

Utilizing Provenance as an Attribute for Visual Data Analysis: A Design Probe with ProvenanceLens



Arpit
Narechania



Shunan
Guo



Eunyee
Koh



Alex
Endert



Jane
Hoffswell

Analytic Provenance

Records the history of analytical actions, showing how data was obtained, transformed, and analyzed.

Ragan et al. (TVCG 2015)

North et al. (CHI EA 2011)

Analytic Provenance – Use Cases

- **Track ongoing** analysis
- **Recall past** interactions
- **Inform new** analytic directions
- ...

Analytic Provenance – Challenge

- Is often hardwired into analytics systems, affording
 - Limited user control
 - Limited opportunities for self-reflection

Website Hyperlinks

Unvisited links (blue)

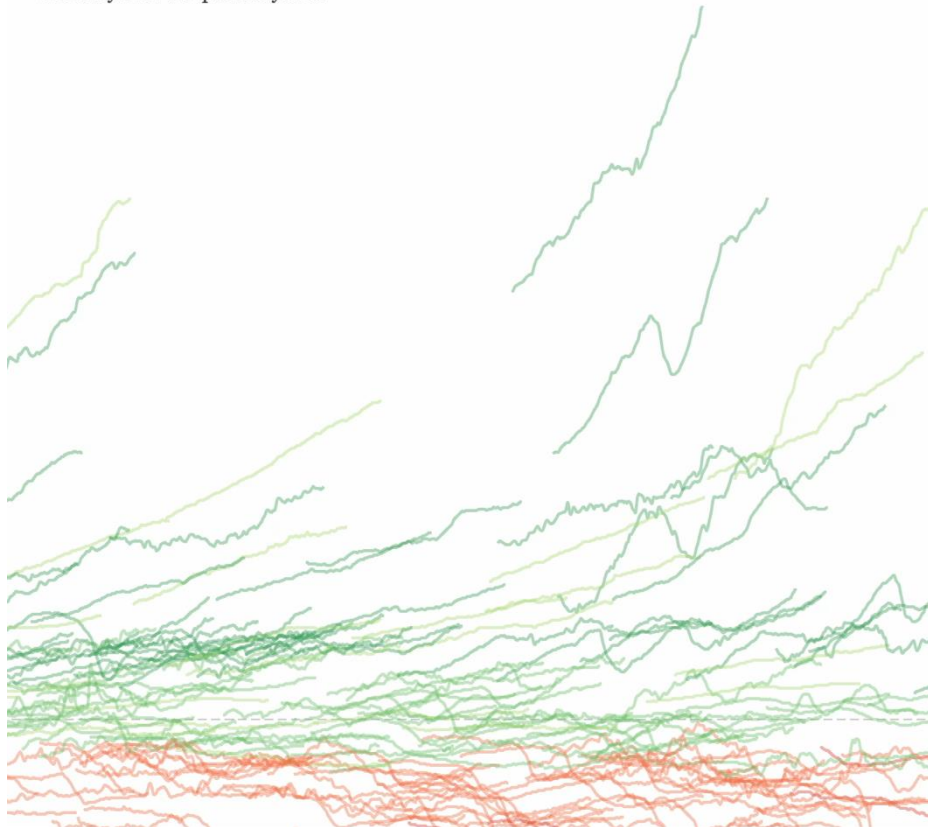
Visited links (purple)

HindSight

M. Feng, C. Chen, E. Peck, L. Harrison

Analyzing trends in the economy

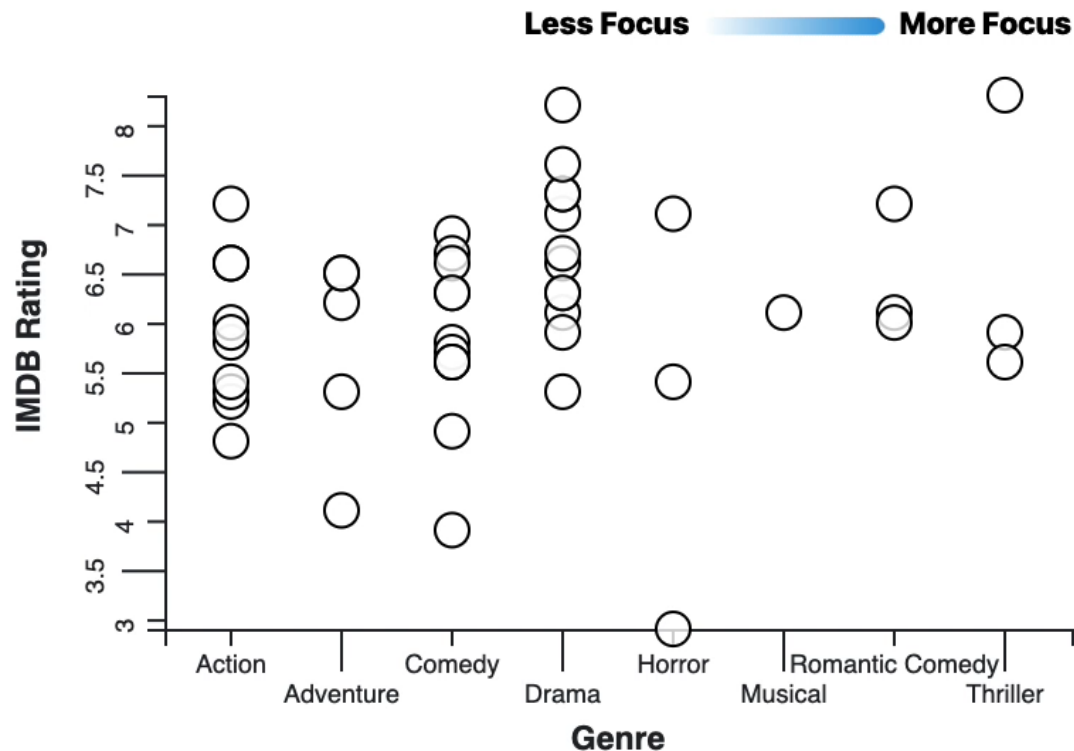
Five years since the end of the recent Recession, the economy has finally regained the nine million jobs it lost. Each line below shows how the number of jobs has changed for a particular industry over the past 10 years.



Visited lines become
darker and **thicker** to encourage
exploration and engagement

Lumos

A. Narechania, A. Coscia, E. Wall, A. Endert



Visited points are colored in **shades of blue**
(more visits = darker shades)

to increase users' awareness
of their Analytic Behavior

Lumos

A. Narechania, A. Coscia, E. Wall, A. Endert

Visited points are colored in
shades of blue
(more visits = darker shades)

to increase users' awareness
of their Analytic Behavior

*“I confused the **blue** colors
with an [attribute] encoding”*

Confusing

P02_A

*[important to make
Lumos color-blind safe]*

P12_A

Inaccessible

Lumos

A. Narechania, A. Coscia, E. Wall, A. Endert

*“I confused the **blue** colors
with an [attribute] encoding”*

Confusing

P02_A

*[What about **RECENCY**
of interactions?]*

Limited metrics

*[important to make
Lumos color-blind safe]*

P12_A

Inaccessible

Problems to Solve...

