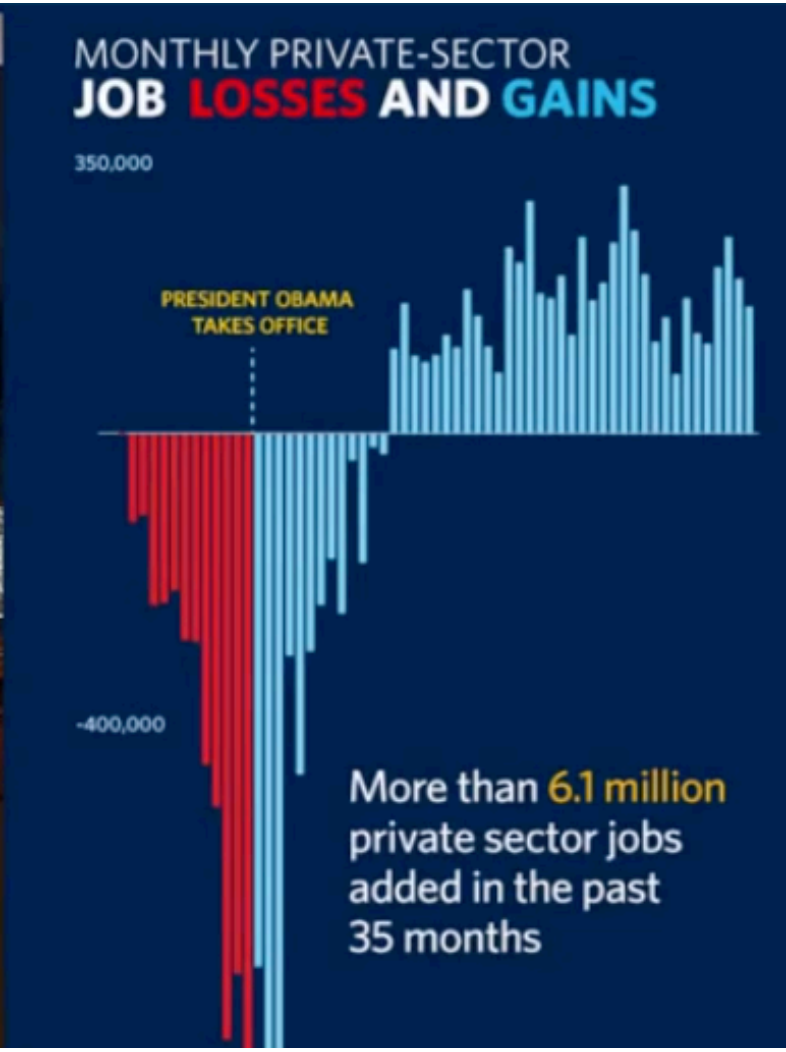
Scorpion

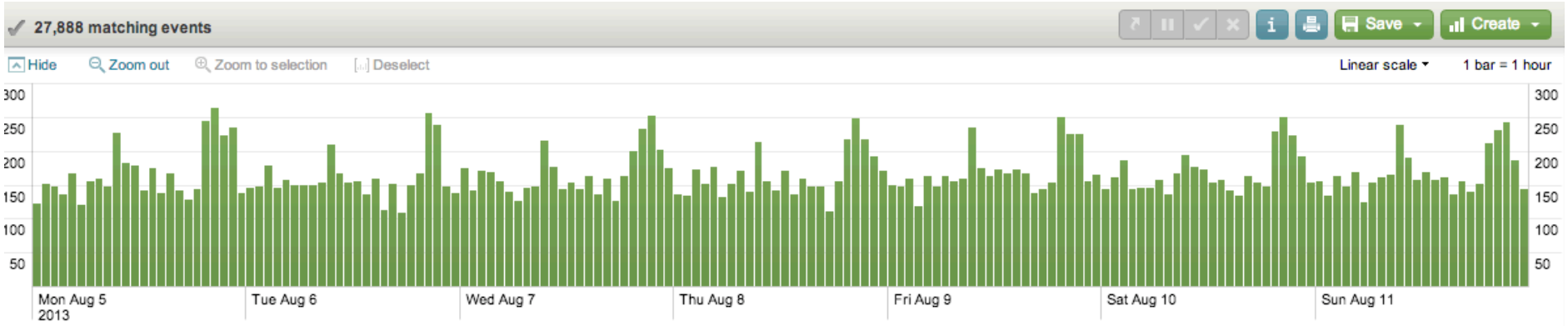
Explaining Away Outliers in Aggregate Queries

eugene wu and sam madden
MIT

http://springfieldpunx.blogspot.com/2010/11/mortal-kombat-ninjas-scorpion.html


Compare:
Enter ticker here
Add
Dow Jones
S&P 500
ORCL
CSCO
INTC
SAP
MSFT
HPQ
ACN
Zoom: 1d 5d 1m 3m 6m YTD 1y 5y 10y All
Sep 25, 2009 - Aug 22, 2013
MSFT +29.91%
ORCL +41.42%
IBM +51.14%
60%
40%
20%
0%
-20%
2010
2011
2012
2013
1980
1985
1990
1995
2000
2005
2010


MONTHLY PRIVATE-SECTOR
JOB LOSSES AND GAINS
350,000
PRESIDENT OBAMA
TAKES OFFICE
-400,000
More than 6.1 million
private sector jobs
added in the past
35 months



Hide

27,888 events over all time

Export | Options | « prev 1 2 3 4 5 6 7 8 9 10 next » | 10 per page

3 selected fields Edit

- host (1)
- source (1)
- sourcetype (1)

27 interesting fields

- action (2)
- # bytes (65)
- category_id (5)
- clientip (≥100)
- file (14)
- ident (1)

1 8/11/13 11:59:34.000 PM
178.19.3.39 - - [11/Aug/2013:23:59:34] "GET /flower_store/category.screen?category_id=CANDY HTTP/1.1" 200 10567 "http://mystore.splunk.com/flower_store/cart.do?action=purchase&itemId=EST-14&JSESSIONID=SD5SL10FF8ADFF3" "Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.8.0.10) Gecko/20070223 CentOS/1.5.0.10-0.1.el4.centos Firefox/1.5.0.10" 3187 1245
host=li-desktop | sourcetype=access_combined_wcookie | source=/Users/sirrice/Downloads/Sampledata/apache3.splunk.com/access_combined.log

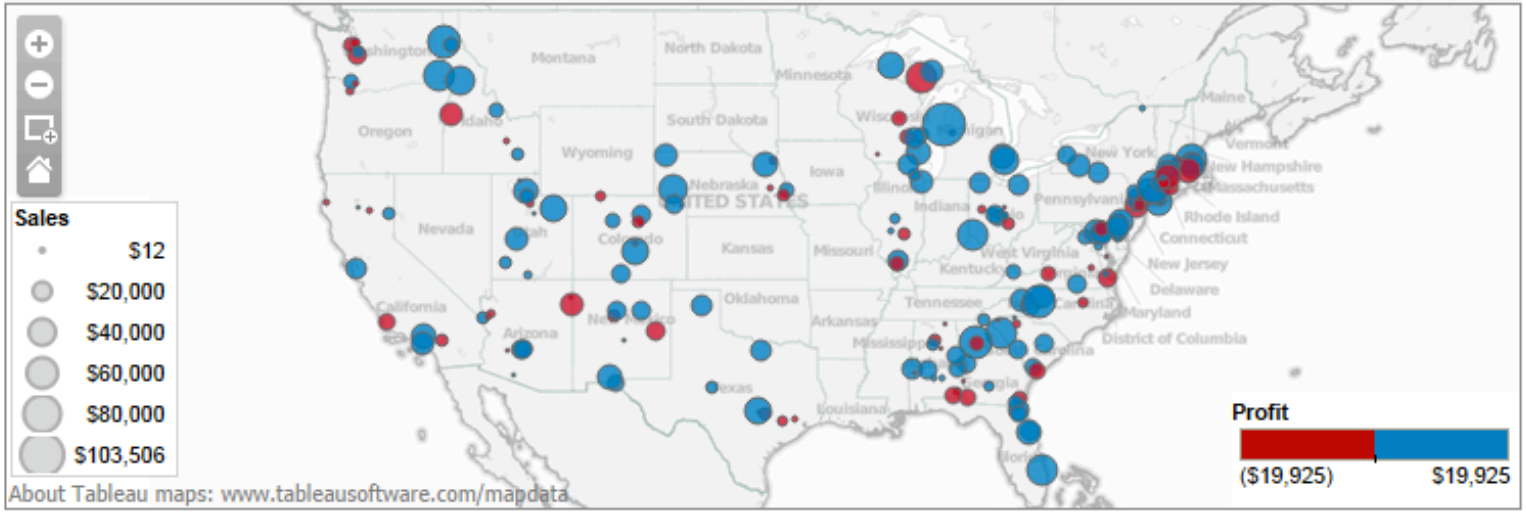
2 8/11/13 11:59:15.000 PM
10.192.1.46 - - [11/Aug/2013:23:59:15] "GET /flower_store/category.screen?category_id=PLANTS HTTP/1.1" 200 10567 "http://mystore.splunk.com/flower_store/cart.do?action=purchase&itemId=EST-27&JSESSIONID=SD5SL10FF8ADFF3" "Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.8.0.10) Gecko/20070223 CentOS/1.5.0.10-0.1.el4.centos Firefox/1.5.0.10" 3352 1920
host=li-desktop | sourcetype=access_combined_wcookie | source=/Users/sirrice/Downloads/Sampledata/apache3.splunk.com/access_combined.log

3 8/11/13 11:59:15.000 PM
10.192.1.46 - - [11/Aug/2013:23:59:15] "GET /flower_store/category.screen?category_id=PLANTS HTTP/1.1" 200 10567 "http://mystore.splunk.com/flower_store/cart.do?action=purchase&itemId=EST-27&JSESSIONID=SD5SL10FF8ADFF3"

Executive Dashboard | Product Dashboard | Customer Dashboard

Executive Dashboard

Sales by Customer Location



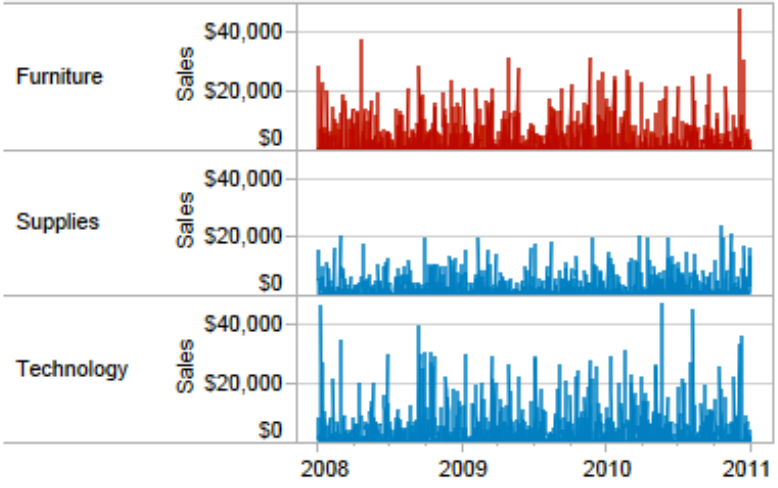
Select Year: (All)

Customer Region: (All)

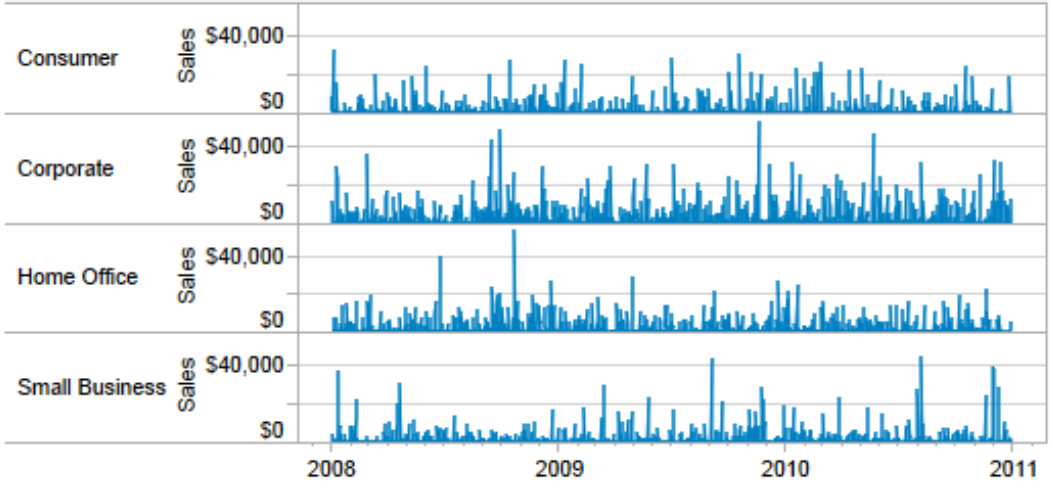
Product Category: (All)

Customer Segment: (All)

Sales by Product Category



Sales by Customer Segment



Find reports & more

MY STUFF

- Dashboards
- Shortcuts
- Intelligence Events

STANDARD REPORTS

- Real-Time
- Audience
  - Overview
  - Demographics
  - Behavior
  - Technology
  - Mobile
  - Custom
  - Visitors Flow
- Traffic Sources
- Content
- Conversions

Audience Overview

Jul 17, 2013 - Aug 16, 2013

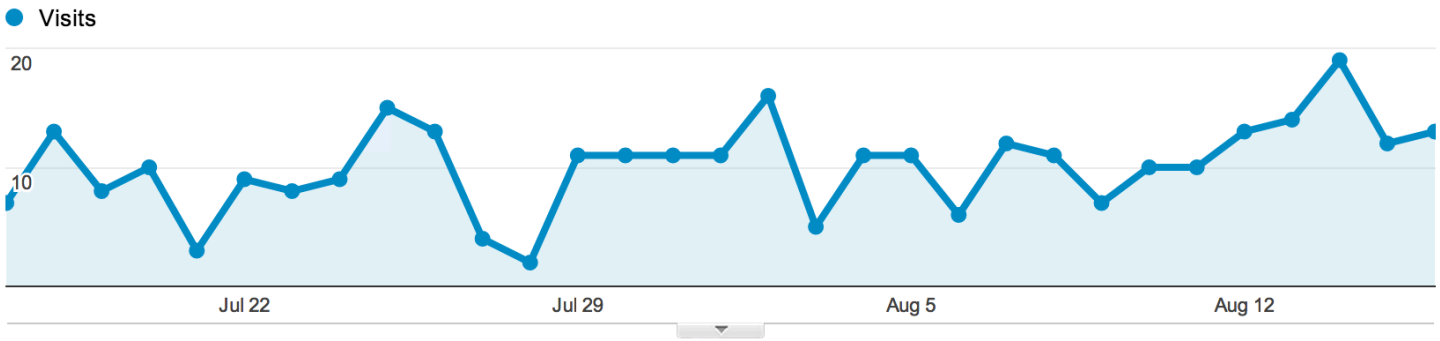
Advanced Segments | Email | Export | Add to Dashboard | Shortcut

