Storytelling with data

02 | purpose of visuals; coordinate systems; data encodings; the grammar of graphics; practice

Scott Spencer | Columbia University

next deliverable, **homework one**

Individual Work

For learning data visualization and written narrative techniques

Group work

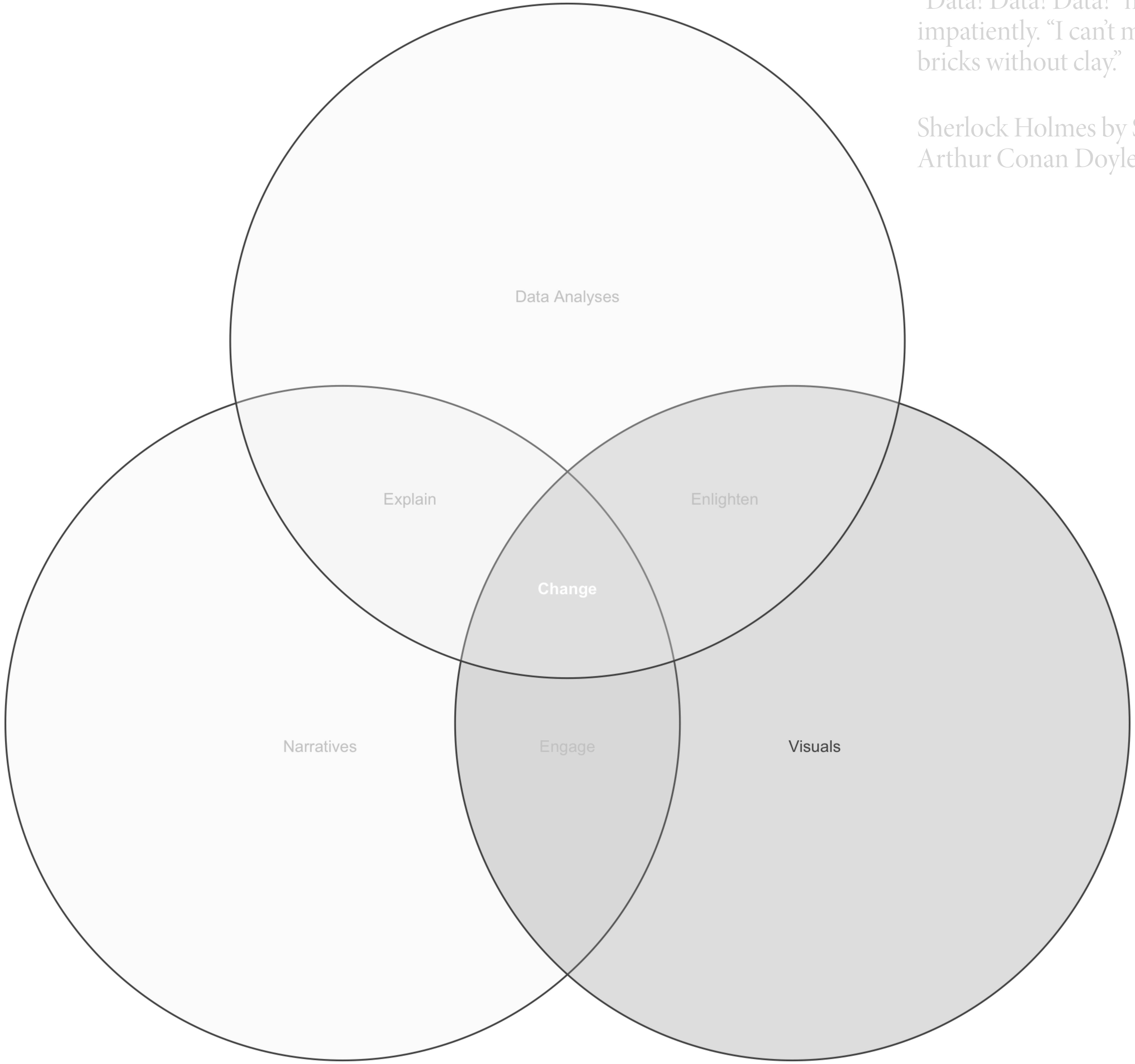
For building graphics and narrative into interactive communications

Sept 30	Oct 14	Oct 28	Nov 18	Nov 18	Dec 11	Dec 13
Homework 1 graphics	Homework 2 graphics	Homework 3 writing	Homework 4 graphics	Proposal	Interactive Communication	Multimodal communication
10%	10%	10%	10%	15%	20%	15%
Participation 10%						

course overview, **learn to drive change using data visuals and narrative**

“Data! Data! Data!” he cried impatiently. “I can’t make bricks without clay.”

Sherlock Holmes by Sir Arthur Conan Doyle, *author*



No one ever made a decision because of a number. They need a story.

Daniel Kahneman, *psychologist, behavioral economist, and author*

The greatest value of a picture is when it forces us to notice what we never expected to see.

John W Tukey, *mathematician*

Why show data graphically?

why data graphics, **one value or strength of visualization**

While text can use different types of content structures, an abstract visualization just presents relationships between data points.

Thus, a single bar, map symbol or shape does not convey information. It only becomes meaningful by its relationship with other elements in the image—in other words, it is *polysemic:* **A data graphic acquires its meaning from comparison**.

— Koponen & Hildén, *The Data Visualization Handbook*

why data graphics, **graphic of a datum — effective? Conveys meaning?**

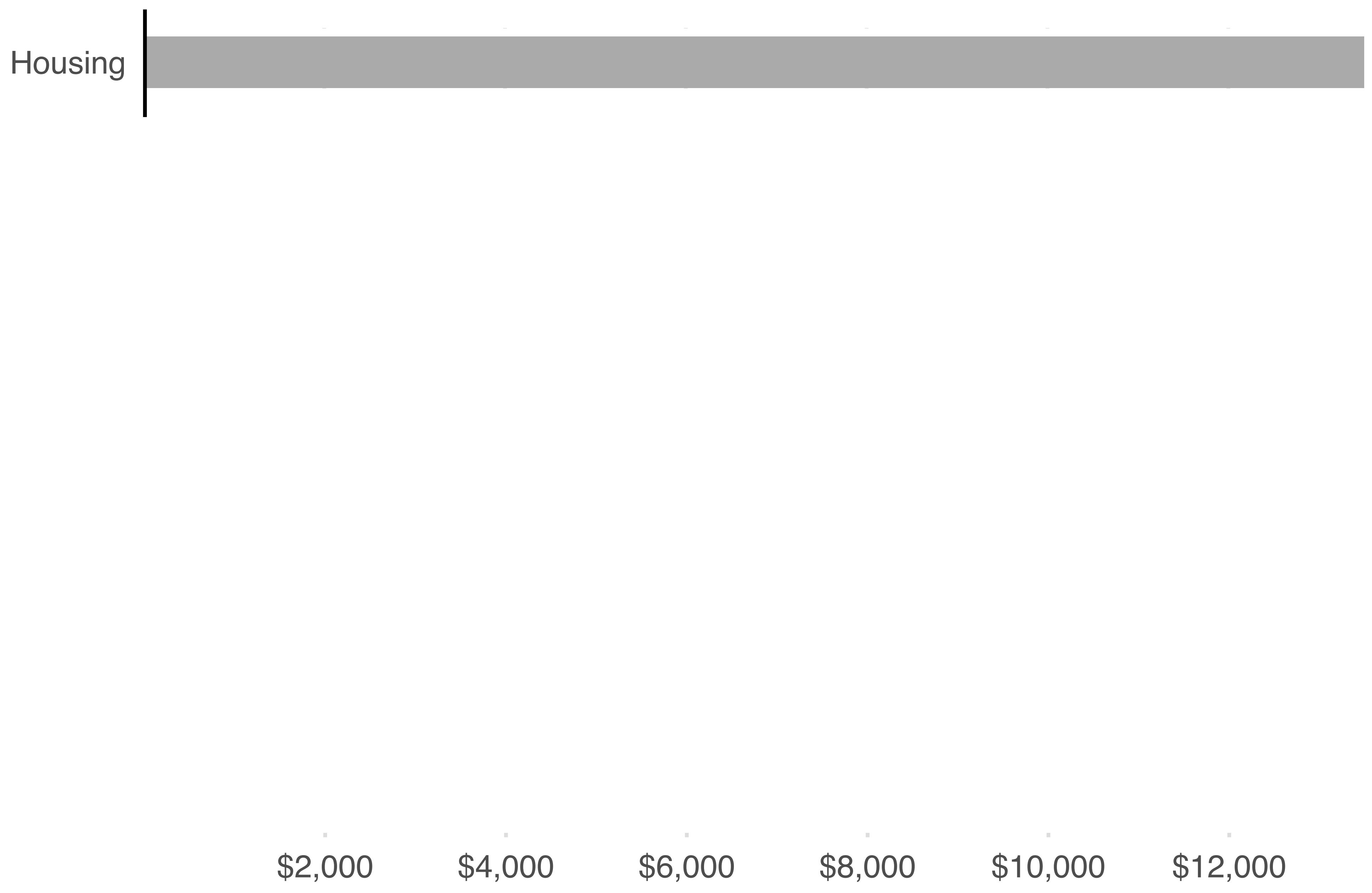


Fig. 3. Major categories of expenditures, descending dollar value, 2002 U.S. Consumer Expenditure Survey

While text can use different types of content structures, an abstract visualization just presents relationships between data points.

Thus, a single bar, map symbol or shape does not convey information. It only becomes meaningful by its relationship with other elements in the image—in other words, it is *polysemic:* **A data graphic acquires its meaning from comparison**.

— Koponen & Hildén, *The Data Visualization Handbook*

why data graphics, graphic comparing data— more effective? Conveys meaning?

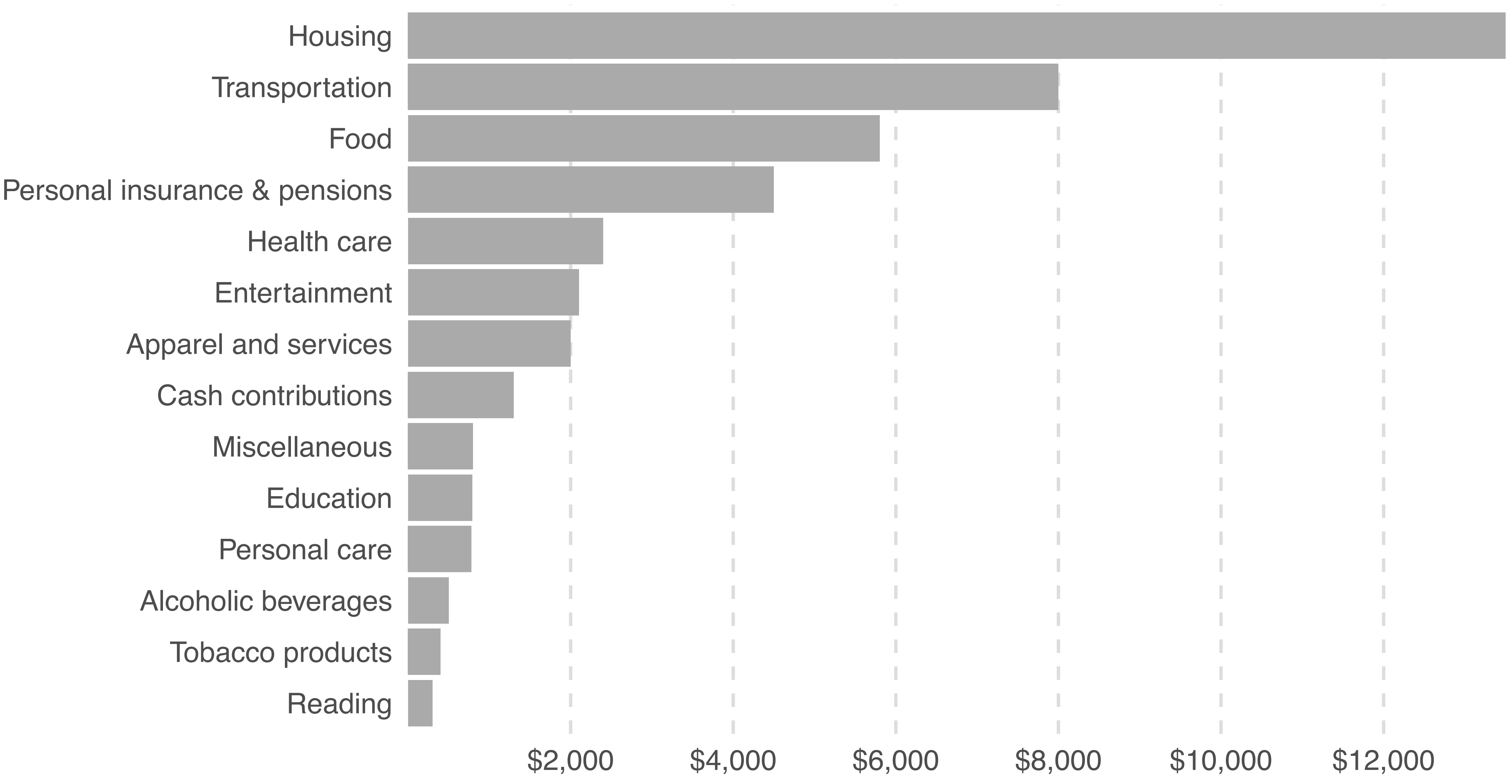


Fig. 3. Major categories of expenditures, descending dollar value, 2002 U.S. Consumer Expenditure Survey

While text can use different types of content structures, an abstract visualization just presents relationships between data points.

Thus, a single bar, map symbol or shape does not convey information. It only becomes meaningful by its relationship with other elements in the image—in other words, it is *polysemic:* **A data graphic acquires its meaning from comparison**.

— Koponen & Hildén, *The Data Visualization Handbook*

why data graphics, example data from Anscombe

1		2		3		4	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Are the 4 data sets the same?

why data graphics, example data from Anscombe

1		2		3		4	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

summary statistics

	1		2		3		4	
	x	y	x	y	x	y	x	y
mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
sd	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

Are the 4 data sets the same?

why data graphics, example data from Anscombe

1		2		3		4	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Are the 4 data sets the same?

summary statistics and regressions (y ~ 1 + x)

	1		2		3		4	
	x	y	x	y	x	y	x	y
mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
sd	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

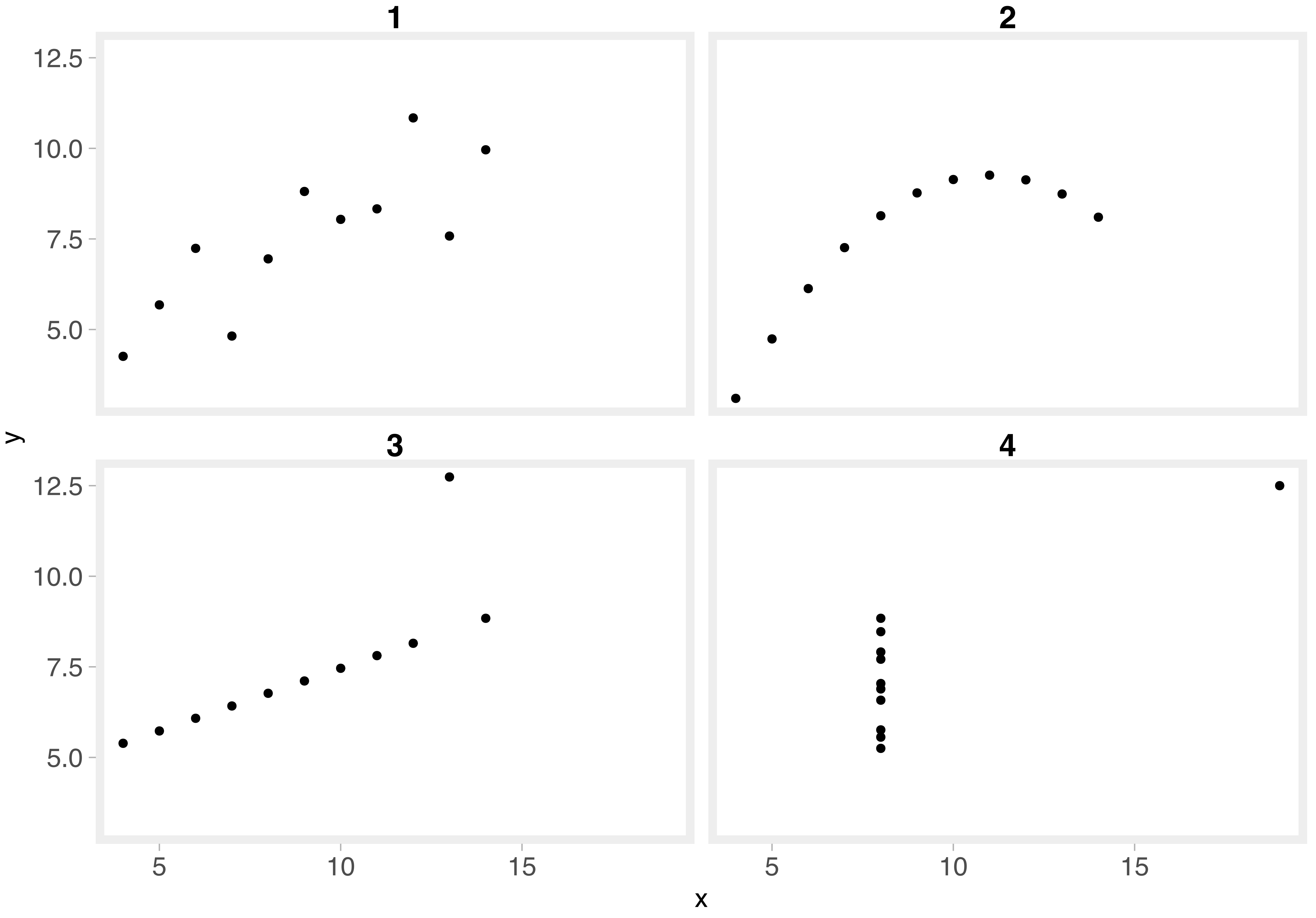
Parameter	Mean	Std Err	t-val	p-val
Dataset 1				
(Intercept)	3.000	1.125	2.667	0.026
x	0.500	0.118	4.241	0.002
Dataset 2				
(Intercept)	3.001	1.125	2.667	0.026
x	0.500	0.118	4.239	0.002
Dataset 3				
(Intercept)	3.002	1.124	2.670	0.026
x	0.500	0.118	4.239	0.002
Dataset 4				
(Intercept)	3.002	1.124	2.671	0.026
x	0.500	0.118	4.243	0.002

why data graphics, example data from Anscombe

1		2		3		4	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

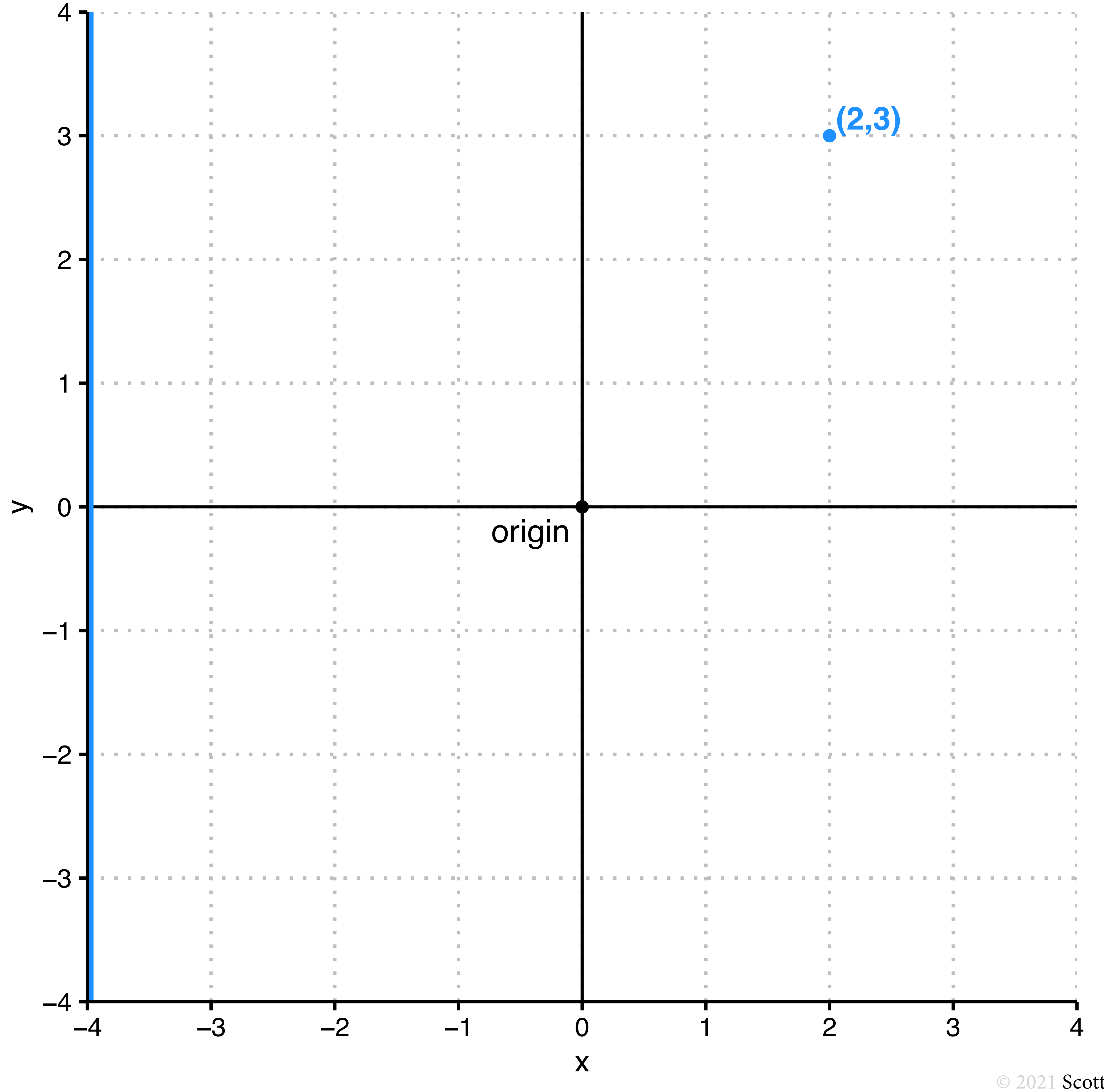
Are the 4 data sets the same?

With graphics we can use our natural ability to see patterns through visual comparison



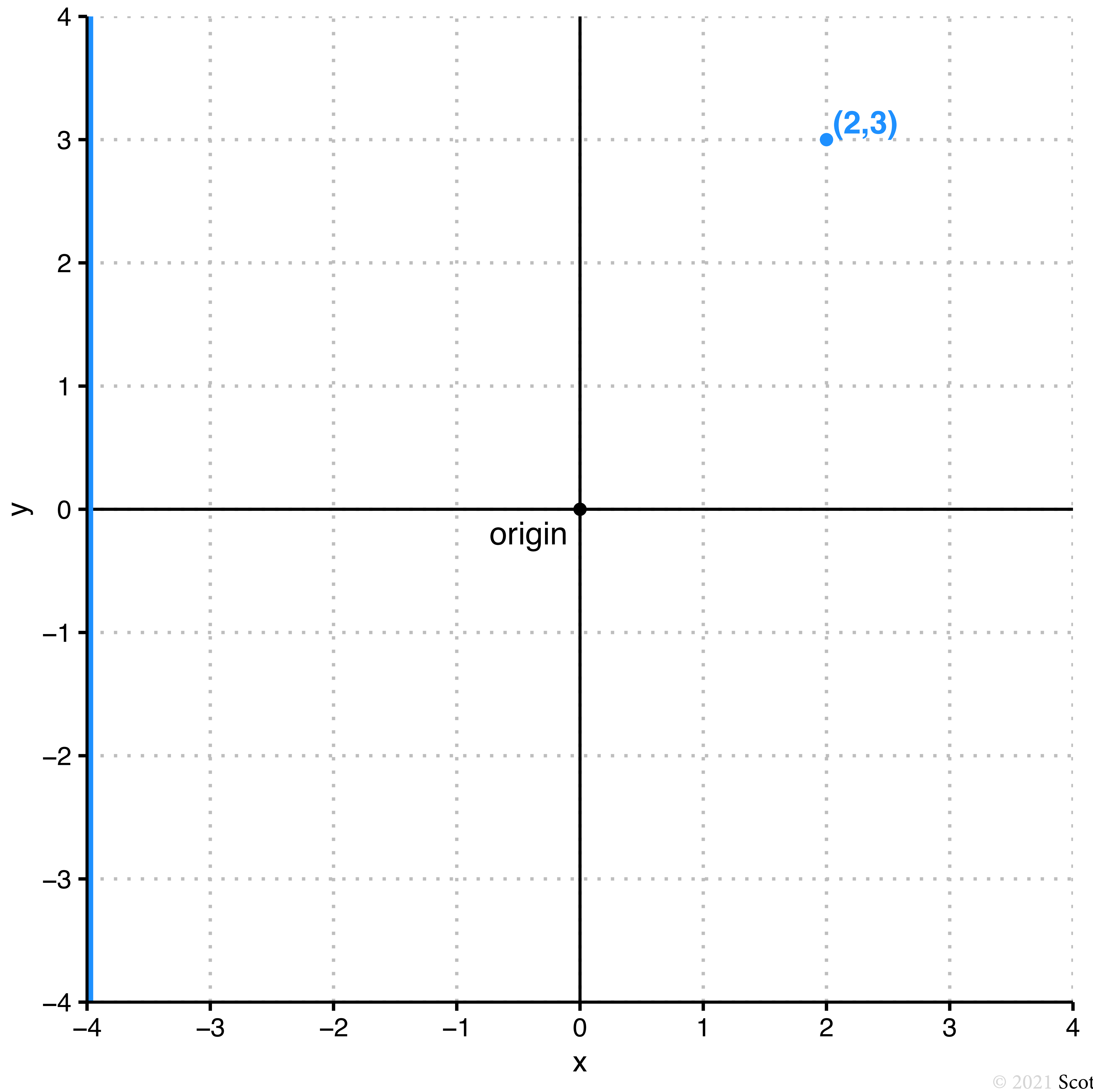
graphs — coordinate systems and scales

coordinates and scales, *two-dimensional Cartesian* coordinates — x and y axes run orthogonally to each other, and data values placed along linear axes