*“I confused the **blue** colors with an [attribute] encoding”*

Confusing

[important to make Lumos color-blind safe]

Inaccessible

*[What about **RECENCY** of interactions?]*

Limited metrics

*[Limited opportunities for **SELF-REFLECTION**]*

Missing use-case

Utilizing **Provenance as an Attribute** for Analysis

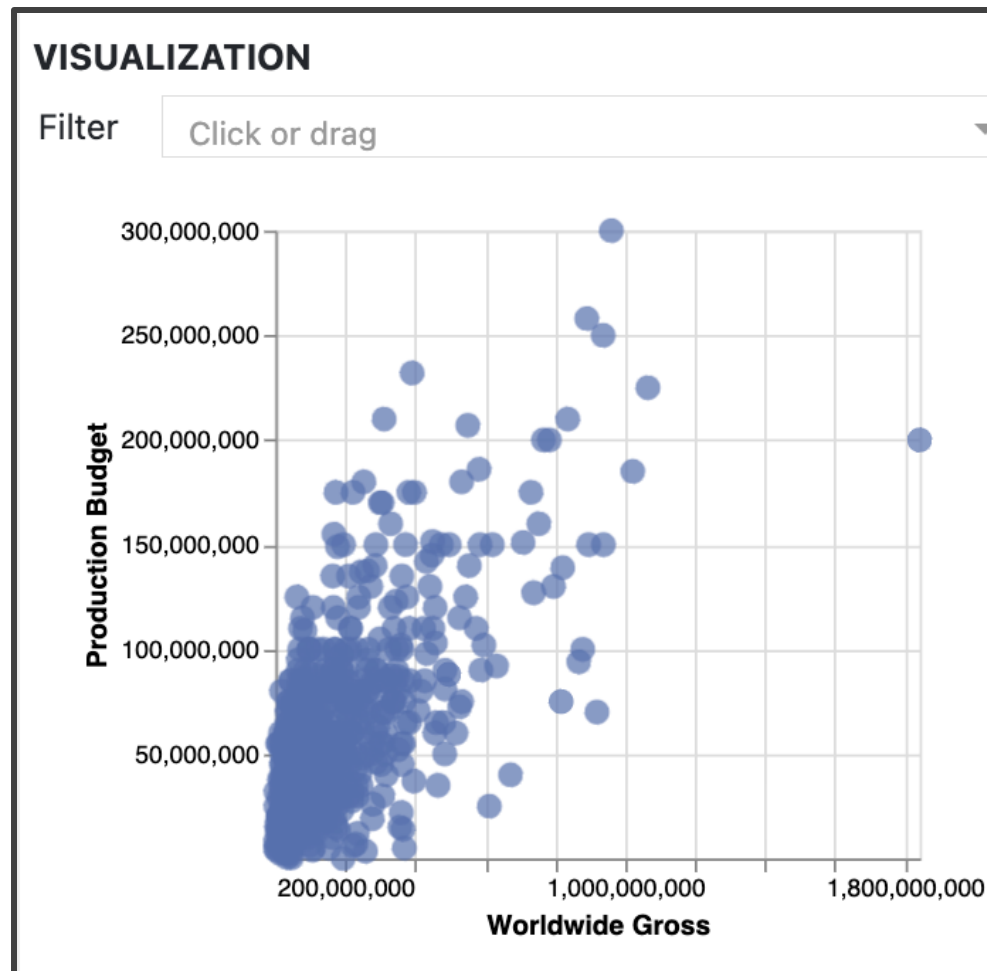
We model two provenance attributes that track the

- **frequency** (how frequent?)
- **recency** (how recent?)

of user interactions with data during analysis.



Consider this Scatterplot Visualization (dataset of movies)



ENCODINGS

i x

Worldwide Gross

i y

Production Budget

Frequency

Tracks the *frequency* of interactions with points.

$$f_x = \frac{n_x}{\max_{i=1}^N n_i}$$

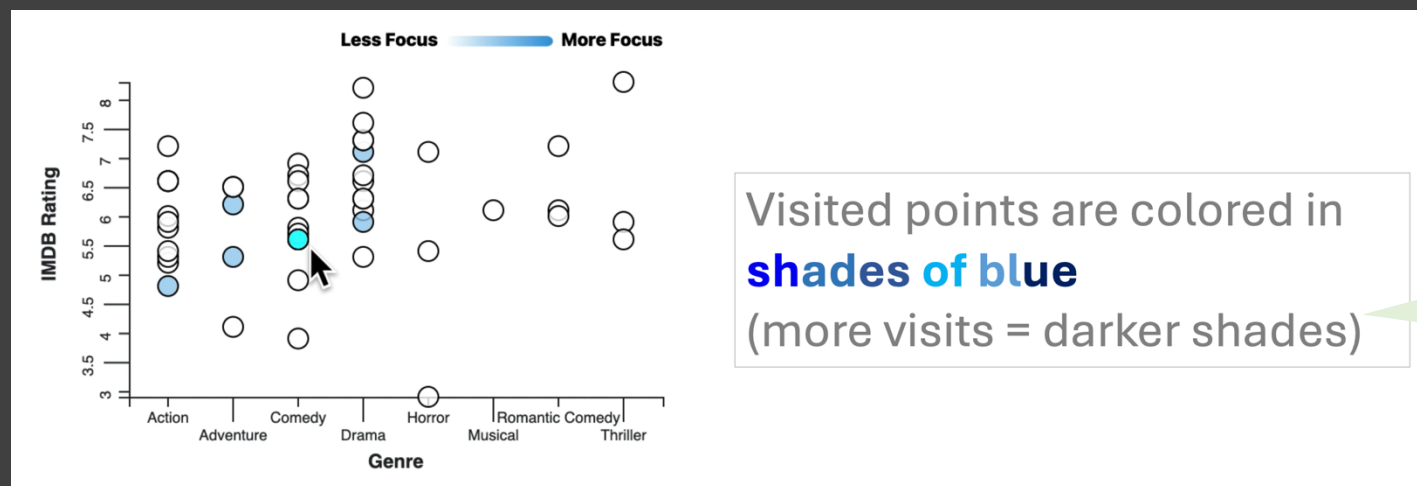
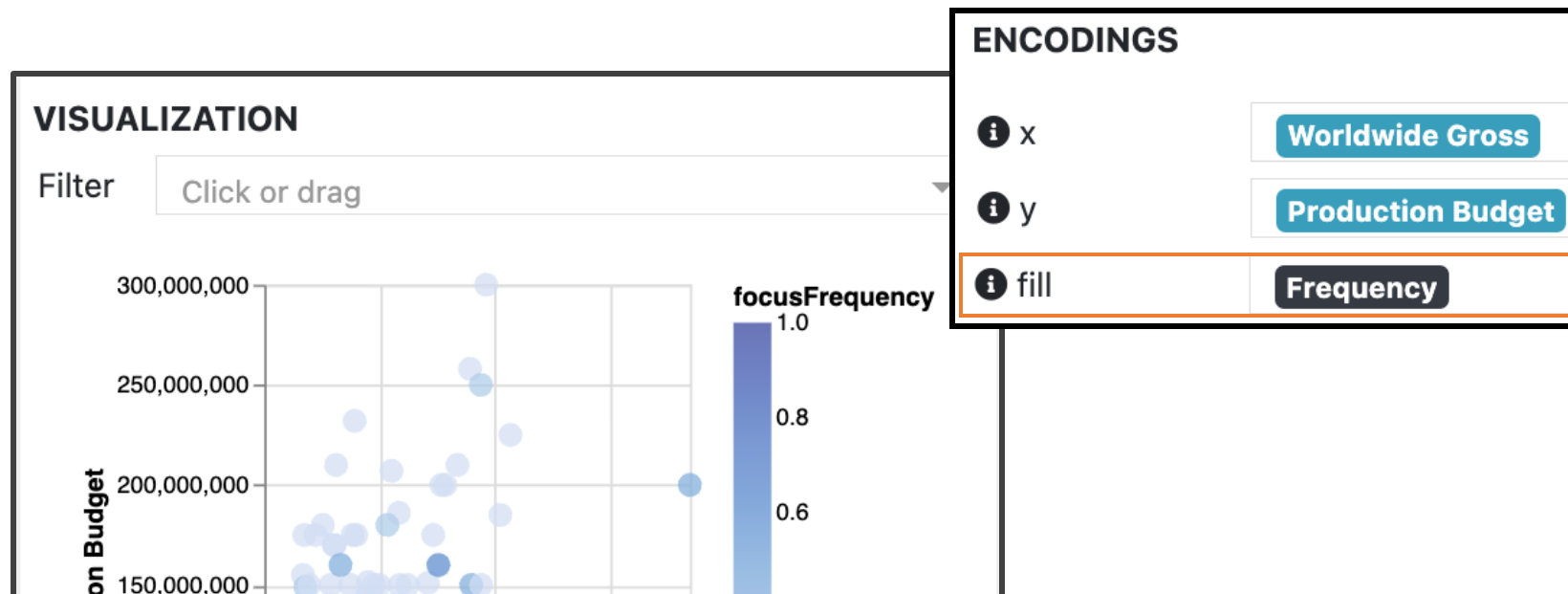
Values $\in [0, 1]$

1 = most frequent.

0 = not interacted.



