# Lecture 12: Case I: Visualization Synthesis

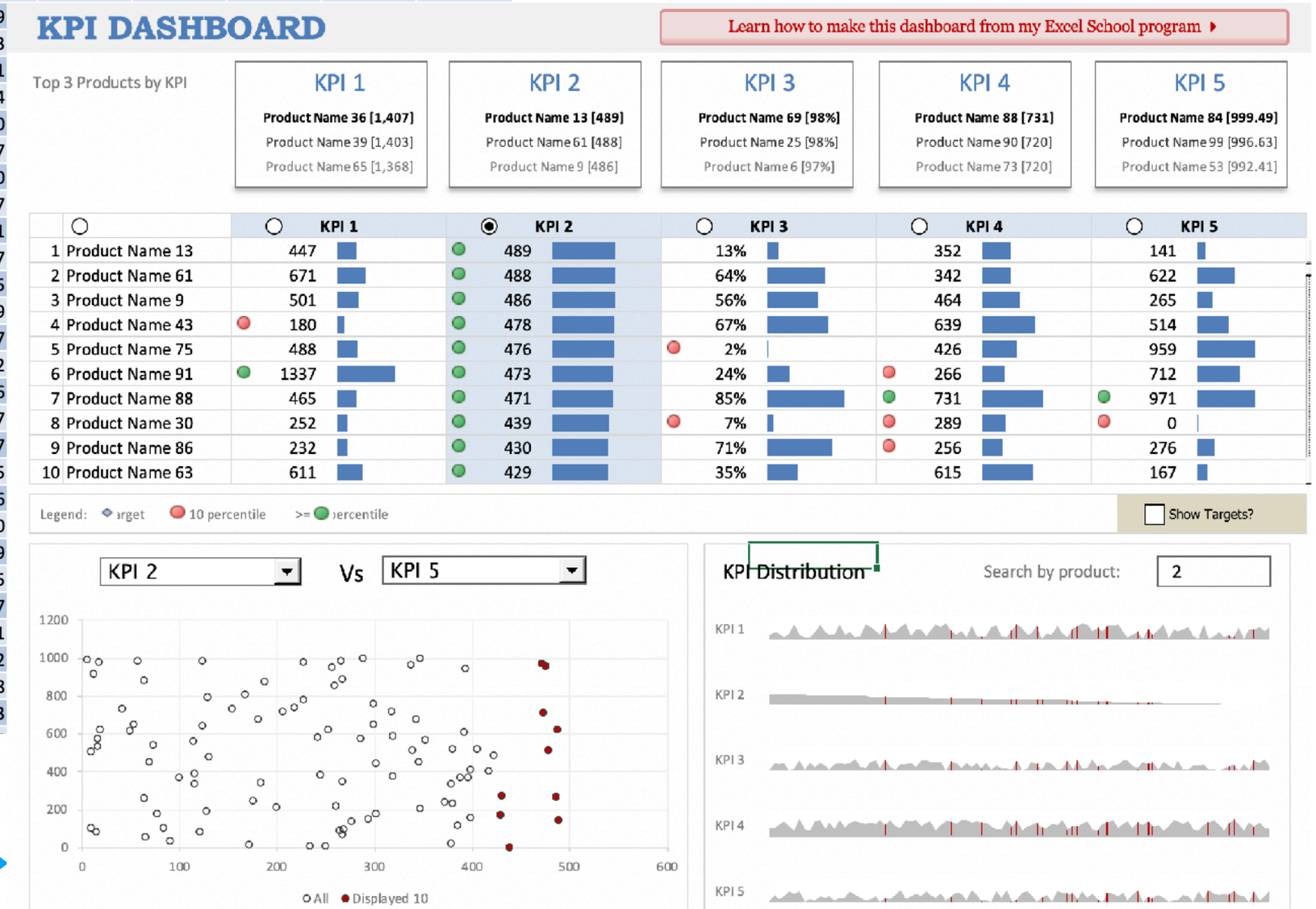
Yu Feng
Spring 2021

# Summary of previous lecture

- Proposal is due today
- Synthesis with statistical models

# Business

| No. | Products | KPI 1 | KPI 2 | KPI 3 | KPI 4 | KPI 5 | tKPI1 | tKPI2 | tKPI3 | tKPI4 | tKPI5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Product Name 1 | 284 | 267 | 0.28 | 318 | 348.83 | 203.4 | 264.1 | 0.29 | 248.1 | 355.2 |
| 2 | Product Name 2 | 170 | 218 | 0.86 | 295 | 734.27 | 46.2 | 112.1 | 0.72 | 315.8 | 262.0 |
| 3 | Product Name 3 | 760 | 9 | 0.95 | 372 | 503.34 | 1132.0 | 13.0 | 1.32 | 375.1 | 411.4 |
| 4 | Product Name 4 | 366 | 388 | 0.35 | 578 | 367.9 | 86.9 | 395.2 | 0.06 | 295.3 | 110.8 |
| 5 | Product Name 5 | 1345 | 130 | 0.58 | 352 | 477.47 | 1373.2 | 184.4 | 0.50 | 151.1 | 655.4 |
| 6 | Product Name 6 | 790 | 181 | 0.97 | 295 | 678.05 | 326.9 | 214.7 | 1.07 | 104.7 | 773.0 |
| 7 | Product Name 7 | 1269 | 319 | 0.78 | 285 | 373.29 | 952.8 | 29.0 | 0.20 | 198.0 | 74.2 |
| 8 | Product Name 8 | 107 | 16 | 0.59 | 280 | 532.21 | 64.1 | 0.1 | 0.80 | 200.5 | 559.9 |
| 9 | Product Name 9 | 501 | 486 | 0.56 | 464 | 265 | 373.6 | 413.3 | 0.35 | 317.2 | 375.4 |
| 10 | Product Name 10 | 953 | 259 | 0.05 | 260 | 855.81 | 315.7 | 257.9 | 0.04 | 245.9 | 974.6 |
| 11 | Product Name 11 | 783 | 299 | 0.18 | 711 | 649 | | | | | |
| 12 | Product Name 12 | 669 | 124 | 0.22 | 609 | 983 | | | | | |
| 13 | Product Name 13 | 447 | 489 | 0.13 | 352 | 141 | | | | | |
| 14 | Product Name 14 | 682 | 417 | 0.41 | 330 | 404 | | | | | |
| 15 | Product Name 15 | 807 | 77 | 0.91 | 643 | 180 | | | | | |
| 16 | Product Name 16 | 1180 | 267 | 0.43 | 318 | 887 | | | | | |
| 17 | Product Name 17 | 725 | 172 | 0.57 | 540 | 10 | | | | | |
| 18 | Product Name 18 | 522 | 227 | 0.22 | 669 | 77 | | | | | |
| 19 | Product Name 19 | 1350 | 398 | 0.68 | 677 | 411 | | | | | |
| 20 | Product Name 20 | 1163 | 168 | 0.08 | 610 | 807 | | | | | |
| 21 | Product Name 21 | 830 | 264 | 0.1 | 396 | 85 | | | | | |
| 22 | Product Name 22 | 1195 | 199 | 0.92 | 439 | 209 | | | | | |
| 23 | Product Name 23 | 482 | 116 | 0.71 | 425 | 337 | | | | | |
| 24 | Product Name 24 | 1024 | 176 | 0.35 | 268 | 242 | | | | | |
| 25 | Product Name 25 | 1339 | 379 | 0.98 | 494 | 336 | | | | | |
| 26 | Product Name 26 | 958 | 206 | 0.1 | 443 | 717 | | | | | |
| 27 | Product Name 27 | 1153 | 250 | 0.57 | 576 | 7 | | | | | |
| 28 | Product Name 28 | 136 | 57 | 0.07 | 607 | 985 | | | | | |
| 29 | Product Name 29 | 923 | 116 | 0.61 | 622 | 386 | | | | | |
| 30 | Product Name 30 | 252 | 439 | 0.07 | 289 | 0 | | | | | |
| 31 | Product Name 31 | 1270 | 64 | 0.53 | 487 | 259 | | | | | |
| 32 | Product Name 32 | 880 | 12 | 0.9 | 395 | 915 | | | | | |
| 33 | Product Name 33 | 1330 | 385 | 0.25 | 550 | 117 | | | | | |
| 34 | Product Name 34 | 255 | 128 | 0.87 | 644 | 191 | | | | | |
| 35 | Product Name 35 | 253 | 266 | 0.19 | 706 | 982 | | | | | |
| 36 | Product Name 36 | 1407 | 302 | 0.02 | 634 | 178 | | | | | |
| 37 | Product Name 37 | 359 | 269 | 0.7 | 571 | 93 | | | | | |