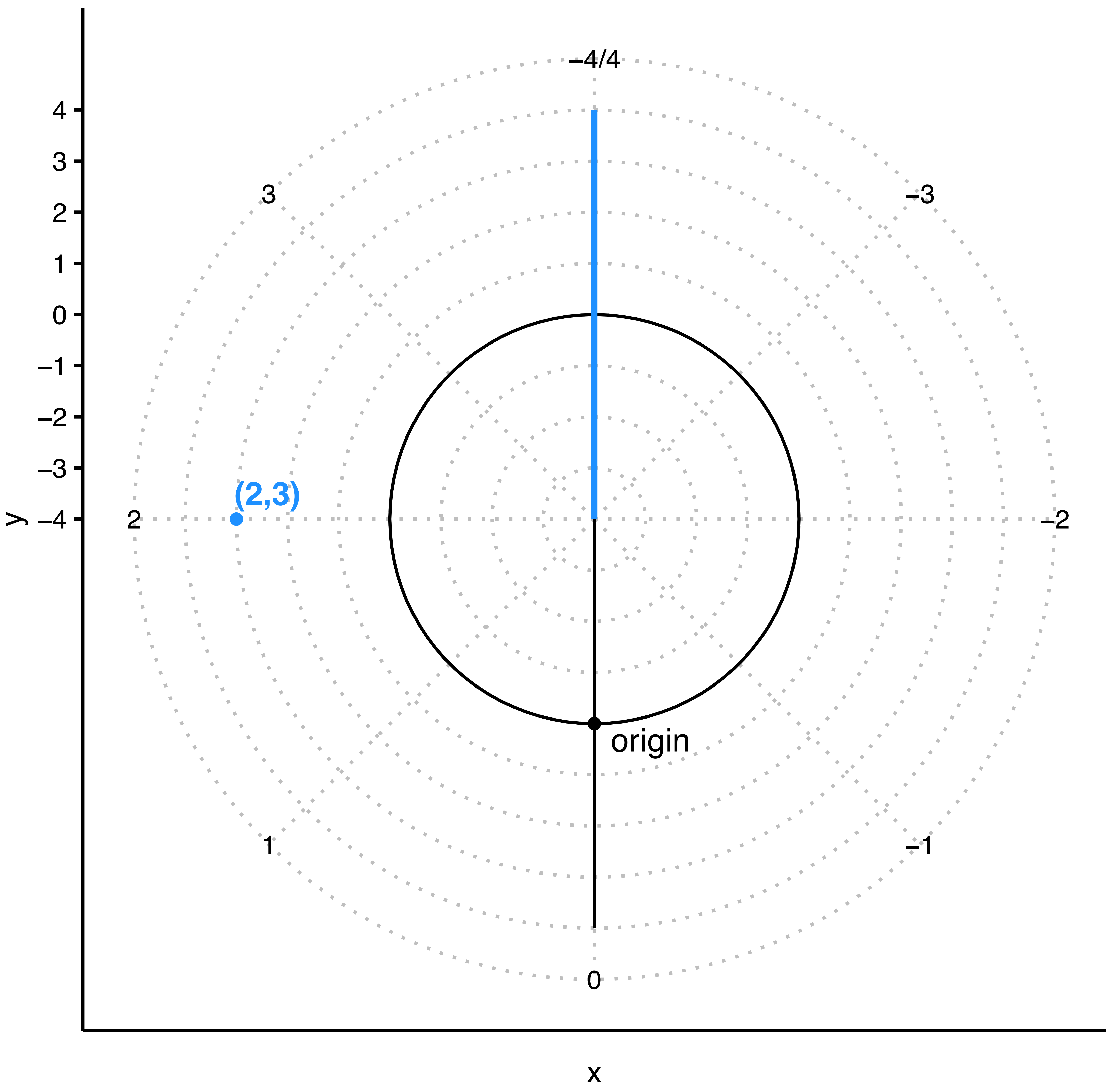


 ***other* coordinate systems are sometimes more effective in conveying information**

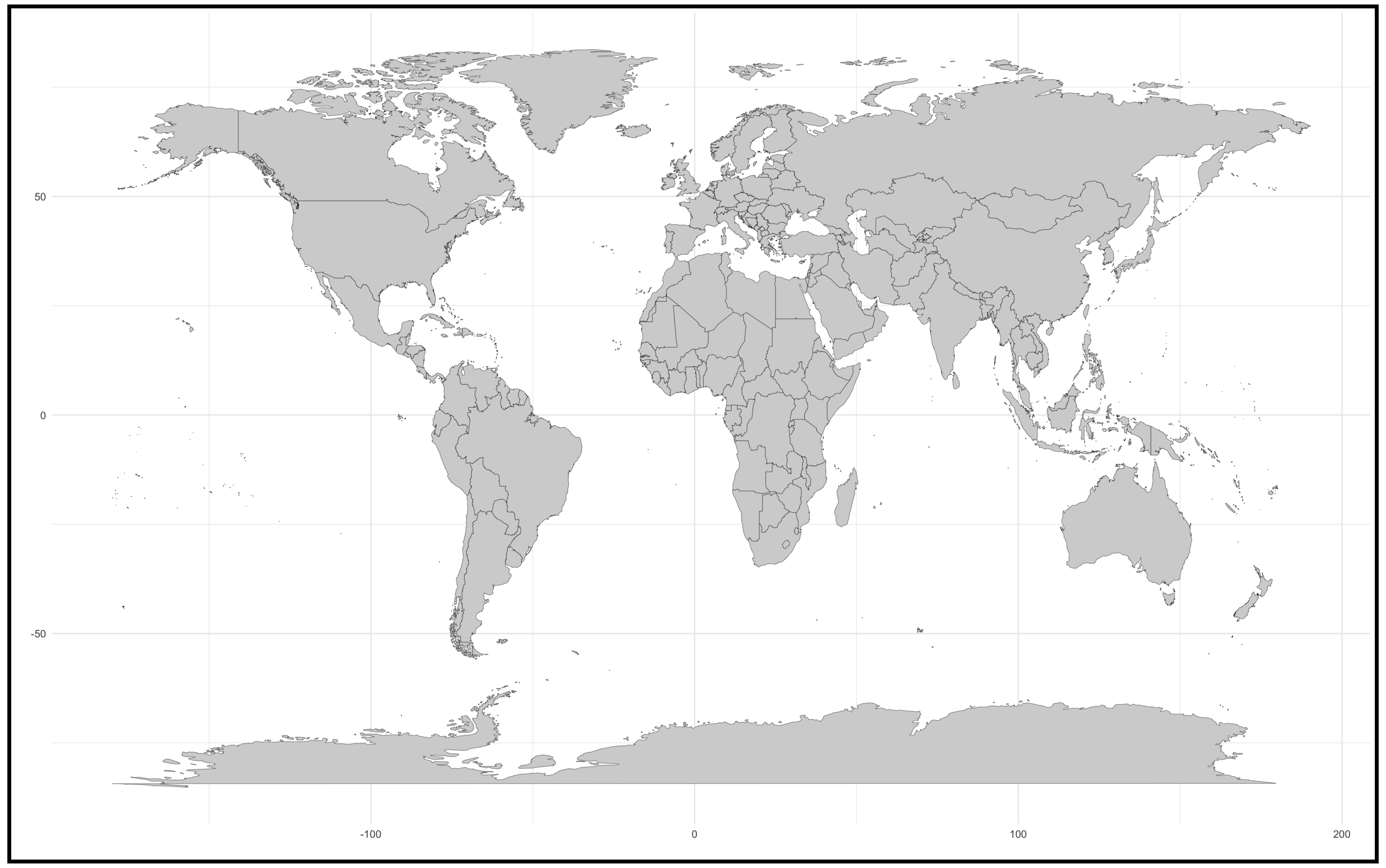
cartesian coordinates



polar coordinates



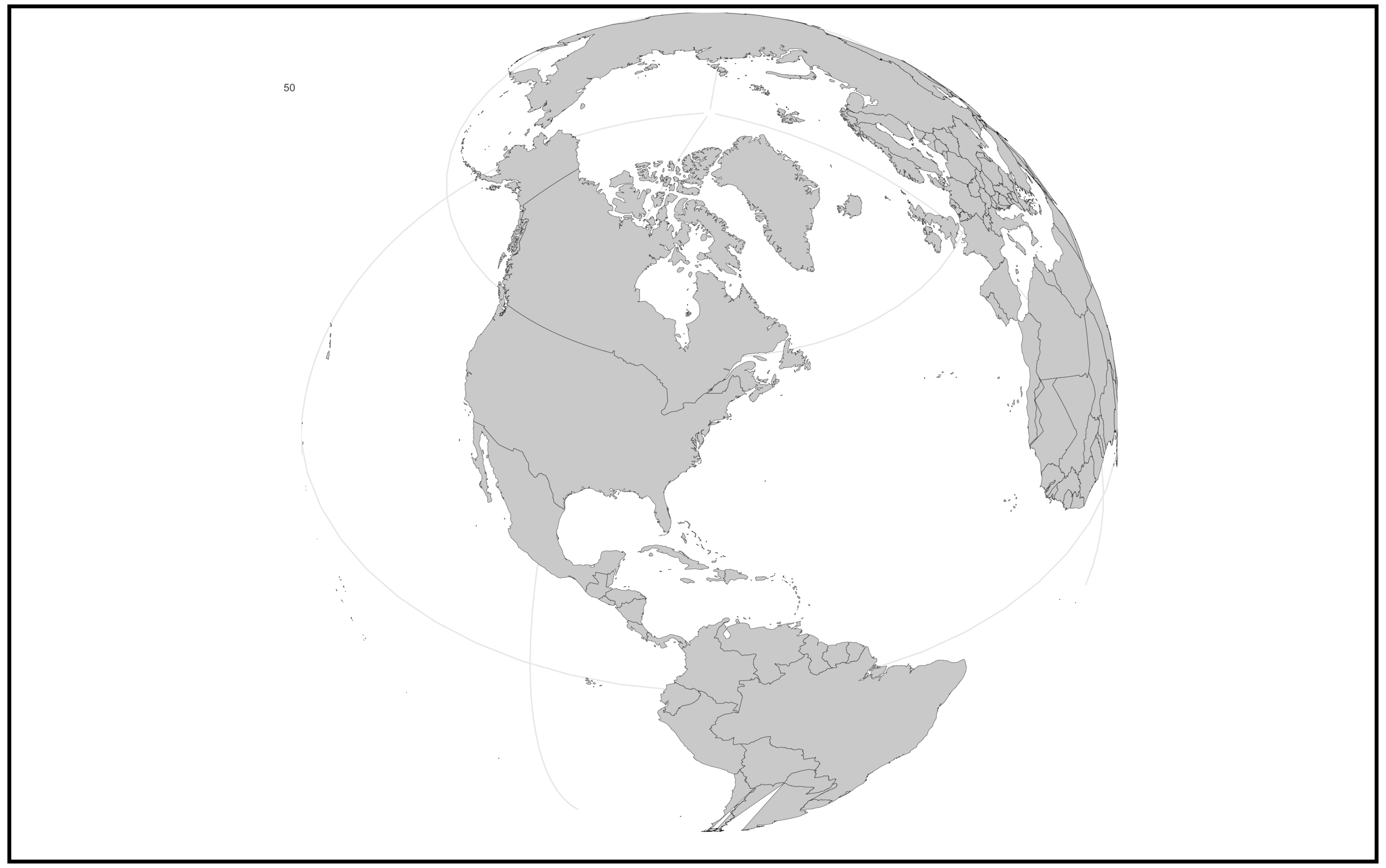
 another example, **projecting spherical surface to a plane**



Cartesian coordinates, x = longitude and y = latitude

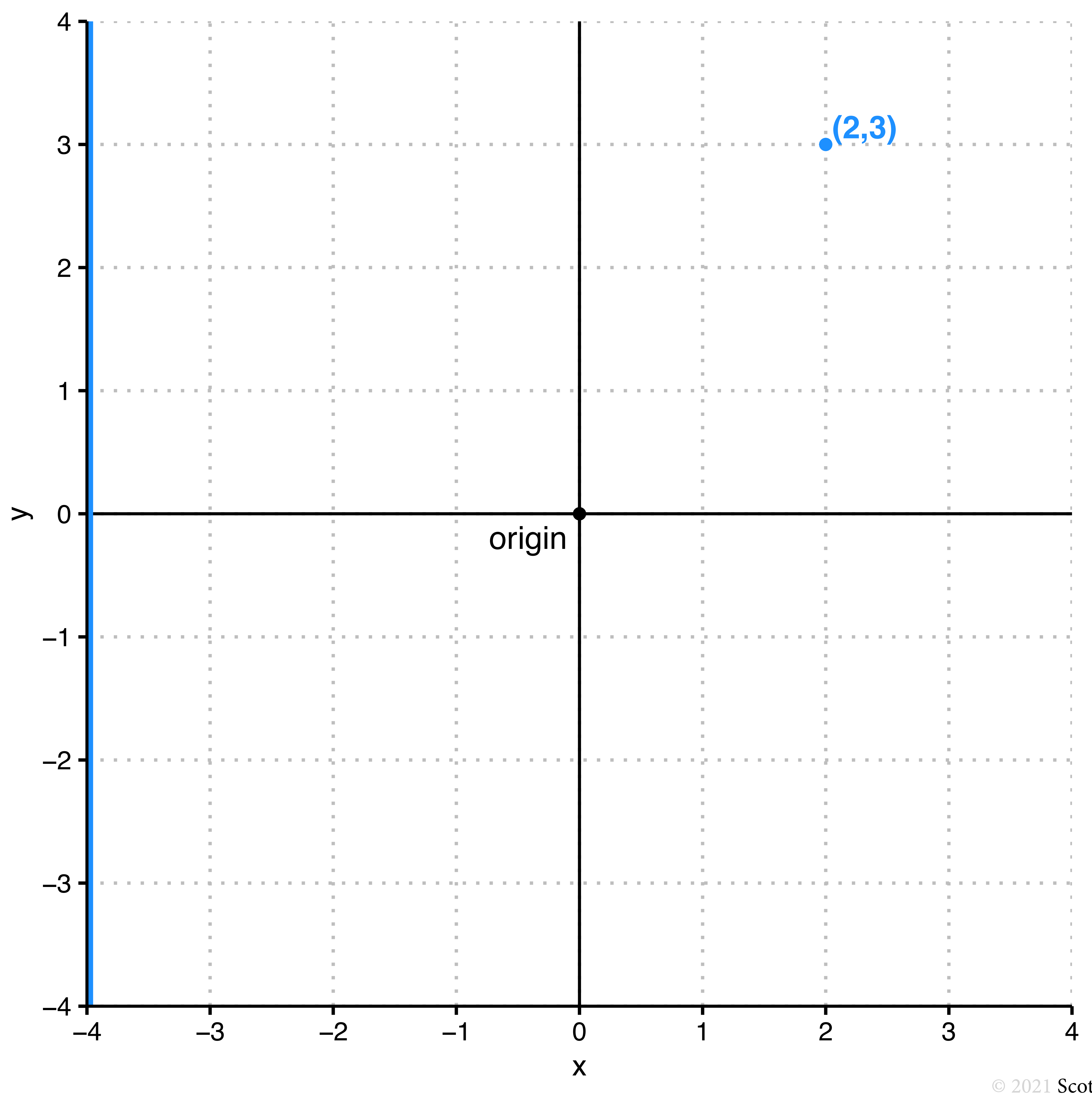
Mercator coordinates, x = longitude and y = latitude

Orthographic coordinates,
oriented towards New York City

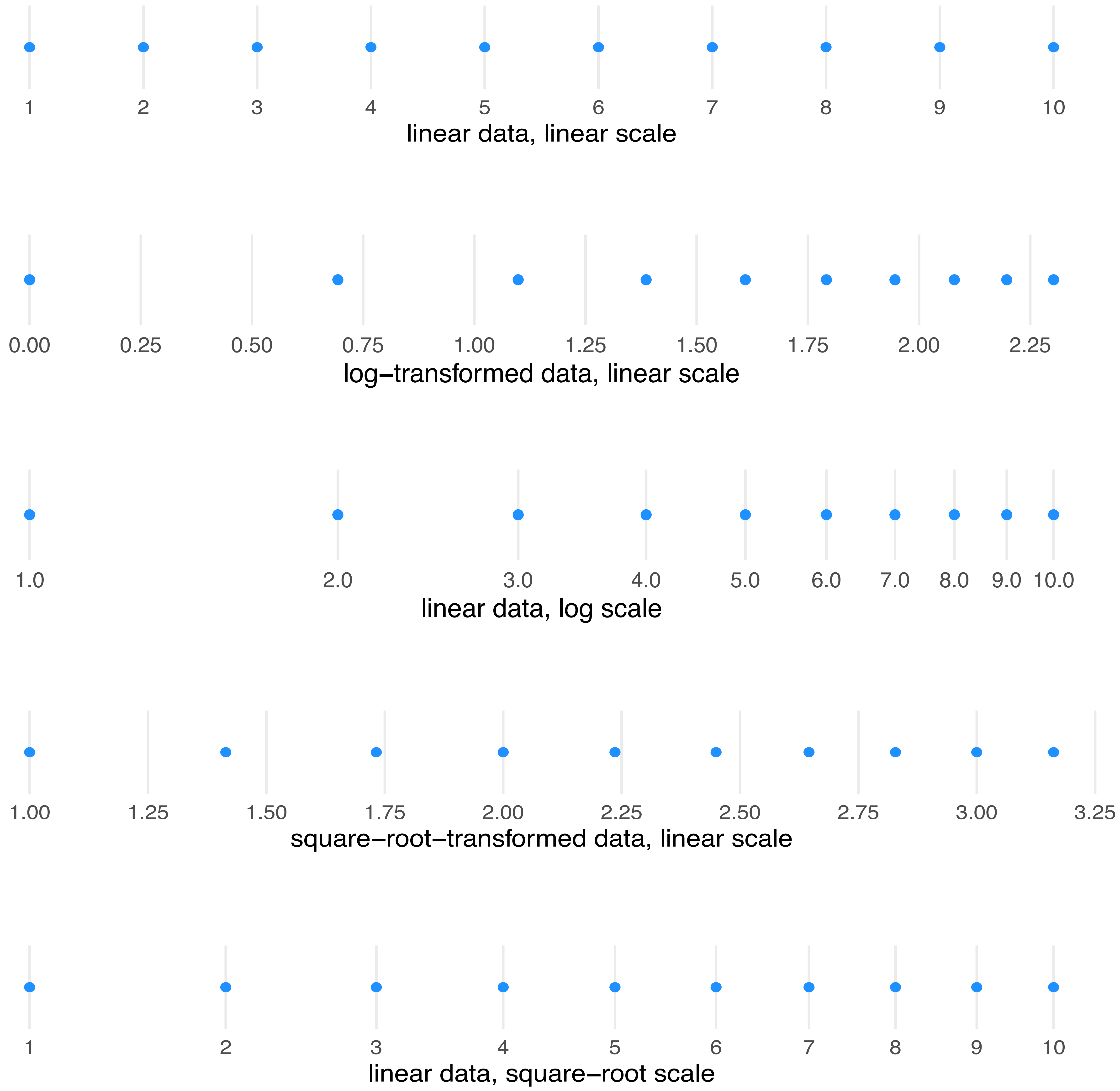


 as with choosing coordinates, we can *transform* ***scales for data or axes*** for better understanding

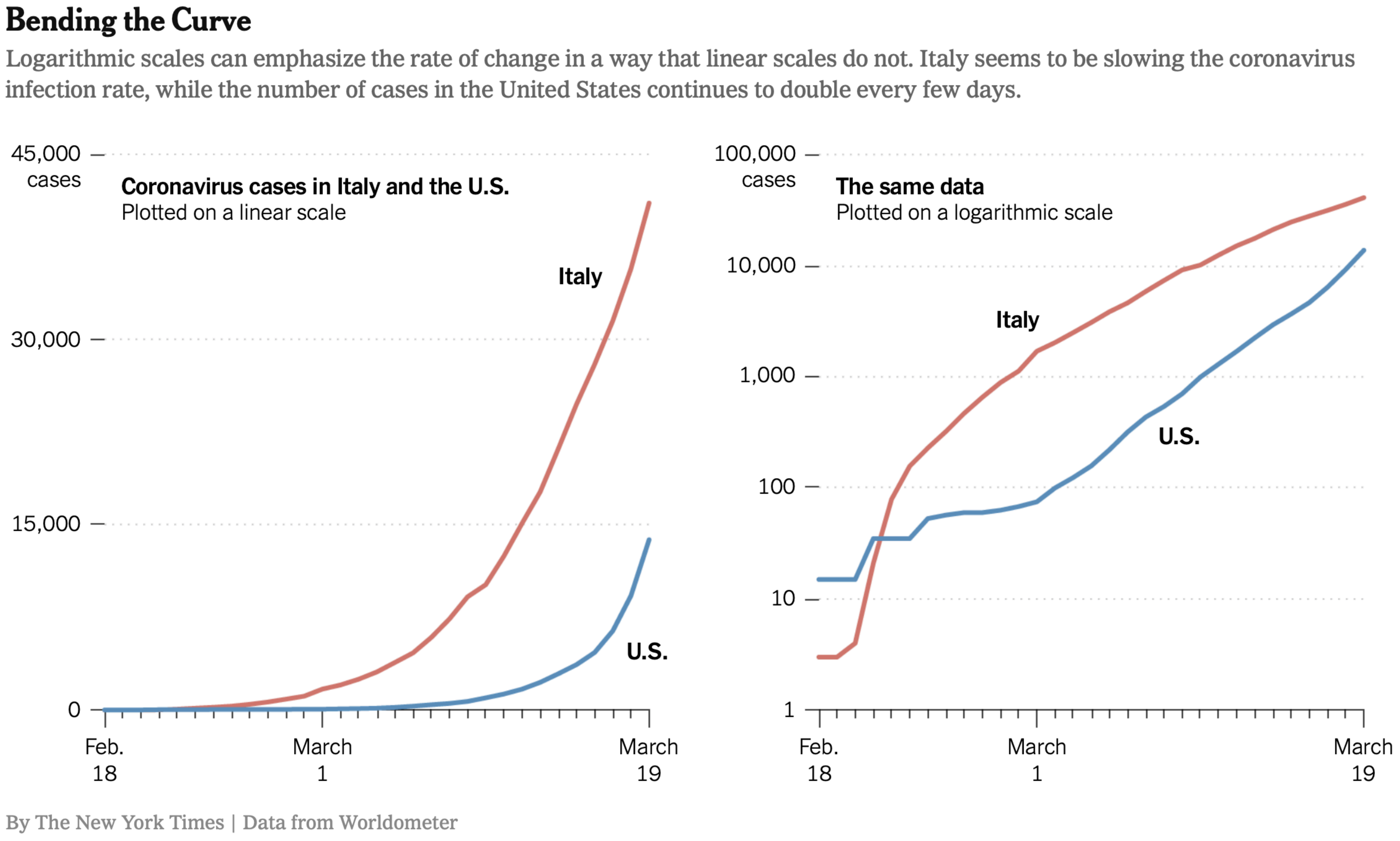
linear ***scales*** on cartesian coordinates



data or scale transformations, examples



 as with choosing coordinates, we can *transform* *scales for data or axes* for better understanding

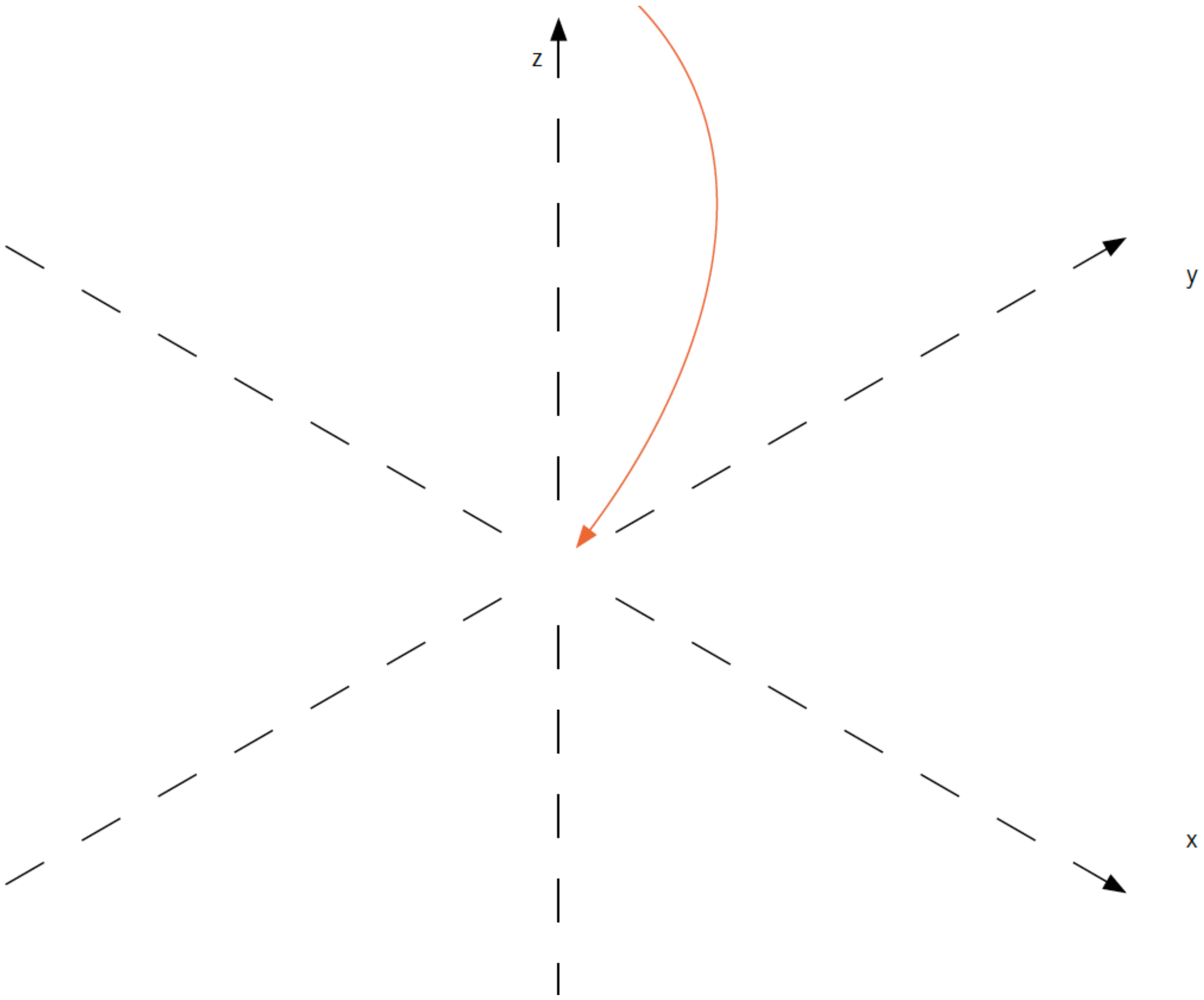


— Chang, Kenneth. *A Different Way to Chart the Spread of Coronavirus*, New York Times. 2020 March 20. https://www.nytimes.com/2020/03/20/health/coronavirus-data-logarithm-chart.html

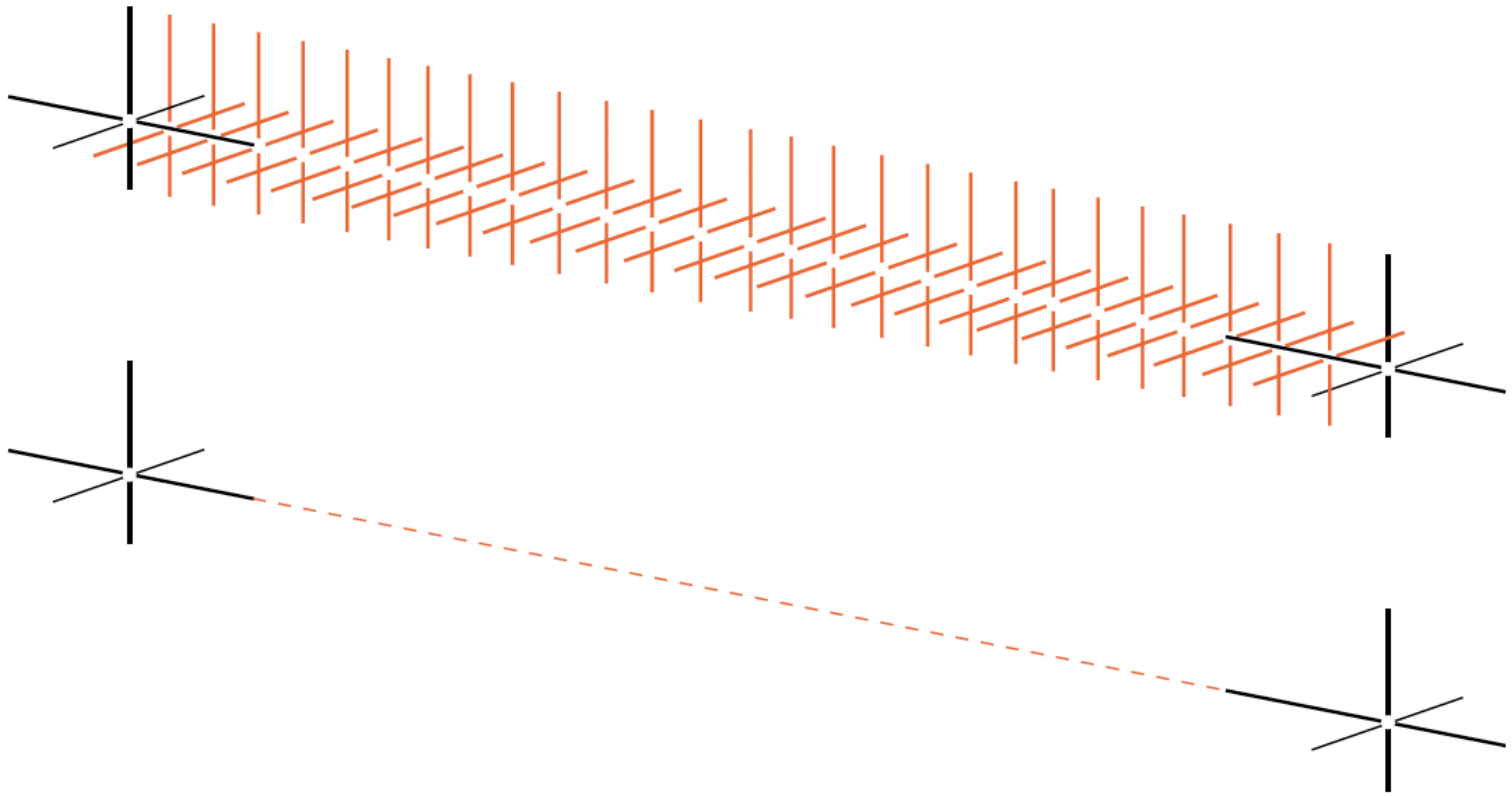
data encodings for visual comparison

 geometry of graphical elements — point

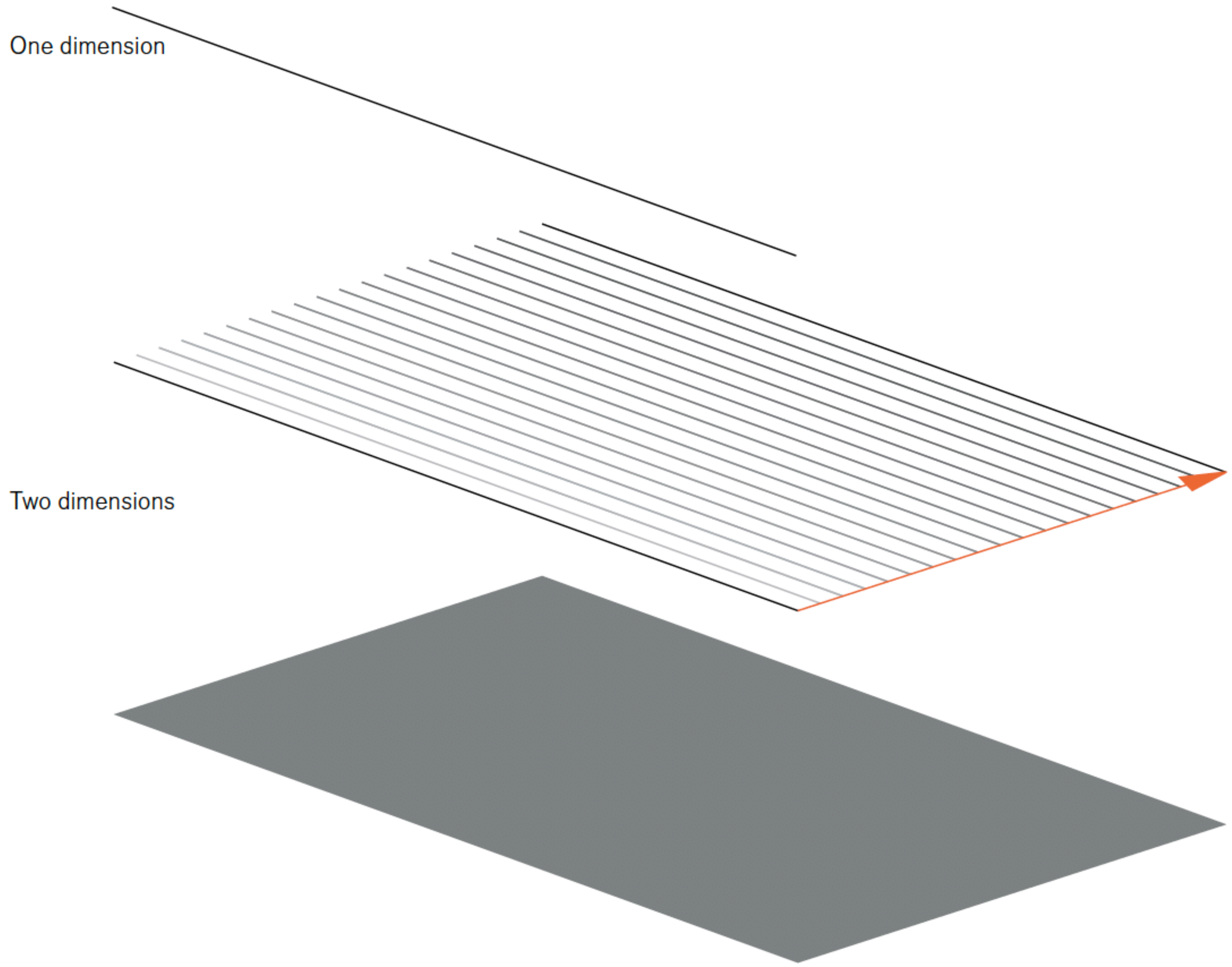
point := (x, y, z)



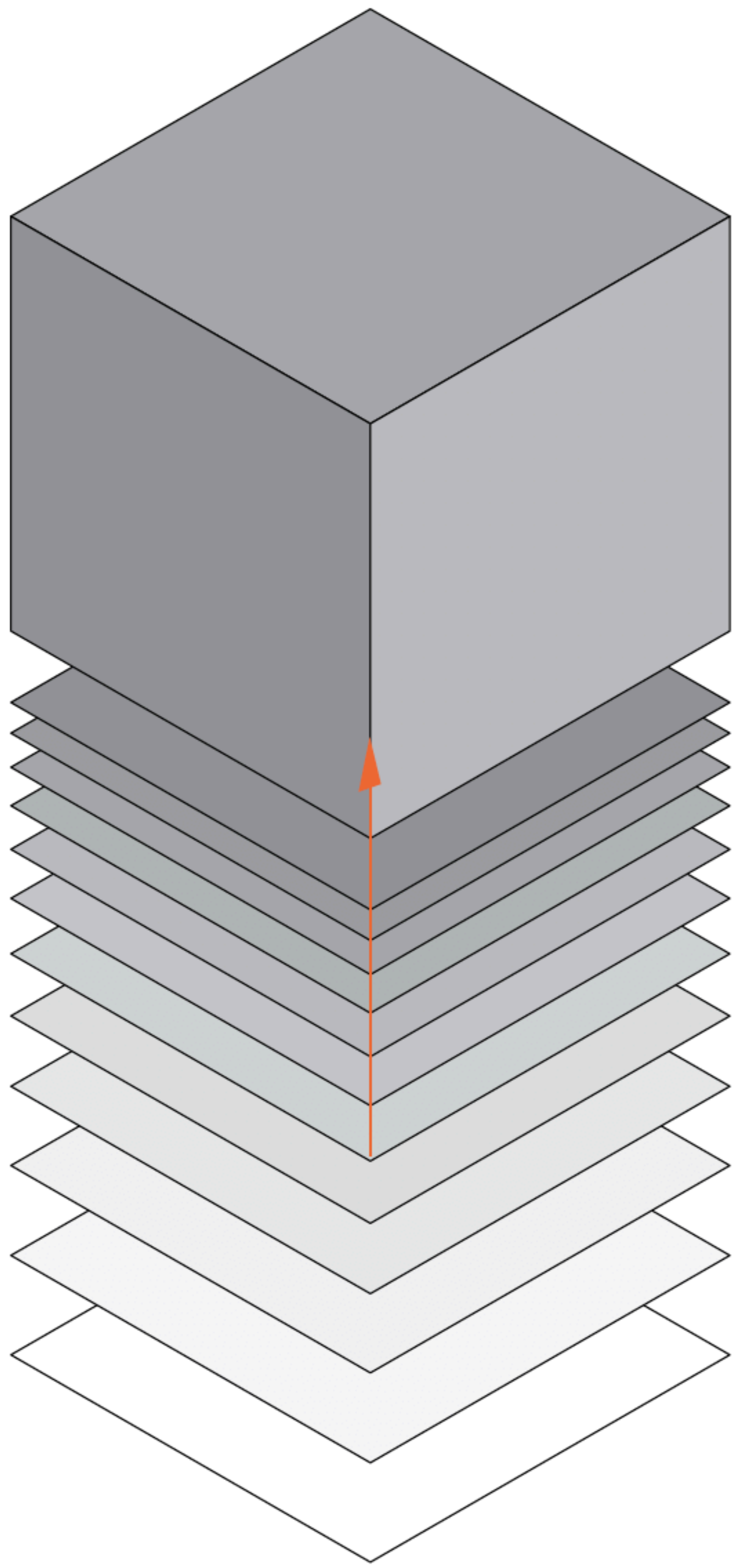
data encodings, geometry of graphical elements — line