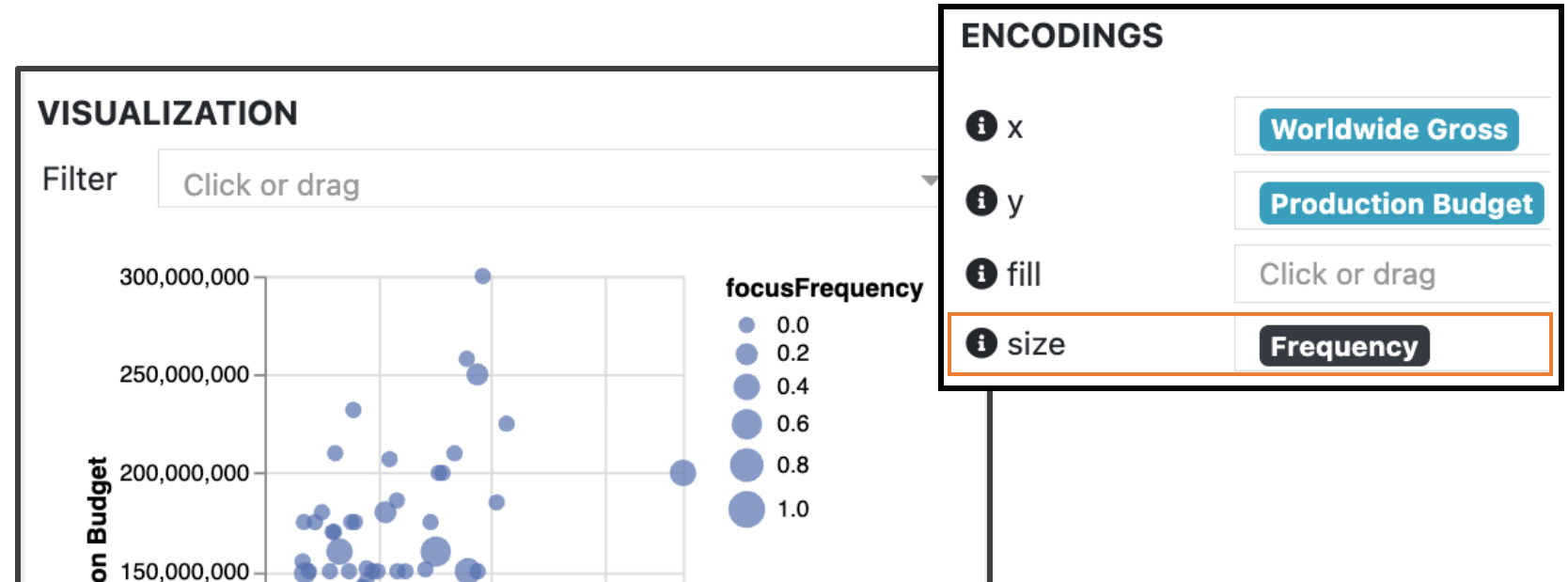
Map **frequency** of my focus to “fill” (color)



We just recreated **Lumos!!!**



If not “fill” (color), how about “Size”?

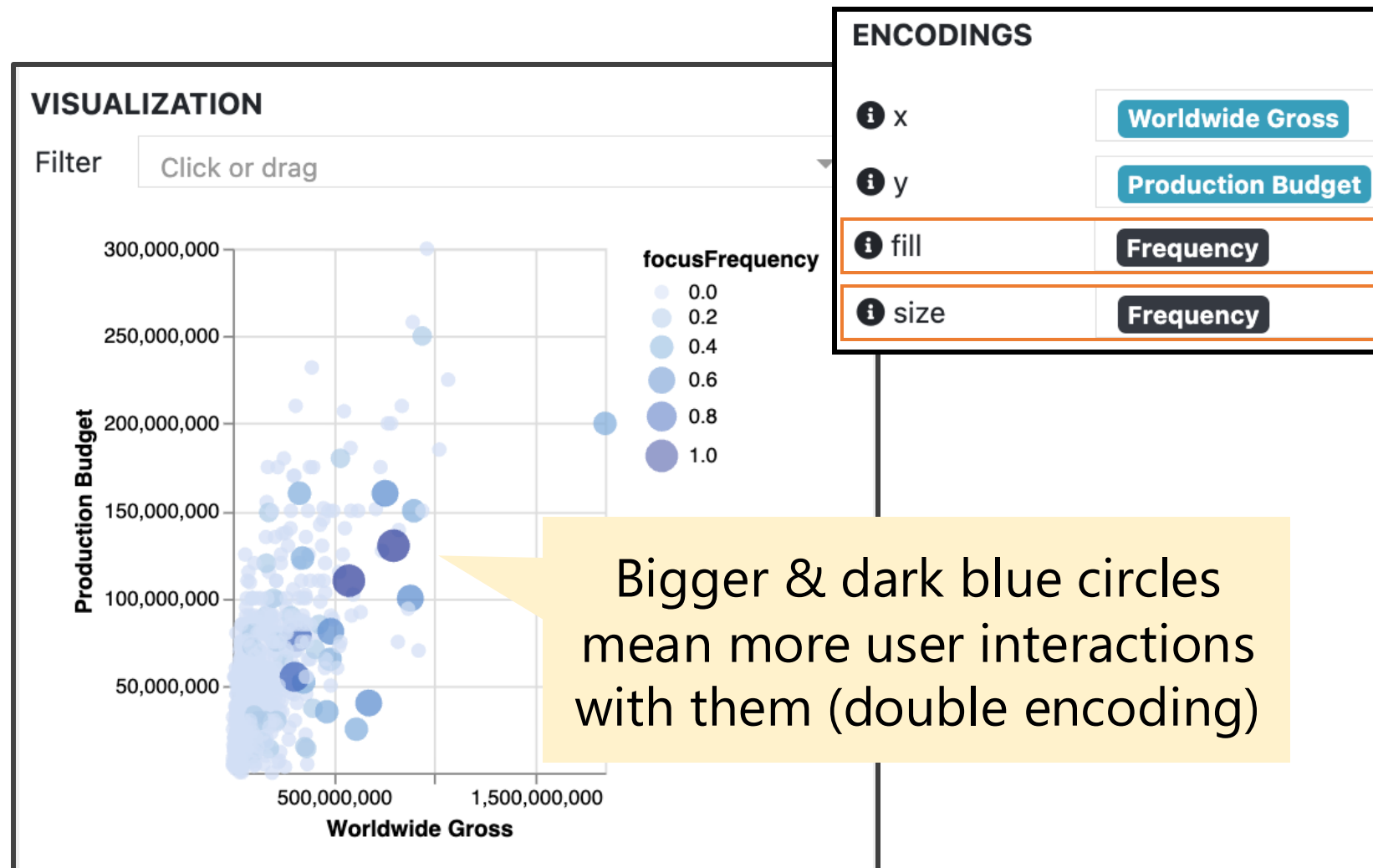


*“I confused the **blue** colors with an [attribute] encoding”*

[important to make Lumos color-blind safe]



How about both “fill” (color) and “Size”?





Not just Datapoints, we also track **Attributes**

DATA ATTRIBUTES

Sort

Filter

Expand/Collapse All

▼ ⓘ (A) id	
▼ ⓘ (A) Title	
▼ ⓘ (#) Worldwide Gross	
▼ ⓘ (#) Production Budget	
▼ ⓘ (#) Release Year	
▼ ⓘ (A) Content Rating	

ENCODINGS

ⓘ x	Worldwide Gross
ⓘ y	Production Budget
ⓘ fill	Frequency
ⓘ size	Frequency

Bigger & dark blue circles mean these attributes have been interacted with a lot.



How about ANY encoding? X or Column

point
x or column



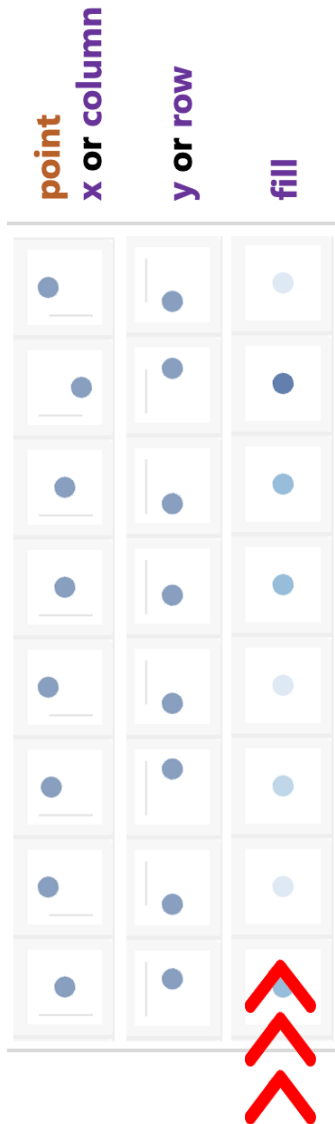


How about ANY encoding? Y or Row





How about ANY encoding? **Fill**





How about ANY encoding? **Fill Opacity**

point x or column	y or row	fill	fillOpacity



How about ANY encoding? Stroke

point x or column	y or row	fill	fillOpacity	stroke



How about ANY encoding? **Stroke Opacity**

point x or column	y or row	fill	fillOpacity	stroke	strokeOpacity



How about ANY encoding?

