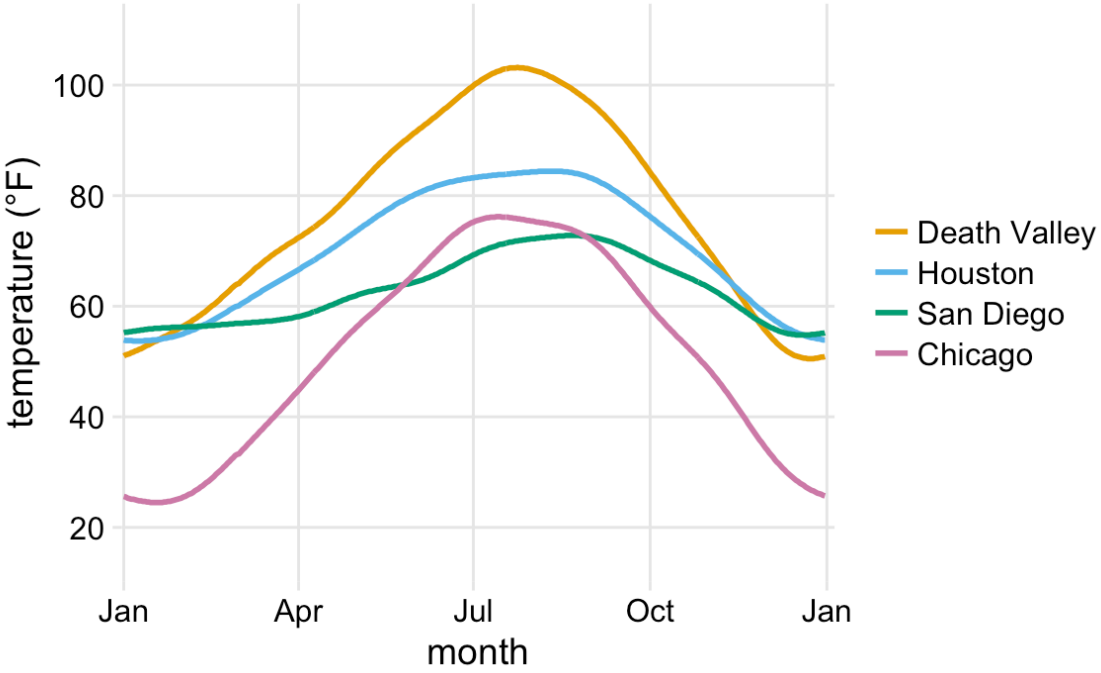


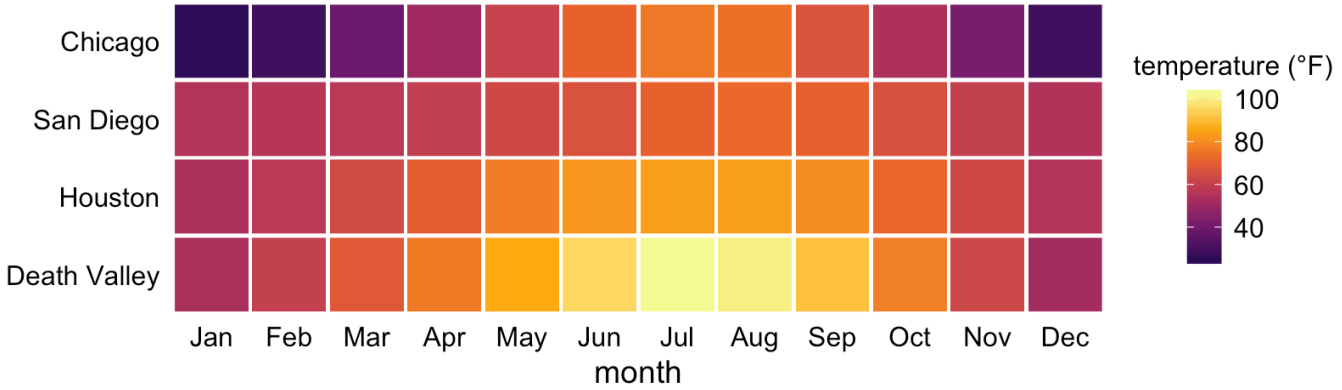
Aesthetics Map Data to Visual Representation



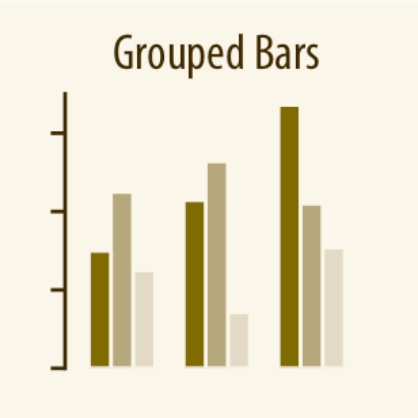
Aesthetics Map Data to Visual Representation

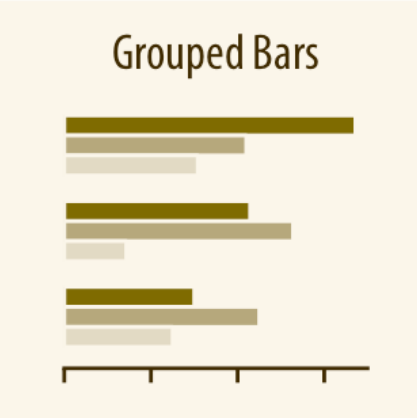
The same data, different ways for visual encoding

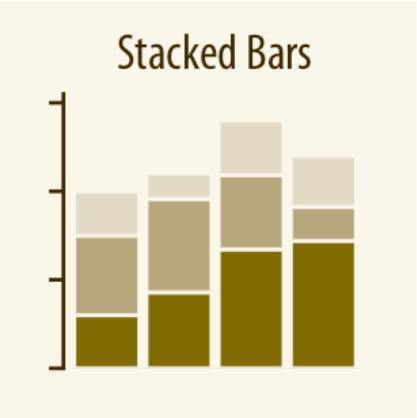




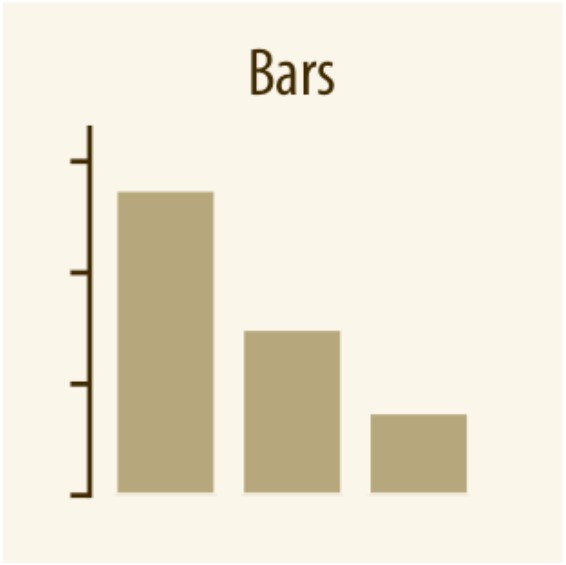
Types of Visualizations: Amounts

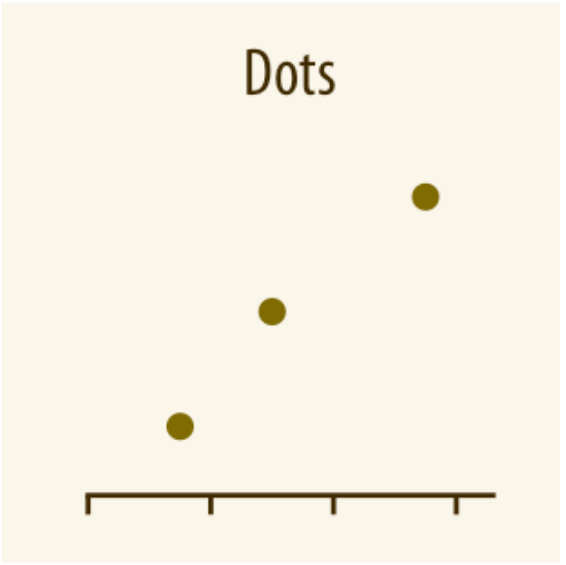




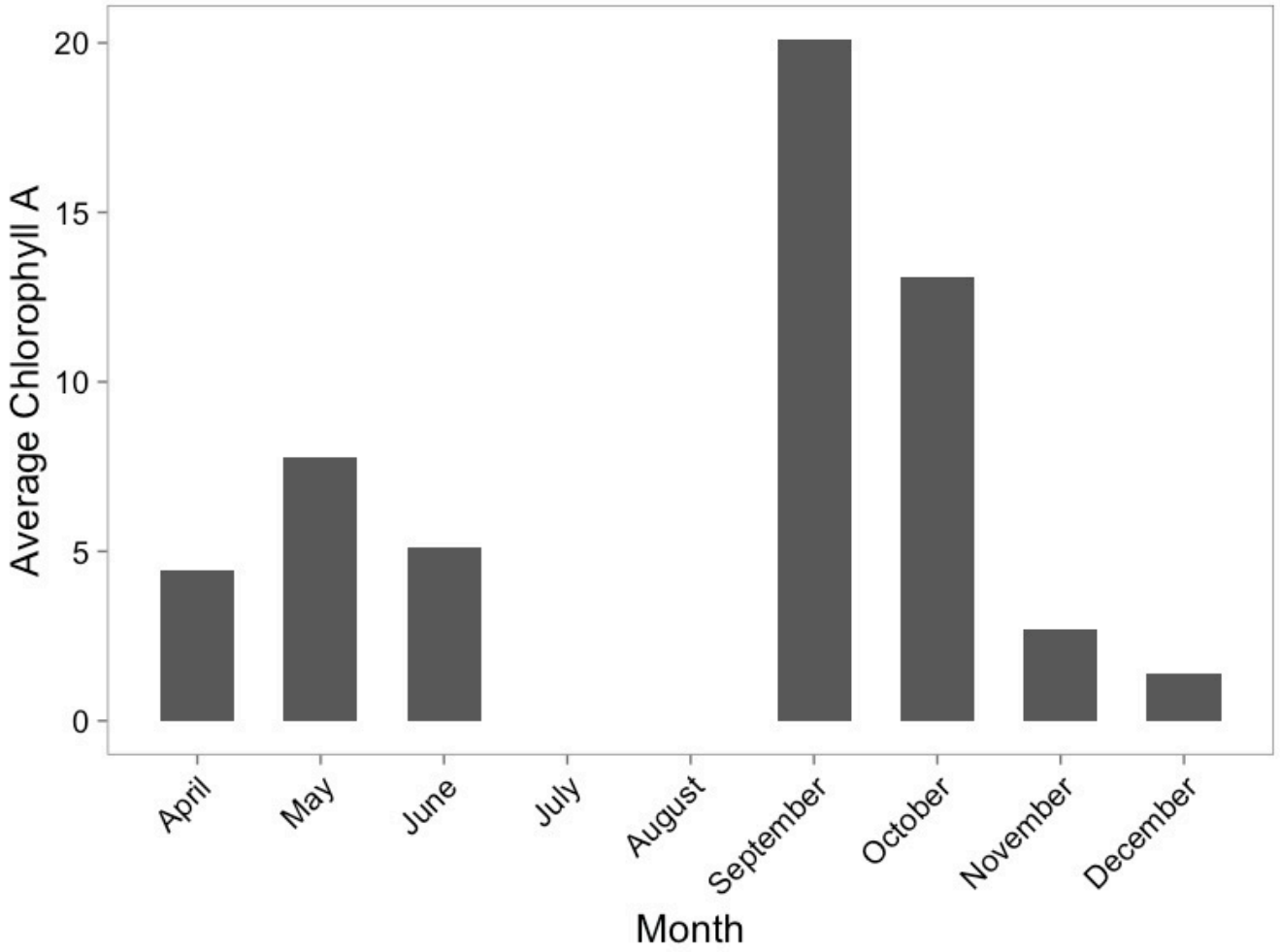


Types of Visualizations
for Amounts, and Comparison

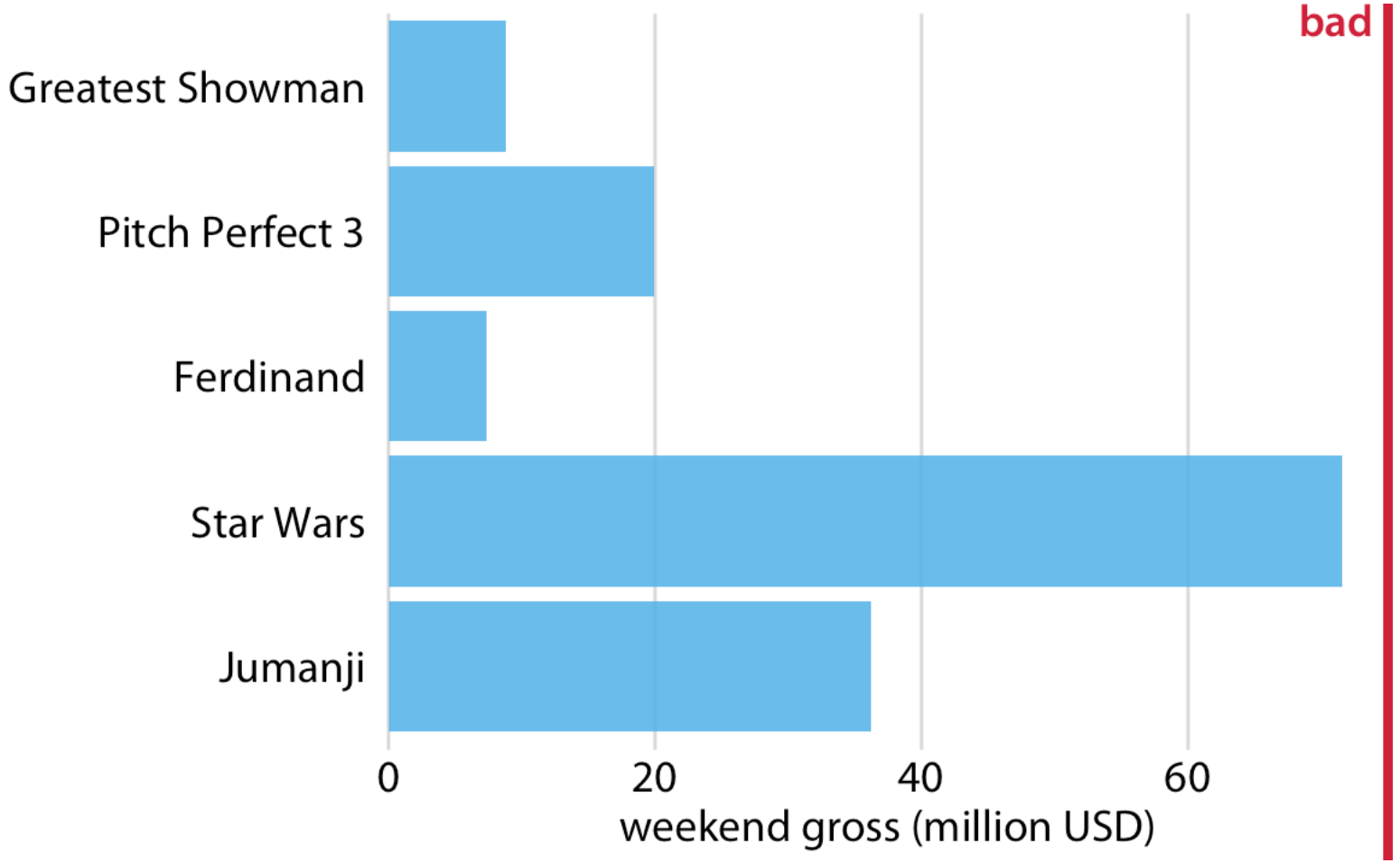




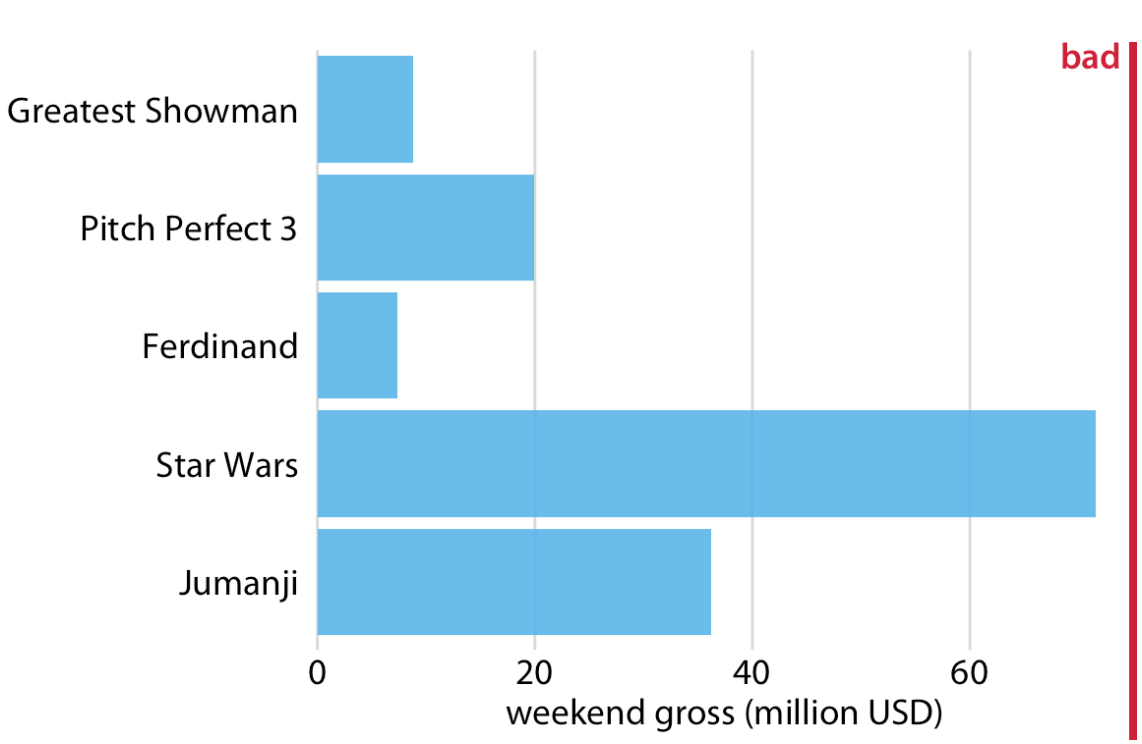
Barplots

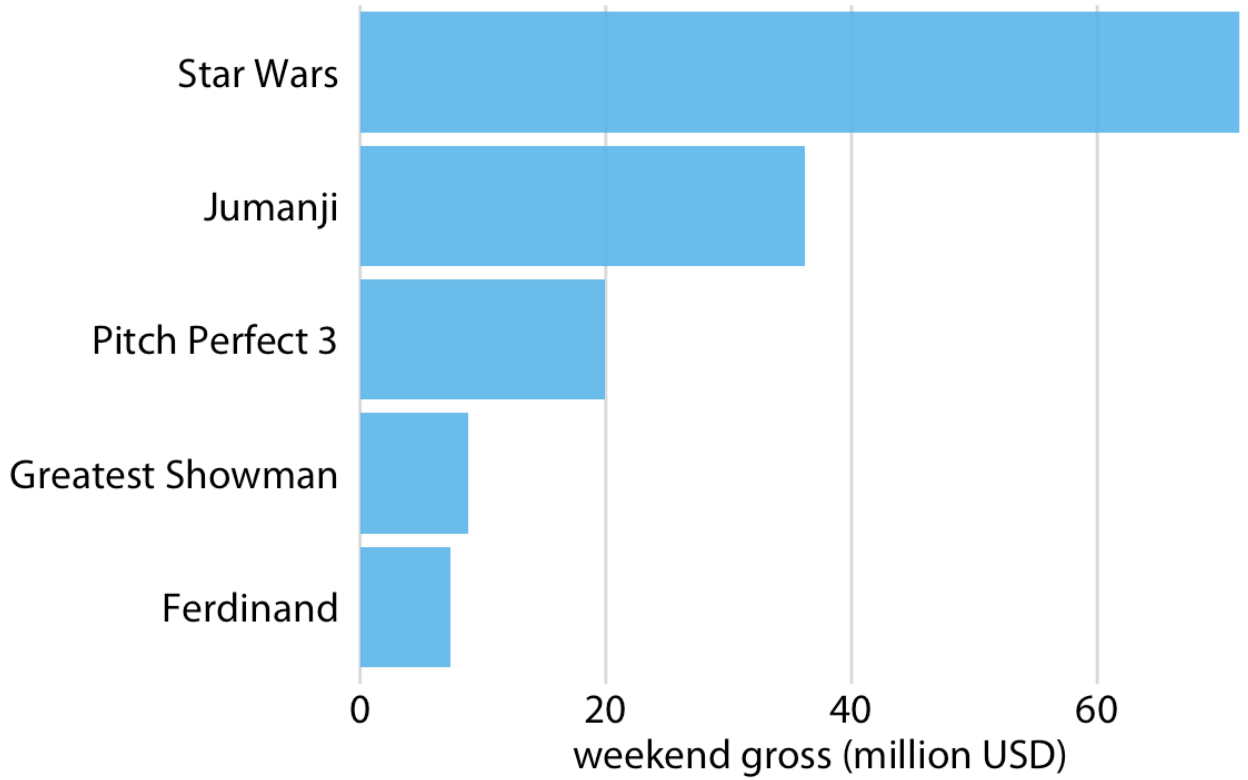

Average Chlorophyll A
20
15
10
5
0
April
May
June
July
August
September
October
November
December
Month

What is Wrong Here?

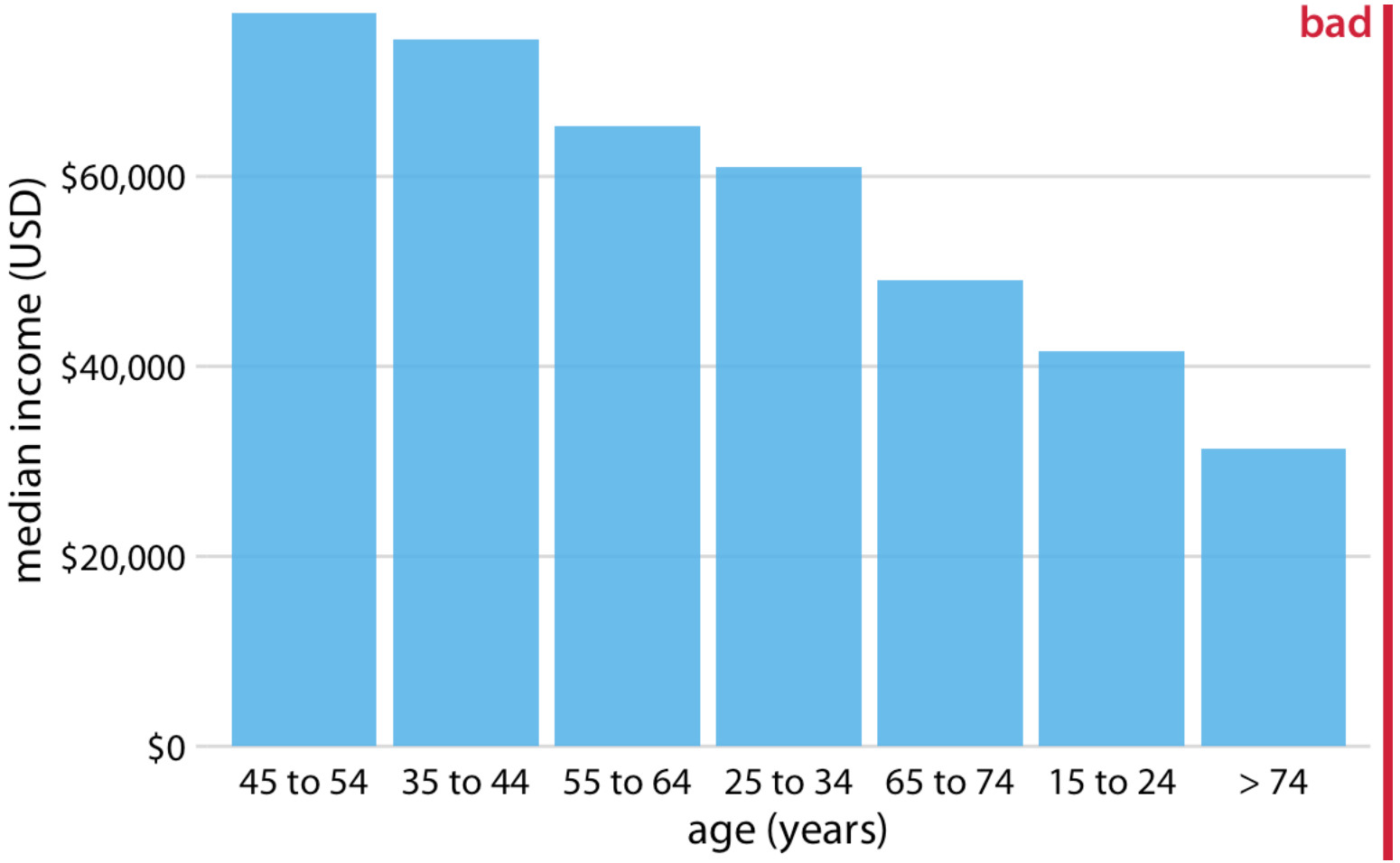


What is Wrong Here?

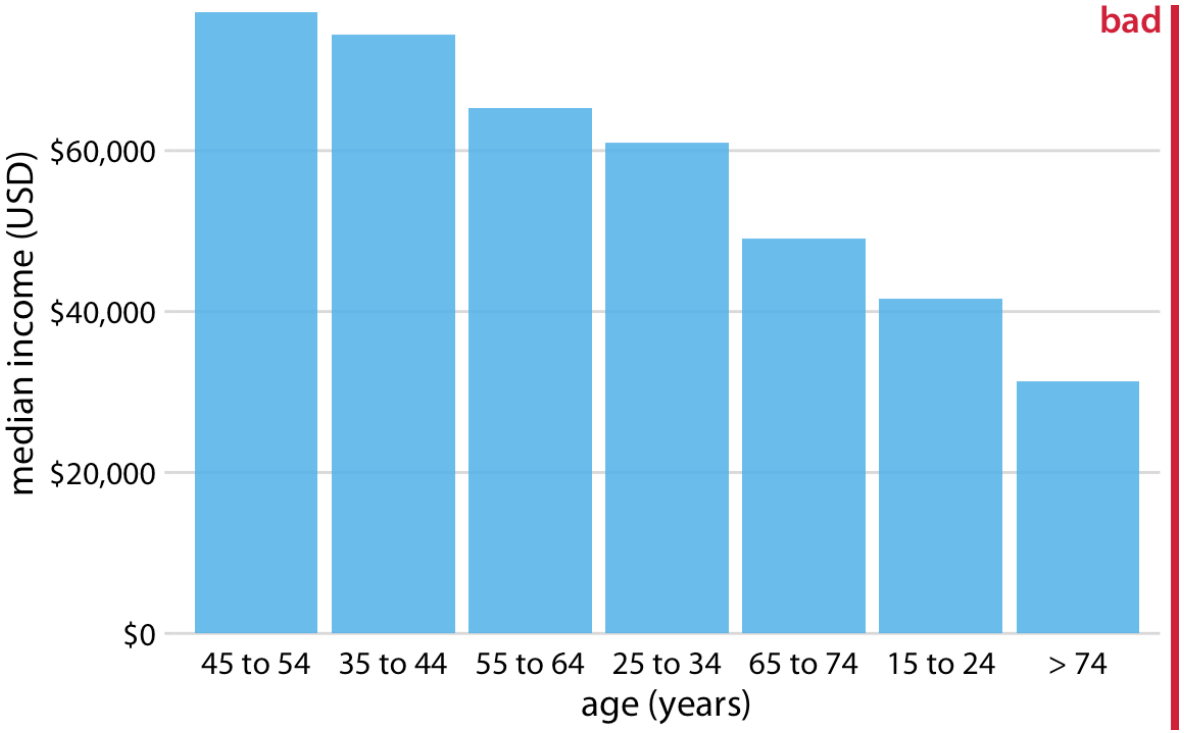


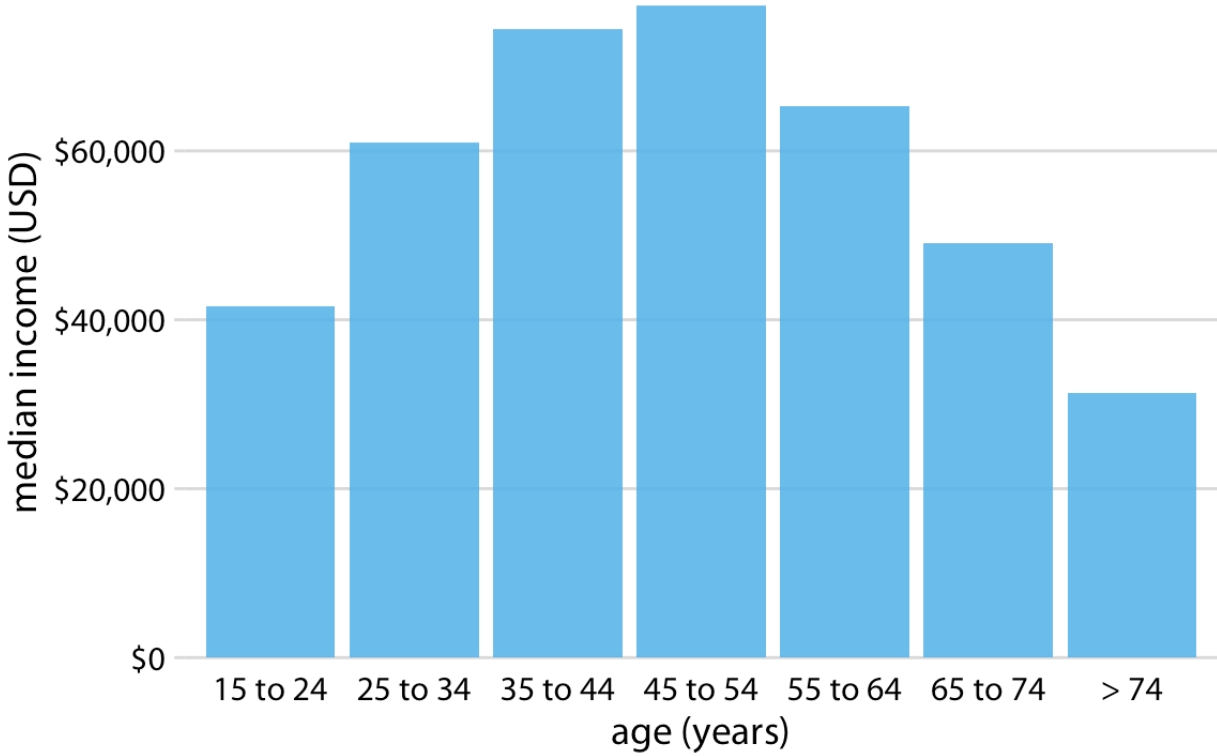


What is Wrong Here?

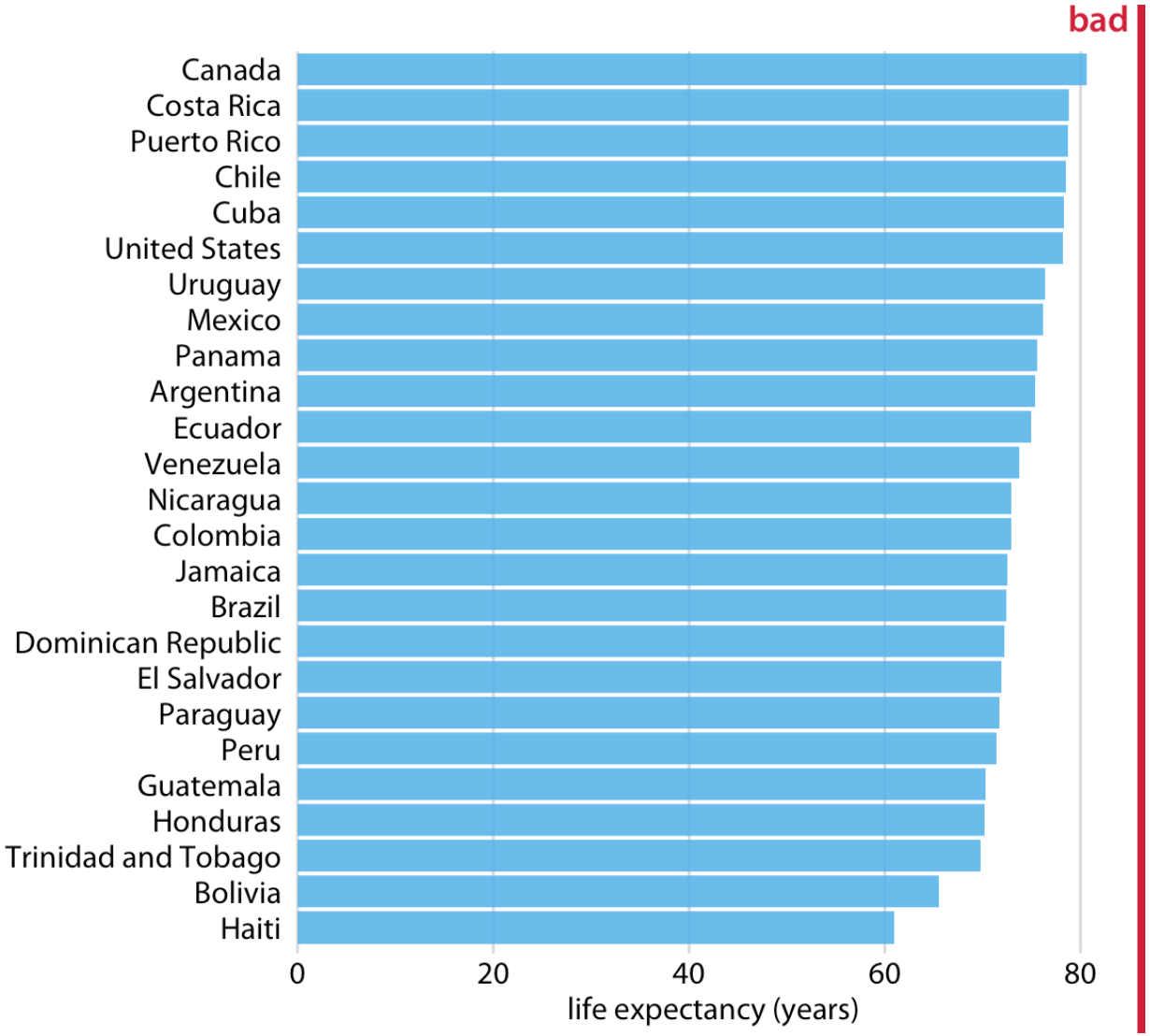


What is Wrong Here?