https://chandoo.org/wp/kpi-dashboard-revisited/

# *Social Science*

## HOUSING MARKET SNAPSHOT

State of Washir

## EXISTING HOME

## MEDIAN HOME PRICES

State of Washington and Counties
Time Trend

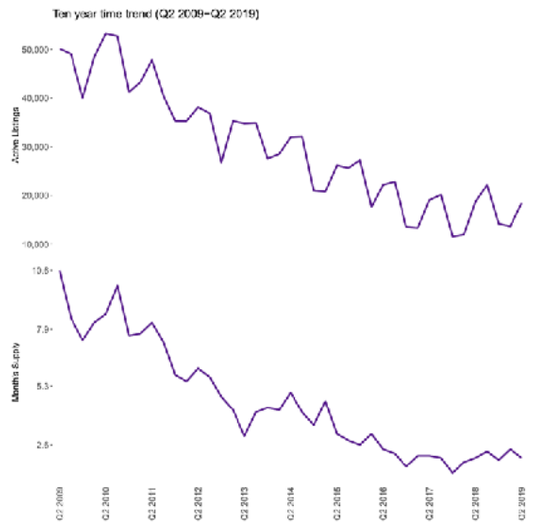
State of Was
Seasonally A

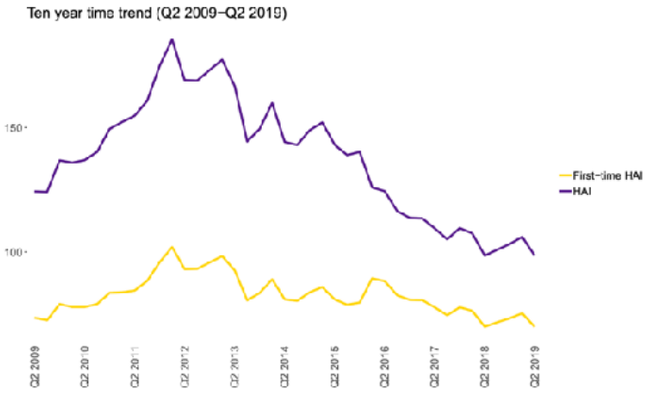
## EXISTING HOME SALES

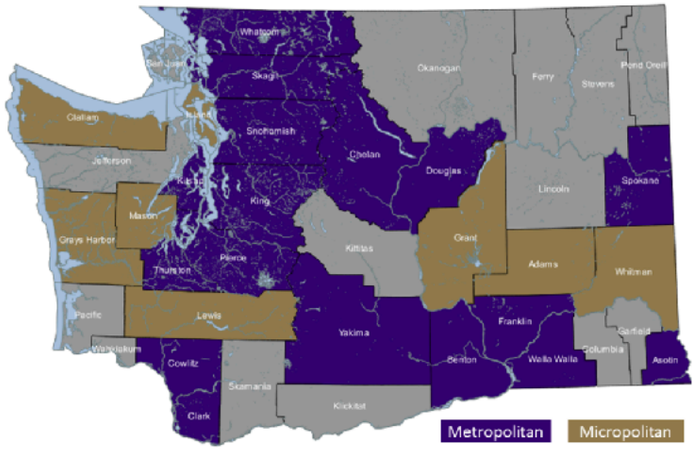
State of Washington and Counties
Seasonally Adjusted Annual Rate

| County | Q2 2018 | Q3 2018 | Q4 2018 | Q1 2019 | Q2 2019 | % Change by qtr | % Change by year |
|---|---|---|---|---|---|---|---|
| Adams | 160 | 170 | 170 | 170 | 170 | 0.0 | 6.2 |
| Asotin | 270 | 270 | 290 | 260 | 240 | -7.7 | -11.1 |
| Benton | 4,100 | 4,100 | 4,270 | 4,070 | 3,740 | -8.1 | -8.8 |
| Chelan | 1,030 | 1,010 | 990 | 960 | 920 | -4.2 | -10.7 |
| Clallam | 1,120 | 1,110 | 1,120 | 1,080 | 1,050 | -2.8 | -6.2 |
| Clark | 7,340 | 7,300 | 7,120 | 6,830 | 6,900 | 1.0 | -6.0 |
| Columbia | 130 | 130 | 120 | 100 | 80 | -20.0 | -38.5 |
| Cowlitz | 1,520 | 1,530 | 1,520 | 1,420 | 1,390 | -2.1 | -8.6 |
| Douglas | 660 | 660 | 640 | 650 | 580 | -10.8 | -12.1 |
| Ferry | 110 | 110 | 100 | 100 | 110 | 10.0 | 0.0 |
| Franklin | 1,380 | 1,380 | 1,430 | 1,370 | 1,250 | -8.8 | -9.4 |
| Garfield | 50 | 50 | 60 | 50 | 50 | 0.0 | 0.0 |
| Grant | 1,060 | 1,080 | 1,090 | 1,070 | 1,030 | -3.7 | -2.8 |
| Grays Harbor | 1,920 | 1,950 | 1,890 | 1,900 | 1,910 | 0.5 | -0.5 |
| Island | 2,170 | 2,090 | 2,040 | 1,970 | 1,960 | -0.5 | -9.7 |
| Jefferson | 700 | 710 | 680 | 640 | 650 | 1.6 | -7.1 |
| King | 27,640 | 27,080 | 26,090 | 25,030 | 25,750 | 2.9 | -6.8 |
| Kitsap | 5,050 | 5,100 | 4,880 | 4,690 | 4,660 | -0.6 | -7.7 |
| Kittitas | 1,220 | 1,200 | 1,140 | 1,090 | 1,130 | 3.7 | -7.4 |
| Klickitat | 280 | 280 | 260 | 250 | 250 | 0.0 | -10.7 |
| Lewis | 1,310 | 1,300 | 1,290 | 1,250 | 1,270 | 1.6 | -3.1 |
| Lincoln | 170 | 190 | 180 | 160 | 150 | -6.2 | -11.8 |
| Mason | 1,410 | 1,420 | 1,380 | 1,320 | 1,350 | 2.3 | -4.3 |
| Okanogan | 450 | 470 | 500 | 490 | 500 | 2.0 | 11.1 |
| Pacific | 520 | 560 | 550 | 530 | 580 | 9.4 | 11.5 |
| Pend | 300 | 320 | 310 | 290 | 300 | 3.4 | 0.0 |
| Pierce | 16,250 | 16,140 | 15,660 | 15,120 | 15,230 | 0.7 | -6.3 |
| San Juan | 360 | 350 | 340 | 310 | 320 | 3.2 | -11.1 |
| Skagit | 2,290 | 2,260 | 2,160 | 2,090 | 2,140 | 2.4 | -6.6 |
| Skamania | 260 | 280 | 260 | 220 | 240 | 9.1 | -7.7 |
| Snohomish | 10,580 | 11,030 | 10,520 | 9,990 | 10,280 | 2.9 | -2.8 |
| Spokane | 9,210 | 9,420 | 9,290 | 8,850 | 8,600 | -2.8 | -6.6 |
| Stevens | 890 | 940 | 930 | 870 | 890 | 2.3 | 0.0 |
| Thurston | 5,520 | 5,600 | 5,400 | 5,230 | 5,290 | 1.1 | -4.2 |
| Wahkiakum | 80 | 90 | 80 | 70 | 90 | 28.6 | 12.5 |
| Walla Walla | 910 | 910 | 890 | 860 | 830 | -3.5 | -8.8 |
| Whatcom | 3,320 | 3,280 | 3,150 | 3,080 | 3,130 | 1.6 | -5.7 |
| Whitman | 460 | 460 | 460 | 450 | 420 | -6.7 | -8.7 |
| Yakima | 1,910 | 1,920 | 1,940 | 1,860 | 1,830 | -1.6 | -4.2 |
| **Statewide** | **114,110** | **114,250** | **111,200** | **106,740** | **107,250** | **0.5** | **-6.0** |

| 2 2019 | % Change by year |
|---|---|
| 153,300 | -3.2 |
| 206,000 | -0.7 |
| 304,400 | 10.2 |
| 361,000 | 8.1 |
| 306,800 | 4.5 |
| 369,300 | 2.2 |
| 173,200 | 18.0 |
| 274,100 | 15.4 |
| 351,000 | 7.0 |
| 162,000 | -7.4 |
| 304,400 | 10.2 |
| 206,000 | -0.7 |
| 226,500 | 9.0 |
| 211,400 | 12.0 |
| 387,700 | 2.6 |
| 383,300 | 1.2 |
| 701,200 | -3.9 |
| 380,800 | 7.1 |
| 345,600 | 2.1 |
| 316,700 | 26.7 |
| 258,300 | 15.2 |
| 140,000 | 0.0 |
| 271,800 | 13.1 |
| 223,600 | 10.9 |
| 216,700 | 18.7 |
| 195,000 | 7.3 |
| 372,300 | 5.3 |
| 353,600 | 17.1 |
| 385,600 | 12.6 |
| 328,600 | 25.2 |
| 500,600 | 1.2 |
| 277,400 | 9.6 |
| 195,000 | 7.3 |
| 340,700 | 6.7 |
| 270,800 | 25.6 |
| 267,800 | 8.0 |
| 407,700 | 5.9 |
| 287,900 | 7.2 |
| 244,700 | 5.8 |
| **10,600** | **10.0** |