Size, Shape, Tooltip, Annotation

point x or column	y or row	fill	fillOpacity	stroke	strokeOpacity	strokeWidth	size	shape	tooltip	annotation
										0
										1
										0.5
										0.5
										0
										0.25
										0
										0.5

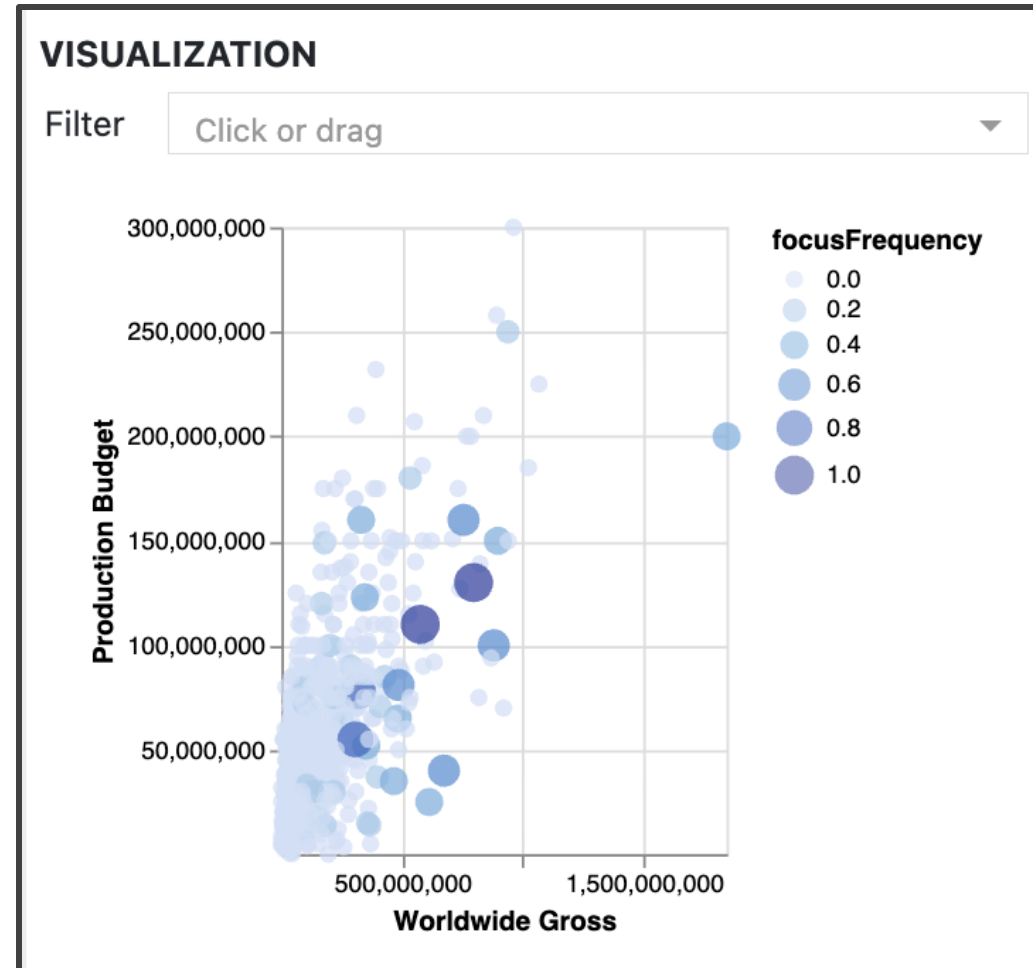


How about ANY **mark**? Point, Text, Bar...

point x or column	y or row	fill	fillOpacity	stroke	strokeOpacity	strokeWidth	size	shape	tooltip	annotation	text text	bar fill
											0	
											1	
											0.5	
											0.5	
											0	
											0.25	
											0	
											0.5	

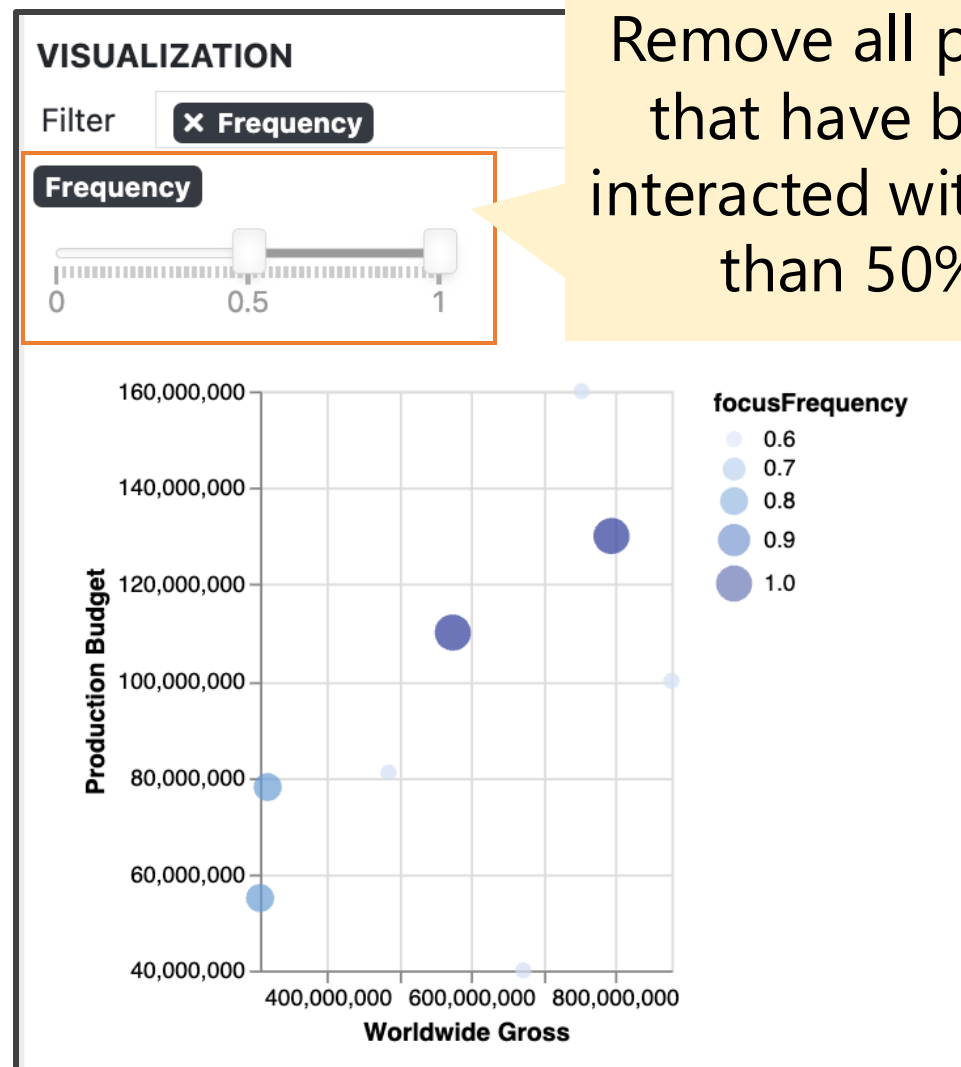
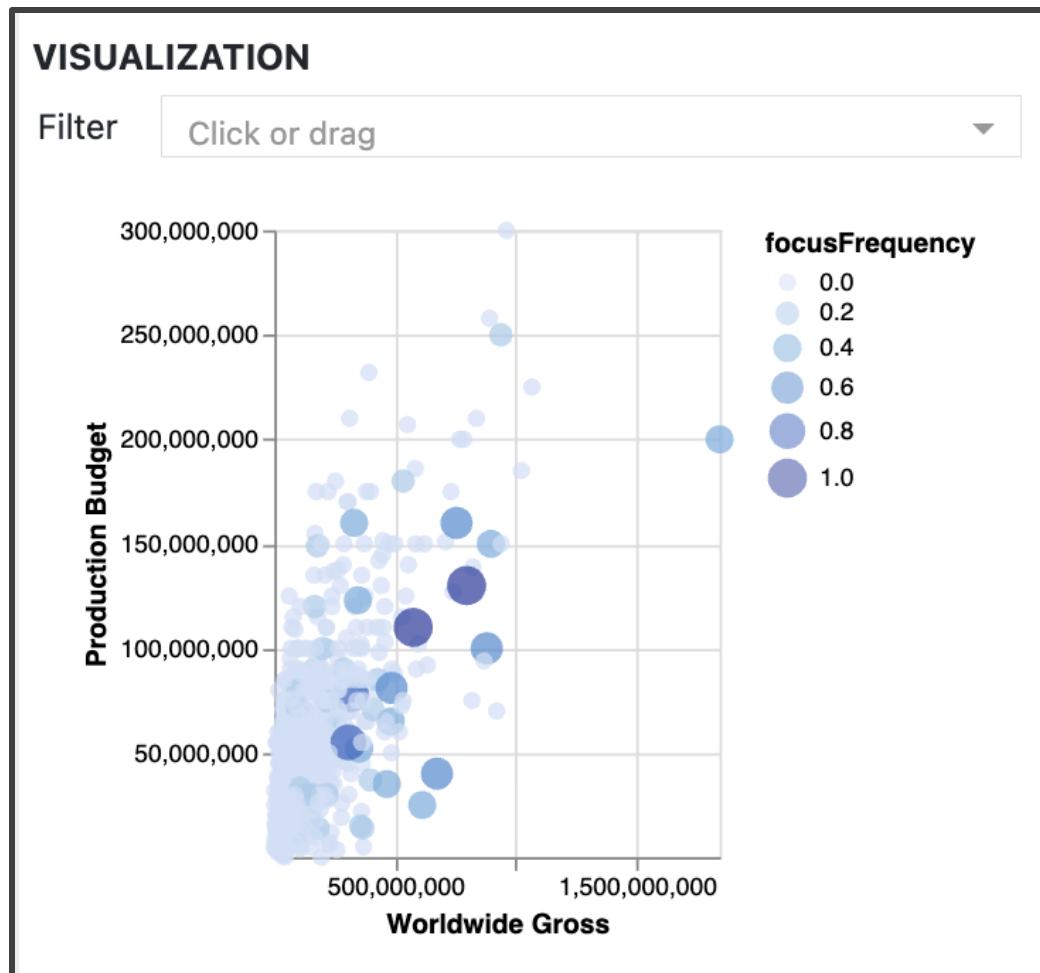