% of visits: 100.00%

Overview

Visits vs. Select a metric

Hourly | Day | Week | Month

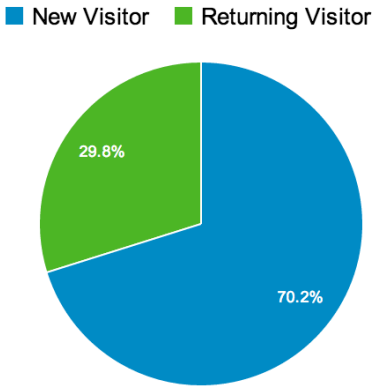


237 people visited this site

Visits
315

Unique Visitors
237

Pageviews
482




+ableau®
SOFTWARE


splunk>®
POWERED

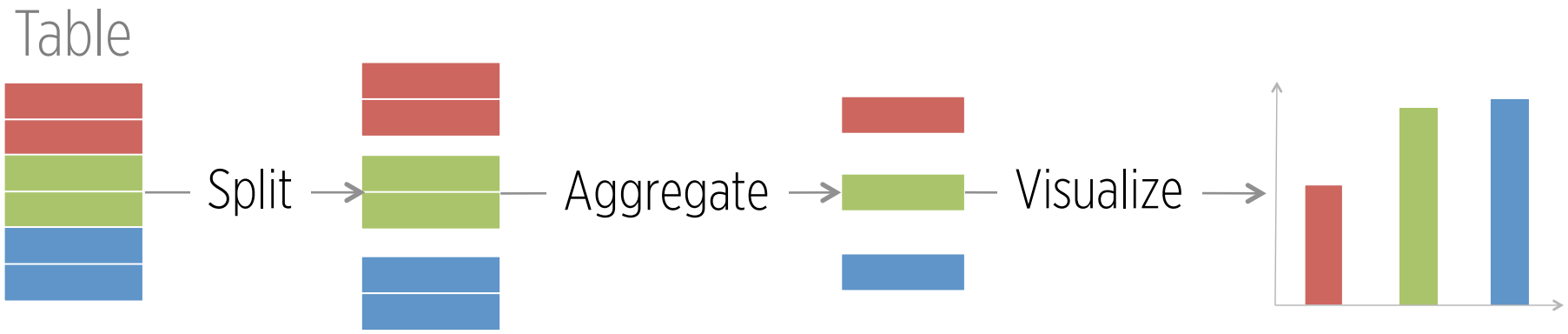
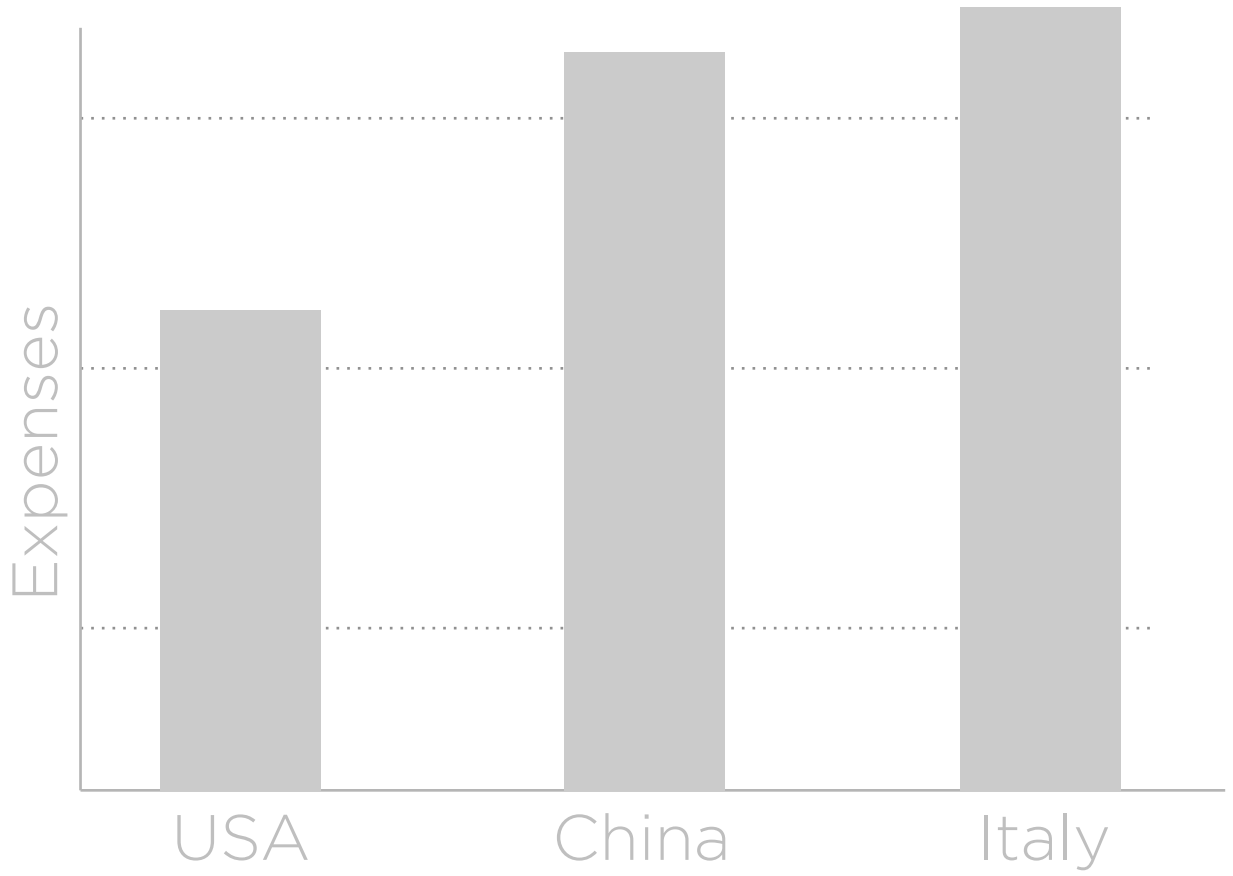
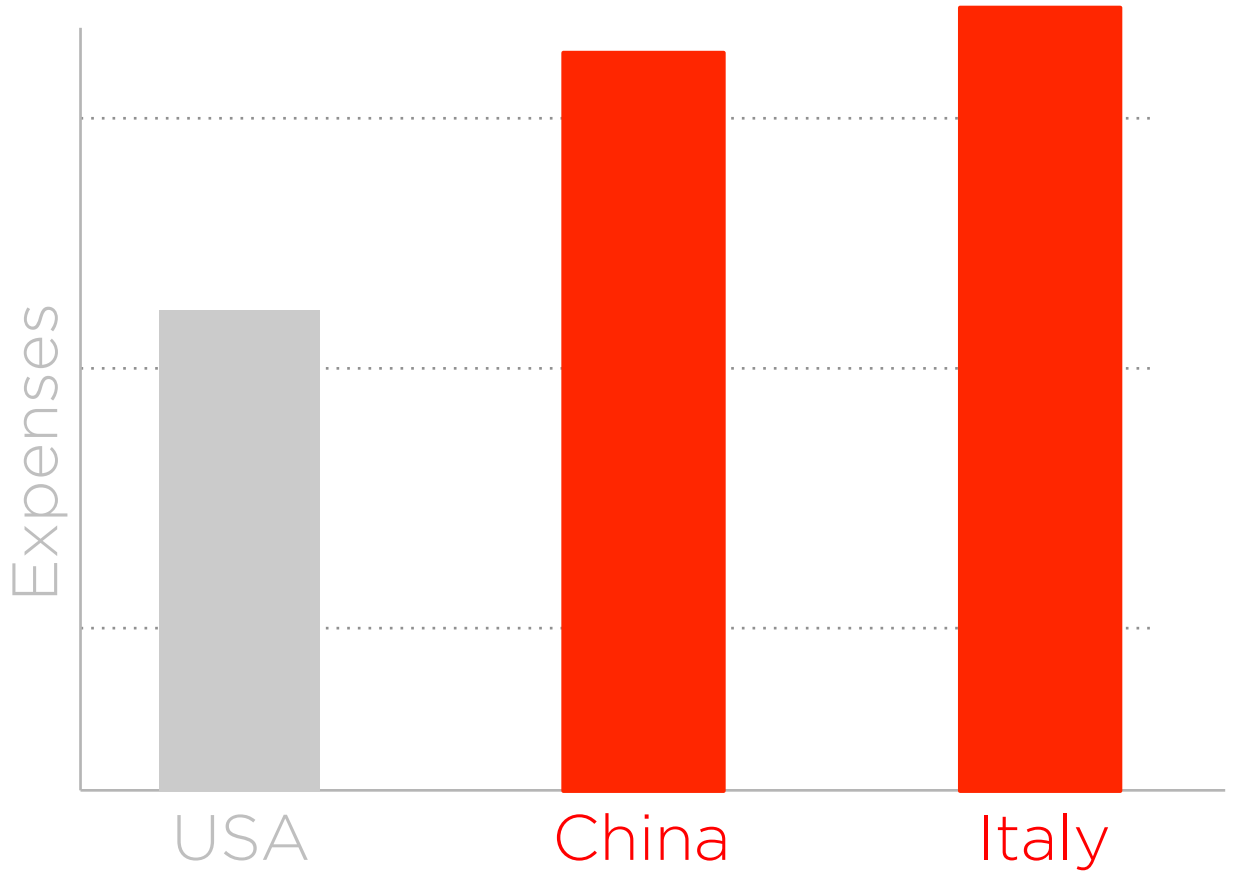


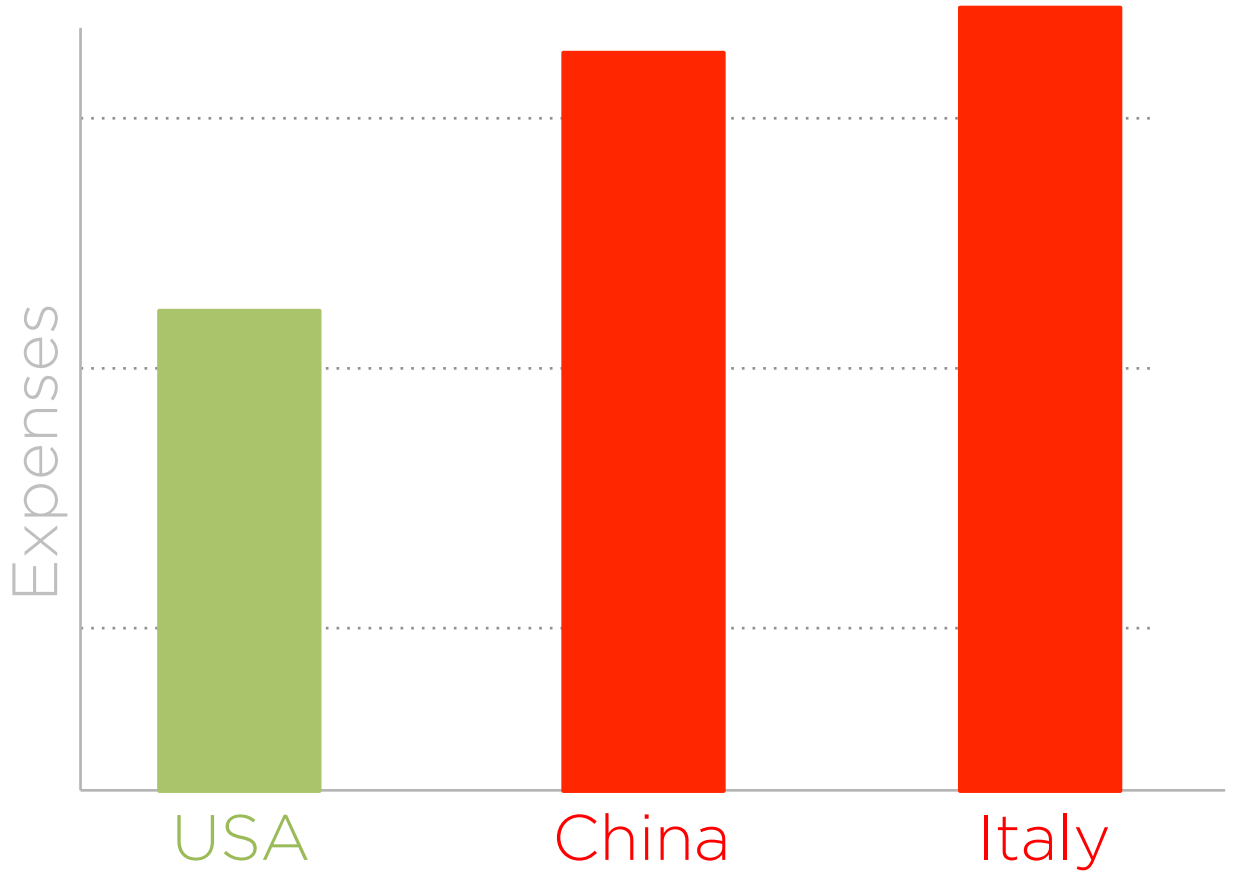
Table
Split
Aggregate
Visualize



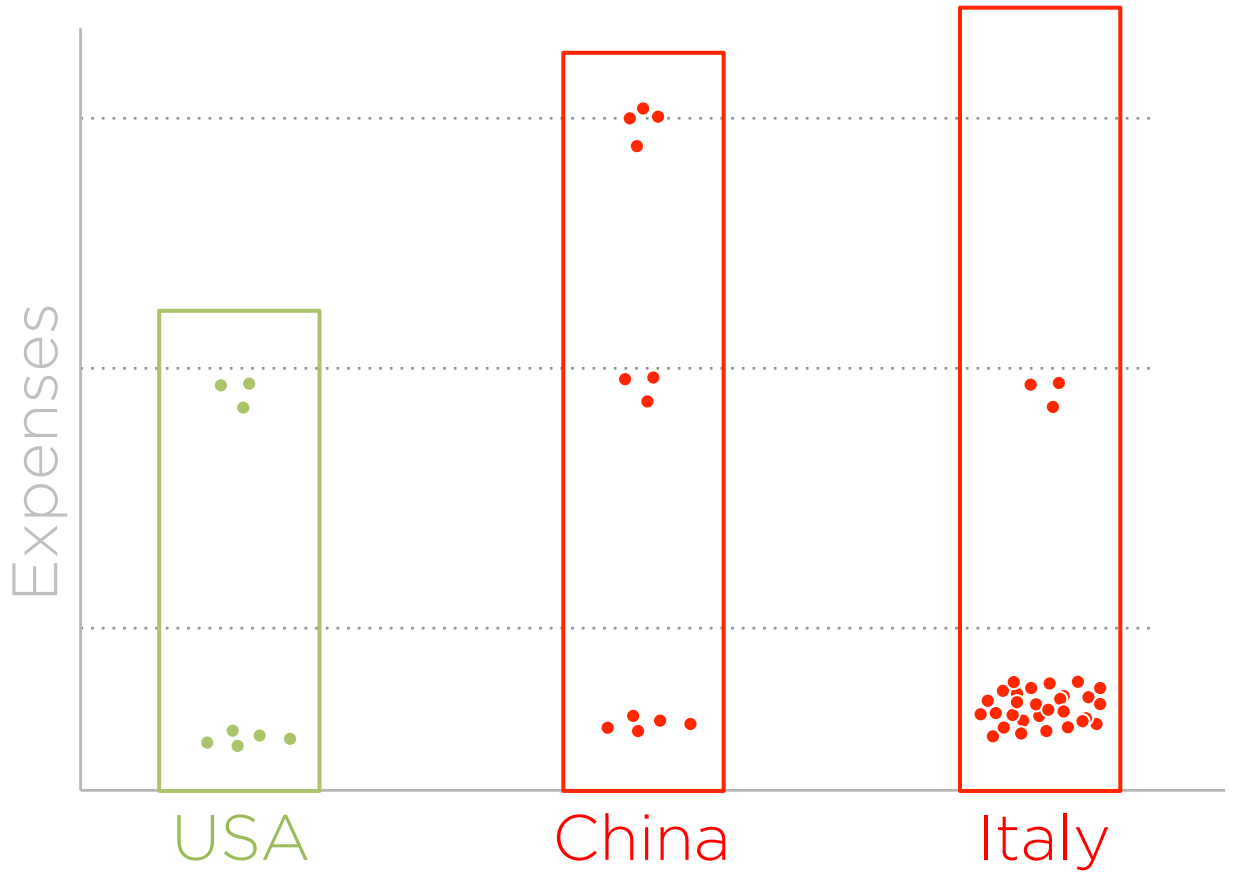

```
SELECT sum(cost)
FROM   expenses
GROUPBY country
```




```
SELECT sum(cost)
FROM   expenses
GROUPBY country
```




```
SELECT sum(cost)
FROM   expenses
GROUPBY country
```