http://realestate.washington.edu/

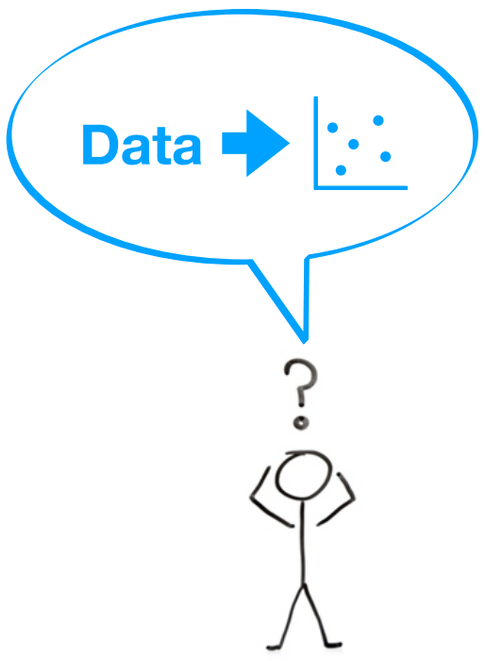
# As a non-programmer…


## Knowledge Gap

*“Most scholars do not receive formal training in data visualization and therefore rely on statistical packages with limited visualization capabilities, e.g., SPSS, STATA, or spreadsheets.”*

## Types and Format of Data

*“One of the challenges that results from the variety of data types is that academics often produce datasets in formats that can only be read by specific software.”*




## Software Diversity

*“Department, generation, university, data type, and personal preferences are just some of the factors that may influence academics’ choice of data analysis software.*

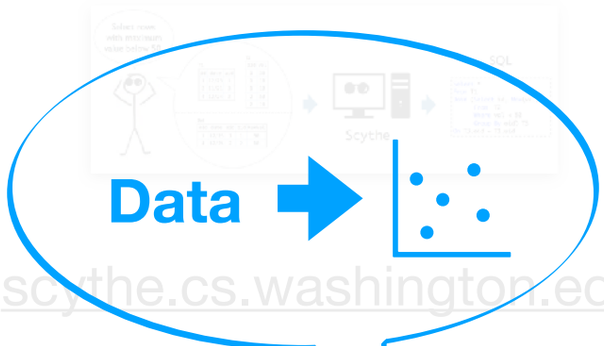
*The most common include SPSS, STATA, MatLab, R, and, NVivo, all of which have different graphics packages”*

Gatto, Malu AC. "Making research useful: Current challenges and good practices in data visualisation." (2015).

# Automated Programming

## Scythe

SQL Query Synthesis
from Input-Output Examples



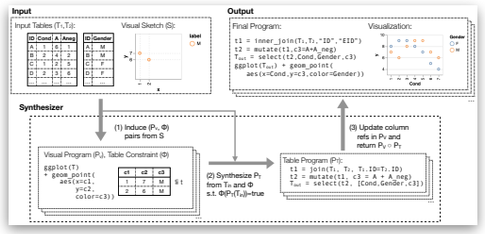

scythe.cs.washington.edu

| id | customer | total |
|---|---|---|
| 1 | Joe | 5 |
| 2 | Sally | 3 |
| 3 | Joe | 2 |
| 4 | Sally | 1 |

```
Select t1.*
From table As t1
Join (Select customer, Max(total)
      From table Group By customer) As t2
Where t1.customer = t2.customer
  And t1.total = t2.max-total
```

## Falx

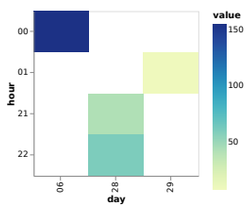
Synthesizing Visualizations
from Demonstration




CHI'21, POPL'20

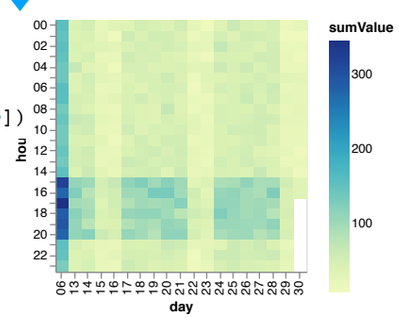
| DayHour | Site | Type | Value |
|---|---|---|---|
| 06_00 | e90e50 | PC | 22 |
| 06_00 | e90e50fx | PC | 8 |
| 06_01 | e90e50 | Mobile | 8 |
| … | … | … | … |

*1986x4*




```
t=group_by(T,[DayHour])
t=summarize(t,sum,Value)
t=separate(t,DayHour)
t=select(t,[Day,Hour,sumValue])

   Scatter[rect](
       x=Day,y=Hour,
       color=sumValue)
```




## Draco

Visualization Recommendation
from Partial Specification



uwdata.github.io/draco

```
{ "data": "cars.csv",
  "encoding": [
   { "channel": "x",
     "bin": true,
     "field": "horsepower"}]}
```

(*partial spec*)




*https://falx.cs.washington.edu/*

# Ideally

Data + Vis Program → Visualization

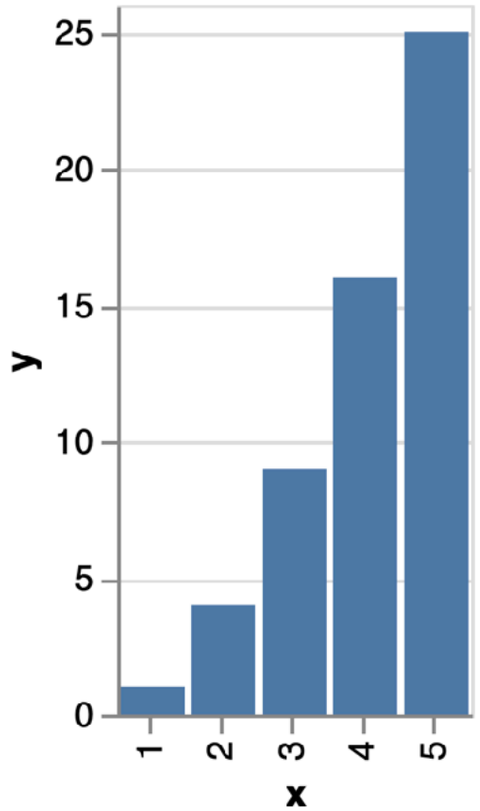
Bar( )

+

| x | y |
|---|---|
| 1 | 1 |
| 2 | 4 |
| 3 | 9 |
| 4 | 16 |
| 5 | 25 |

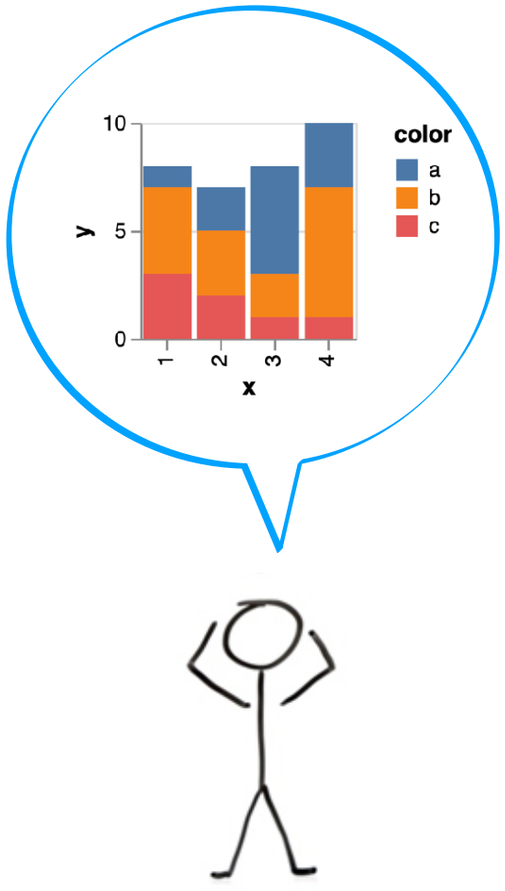
Bar( )

| x | y |
|---|---|
| 1 | 1 |
| 2 | 4 |
| 3 | 9 |
| 4 | 16 |
| 5 | 25 |



# In practice

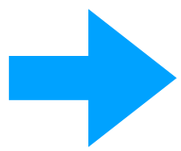
Data + Vis Program → ?