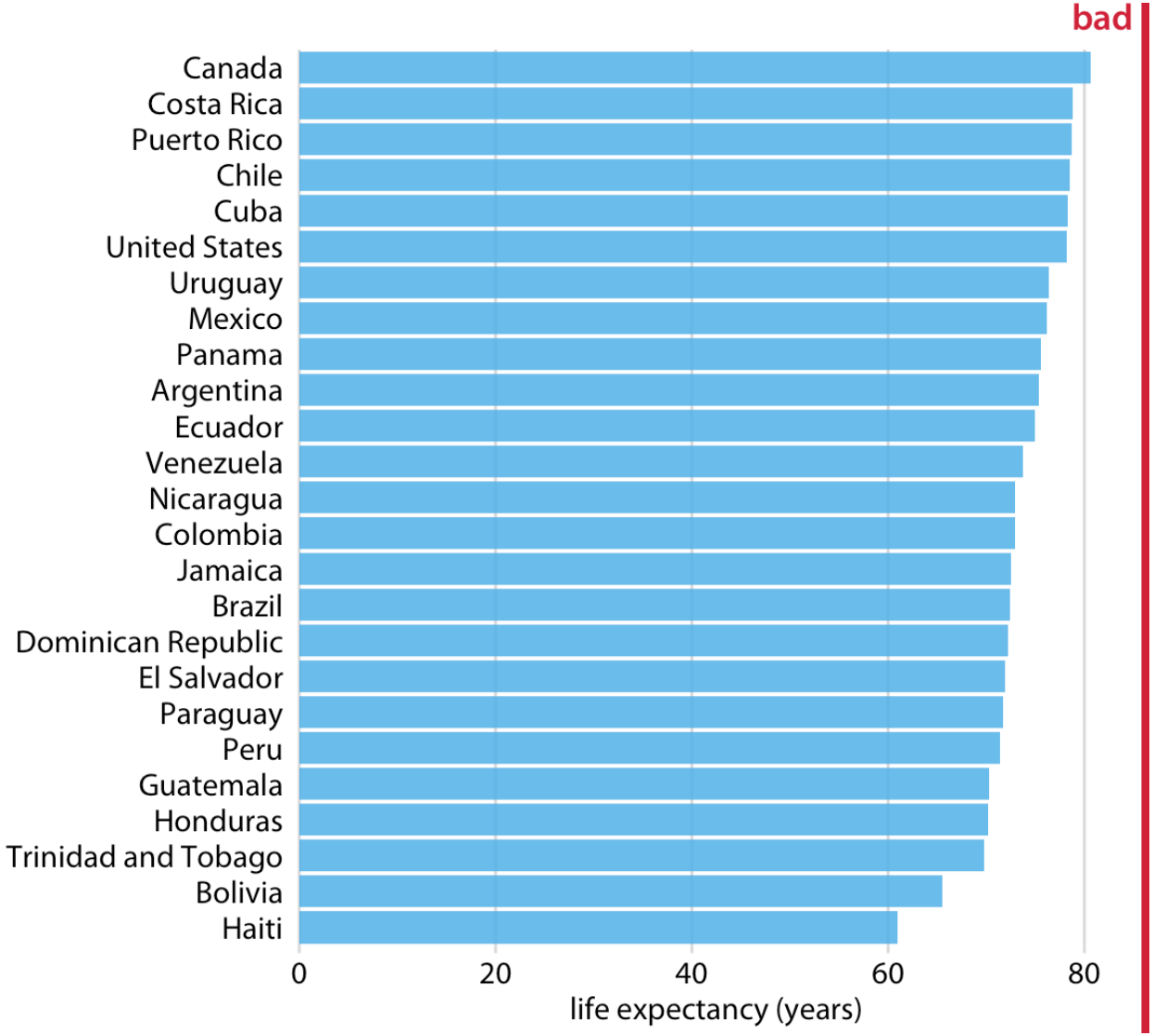


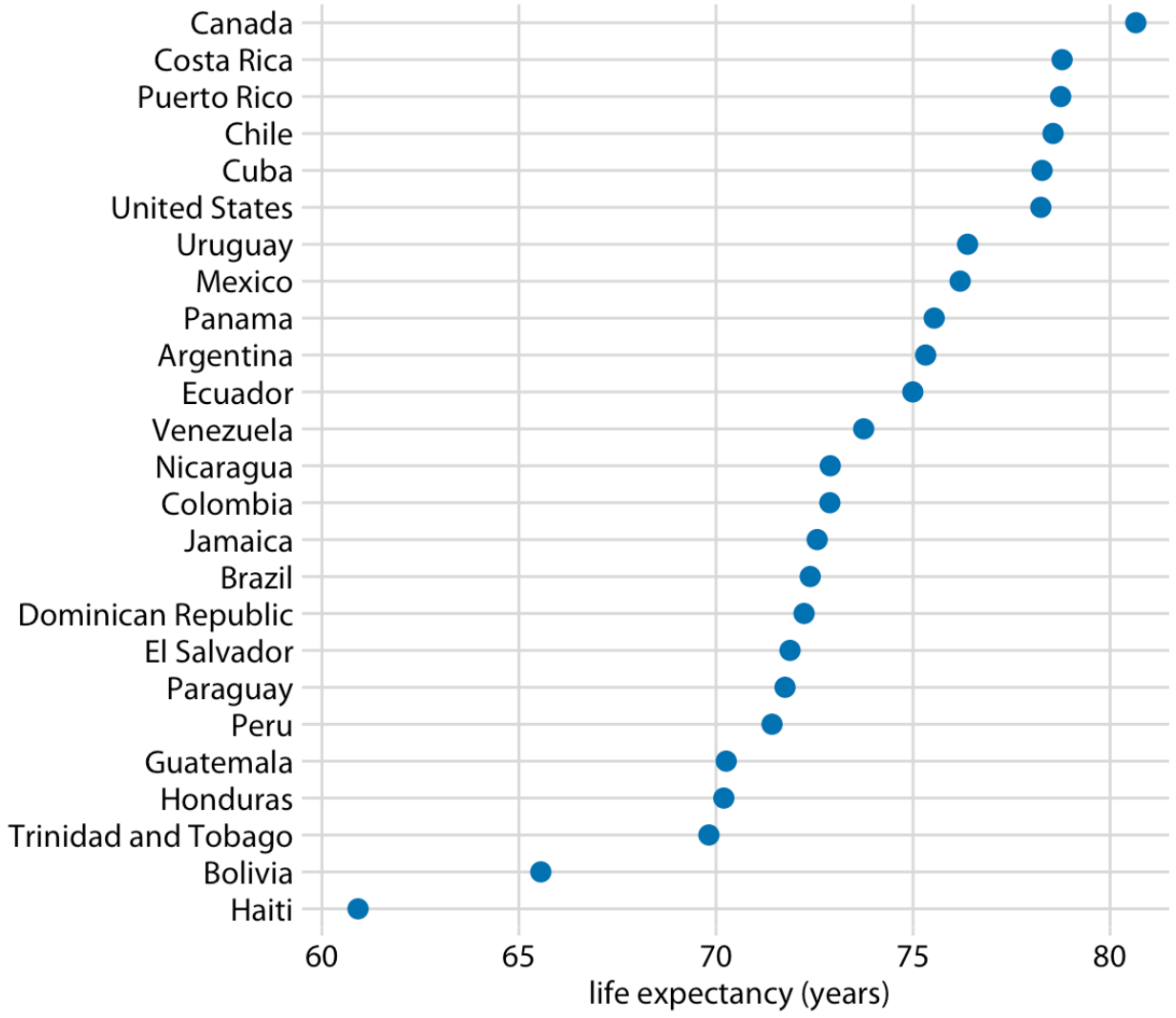


Bars are Not Always Right

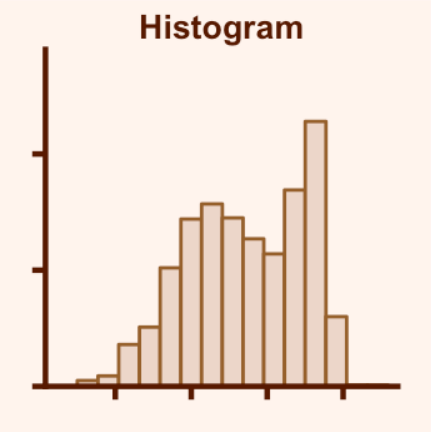


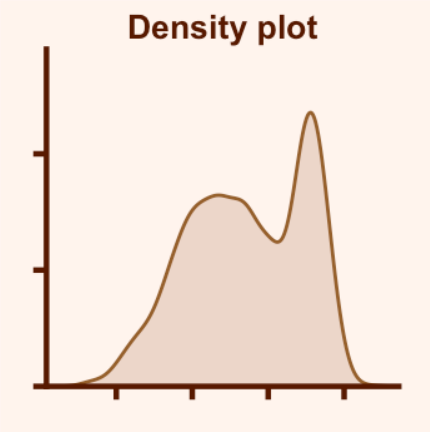
Bars are Not Always Right

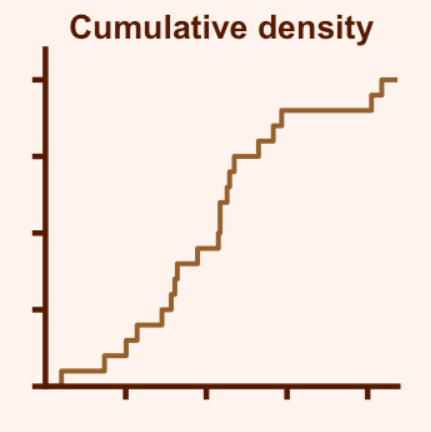


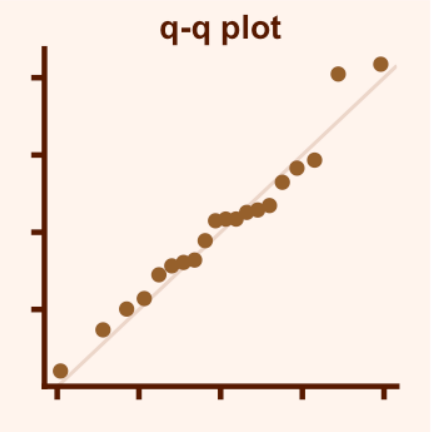


Types of Visualizations *for Distributions*

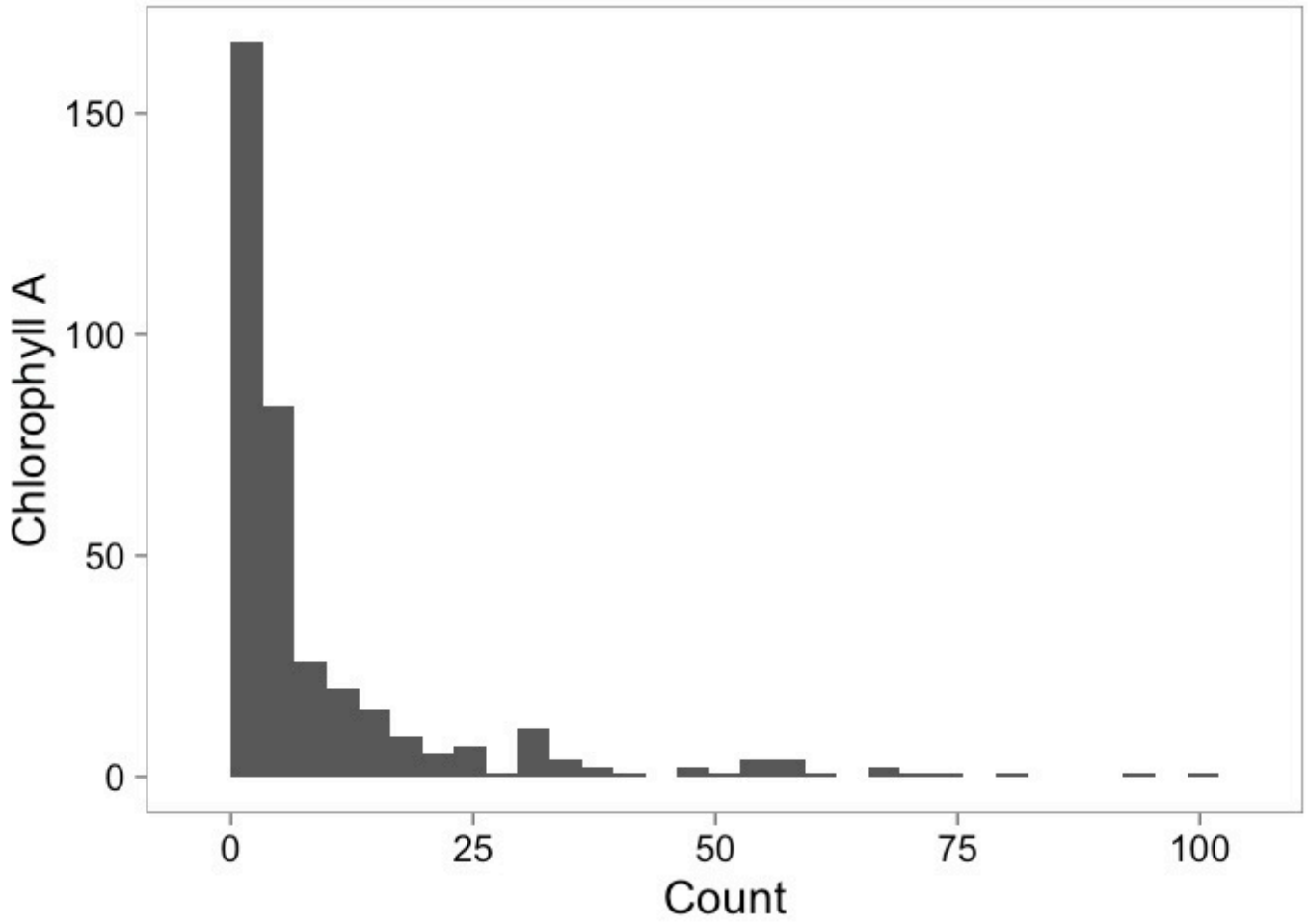




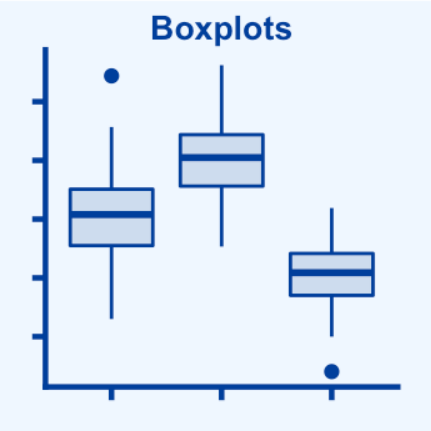


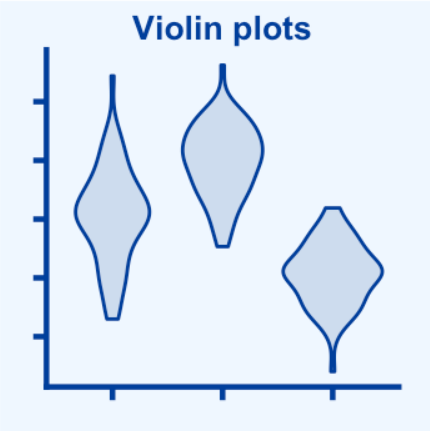


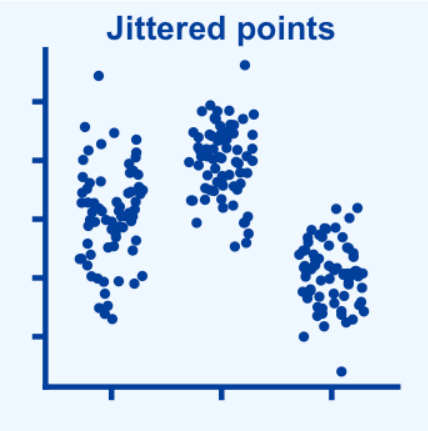
Histograms Show Frequency

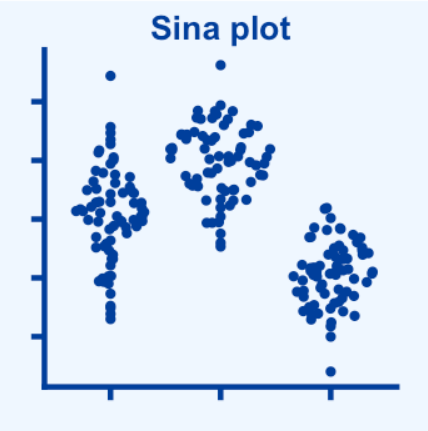


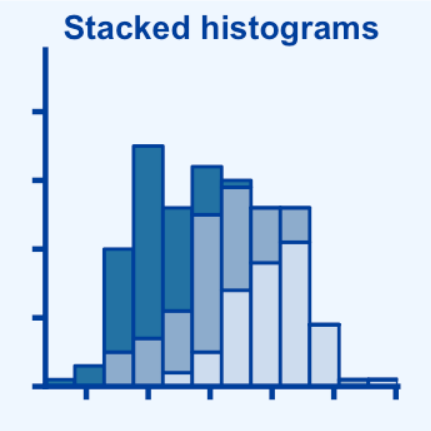
Types of Visualizations: Multiple Distributions