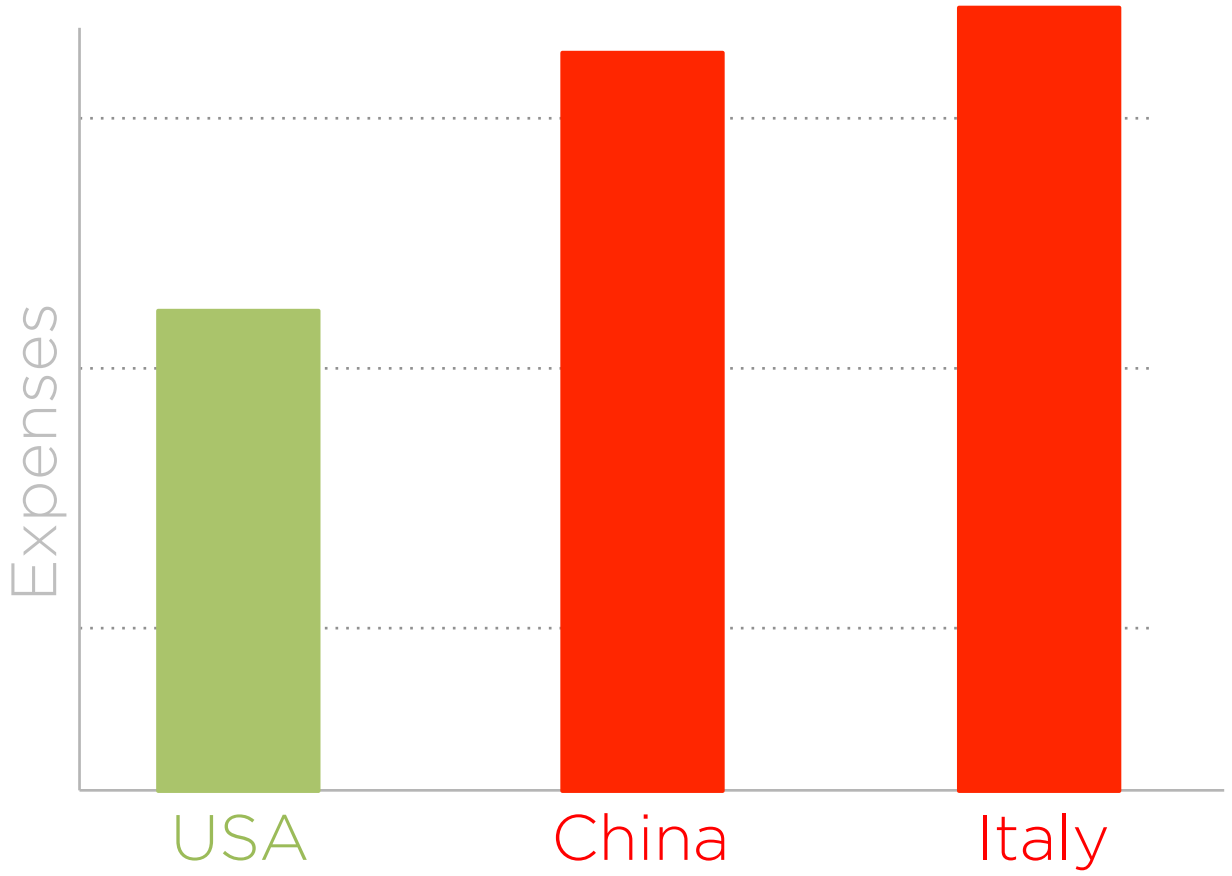



```
SELECT sum(cost)
FROM   expenses
GROUPBY country
```

Given

Outlier and normal results

Understand Why




```
SELECT sum(cost)
FROM   expenses
GROUPBY country
```

Given

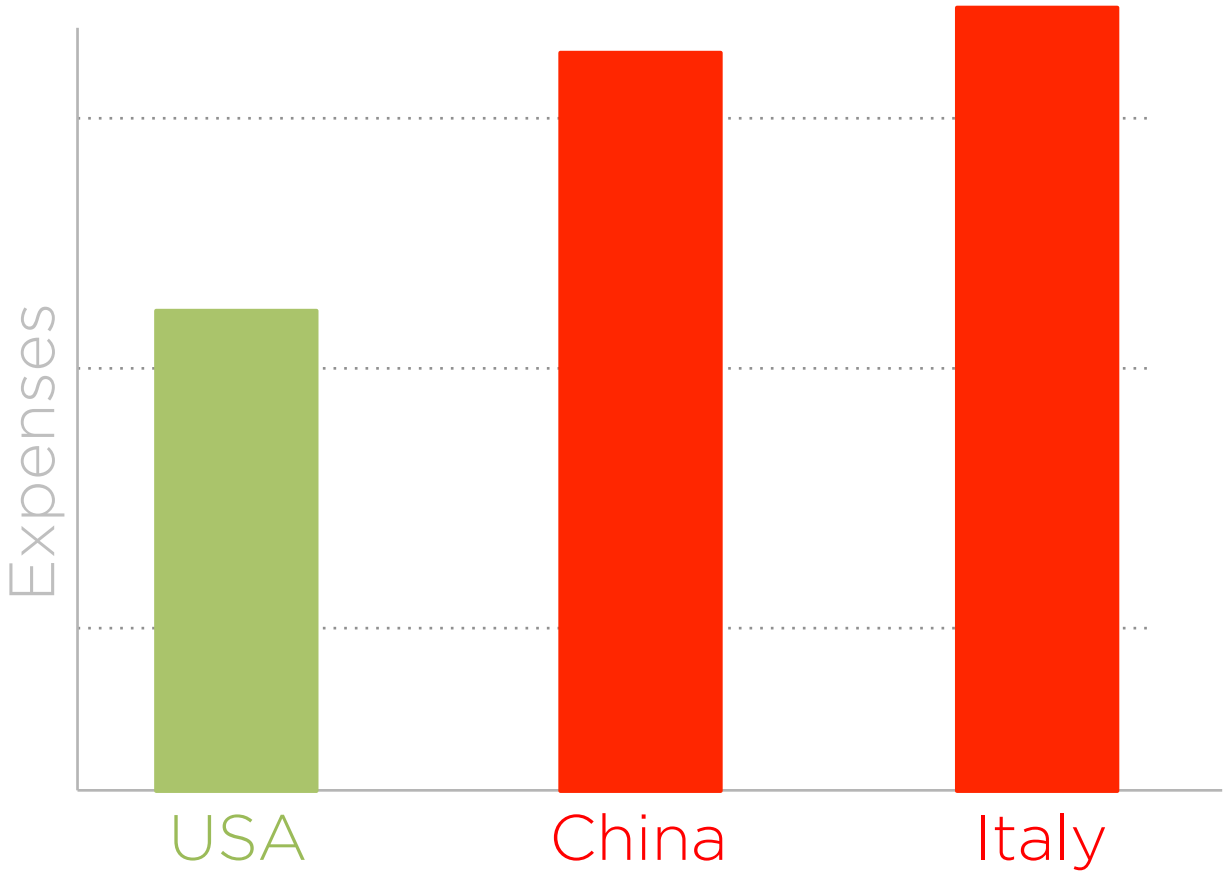
Outlier and normal results

What input properties

caused the outliers?

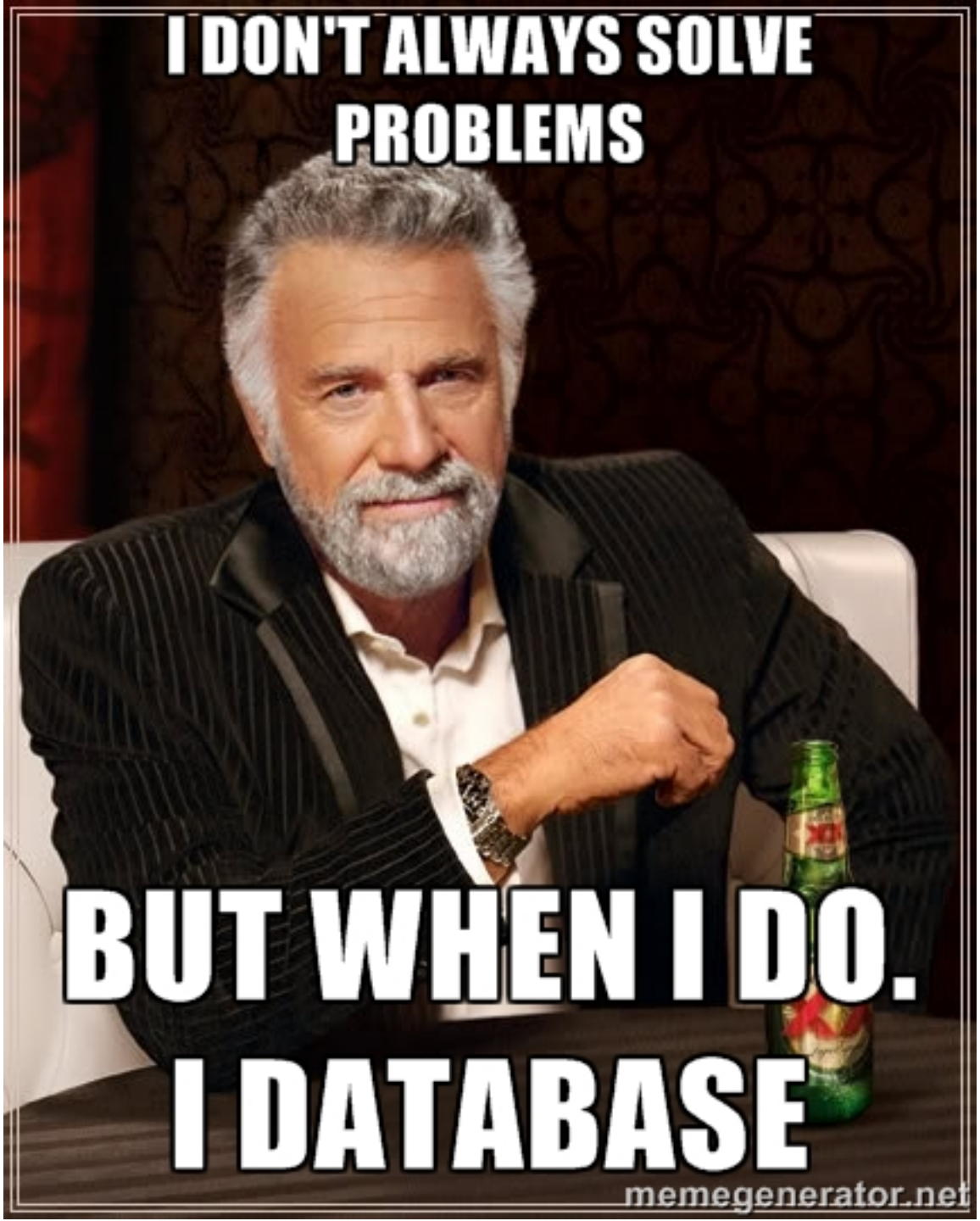
most caused the outliers?

caused outliers but didn't affect normal outputs?




```
SELECT sum(cost)
FROM   expenses
GROUPBY country
```

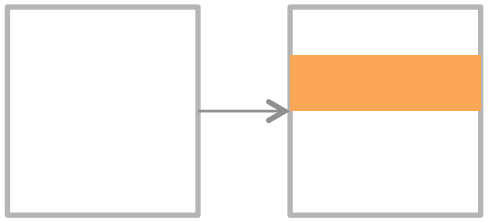
Can’t Touch This


I DON'T ALWAYS SOLVE PROBLEMS
BUT WHEN I DO.
I DATABASE
memegenerator.net

Provenance

Provenance

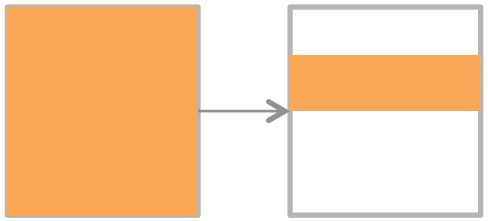
SELECT SUM(cost)
FROM sam's bank account



$$$

Provenance

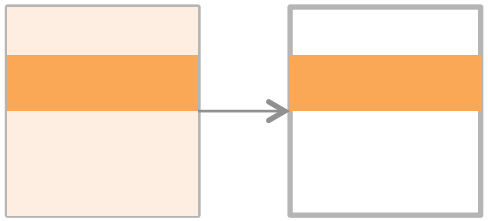
SELECT SUM(cost)
FROM sam's bank account



$$$

Provenance

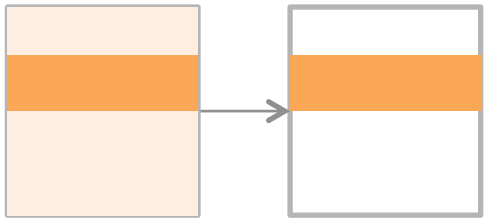
SELECT SUM(cost)
FROM sam's bank account



$$$

Provenance

SELECT SUM(cost)
FROM sam's bank account



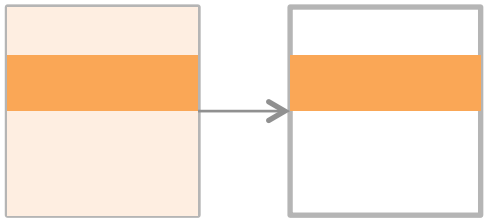
$$$

Proven

http://weknowmemes.com/2012/04/whats-the-point/

Provenance

SELECT SUM(cost)
FROM sam's bank account

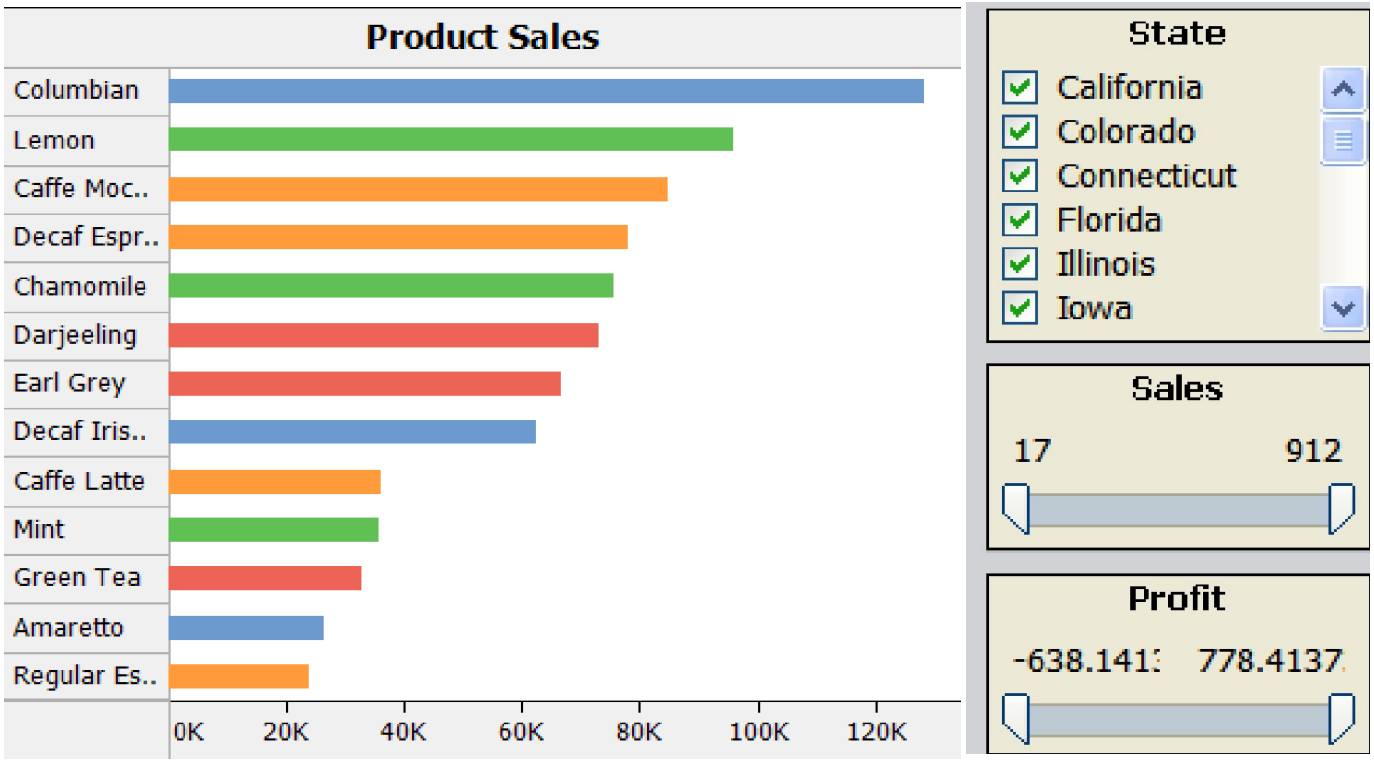


Filter for
"most influential"