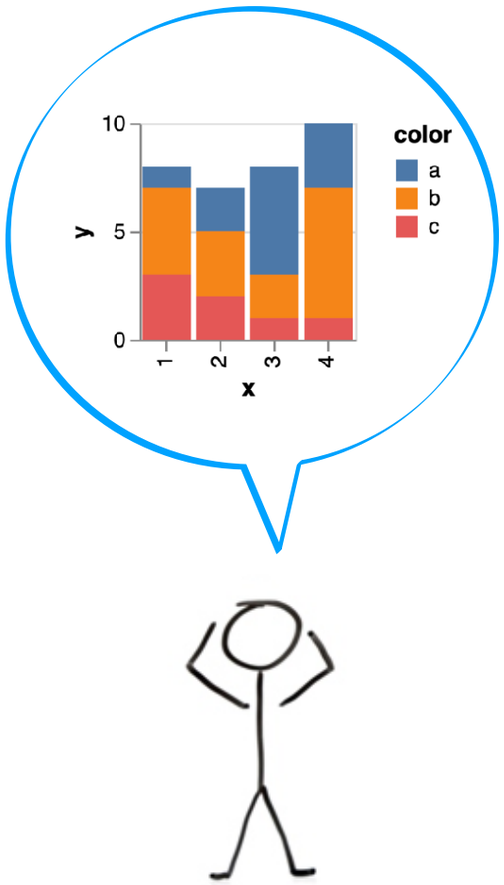

StackedBar( )

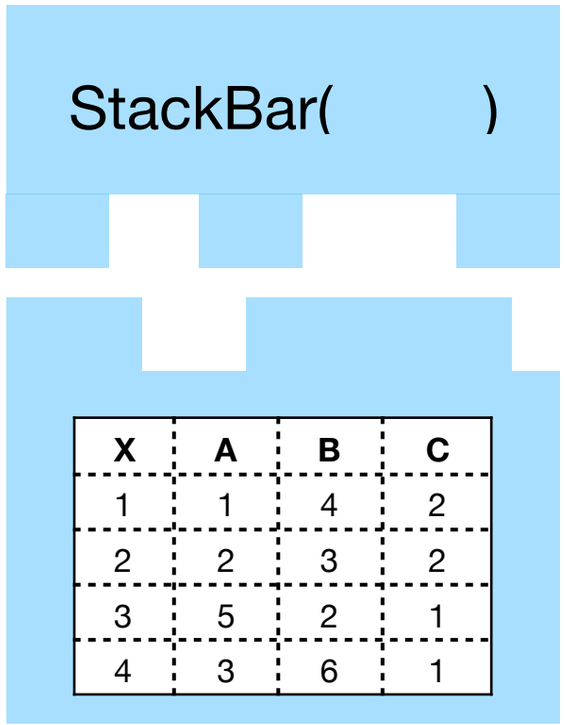
+

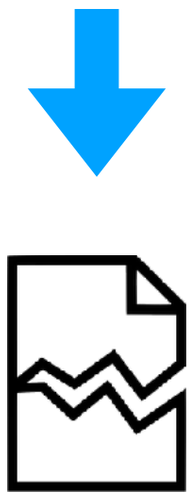
| X | A | B | C |
|---|---|---|---|
| 1 | 1 | 4 | 2 |
| 2 | 2 | 3 | 2 |
| 3 | 5 | 2 | 1 |
| 4 | 3 | 6 | 1 |

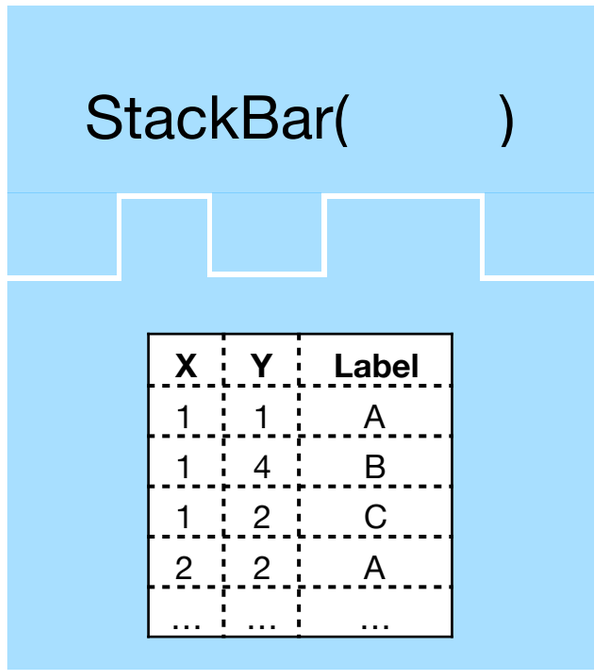


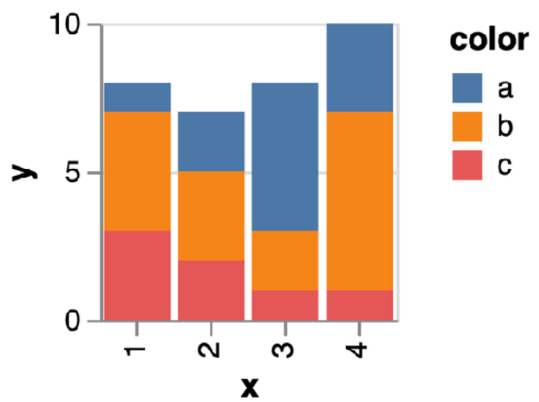
**Problem:** the input data comes with different ***shapes***, but the visualization library expects only one of them