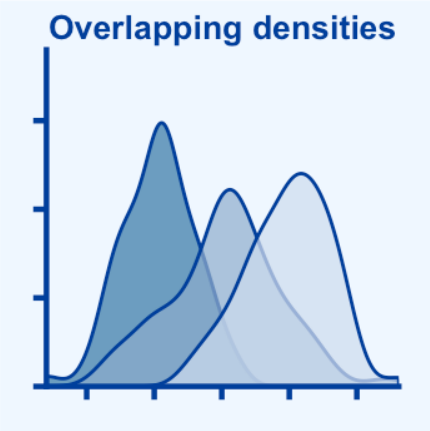


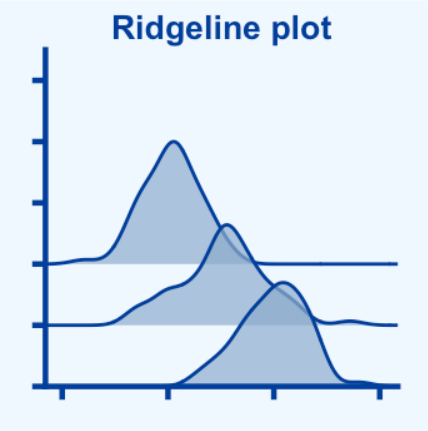




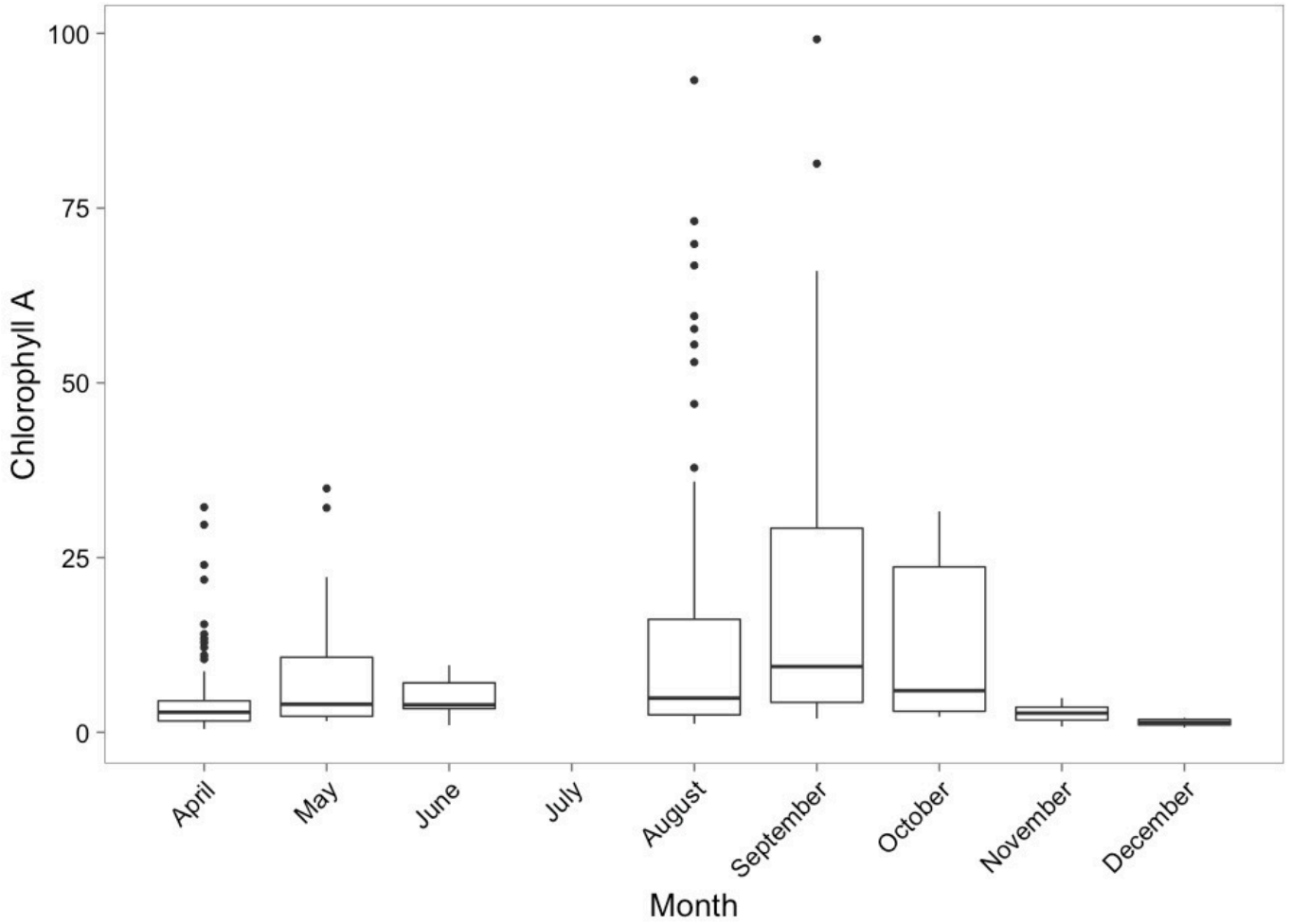




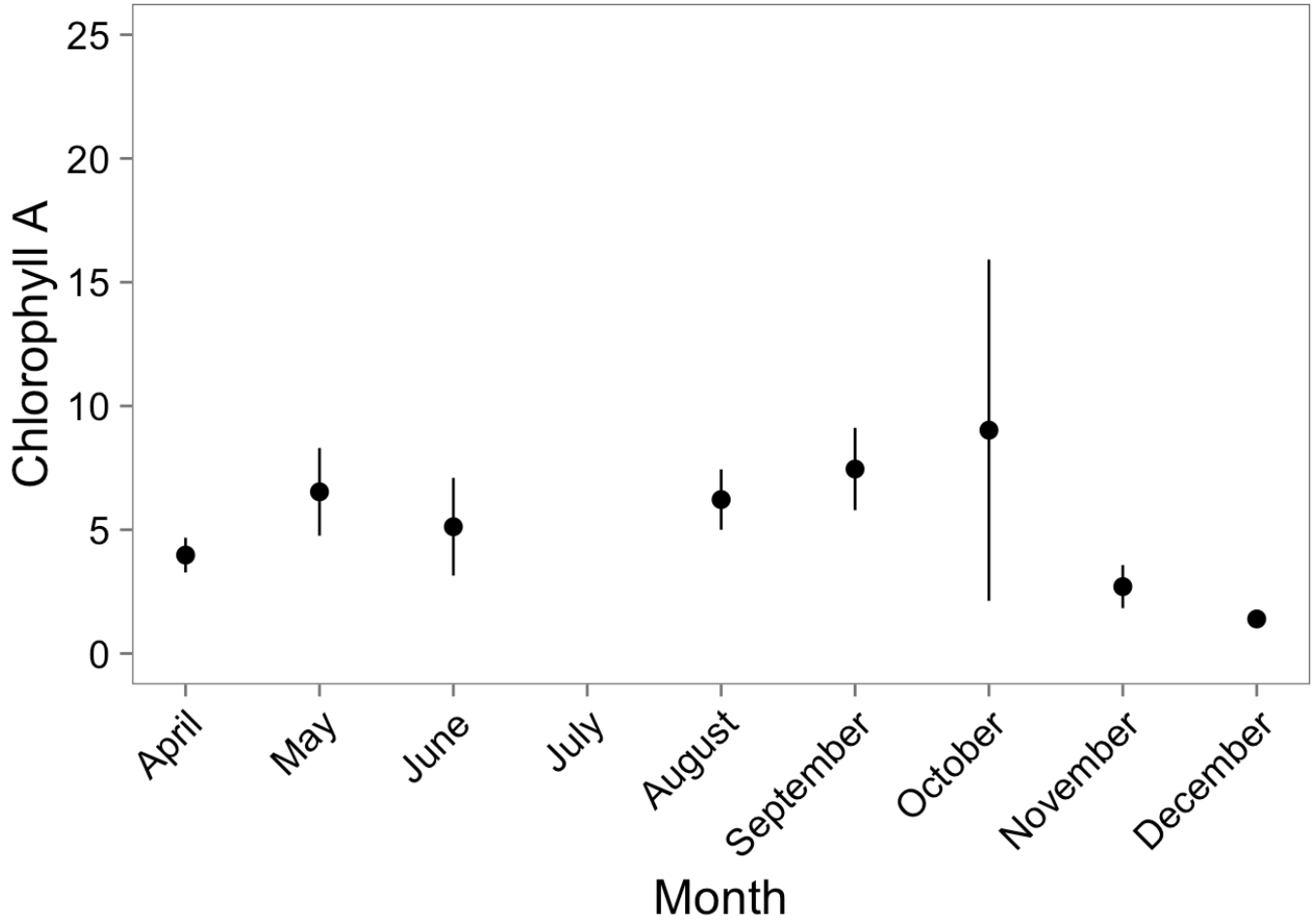




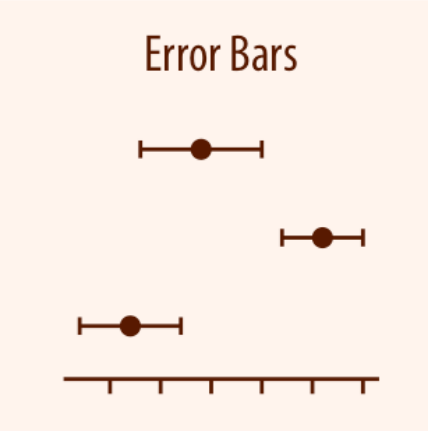
Boxplots to Show Variation

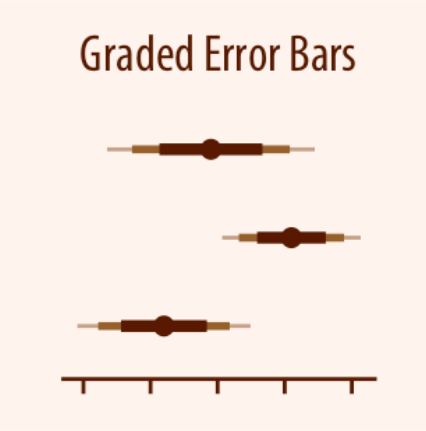


…Or Point-Ranges

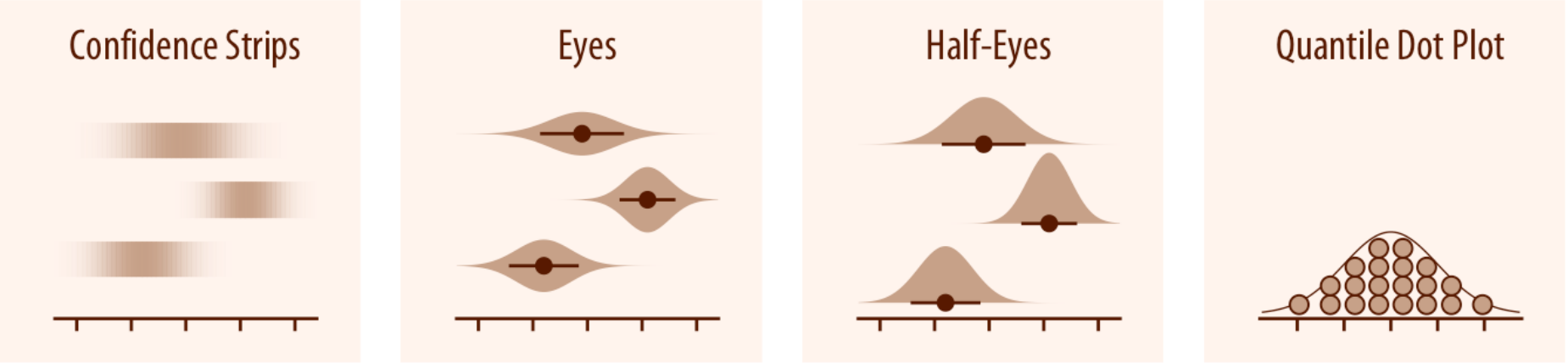


Visualizing Uncertainty is Difficult

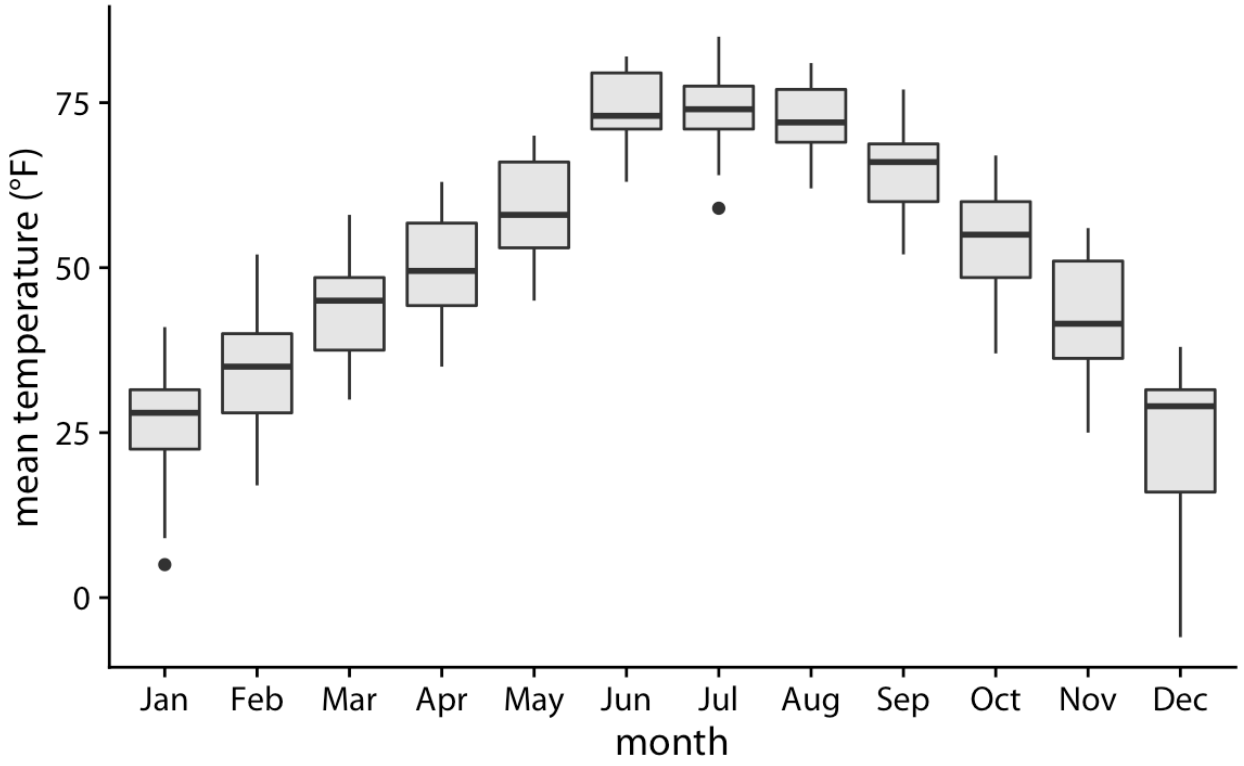


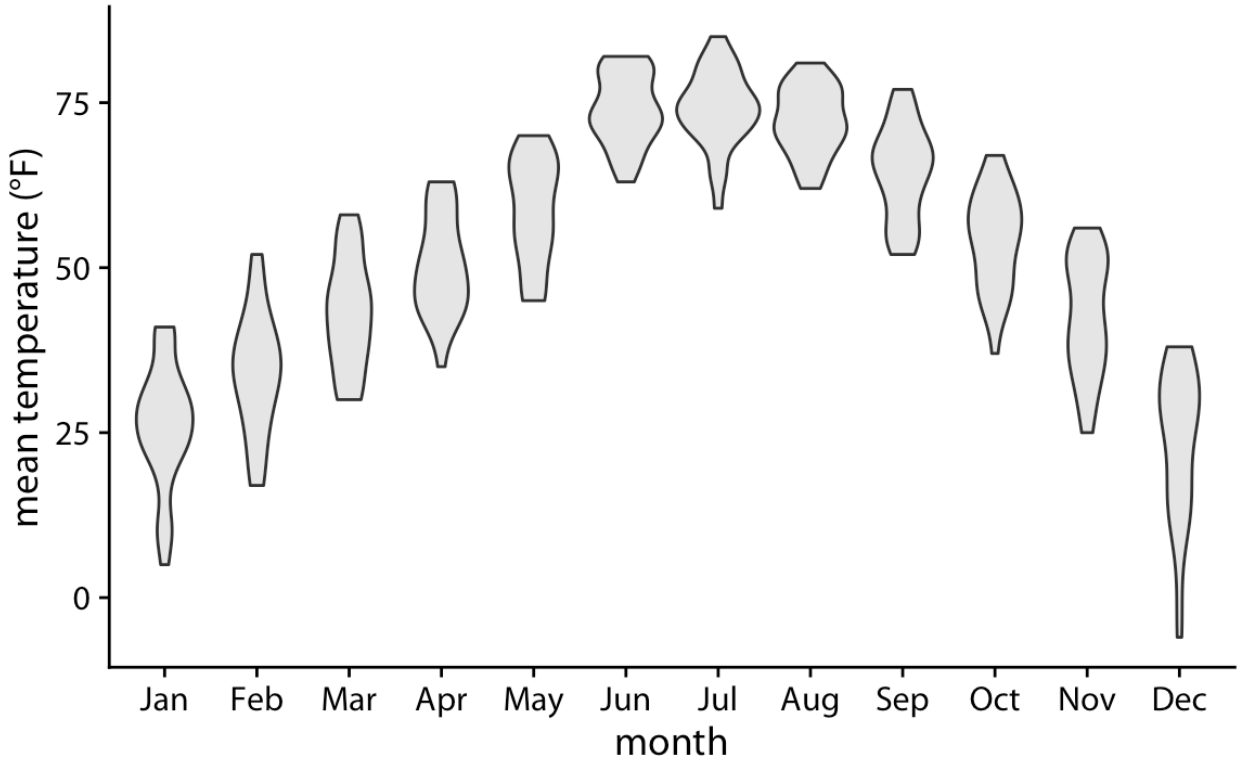


Visualizing Distributions versus Uncertainty

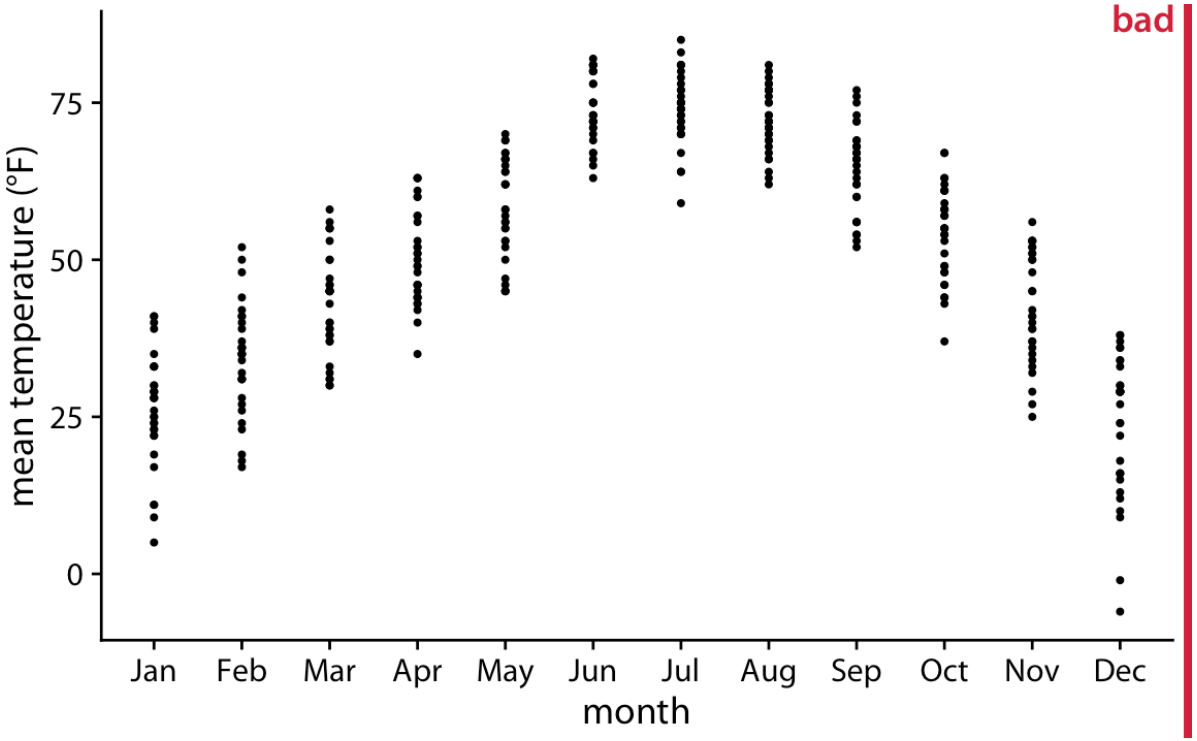


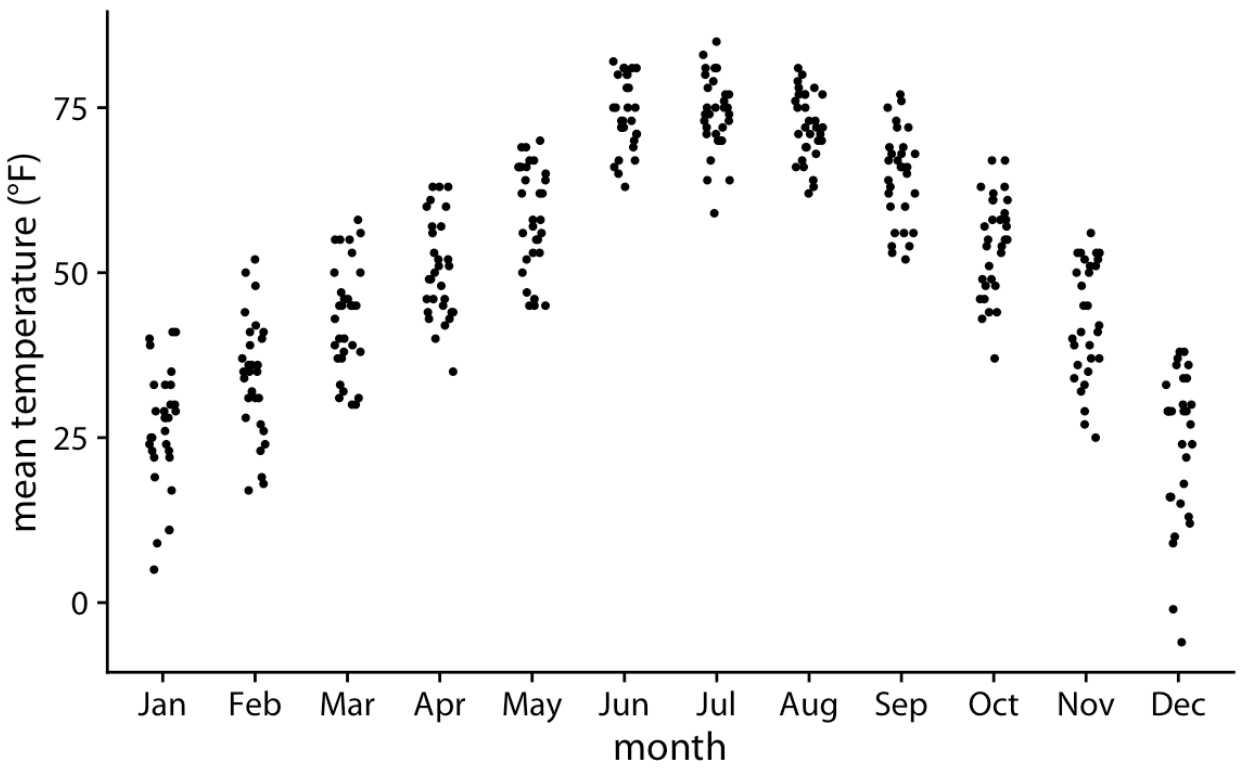
How Does this View Differently?



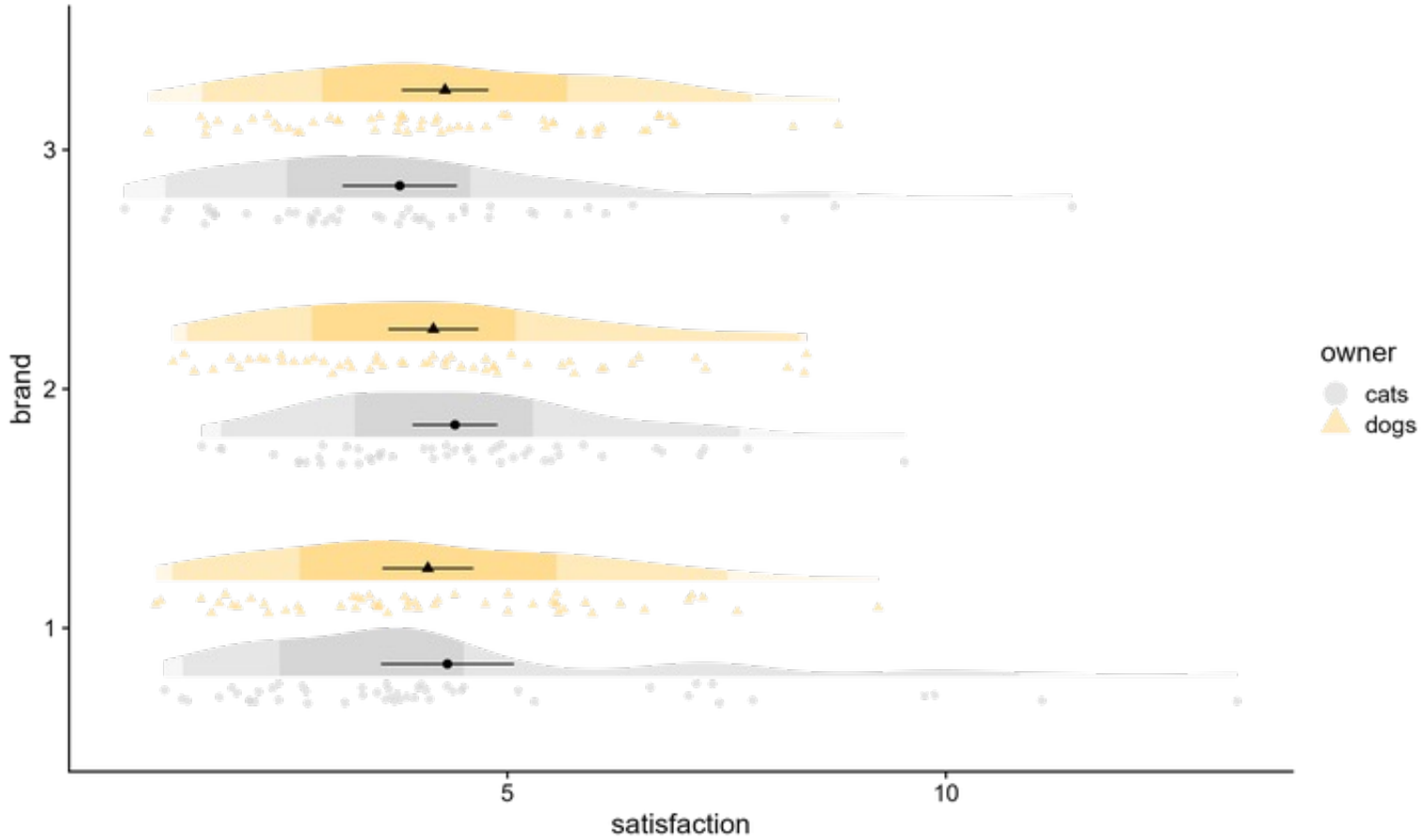


It’s Tempting to Plot Raw Data, but, Be Careful…



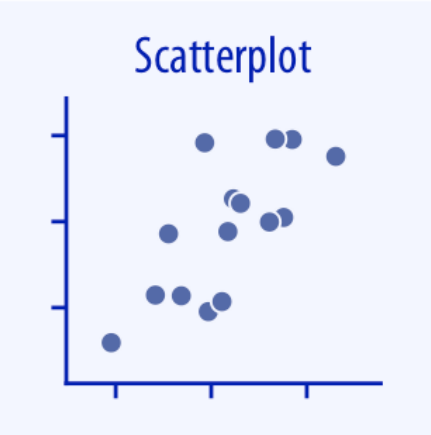


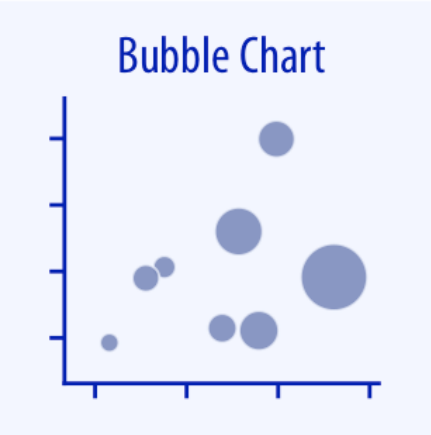
Lots of Ways to Visualize Data, Distribution, and Uncertainty

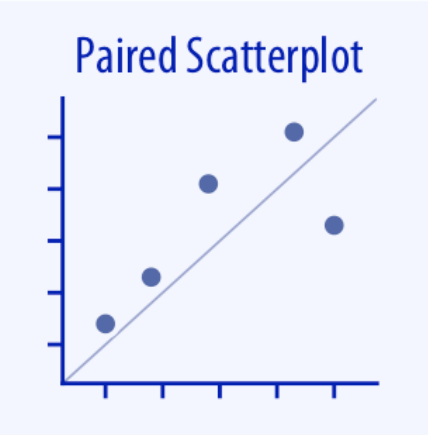


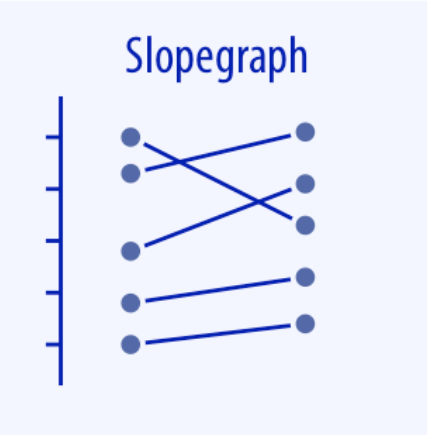
https://dallasnova.rbind.io/post/efficient-data-visualization-with-faded-raincloud-plots-delete-boxplot/
From rweekly.org 9/5 edition https://rweekly.org/2022-W36.html

Relationships Between Variables

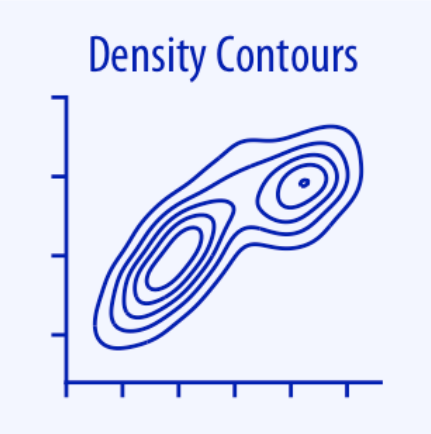


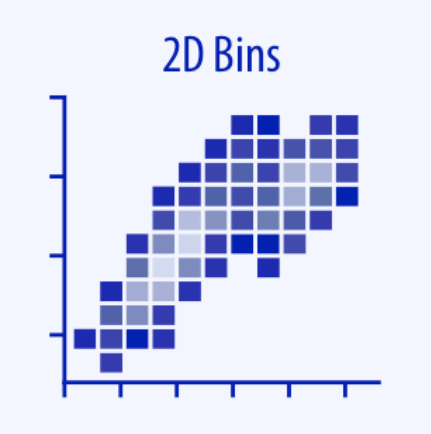


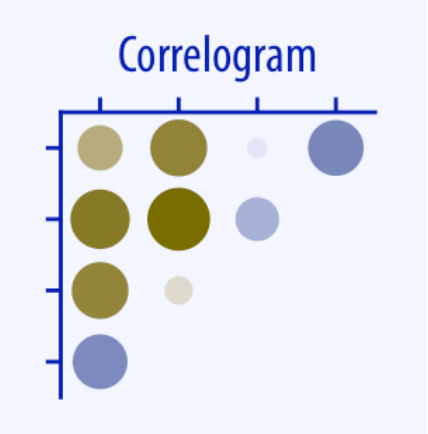




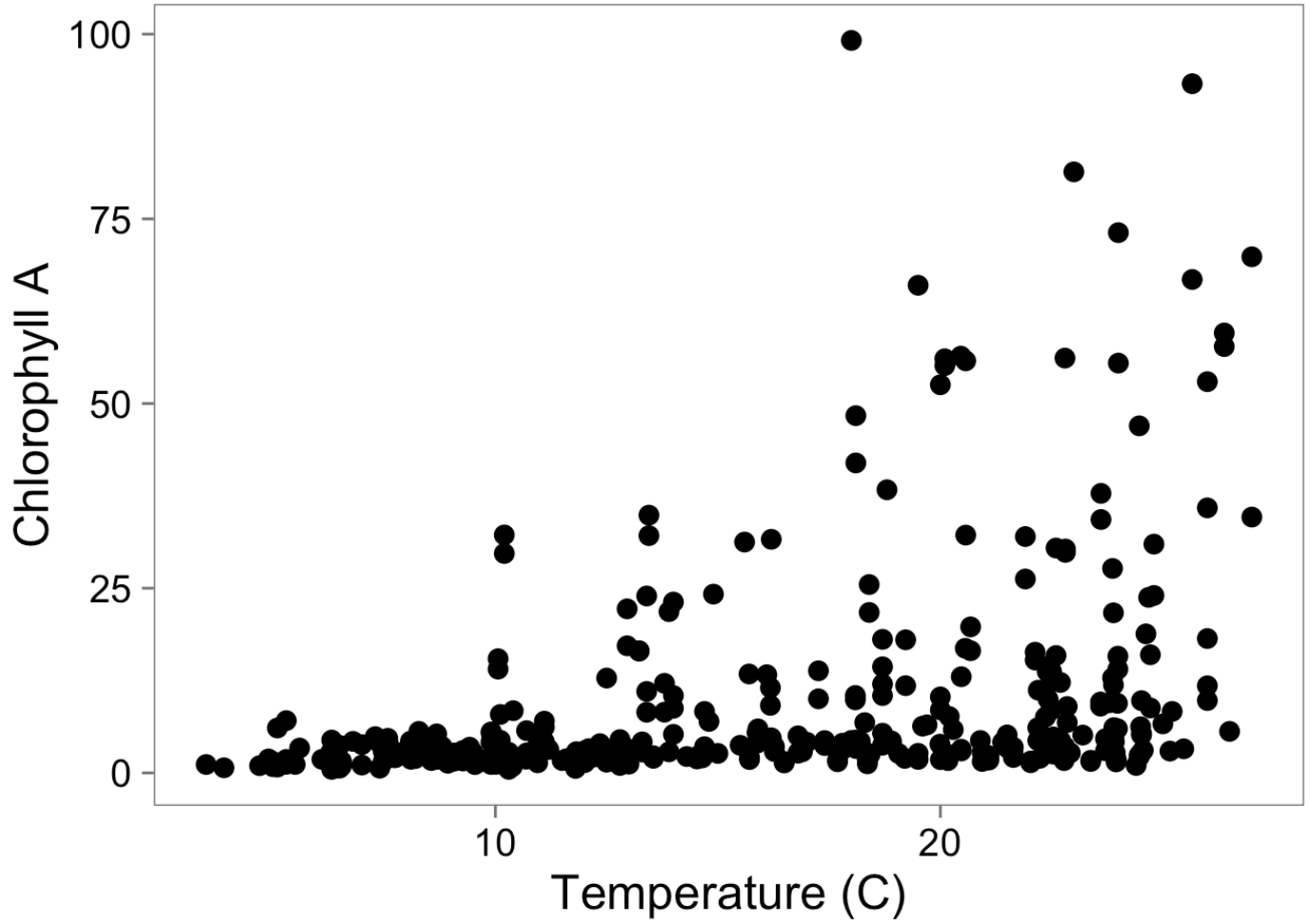
Relationships Between Variables With a Lot of Data



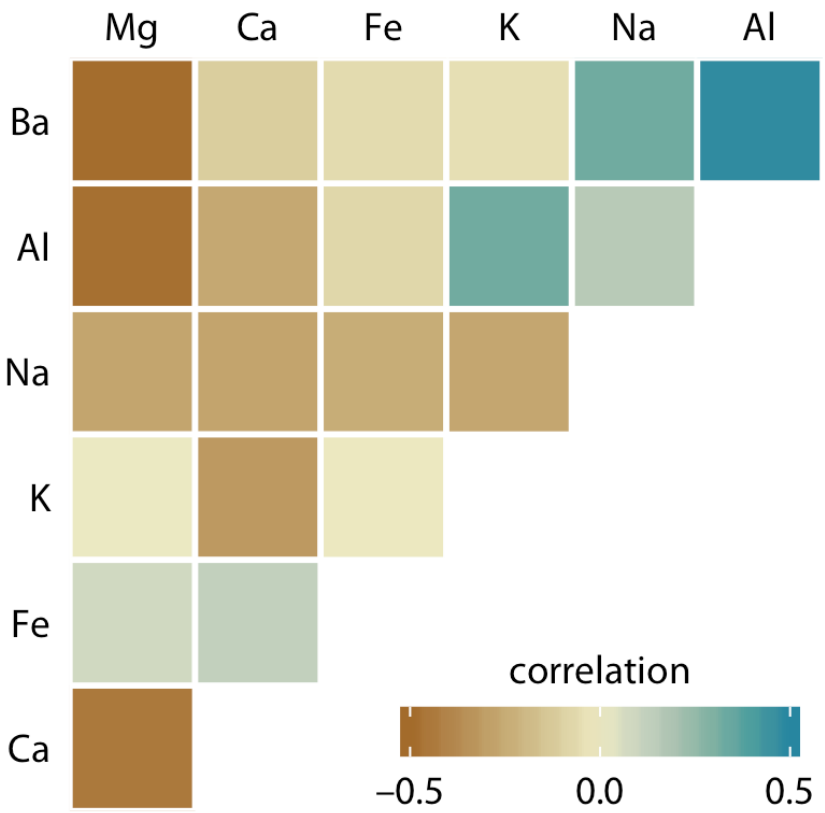


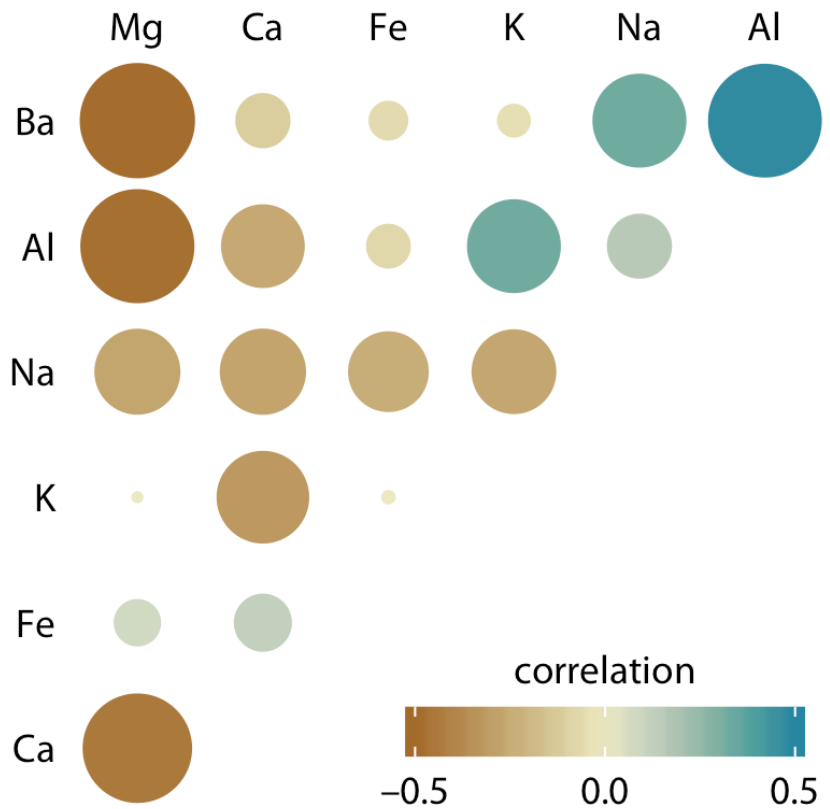


Scatterplots Show Relationships

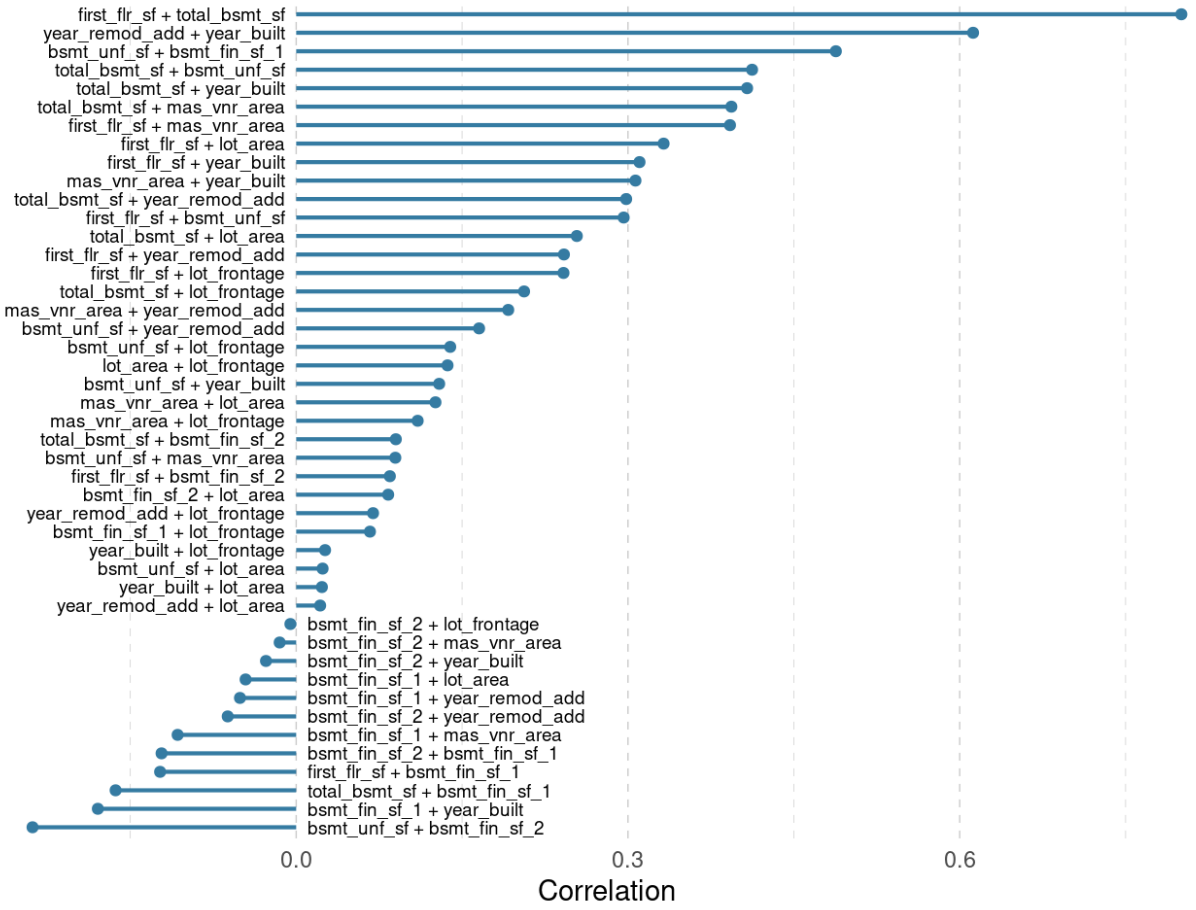


Correlations to Understand Relationships



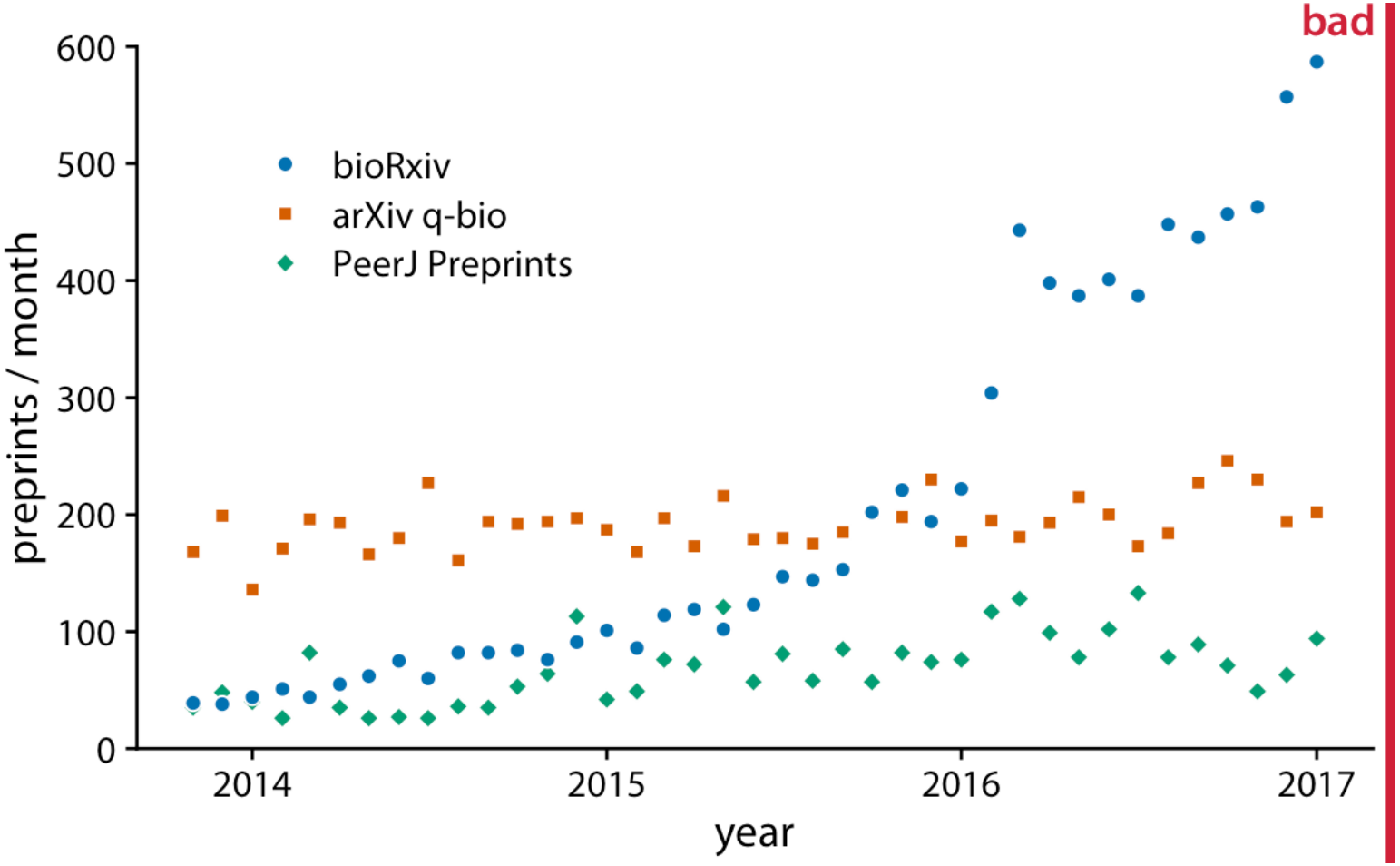


Alternate Ways to View Correlations Efficiently

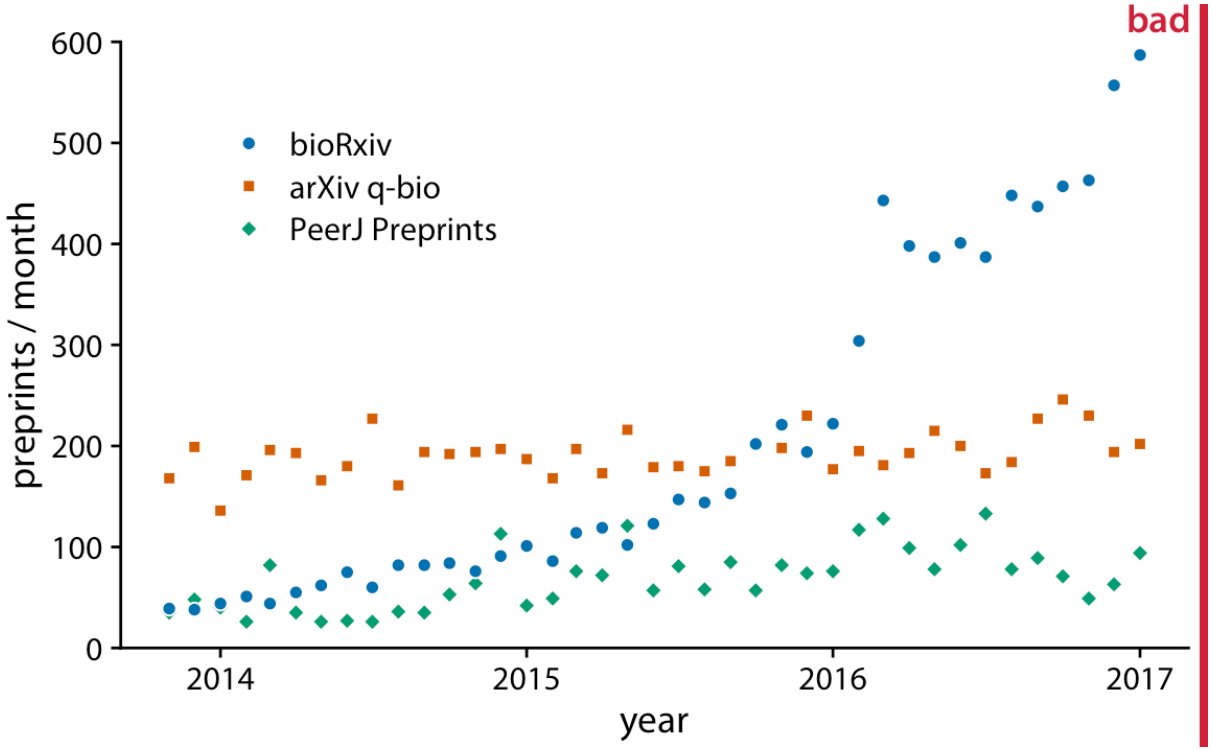


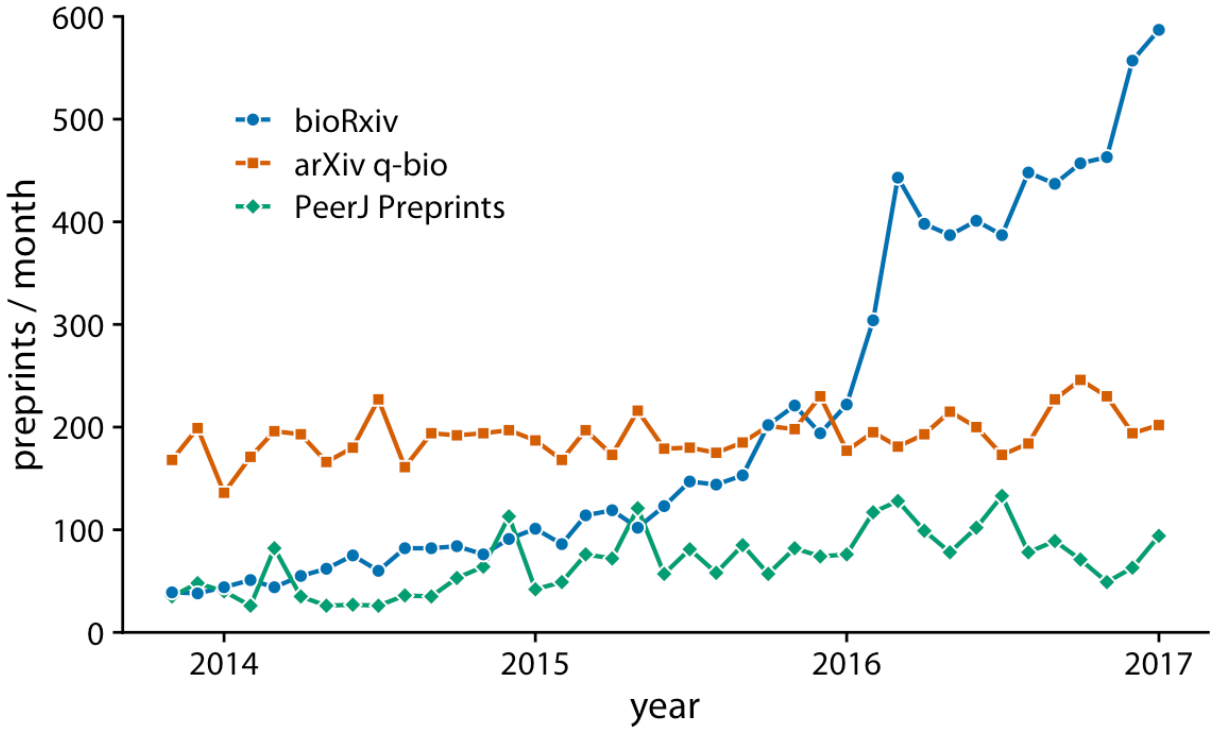
https://albert-rapp.de/posts/ggplot2-tips/13_alternative_corrplots/13_alternative_corrplots.html

Should Your Data Be Connected?