~~Provenance~~

~~Provenance~~

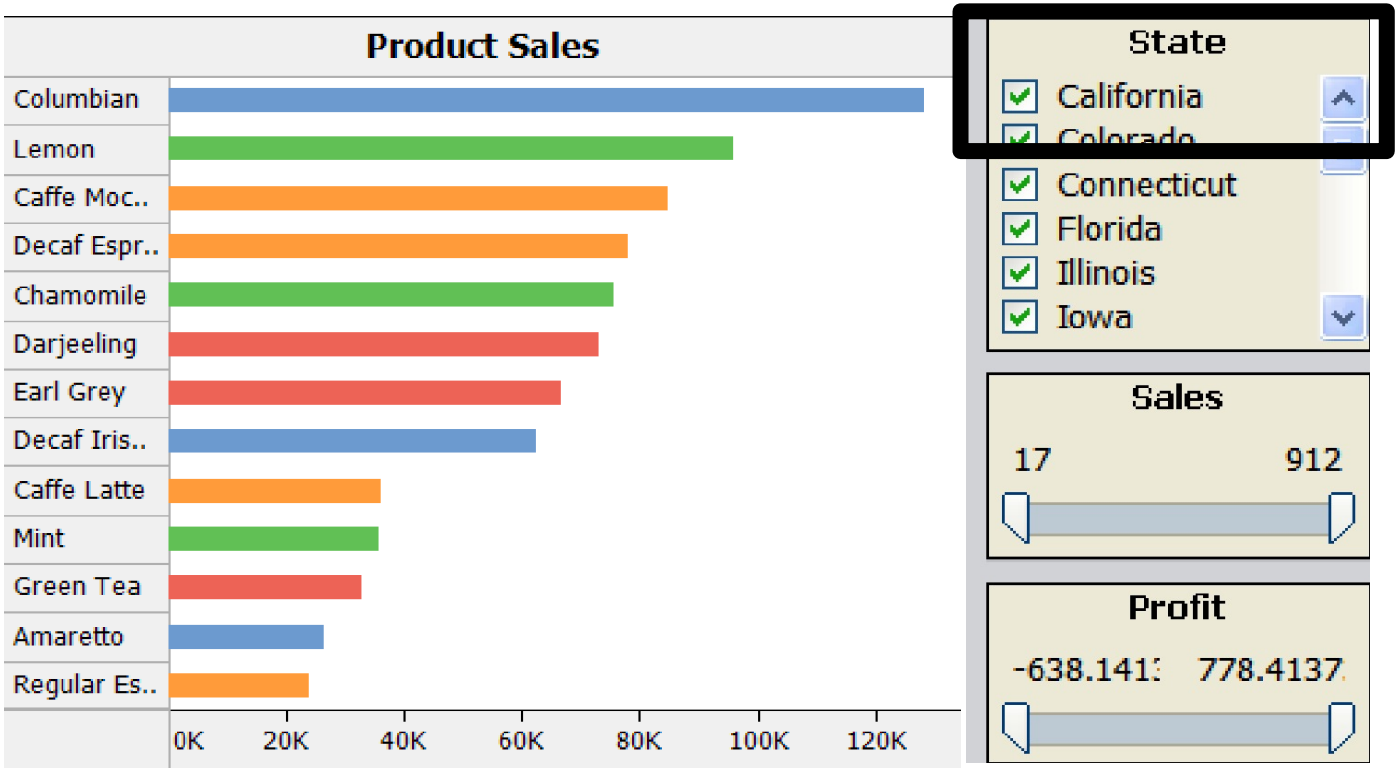
Faceting



http://www.perceptualedge.com/articles/Whitepapers/Three_Blind_Men.pdf

~~Provenance~~

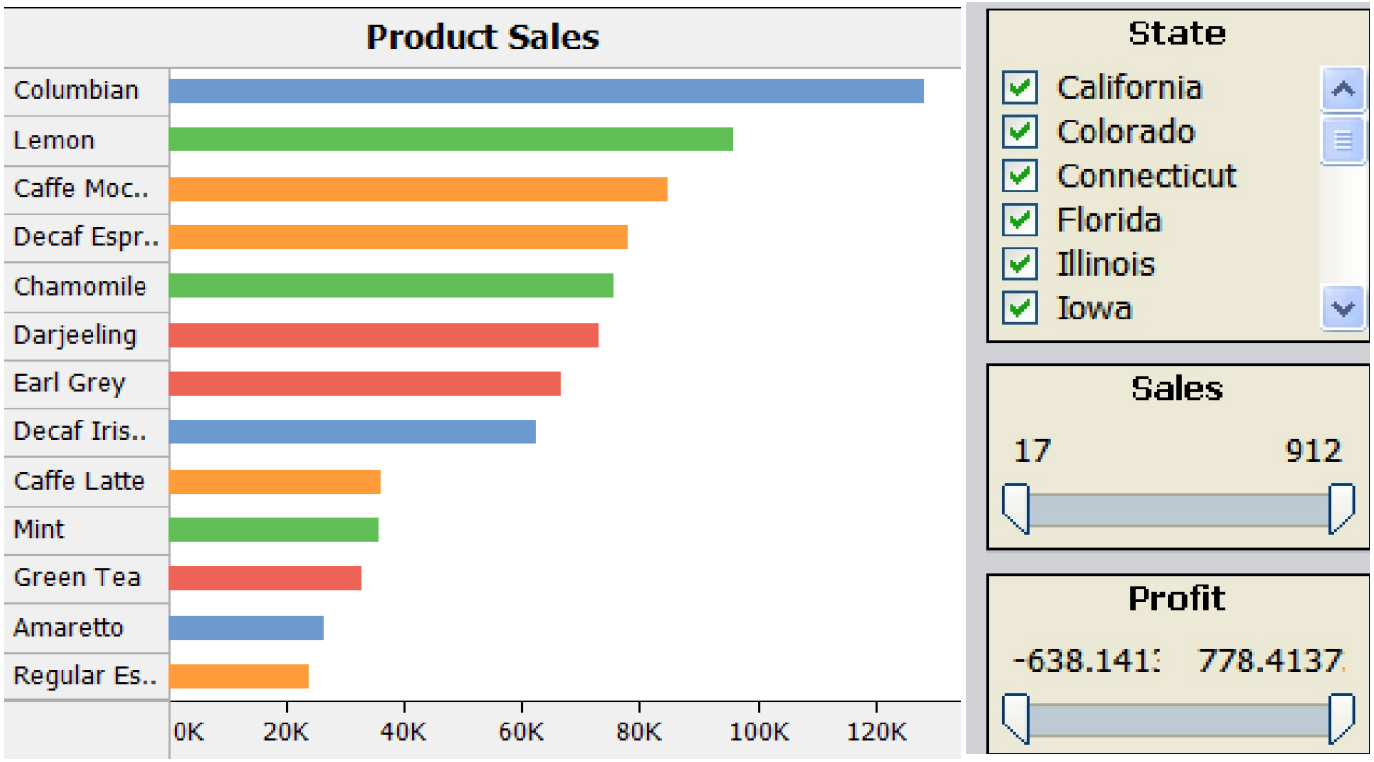
Faceting



http://www.perceptualedge.com/articles/Whitepapers/Three_Blind_Men.pdf

~~Provenance~~

Faceting



Dimensionality :(

Dealing with multiple outliers?

http://www.perceptualedge.com/articles/Whitepapers/Three_Blind_Men.pdf

~~Provenance~~

~~Faceting~~

~~Provenance~~

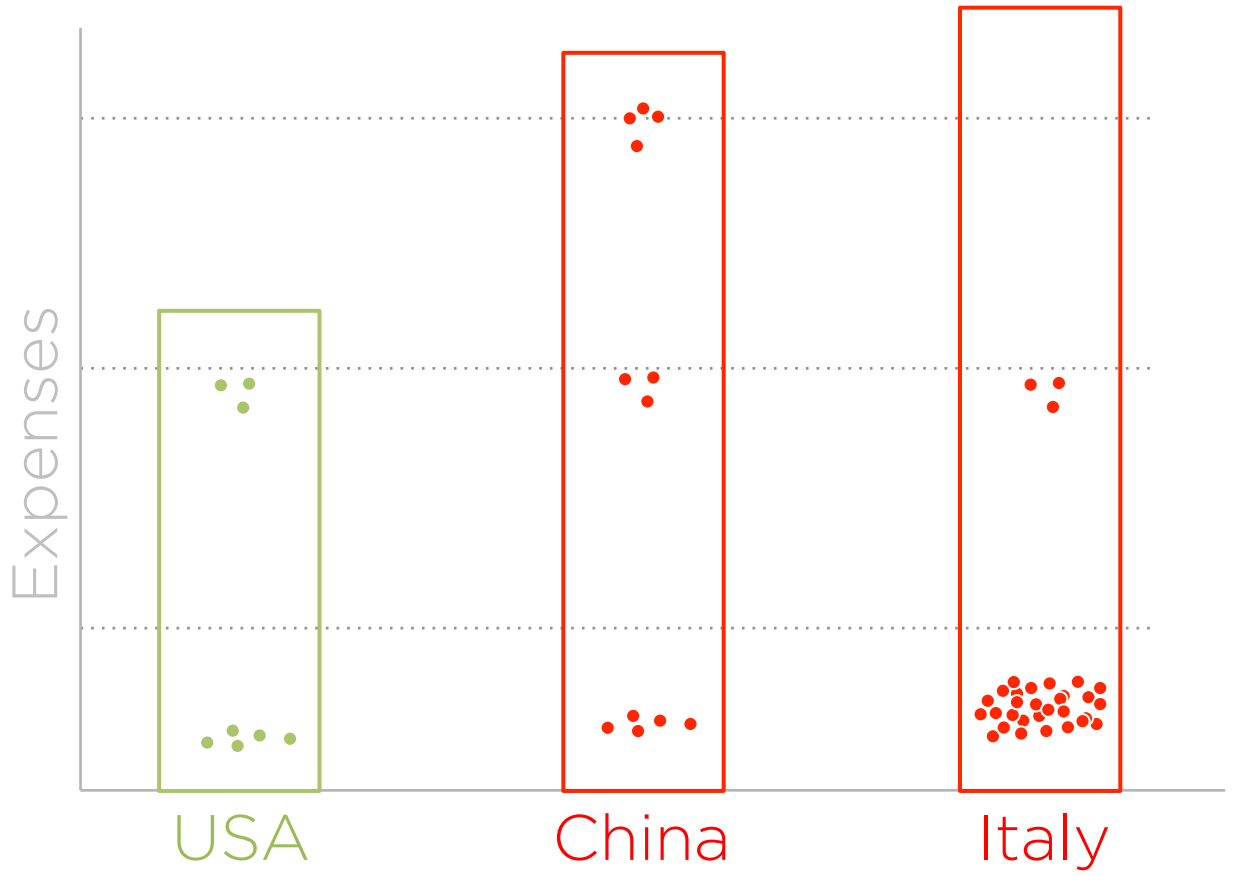
~~Faceting~~

Scorpion!

Given

Outlier and normal results

Understand Why

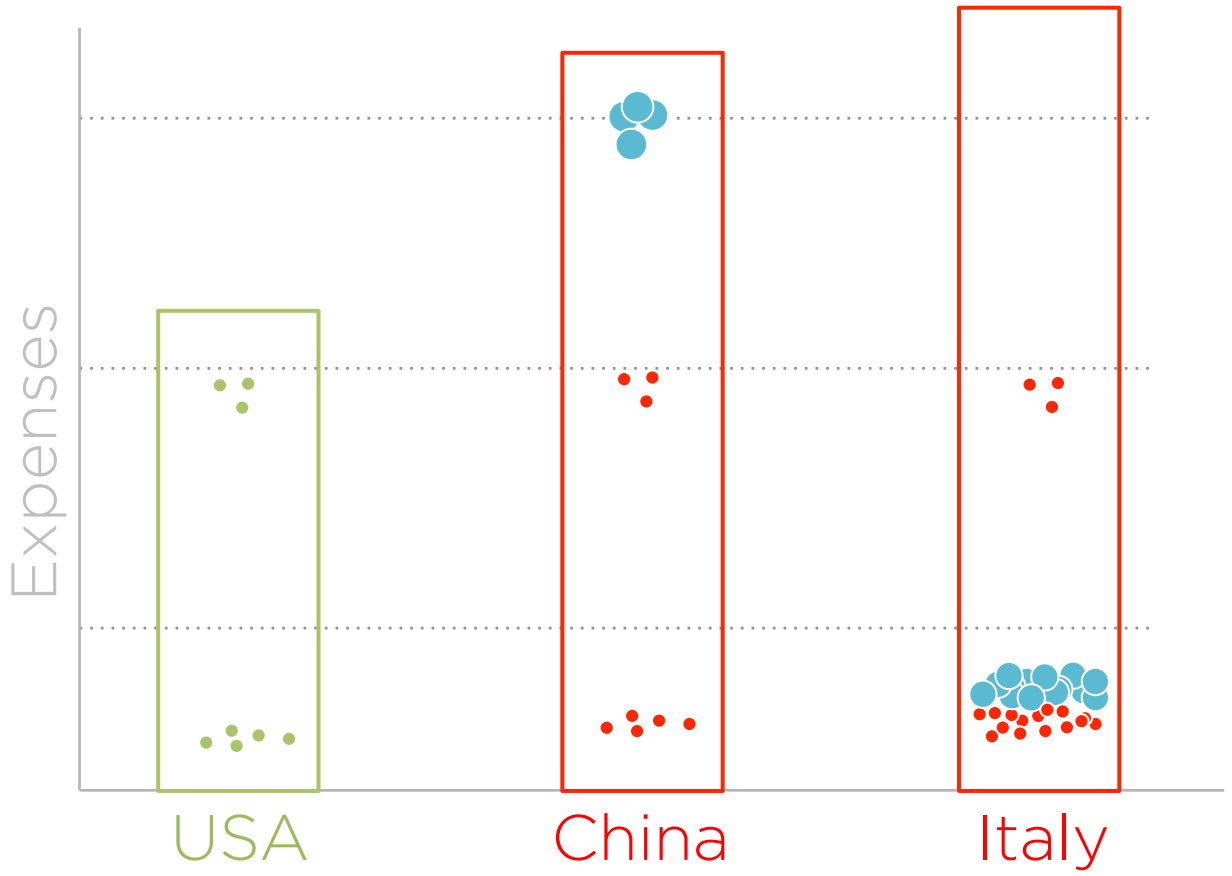


Given

Outlier and normal results

Find

Predicates correlated with outliers



Desc = “toilets”

Given

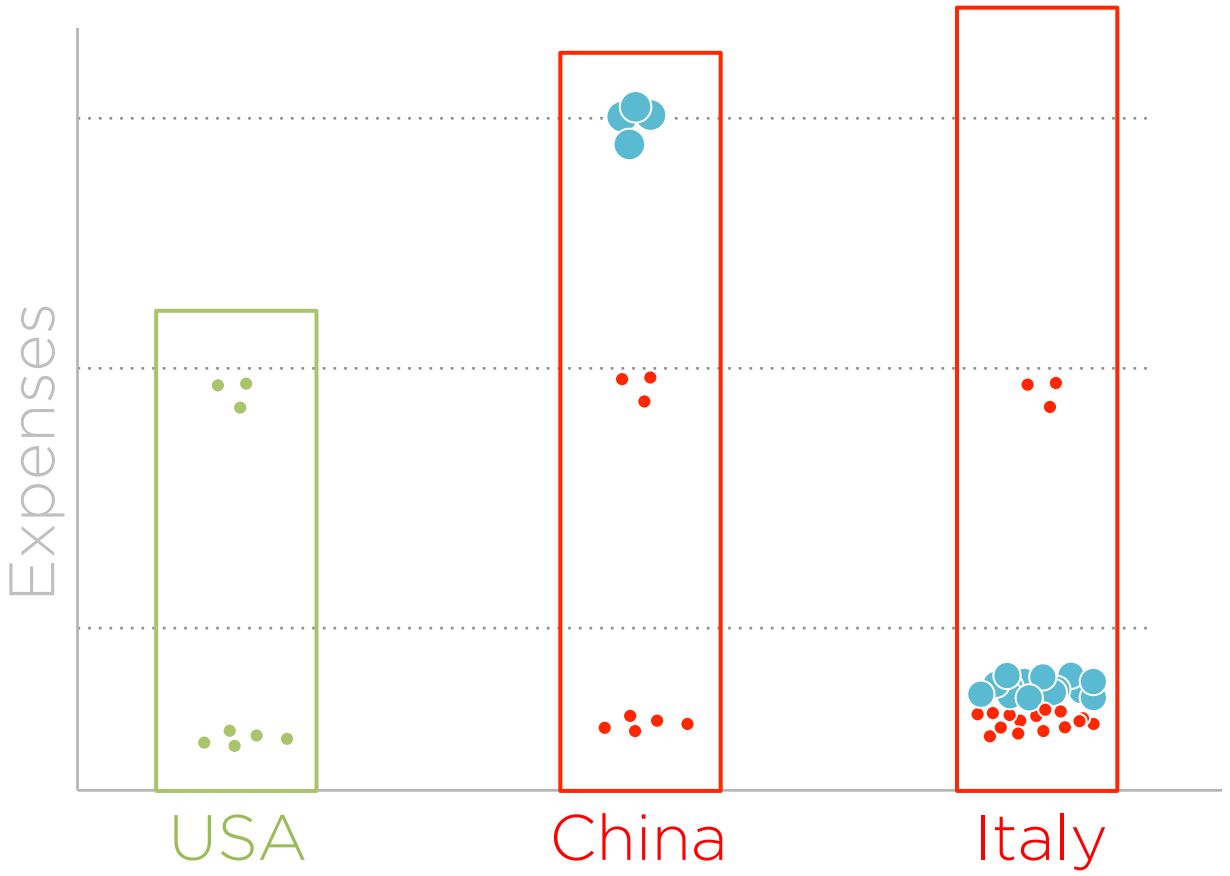
Outlier and normal results

Find

Predicates correlated with outliers

s.t.