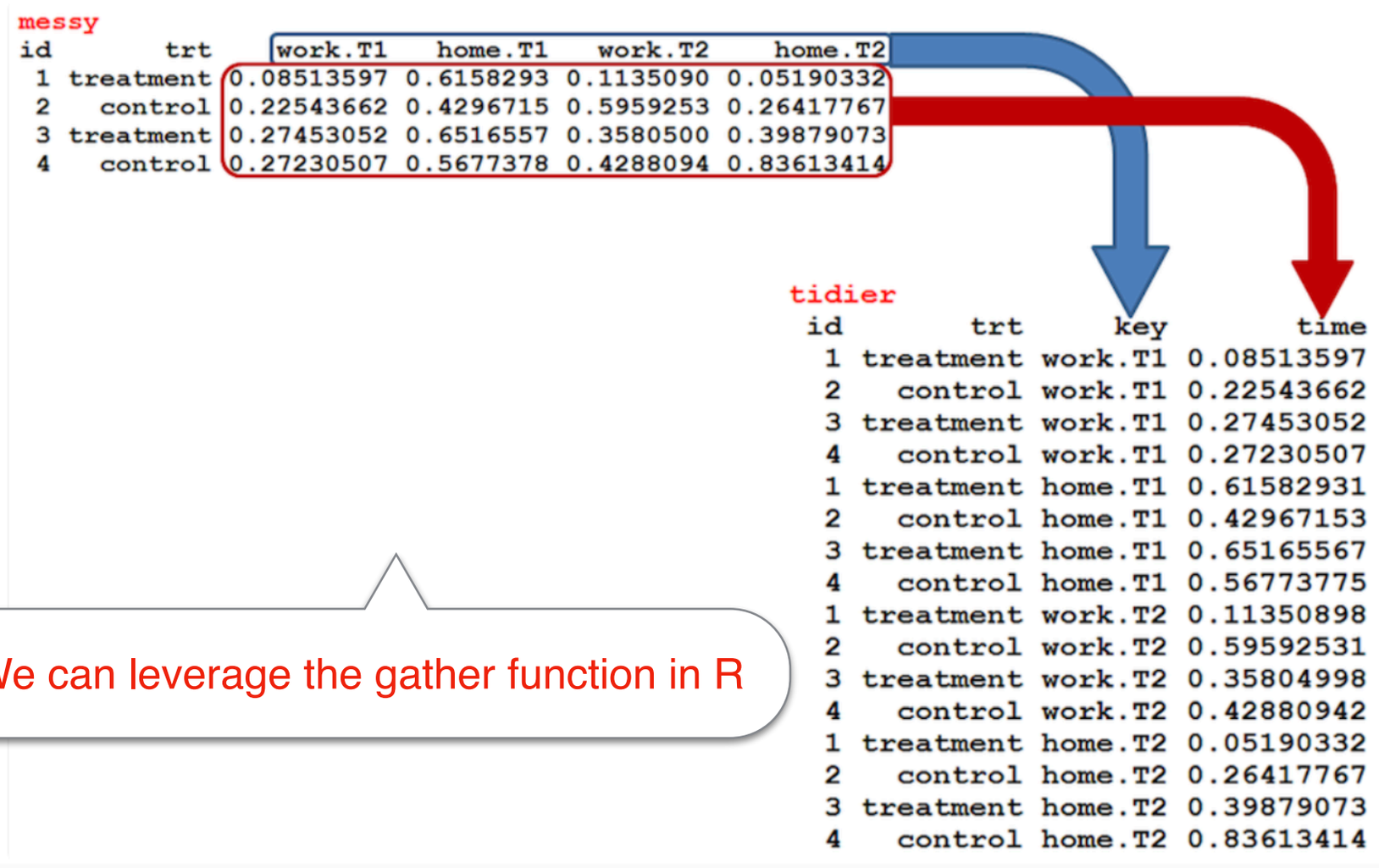


## gather( ) function:

**Objective:** Reshaping wide format to long format

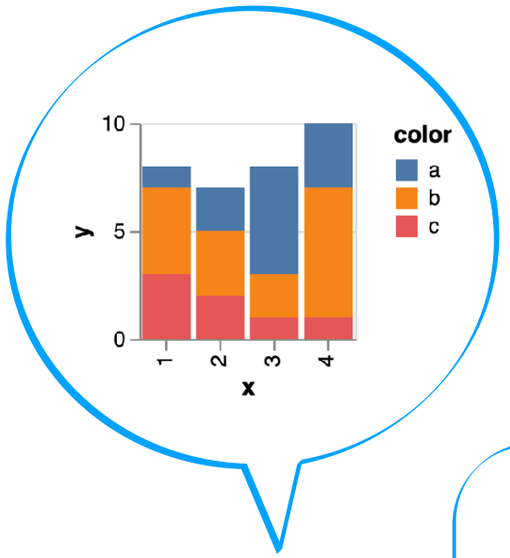
**Description:** There are times when our data is considered unstacked and a common attribute of concern is spread out across columns. To reformat the data such that these common attributes are *gathered* together as a single variable, the `gather()` function will take multiple columns and collapse them into key-value pairs, duplicating all other columns as needed.




https://uc-r.github.io/tidyr

# In practice

Data + Data Program + Vis Program → Visualization



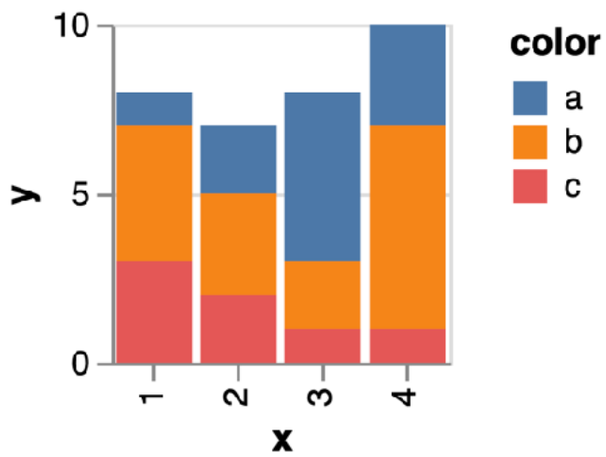

| X | A | B | C |
|---|---|---|---|
| 1 | 1 | 4 | 2 |
| 2 | 2 | 3 | 2 |
| 3 | 5 | 2 | 1 |
| 4 | 3 | 6 | 1 |

| X | Y | Label |
|---|---|---|
| 1 | 1 | A |
| 1 | 4 | B |
| 1 | 2 | C |
| 2 | 2 | A |
| … | … | … |

+ StackedBar( )

*Transform the raw data T into T':*

T' ← gather(T, key=X, value=[A, B, C])



# In practice

Data + DataProgram + VisProgram




| X | A | B | C |
|---|---|---|---|
| 1 | 1 | 4 | 2 |
| 2 | 2 | 3 | 2 |
| 3 | 5 | 2 | 1 |
| 4 | 3 | 6 | 1 |

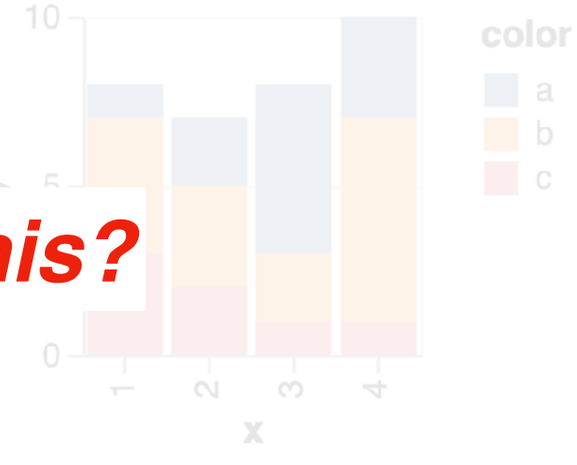
| X | Y | Label |
|---|---|---|
| 1 | 1 | A |
| 1 | 4 | B |
| 1 | 2 | C |
| 2 | 2 | A |
| ... | ... | ... |

1. Requires knowledge about the library and their expected data shapes

\+ StackedBar( )

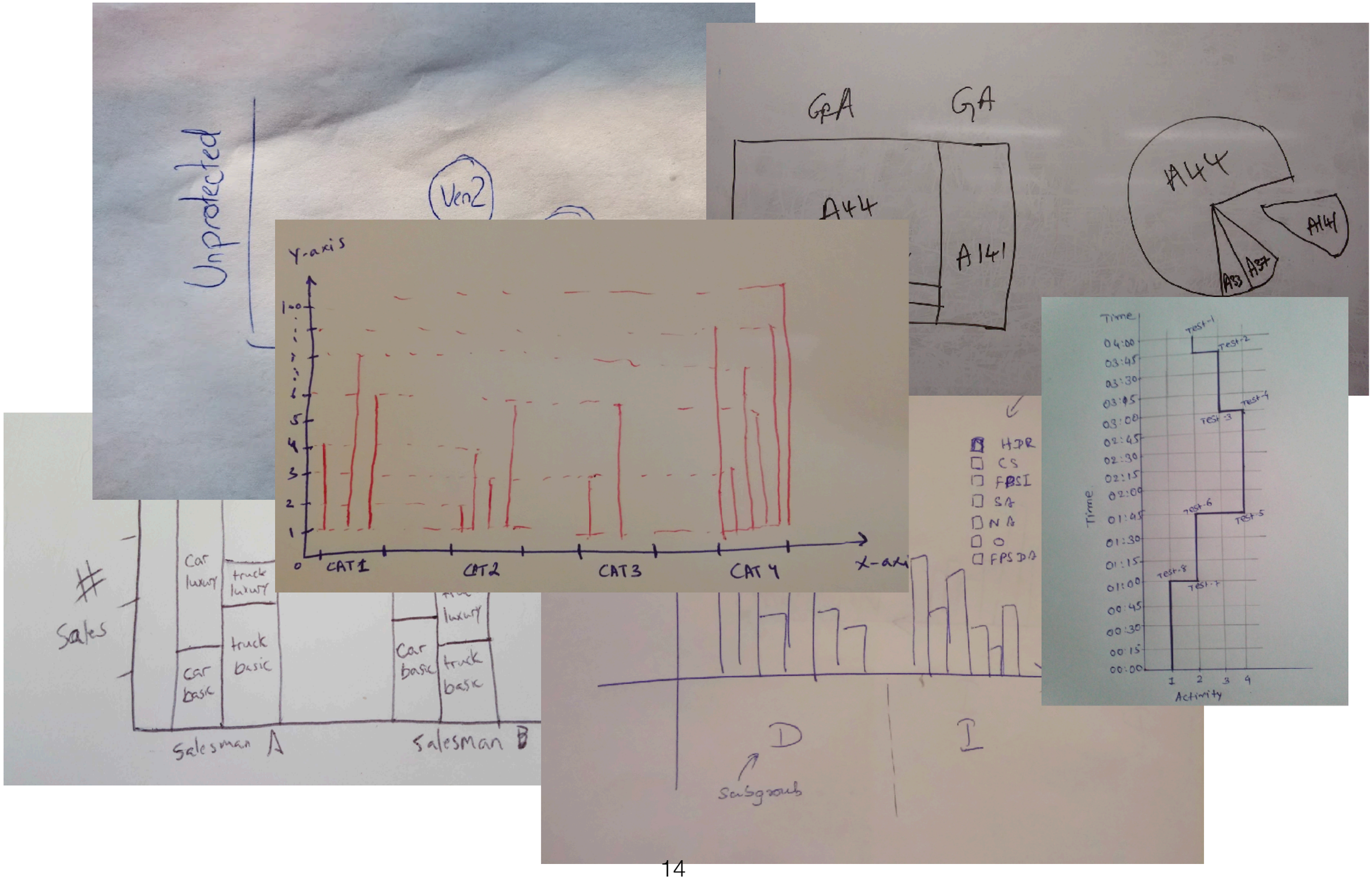
2. Requires knowledge of data preparation library to do transformation