How about “FILTER”ing by **Frequency**?





How about "FILTER"ing by **Frequency**?



Remove all points that have been interacted with less than 50%.



What about Data Transformations (Filter, Sort) ?

point x or column	y or row	fill	fillOpacity	stroke	strokeOpacity	strokeWidth	size	shape	tooltip	annotation	text text	bar fill
											0	
											1	
											0.5	
											0.5	
											0	
											0.25	
											0	
											0.5	

0.5

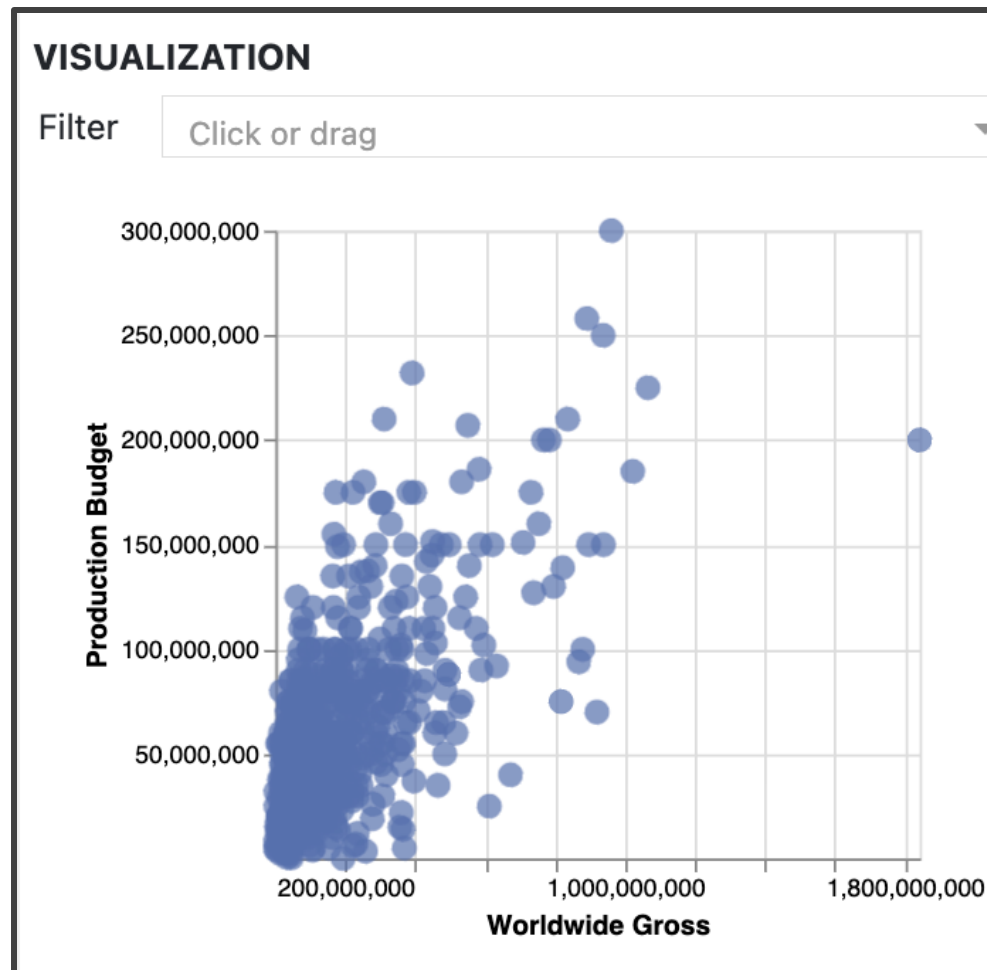


point x, sort (desc)	x, filter (val > 0.5)	-x (reverse), sort





Frequency aside, how about Recency?



ENCODINGS

i x

Worldwide Gross

i y

Production Budget

Recency

Tracks the *recency* of my interactions with points.