Removing predicate from inputs “fixes” outliers & maintains normal results



Desc = “toilets”

Given

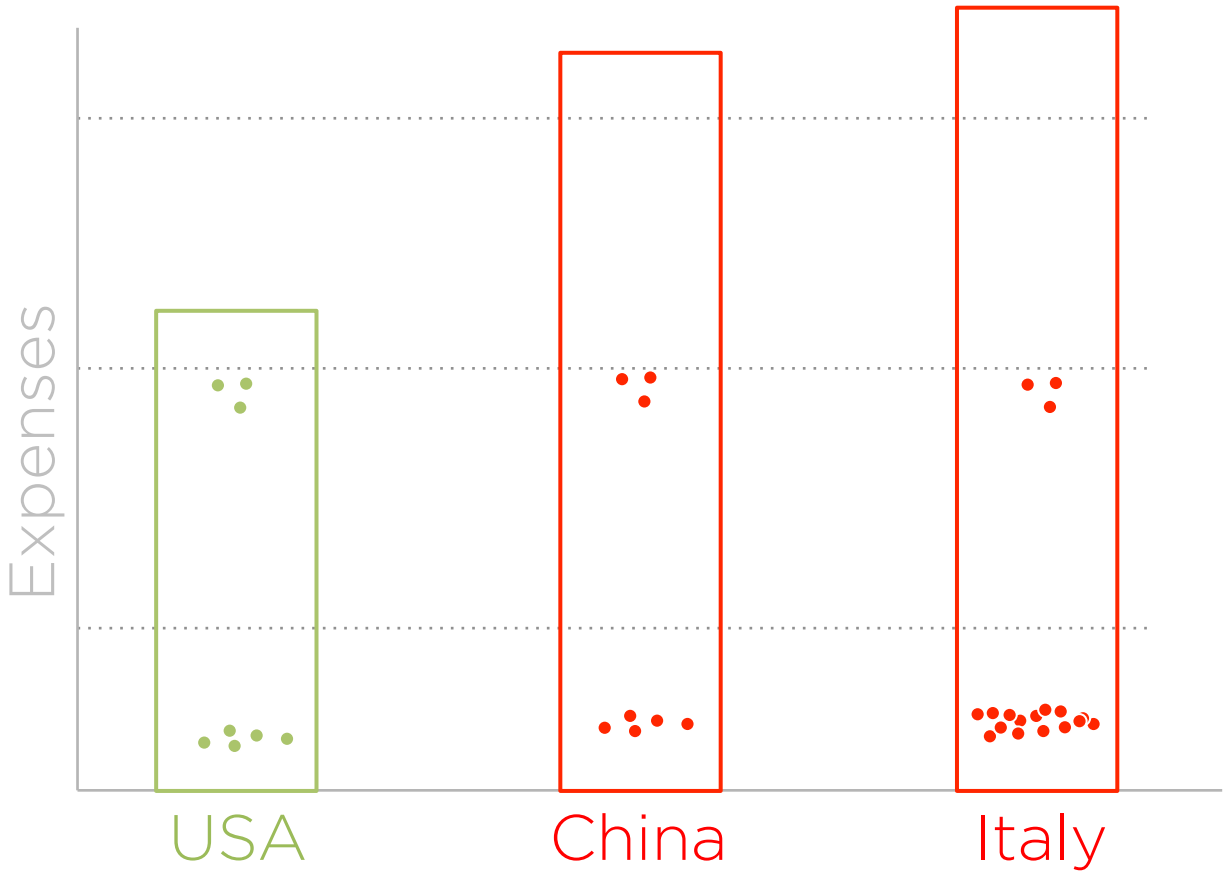
Outlier and normal results

Find

Predicates correlated with outliers

s.t.

Removing predicate from inputs “fixes” outliers & maintains normal results



~~Dese = “toilets”~~

Given

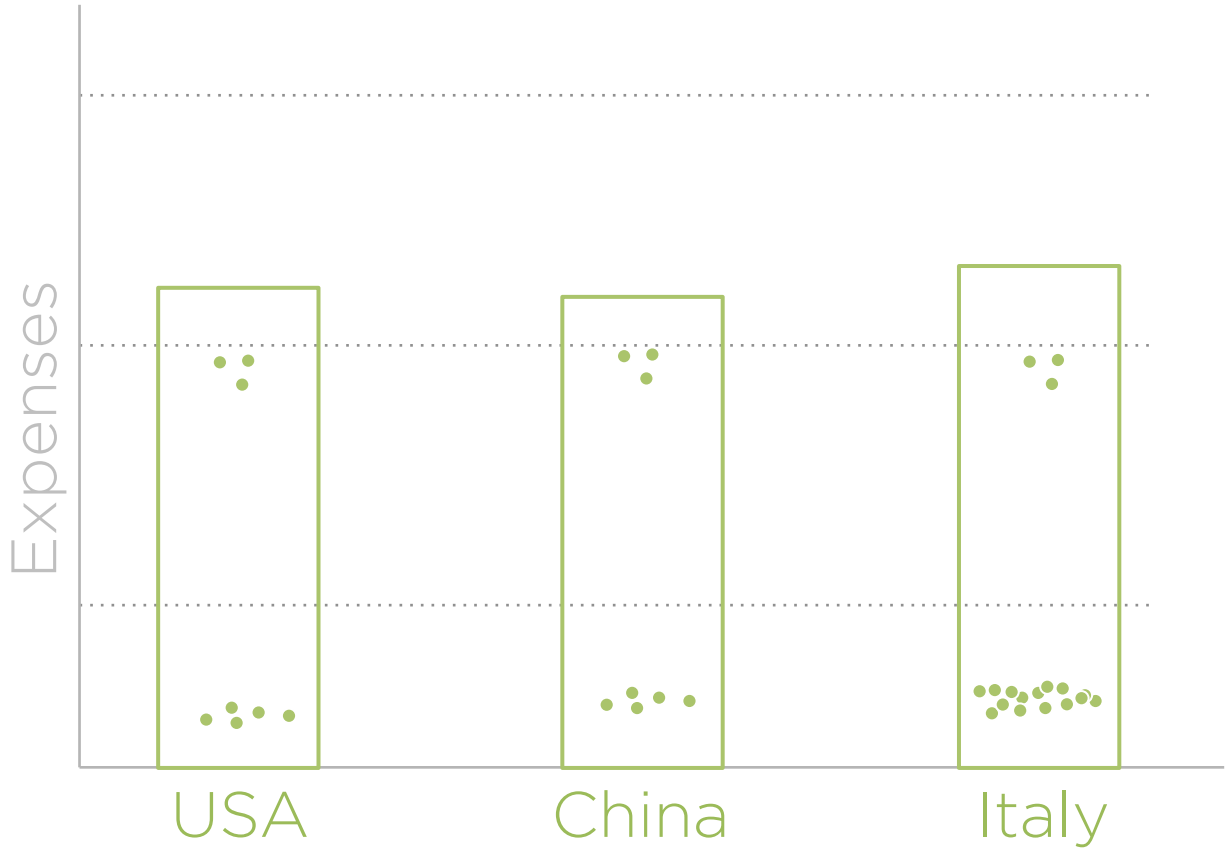
Outlier and normal results

Find

Predicates correlated with outliers

s.t.

Removing predicate from inputs “fixes” outliers & maintains normal results



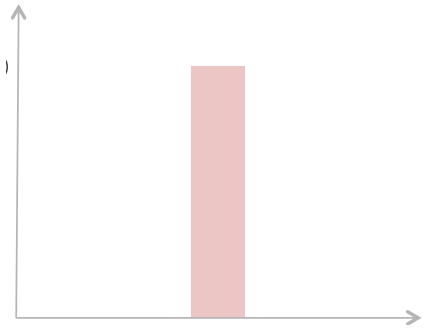
~~Dese = “toilets”~~

Formalize “influence” as metric

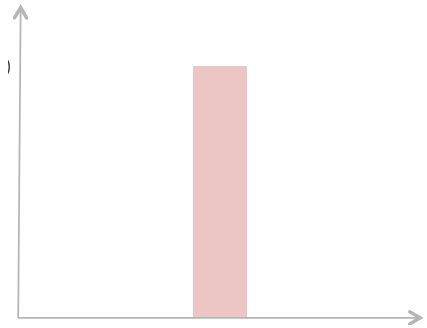
Predicate search heuristics

Some results


T

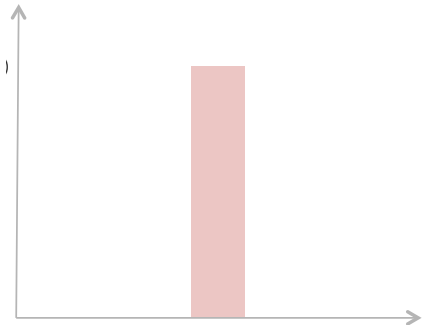


T



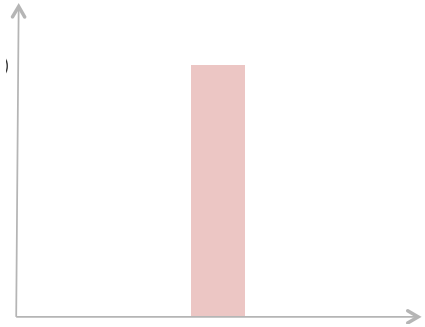
p Desc = “toilet”

T



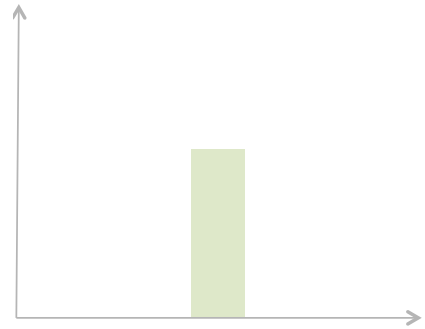
p(T)

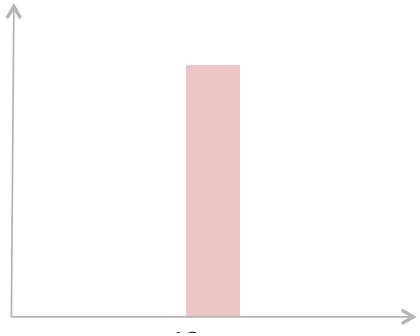
T



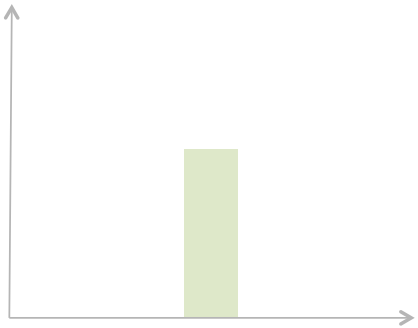
p(T)

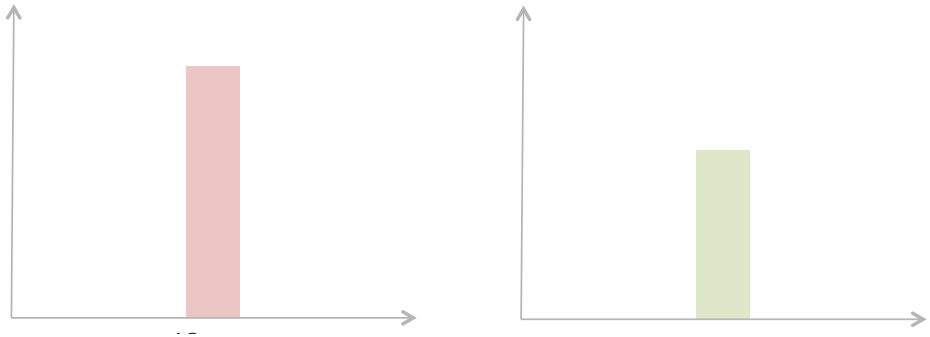
T – p(T)



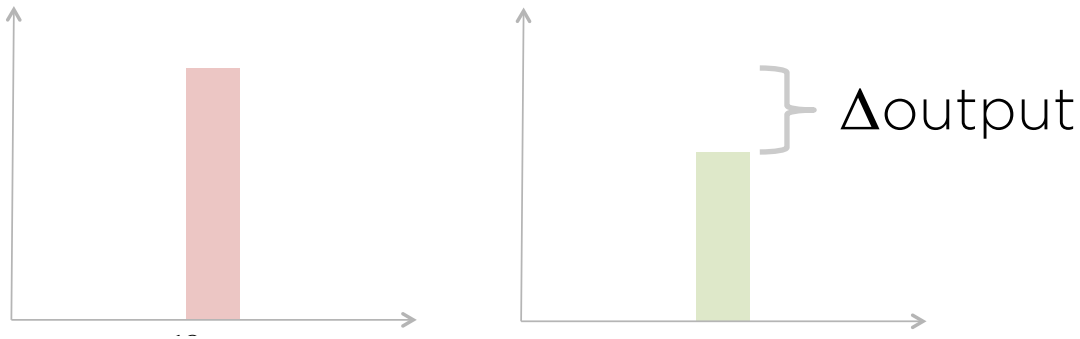


p(T)

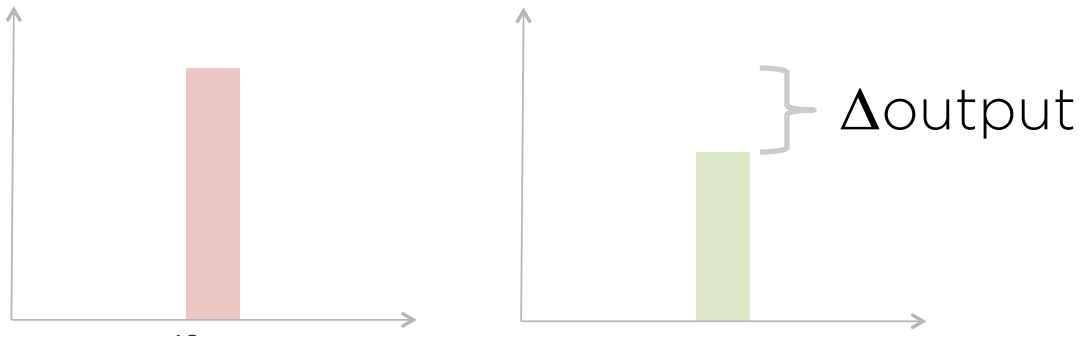




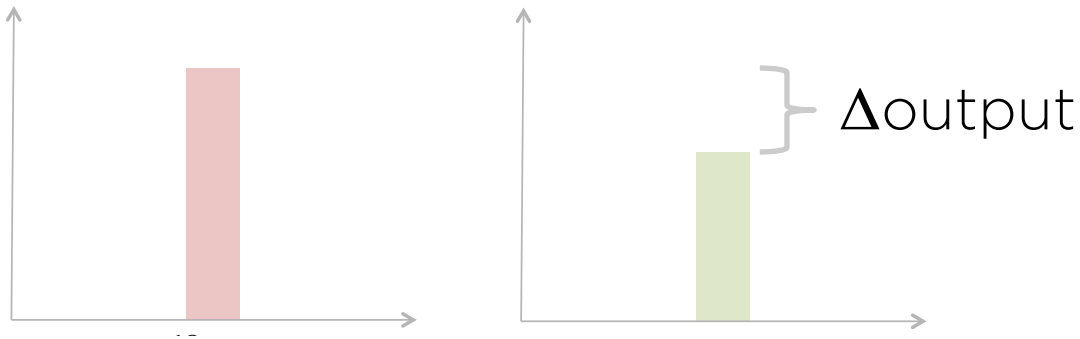
p(T)



p(T)