***Can we automate this?***

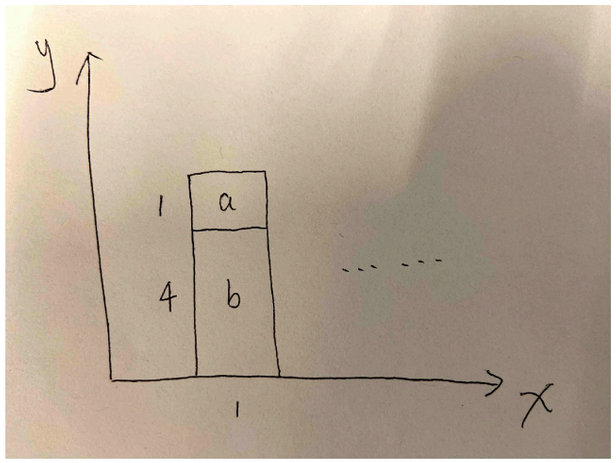
# Insight: Explain task using *partial visualization*



# Visualization by Example

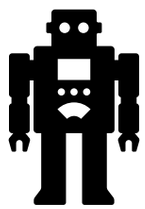
*User Input*

| X | A | B | C |
|---|---|---|---|
| 1 | 1 | 4 | 2 |
| 2 | 2 | 3 | 2 |
| 3 | 5 | 2 | 1 |
| 4 | 3 | 6 | 1 |




How to model examples?

Synthesize



*DataProgram*

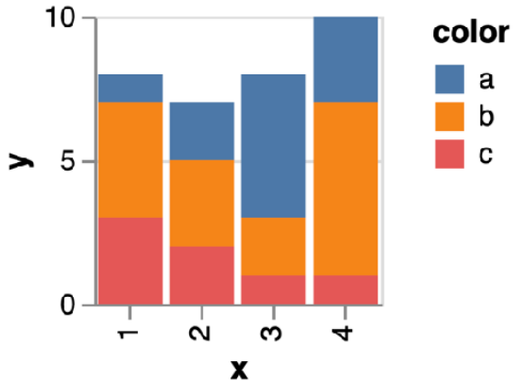
*T' = gather(T, key=c1, values=[A,B,C])*

+

*VisProgram*

*StackedBar(*
*x → c1,*
*y → c2,*
*color → c3)*

Render

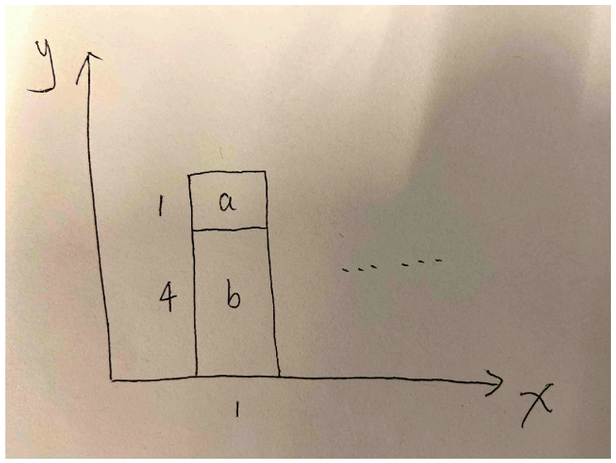



How to solve the compositional synthesis problem efficiently?

# 1. Model Visualization using Visual Trace

Visualization = a set of visual elements that appear on the canvas
(bars, lines, points etc)

| X | A | B | C |
|---|---|---|---|
| 1 | 1 | 4 | 2 |
| 2 | 2 | 3 | 2 |
| 3 | 5 | 2 | 1 |
| 4 | 3 | 6 | 1 |




bar(x=1, y=1, label=a)
bar(x=1, y=4, label=b)

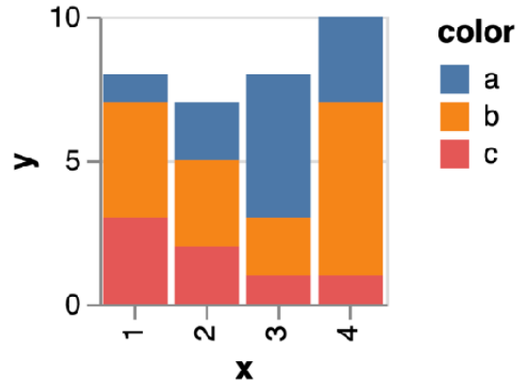
*DataProgram*

*T' = gather(T, key=c1, values=[A,B,C])*

+

*VisProgram*

*StackedBar(*
*x → c1,*
*y → c2,*
*color → c3)*



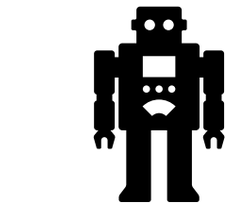

⊆

bar(x=1, y=1, label=a), bar(x=1, y=4, label=b)
bar(x=1, y=2, label=c), bar(x=2, y=2, label=a)
bar(x=2, y=3, label=b), bar(x=2, y=2, label=c)
bar(x=3, y=5, label=a), bar(x=3, y=2, label=b)
bar(x=3, y=1, label=c), bar(x=4, y=3, label=a)
bar(x=4, y=6, label=b), bar(x=4, y=1, label=c)

# 2. Synthesis Algorithm

| X | A | B | C |
|---|---|---|---|
| 1 | 1 | 4 | 2 |
| 2 | 2 | 3 | 2 |
| 3 | 5 | 2 | 1 |
| 4 | 3 | 6 | 1 |

bar(x=1, y=1, label=a)
bar(x=1, y=4, label=b)



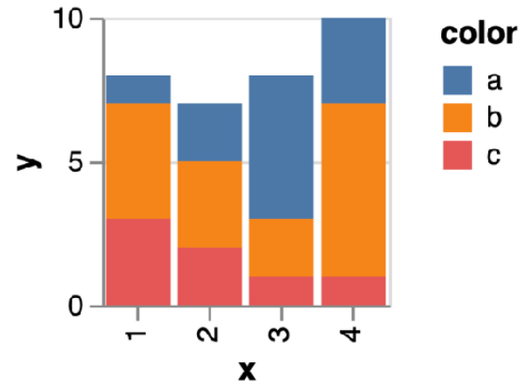
*DataProgram*

*T' = gather(T, key=c1, values=[A,B,C])*

+

*VisProgram*

*StackedBar(*
*x → c1,*
*y → c2,*
*color → c3)*




*Visual Sketch*

bar(x=1, y=1, label=a)
bar(x=1, y=4, label=b)

Decompile visual sketch

*VisProgram*

*StackedBar(*
*x → c1,*
*y → c2,*
*color → c3)*

+

*Table Sketch*

| c1 | c2 | c3 |
|---|---|---|
| 1 | 1 | A |
| 1 | 4 | B |

*Input table*

| X | A | B | C |
|---|---|---|---|
| 1 | 1 | 4 | 2 |
| 2 | 2 | 3 | 2 |
| 3 | 5 | 2 | 1 |
| 4 | 3 | 6 | 1 |

+

*Table Sketch*

| c1 | c2 | c3 |
|---|---|---|
| 1 | 1 | A |
| 1 | 4 | B |

Synthesize data program

$T_{sketch} \subset T'$

*DataProgram*

*T' = gather(T, key=c1, values=[A,B,C])*

# 2. Synthesis Algorithm

## (1) Decompile Visualization

*Visual Sketch*

bar(x=1, y=1, label=a)
bar(x=1, y=4, label=b)

*VisProgram*

*StackedBar(*
*x → c1,*
*y → c2,*
*color → c3)*

\+

*Table Sketch*

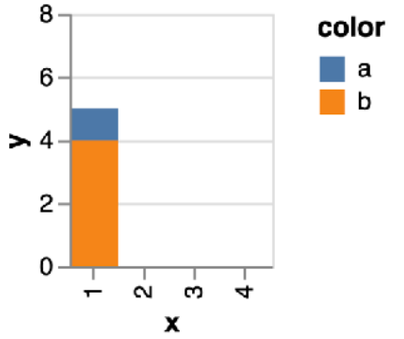
| c1 | c2 | c3 |
|---|---|---|
| 1 | 1 | A |
| 1 | 4 | B |

bar(x=1, y=1, label=a)
bar(x=1, y=4, label=b)