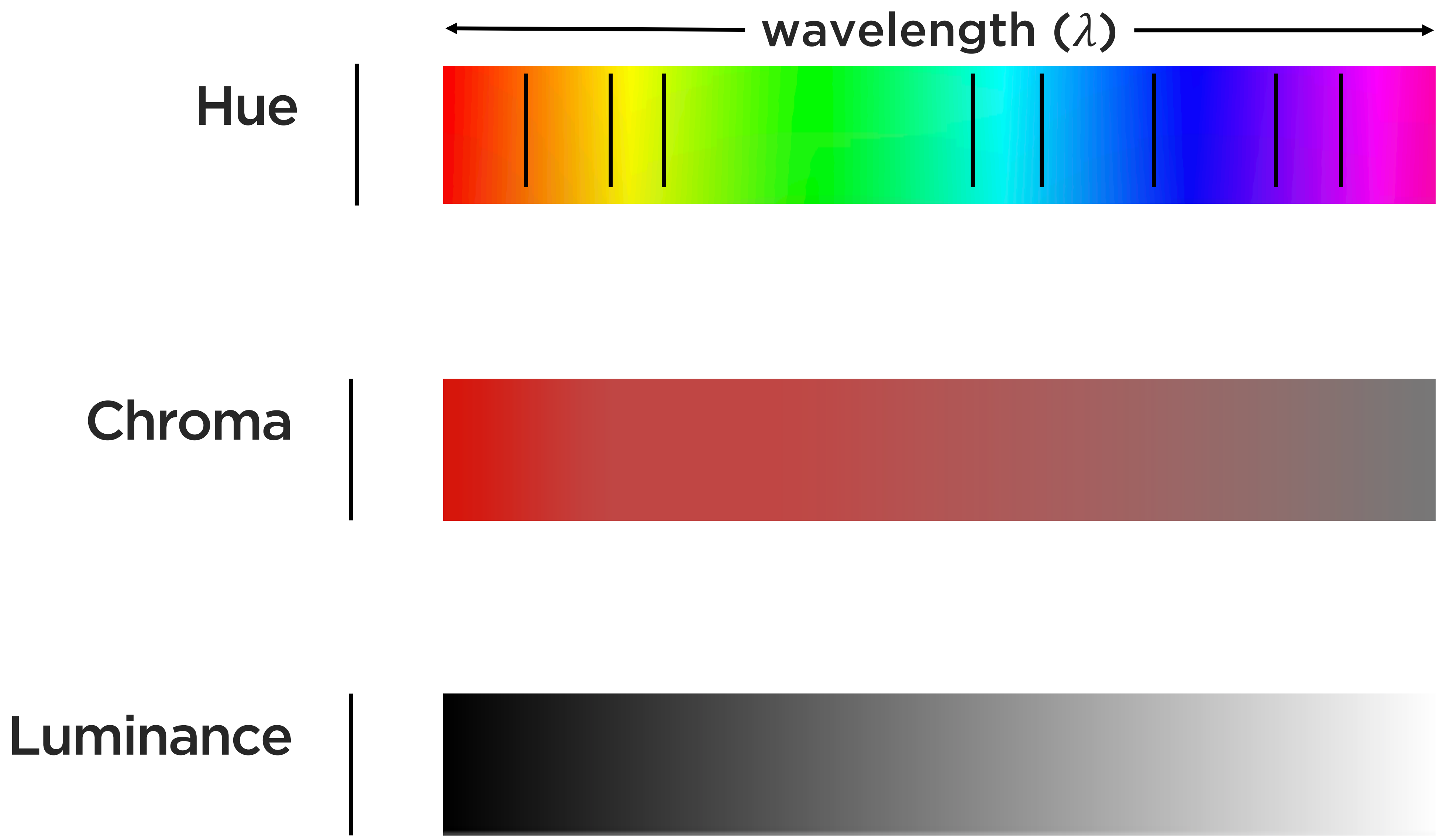
data encodings, geometry of graphical elements — surface



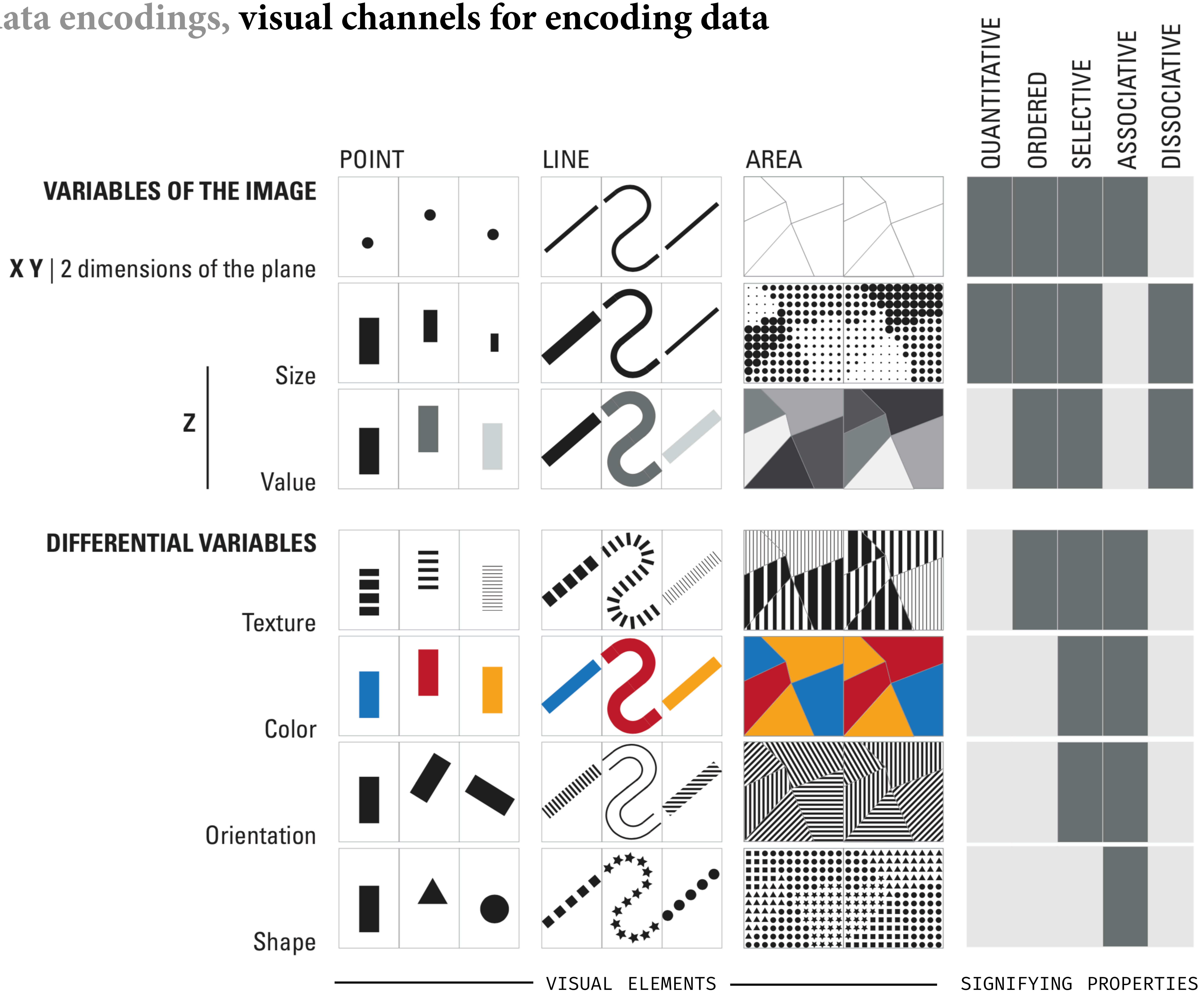
data encodings, geometry of graphical elements — volume



data encodings, geometry of graphical elements — color channels (hue, chroma, luminance)



data encodings, visual channels for encoding data

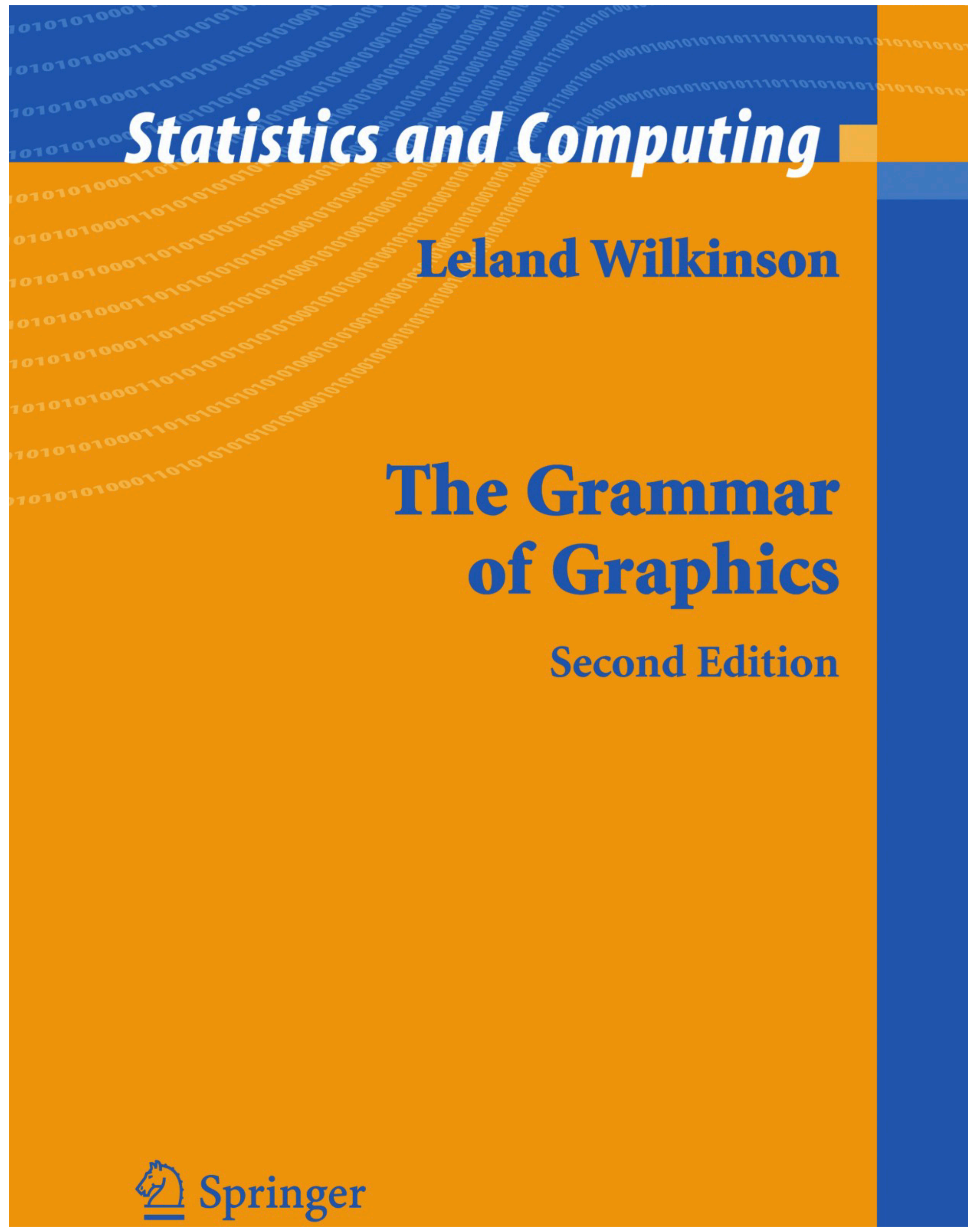


the grammar of graphics

the grammar of graphics, grammar — describes the form of relationships between [things]

GRAMMAR **:** 1.a. That department of the study of a [thing] which deals with its inflectional forms or other means of indicating the relations of [parts in things], and with the rules for employing these in accordance with established usage…

think data encodings, *not* charts …

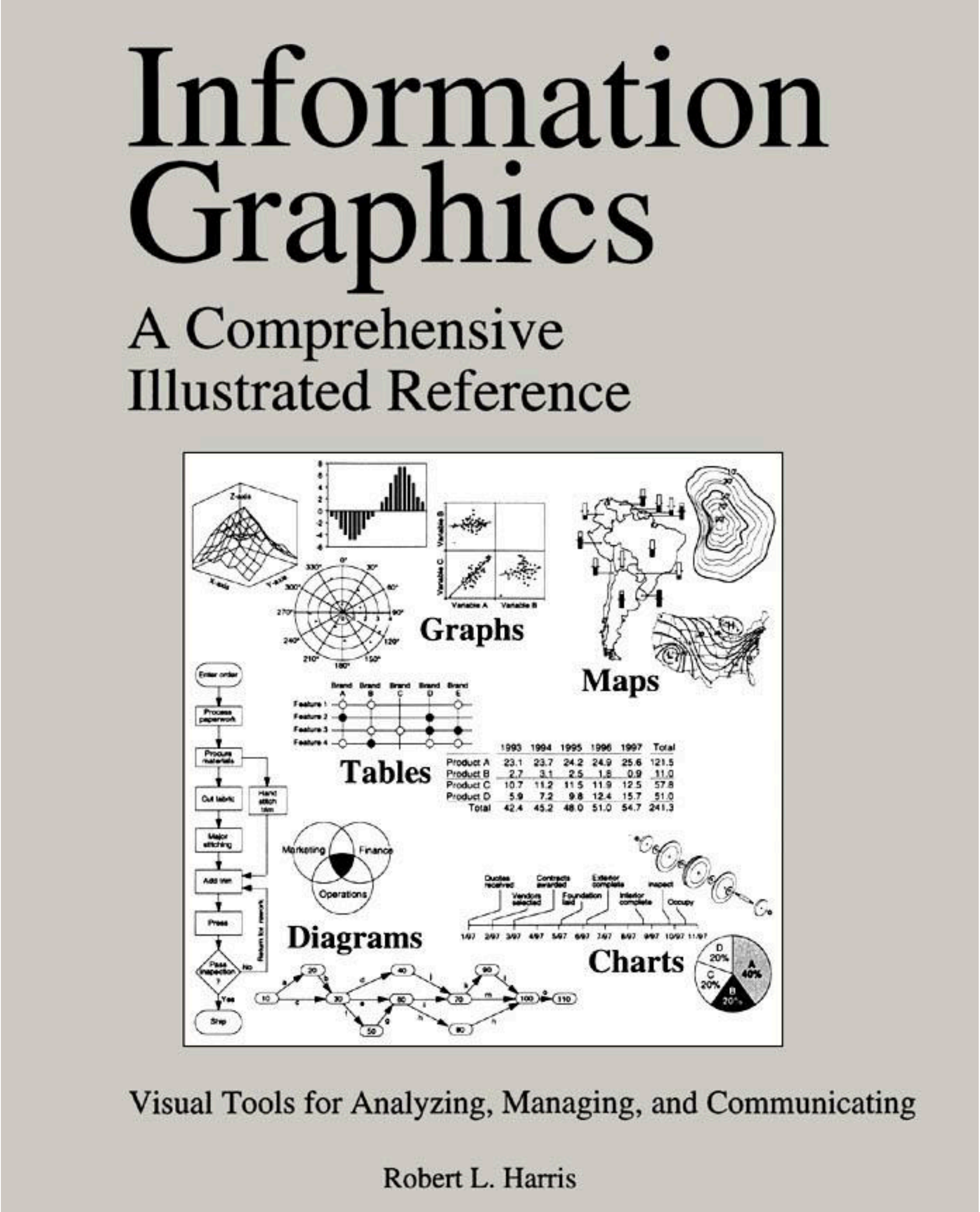


"We often call graphics charts. There are pie charts, bar charts, line charts, and so on. [We should] **shun chart typologies. Charts are usually instances of much more general objects.**

Once we understand that a pie is a divided bar in polar coordinates, we can construct other polar graphics that are less well known. We will also come to realize why a histogram is not a bar chart and why many other graphics that look similar nevertheless have different grammars…. **Elegant design requires us to think about a theory of graphics, not charts.**"

— Leland Wilkinson, *The Grammar of Graphics, Second.*

but chart typologies *can* help us learn and discuss encodings



the grammar of graphics, **statistical graphic specifications are expressed in six statements**

DATA : a set of data operations that create variables from datasets

TRANSFORMATIONS : variable transformations (*e.g., rank, log, square root*)

SCALES : scale transformations (*e.g., linear, log, square root*)

COORDINATES : a coordinate system (*e.g., cartesian, polar*)

ELEMENTS : graphs (*e.g., points, lines, areas*) and their aesthetic attributes (*e.g., position, size, hue, saturation, luminance, opacity, orientation, shape*)

GUIDES : one or more guides (*axes, legends, etc.*)

the grammar of graphics, **implementation example — ggplot2 (grammar of graphics)**

```
# load grammar of graphics
library(ggplot2)
```

Wilkinson's grammar

Data
Transformations

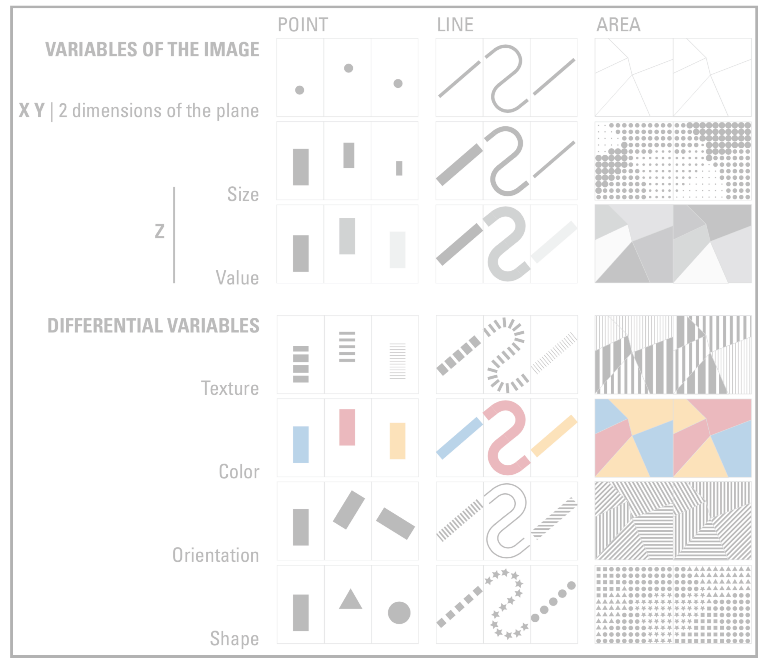
Elements (with data)
Scales & Guides
Coordinates

```
# functions for data ink

ggplot(data = <data>,
       mapping = aes(<aesthetic> = <variable>,
                     <...> = <...>) +
geom_<type>(data = <...>, mapping = aes(<...>), <...>) +
scale_<mapping>_<type>(<...>) +
coord_<type>(<...>) +
facet_<type>(<...>) +
<...> +
```

Guides

```
# functions for non-data ink

labs(<...>) +
theme(<...> = <...>) +
annotate(<...>) +
<...>
```

the grammar of graphics, implementation example — ggplot2 (grammar of graphics)




```
# load grammar of graphics
library(ggplot2)
```

WILKINSON'S GRAMMAR

```
# functions for data ink
```

DATA
TRANSFORMATIONS

```
ggplot(data = <data>,
       mapping = aes(<aesthetic> = <variable>,
                     <...> = <...>) +
```

ELEMENTS (WITH DATA)
SCALES & GUIDES
COORDINATES

```
geom_<type>(data = <...>, mapping = aes(<...>), <...>) +
scale_<mapping>_<type>(<...>) +
coord_<type>(<...>) +
facet_<type>(<...>) +
<...> +
```

BERTIN'S VARIABLES +

x, y, z, size, color, fill, alpha, angle, shape, group ... *and more, depends on the geometry used.*

BERTIN'S ELEMENTS +

point, line, segment, circle, rect, polygon, path ... *and many more!*

```
# functions for non-data ink
```

GUIDES

```
labs(<...>) +
theme(<...> = <...>) +
annotate(<...>) +
<...>
```

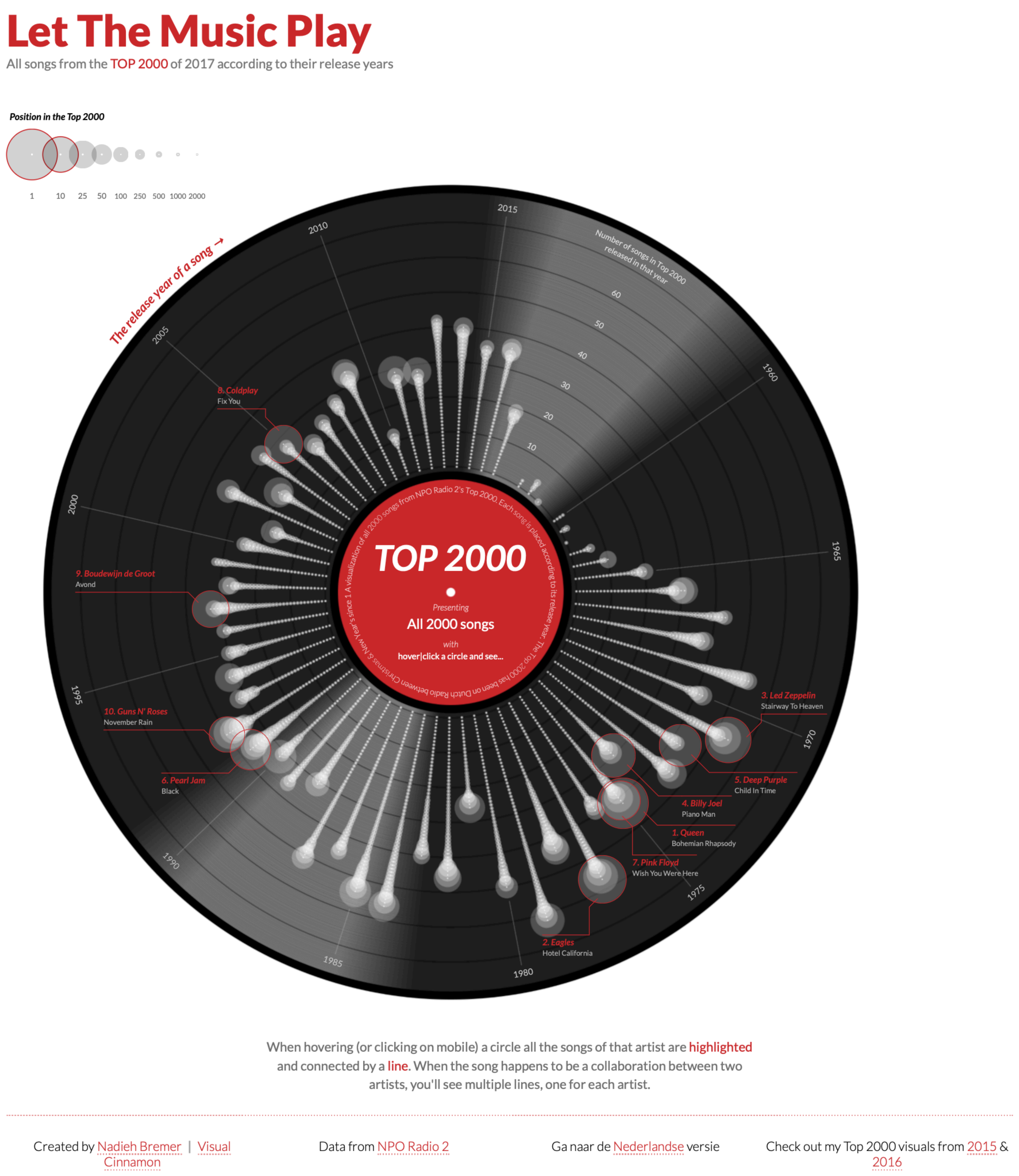
```
element_blank()
element_line(<...> = <...>)
element_rect(<...> = <...>)
element_text(<...> = <...>)
```

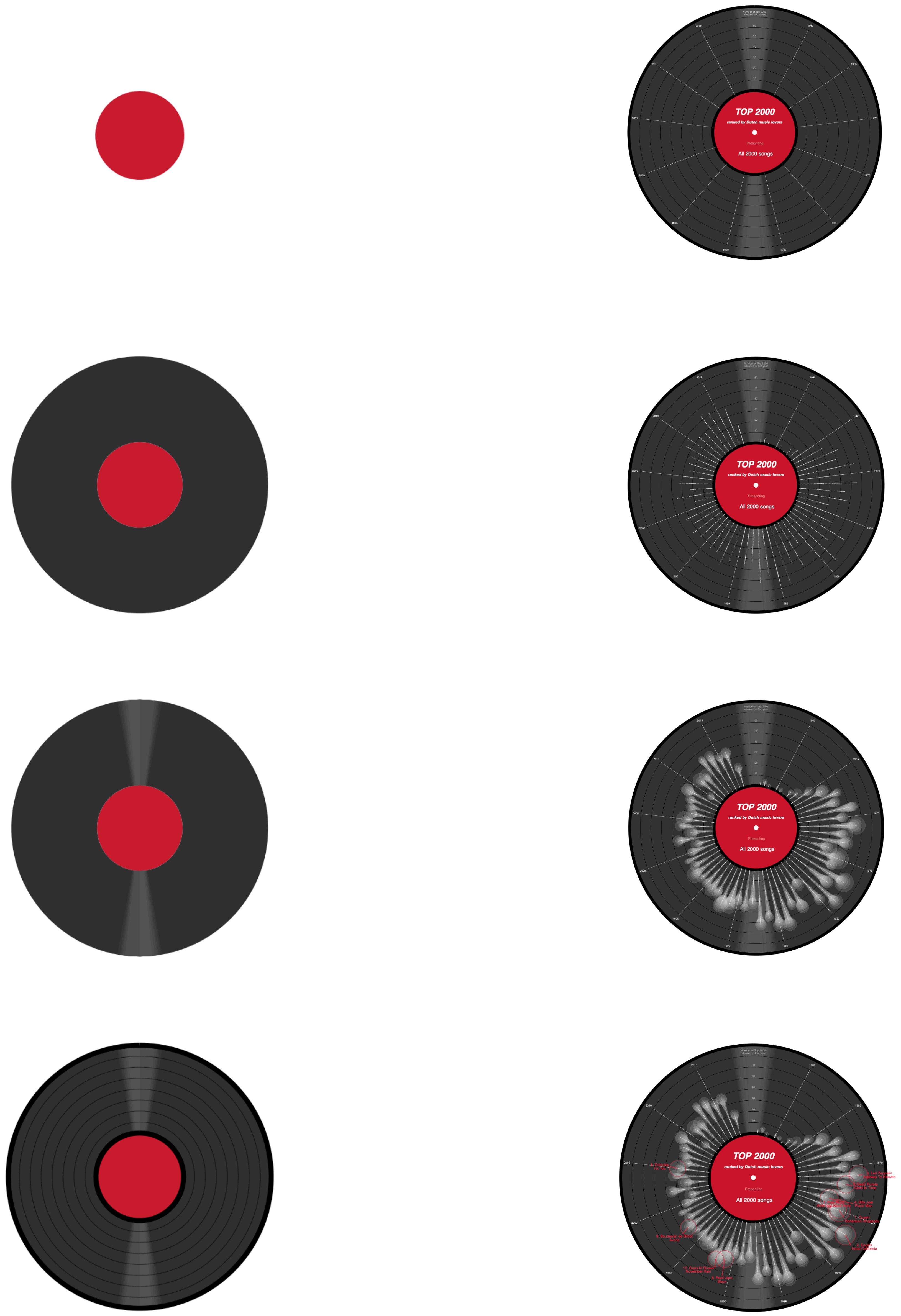
 layering — order of elements determines position towards reader and when overlapping, occlude

```
ggplot() +
  theme_void() +
  scale_x_continuous(limits = c(-5, 5)) +
  scale_y_continuous(limits = c(-5, 5)) +
  geom_point(
    mapping = aes(
      x = 0,
      y = 0),
    size = 50,
    color = "orange") +
  geom_point(
    mapping = aes(
      x = 1,
      y = 1),
    size = 50,
    color = "dodgerblue")
```

```
ggplot() +
  theme_void() +
  scale_x_continuous(limits = c(-5, 5)) +
  scale_y_continuous(limits = c(-5, 5)) +
  geom_point(
    mapping = aes(
      x = 1,
      y = 1),
    size = 50,
    color = "dodgerblue") +
  geom_point(
    mapping = aes(
      x = 0,
      y = 0),
    size = 50,
    color = "orange")
```

the grammar of graphics, **elements in *layers* — example**

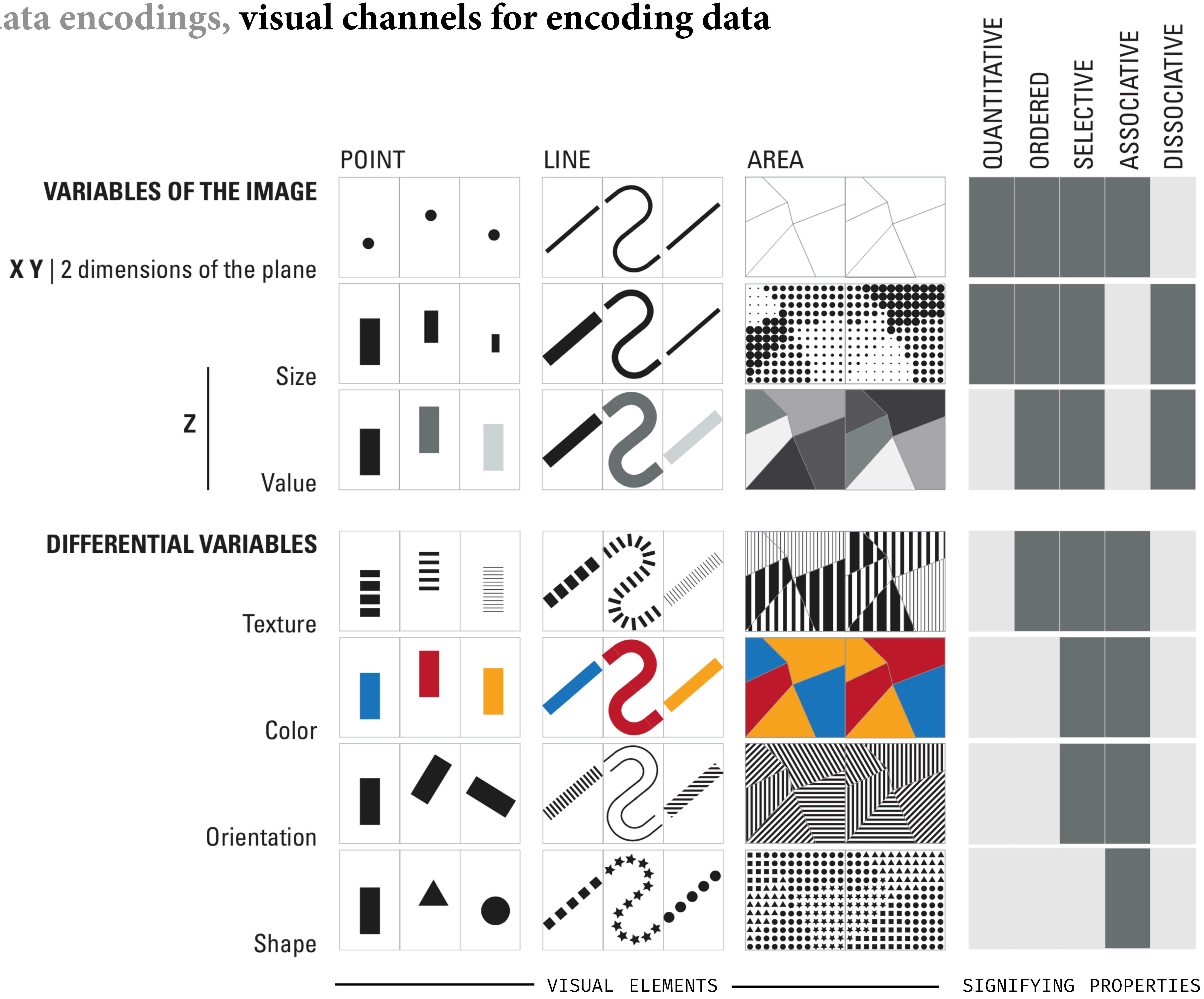




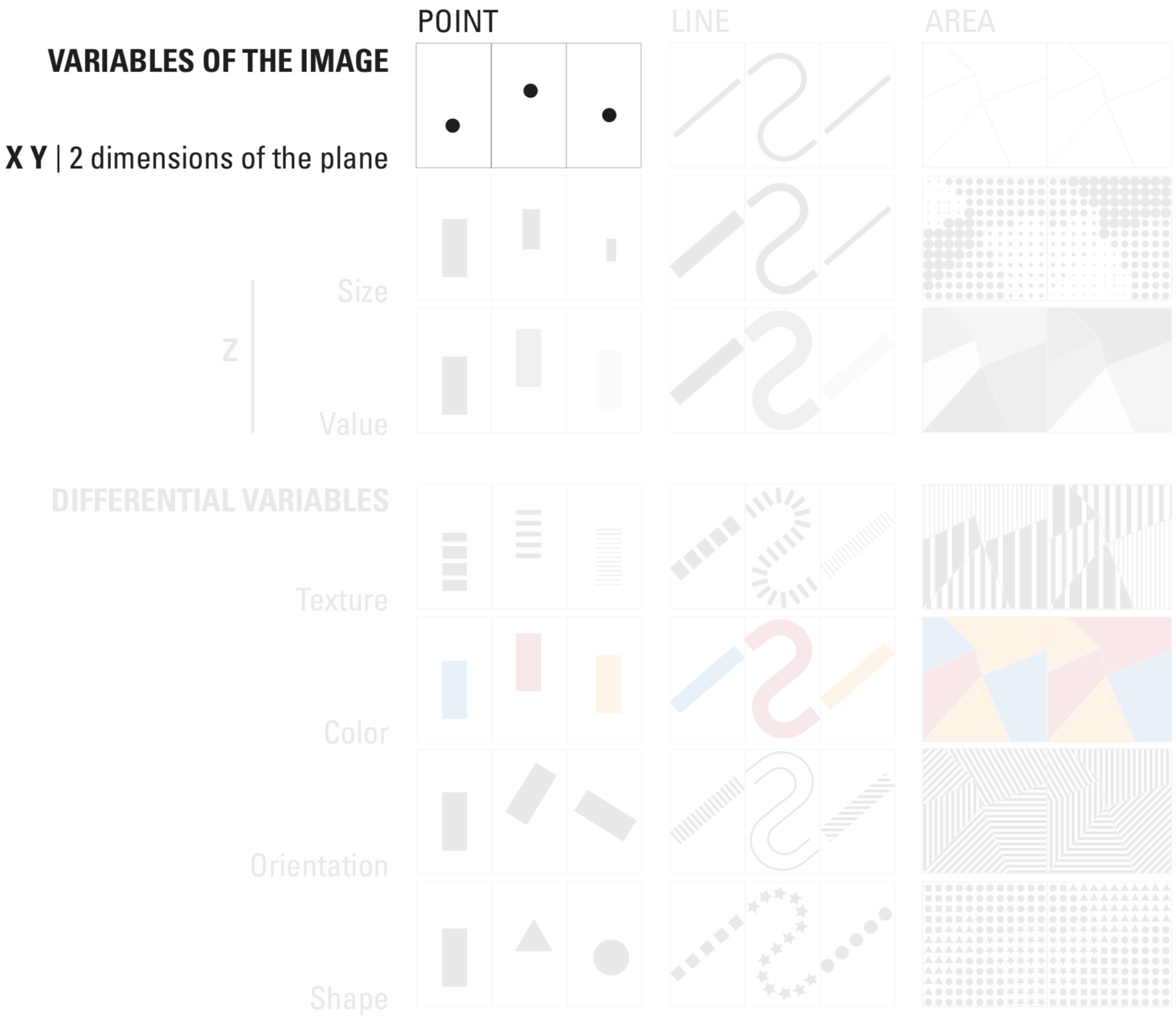
— **Spencer**, Scott. "*Demonstration of Layers in Graphics*." Publications, March 6, 2020. https://ssp3nc3r.github.io/publications/Spencer-2020-Demonstration-of-layers-in-graphics.pdf.