Δoutput


p(T)
|p(T)|



$$\frac{\Delta \text{output}}{|p(T)|}$$

Influence Metric

$$\frac{\Delta f(x)}{\Delta x}$$

Sensitivity Analysis

$$\frac{\Delta \text{output}}{|p(T)|}$$

Influence Metric

Δ Output

"High vs Low"

|p(T)|

Δ Normal

Multiple Outputs

$$\frac{\Delta \text{output}}{|p(T)|}$$

Δ Output

“High vs Low”

|p(T)|

Δ Normal

Multiple Outputs

$$\frac{\Delta output}{|p(T)|}$$

$$\frac{\Delta output \bullet V}{|p(T)|}$$

Δ Output

$$\frac{\Delta output}{|p(T)|}$$

"High vs Low"

$$\frac{\Delta output \bullet V}{|p(T)|}$$

|p(T)|

$$\frac{\Delta output \bullet V}{|p(T)|^c}$$

Δ Normal

Multiple Outputs

Δ Output

$$\frac{\Delta output}{|p(T)|}$$

"High vs Low"

$$\frac{\Delta output \bullet V}{|p(T)|}$$

|p(T)|

$$\frac{\Delta output \bullet V}{|p(T)|^c}$$

Δ Normal

$$\frac{\Delta outlier \bullet V}{|p(T)|^c} - |\Delta Normal|$$

Multiple Outputs

Δ Output

$$\frac{\Delta \text{output}}{|p(T)|}$$

"High vs Low"

$$\frac{\Delta \text{output} \cdot V}{|p(T)|}$$

$|p(T)|$

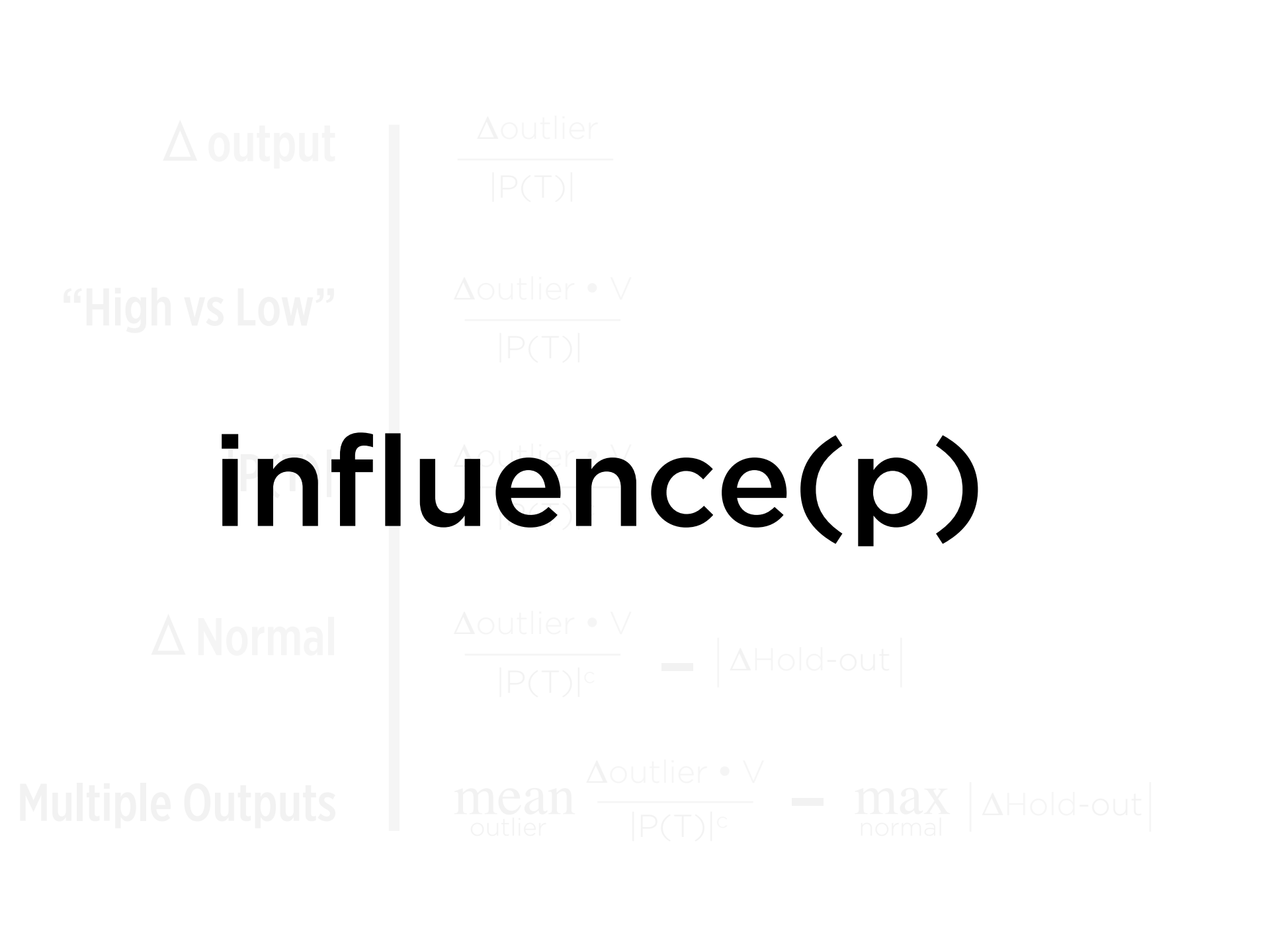
$$\frac{\Delta \text{output} \cdot V}{|p(T)|^c}$$

Δ Normal

$$\frac{\Delta \text{outlier} \cdot V}{|p(T)|^c} - |\Delta \text{Normal}|$$

Multiple Outputs

$$\underset{\text{outlier}}{\text{mean}} \frac{\Delta \text{outlier} \cdot V}{|p(T)|^c} - \underset{\text{normal}}{\max} |\Delta \text{Normal}|$$


Δ output
"High vs Low"
Δ Normal
Multiple Outputs
Δoutlier / |P(T)|
Δoutlier • V / |P(T)|
Δoutlier • V / |P(T)|^c − |ΔHold-out|
mean_outlier Δoutlier • V / |P(T)|^c − max_normal |ΔHold-out|
influence(p)

Formalize “influence” as metric

Predicate search heuristics

Some results

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \ \text{influence}(p)$$

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \; \text{influence}(p)$$

$O(\text{agg}(T - p(T)))$

SUM({1,2,3,4,5}) = 15

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{influence}(p)$$

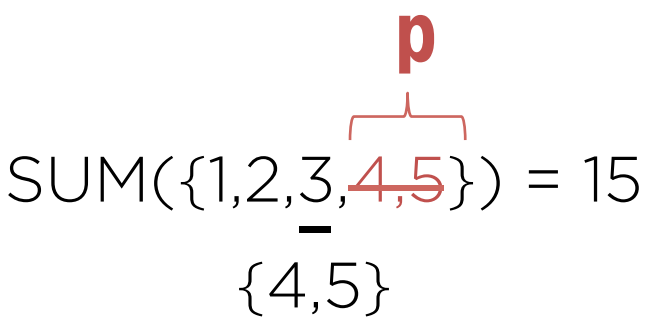
O(agg(T-p(T)))

p

SUM({1,2,3,4,5}) = 15

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{influence}(p)$$

O(agg(T-p(T)))



$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{influence}(p)$$

$O(\text{agg}(T\text{-}p(T)))$

p

SUM({1,2,3,4,5}) = 15

{4,5}

SUM({1,2,3}) = 6

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{influence}(p)$$

O(agg(T-p(T)))

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \; \text{influence}(p)$$

O(exponential) O(agg(T-p(T)))

Operator Properties

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \; \text{influence}(p)$$

O(exponential) O(agg(T-p(T)))

Operator Properties

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \; \text{influence}(p)$$

O(exponential)

O(agg(p(T)))

Incrementally removable

p

SUM({1,2,3,~~4,5~~}) = 15

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{ influence}(p)$$

O(exponential)

O(agg(p(T)))

Incrementally removable

p

SUM({1,2,3,~~4,5~~}) = 15

15 - SUM({ 4,5}) = 6

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{ influence}(p)$$

O(exponential)

O(agg(p(T)))

Incrementally removable

SUM
COUNT
AVG
STDDEV

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{influence}(p)$$

O(exponential)

O(agg(p(T)))

Incrementally removable

~~MEDIAN~~
~~MODE~~

SUM
COUNT
AVG
STDDEV

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{influence}(p)$$

O(exponential)

O(agg(p(T)))

Incrementally removable

Least influence

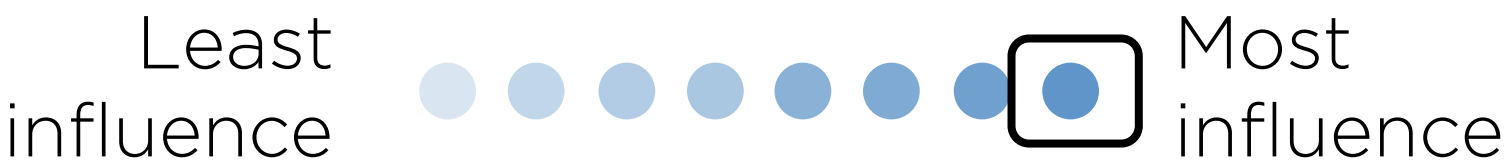
Most influence

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{influence}(p)$$

O(exponential) O(agg(p(T)))

Independent

Incrementally removable



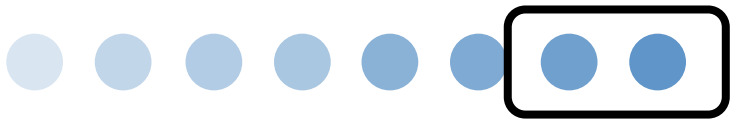
$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{ influence}(p)$$

O(exponential) O(agg(p(T)))

Independent

Incrementally removable

Least influence



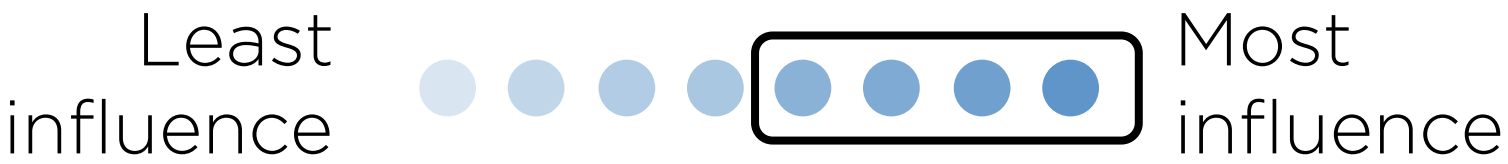
Most influence

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{influence}(p)$$

O(exponential) O(agg(p(T)))

Independent

Incrementally removable



$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{ influence}(p)$$

O(exponential) O(agg(p(T)))

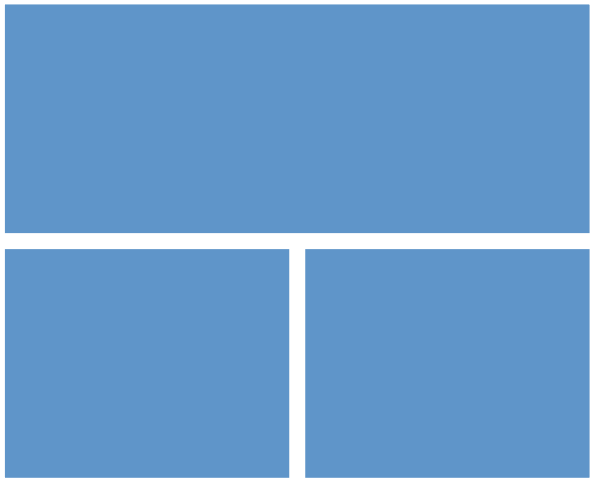
Independent

Incrementally removable

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \; \text{influence}(p)$$

O(exponential) O(agg(p(T)))

Top Down **Independent** **Incrementally removable**



$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{influence}(p)$$

O(exponential) O(agg(p(T)))

Top Down **Independent** **Incrementally removable**

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \; \text{influence}(p)$$

O(exponential) O(agg(p(T)))

Top Down **Independent** **Incrementally removable**

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \; \text{influence}(p)$$

O(exponential) O(agg(p(T)))

Top Down **Independent** **Incrementally removable**

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{influence}(p)$$

O(exponential) O(agg(p(T)))

Top Down

Independent

Anti-monotonic

Incrementally removable

$p' \subset p$

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{ influence}(p)$$

O(exponential) O(agg(p(T)))

Top Down **Independent** **Incrementally removable**

Anti-monotonic

$$p' \subset p$$

$$\updownarrow$$

$$\text{influence}(p') \leq \text{influence}(p)$$

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \ \text{influence}(p)$$

$O(\text{exponential})$ $O(\text{agg}(p(T)))$

Top Down **Independent** **Incrementally removable**

Anti-monotonic

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{influence}(p)$$

O(exponential) O(agg(p(T)))

Top Down **Independent** **Incrementally removable**

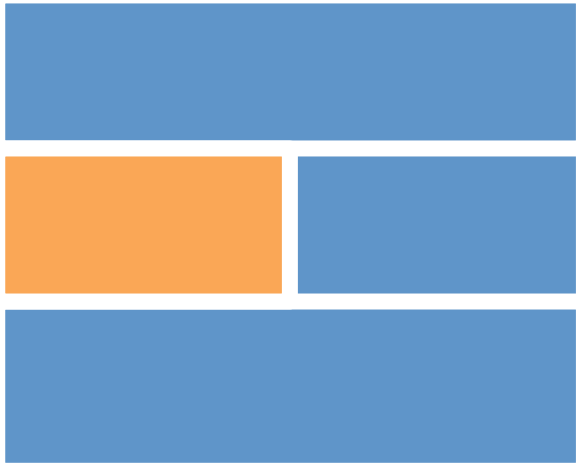
Bottom Up **Anti-monotonic**

$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{influence}(p)$$

O(exponential) O(agg(p(T)))

Top Down **Independent** **Incrementally removable**

Bottom Up **Anti-monotonic**



$$p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{ influence}(p)$$

O(exponential) O(agg(p(T)))