*StackedBar(*
*x → c1,*
*y → c2,*
*color → c3)*

\+

| c1 | c2 | c3 |
|---|---|---|
| 1 | 1 | A |
| 1 | 4 | B |

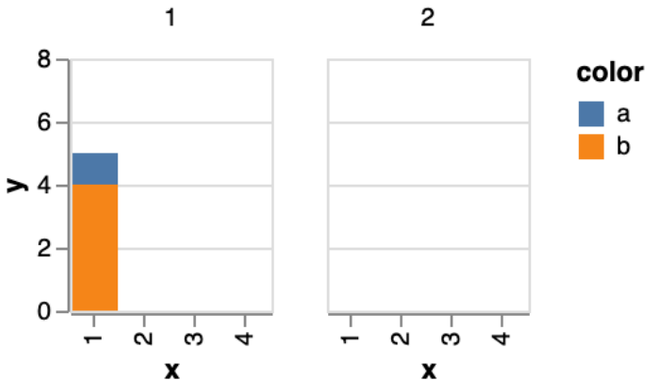



bar(x=1, y=1, label=a)
bar(x=1, y=4, label=b)

*StackedBar(*
*x → c1,*
*y → c2,*
*color → c3,*
*column → ?)*

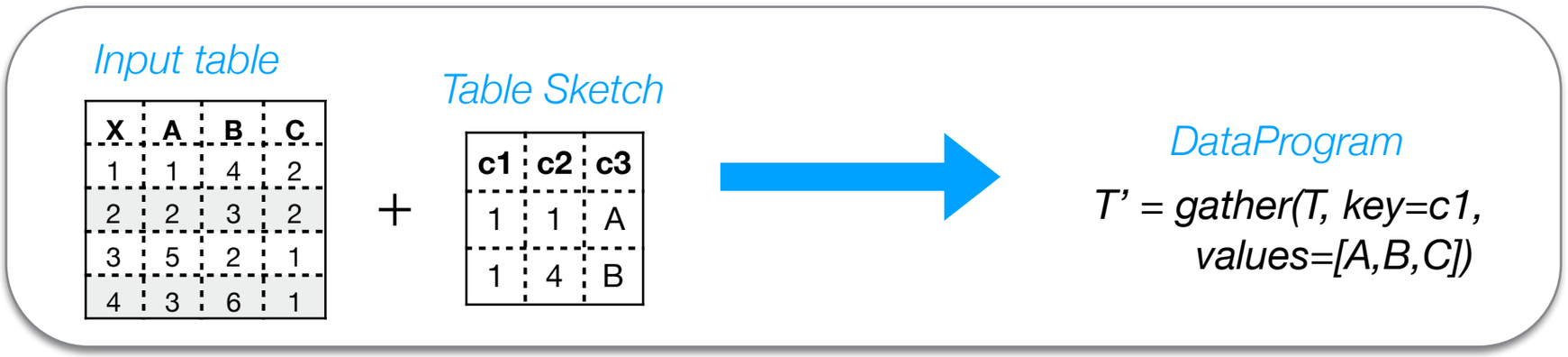
\+

| c1 | c2 | c3 | c4 |
|---|---|---|---|
| 1 | 1 | A | ? |
| 1 | 4 | B | ? |



# 2. Synthesis Algorithm

(2) Synthesize data program $P_T$




*Requires* $T_{sketch} \subset P_T(T_{in})$

Similar to SQL synthesis task:
Given $T_{in}$ and $T_{out}$, synthesize $Q$, s.t.
$Q(T_{in}) = T_{out}$

**Two-phase synthesis algorithm**

1. Enumerate program sketch
2. Complete parameters

**Insight**: Use bidirectional deductive reasoning to prune program sketches

**Weaker specification**

1. "=" → "⊆"
2. No longer assume small input table
3. Larger program space

***Requires stronger pruning algorithm***

# 2. Synthesis Algorithm

(2) Synthesize data program $P_T$

*Input table*

| X | A | B | C |
|---|---|---|---|
| 1 | 1 | 4 | 2 |
| 2 | 2 | 3 | 2 |
| 3 | 5 | 2 | 1 |
| 4 | 3 | 6 | 1 |

+

*Table Sketch*

| c1 | c2 | c3 |
|---|---|---|
| 1 | 1 | A |
| 1 | 4 | B |

→

*DataProgram*

*T' = gather(T, key=c1, values=[A,B,C])*

*Requires $T_{sketch} \subset P_T(T_{in})$*

---

*Can $P_T$ be instantiated from the following sketch?*

*T' = gather(*
*T,*
*key=**C**,*
*values = □)*

Forward reasoning:
What's the property of the output ***T'*** given input $T_{in}$ and $P_T$?

Since ***key=C***, column **C** in $T_{in}$ should be the first column in ***T'***

*i.e., the first column of $T_{sketch}$ should be contained by column **C***

*Cannot prune*

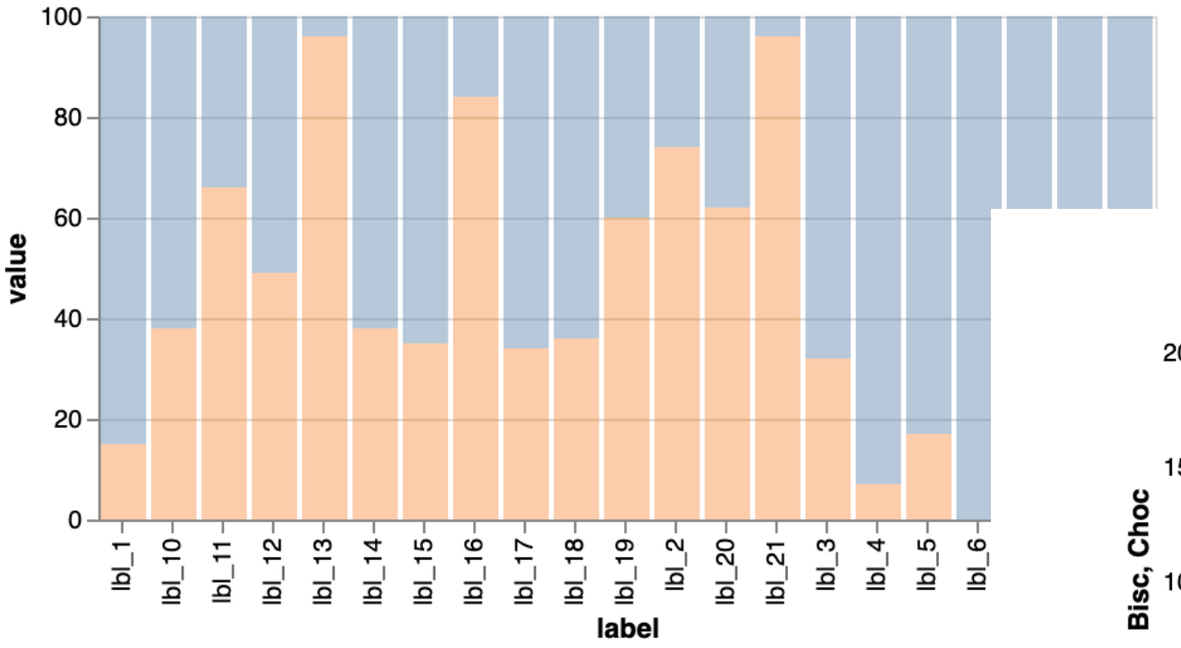
Backward reasoning:
What's the property of the input ***T*** given output $T_{sketch}$ and $P_T$?