$$r_x = \frac{\text{rank}_N(\max(t_x))}{\sum_{i=1}^N n_i}$$

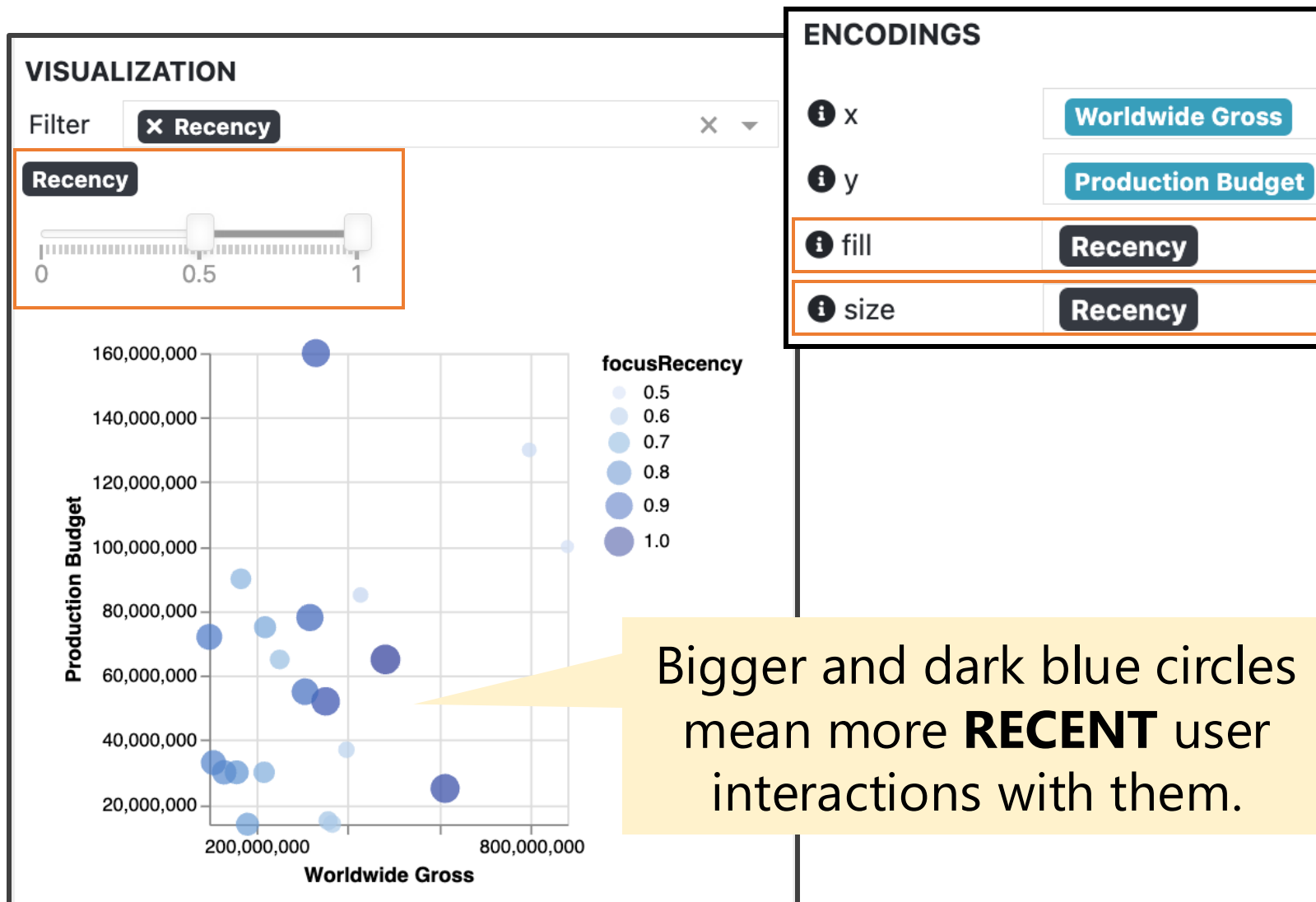
Values $\in [0, 1]$

1 = most recent.

0 = not interacted.



Frequency aside, how about **Recency**?





DATA TRANSFORMATIONS



ProvenanceLens

Main Task

09:03:02 PM, 29 Aug 24 (GMT-4)

0h 2m 49s (0h 2m 49s)

Prev

End Task

DATA ATTRIBUTES

Sort Click or drag

Filter Click or drag

Expand/Collapse All

(A) id

(A) Title

(#) Worldwide Gross

(#) Production Budget

(#) Release Year

(A) Content Rating

(#) Running Time

(A) Genre

(#) Rotten Tomatoes Rating

(#) IMDB Rating

MARK

ENCODINGS

x Worldwide Gross

y Production Budget

fill Frequency

fillOpacity Click or drag

stroke Click or drag

strokeWidth Click or drag

strokeOpacity Click or drag

size Click or drag

shape Click or drag

row Click or drag

column Click or drag

tooltip Title

text Click or drag

annotation Click or drag

PROVENANCE ATTRIBUTES

(#) Frequency

(#) Recency

Frequency: 0 = not interacted at all, 1 = most frequently interacted

Recency: 0 = not interacted at all, 1 = most recently interacted

MARK

point



ENCODINGS

Swap xy



x

Worldwide Gross



Agg



y

Production Budget



Agg



fill

Frequency



Agg



fillOpacity

Click or drag



Agg



stroke

Click or drag



Agg



strokeWidth

Click or drag



Agg



strokeOpacity

Click or drag



Agg



size

Click or drag



Agg



shape

Click or drag



Agg



row

Click or drag



Agg



column

Click or drag



Agg



tooltip

Title



Agg



text

Click or drag



Agg



annotation

Click or drag



Agg

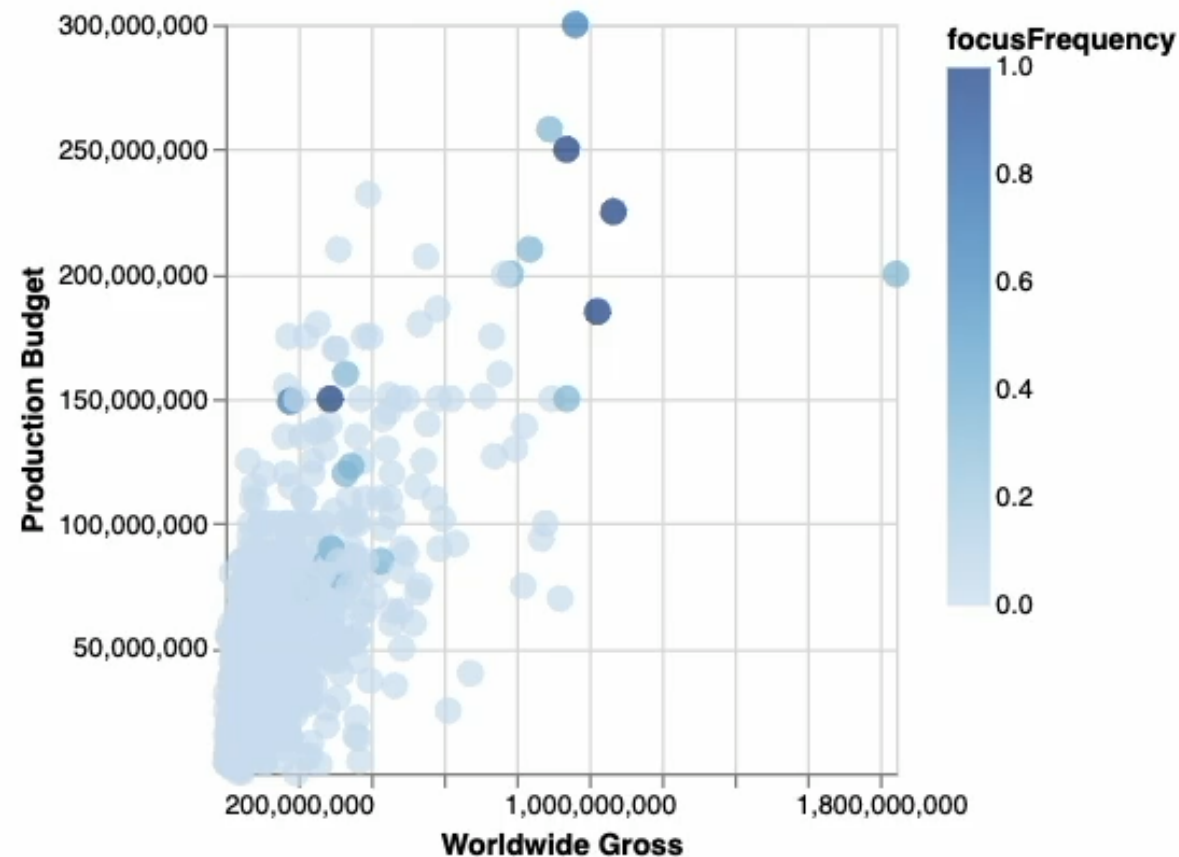


VISUALIZATION



Filter

Click or drag



DATA RECORDS

MARK

point



ENCODINGS

Swap xy



x

Click or drag



Agg



y

Production Budget



Agg



fill

Click or drag



Agg



fillOpacity

Click or drag



Agg



stroke

Click or drag



Agg



strokeWidth

Click or drag



Agg



strokeOpacity

Click or drag



Agg



size

Click or drag



Agg



shape

Click or drag



Agg



row

Click or drag



Agg



column

Click or drag



Agg



tooltip

Title



Agg



text

Click or drag



Agg



annotation

Click or drag



Agg



VISUALIZATION



Filter

Click or drag



300,000,000

250,000,000

200,000,000

150,000,000

100,000,000

50,000,000

Production Budget

DATA RECORDS

DATA ATTRIBUTES



Sort Click or drag

Filter Click or drag

Expand/Collapse All

- (A) id
- (A) Title
- (#) Worldwide Gross
- (#) Production Budget
- (#) Release Year
- (A) Content Rating
- (#) Running Time
- (A) Genre
- (#) Rotten Tomatoes Rating