Top Down **Independent** **Incrementally removable**

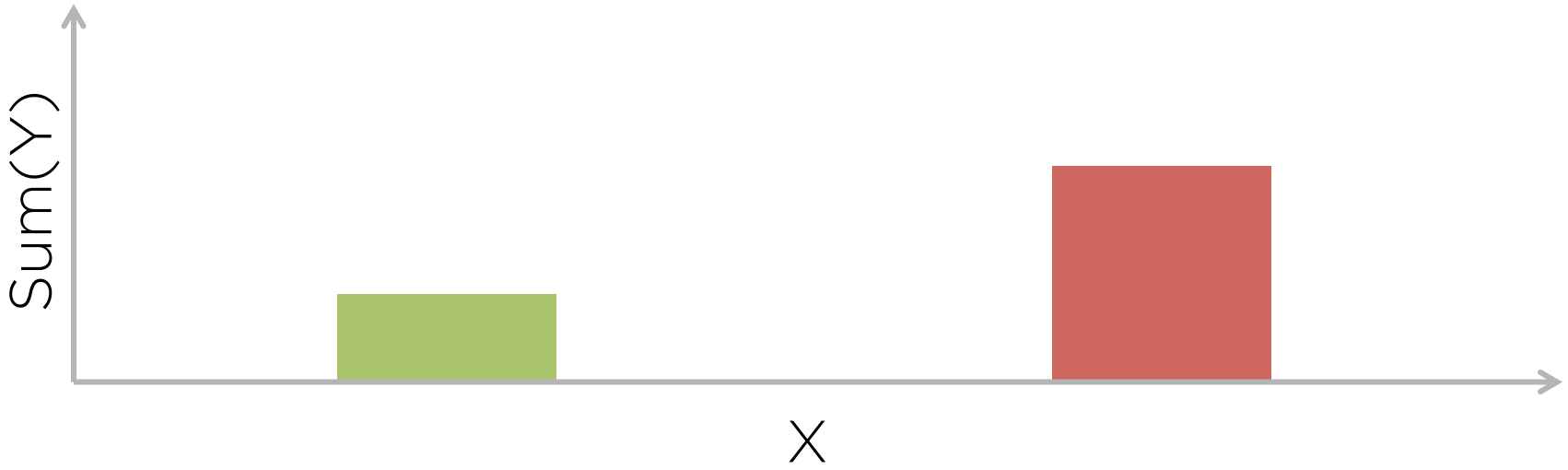
Bottom Up **Anti-monotonic**

Formalize “influence” as metric

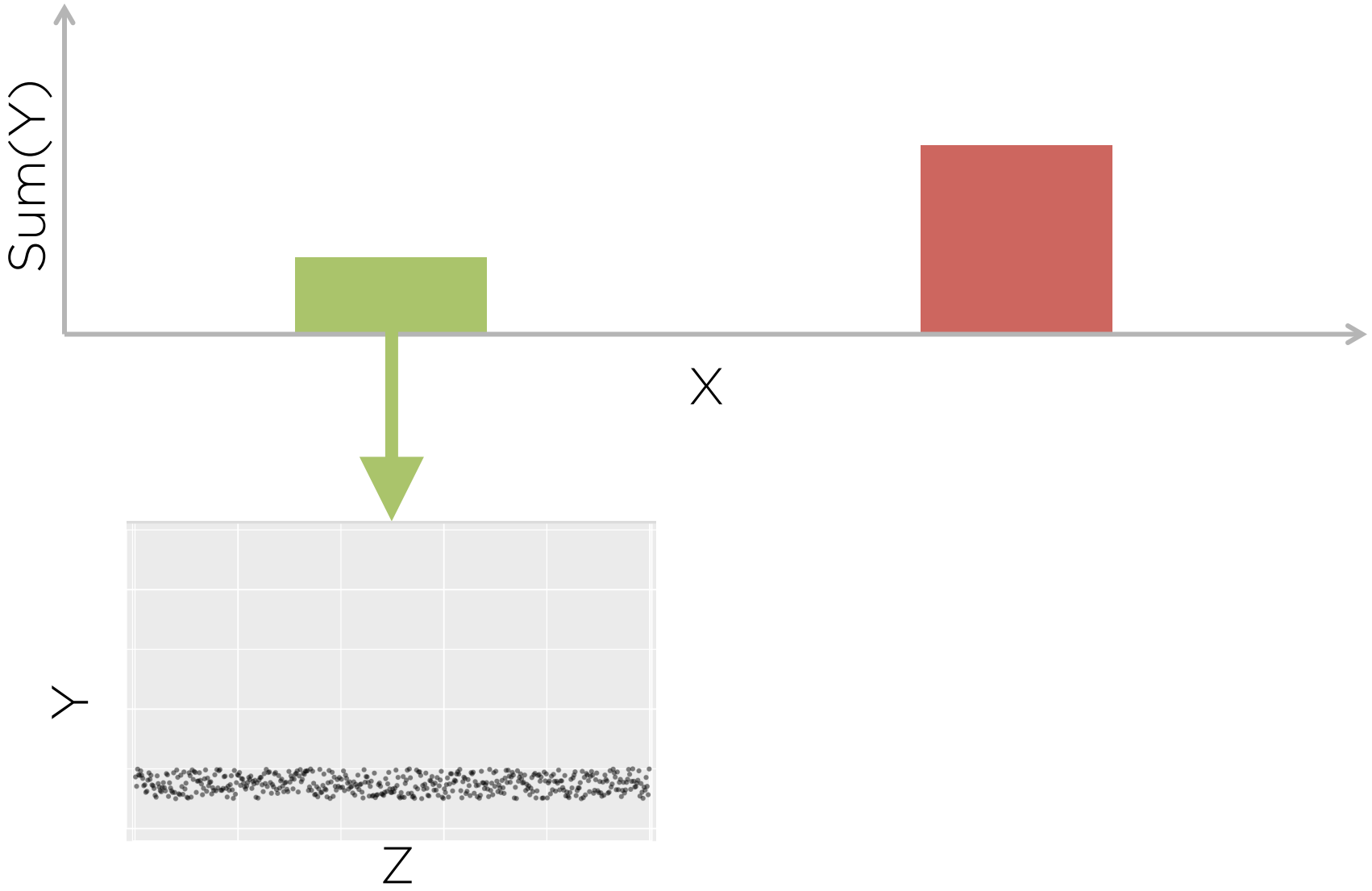
Predicate search heuristics

Some results

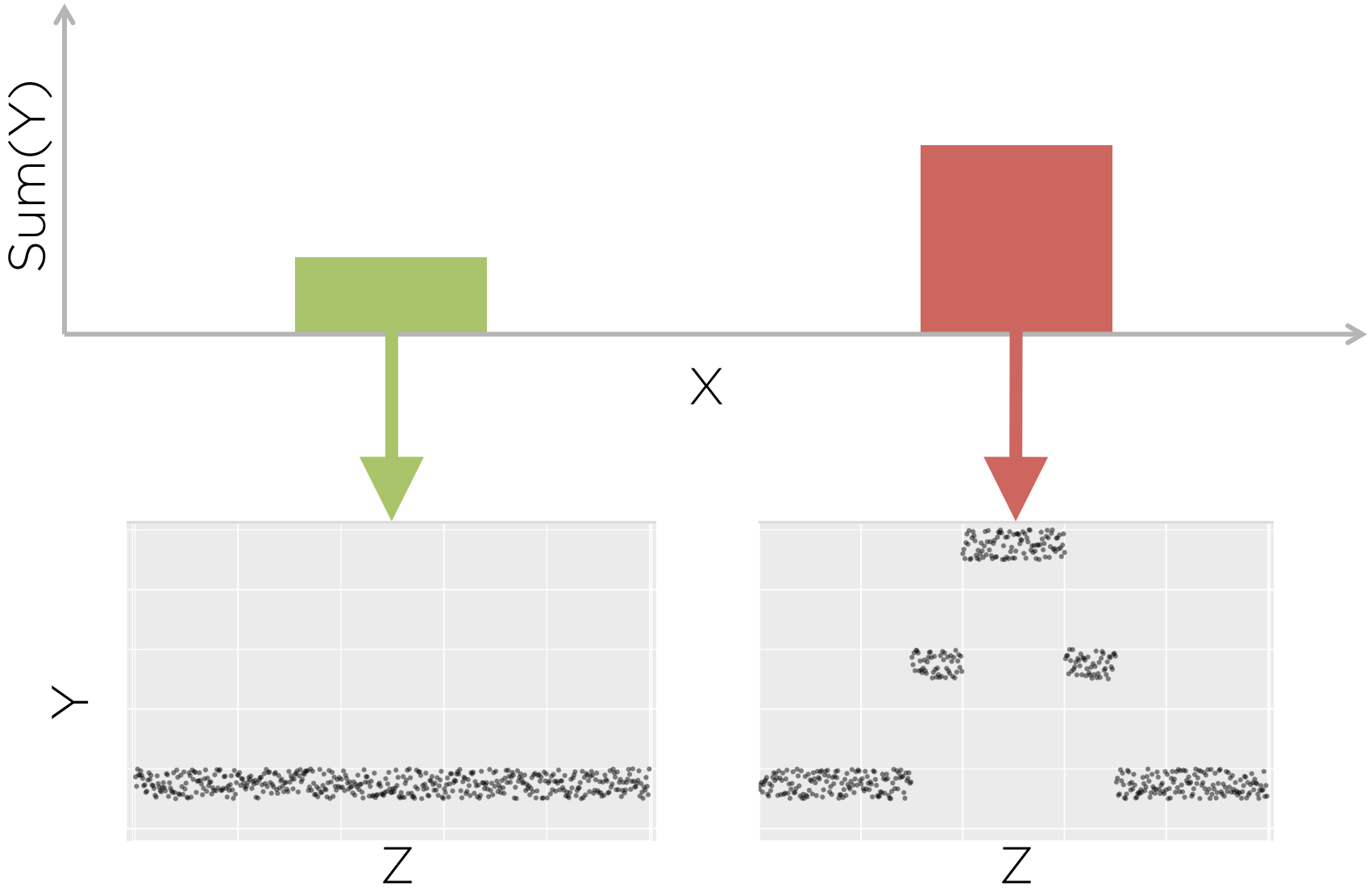
SELECT sum(Y) GROUPBY X



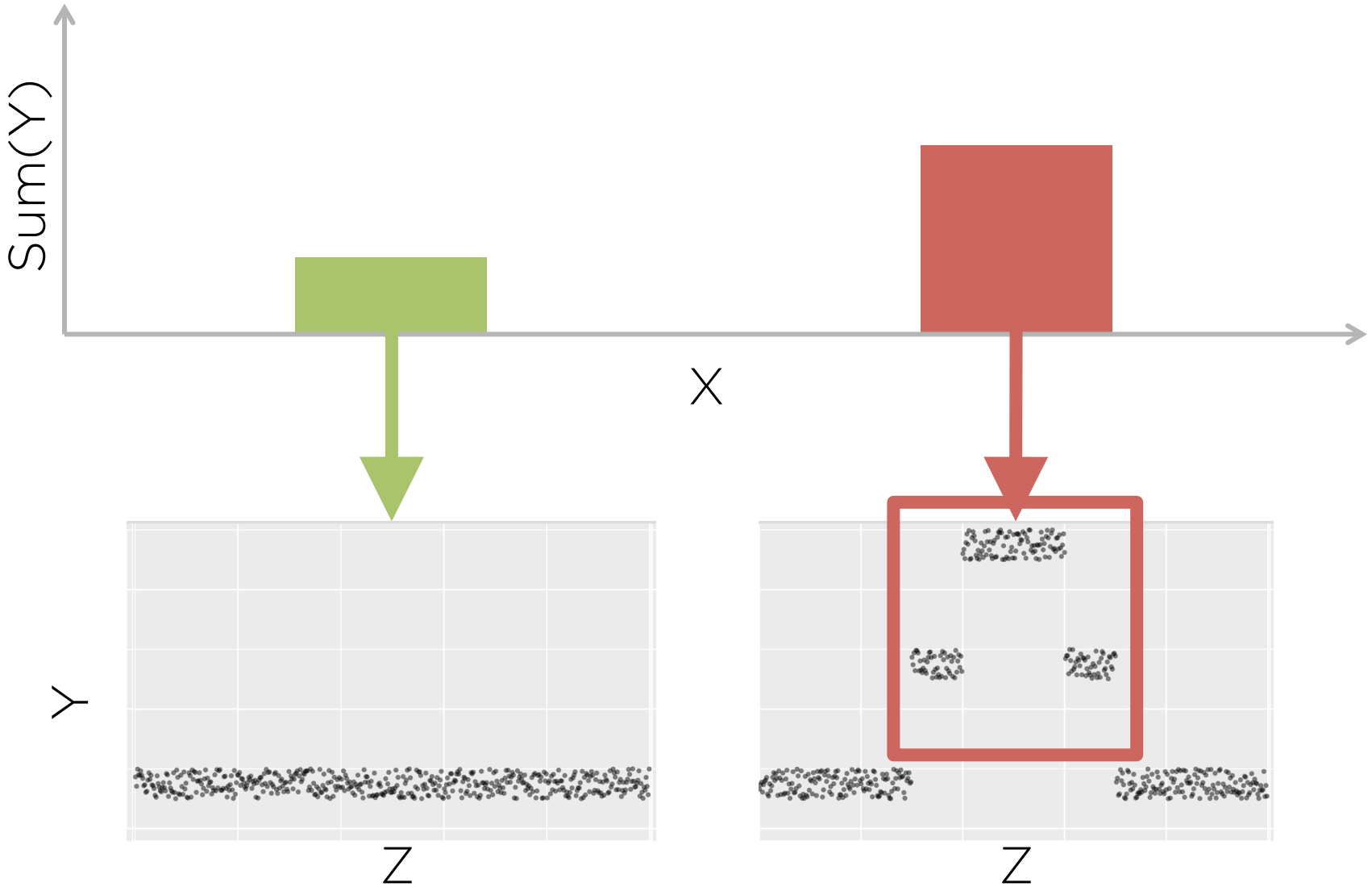
SELECT sum(Y) GROUPBY X

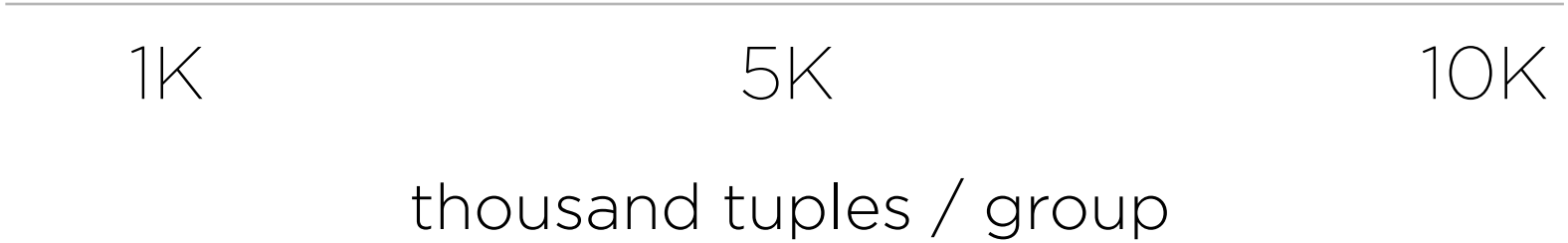


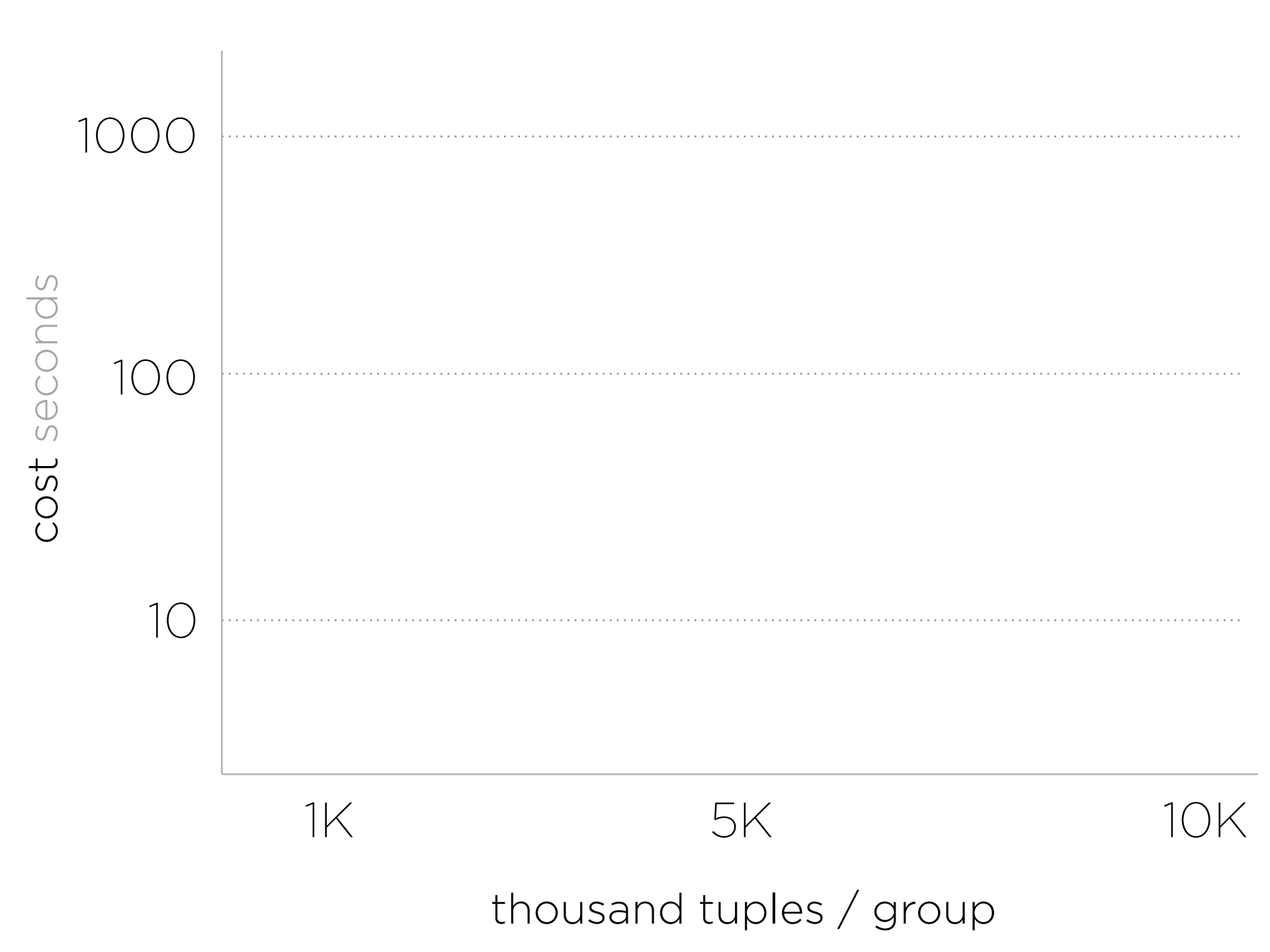
SELECT sum(Y) GROUPBY X

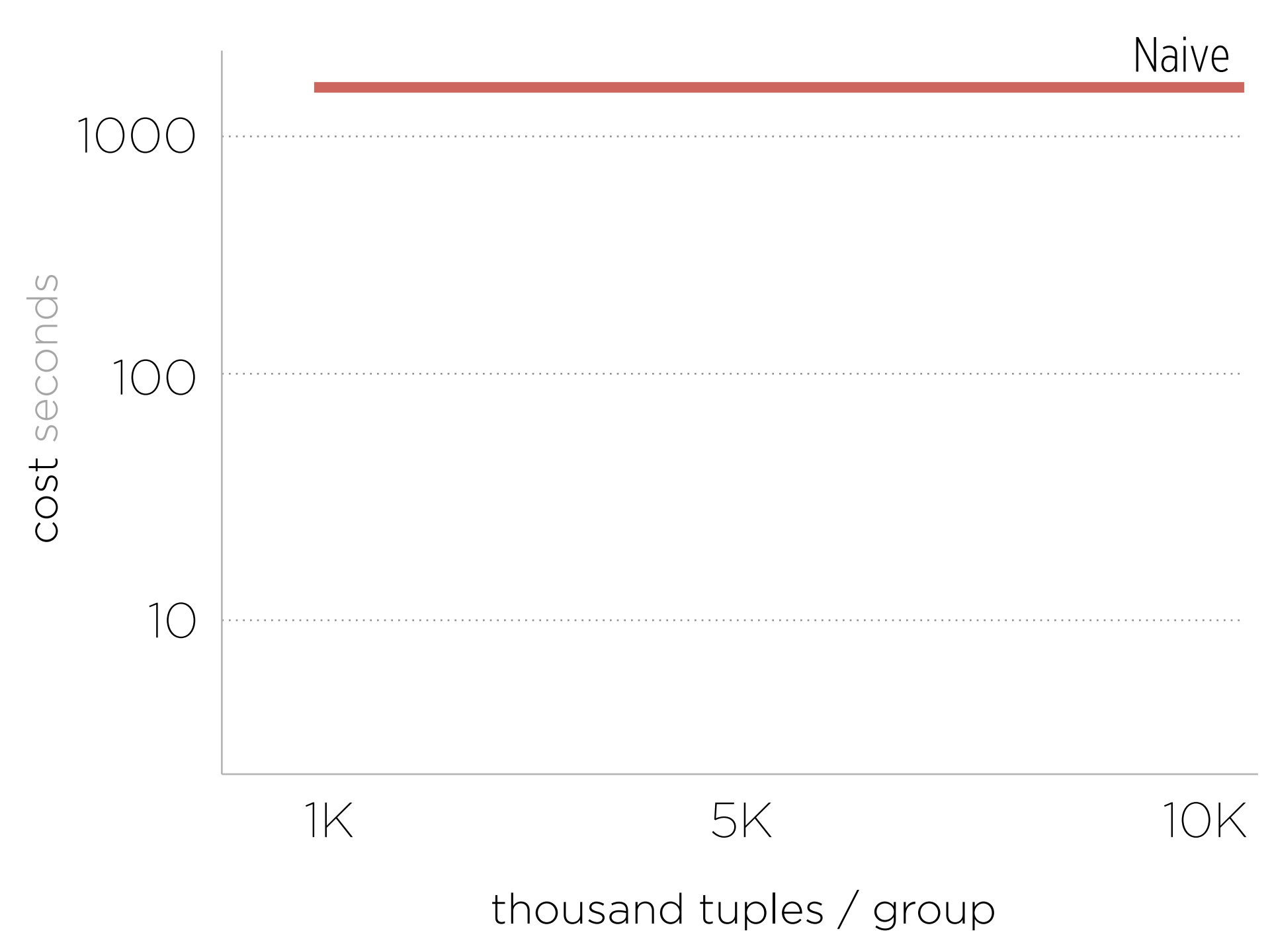


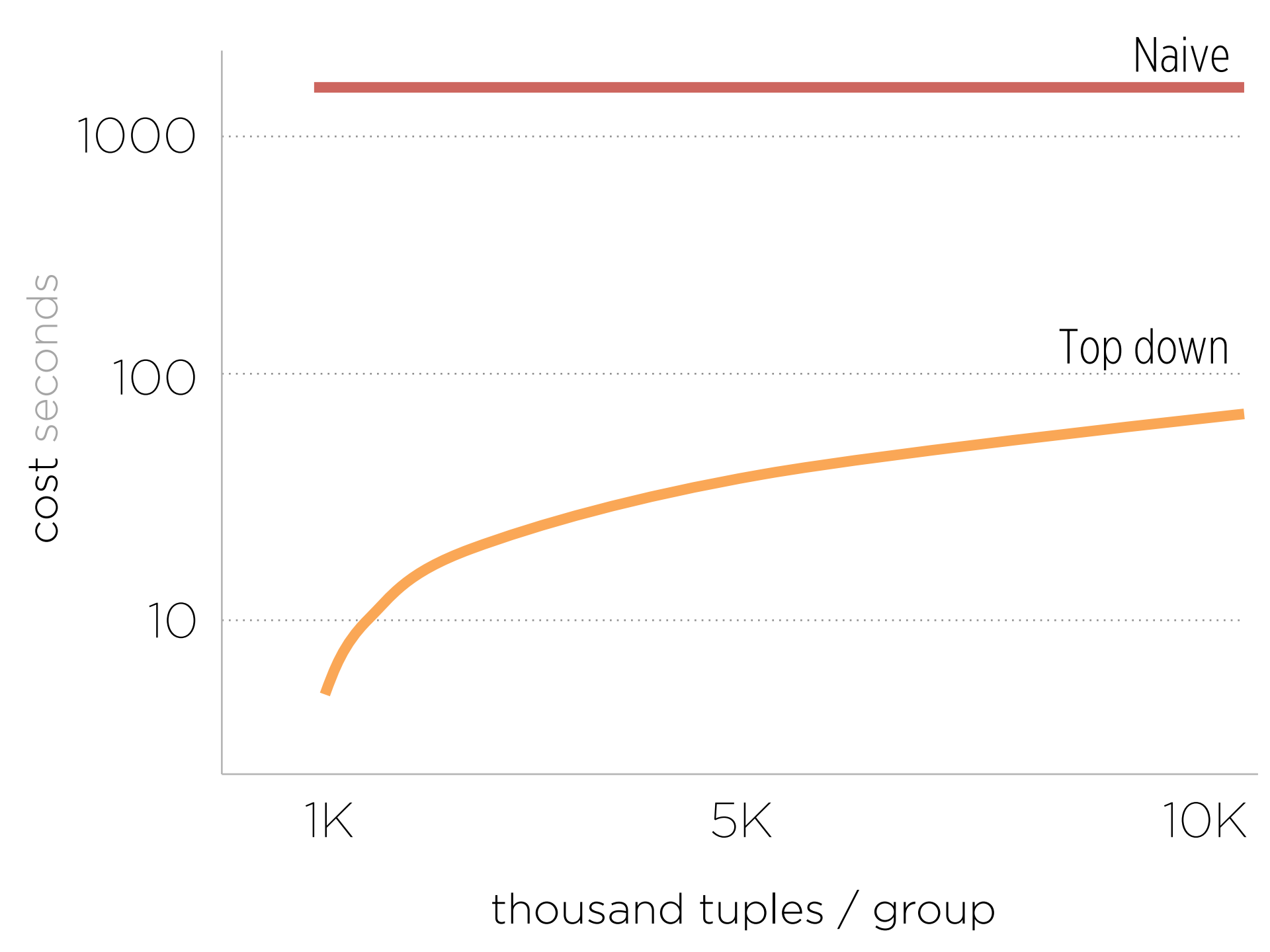
SELECT sum(Y) GROUPBY X

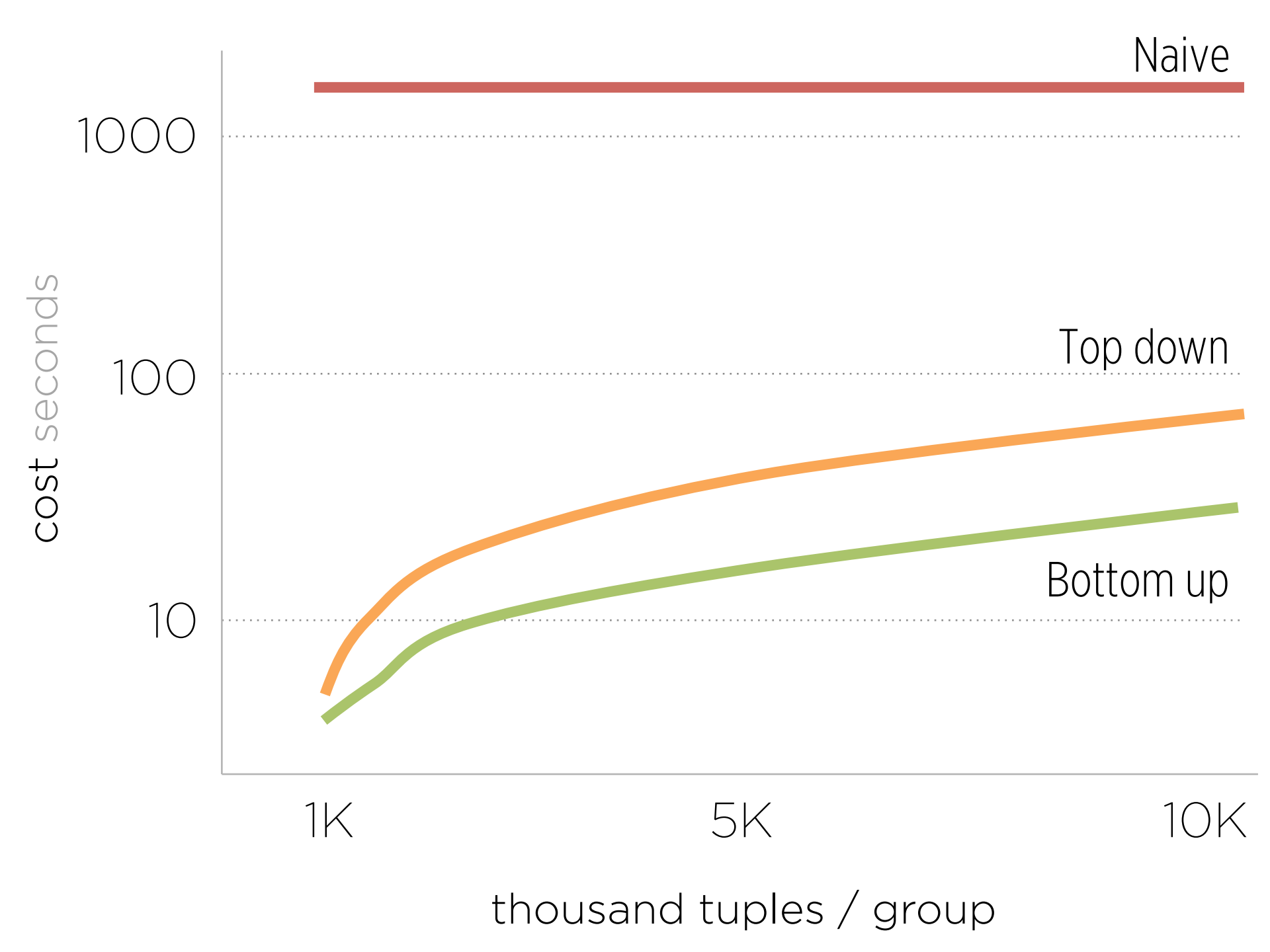



1K
5K
10K
thousand tuples / group


1000
100
10
cost seconds
1K
5K
10K
thousand tuples / group


Naive
1000
100
10
cost seconds
1K
5K
10K