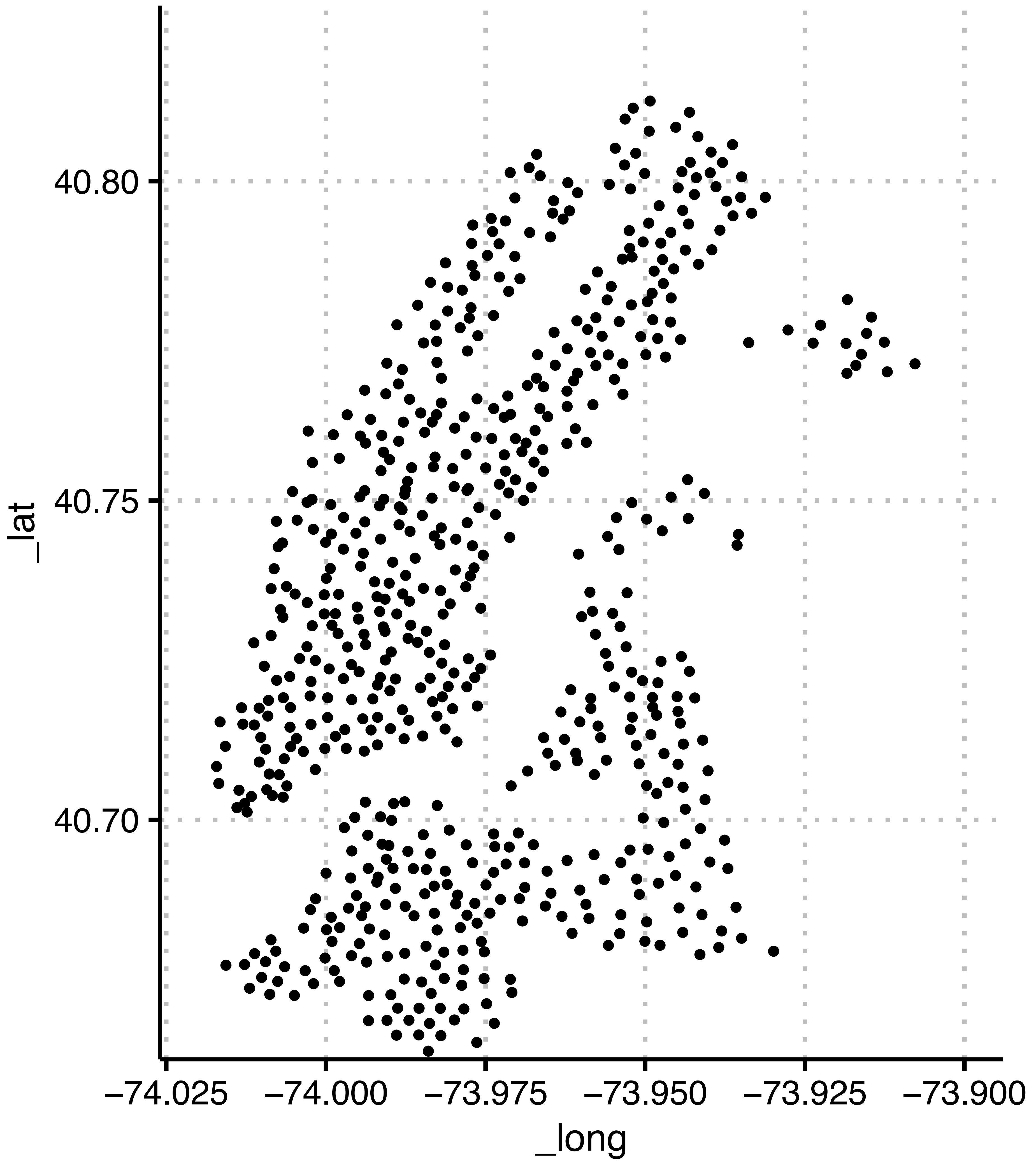
encodings and grammar — exploratory practice with our Citi Bike case study

data encodings, **visual channels for encoding data**

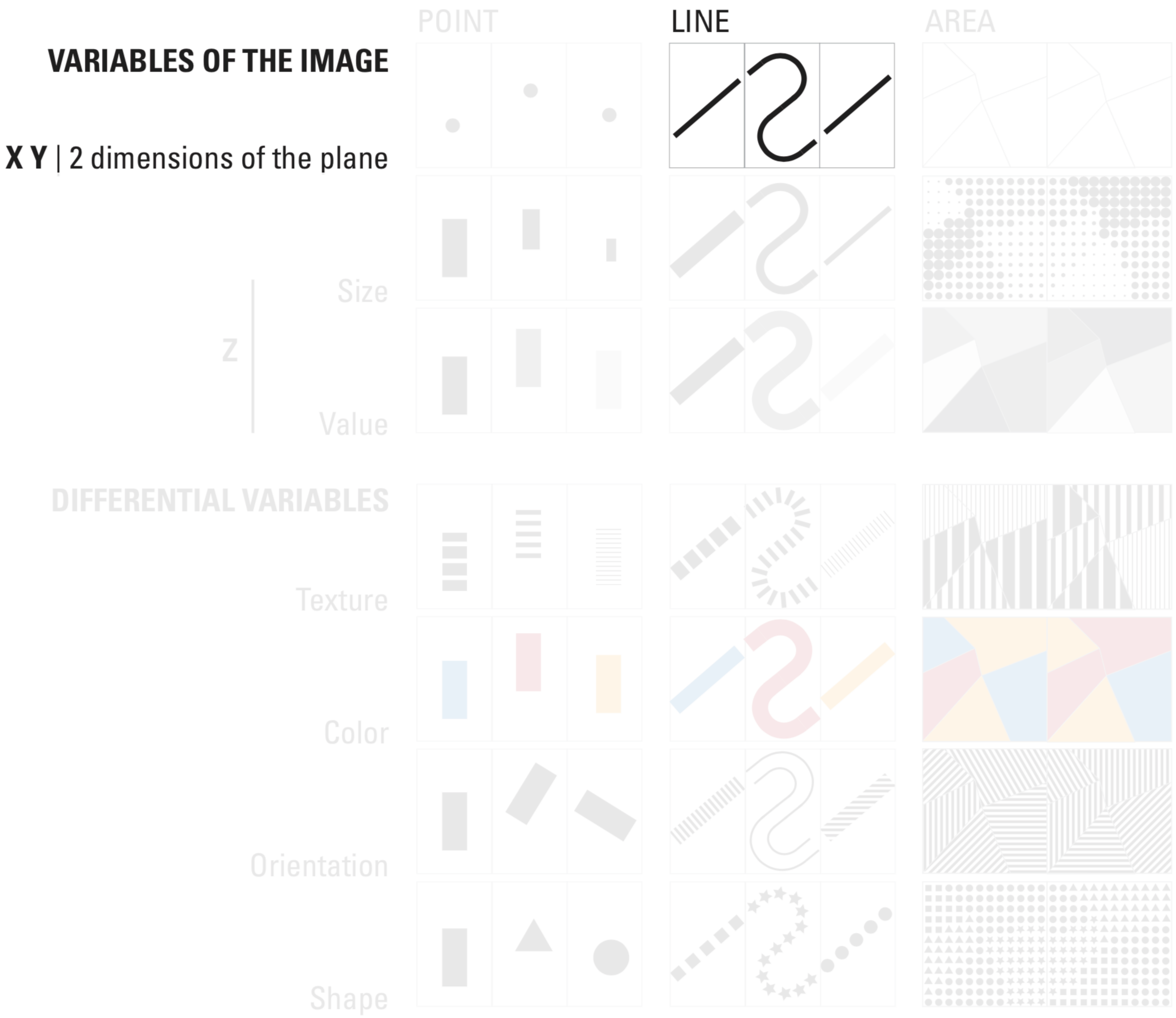


Citi Bike example — *exploratory data analysis*

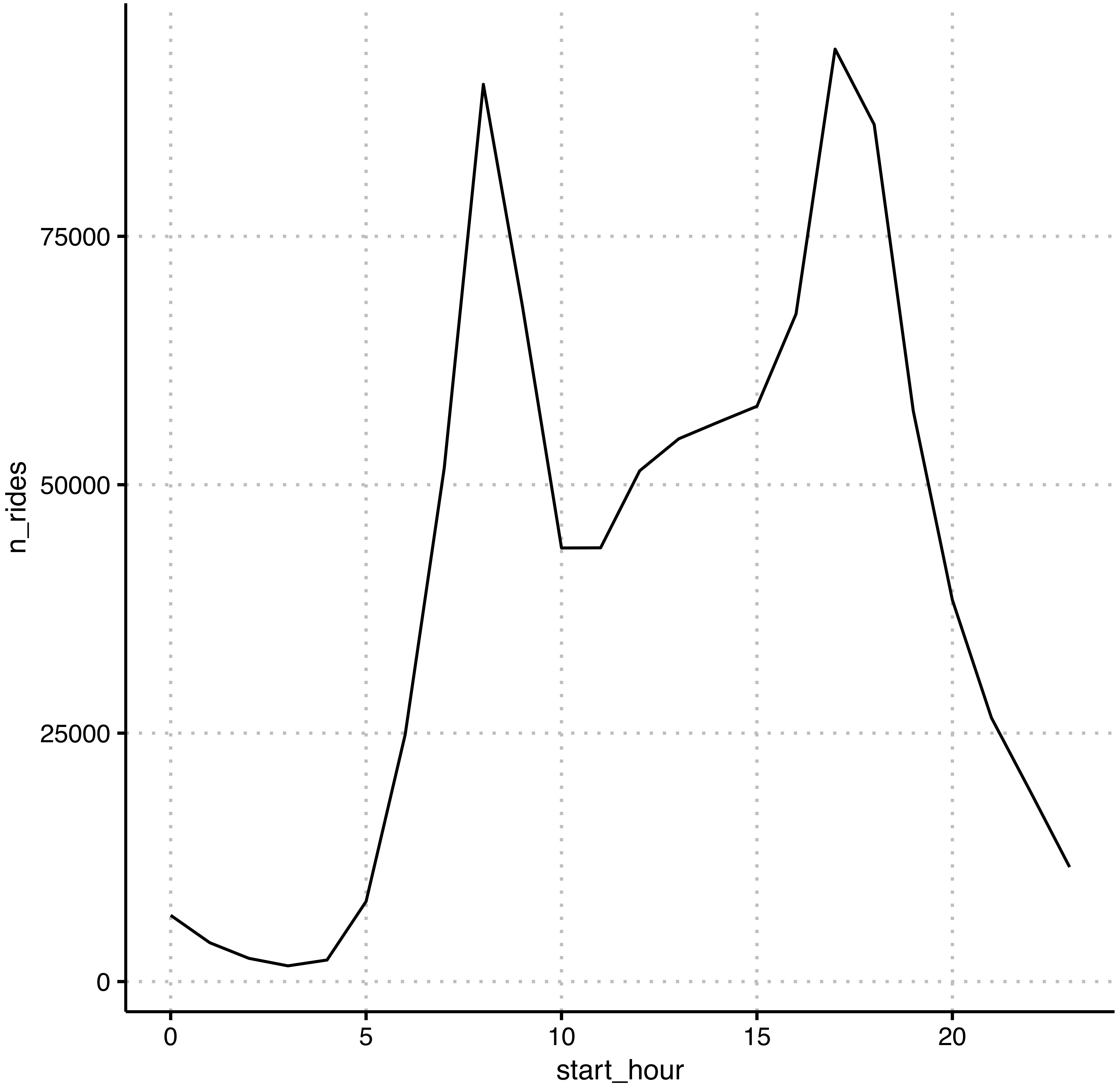




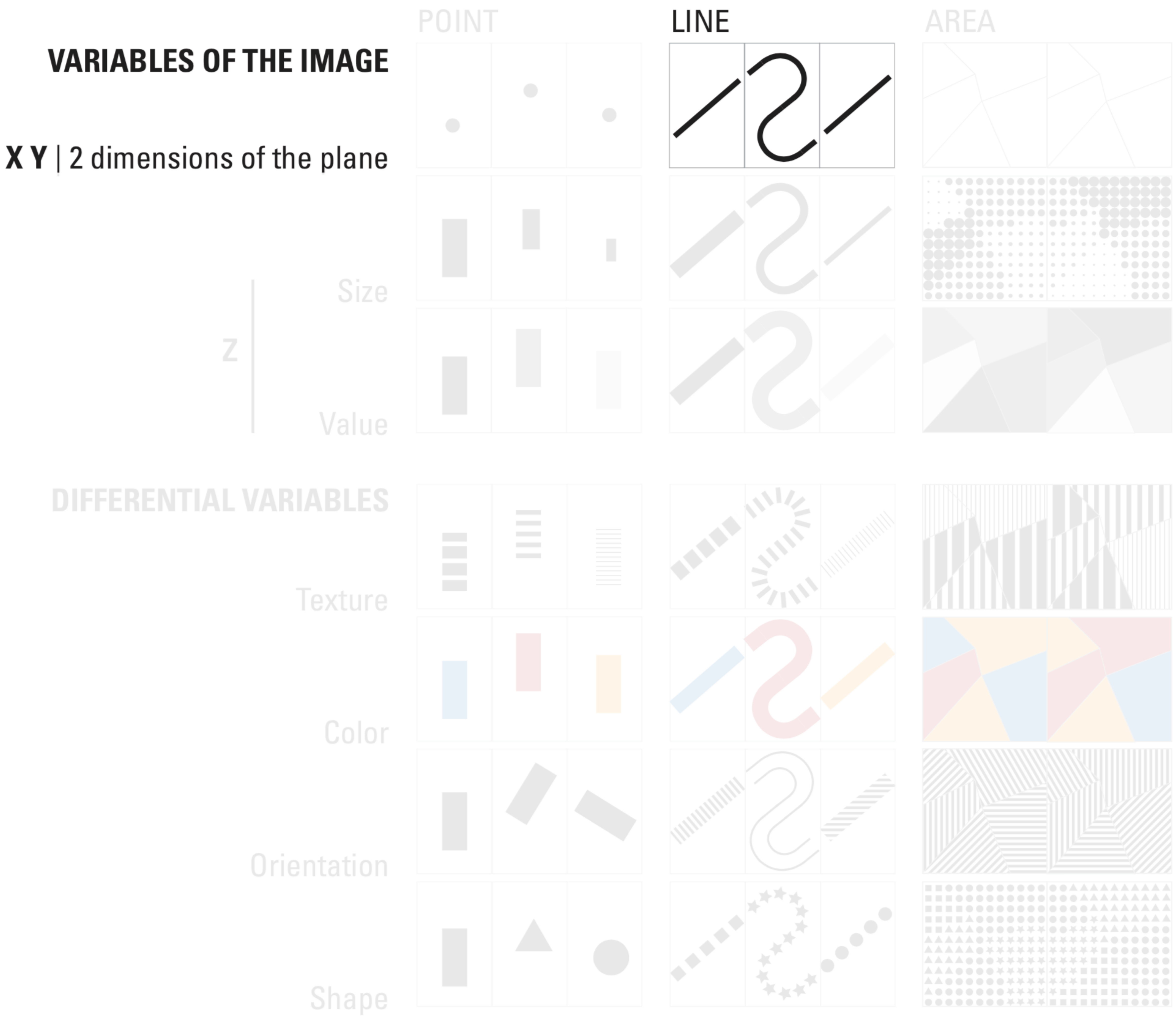
data encodings, visual channels for encoding data



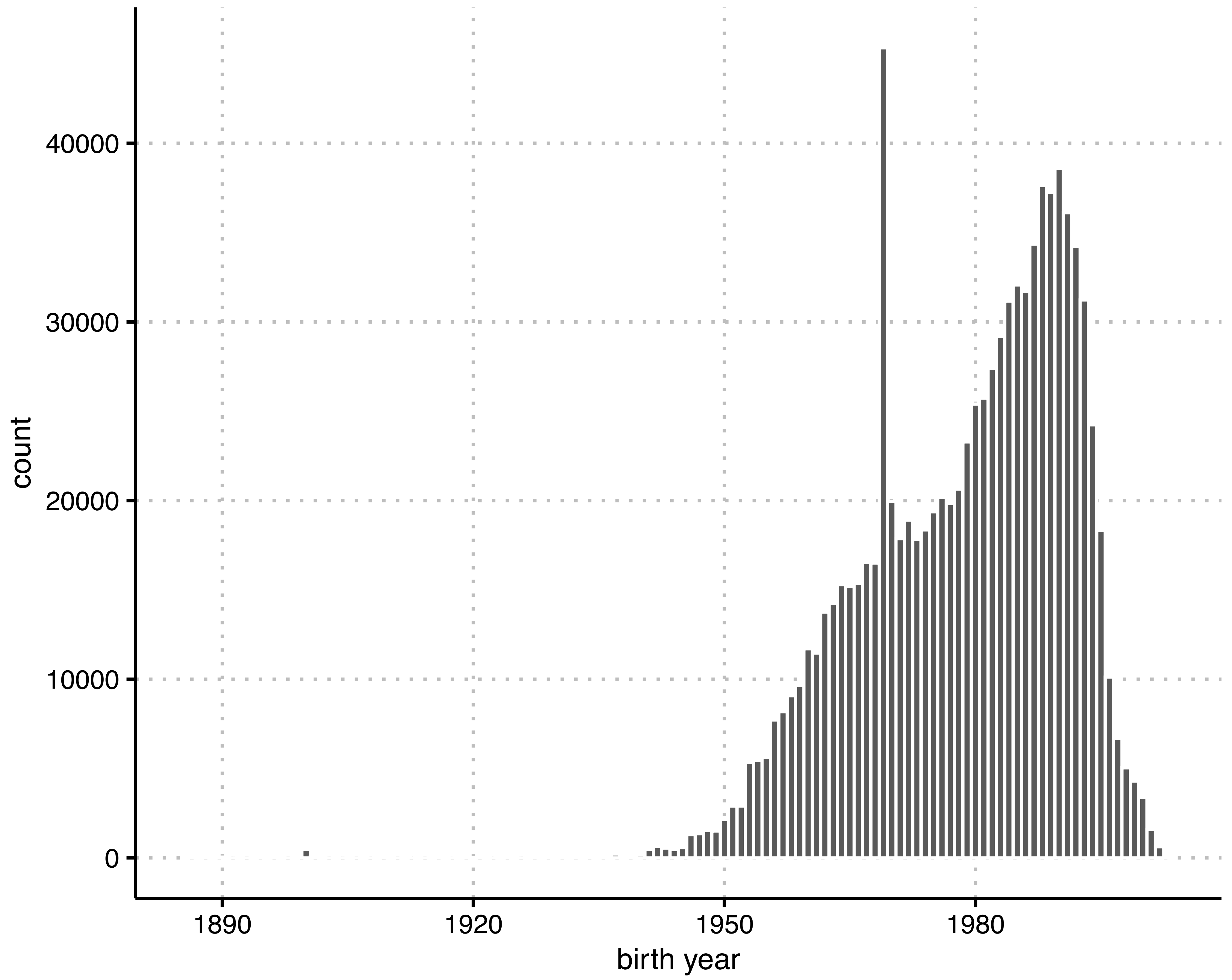
Citi Bike example — *exploratory data analysis*



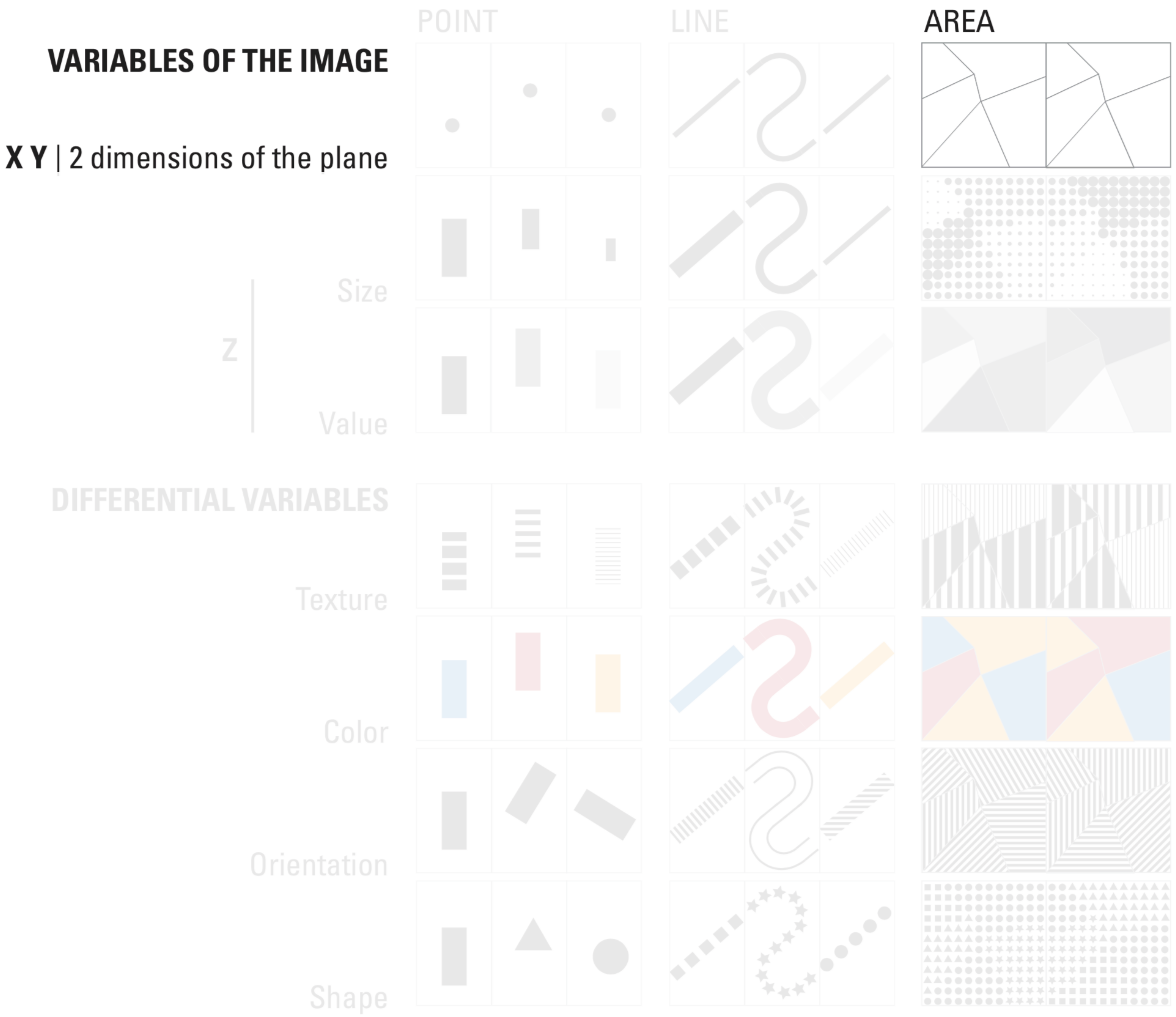
data encodings, visual channels for encoding data