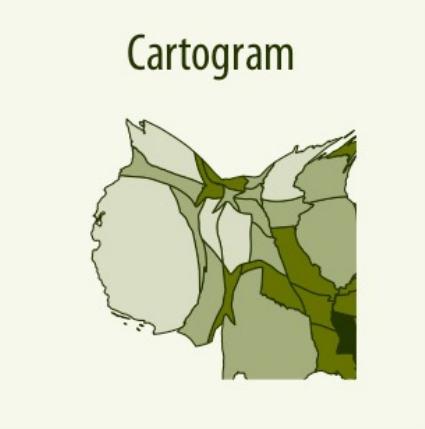


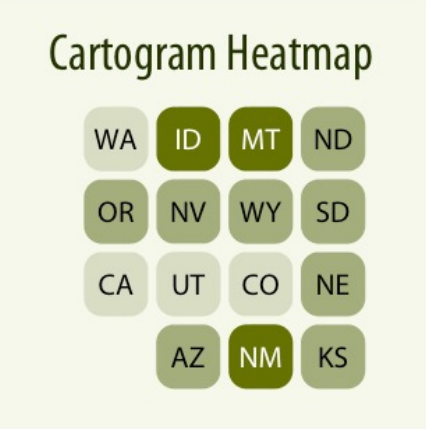
Should Your Data Be Connected?



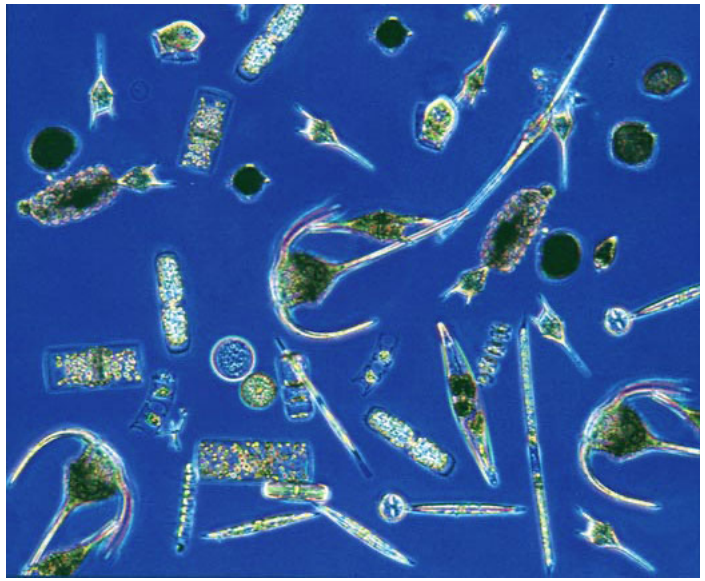


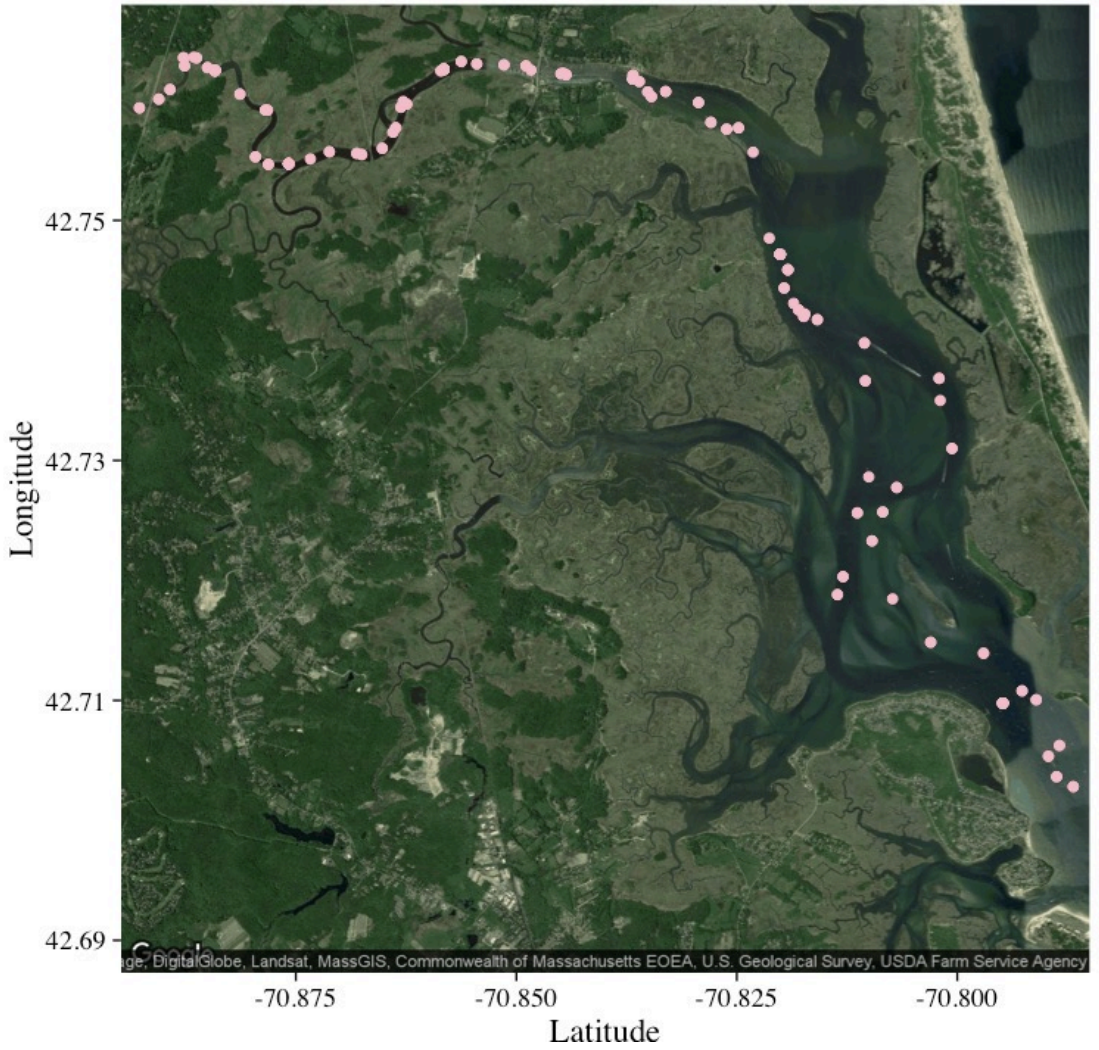
Geospatial Visualization





Combining Data Sources and Maps

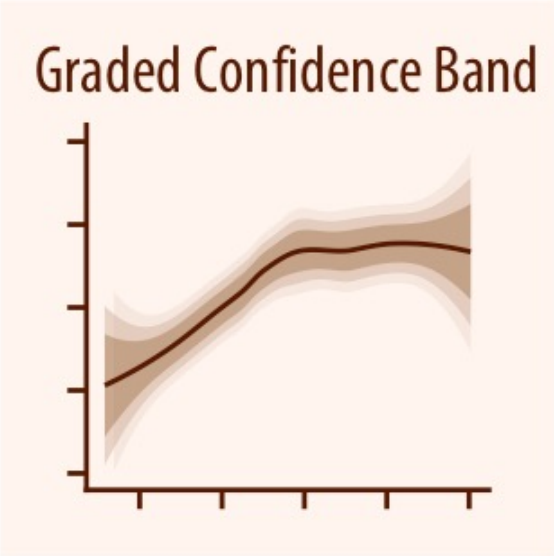


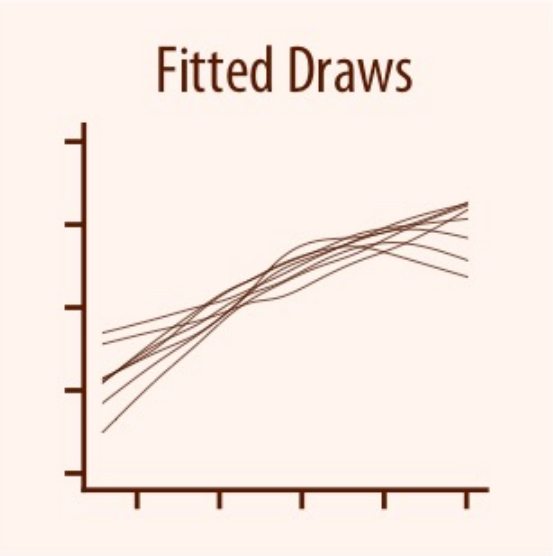


- Chlorophyll *a*
- Abundance of taxonomic groups
- Temperature
- Salinity

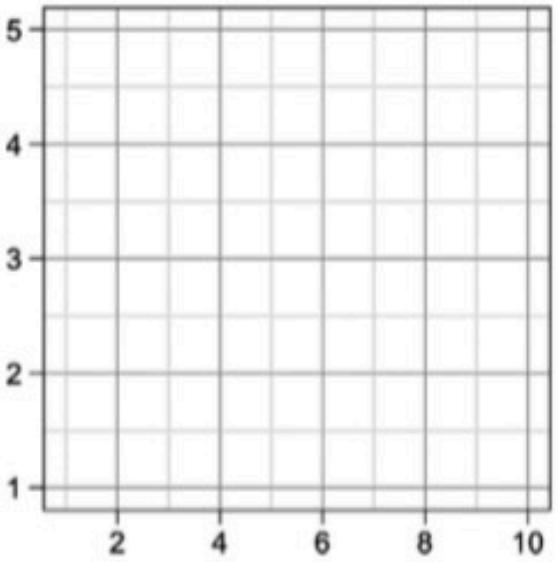
Hopkinson and Hobbie 2012 http://ecosystems.mbl.edu/PIE/data/EST/EST-PR-ChemTax.html

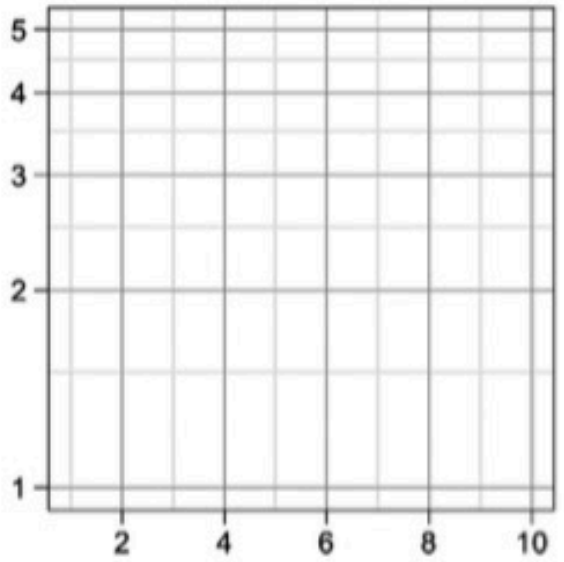
Visualizing Uncertainty in Relationships is Difficult

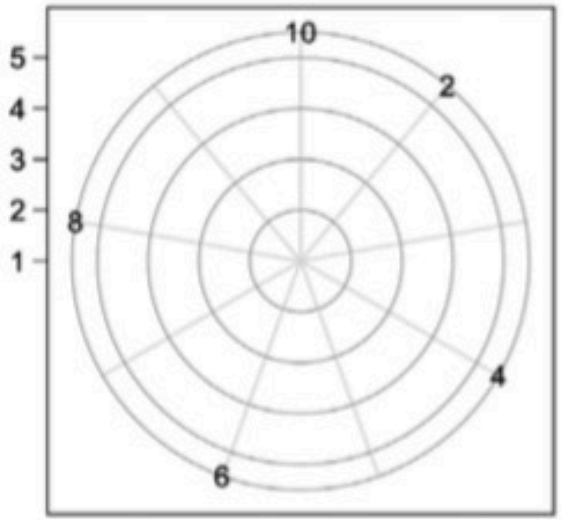




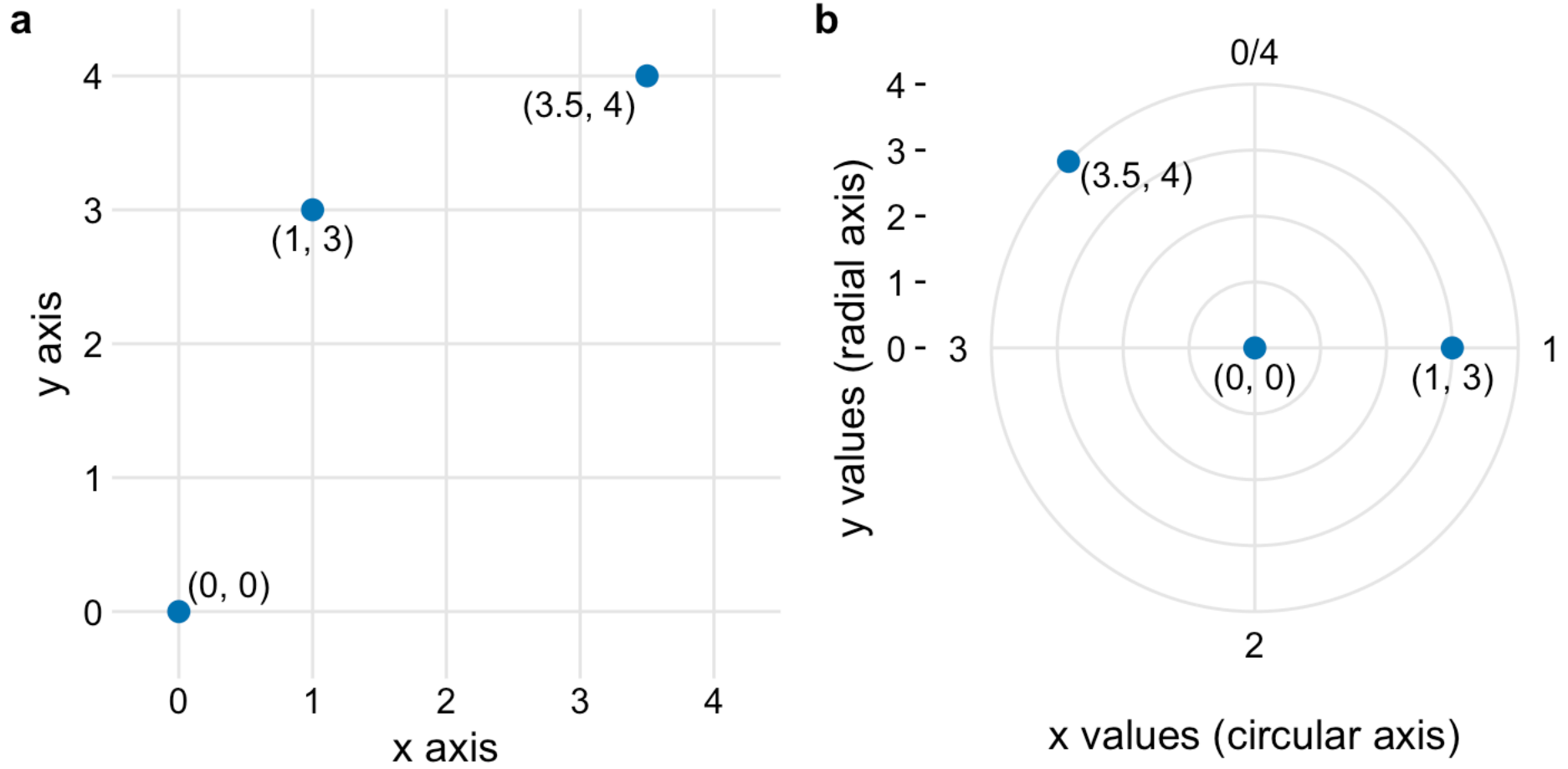
Coordinate Systems Transform Relationships



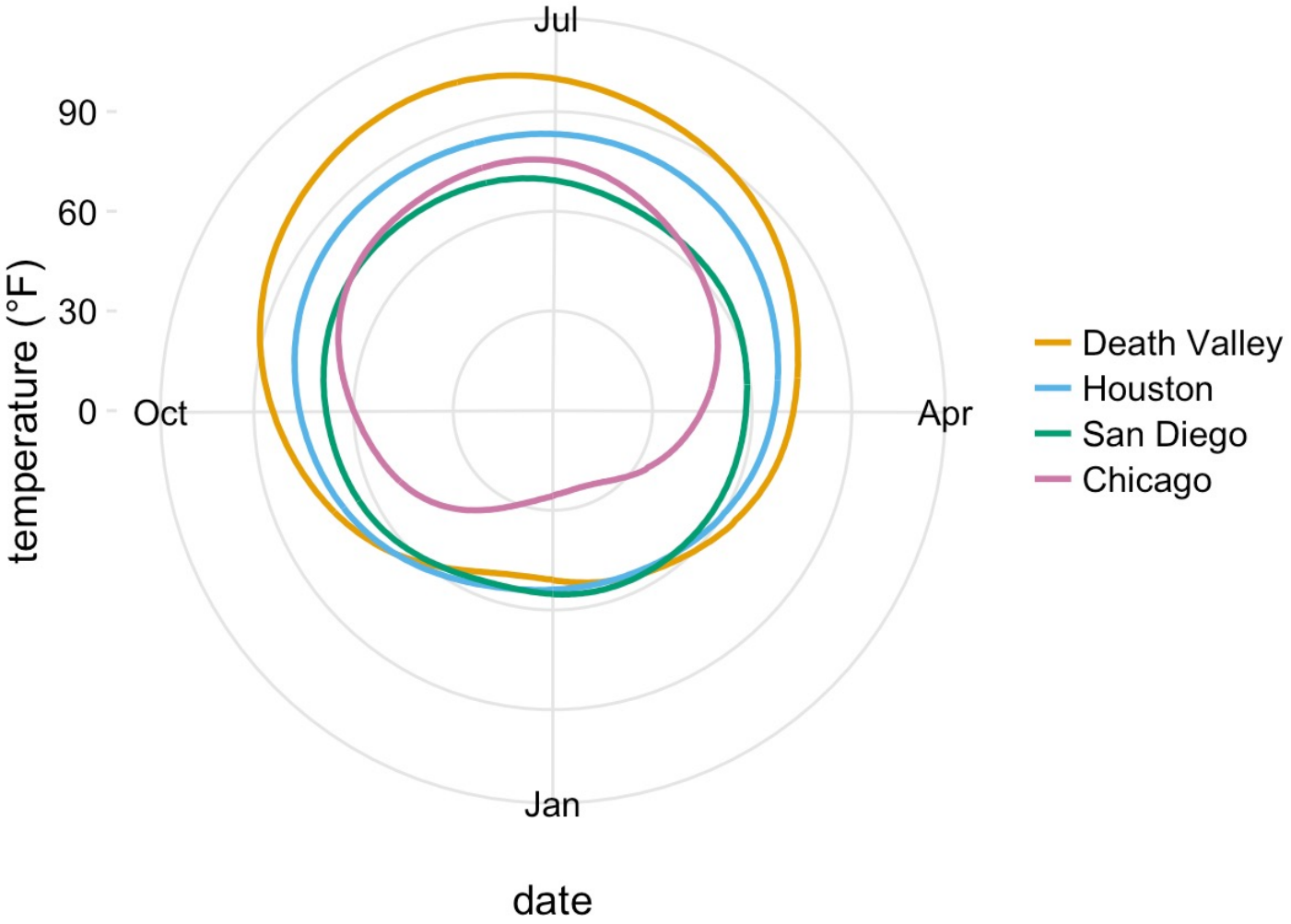




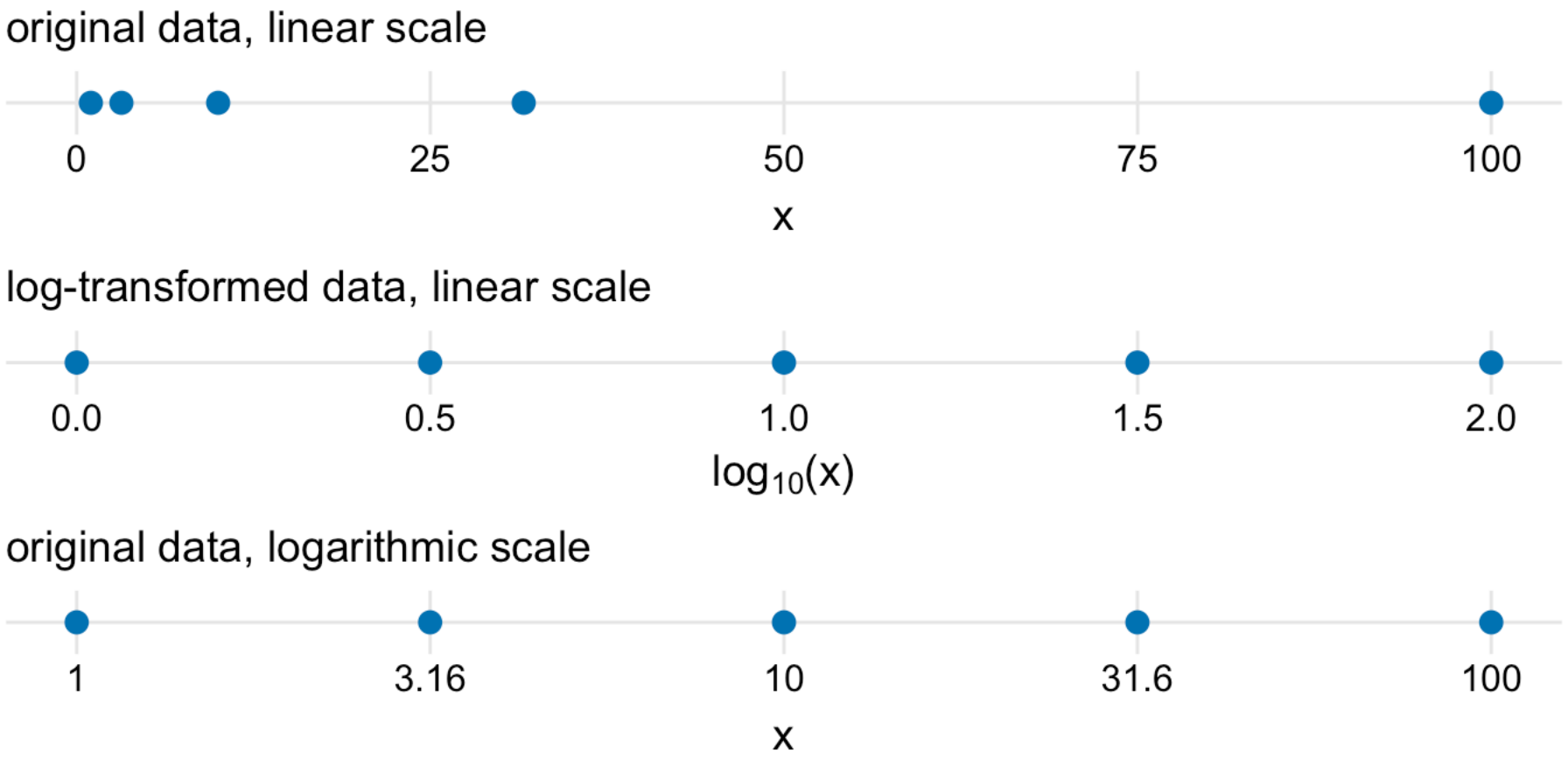
The Polar Coordinate System is Useful!



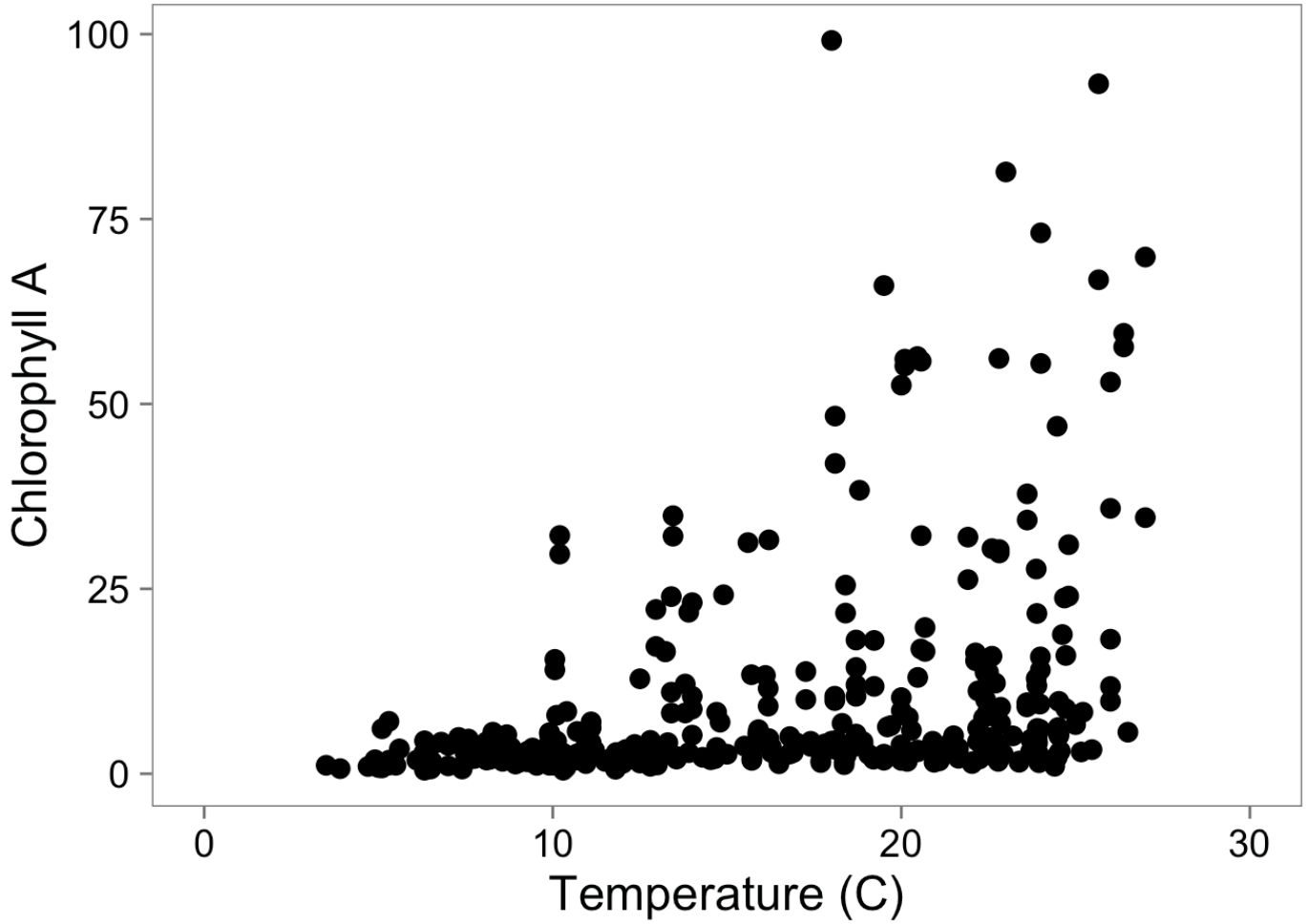
The Polar Coordinate System is Useful!



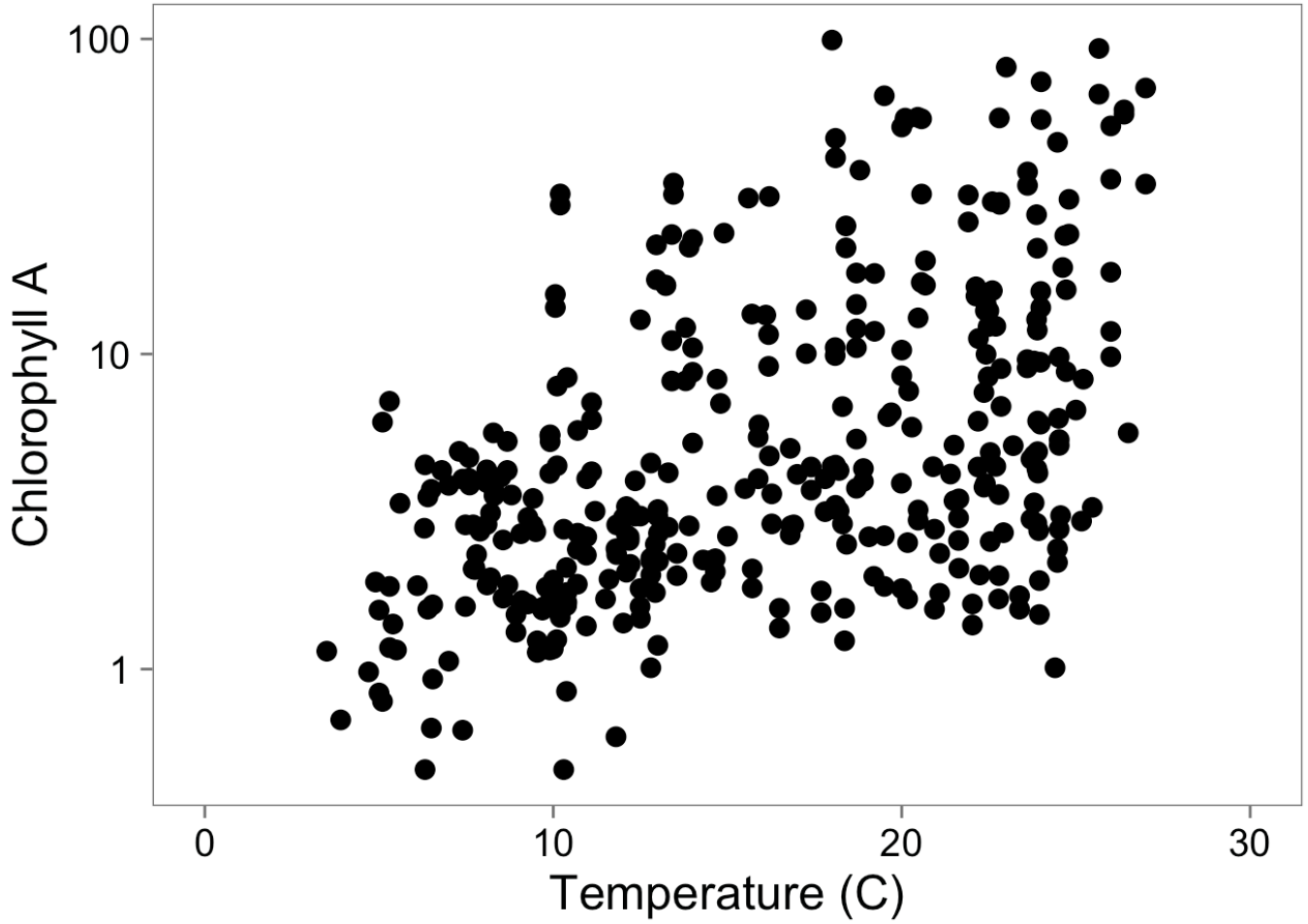
Why Transform?