thousand tuples / group


Naive
1000
Top down
100
cost seconds
10
1K
5K
10K
thousand tuples / group


Naive
1000
Top down
100
cost seconds
Bottom up
10
1K
5K
10K
thousand tuples / group

influence metric

that is

accessible to end-users

for

Data cleaning

Data exploration

Provenance reduction

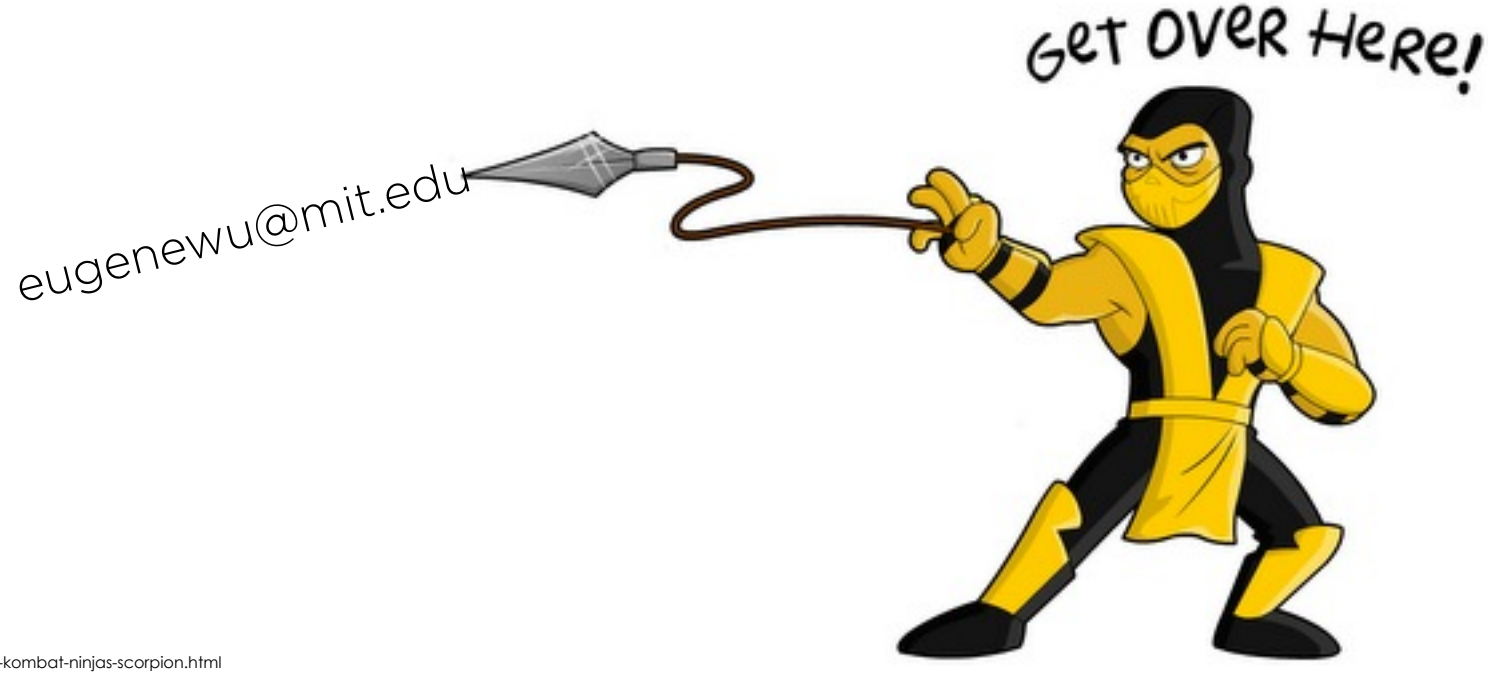
scorpion

eugenewu@mit.edu

http://springfieldpunx.blogspot.com/2010/11/mortal-kombat-ninjas-scorpion.html

scorpion

eugenewu@mit.edu



http://springfieldpunx.blogspot.com/2010/11/mortal-kombat-ninjas-scorpion.html

scorpion

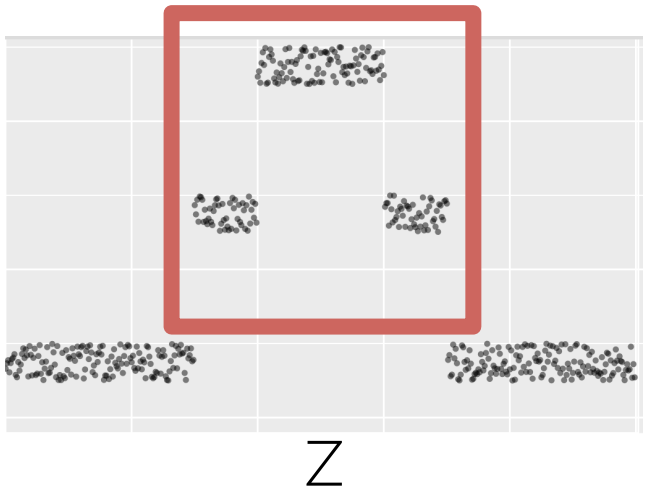
eugenewu@mit.edu

http://springfieldpunx.blogspot.com/2010/11/mortal-kombat-ninjas-scorpion.html

C-parameter

$$\frac{\Delta \text{output} \bullet V}{|p(T)|^{c}}$$

Low C



High C