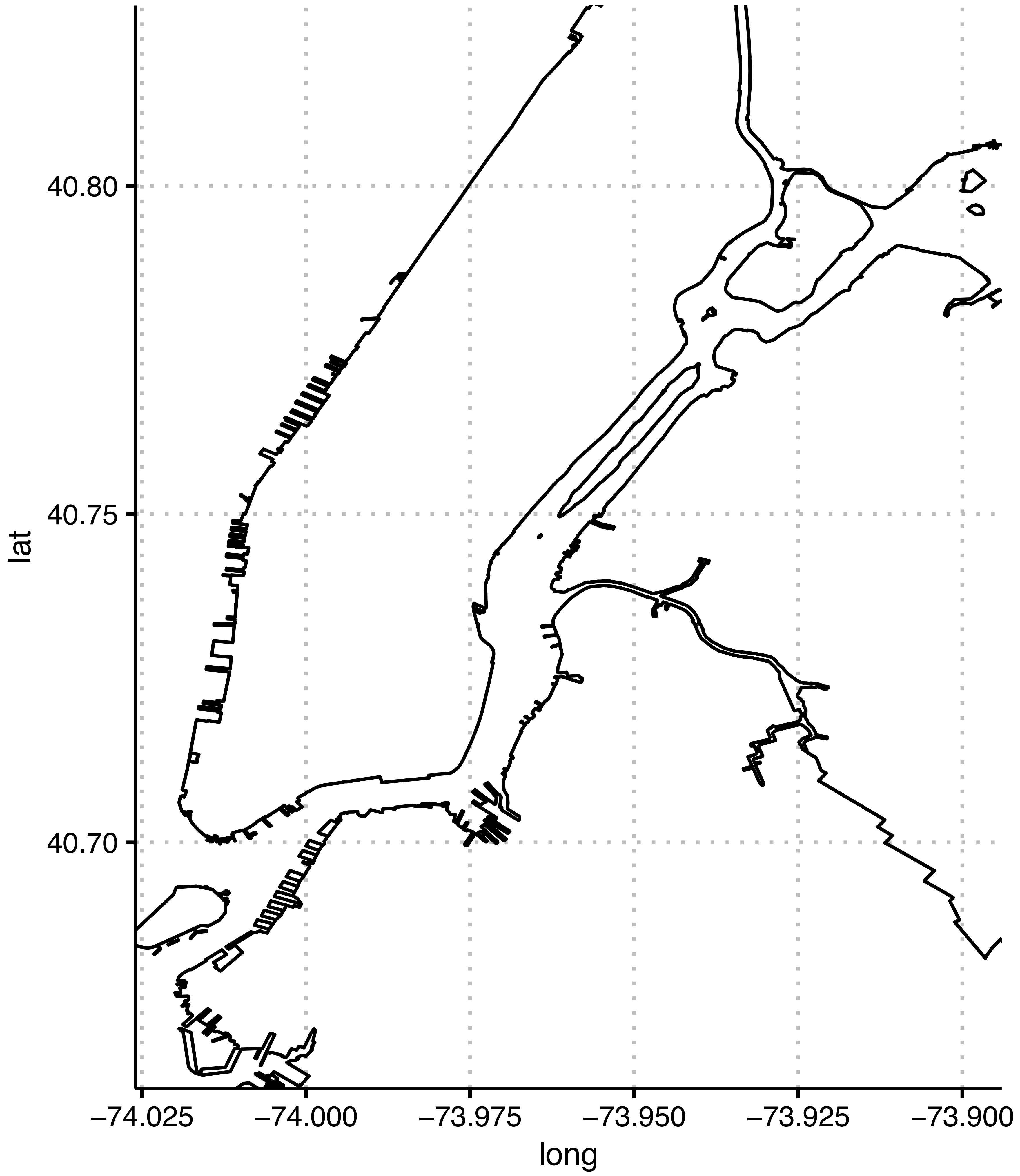


Citi Bike example — *exploratory data analysis*

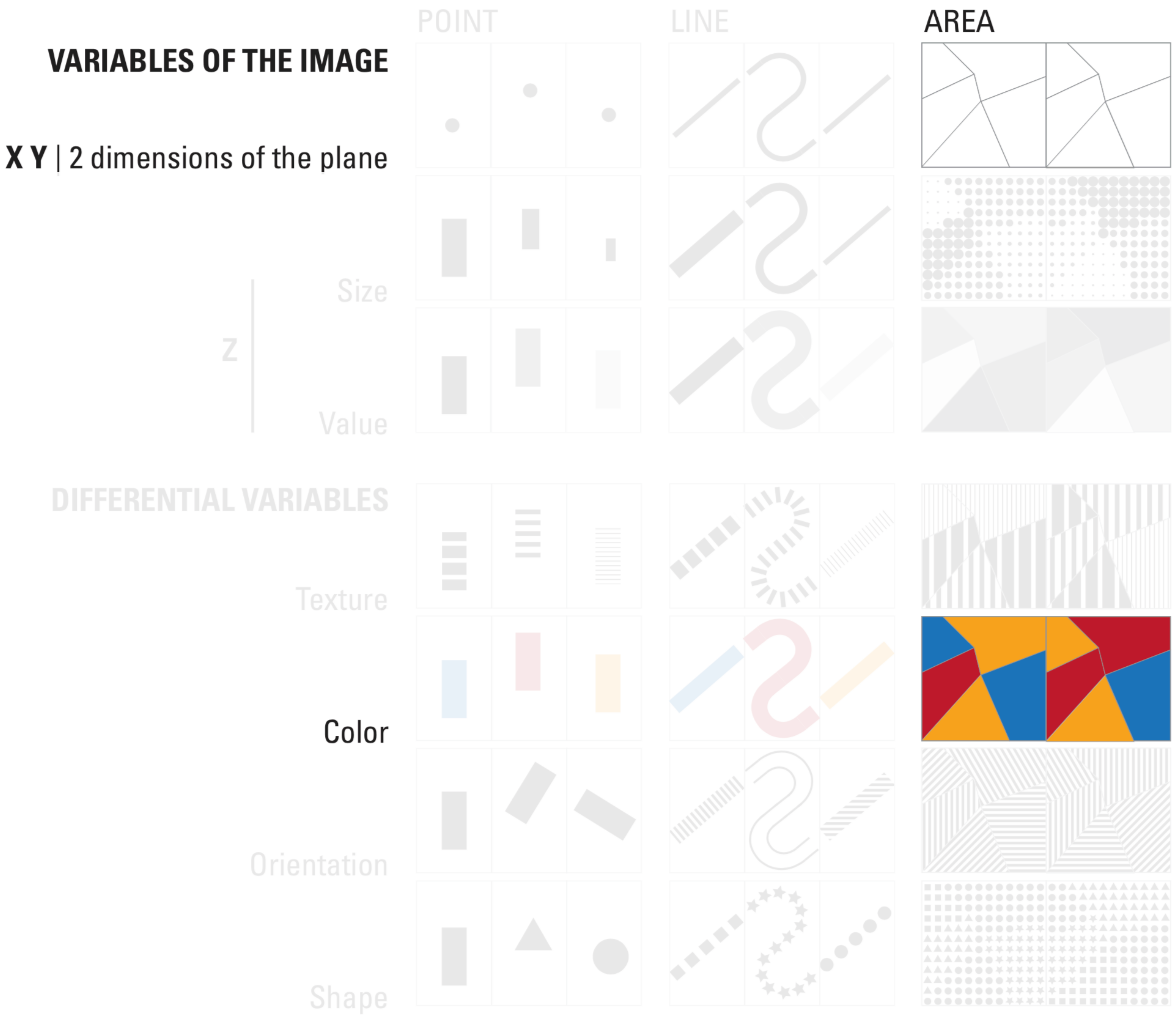


Citi Bike example — *exploratory data analysis*

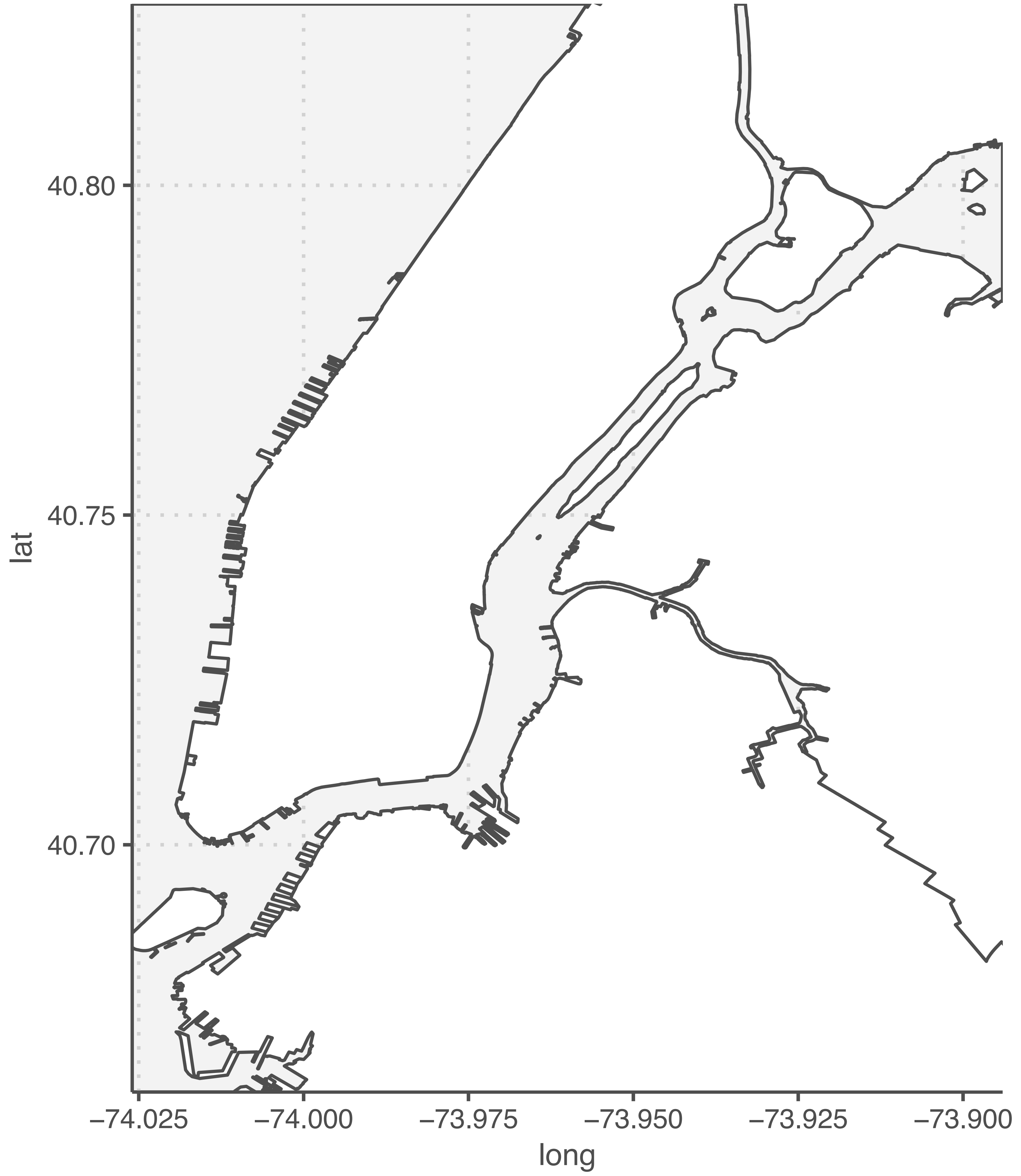




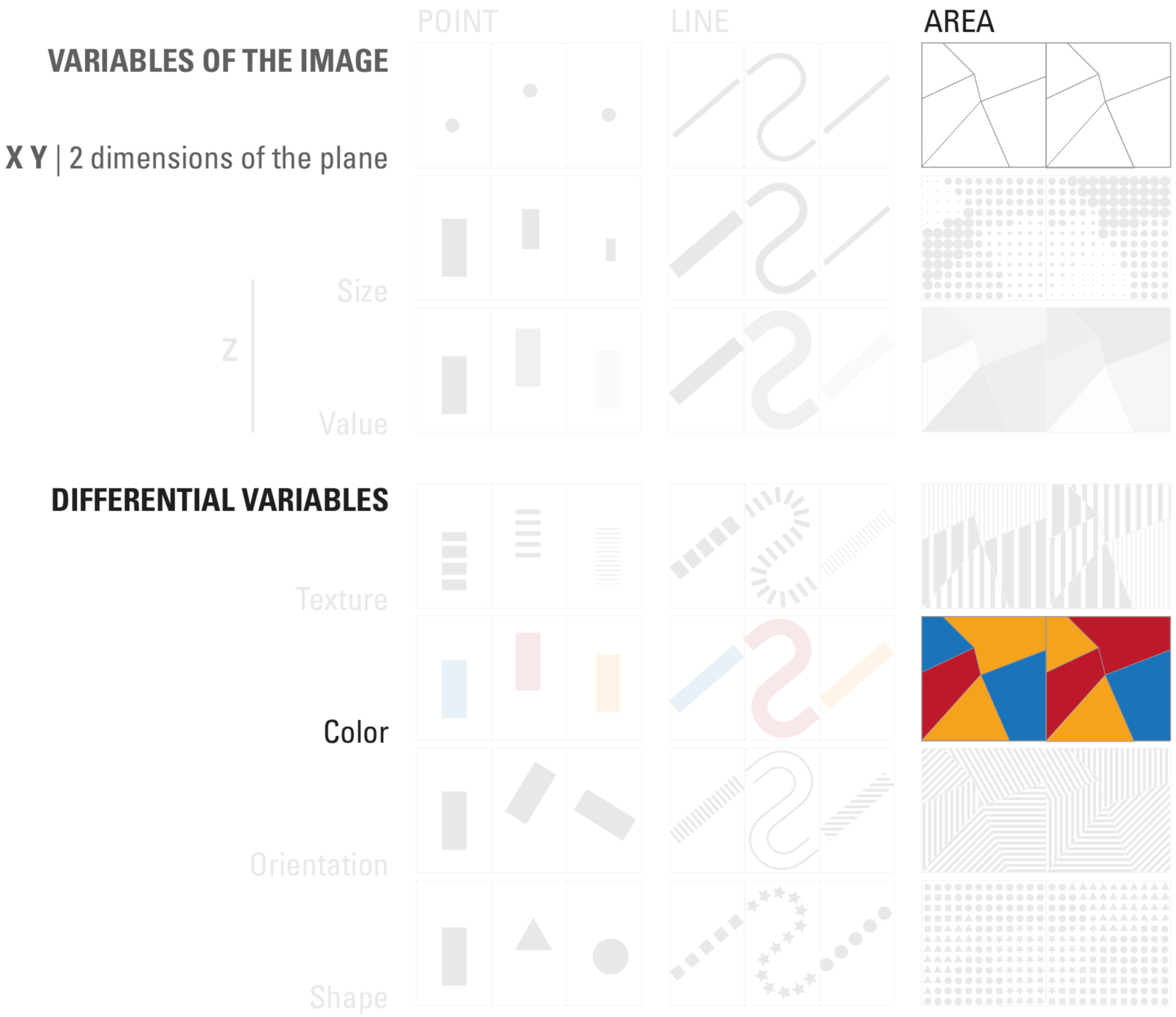
data encodings, visual channels for encoding data



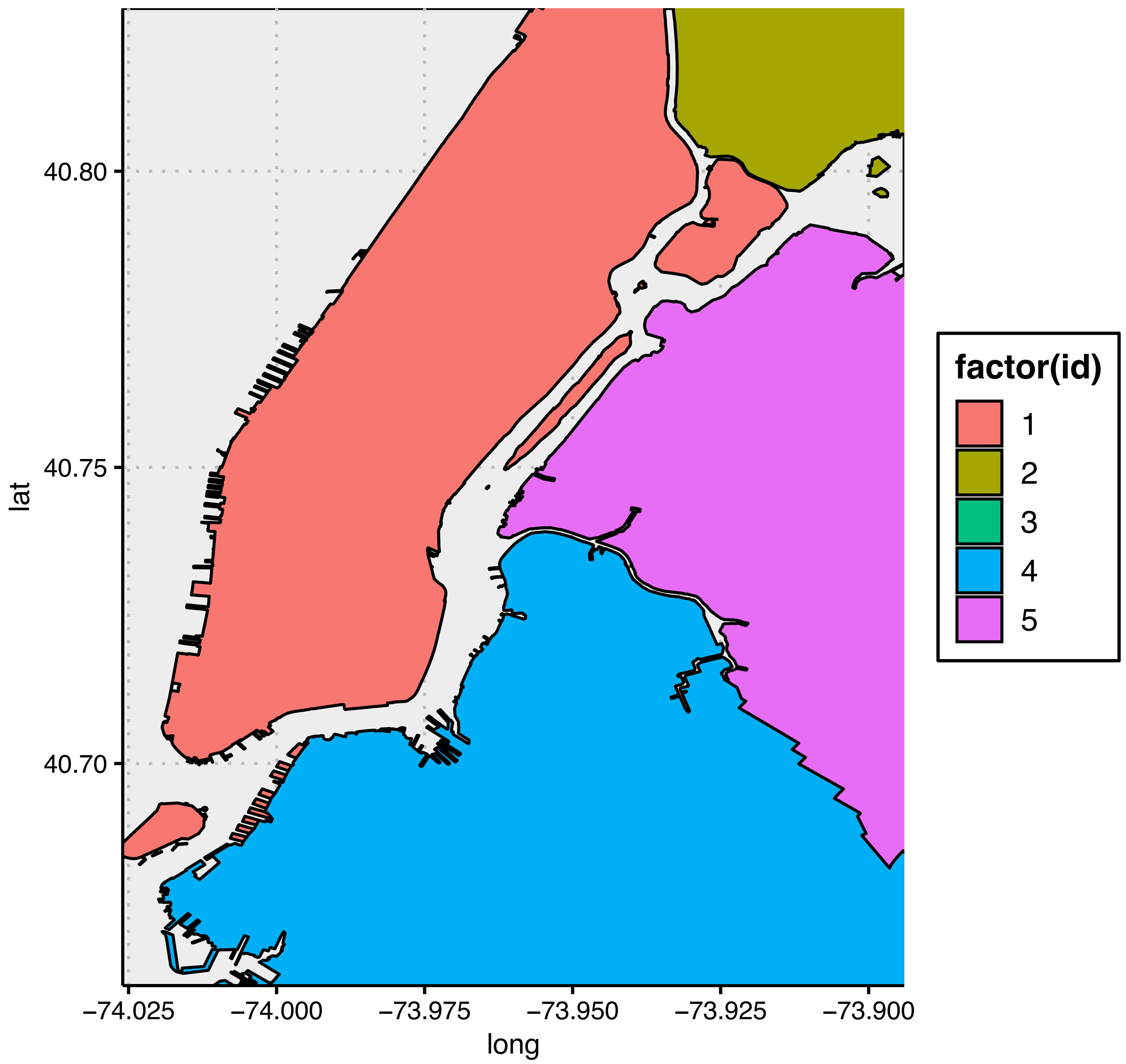
Citi Bike example — *exploratory data analysis*



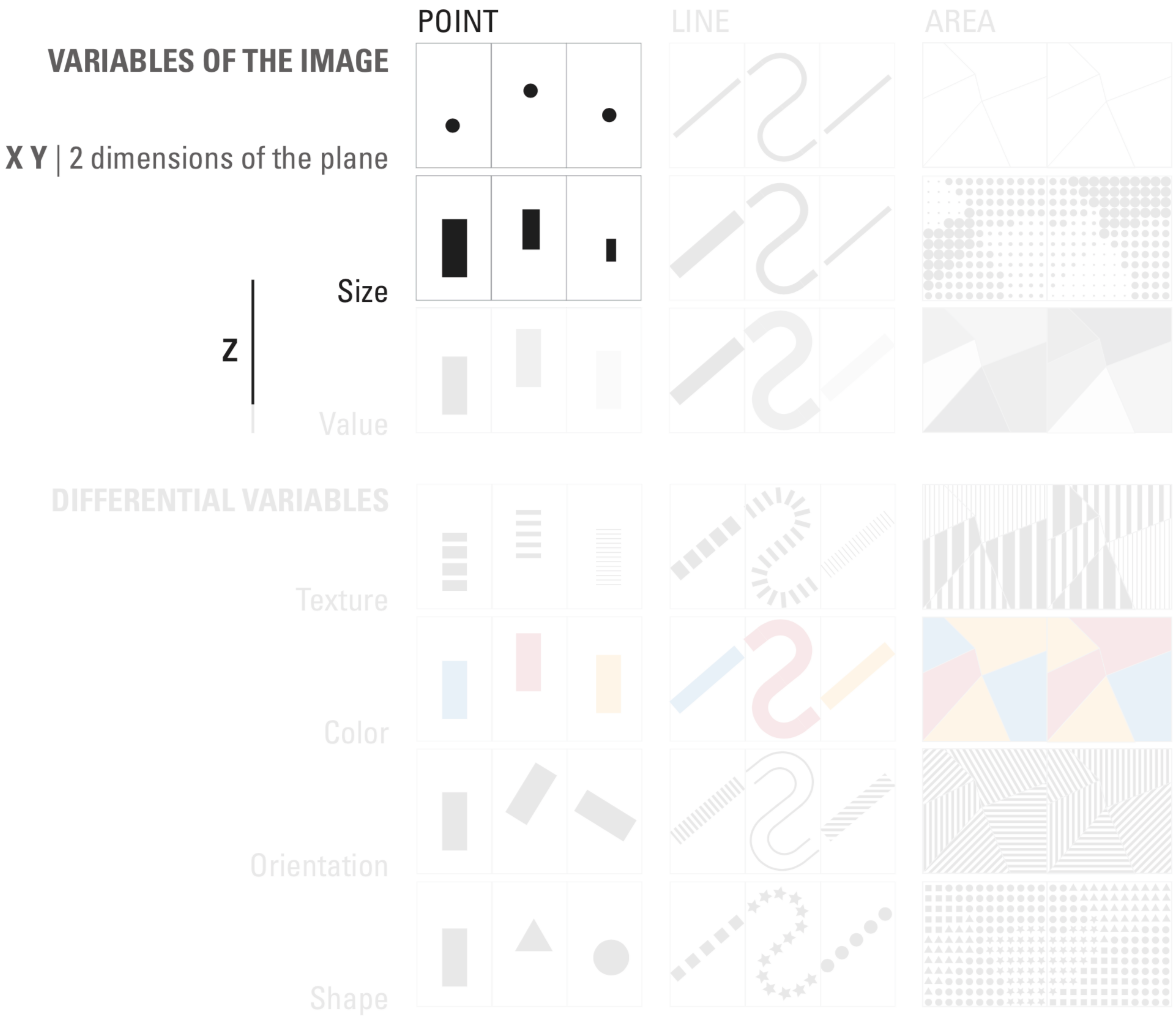
data encodings, visual channels for encoding data



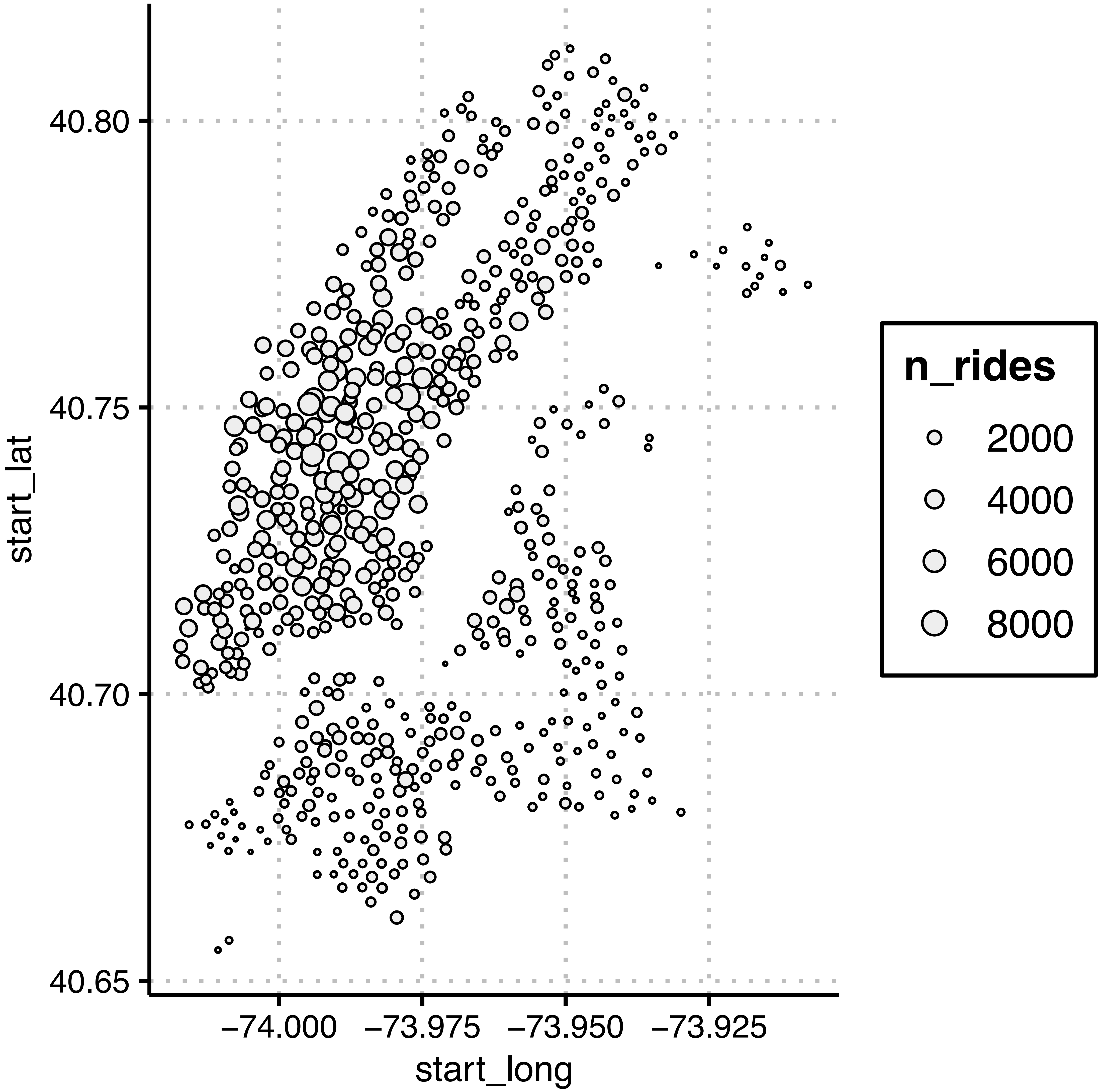
Citi Bike example — *exploratory data analysis*



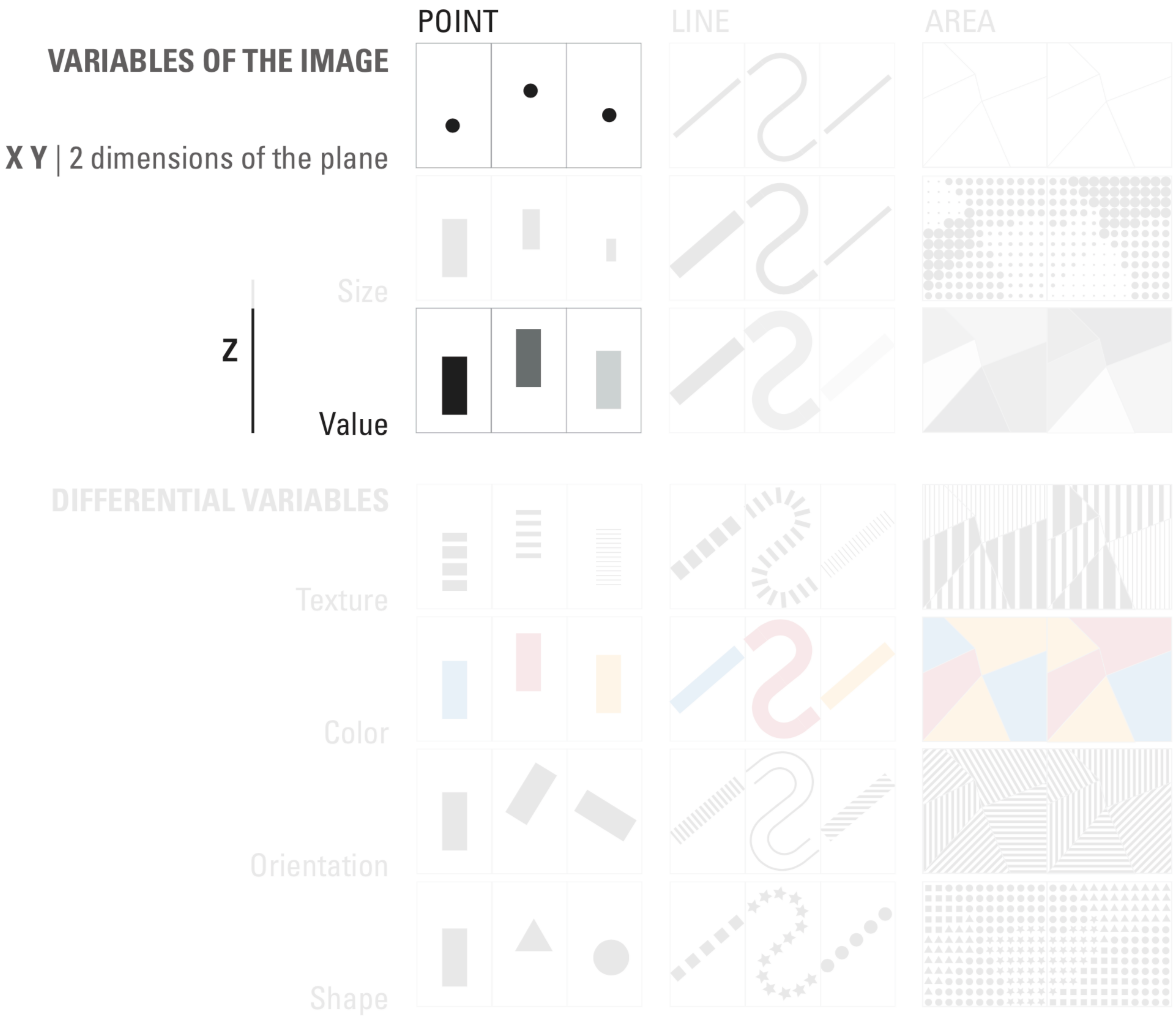
data encodings, visual channels for encoding data



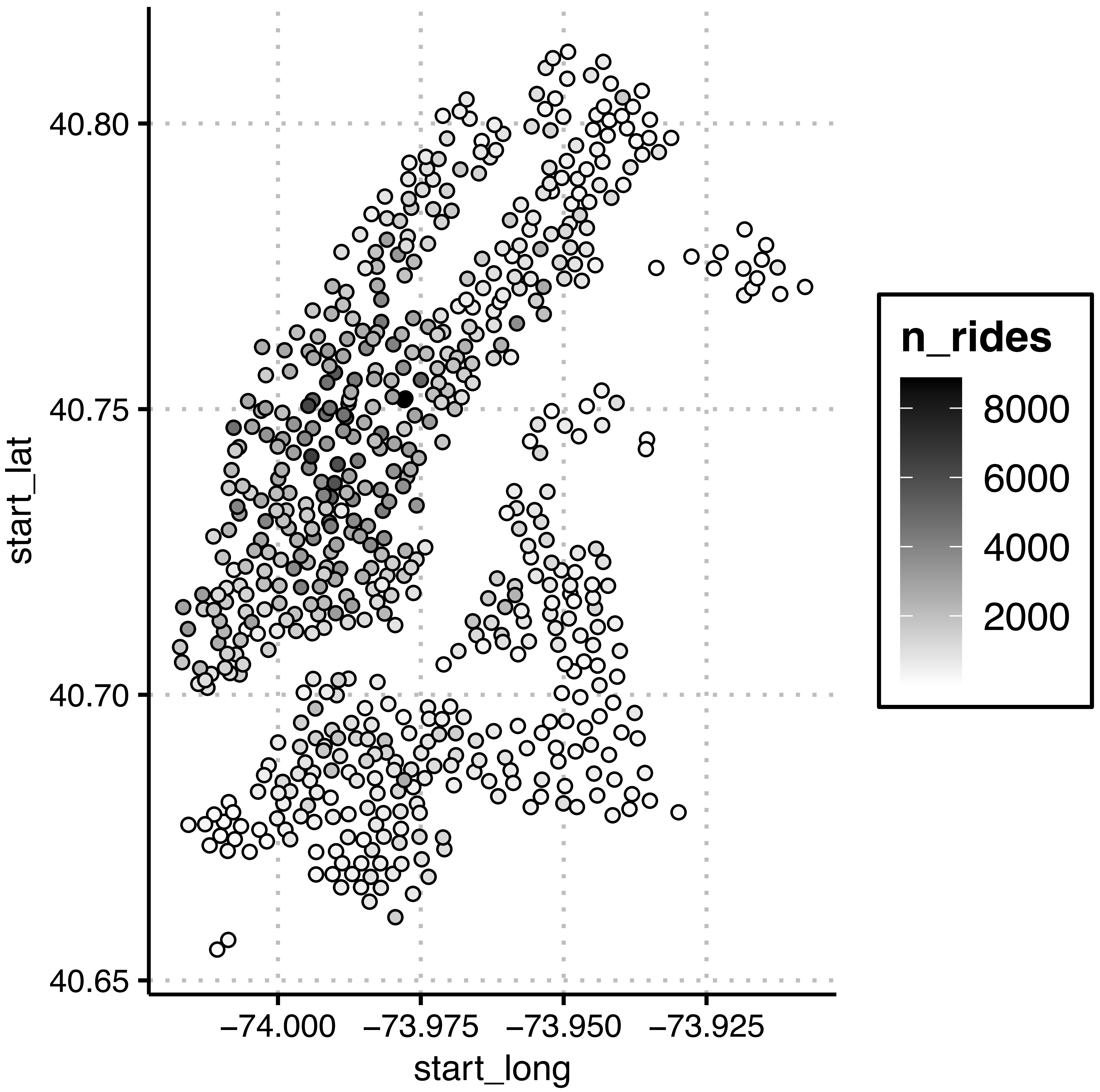
Citi Bike example — *exploratory data analysis*



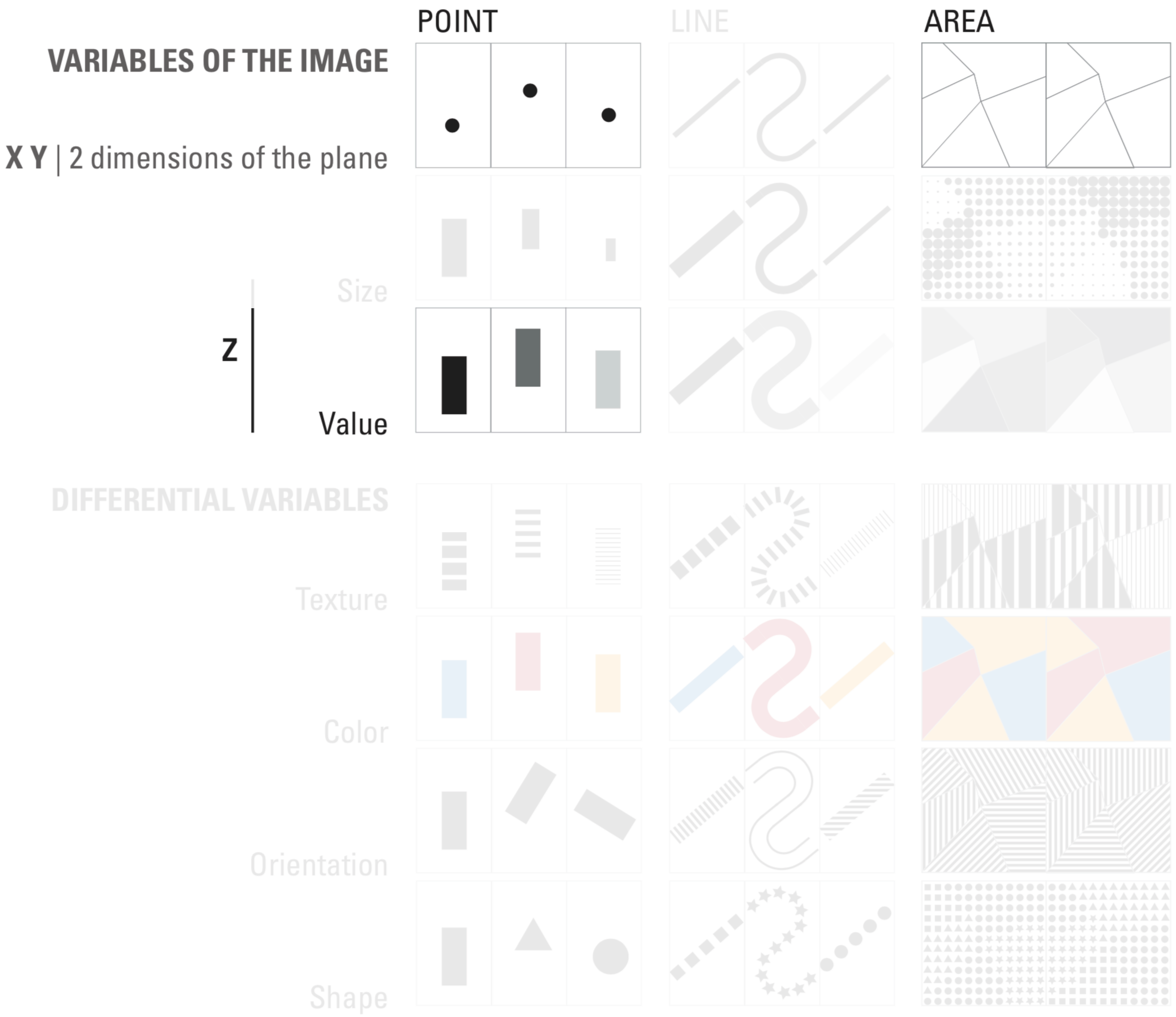
data encodings, visual channels for encoding data