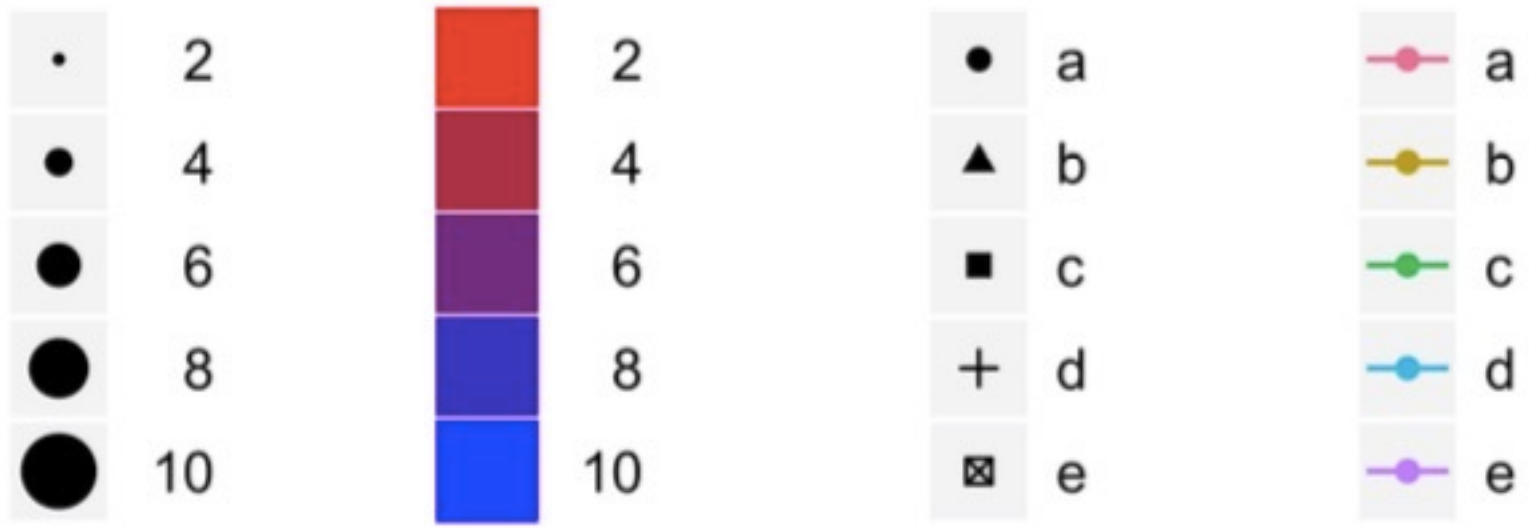
Adding Full Scale to 0



Log-Transformation To See Relationship



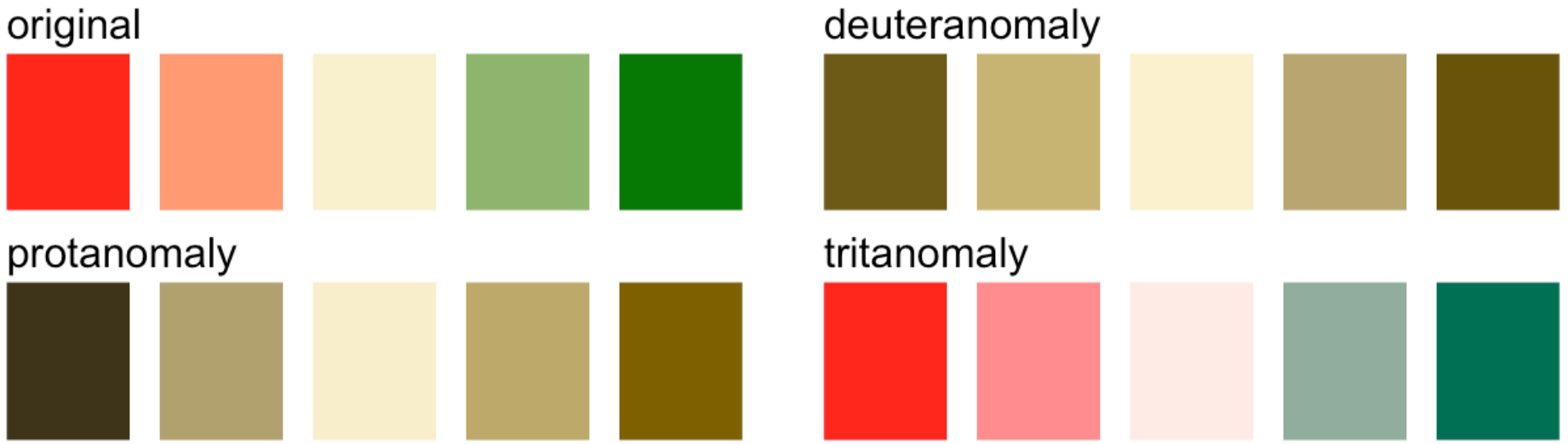
Scales to Add Dimensions of Data



Colors Can Distinguish Groups

Colors Can Show Data

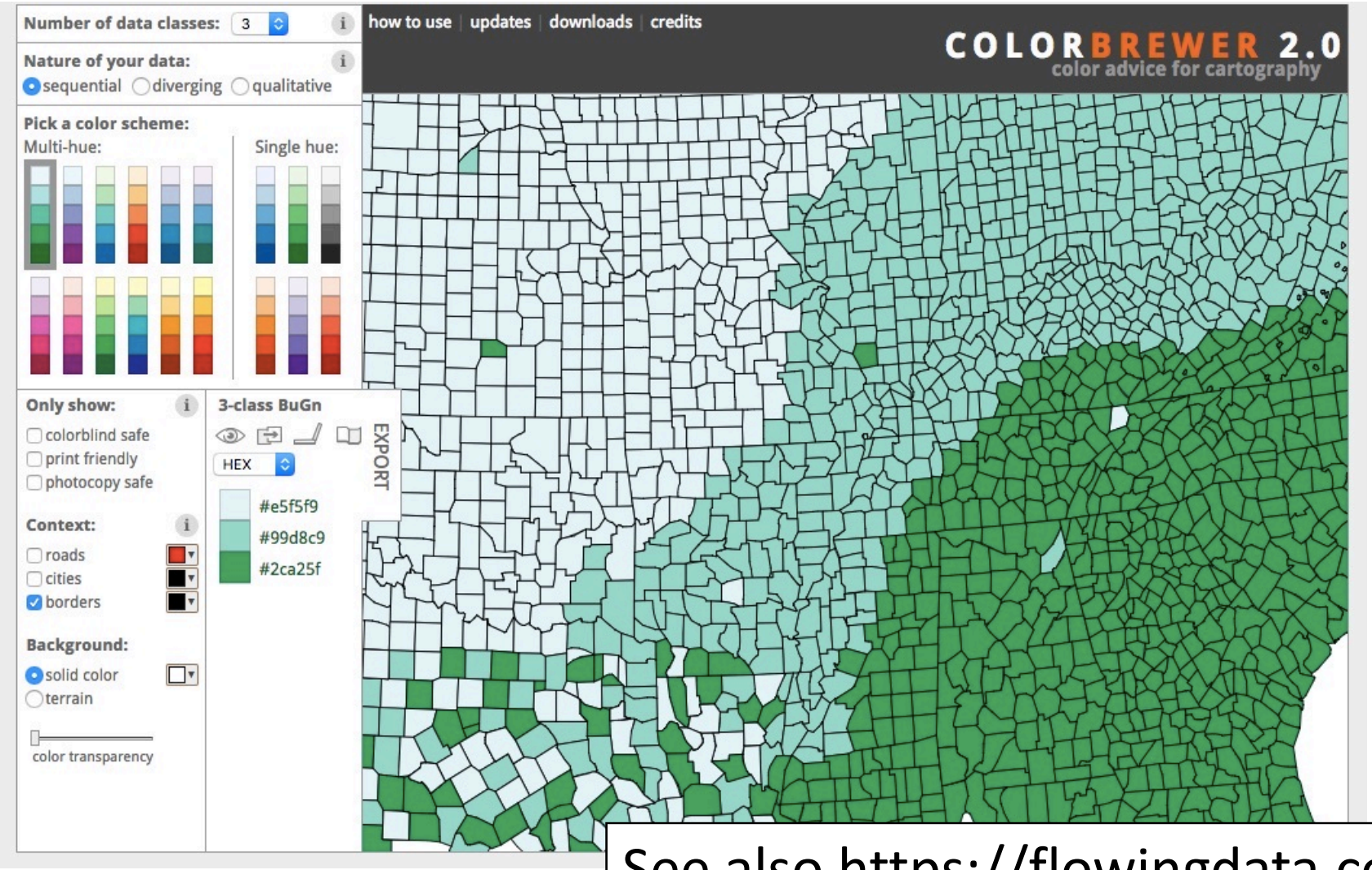
Beware Not Thinking About Color Blindness



Never use Red-Green!

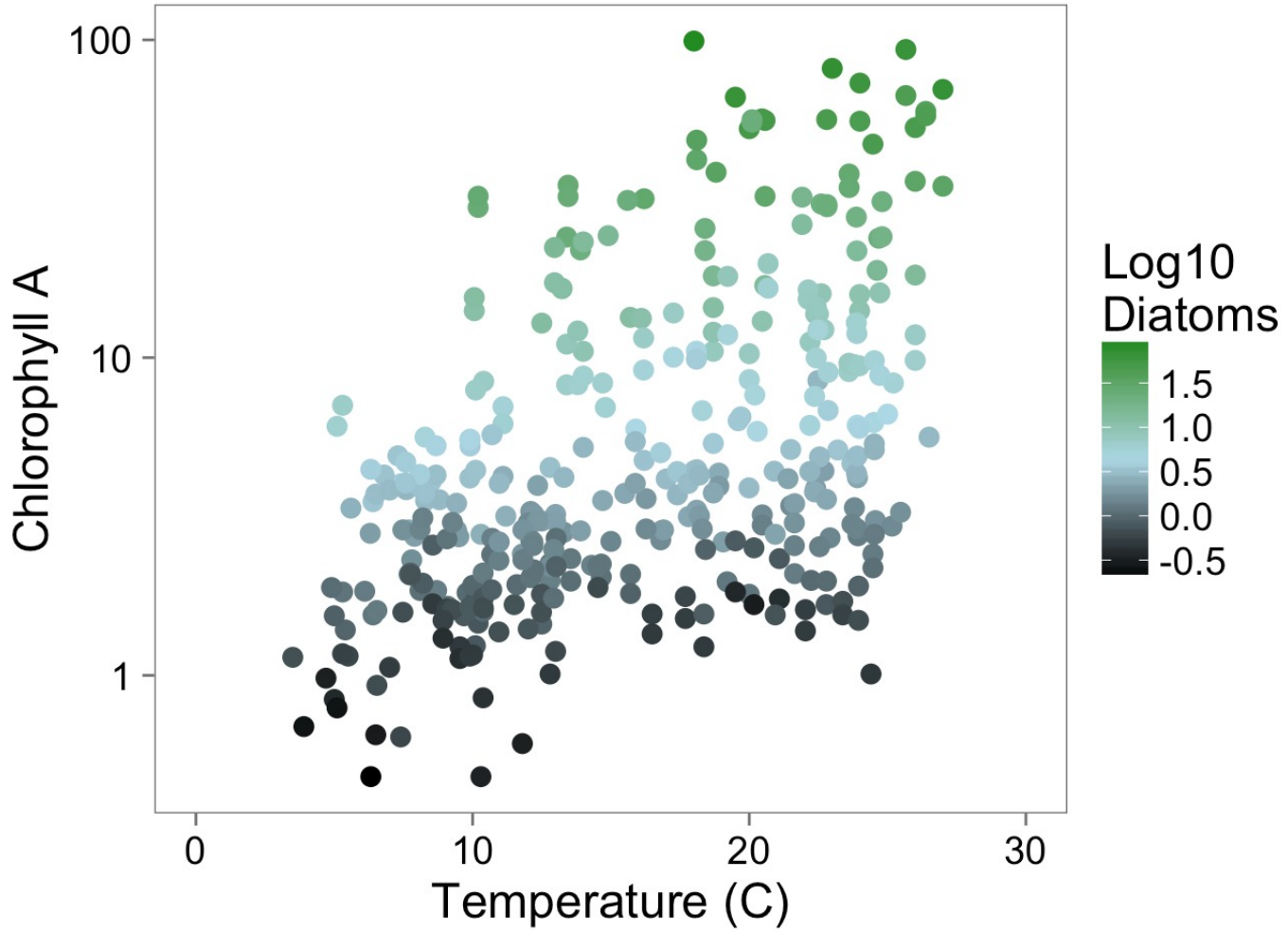
Use redundant coding: shapes, sizes, etc.

Colorbrewer.org

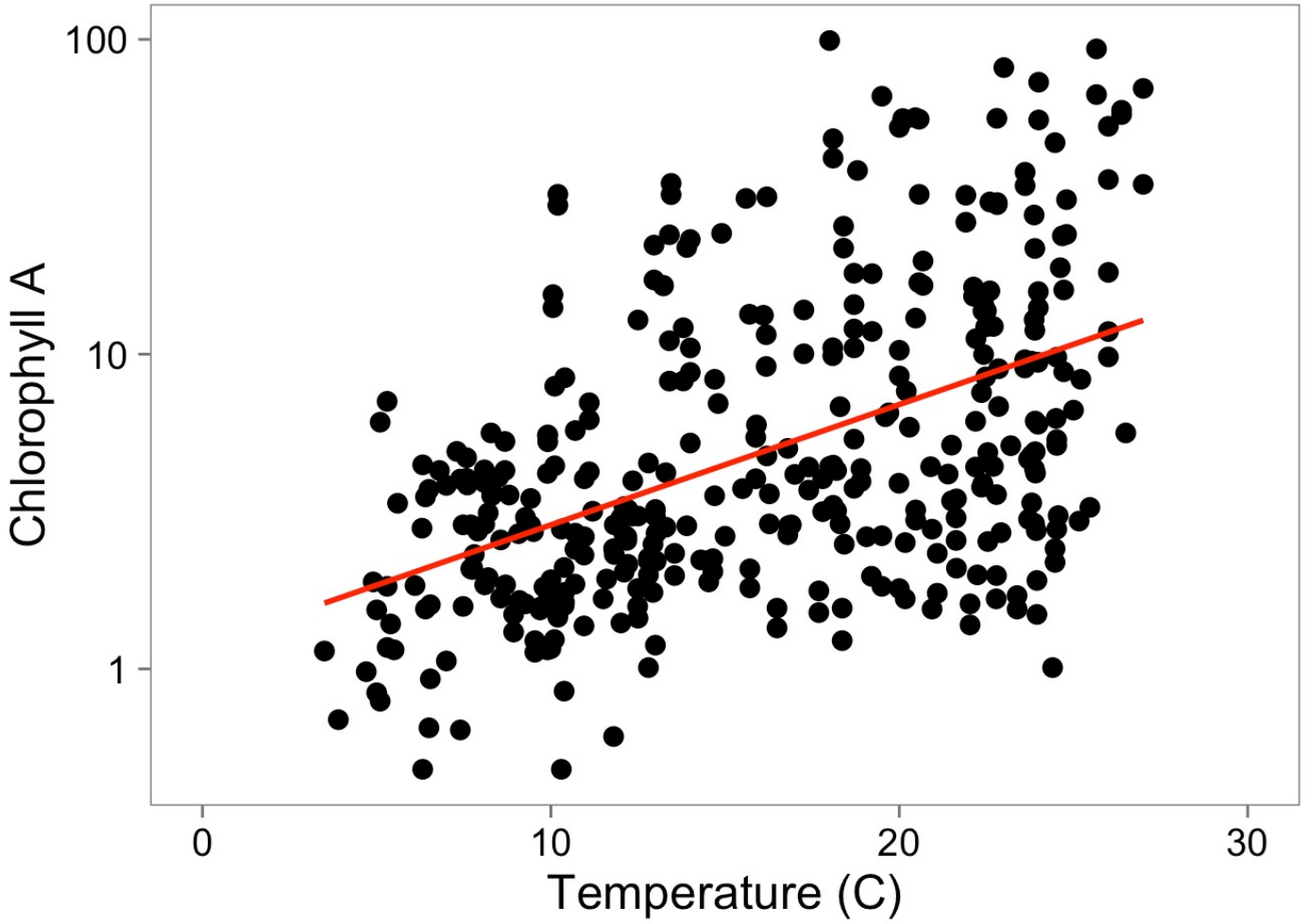


See also https://flowingdata.com/tag/color/

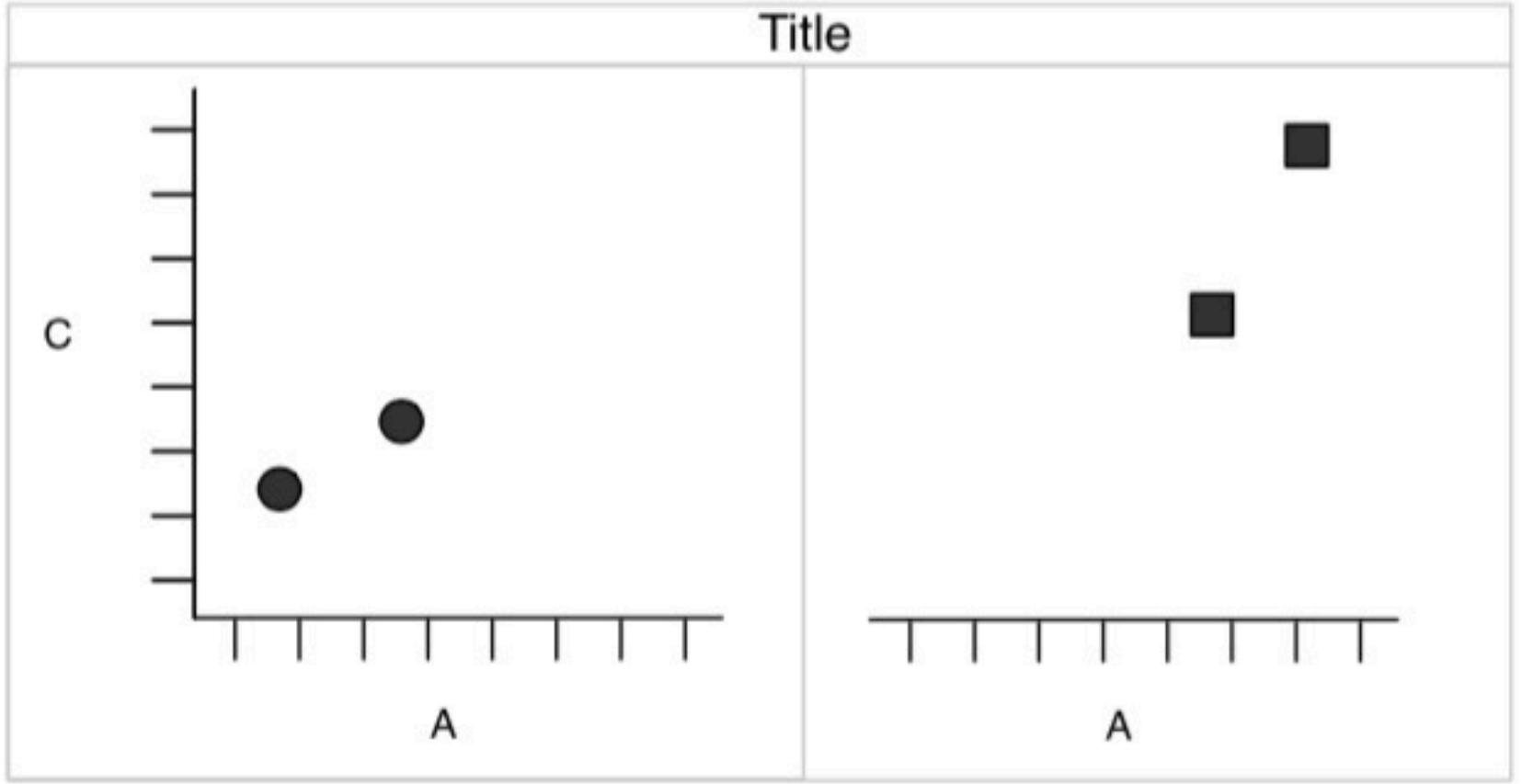
Color Can Bring in Another Dimension



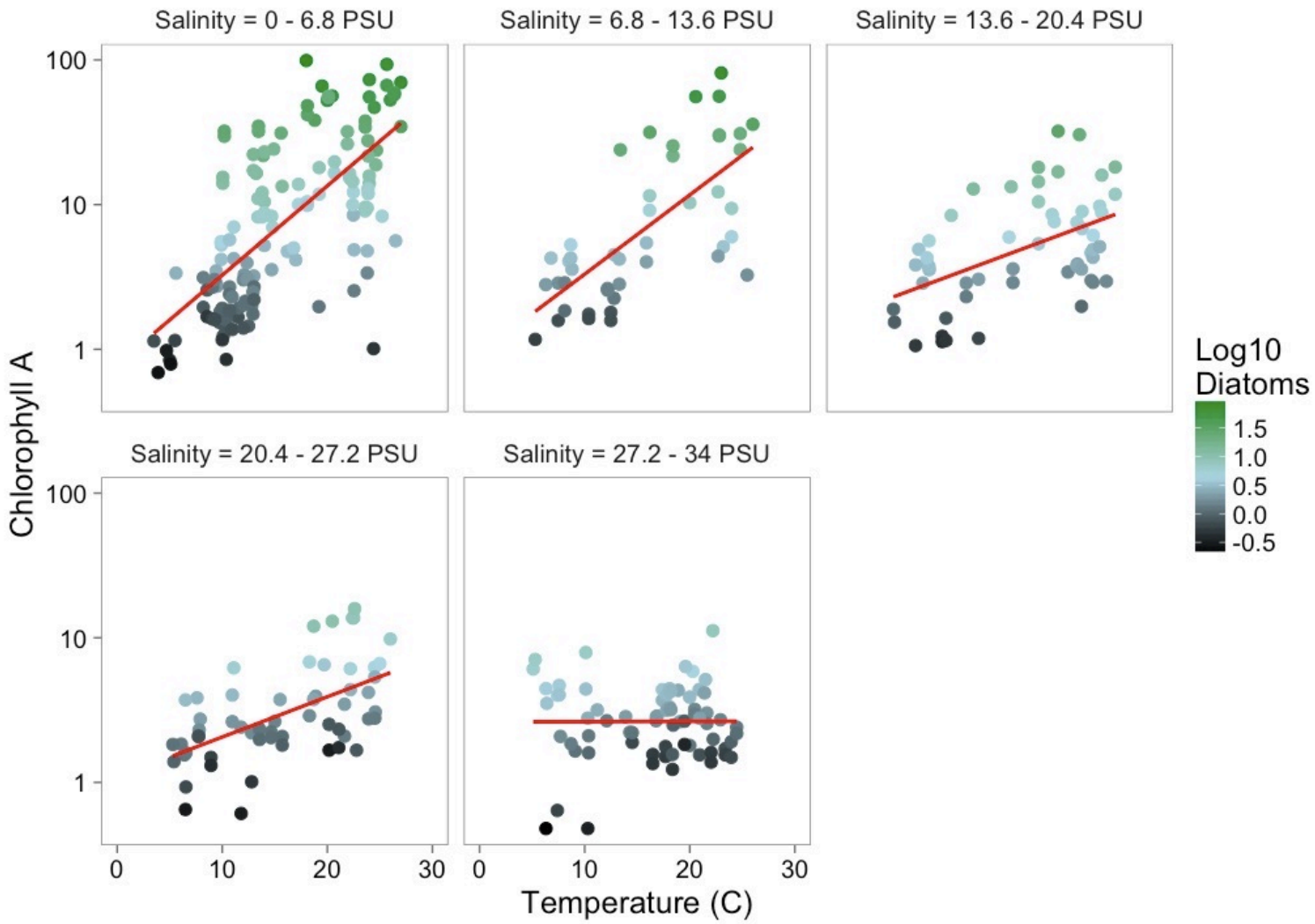
Statistical Fit to Aid Understanding



Facets to Add Fine-Grained Information or New Dimensions



Facets Add Information



Outline

- Background
- Visualization Principles
- **Automatic Visualization**

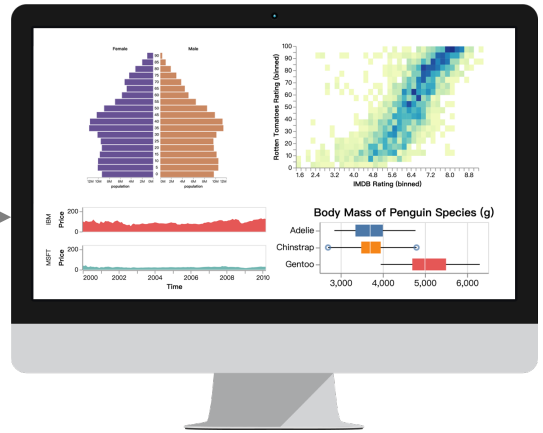
Why Automatic Data Visualization?

Human-powered Data Visualization

1 hour for good visualizations!

Data

Select Data Fields → Data Transformation → Visual Encoding



Visualizations

pandas, R, plotly, ggplot2, +ableau, Qlik, Power BI, Excel

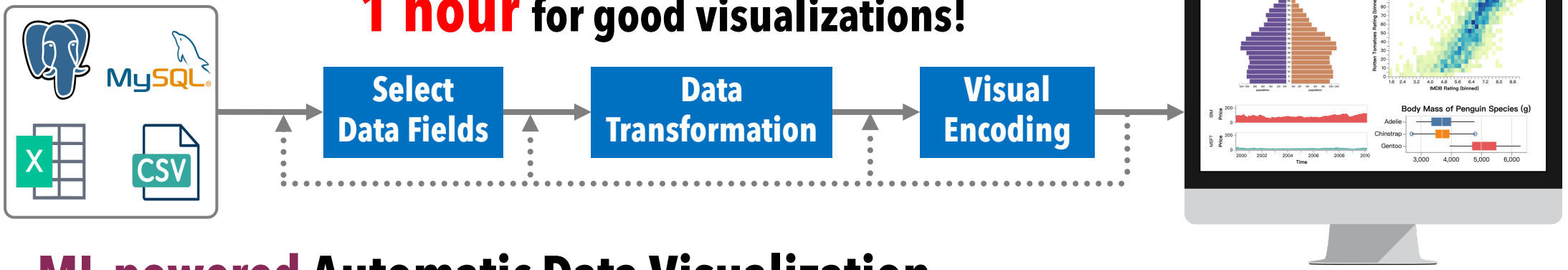
Manually specify through Code or Clicks

- *Require human and domain expertise*
- *Tedious and time-consuming (even for experts)*

Why Automatic Data Visualization?

Human-powered Data Visualization

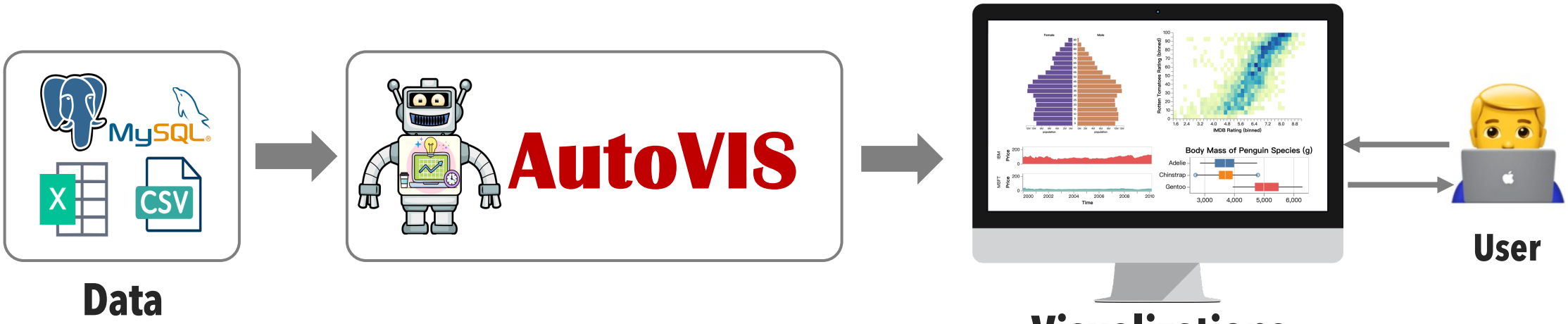
1 hour for good visualizations!



ML-powered Automatic Data Visualization

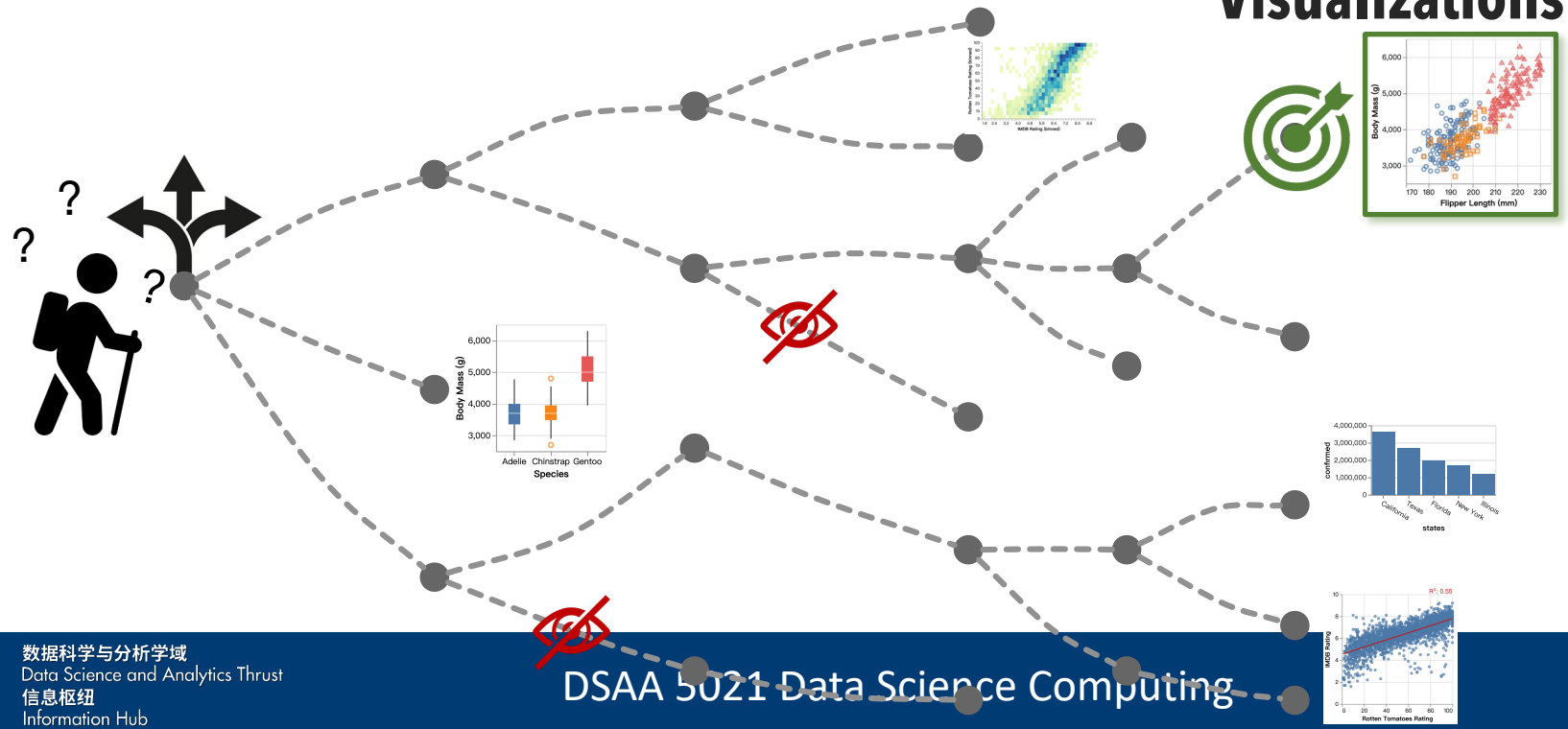
1 second for good visualizations!