MARK

point

ENCODINGS

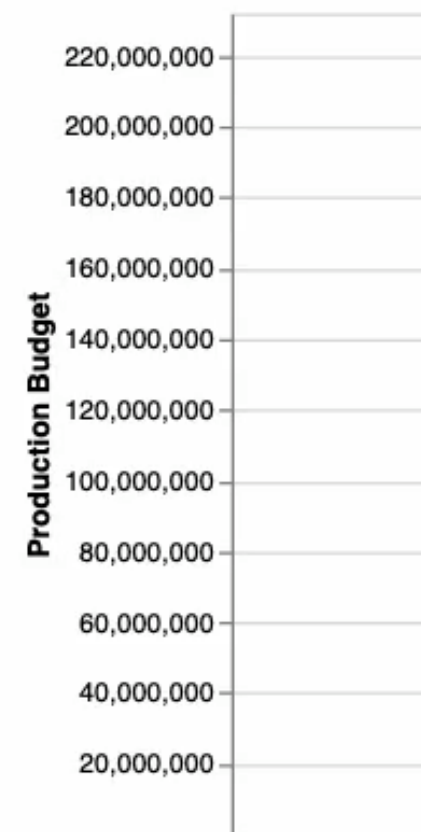
Swap xy

x	Click or drag	Agg
y	Production Budget	Agg
fill	Click or drag	Agg
fillOpacity	Click or drag	Agg
stroke	Click or drag	Agg
strokeWidth	Click or drag	Agg
strokeOpacity	Click or drag	Agg
size	Click or drag	Agg
shape	Click or drag	Agg
row	Click or drag	Agg
column	Click or drag	Agg
tooltip	Title	Agg
text	Click or drag	Agg
annotation	Click or drag	Agg

VISUALIZATION

Filter **Frequency**

Frequency



DATA RECORDS



Evaluation: **Design Probe**

Participants

16 (e.g., university graduate students)

Duration

75-90 minutes

Study Objective

Pick 10 movies to take with you on a family vacation.



Evaluation: **Design Probe**

Catch 1: Family members have preferences!

- 1) Dad: *“I like thrillers and comedies.”*
- 2) Uncle: *“Let’s watch a heartfelt drama.”*
- 3) Cousin: *“I want to re-watch a Harry Potter movie.”*
- 4) Grandpa: *“I would love to watch a 90s’ action film.”*
- 5) Aunt: *“I want to watch a hidden gem: a highly-rated movie that didn’t do well commercially.”*

Catch 2: *Your mom has already picked 5 movies but then she got called to do something else...*



Evaluation: Design Probe

Phase 1: Answer questions about **YOUR MOM's** provenance!

Task	Que	Description
T1 (Review)	Q1	Review YOUR MOM's analysis and write three insights.
T2 (Recall)	Q1	Select the movie characteristic(s) that YOUR MOM interacted with the MOST.
	Q2	Which was the last (i.e. most recent) movie YOUR MOM interacted with?
	Q3	Did YOUR MOM ever interact with the movie 'Titanic'?
	Q4	Did YOUR MOM interact with at least one movie from all 'Content Rating's?
	Q5	Did YOUR MOM interact with all attributes at least once?
T3 (Visualize)	Q1	What was the distribution of YOUR MOM's focus across different movie 'Genre's?
	Q2	How did YOUR MOM's focus on 'Drama' Movies evolve over time?
	Q3	Which were YOUR MOM's most FREQUENTLY interacted movies? Try to show five.



Evaluation: Design Probe

Phase 2: Answer questions about **YOUR OWN** provenance!

T4 (Analyze)	Q1	Select the five remaining movies.
--------------	----	-----------------------------------

T5 (Recall)	Q1	Select three movie characteristic(s) YOU most RECENTLY interact with.
	Q2	Which was the first (i.e. earliest) movie YOU interacted with?
	Q3	Did YOU interact with the movie 'Pearl Harbor'?
	Q4	Which 'Content Rating' category did YOU interact with the LEAST or NONE AT ALL?
	Q5	Did YOU interact with at least half of the movie characteristics available?

T6 (Visualize)	Q1	How similar were YOUR interaction patterns for 'Comedy' and 'Thriller' movies?
	Q2	Which were YOUR most RECENTLY interacted movies? Try to show THREE movies.
	Q3	Given an opportunity, which (kinds of) movies would YOU like to go back and interact with?



Findings: **General Feedback** (SUS=80.94)

“Tools like Tableau [could] really overlay communities’ or experts’ interaction stats such as frequency and recency in their system.

It is one of those features that only adds value and is not necessarily a hindrance.”



Findings: **Task Performance**



High accuracy (99.21% for 16 questions)



High confidence ($\mu=6.48$, s.d.=0.95
on a 7 pt Likert scale (Low to High))

...and also expressed several instances of being:



Surprised ($\mu=2.54$, s.d.=1.62)



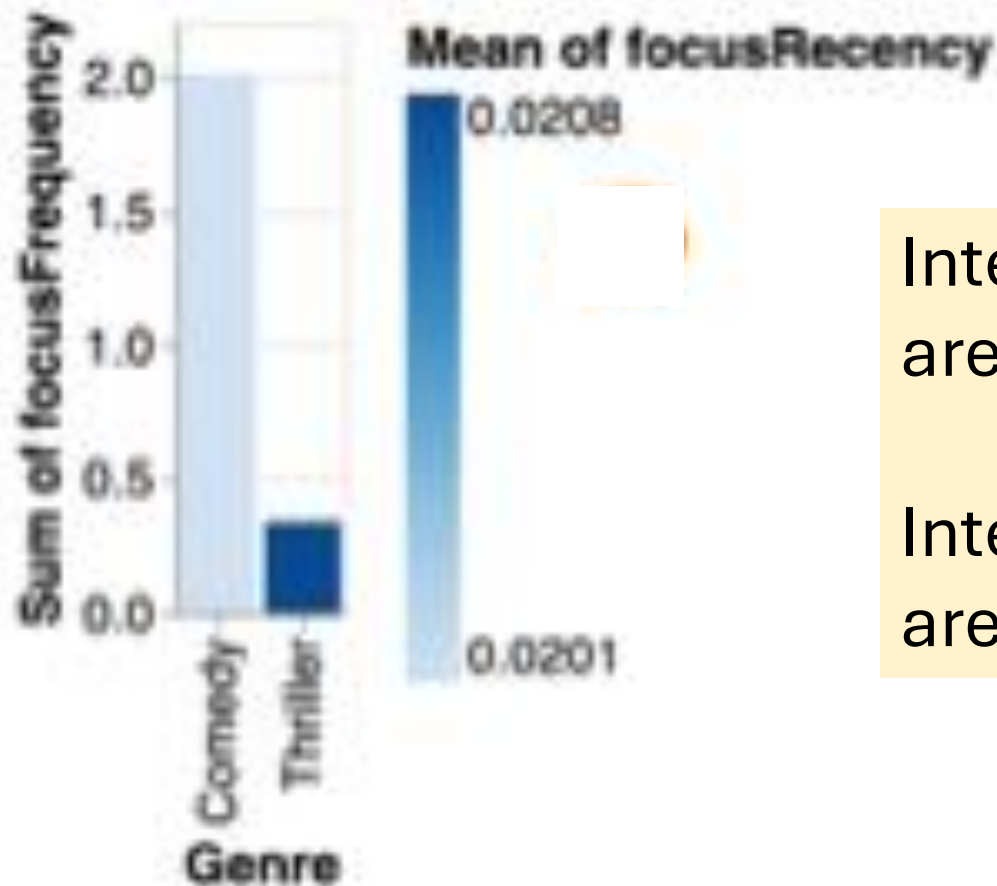
Findings: **Provenance Attributes Use-Cases**

- **Fun and feel good** about making progress in the task.
- **Answer questions** when unable to recall analysis behavior.
- **Verify and build confidence.**
- **Increase awareness.**
- **Improve exploration coverage** and uncover new insights.



Variety of User Strategies to Answer Questions

“How similar were your interaction patterns for “Comedy” and “Thriller” movies?”