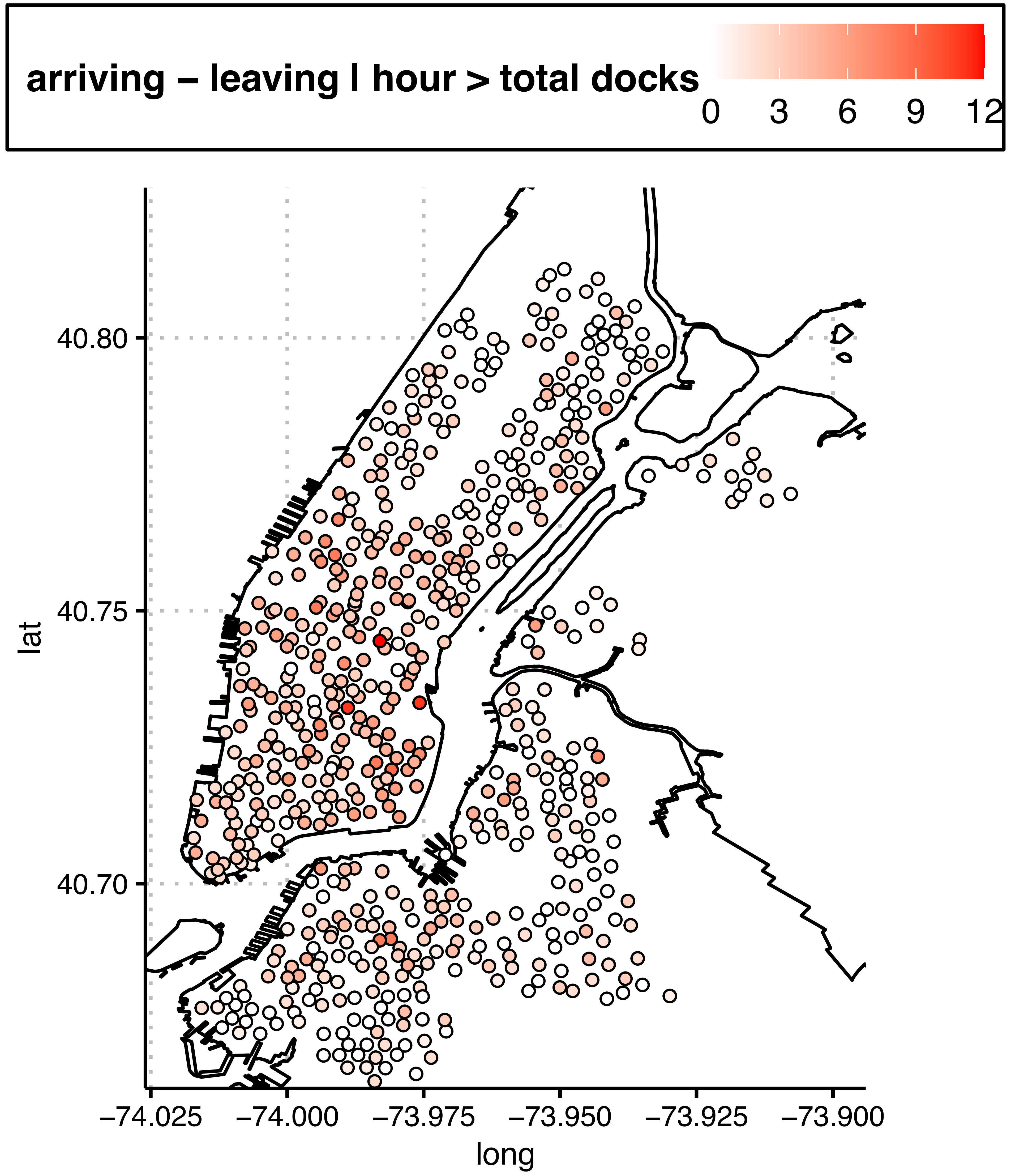
Citi Bike example — *exploratory data analysis*



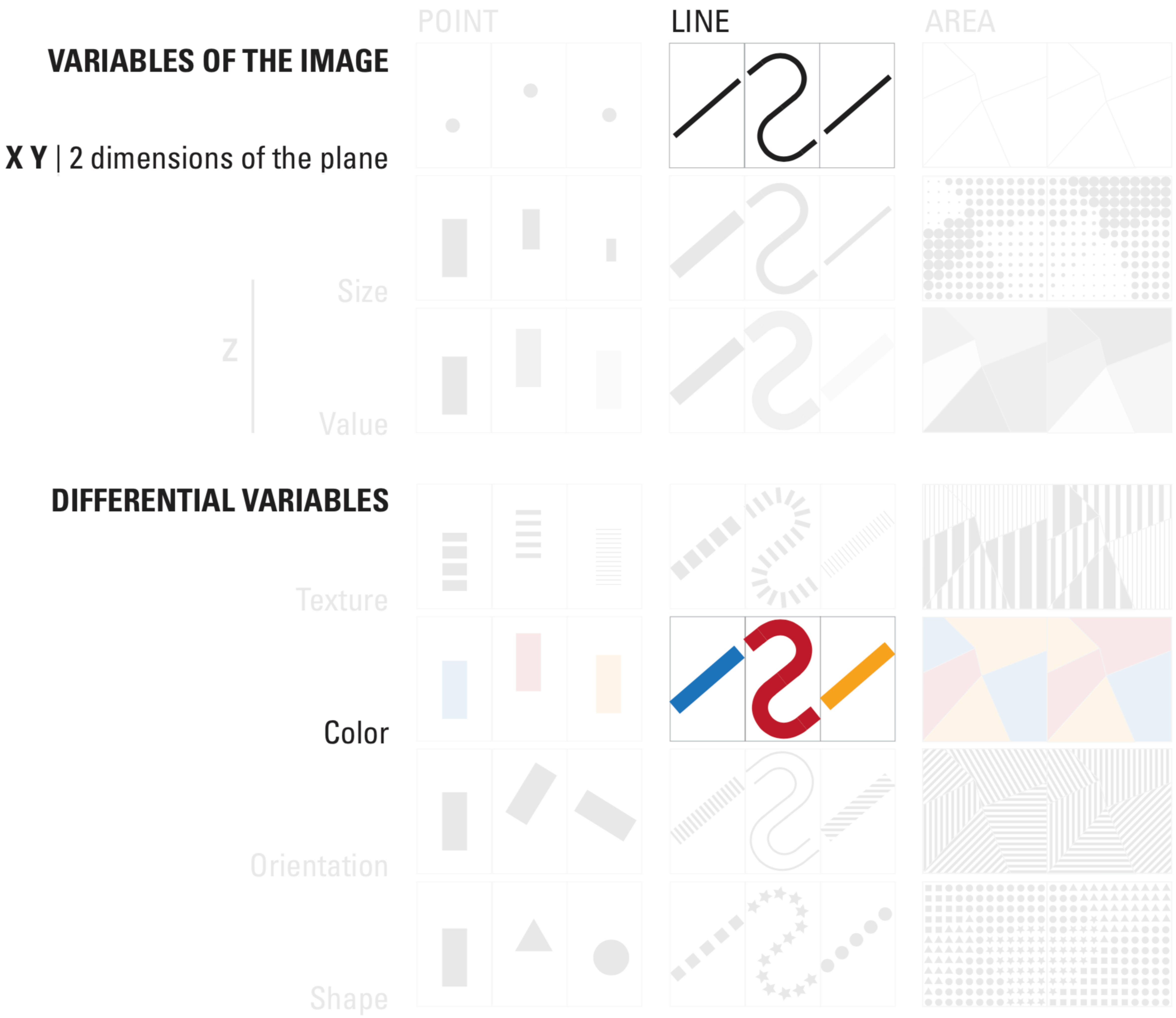
data encodings, visual channels for encoding data



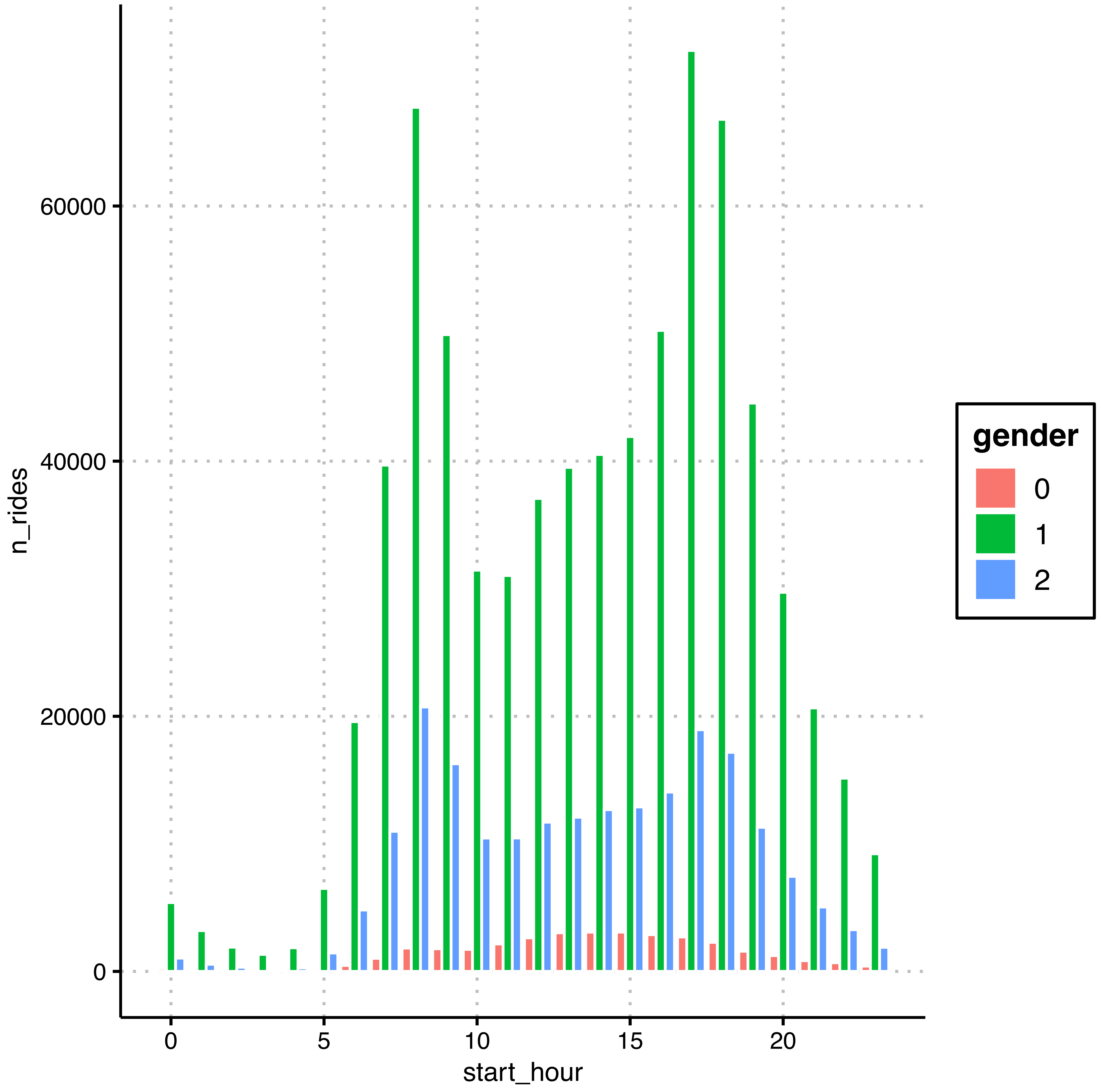
Citi Bike example — *exploratory data analysis*



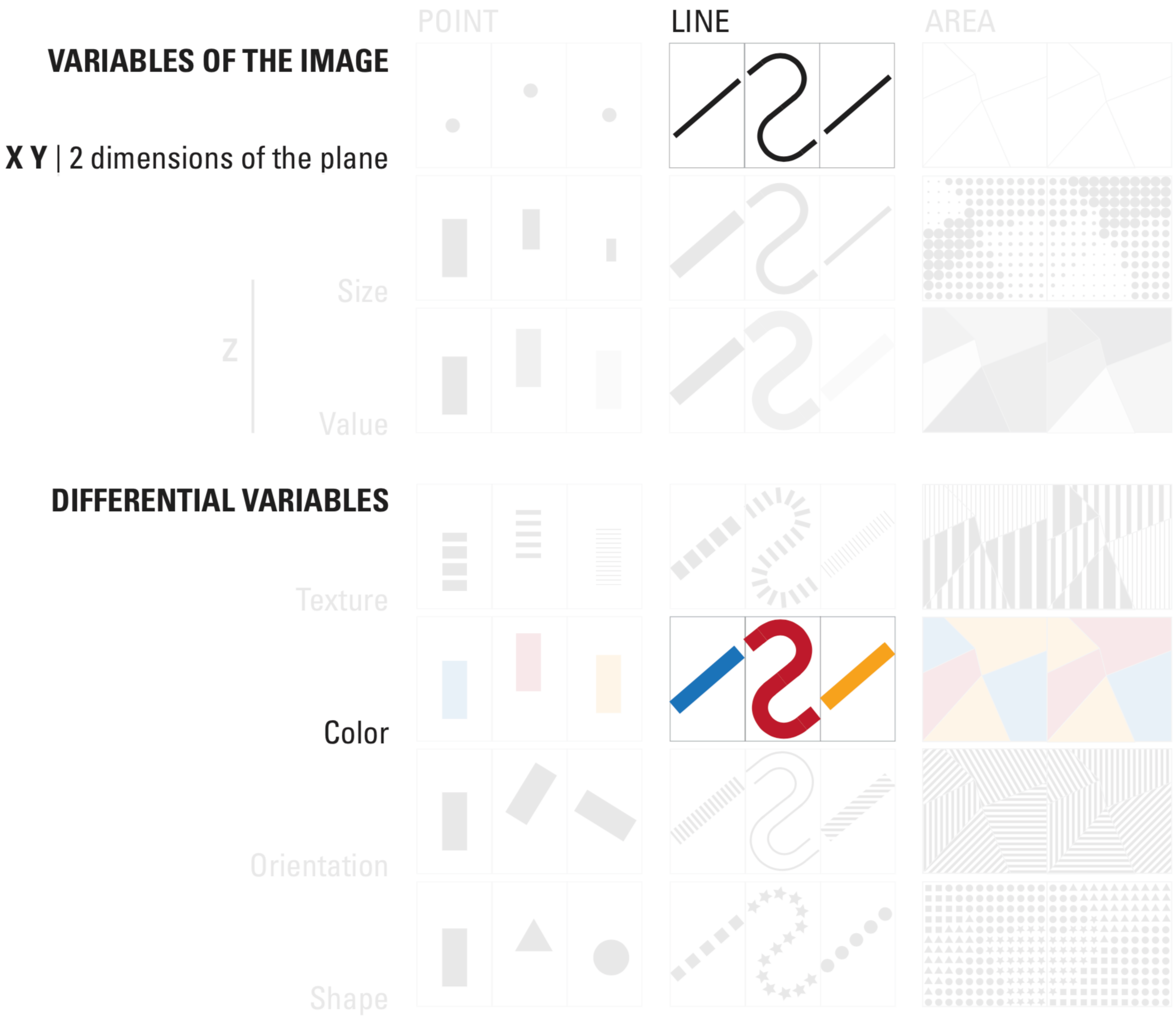
data encodings, visual channels for encoding data

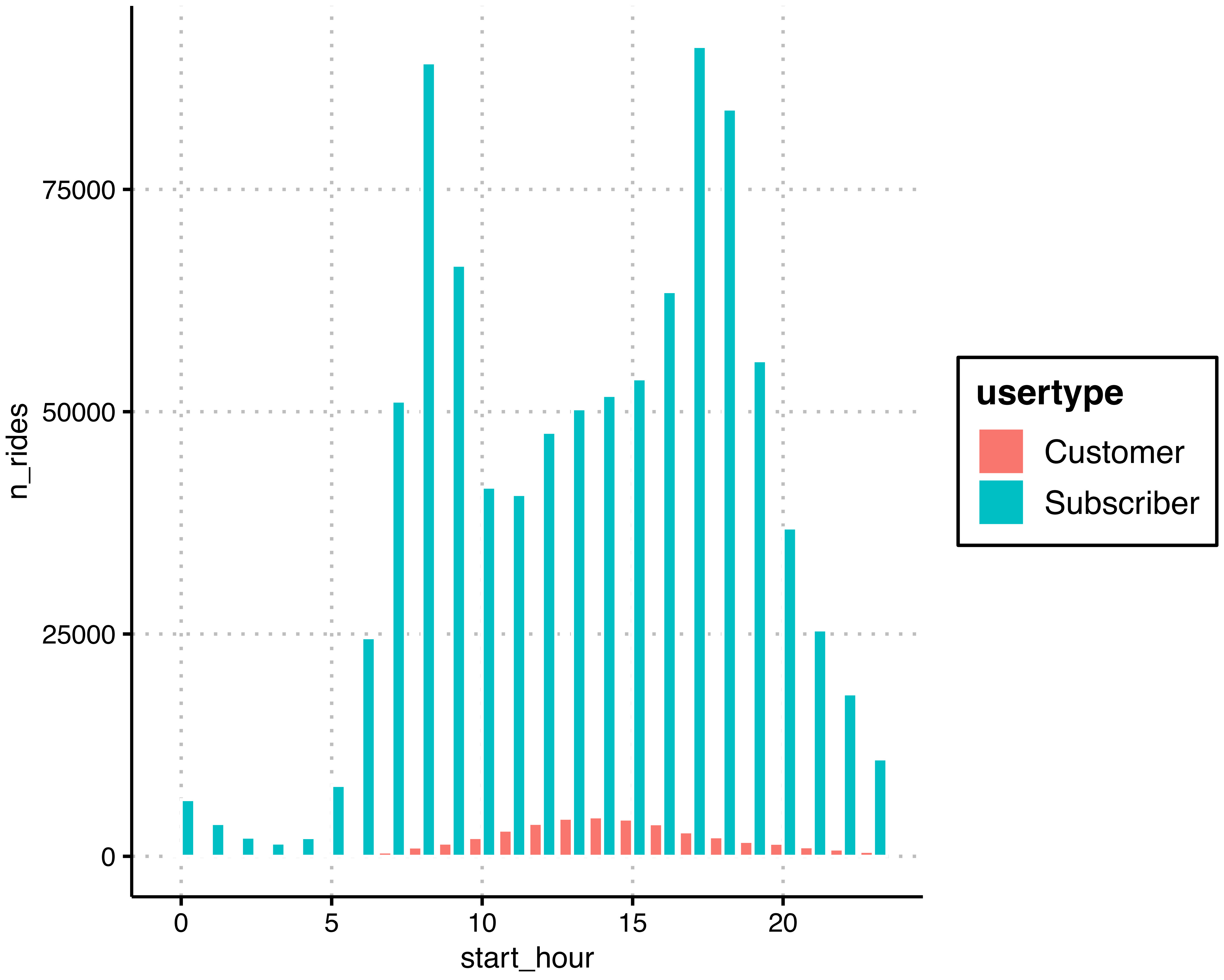


Citi Bike example — *exploratory data analysis*

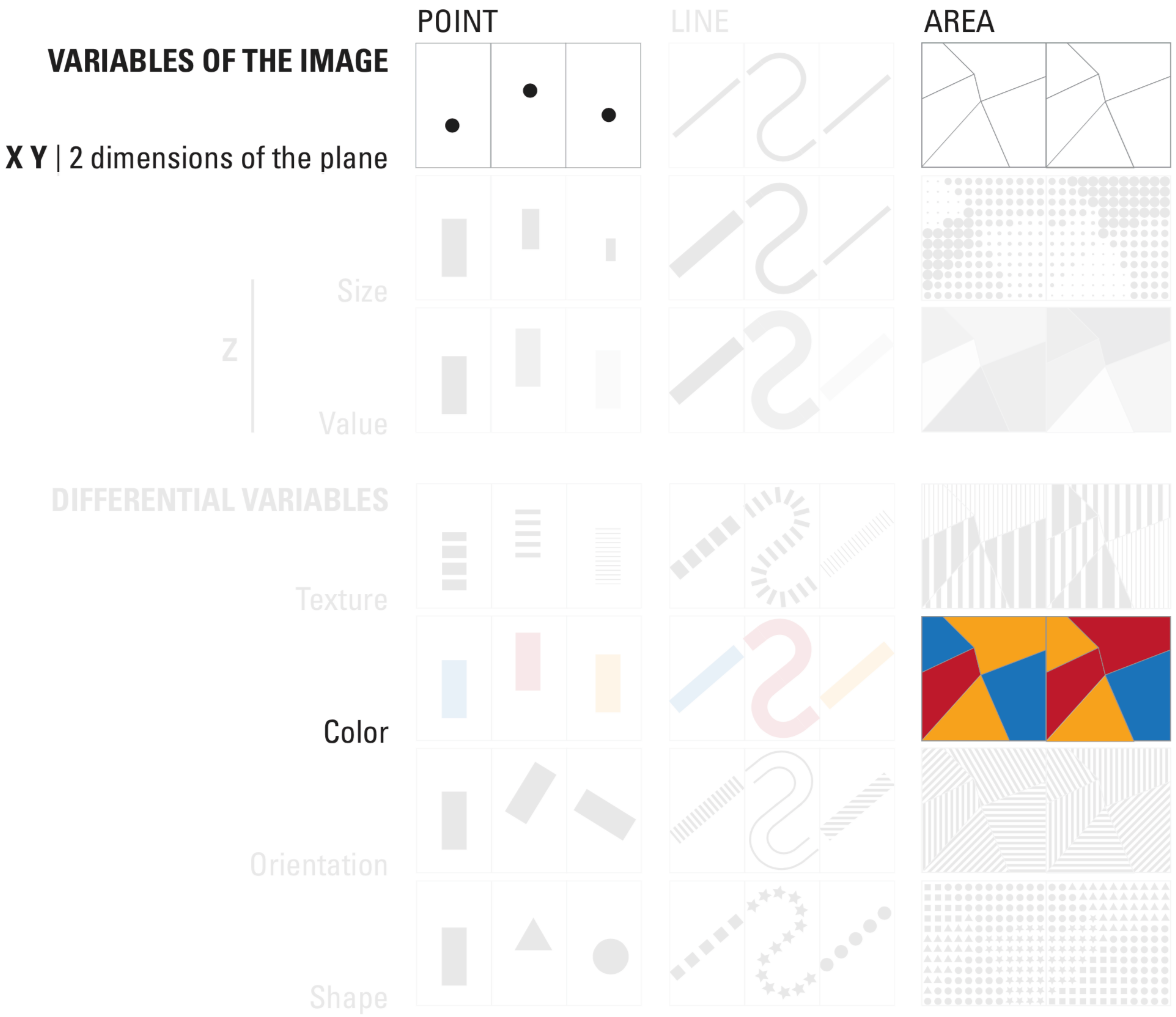


Citi Bike example — *exploratory data analysis*

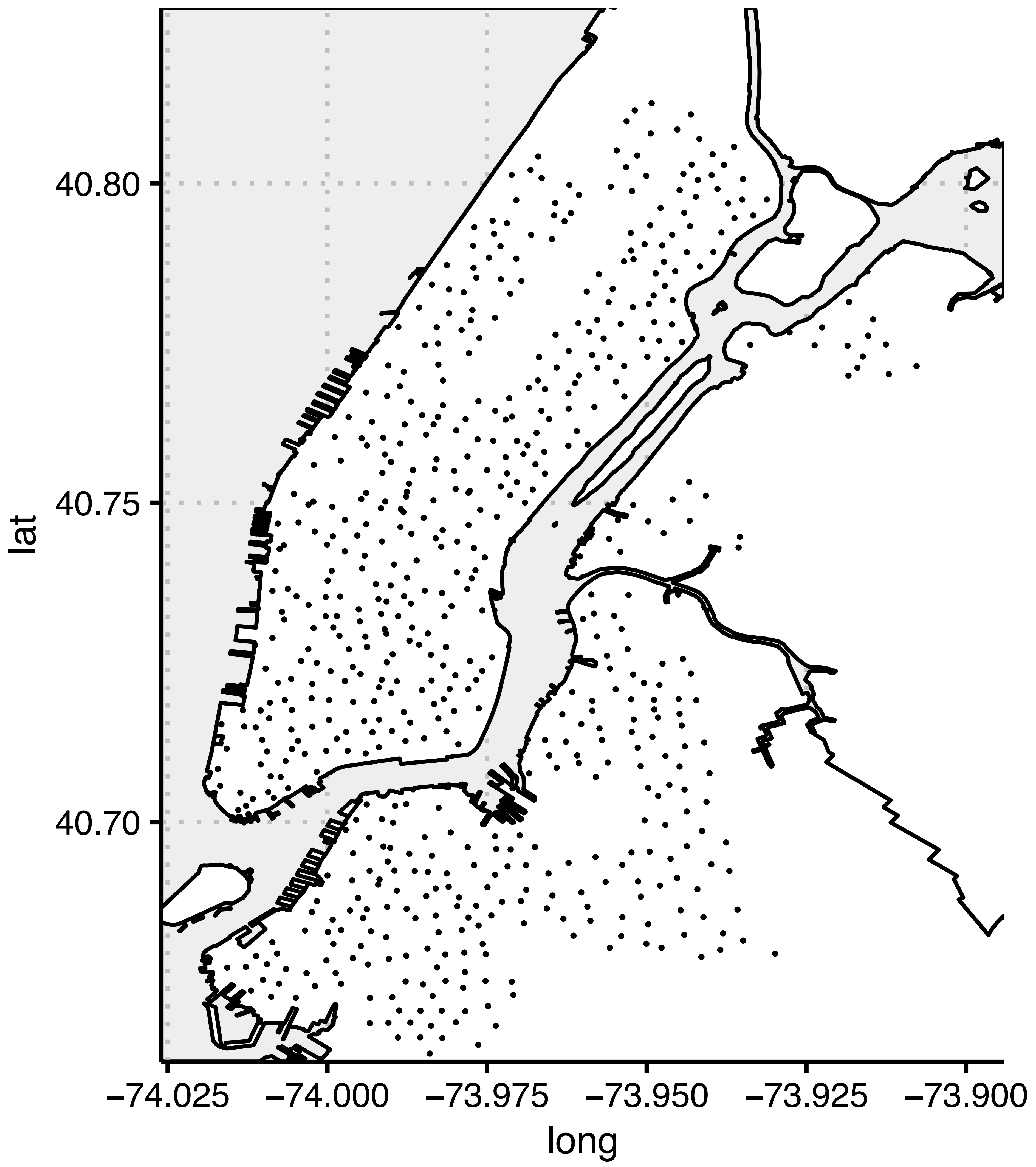




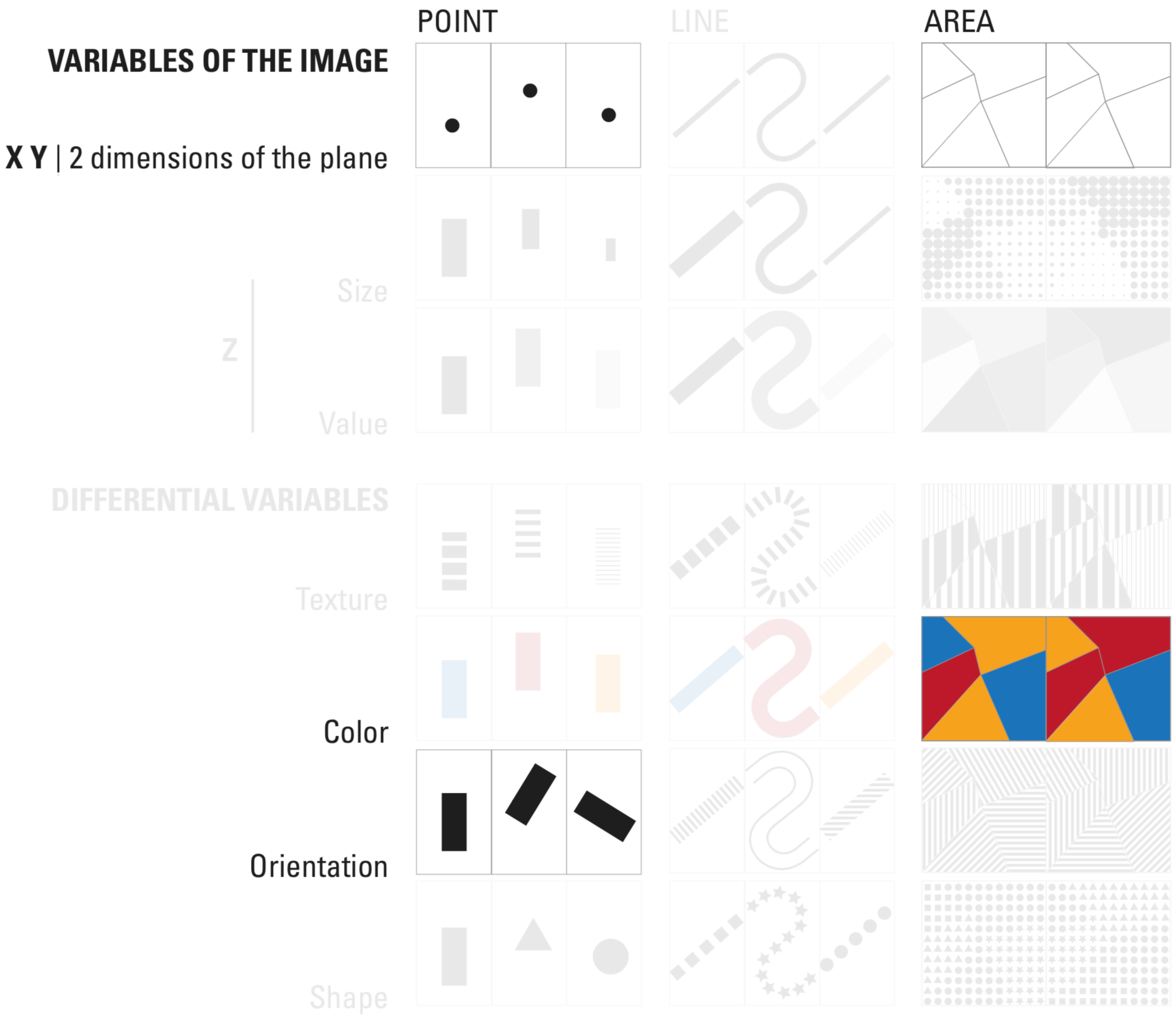
data encodings, visual channels for encoding data