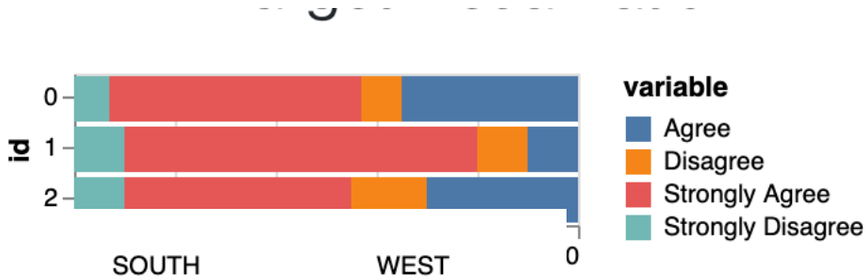
Since ***key=C***, we have

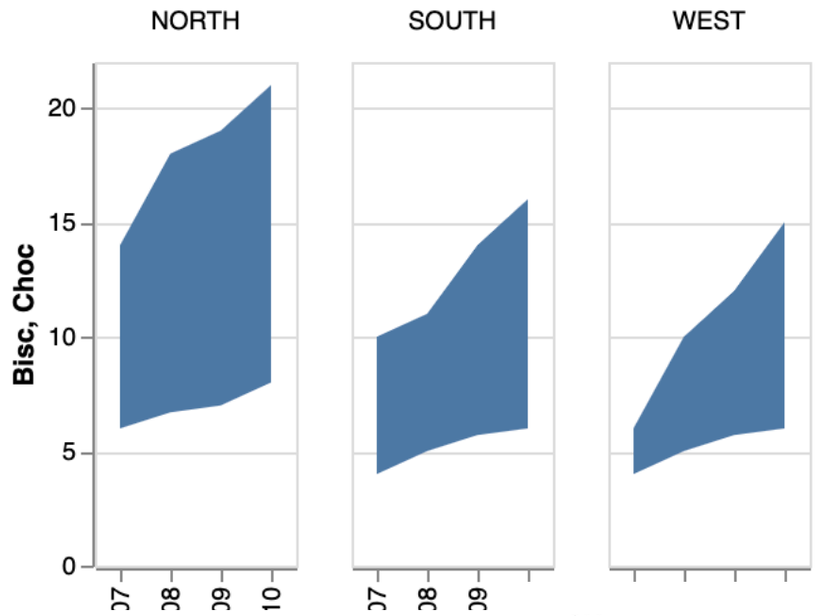
| C | A | B |
|---|---|---|
| 1 | 1 | 4 |

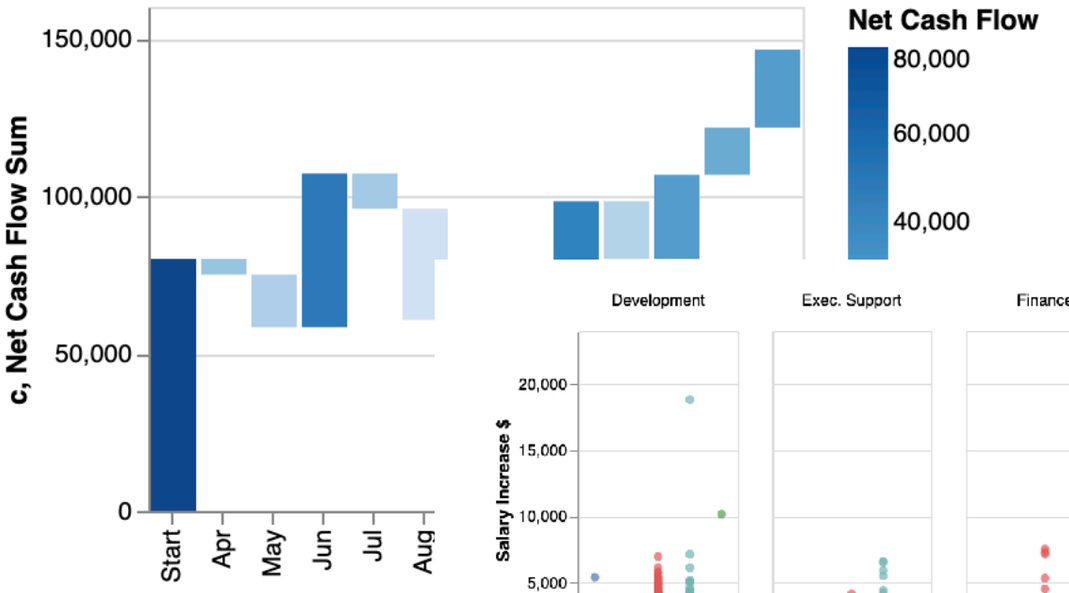
$\subseteq T_{in}$

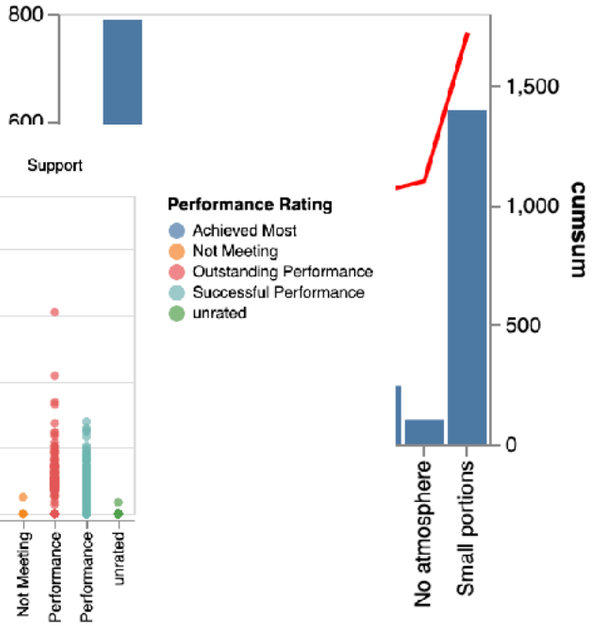
*Prune!*

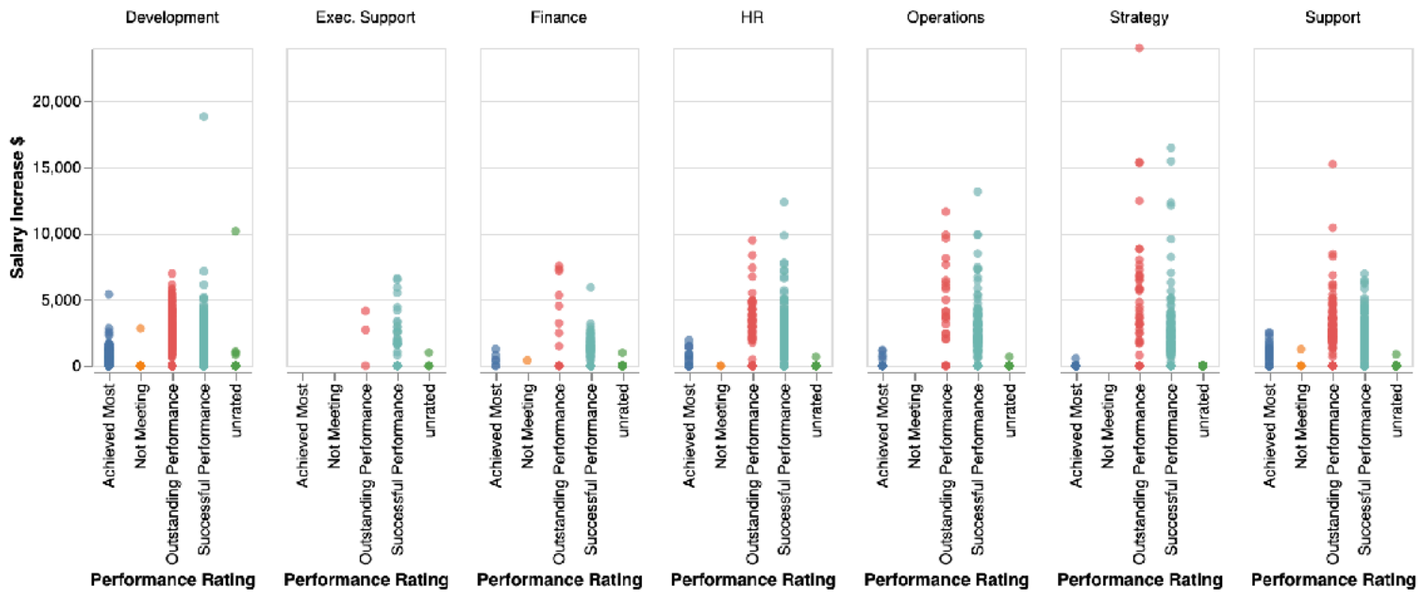

value
label
lbl_1
lbl_10
lbl_11
lbl_12
lbl_13
lbl_14
lbl_15
lbl_16
lbl_17
lbl_18
lbl_19
lbl_2
lbl_20
lbl_21
lbl_3
lbl_4
lbl_5
lbl_6
variable
mutate_a
values


id
variable
Agree
Disagree
Strongly Agree
Strongly Disagree


NORTH
SOUTH
WEST
Bisc, Choc


Net Cash Flow
c, Net Cash Flow Sum
Start
Apr
May
Jun
Jul
Aug


Volume
cumsum
No atmosphere
Small portions


Development
Exec. Support
Finance
HR
Operations
Strategy
Support
Salary Increase $
Performance Rating
Achieved Most
Not Meeting
Outstanding Performance
Successful Performance
unrated

# TODOs by next lecture

- Start to work on your final report/project! (30%)