AutoVIS: Automatic Data Visualization



Data

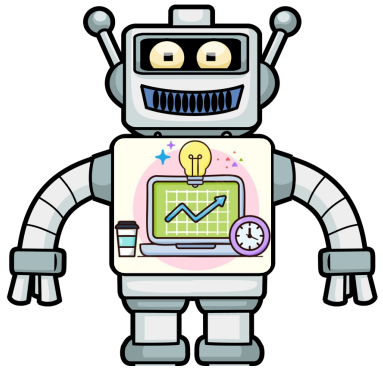
Visualizations



AutoVIS: A Path for Democratizing Data Analytics

Manual → Automatic

Cost Saving



AutoVIS

Barrier Lowering

Capability Improving

Expert → Everyone

Empirical → ML-powered

DEEPEYE: An End-to-End and ML-powered AutoVIS System

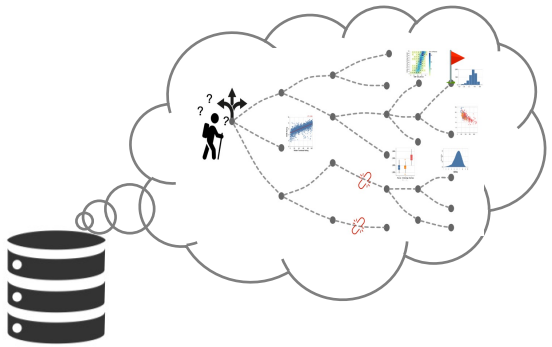
DEEPEYE System = AutoVIS + User Intent + Cleaned Data

Fully Automatic
Data Visualization
[ICDE'18, SIGMOD'23]

Intent-based
Data Visualization
[SIGMOD'21, IEEE VIS'21]

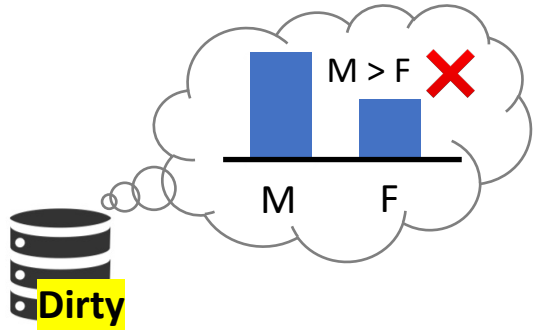
Quality-aware
Data Visualization
[ICDE'20, VLDB'20 demo]

How to lower the barriers?



Missing User Intent

Visualizations Meet Data Errors



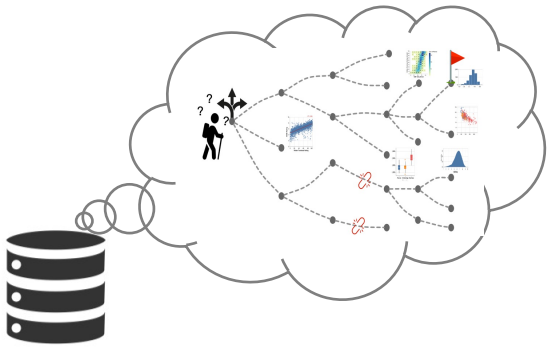
Visualization System = Visualization Process + User + Data

DEEPEYE: An End-to-End and ML-powered AutoVIS System

DEEPEYE System = AutoVIS + User Intent + Cleaned Data

Fully Automatic
Data Visualization
[ICDE'18, SIGMOD'23]

How to lower the barriers?



Intent-based
Data Visualization
[SIGMOD'21, IEEE VIS'21]

Missing User Intent

Quality-aware
Data Visualization
[ICDE'20, VLDB'20 demo]

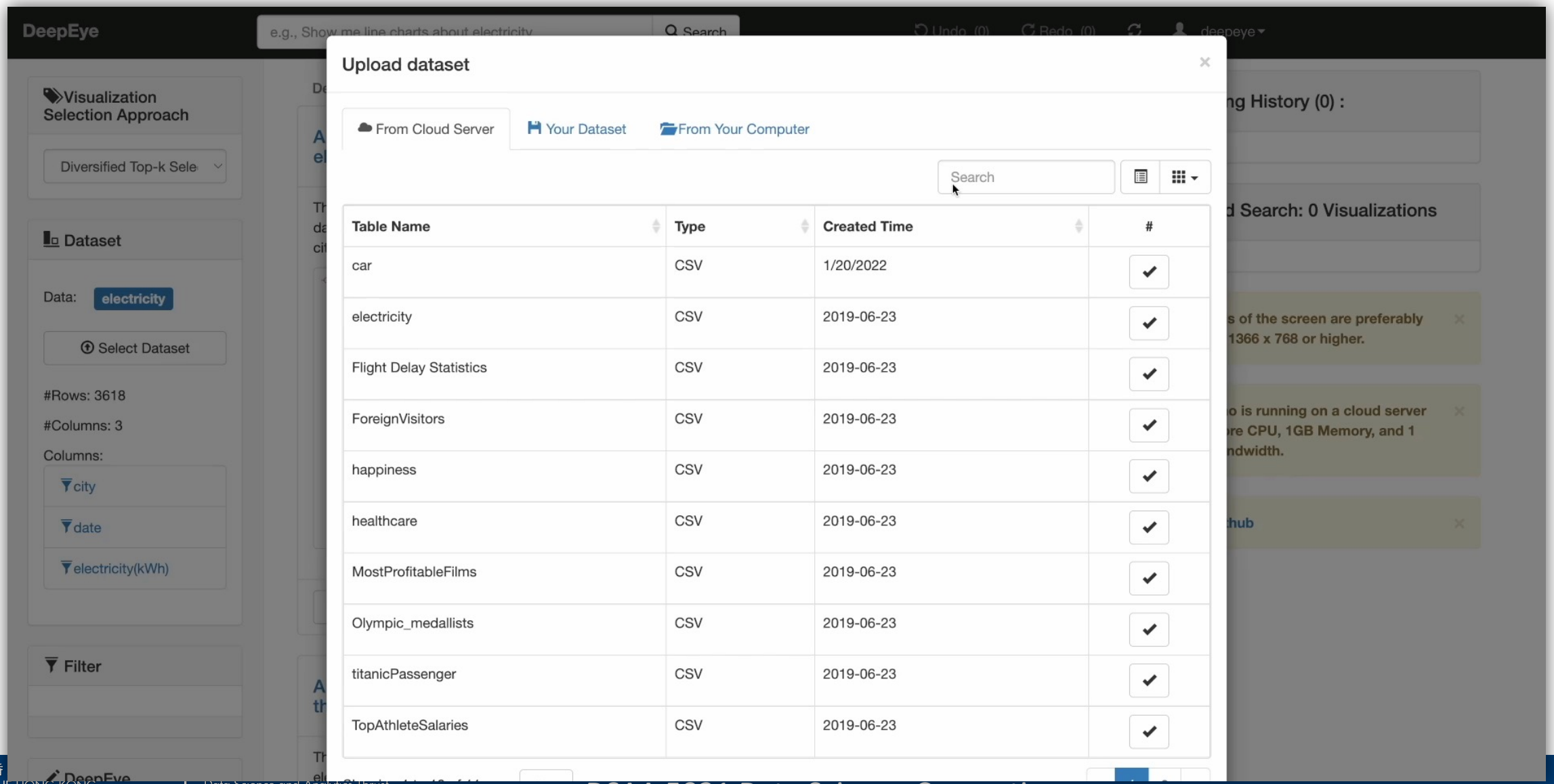
Visualizations Meet Data Errors

Fully AutoVIS

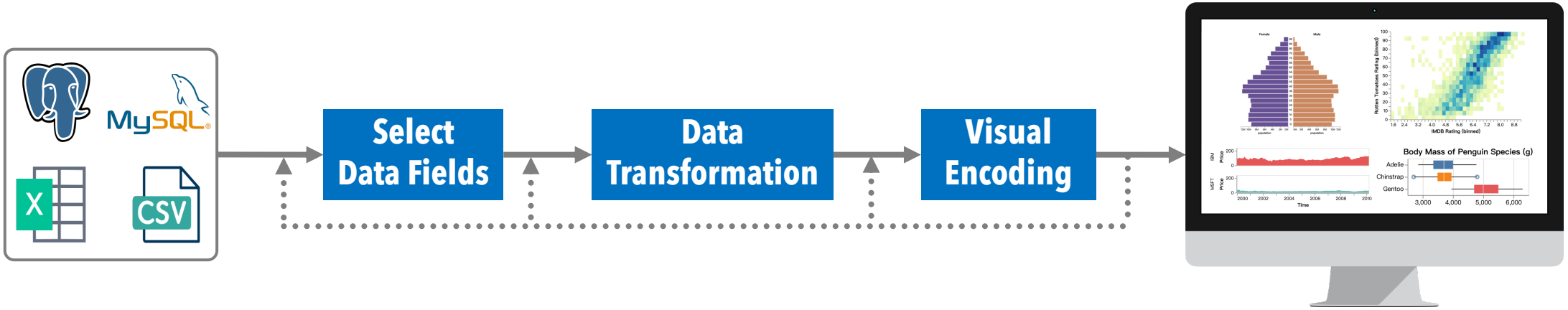
year	month	carrier	carrier_name	airport	airport_name	arr_flights	arr_del15	carrier_ct	weather_ct	nas_ct	security_ct	late_aircraft_ct	arr_cancelled	arr_diverted	arr_delay	carrier_delay	weather_delay	nas_delay	security_delay	late_aircraft_delay
2020	12	9E	Endeavor Air Inc.	ABE	Allentown/Bethlehem/Easton, PA: Lehigh Valley International	44	3	1.63	0	0.12	0	1.25	0	1	89	56	0	3	0	30
2020	12	9E	Endeavor Air Inc.	ABY	Albany, GA: Southwest Georgia Regional	90	1	0.96	0	0.04	0	0	0	0	23	22	0	1	0	0
2020	12	9E	Endeavor Air Inc.	AEX	Alexandria, LA: Alexandria International	88	8	5.75	0	1.6	0	0.65	0	1	338	265	0	45	0	28
2020	12	9E	Endeavor Air Inc.	AGS	Augusta, GA: Augusta Regional at Bush Field	184	9	4.17	0	1.83	0	3	0	0	508	192	0	92	0	224
2020	12	9E	Endeavor Air Inc.	ALB	Albany, NY: Albany International	76	11	4.78	0	5.22	0	1	1	0	692	398	0	178	0	116
2020	12	9E	Endeavor Air Inc.	ATL	Atlanta, GA: Hartsfield-Jackson Atlanta International	5985	445	142.89	11.96	161.37	1	127.79	5	0	30756	16390	1509	5060	16	7781
2020	12	9E	Endeavor Air Inc.	ATW	Appleton, WI: Appleton International	142	14	5.36	0	7.7	0	0.94	1	0	436	162	0	182	0	92
2020	12	9E	Endeavor Air Inc.	AVL	Asheville, NC: Asheville Regional	147	10	6.04	1	1	0	1.96	0	1	1070	838	141	24	0	67
2020	12	9E	Endeavor Air Inc.	AZO	Kalamazoo, MI: Kalamazoo/Battle Creek International	84	14	6.24	0.96	6.8	0	0	1	1	2006	1164	619	223	0	0
2020	12	9E	Endeavor Air Inc.	BDL	Hartford, CT: Bradley International	150	19	5.7	0	12.07	0	1.23	3	0	846	423	0	389	0	34
2020	12	9E	Endeavor Air Inc.	BHM	Birmingham, AL: Birmingham-Shuttlesworth International	123	9	6.82	0	1.18	0	1	0	0	572	527	0	26	0	19
2020	12	9E	Endeavor Air Inc.	BIS	Bismarck/Mandan, ND: Bismarck Municipal	25	7	0.34	0	3.54	0	3.11	2	0	278	13	0	93	0	172
2020	12	9E	Endeavor Air Inc.	BMI	Bloomington/Normal, IL: Central Il Regional Airport at Bloomington	59	5	3.15	0	1.48	0	0.37	0	0	218	155	0	47	0	16
2020	12	9E	Endeavor Air Inc.	BNA	Nashville, TN: Nashville International	2	1	0	0	1	0	0	0	0	20	0	0	20	0	0
2020	12	9E	Endeavor Air Inc.	BOS	Boston, MA: Logan International	21	3	1	0.71	1.29	0	0	0	0	158	17	81	60	0	0
2020	12	9E	Endeavor Air Inc.	BQK	Brunswick, GA: Brunswick Golden Isles	90	6	3.8	0	1.63	0	0.57	0	0	374	249	0	88	0	37
2020	12	9E	Endeavor Air Inc.	BTR	Baton Rouge, LA: Baton Rouge Metropolitan/Ryan Field	115	5	4.82	0	0.18	0	0	0	1	244	236	0	8	0	0
2020	12	9E	Endeavor Air Inc.	BTV	Burlington, VT: Burlington International	90	10	3.06	0.85	5.4	0	0.68	0	0	759	220	258	211	0	70
2020	12	9E	Endeavor Air Inc.	BUF	Buffalo, NY: Buffalo Niagara International	14	1	1	0	0	0	0	1	0	18	18	0	0	0	0
2020	12	9E	Endeavor Air Inc.	BWI	Baltimore, MD: Baltimore/Washington International Thurgood Marshall	96	17	8.5	0	5.11	0	3.4	1	0	1064	558	0	200	0	306
2020	12	9E	Endeavor Air Inc.	CAE	Columbia, SC: Columbia Metropolitan	106	5	3.7	0	1.3	0	0	0	0	230	205	0	25	0	0
2020	12	9E	Endeavor Air Inc.	CHA	Chattanooga, TN: Lovell Field	213	12	6.42	2.96	0.92	0	1.71	0	0	542	279	202	25	0	36
2020	12	9E	Endeavor Air Inc.	CHO	Charlottesville, VA: Charlottesville Albemarle	86	1	0.59	0	0	0	0.41	1	1	17	10	0	0	0	7
2020	12	9E	Endeavor Air Inc.	CHS	Charleston, SC: Charleston AFB/International	61	6	3.03	0	2.65	0	0.32	1	0	209	121	0	67	0	21
2020	12	9E	Endeavor Air Inc.	CID	Cedar Rapids/Iowa City, IA: The Eastern Iowa	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2020	12	9E	Endeavor Air Inc.	CLE	Cleveland, OH: Cleveland-Hopkins International	74	12	2.57	0.96	7.47	0	1	0	0	677	118	121	398	0	40
2020	12	9E	Endeavor Air Inc.	CLT	Charlotte, NC: Charlotte Douglas International	42	1	1	0	0	0	0	0	0	64	64	0	0	0	0
2020	12	9E	Endeavor Air Inc.	CMH	Columbus, OH: John Glenn Columbus International	84	8	2.44	0	4.3	0	1.26	2	0	446	127	0	190	0	129
2020	12	9E	Endeavor Air Inc.	CRW	Charleston/Dunbar, WV: Yeager	113	8	6	1	0.16	0	0.84	0	0	376	305	22	8	0	41
2020	12	9E	Endeavor Air Inc.	CSG	Columbus, GA: Columbus Airport	99	6	2.45	0.97	2.58	0	0	0	0	162	74	33	55	0	0
2020	12	9E	Endeavor Air Inc.	CVG	Cincinnati, OH: Cincinnati/Northern Kentucky International	381	43	6.09	1.33	23.77	0	11.82	4	0	1689	216	100	685	0	688
2020	12	9E	Endeavor Air Inc.	CWA	Mosinee, WI: Central Wisconsin	88	14	6.16	0	5.82	0	2.02	2	0	765	291	0	245	0	229
2020	12	9E	Endeavor Air Inc.	DAL	Dallas, TX: Dallas Love Field	39	6	0.09	0	3.76	0	2.15	0	0	420	5	0	296	0	119
2020	12	9E	Endeavor Air Inc.	DAY	Dayton, OH: James M Cox/Dayton International	76	7	4.89	0	1.54	0	4.57	0	0	850	257	0	232	0	361
2020	12	9E	Endeavor Air Inc.	DCA	Washington, DC: Ronald Reagan Washington National	0	0	0	0	0	0	0	1	0	15	0	0	2	0	13
2020	12	9E	Endeavor Air Inc.	DFW	Dallas/Fort Worth, TX: Dallas/Fort Worth International	83	8	1.5	0	5.4	0	1.1	0	0	454	134	0	260	0	60

Flight Delay Dataset

Fully AutoVIS



How

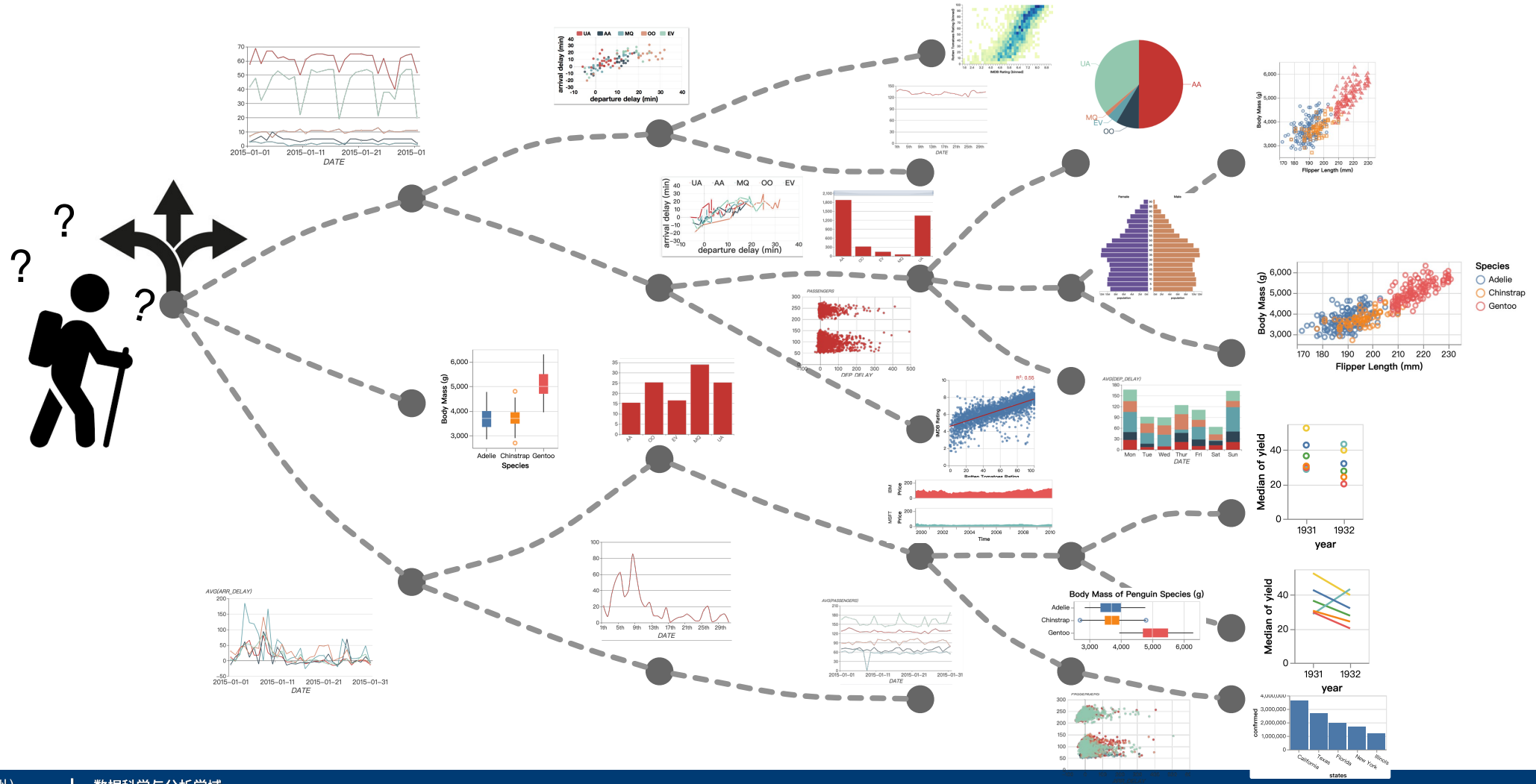


Users manually try the right **combination of data fields**
Users manually try the right **data transformation**
Users manually try the right **visual encoding**

Machines automatically run the right **combination of data fields**
Machines automatically run the right **data transformation**
Machines automatically run the right **visual encoding**

AutoVIS Pipeline

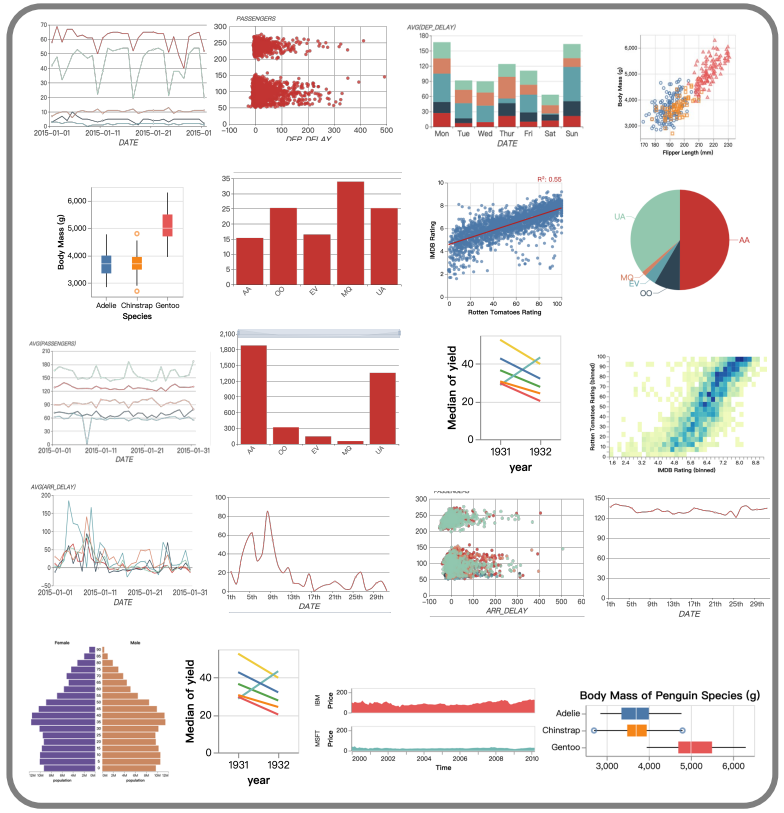
Candidates Generation → **Visualization Recognition** → **Ranking& Recommendation**



AutoVIS Pipeline

Candidates Generation → **Visualization Recognition** → **Ranking& Recommendation**

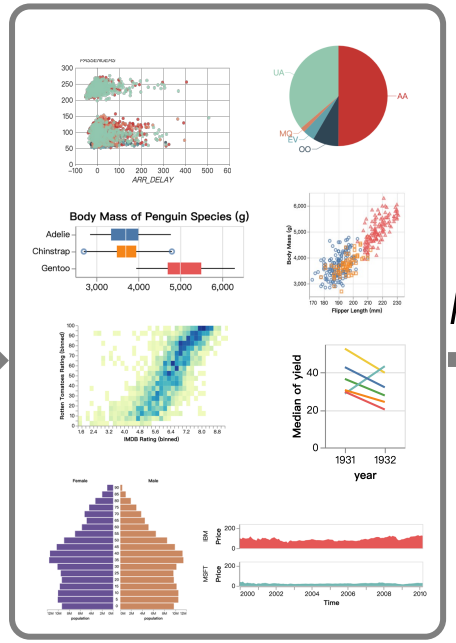
1. What is the search space?



All Visualization Candidates

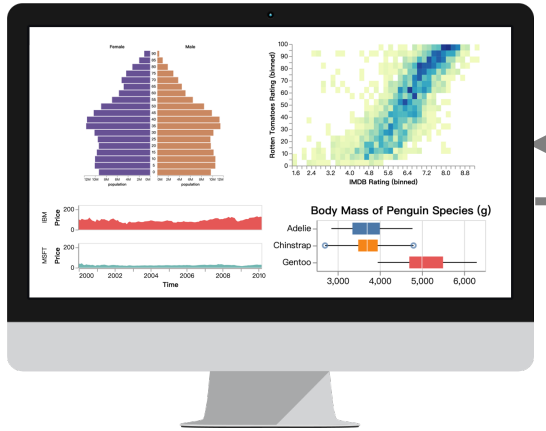
Select Good

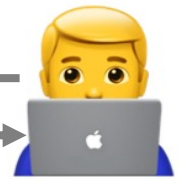
2. Which one is good?



Good Visualizations

Rank & Recommend

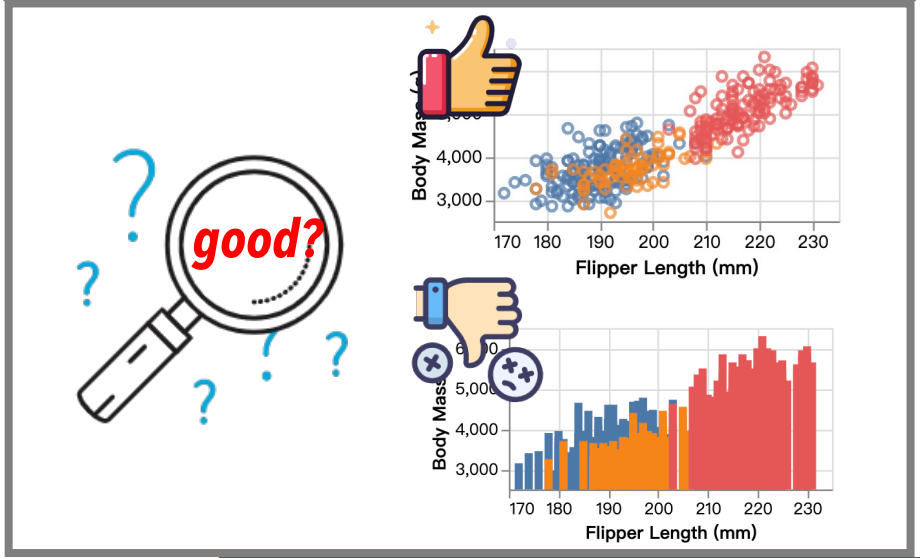




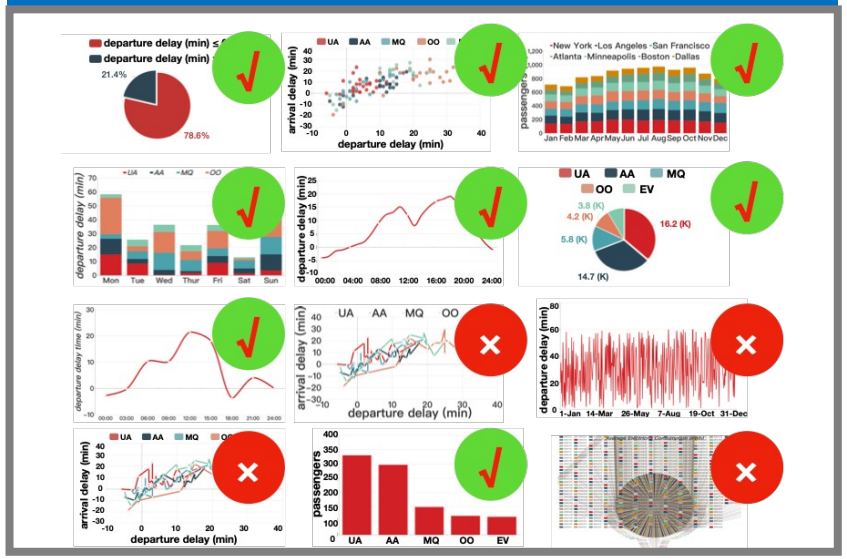
Recommended Visualizations

Three Key Challenges

How to quantify the goodness?



Lack of ground truth

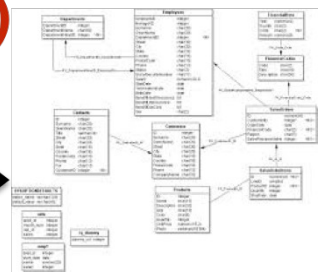


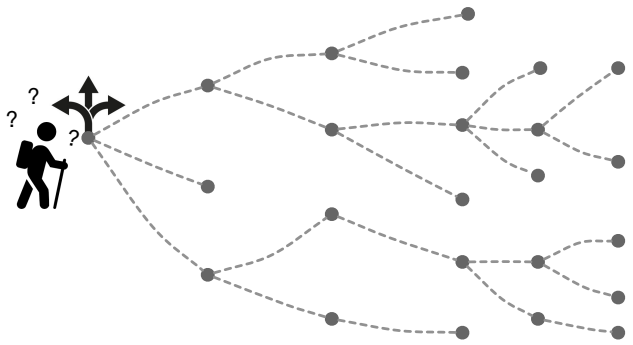
Huge search space for candidate visualizations

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

(DB problem)

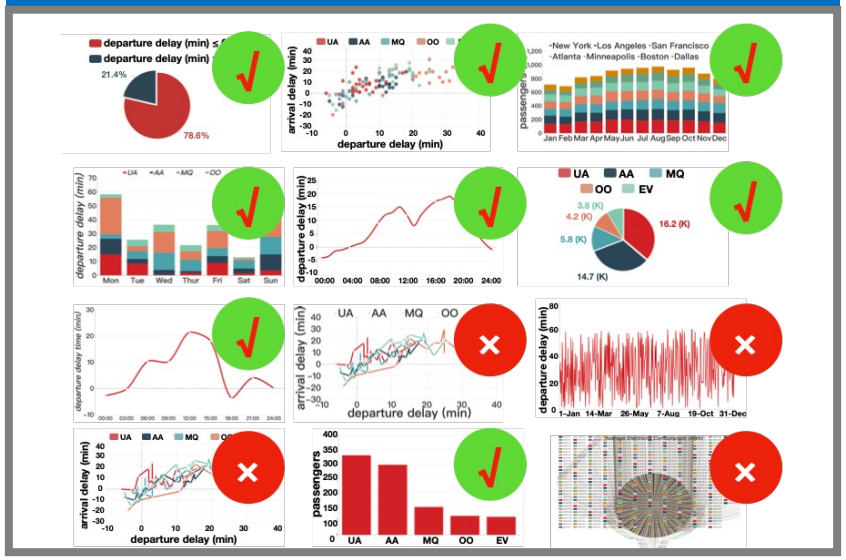
transform



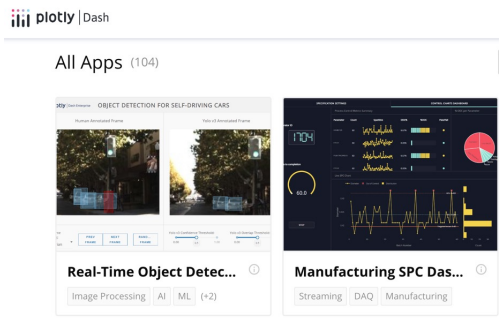


Basic Ideas

Lack of ground truth

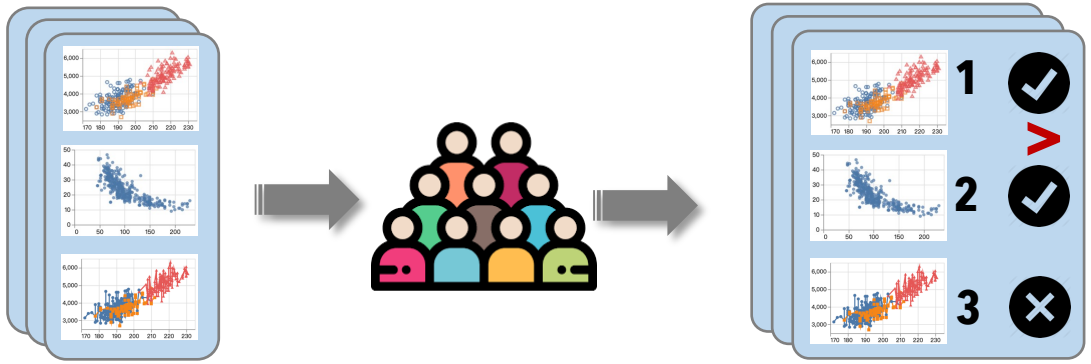


- **Collecting from the internet**



Scale ↑ Coverage ↓ Quality ↓

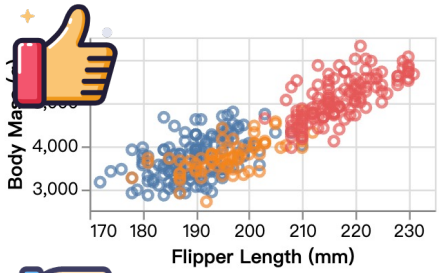
- **Collecting by crowdsourcing**

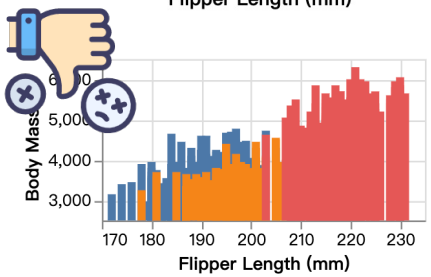


Scale ↑ Coverage ↑ Quality ↑

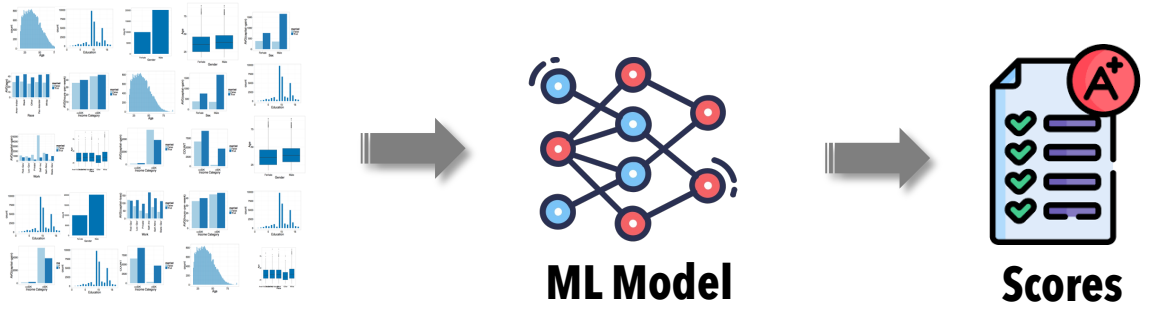
Basic Ideas

How to quantify the goodness?