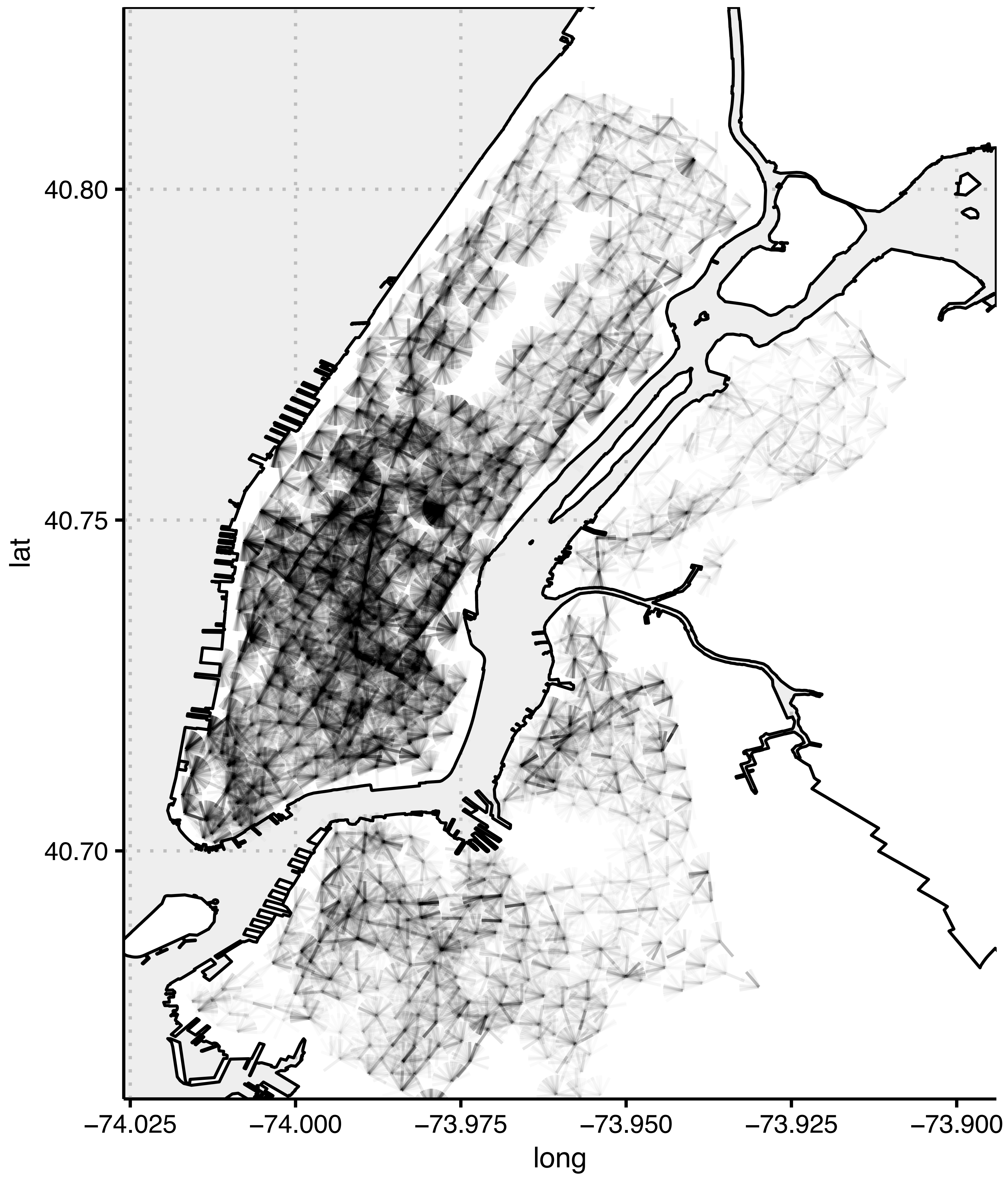
Citi Bike example — *exploratory data analysis*



data encodings, visual channels for encoding data

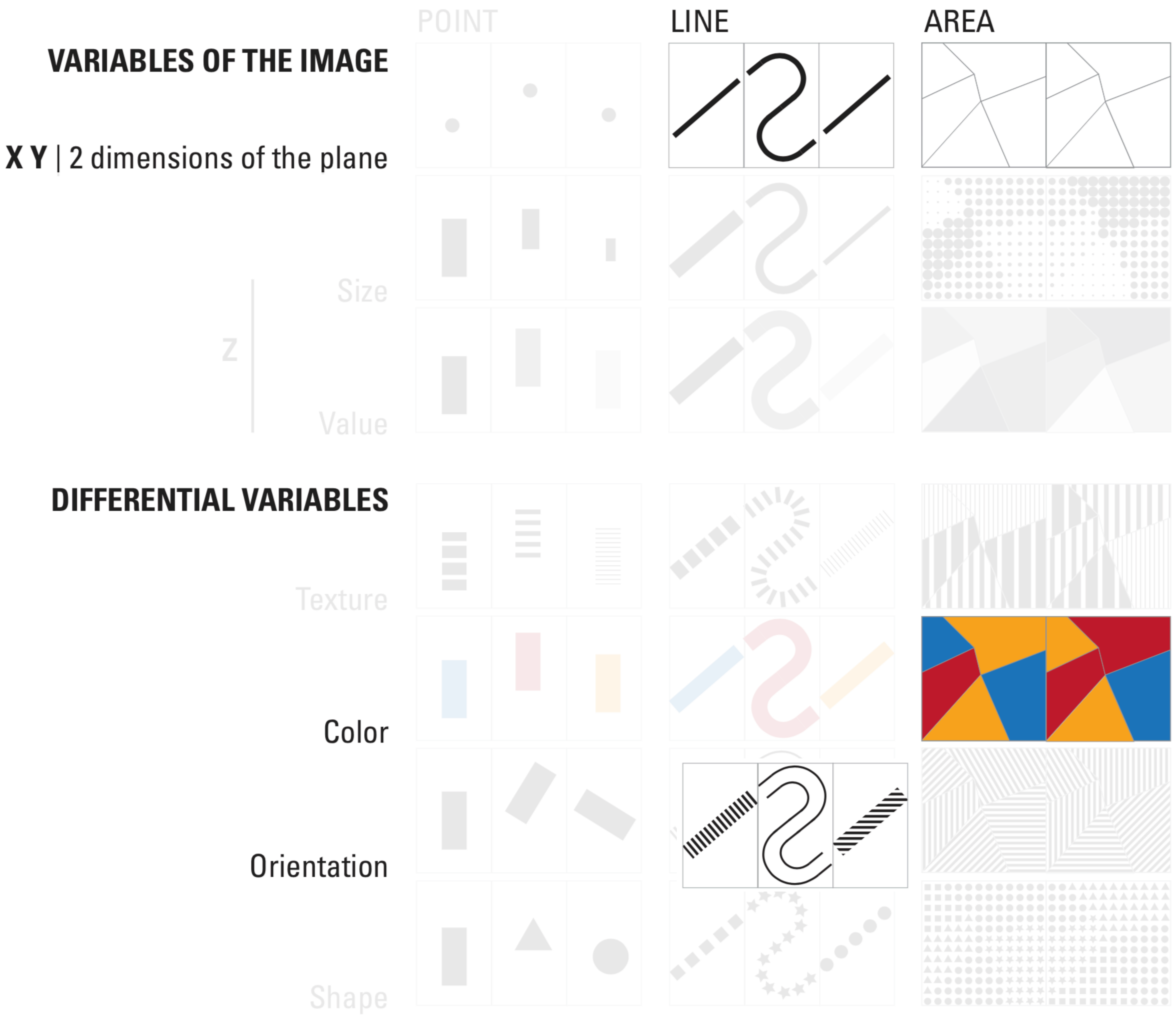


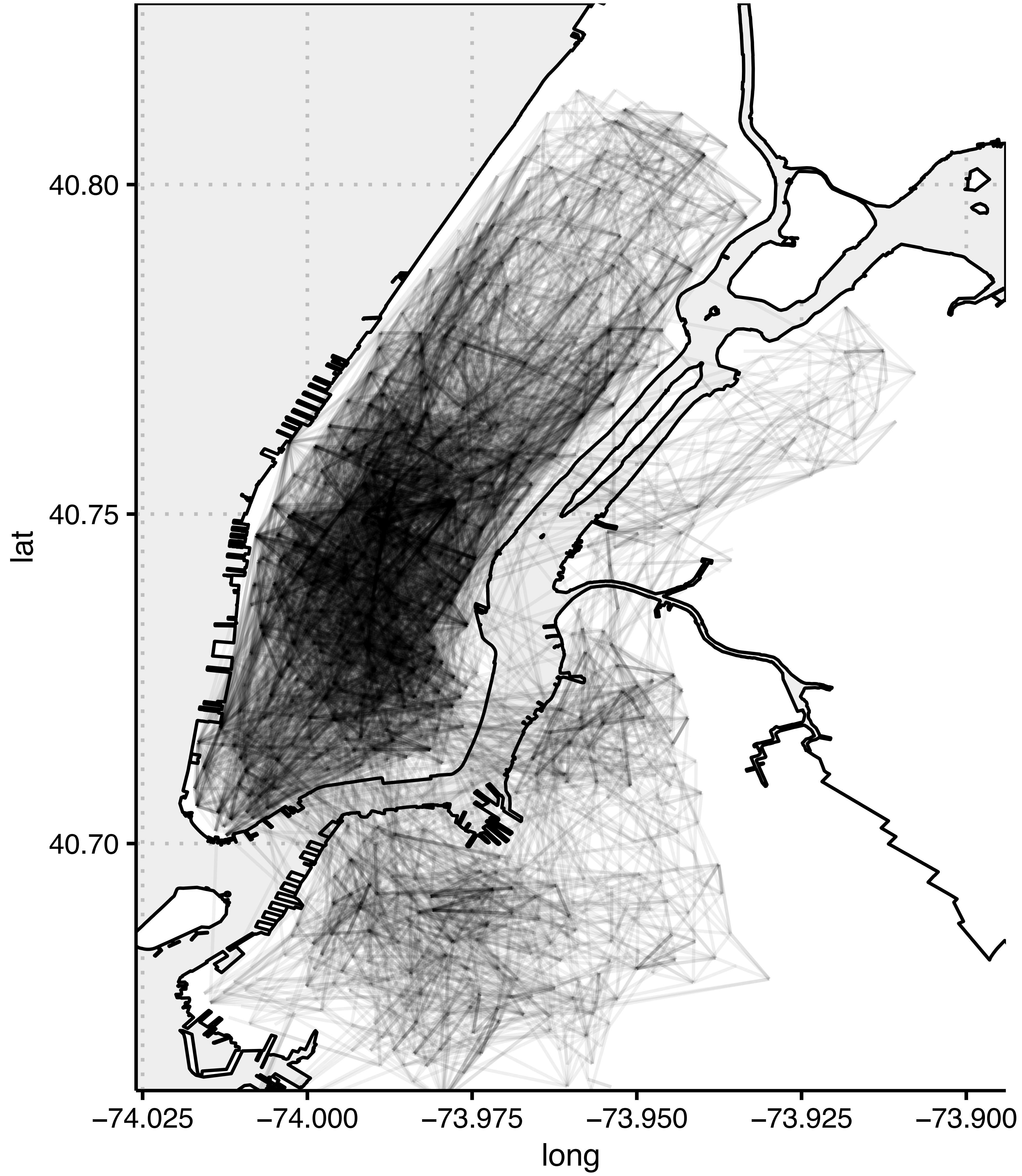
Citi Bike example — *exploratory data analysis*



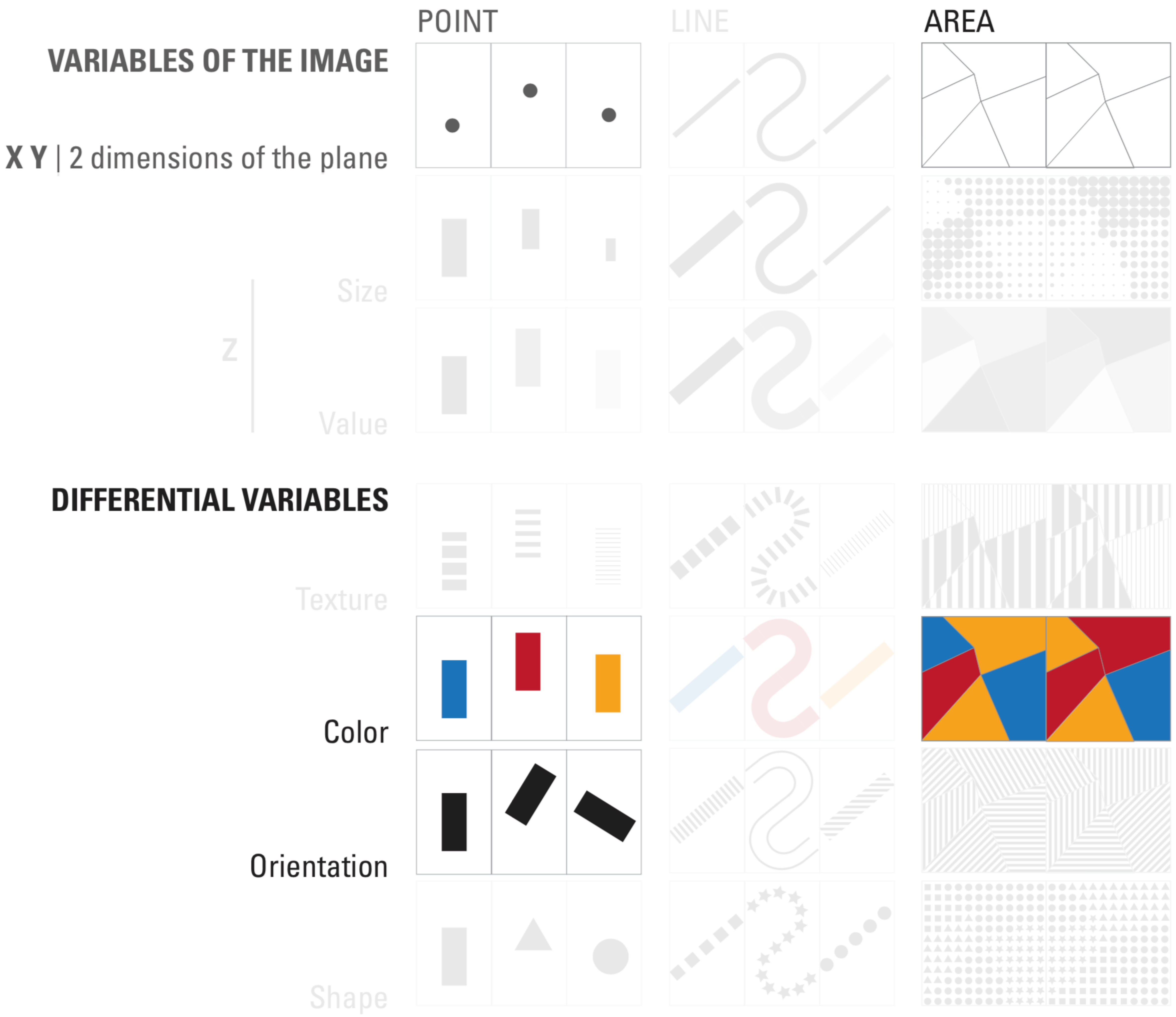
data encodings, **visual channels for encoding data**

Citi Bike example — *exploratory data analysis*

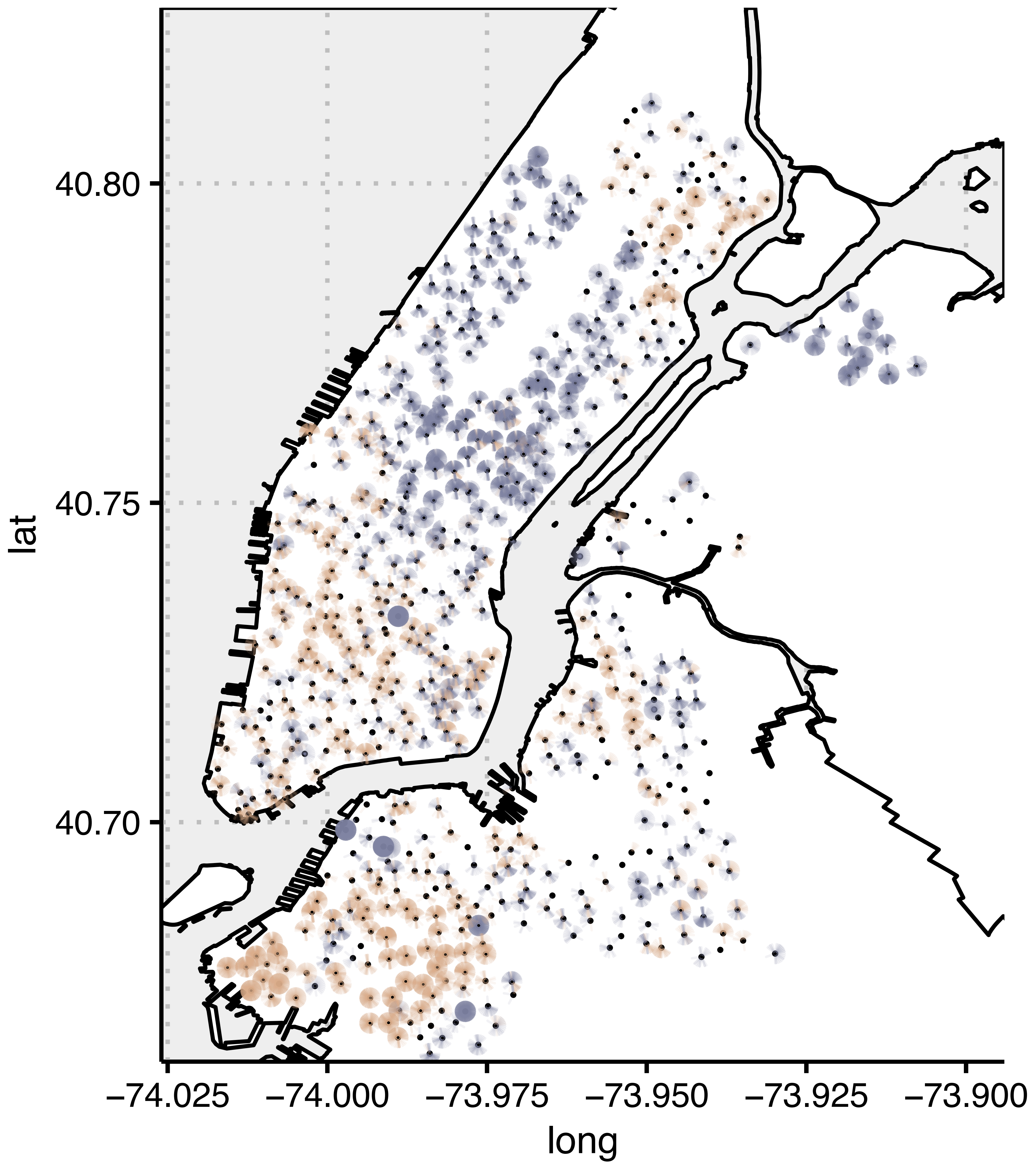




data encodings, visual channels for encoding data

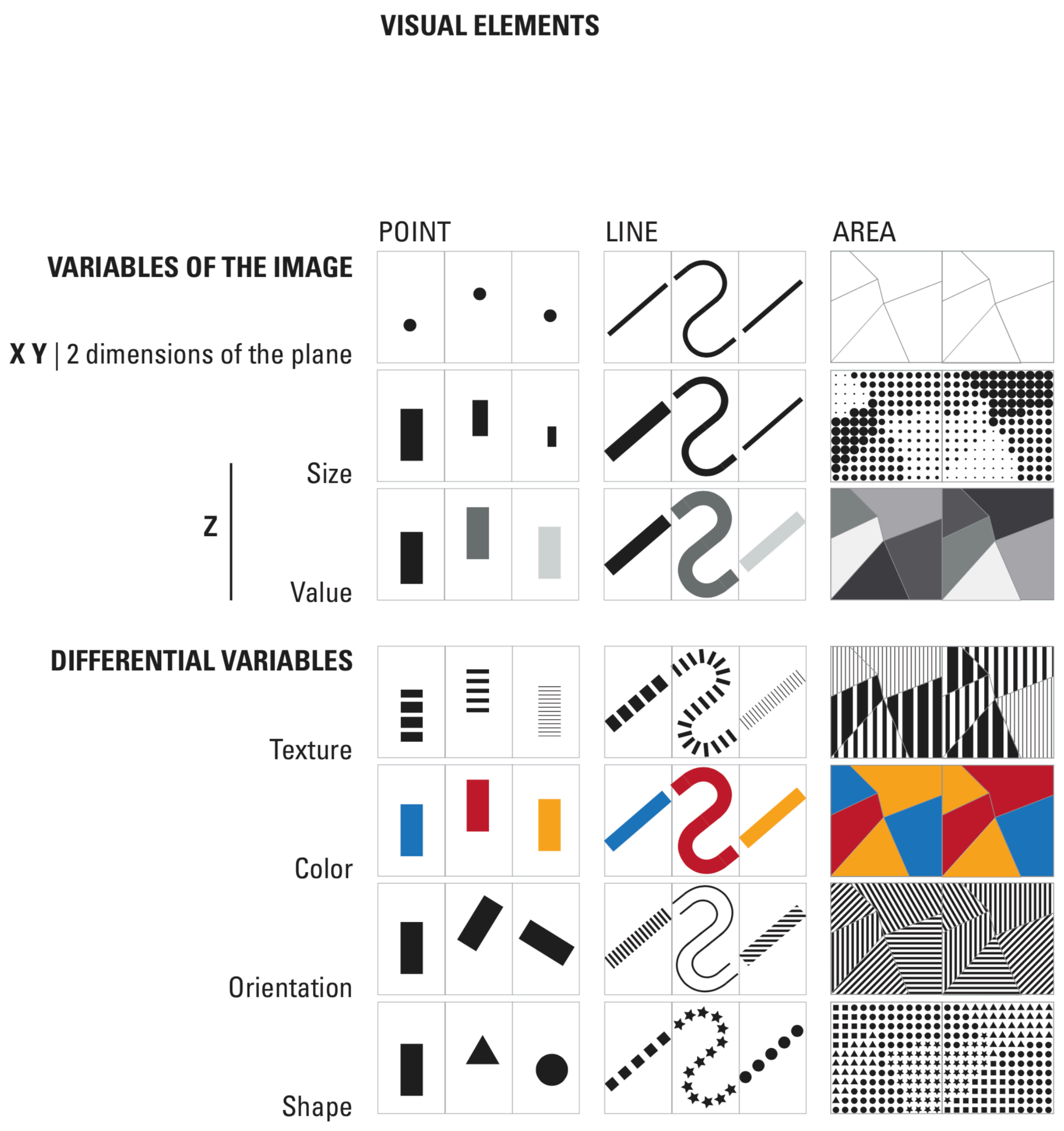


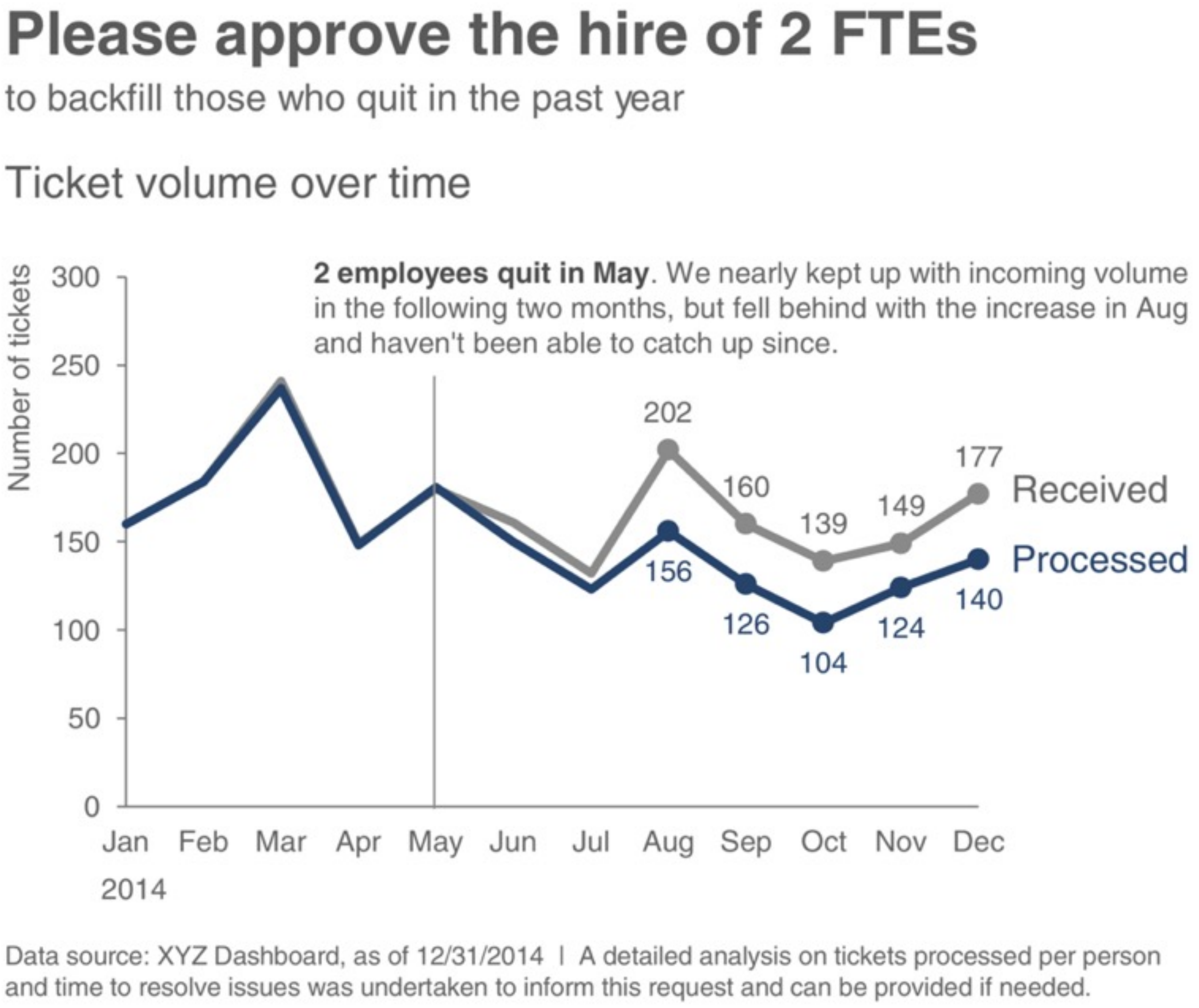
Citi Bike example — *exploratory data analysis*



class exercises

exercise, identify data encodings in visual channels

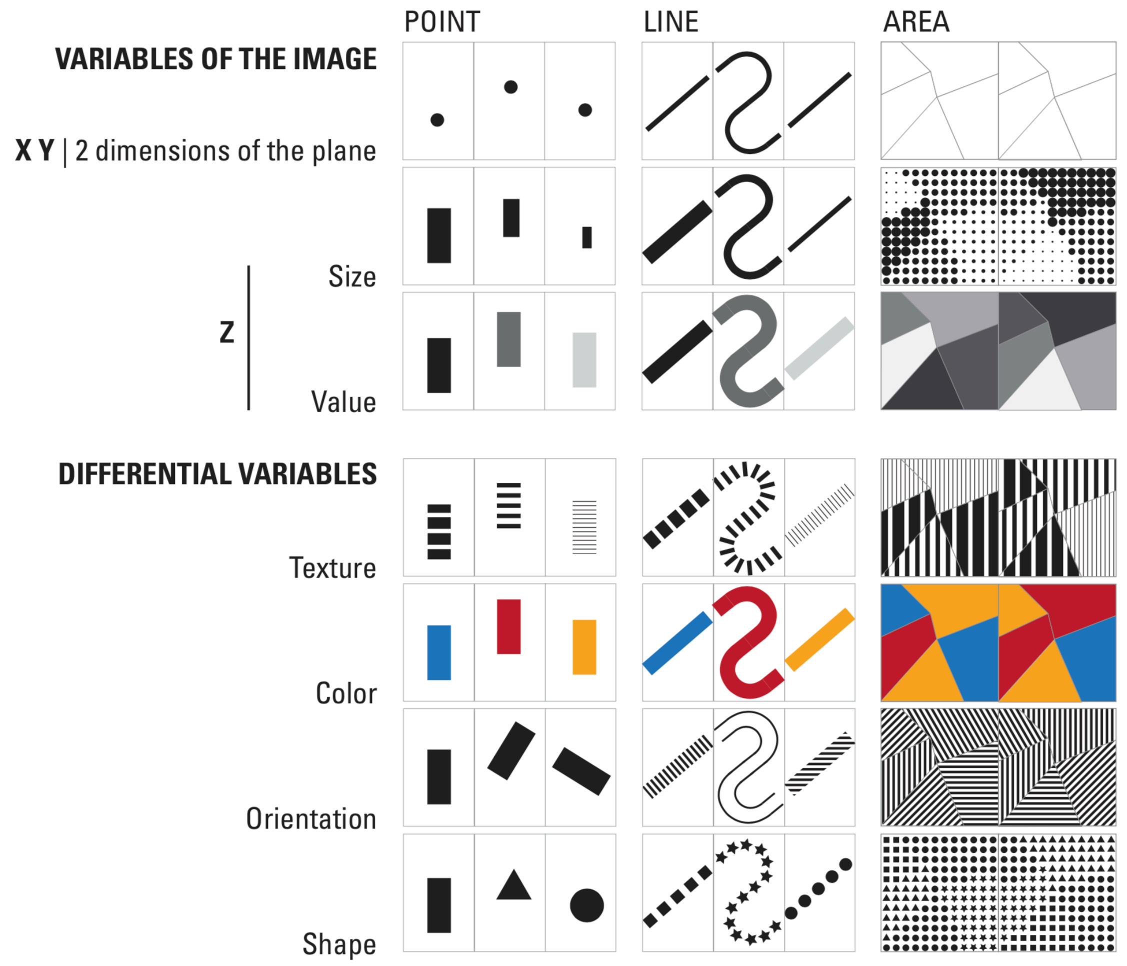


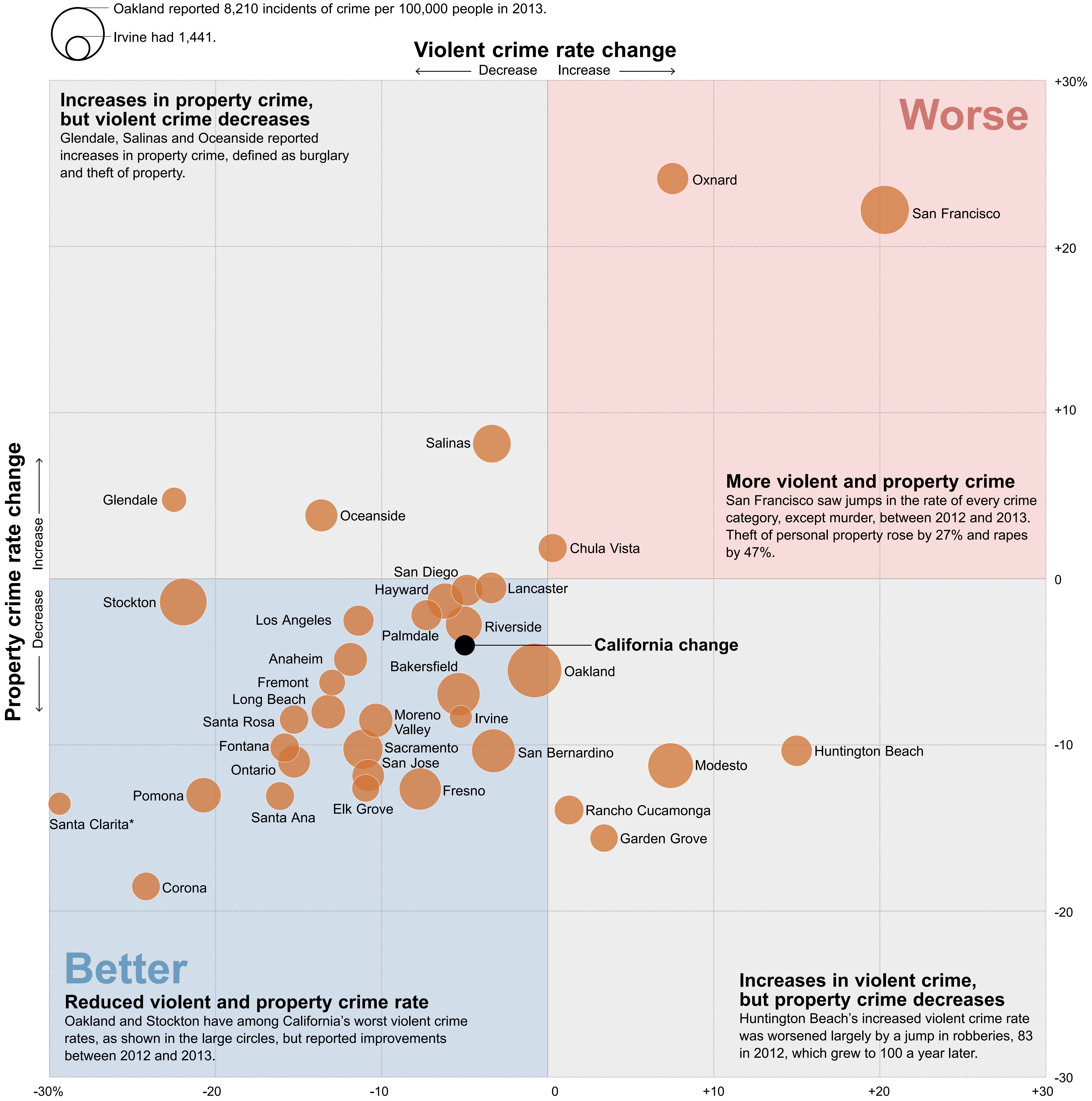


— Knaflic, Cole Nussbaumer. *Storytelling with Data. A Data Visualization Guide for Business Professionals*. Wiley, 2015.

exercise, **identify data encodings in visual channels**

VISUAL ELEMENTS





Schleuss, Jon, and Rong-Cong Lin II. 2013. "*California Crime 2013.*" Los Angeles Times.