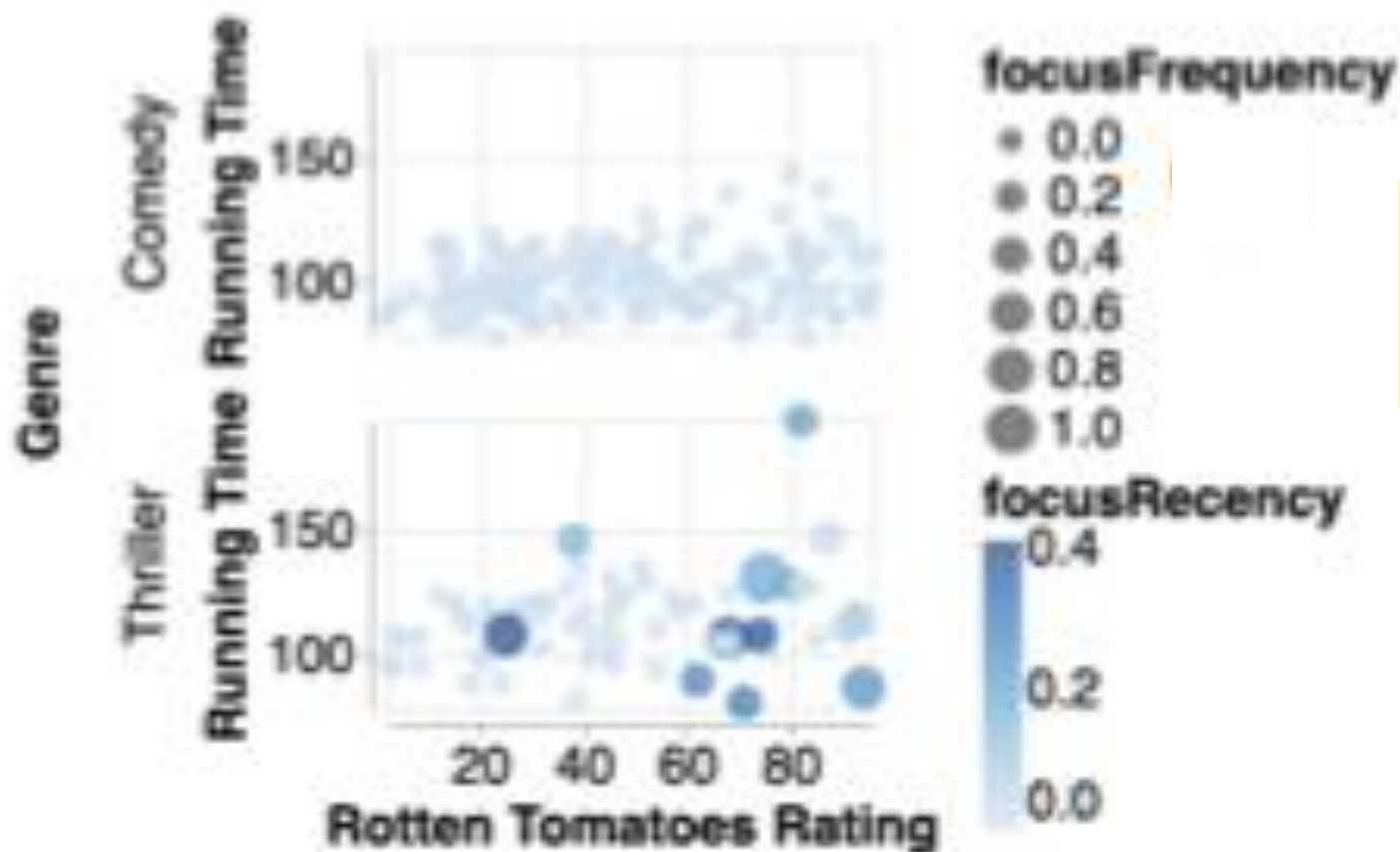
Interactions with **Thriller** movies are **MORE RECENT** (dark blue).

Interactions with **Comedy** movies are **MORE FREQUENT** (taller bar).



Variety of User Strategies to Answer Questions

“How similar were your interaction patterns for “Comedy” and “Thriller” movies?”



Did **NOT** interact with
Comedy movies!

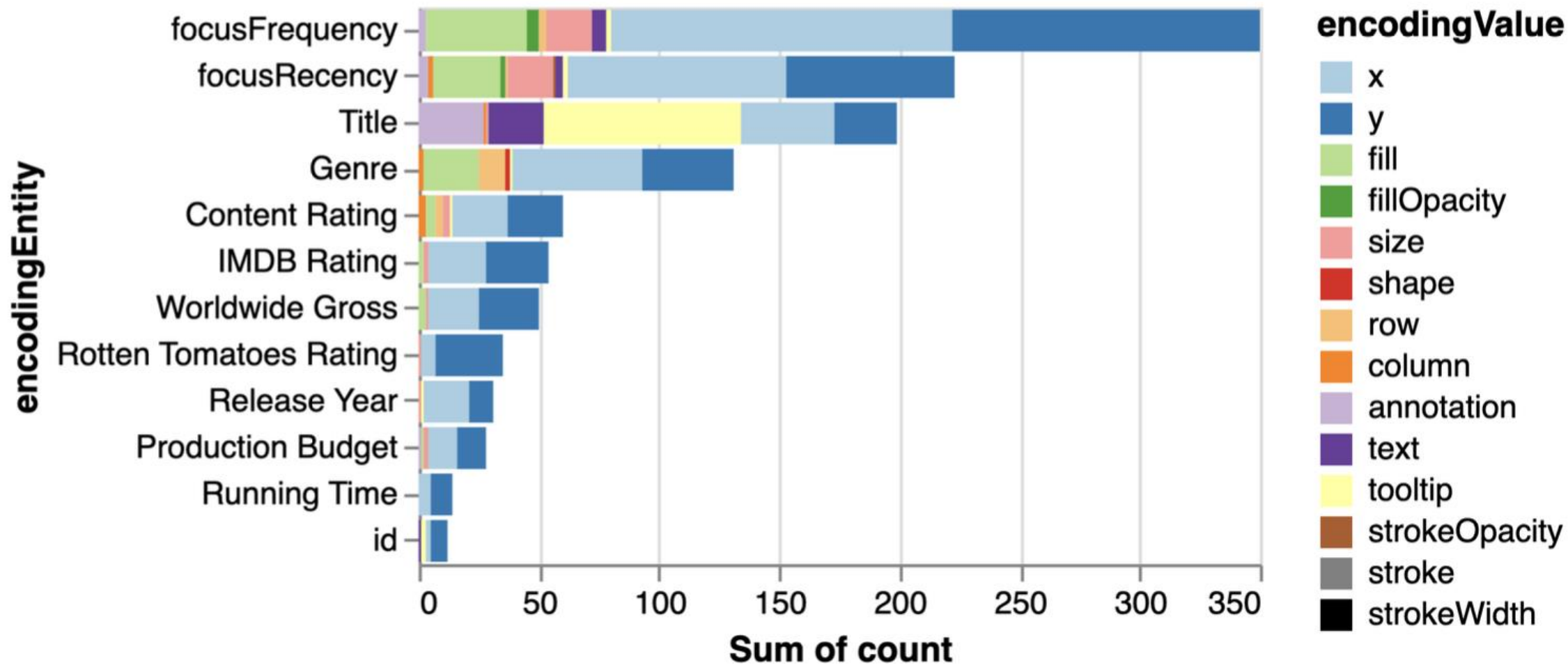


Findings: **Moments of Surprise**

- System's answer did not match participants' recollection.
- System's answer did not match original analytic intention.
- System logged interactions that were perhaps accidental (even though we had a lot of safeguards)



Findings: **Attribute – Visual Encoding Preferences**



Problems to Solve...

*“I confused the **blue** colors with an [attribute] encoding”*

Confusing

[important to make Lumos color-blind safe]

Inaccessible

*[What about **RECENCY** of interactions?]*

Limited metrics

*[Limited opportunities for **SELF-REFLECTION**]*

Missing use-case

SOLVED

Takeaway – Make provenance a core aspect of analysis

- **Integrate Provenance-Tracking & Visual Data Analysis**
(Users used data attributes together with provenance attributes)
- **Foster Provenance-driven (not just Data-driven) Analysis**
(Users used provenance attributes to track their analytic progress)

Limitations

- Did not have a solid baseline.
- Single 90-minute session may not fully capture long-term benefits of visualizing provenance.
- Equally weighting all interactions to model focus can be improved, e.g., recent ones more than the older ones.

Utilizing Provenance as an Attribute for Visual Data Analysis: A Design Probe with ProvenanceLens



Arpit
Narechania



Shunan
Guo



Eunyee
Koh



Alex
Endert



Jane
Hoffswell