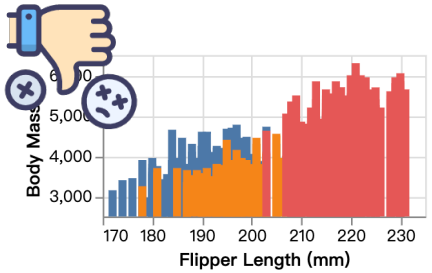


- **Learning from visualization examples**



- **Learning from expert knowledge**

a bar chart with more than 50 bars is hard to get insights



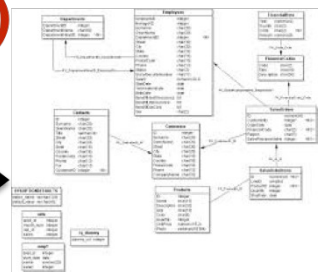
Basic Ideas

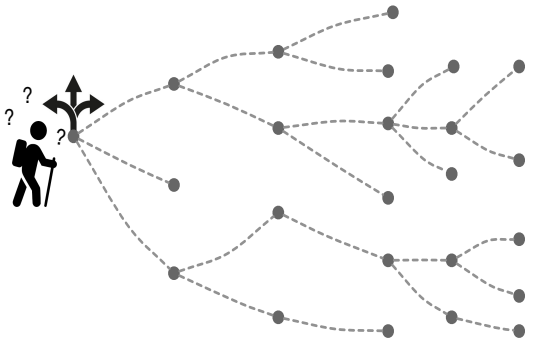
Huge search space for candidate visualizations

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itottaw	28

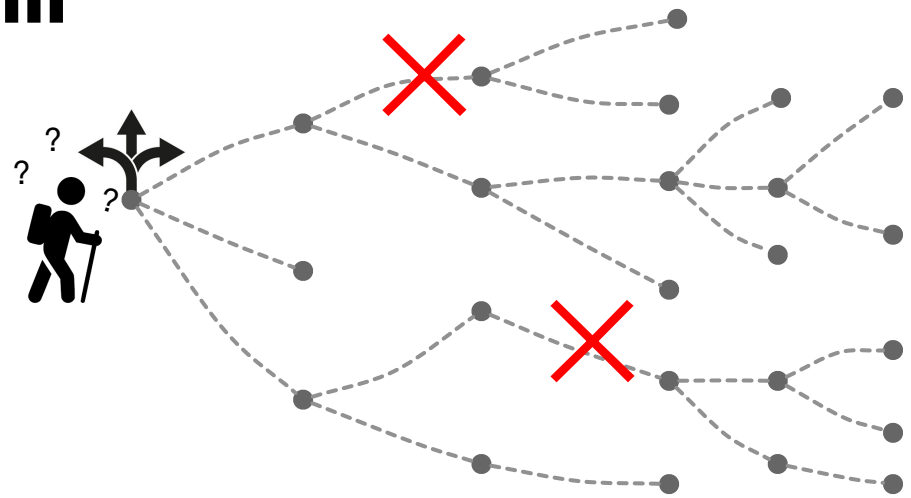
(DB problem)

transform



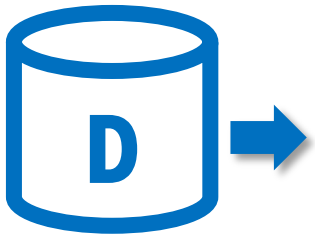


- **DB Optimization Problem**
  - Pruning techniques, …



DeepEye Framework

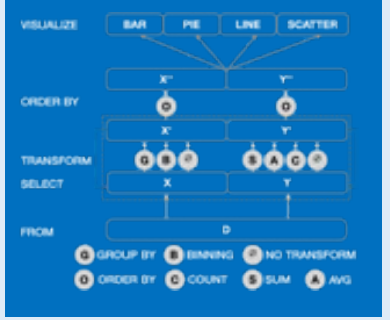
Input

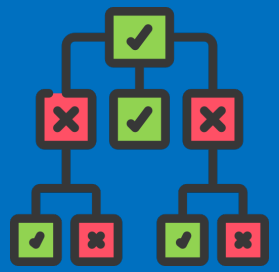


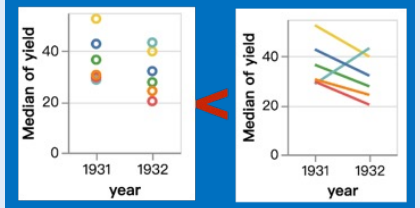
DeepEye

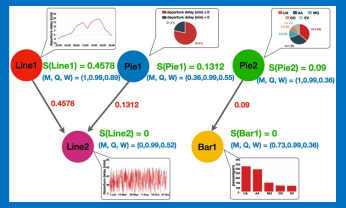
Online

Search Space	Recognition	Ranking	Optimizations
All Visualizations	Decision Tree	Learning to Rank Partial Orders	Rule-based Optimizations



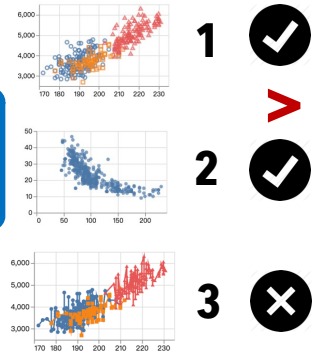






Offline

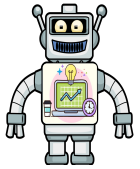
Examples

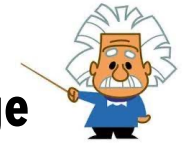


1 ✓ > 2 ✓

3 ✗

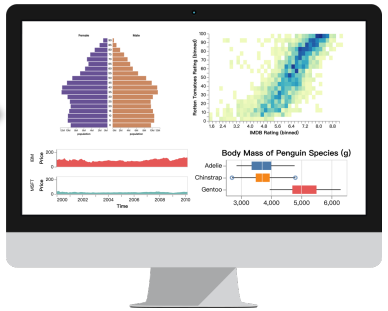
Machine Learning	Decision Tree, Learning to Rank
Expert Knowledge	Partial Orders, Search Rules





Output

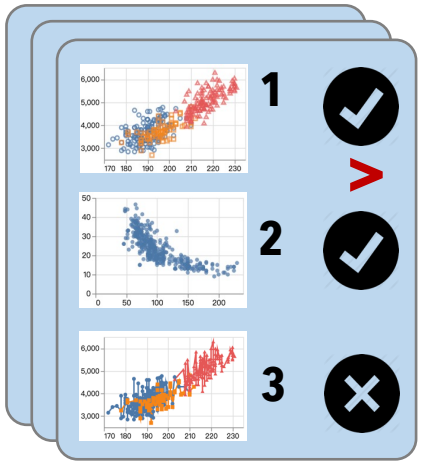
top-k visualizations



[Deepeye, ICDE 2018]

ML-powered AutoVIS

Learning to create and recommend good visualizations
based on good visualization examples

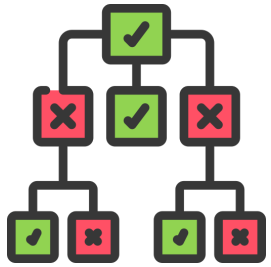


Visualization Corpus

- **42** real-world **datasets**
- 2520/30892 **good/bad labels**
- 285,236 **visualization pairs**

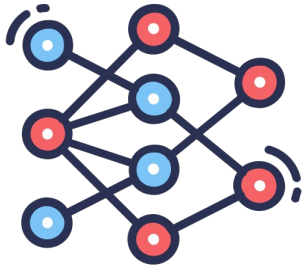
- #-distinct values
- #-tuples
- ratio of unique values
- max() and min() of values
- the data type,
- attribute correlation
- chart type

14 types of features



Binary Classifier
Decision Tree

Visualization Recognition

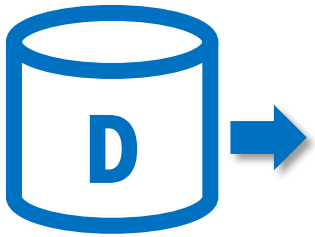


Learning-to-Rank
LambdaMART

Ranking & Recommendation

DeepEye Framework

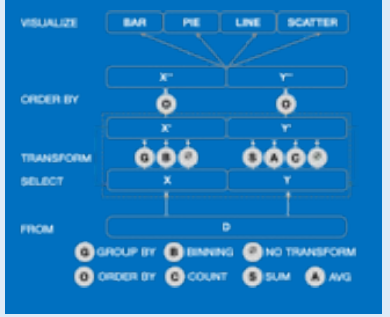
Input

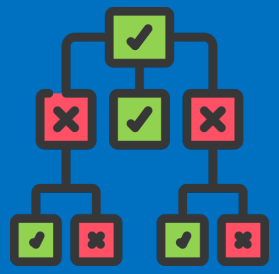


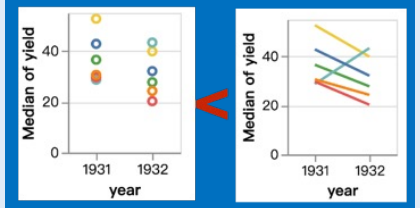
DEEPEYE

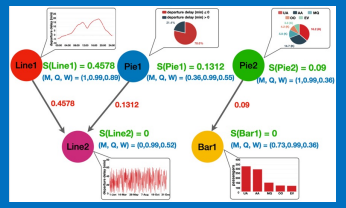
Online

Search Space	Recognition	Ranking	Optimizations
All Visualizations	Decision Tree	Learning to Rank Partial Orders	Rule-based Optimizations



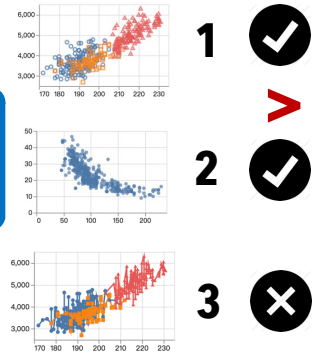




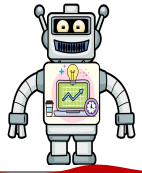


Offline

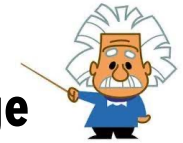
Examples



Machine Learning



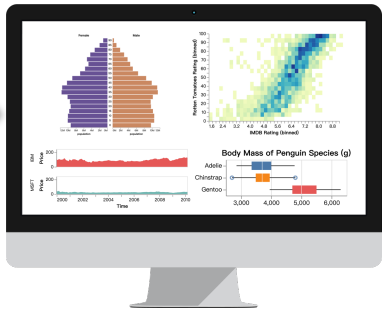
Expert Knowledge



- Decision Tree
- Learning to Rank
- Partial Orders
- Search Rules

Output

top-k visualizations



[Deepeye, ICDE 2018]

Partial Order-based Ranking & Recommendation

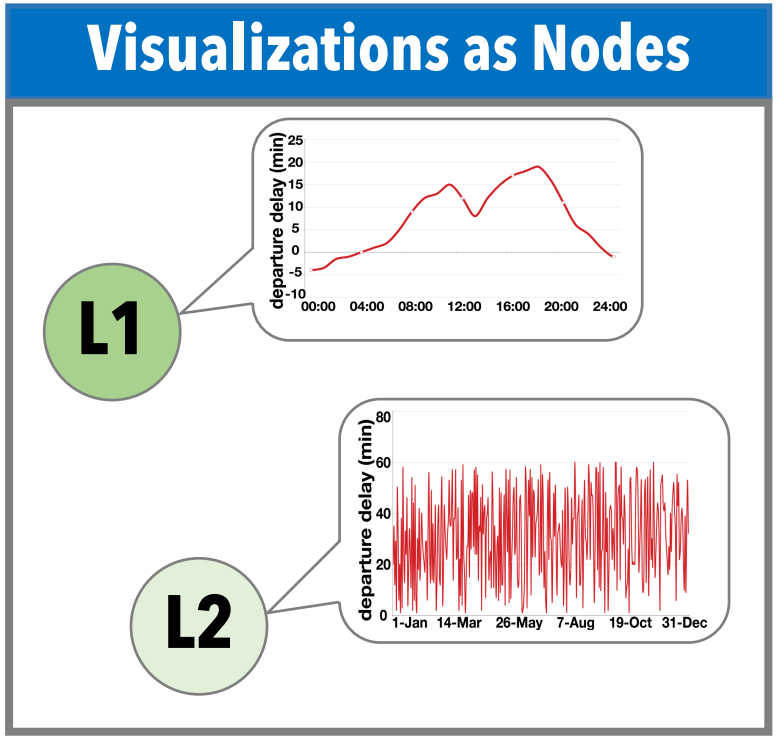
Why incorporate expert knowledge?

(1) **ML models** may **work poorly** in some aspects, **especially in domain-specific datasets**.

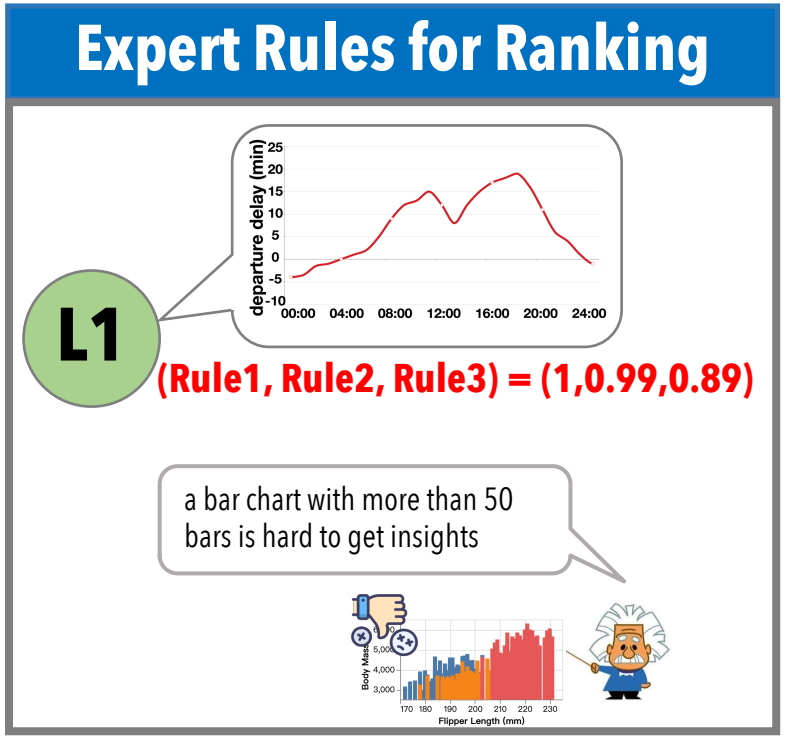
(2) **Hard to improve recommendation performance** due **to the black box of ML** models.

How does partial order work for visualization ranking?

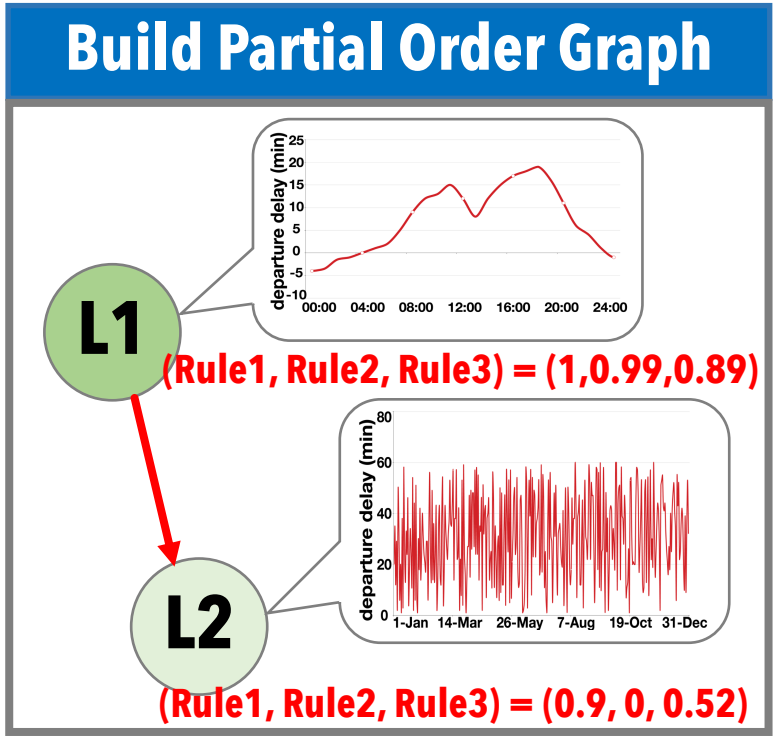
Visualizations as Nodes



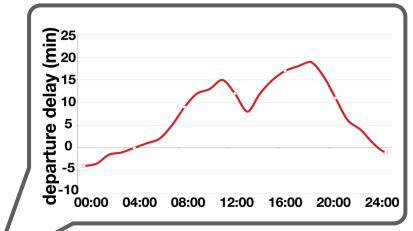
Expert Rules for Ranking



Build Partial Order Graph

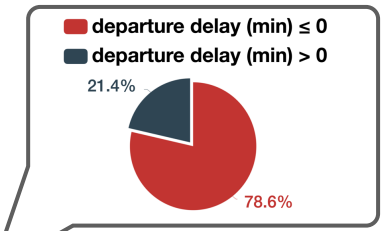


Partial Order-based Ranking & Recommendation



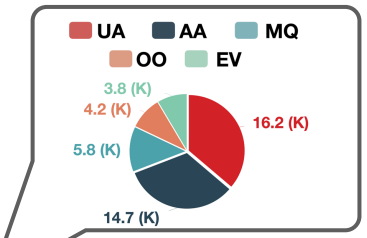
L1 S(L1) = 0.4578

(M, Q, W) = (1,0.99,0.89)



P1 S(P) = 0.1312

(M, Q, W) = (0.36,0.99,0.55)



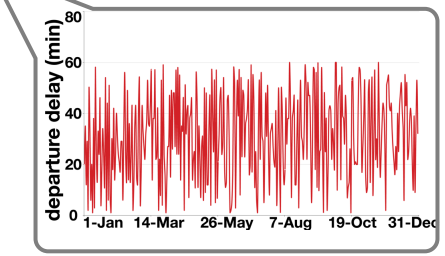
P2 S(P2) = 0.09

(M, Q, W) = (1,0.99,0.36)

"M(u) > M(v), Q(u) > Q(v), W(u) > W(v)" → u > v

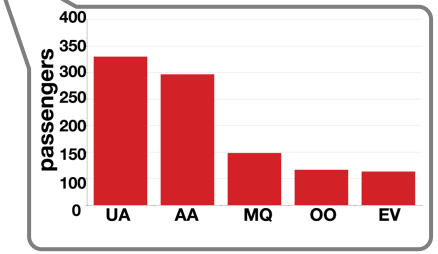
L2 S(L2) = 0

(M, Q, W) = (0.9, 0, 0.52)



B1 S(B1) = 0

(M, Q, W) = (0.73,0.99,0.36)



Top-3 Visualizations: L1, P1, P2

DEEPEYE: An End-to-End and ML-powered AutoVIS System

DEEPEYE System = AutoVIS + User Intent + Cleaned Data

Fully Automatic Data Visualization [ICDE'18, SIGMOD'23]

Intent-based Data Visualization [SIGMOD'21, IEEE VIS'21]

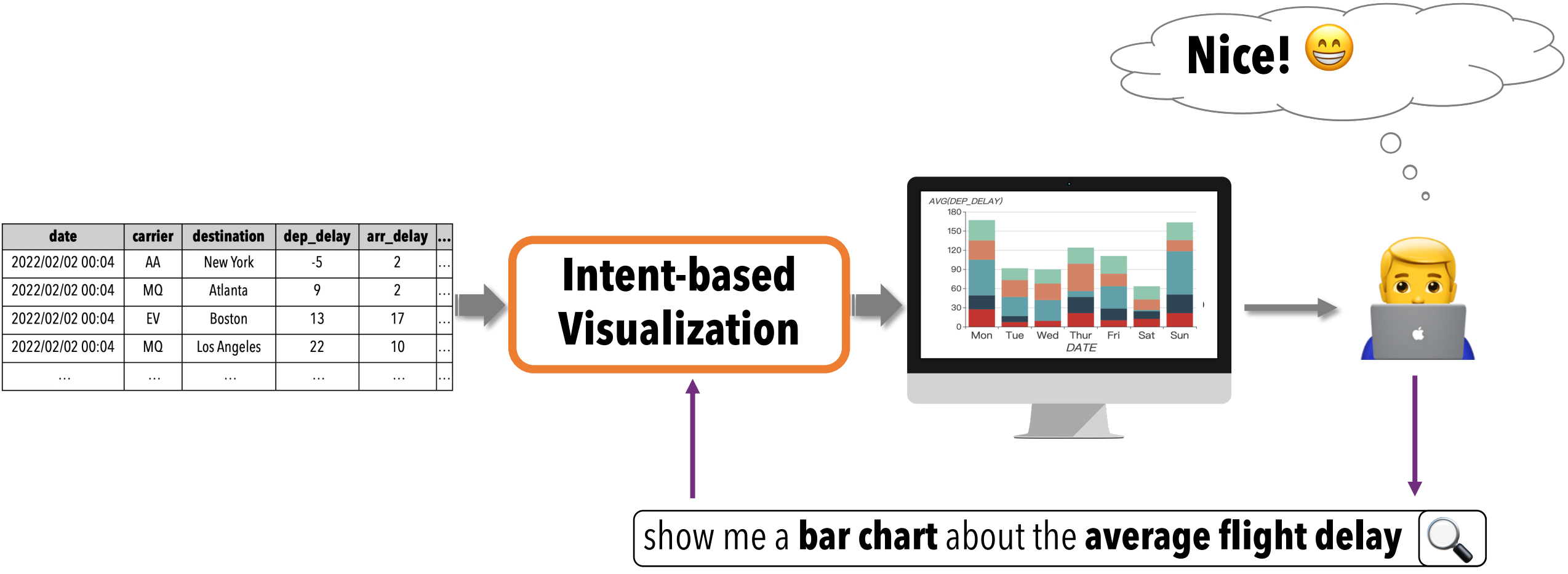
Quality-aware Data Visualization [ICDE'20, VLDB'20 demo]

How to lower the barriers?

Missing User Intent

Visualizations Meet Data Errors

Intent-based Data Visualization



Expressing **User Intent** through **Natural Language Query**

Intent-based Visualization

DeepEye | e.g., Show me line charts about electricity | Search | Undo (1) | Redo (0) | deepeye

Visualization Selection Approach

Diversified Top-k Sele

Dataset

Data: Flight Delay Statistics

Select Dataset

#Rows: 3757

#Columns: 6

Columns:

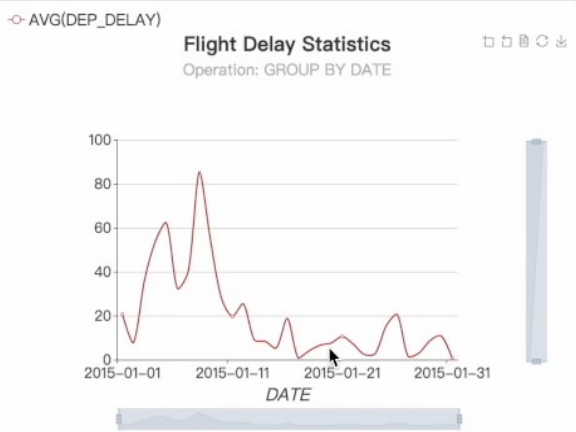
- DATE
- CARRIER
- DEST_CITY
- DEP_DELAY
- ARR_DELAY

Filter

DeepEye Recommendation: 53 visualizations (0.572 seconds)

A line chart with the x-axis DATE and y-axis the average DEP_DELAY

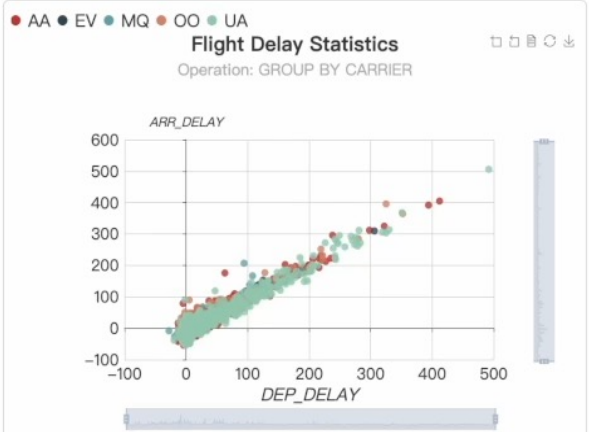
This line chart shows the change of the average DEP_DELAY over DATE, where the the average DEP_DELAY is grouped by attribute DATE



Zoom | Faceted

A scatter chart with the x-axis DEP_DELAY and y-axis ARR_DELAY

This scatter chart shows the distribution of DEP_DELAY and ARR_DELAY, where the ARR_DELAY is grouped by attribute CARRIER



Zoom | Faceted

A bar chart with the x-axis DATE and y-axis the amount of DATE

This bar chart shows the distribution of DATE and the

A bar chart with the x-axis DATE and y-axis the sum of PASSENGERS

This bar chart shows the distribution of DATE and the sum

Browsing History (0) :

Faceted Search: 0 Visualizations

The pixels of the screen are preferably better at 1366 x 768 or higher.

This demo is running on a cloud server with 1 Core CPU, 1GB Memory, and 1 Mbps Bandwidth.

APIs: Github

Expressing User Intent through Natural Language Query

show me a **bar chart** about the **average flight delay**

date	carrier	destination	dep_delay	arr_delay	...
2022/02/02 00:04	AA	New York	-5	2	...
2022/02/02 00:04	MQ	Atlanta	9	2	...
2022/02/02 00:04	EV	Boston	13	17	...
2022/02/02 00:04	MQ	Los Angeles	22	10	...
...	...	...	...	...	...

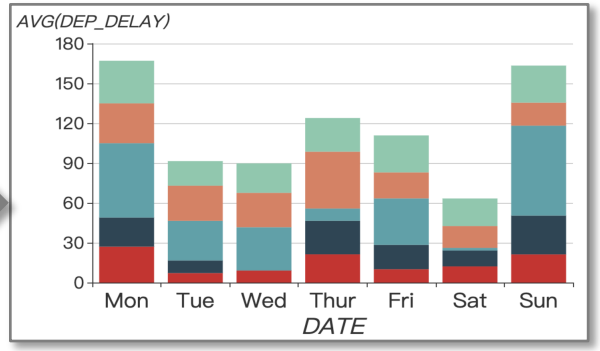
Tabular Data (e.g., flight delay)

Natural Language Query
(NL Query)

NL2VIS

```
{
  "data": {"url": "data/flightdelay.csv"},
  "mark": {"type": "bar"},
  "encoding": {
    "x": {
      "field": "date"
    },
    "y": {
      "aggregate": "avg",
      "field": "dep_delay"
    },
    "transform":{
      "bin": "weekday",
      "field": "date"
    },
    "color": {"field": "carrier"}
  }
}
```

Visualization Query
(VIS Query)



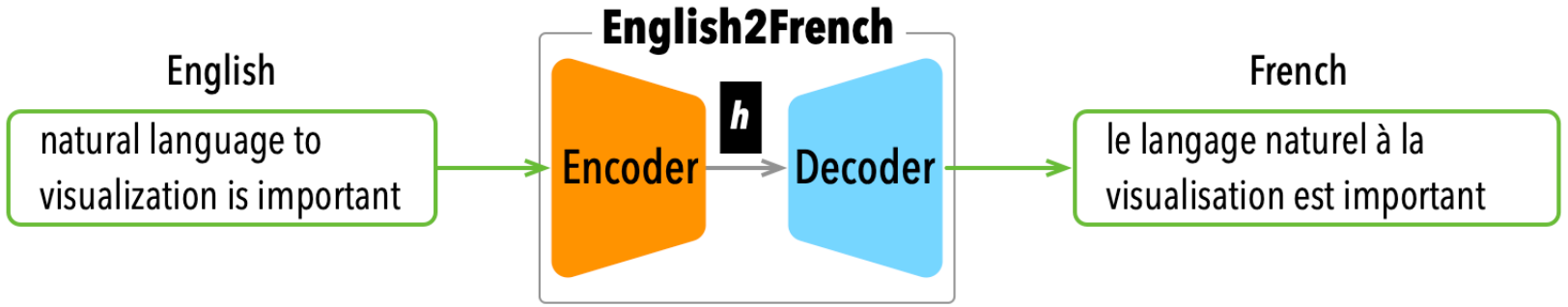
Visualization

Challenges:

- How to **understand the user intent** from the natural language query?
- How to **accurately create** corresponding **visualizations based on intent**?

NL2VIS by Neural Machine Translation

- **A Transformer-based sequence-to-sequence (seq2seq) model**



- **Hierarchical** visualization language
- **Hard** for model processing

Table: COVID-19 in United States

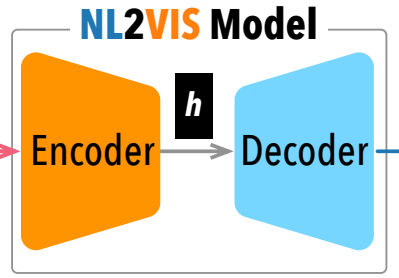
date	states	cases	number
2021-03-08	California	confirmed	3599250
2021-03-08	California	deaths	54220
2021-03-08	New York	confirmed	1694651
2021-03-08	New York	deaths	48335

D

Natural Language Query

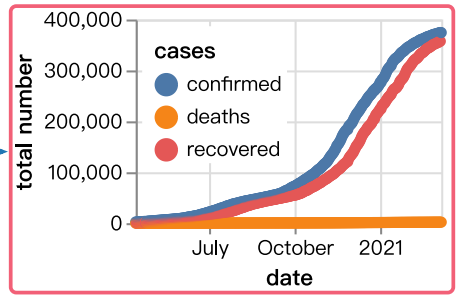
N

Show me the trend of confirmed, died, and recovered cases in Utah




```
{
  "data": {"url": "US_States.csv"},
  "mark": "line",
  "encoding": {
    "x": {"field": "date", "type": "date"},
    "y": {"field": "number", "type": "quantitative"},
    "color": {"field": "cases", "type": "nominal"},
  }
}
```

Vega-Lite Specification



Visualization Result

***Yuyu Luo**, et al. Natural Language to Visualization by Neural Machine Translation. **IEEE VIS 2021** (**Full** Papers, **50 Citations**)*

NL2VIS by Neural Machine Translation

Vega-Zero: a sequence-based grammar for model-friendly, by simplifying Vega-Lite

Vega-Lite

```
{
  "data": {"url": "US_States.csv"},
  "mark": "line",
  "encoding": {
    "x": {"field": "date", "type": "date"},
    "y": {"field": "number", "type": "quantitative"},
    "color": {"field": "cases", "type": "nominal"},
   }
}
```

Vega-Zero

mark line data US_States.csv encoding x date y aggregate none number color cases

Table: COVID-19 in United States

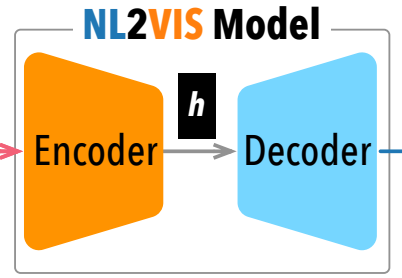
date	states	cases	number
2021-03-08	California	confirmed	3599250
2021-03-08	California	deaths	54220
2021-03-08	New York	confirmed	1694651
2021-03-08	New York	deaths	48335

D

Natural Language Query

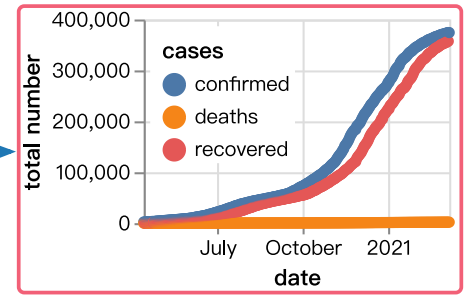
N

Show me the trend of confirmed, died, and recovered cases in Utah



mark line data US_States.csv encoding x date y aggregate none number color cases

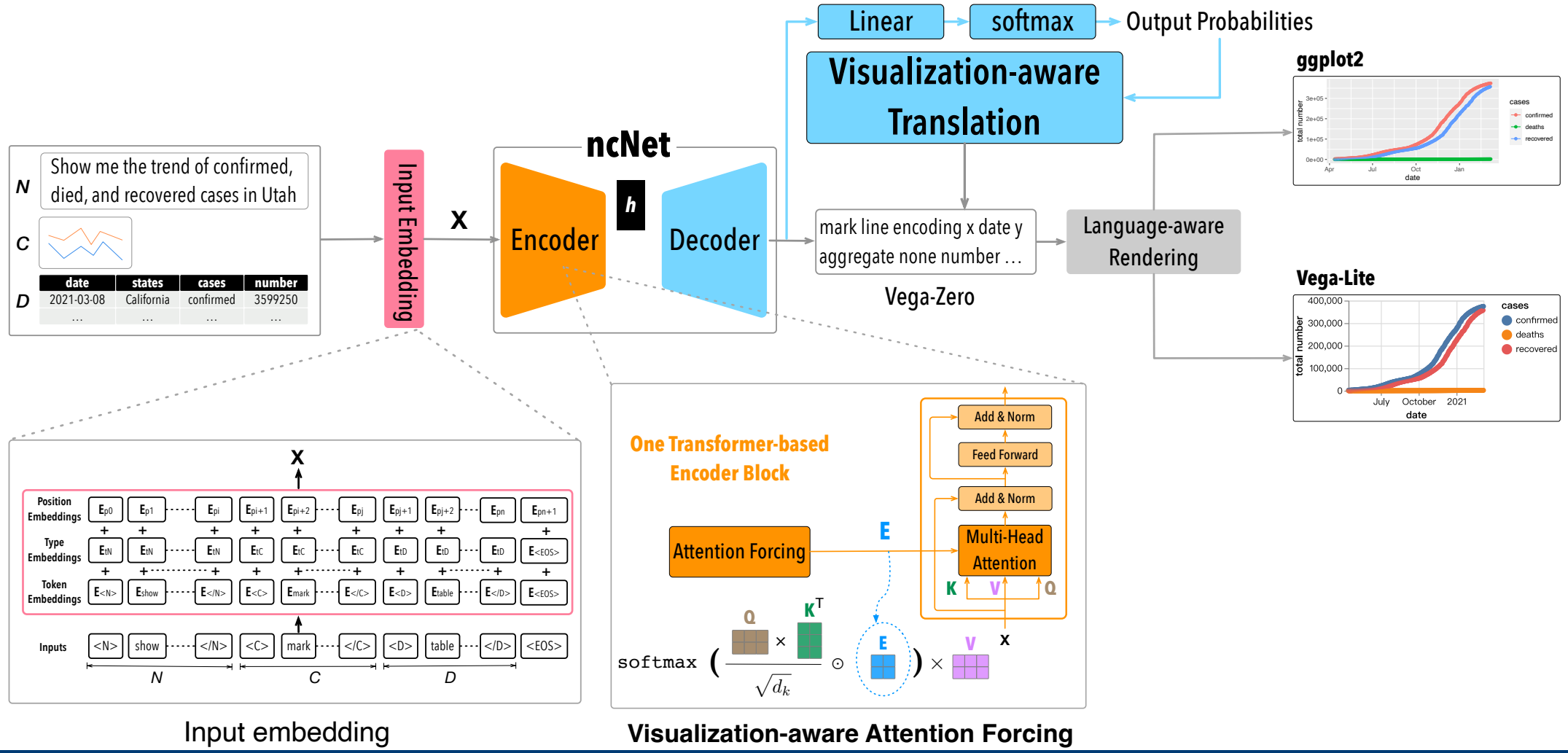
Vega-Zero Query



Visualization Result

NL2VIS by Neural Machine Translation

- **A Transformer-based model with several optimizations**



Input embedding

Visualization-aware Attention Forcing

nvBench: NL2VIS Benchmark

We develop **the first large-scale** benchmark for **cross-domain natural language to visualization task**

- 780 Tables from 153 databases, 25,750 (NL, VIS) pairs

Database: products_for_hire

Table: payments

...	payment_type_code	amount_paid_in_full_yn	payment_date	aoumnt_due	amount_paid
...	Check	1	2018-03-09 16:28:00	369.52	206.27
...	Cash	1	2018-03-03 13:39:44	278.60	666.45
...	Credit Card	0	2018-03-22 15:00:23	840.06	135.70
...	Check	0	2018-03-22 02:28:11	678.29	668.40
...	...	...	...	...	...

NL Query:

① What are the payment date of the payment with amount paid higher than 300 or with payment type is 'Check, and count them by a line chart

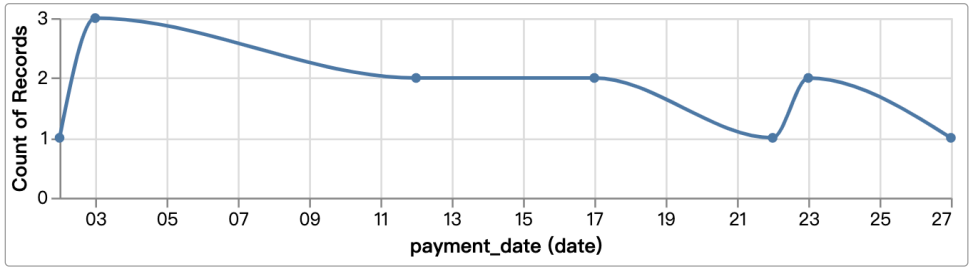
Vega-Zero Query:

```
mark      line
data      payments
encoding  x payment_date y aggregate count payment_date
transform filter amount_paid > 300 or payment_type_code='check' group x
```

Vega-Lite Query:

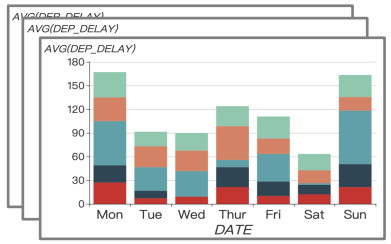
```
{
  "data": {"values": payments},
  "mark": {
     "type": "line", "point": true, "interpolate": "monotone"},
  "transform": [{
    "filter": "datum.amount_paid > 300 |
            datum.payment_type_code == 'check'"
  }],
  "encoding": {
    "x": {"field": "payment_date", "type": "temporal", "timeUnit": "date"},
    "y": {"field": "payment_date", "aggregate":"count", "type": "temporal"}
  }
}
```

Visualization:

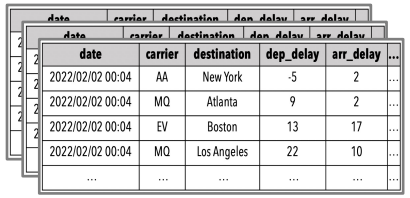


NL2VIS Benchmark

- **How to build the NL2VIS benchmark?**

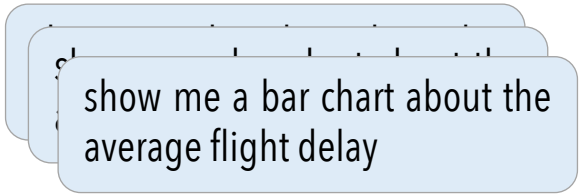


Visualization



date	carrier	destination	dep_delay	arr_delay	...
2022/02/02 00:04	AA	New York	-5	2	...
2022/02/02 00:04	MQ	Atlanta	9	2	...
2022/02/02 00:04	EV	Boston	13	17	...
2022/02/02 00:04	MQ	Los Angeles	22	10	...
...	...	...	...	...	...

Dataset



Natural Language Query

- **Low coverage**
- **High human cost**
- **Not extensible**

***Yuyu Luo**, et al. Synthesizing Natural Language to Visualization (NL2VIS) Benchmarks from NL2SQL Benchmarks. **SIGMOD 2021** (**Full** Papers, **16 Citations**)*

Synthesizing NL2VIS Benchmarks from NL2SQL Benchmarks.

There are dozens of **NL2SQL** benchmarks.

NL2SQL Benchmarks	#-Questions	#-SQL	#-Databases	#-Domains
ATIS	5,280	947	1	1
GeoQuery	877	247	1	1
Scholar	817	193	1	1
Academic	196	185	1	1
IMDB	131	89	1	1
Yelp	128	110	1	1
Advising	3,898	208	1	1
Restaurants	378	378	1	1
WikiSQL	80,654	77,840	26,521	/
Spider	10,181	5,693	200	138

Can we reuse these NL2SQL benchmarks that have spent a lot of effort annotated by experts to synthesize the NL2VIS benchmark?

Synthesizing NL2VIS Benchmarks from NL2SQL Benchmarks

SQL Query and **Visualization Query** have strong semantic connections.

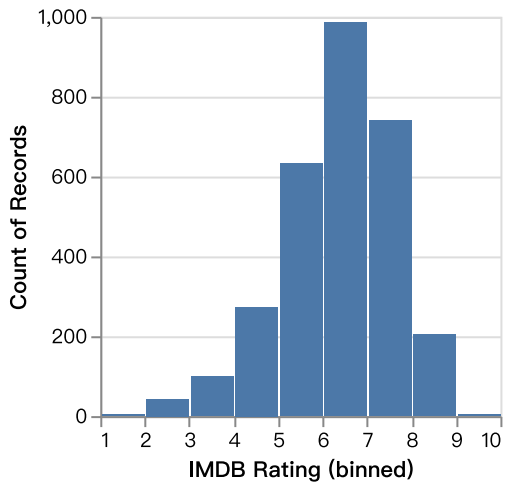
- **SQL: What data is needed?**
- **VIS: What data is needed? + How to visualize the data?**

Examples:

"What" data is needed

"How" to visualize the data

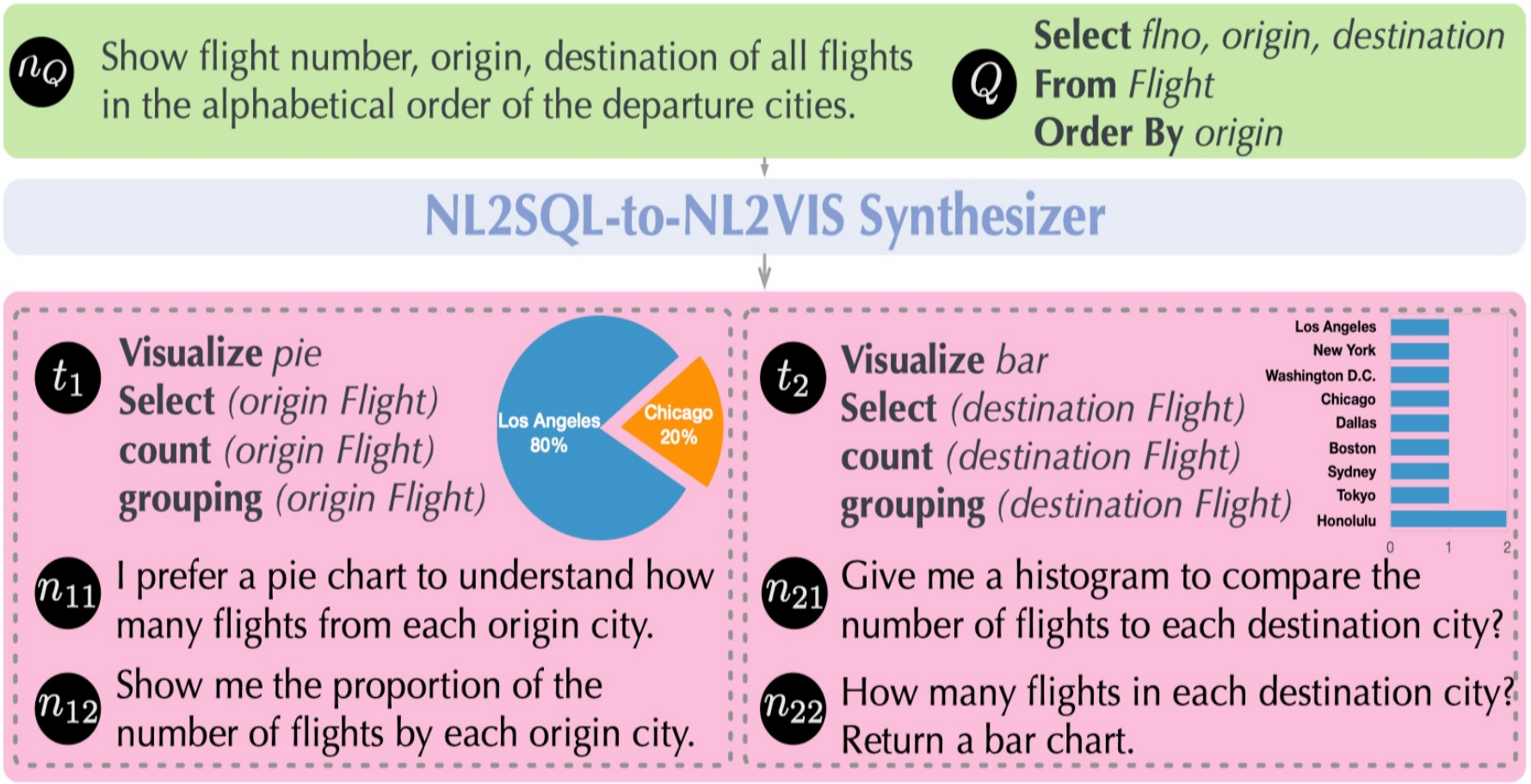
```
{
  "data": {"url": "data/movies.json"},
  "transform": [{
    "bin": true,
    "field": "IMDB Rating",
    "as": "binned rating"
    }],
  "mark": "bar",
  "encoding": {
    "x": {
      "field": "binned rating",
      "title": "IMDB Rating (binned)",
      "bin": {"binned": true, "step": 1}
    },
    "x2": {"field": "binned rating_end"},
    "y": {"aggregate": "count"}
  }
}
```



(a) Vega-Lite Visualization Specification

Piggyback the NL2SQL Benchmark

NL2SQL Benchmark → **NL2SQL-to-NL2VIS Synthesizer** → **NL2VIS Benchmark**



Semi-automatic

DEEPEYE: An End-to-End and ML-powered AutoVIS System

DEEPEYE System = AutoVIS + User Intent + Cleaned Data

Fully Automatic
Data Visualization
[ICDE'18, SIGMOD'23]

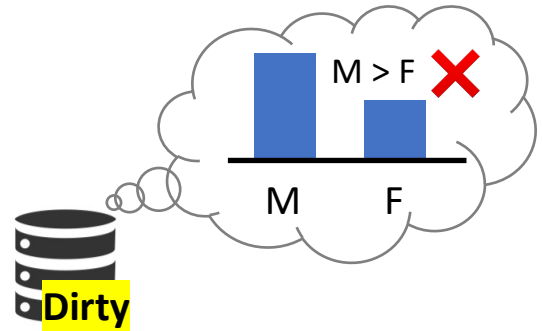
Intent-based
Data Visualization
[SIGMOD'21, IEEE VIS'21]

Quality-aware
Data Visualization
[ICDE'20, VLDB'20 demo]

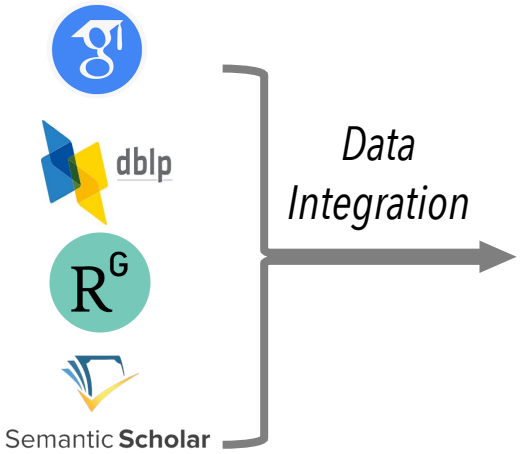
How to lower the barriers?

Missing User Intent

Visualizations Meet Data Errors



Visualizations Meet Data Errors



Data Integration

An Excerpt of **Publications Dataset** (Dirty)

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

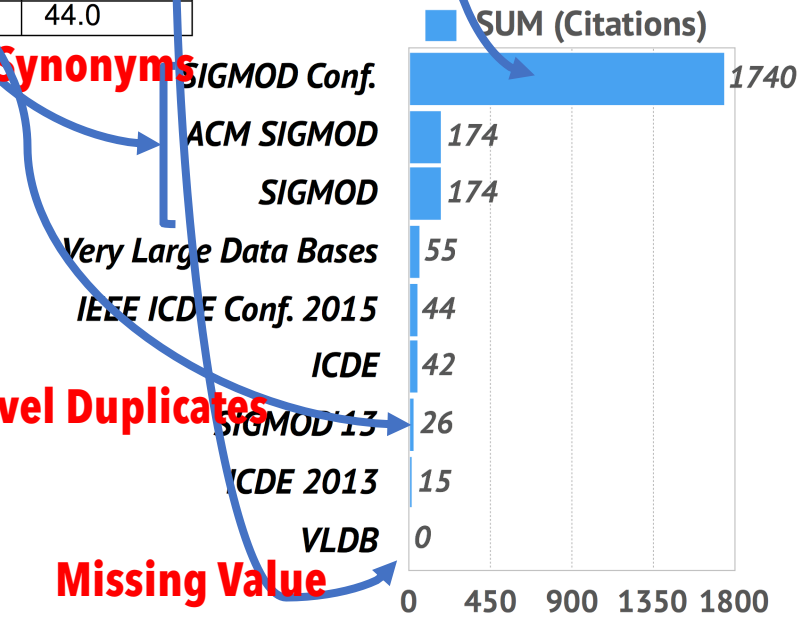
Visualize a ***bar chart*** that
X-axis is the ***Venue*** and
Y-axis is the ***total Citations*** of each Venue.

Outlier

Synonyms

Tuple-level Duplicates

Missing Value



Visualizations Meet Data Errors

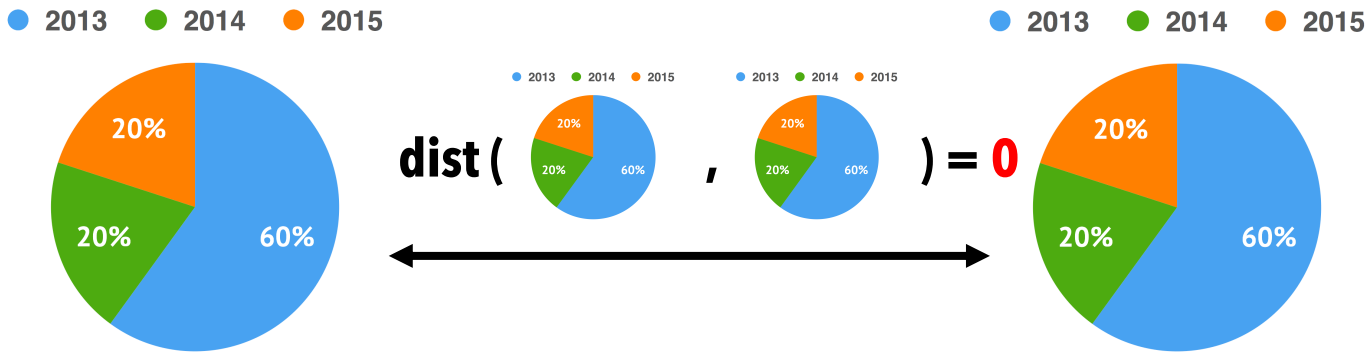
An Excerpt of **Publications Dataset** (Dirty)

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

Fully Cleaned Data

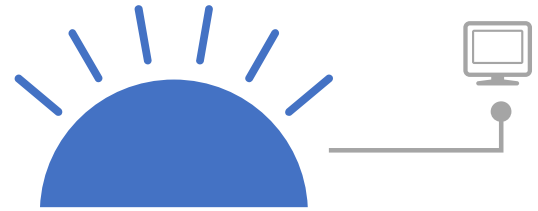
Id	Year	Title (abbr.)	Venue	Affiliation	Citations
t_{123}	2013	NADEEF	SIGMOD	QCRI	***174.0***
t_4	2013	KuaFu	ICDE	Microsoft	15.0
t_{56}	2013	TsingNUS	SIGMOD	Tsinghua	13.0
t_{78}	2014	SeeDB	VLDB	Stantford Univ.	***55.0***
t_{910}	2015	Elaps	ICDE	NUS	43.0

Visualize a *pie chart* that shows the proportions of *#-Papers* by *Year*.



To Clean or Not to Clean?

To Clean or Not to Clean?

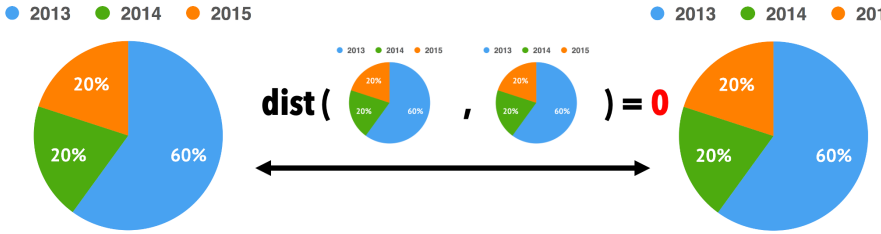


Observations

Data cleaning is expensive

- *Human cost*
- *Time-consuming*

Some data errors do not impact visualizations



How about just cleaning up visualization-aware data errors?

- is cheaper and more flexible than cleaning the entire dataset

Visualization-aware Data Cleaning

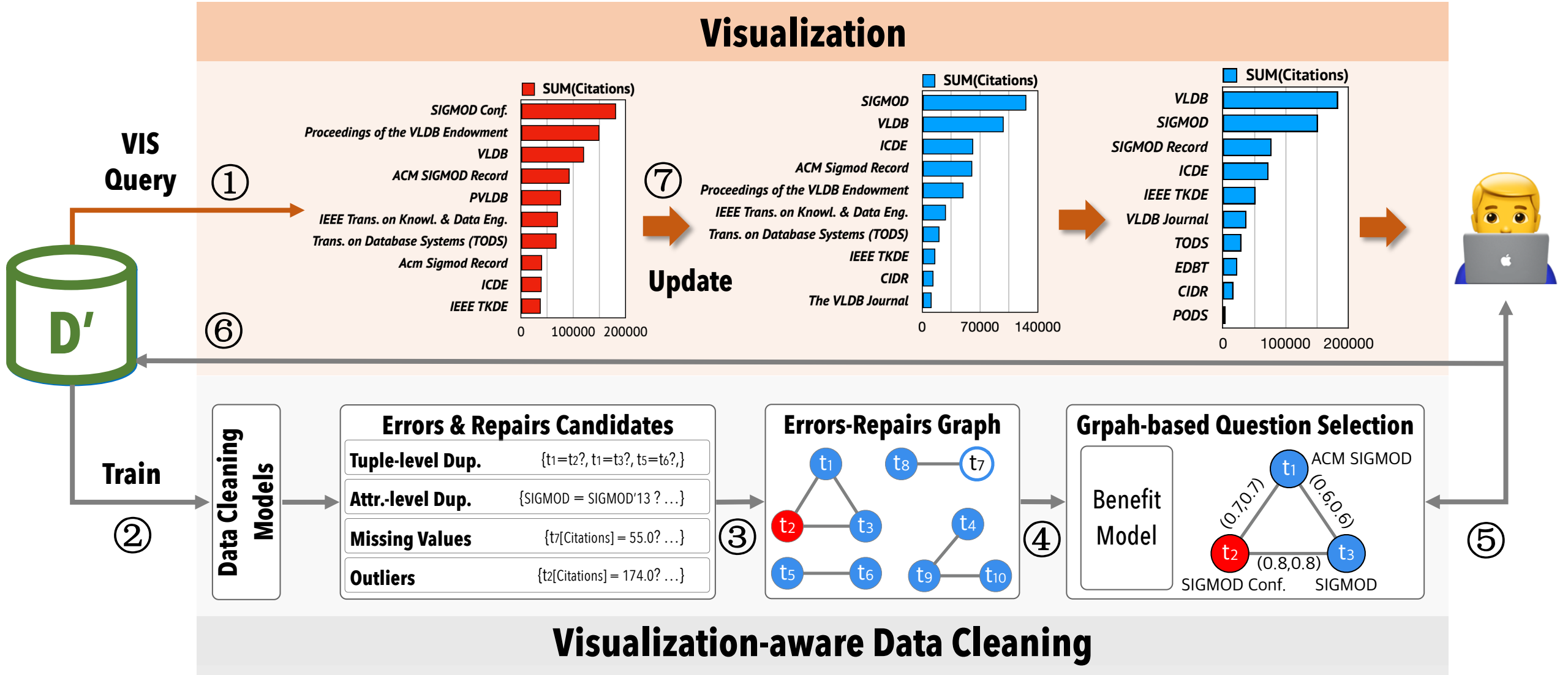
Goal 1: progressively improve the visualization quality ↑

Goal 2: minimize the human cost ↓

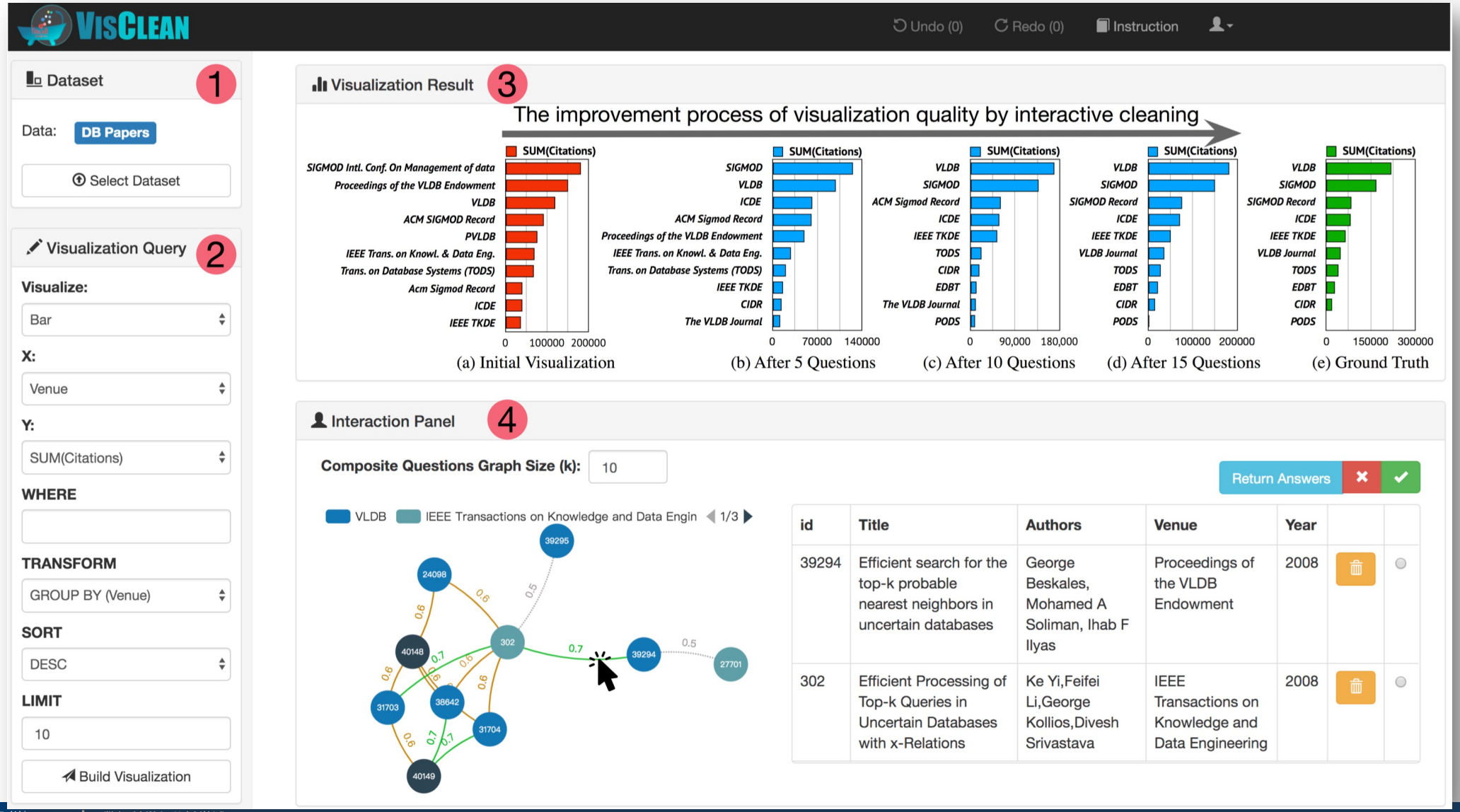
Problem: ***Visualization-aware data cleaning***

- to **progressively improve the quality of visualization**
- by ***minimizing the cost*** **of interacting with the user** to *clean the visualization relevant data*.

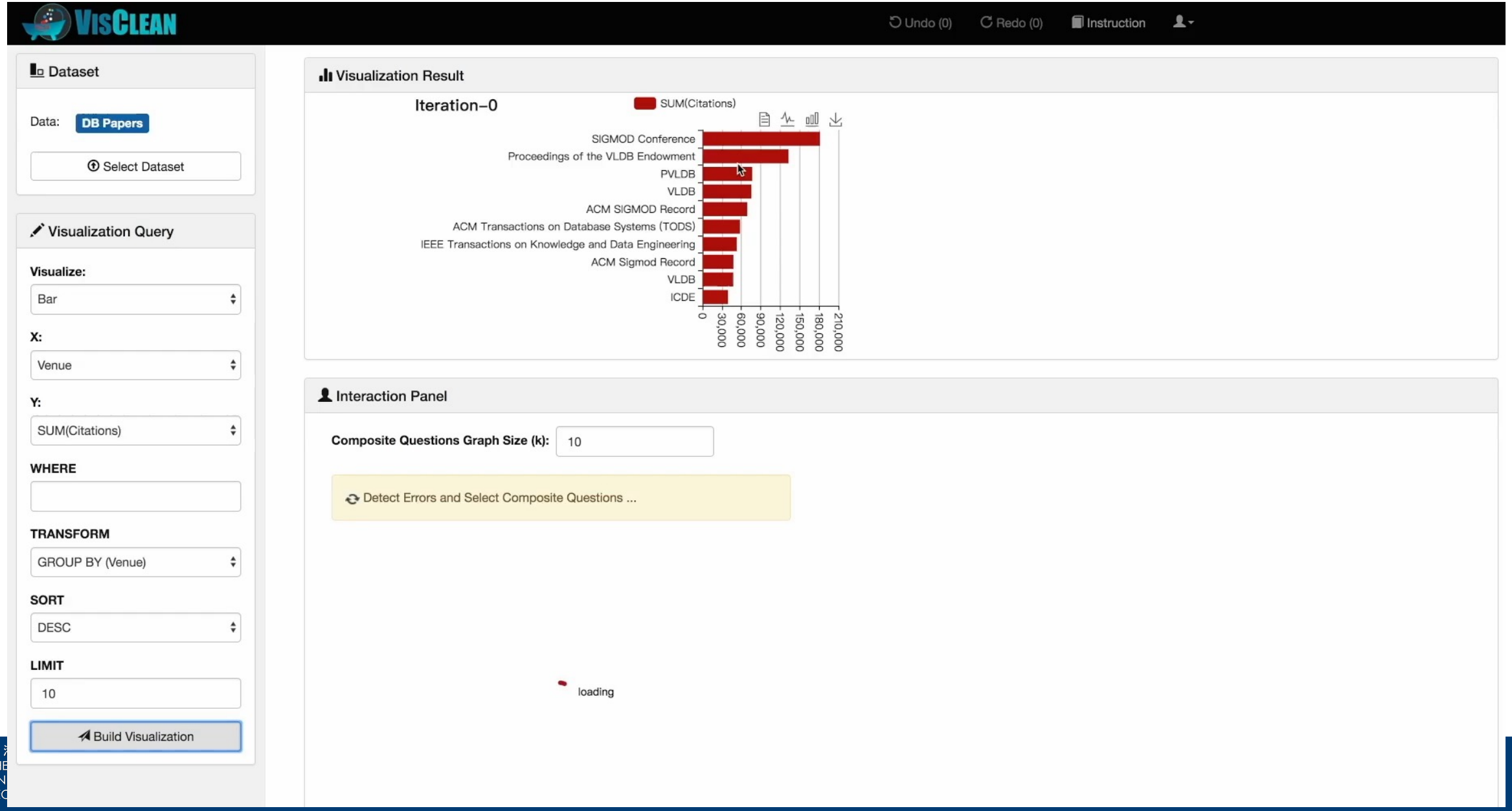
VisClean: Visualization-aware Data Cleaning



VisClean: Visualization-aware Data Cleaning



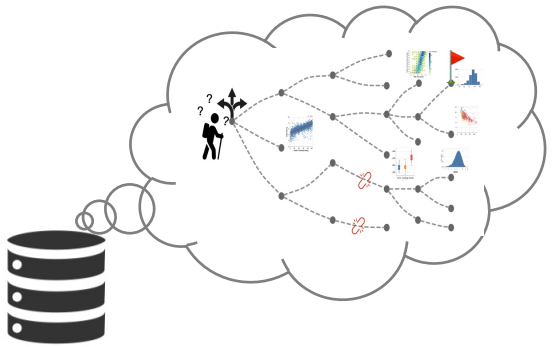
VisClean: Visualization-aware Data Cleaning

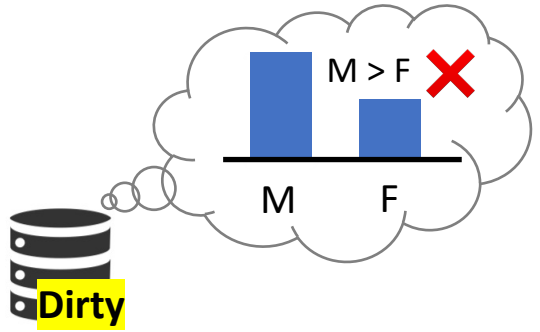


DEEPEYE: An End-to-End and ML-powered AutoVIS System

DEEPEYE System = AutoVIS + User Intent + Cleaned Data

Fully Automatic Data Visualization [ICDE'18, SIGMOD'23]	Intent-based Data Visualization [SIGMOD'21, IEEE VIS'21]	Quality-aware Data Visualization [ICDE'20, VLDB'20 demo]
How to lower the barriers?	Missing User Intent	Visualizations Meet Data Errors





Visualization System = Visualization Process + User + Data

Thanks!

Have fun, learn stuff

Dr. Yuyu LUO

Data Science and Analytics Thrust

Information Hub, HKUST(GZ)

yuyuluo@hkust-gz.edu.cn

http://luoyuyu.vip