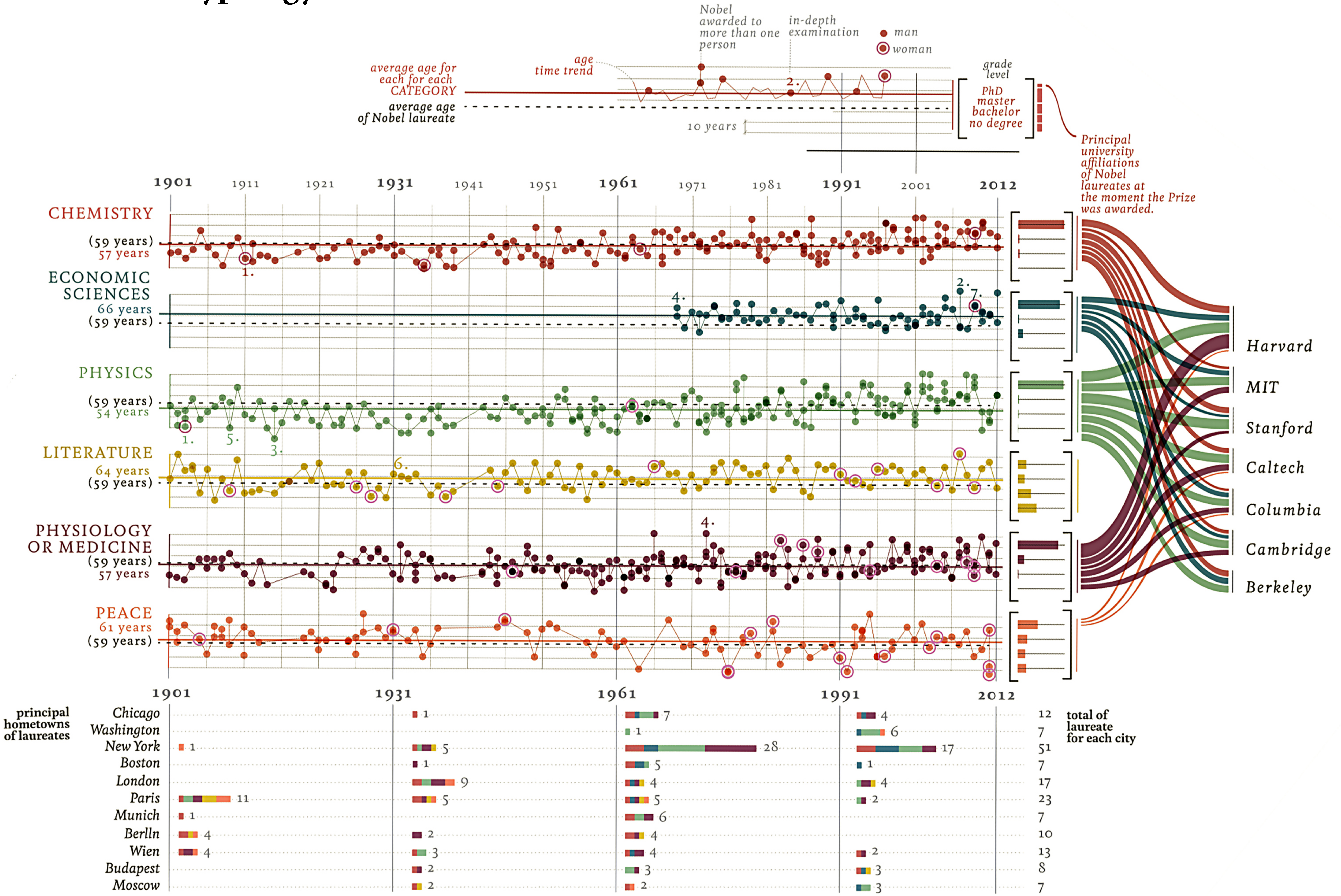
a graphics study — deconstructing Lupi's *Nobels, no degrees*, identifying typologies

a graphics study through deconstruction and typology

If at first, this seems complex, Lupi's graphic is just organized groups of layered data encodings. These even follow typologies commonly used in business communications. We can make something complex like this by creating component parts and carefully arranging them.

Don't be intimidated! — Just methodically experiment with encodings for each data type, then organize them.

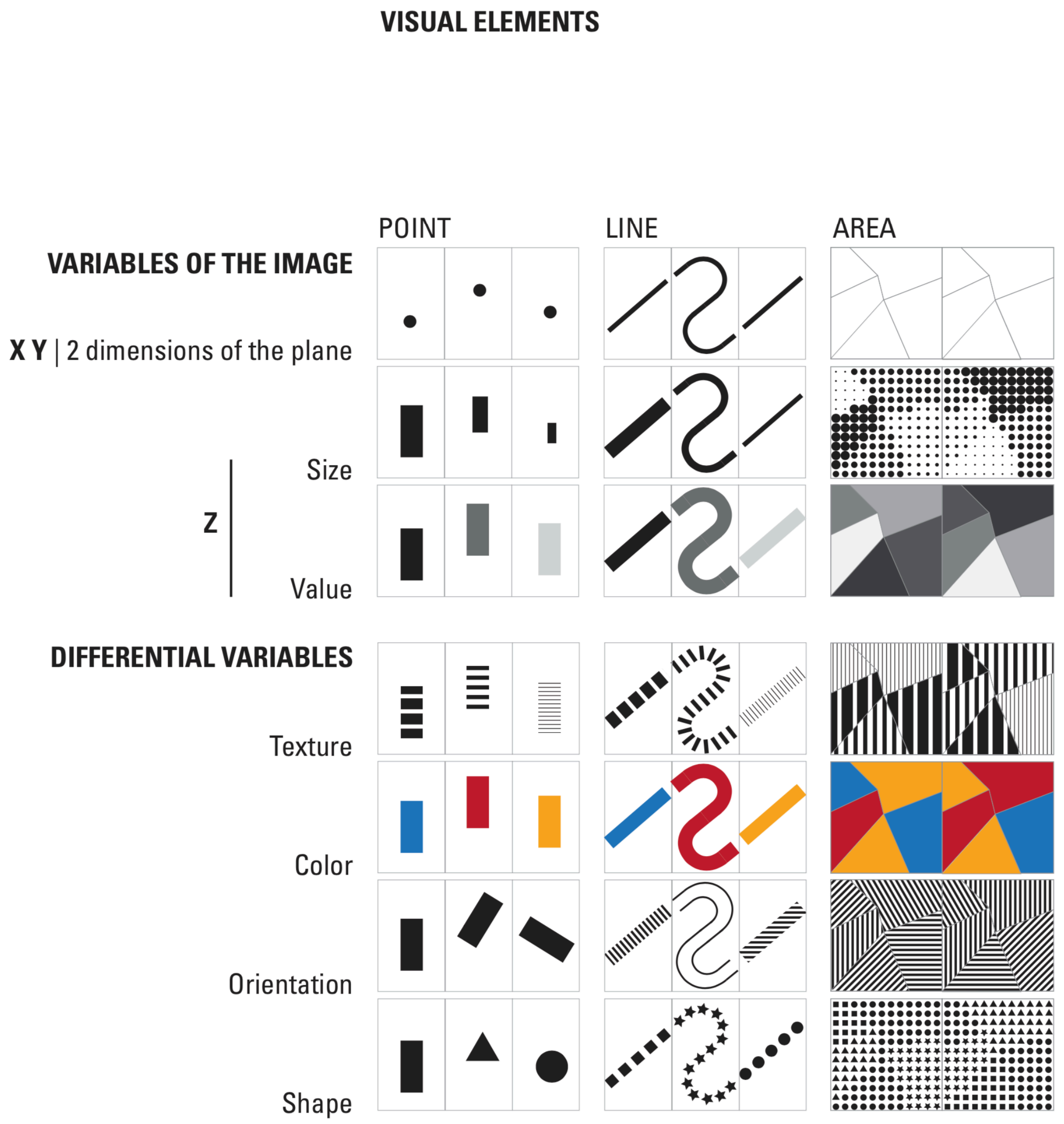
Of note: in Lupi's organization, she aligns graphics by common axis scales. We'll discuss this idea more later.

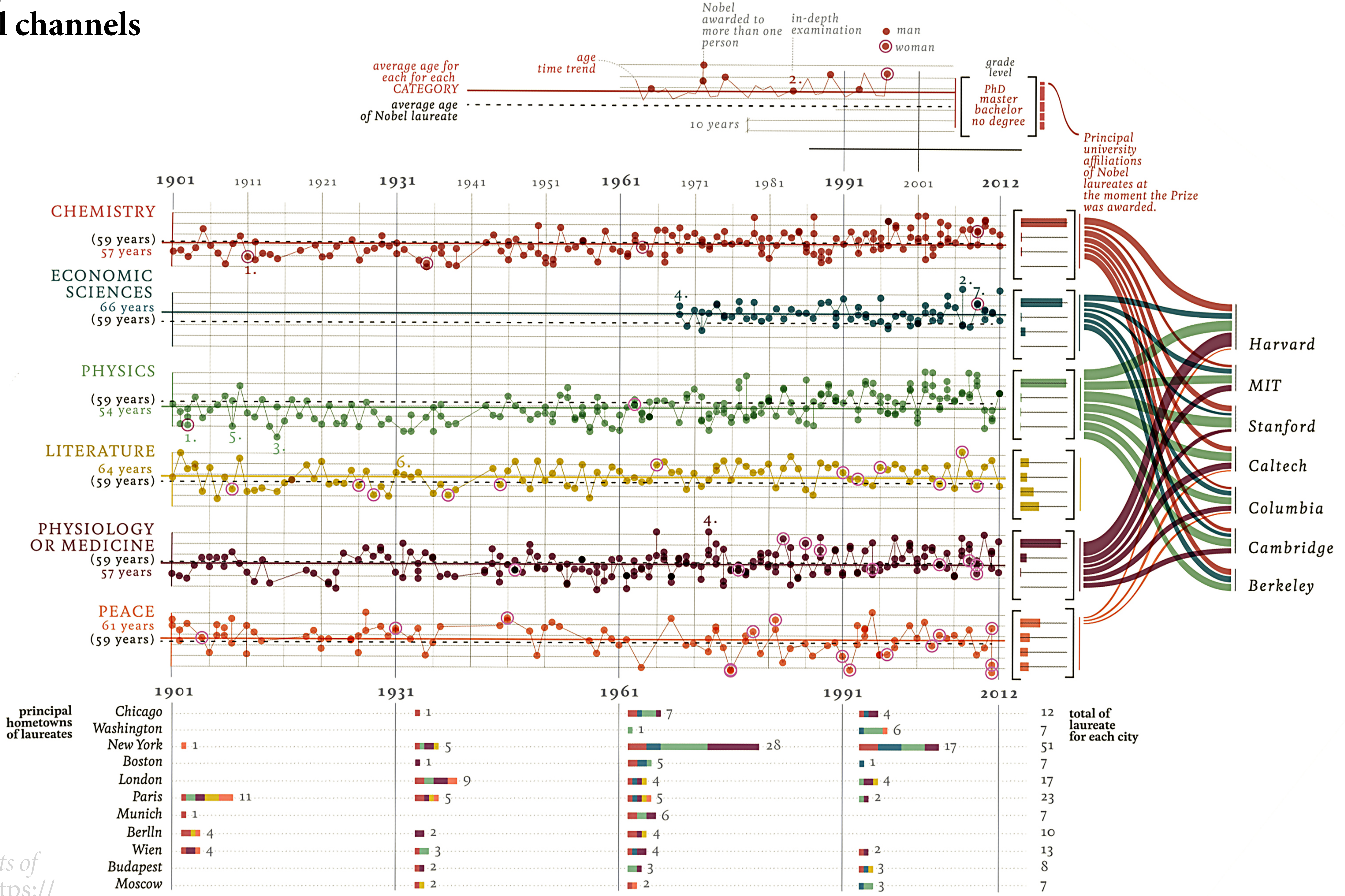


Lupi, Giorgia. 2016. *Visual Data - La Lettura.* http://giorgialupi.com/lalettura

decomposition & topology study,

identify data encodings in visual channels

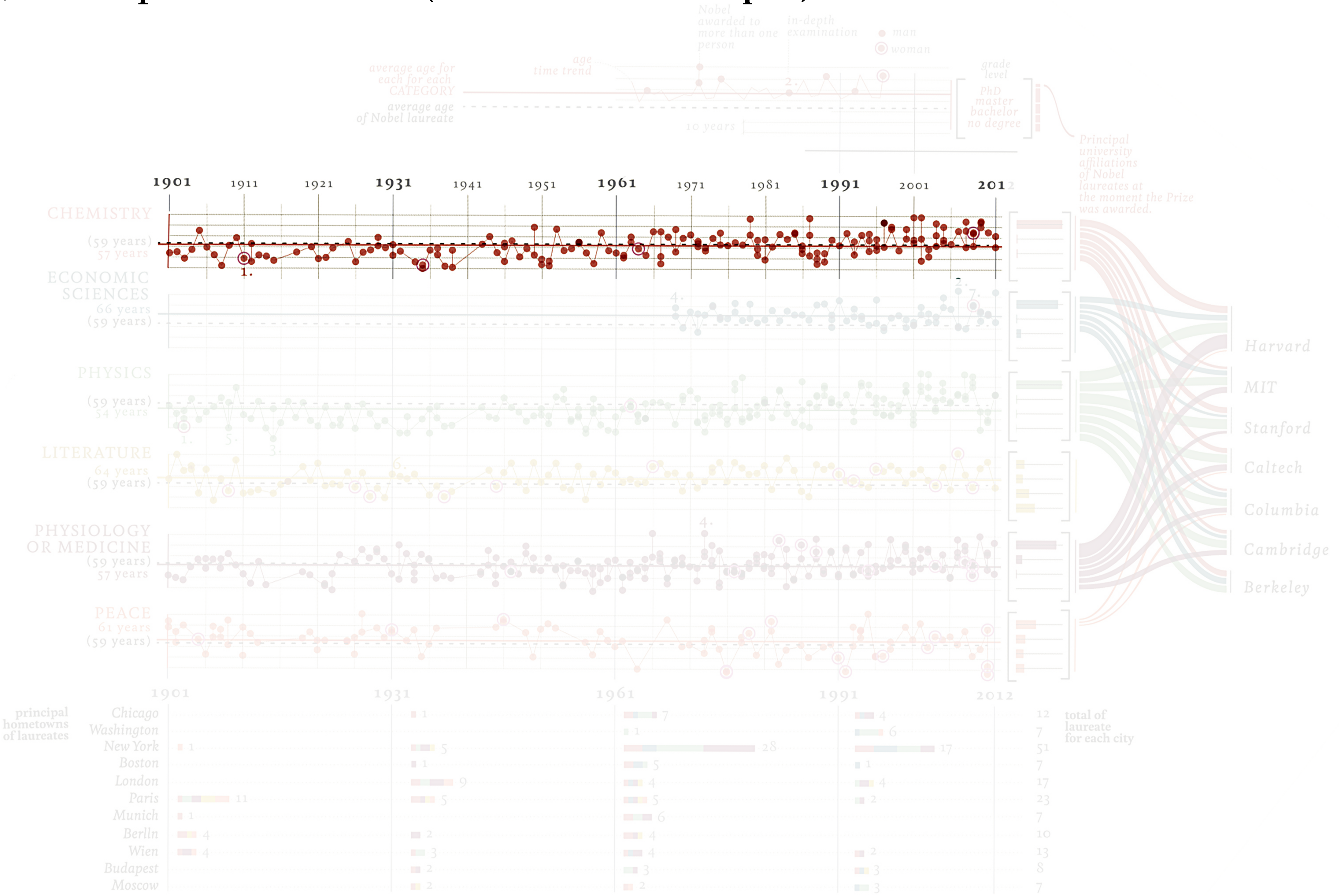




Spencer, Scott. *Approximating the Components of Lupi's Nobels, No Degrees*. March 15, 2019. https://ssp3nc3r.github.io/post/approximating-the-components-of-lupi-s-nobel-no-degrees/.

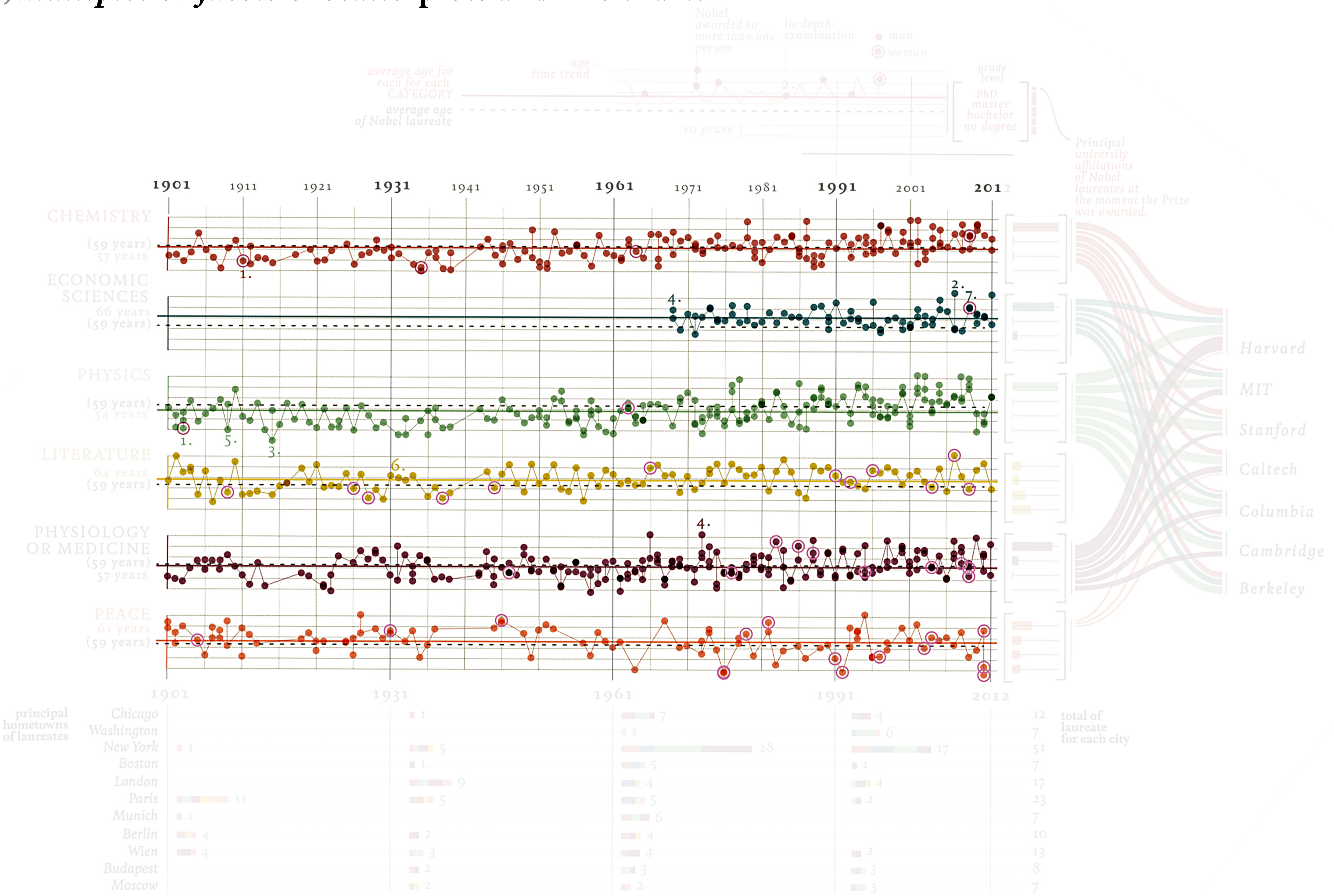
decomposition & topology study, a scatterplot and line charts (aka connected scatterplot)

```
ggplot(filter(df, Category == "Chemistry")) +

  scale_color_manual(
    values = category_colors,
    breaks = category_names) +

  scale_alpha_manual(
    values = c(1, 0),
    breaks = c("Female", "Male")) +

  scale_x_continuous(
    breaks = seq(1901, 2016, by = 30),
    minor_breaks = seq(1901, 2016, by = 10),
    position = "top") +

  geom_hline(
    mapping = aes(
      yintercept = mean(Age, na.rm = TRUE)),
    lwd = 0.2,
    color = "black",
    linetype = "dashed") +

  geom_hline(
    mapping = aes(
      yintercept = cat_avg_age,
      color = Category) ) +

  geom_line(
    mapping = aes(
      x = Year,
      y = Age,
      color = Category),
    lwd = 0.2) +

  geom_point(
    mapping = aes(
      x = Year,
      y = Age,
      color = Category),
    size = 1.5,
    alpha = 0.5) +

  geom_point(
   mapping = aes(
      x = Year,
      y = Age,
      alpha = Sex),
    color = "pink",
    shape = 21,
    size = 4)
```



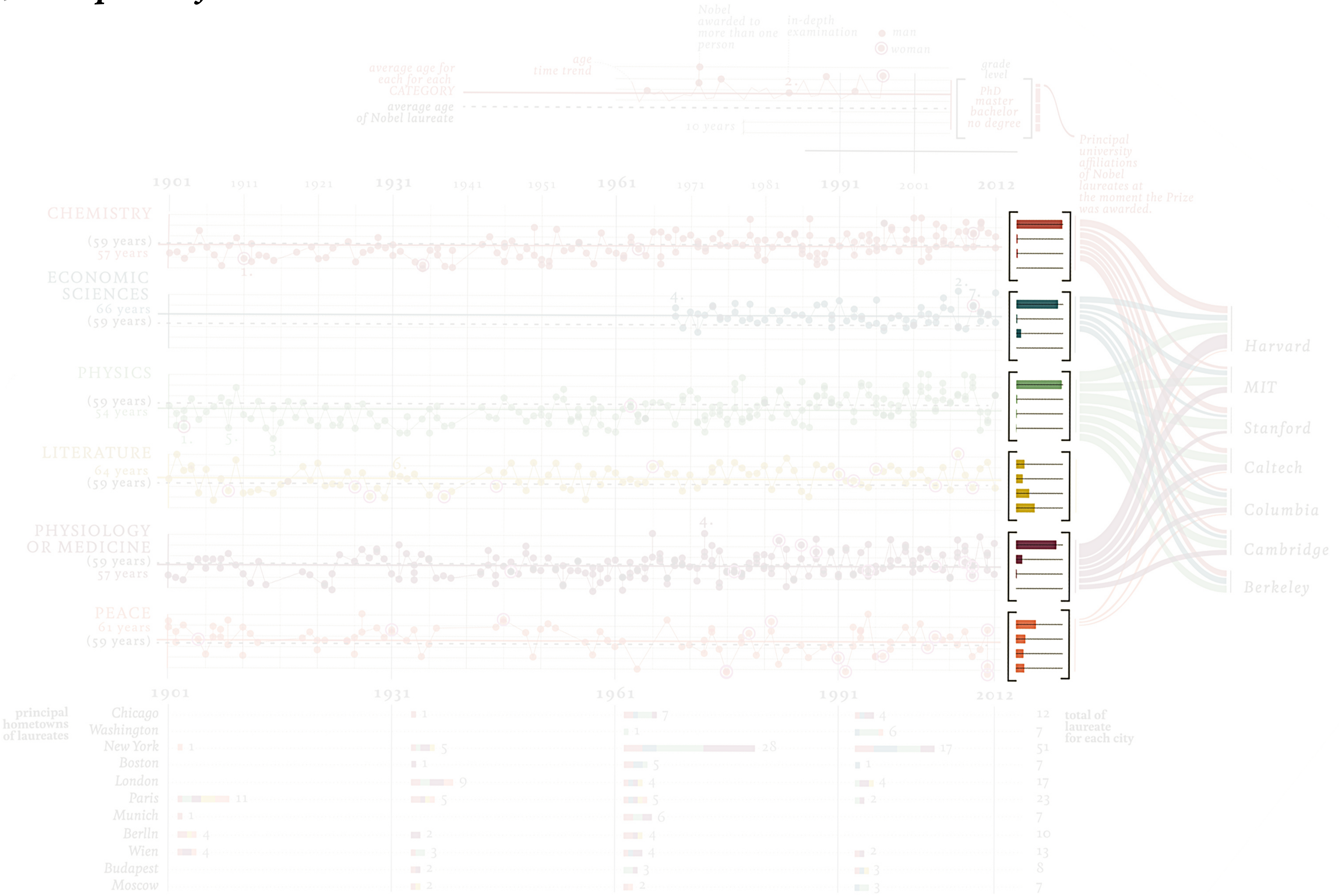
decomposition & topology study, ***multiples or facets*** **of scatterplots and line charts**

```
ggplot(df) +

  facet_wrap(
    Category ~ .,
    nrow = 6,
    strip.position = "left") +

  scale_color_manual(
    values = category_colors,
    breaks = category_names) +

  scale_alpha_manual(
    values = c(1, 0),
    breaks = c("Female", "Male")) +

  scale_x_continuous(
    breaks = seq(1901, 2016, by = 30),
    minor_breaks = seq(1901, 2016, by = 10),
    position = "top") +

  geom_hline(
    mapping = aes(
      yintercept = mean(Age, na.rm = TRUE)),
    lwd = 0.2,
    color = "black",
    linetype = "dashed") +

  geom_hline(
    mapping = aes(
      yintercept = cat_avg_age,
      color = Category) ) +

  geom_line(
    mapping = aes(
      x = Year,
      y = Age,
      color = Category),
    lwd = 0.2) +

  geom_point(
    mapping = aes(
      x = Year,
      y = Age,
      color = Category),
    size = 1.5,
    alpha = 0.5) +

  geom_point(
   mapping = aes(
      x = Year,
      y = Age,
      alpha = Sex),
    color = "pink",
    shape = 21,
    size = 4)
```



decomposition & topology study, ***multiples or facets*** **of bar charts**

```
ggplot(df) +


  facet_wrap(
    . ~ Category,
    ncol = 1) +


  scale_fill_manual(
    values = category_colors,
    breaks = category_names) +


  geom_bar(
    mapping = aes(
      x = Percent,
      y = Education,
      fill = Category),
    stat = "identity")
```

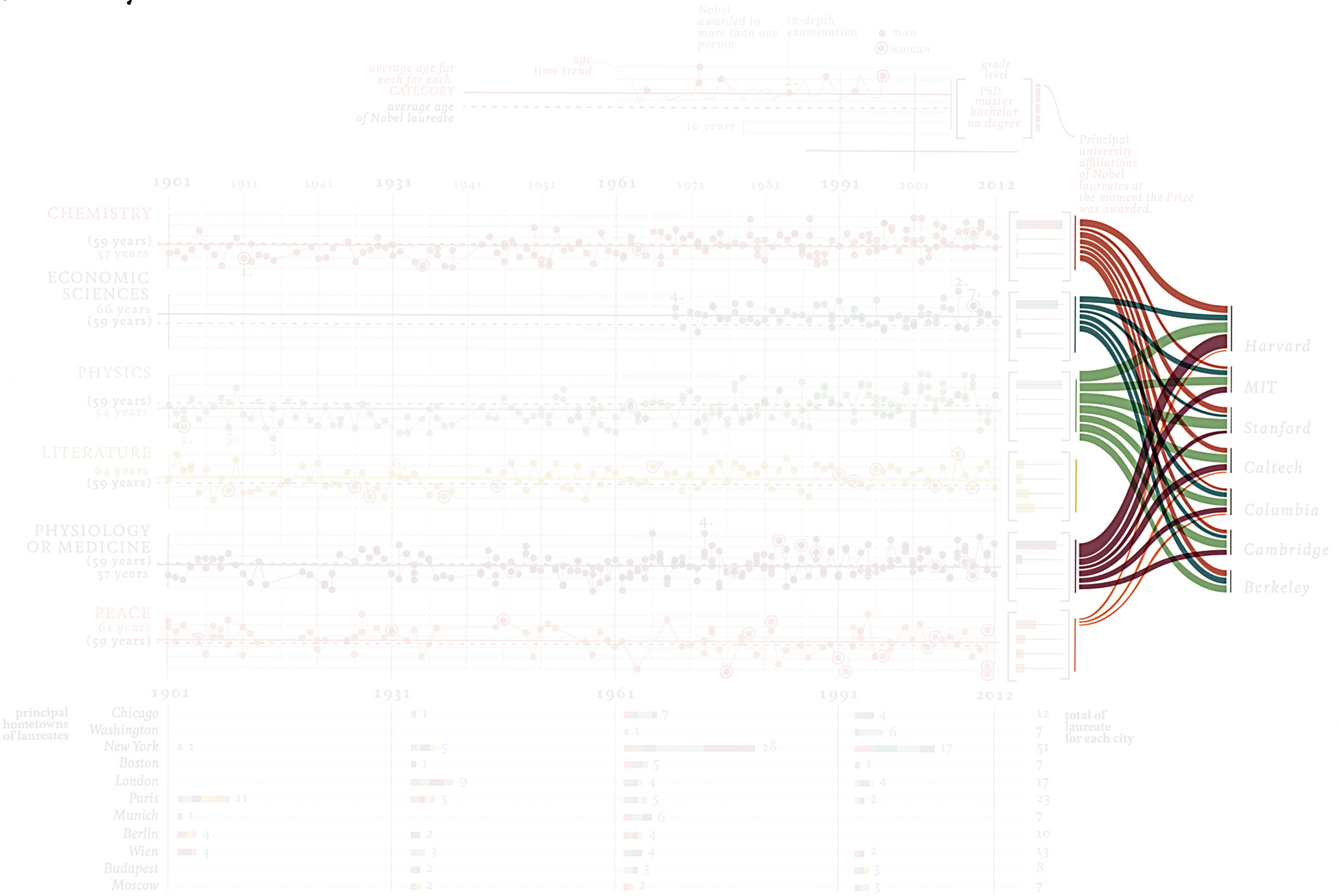


decomposition & topology study, ***multiples or facets*** **of stacked bar charts**

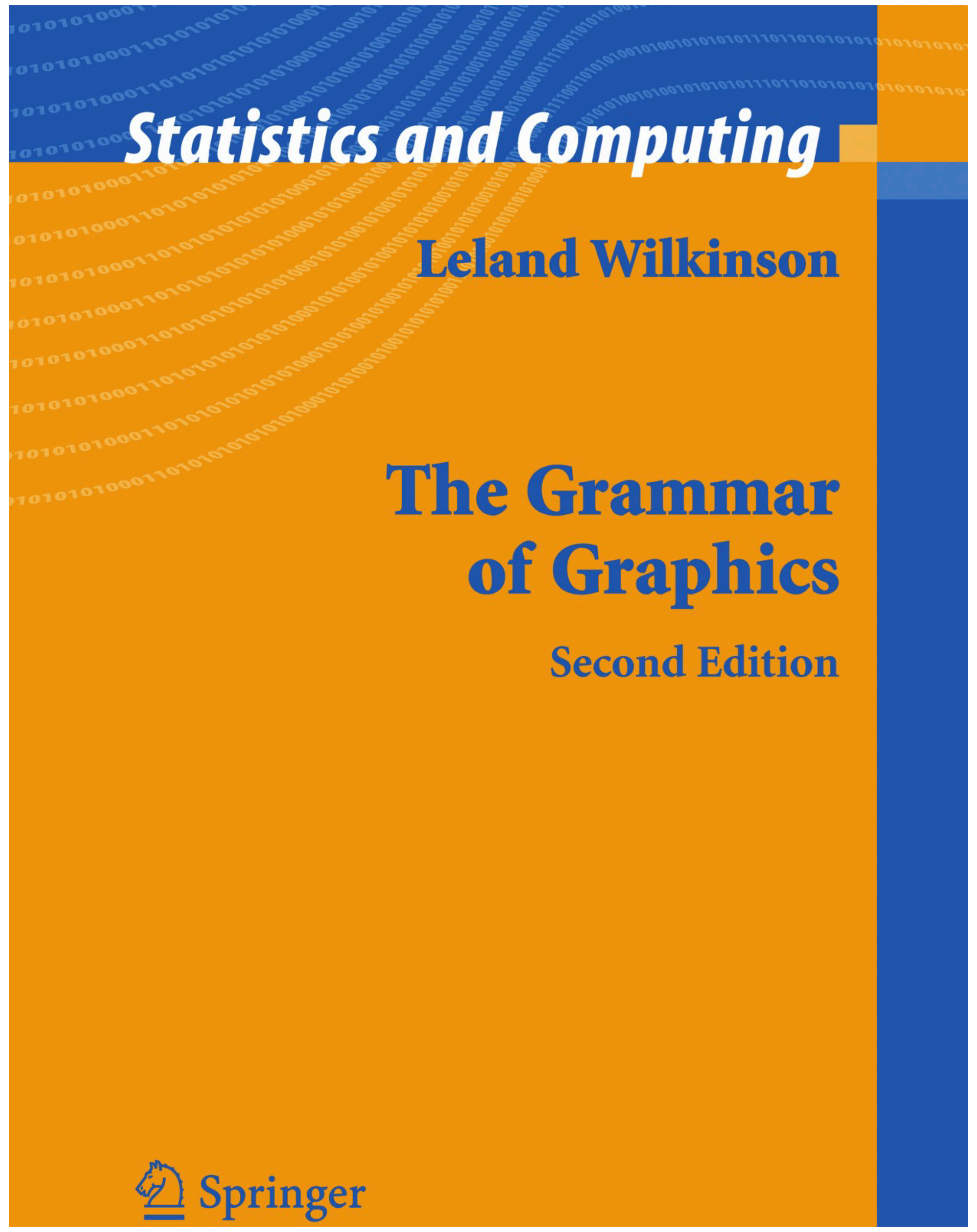
```
ggplot(df) +


  facet_wrap(
    era~.,
    nrow = 1) +


  scale_fill_manual(
    values = category_colors,
    breaks = category_names) +


  geom_bar(
    mapping = aes(
      x = n,
      y = Birth.City,
      fill = Category),
    stat = 'identity',
    width = 0.2) +


  geom_text(
    mapping = aes(
      x = n_era_city + 2,
      y = Birth.City,
      label = n_era_city),
    stat = 'identity',
    width = 0.2)
```

decomposition & topology study, **a sankey chart**

```
ggplot(
  data = data,
  mapping = aes(
    x = x,
    id = id,
    split = y,
    value = n)) +


  scale_fill_manual(
    values = category_colors,
    breaks = category_names) +


  geom_parallel_sets(
    mapping = aes(
      fill = Category),
    alpha = 0.6,
    axis.width = 0.05,
    sep = 0.1) +


  geom_parallel_sets_axes(
    axis.width = 0.01,
    fill = "gray80",
    sep = 0.1)
```



think data encodings, *not* charts …

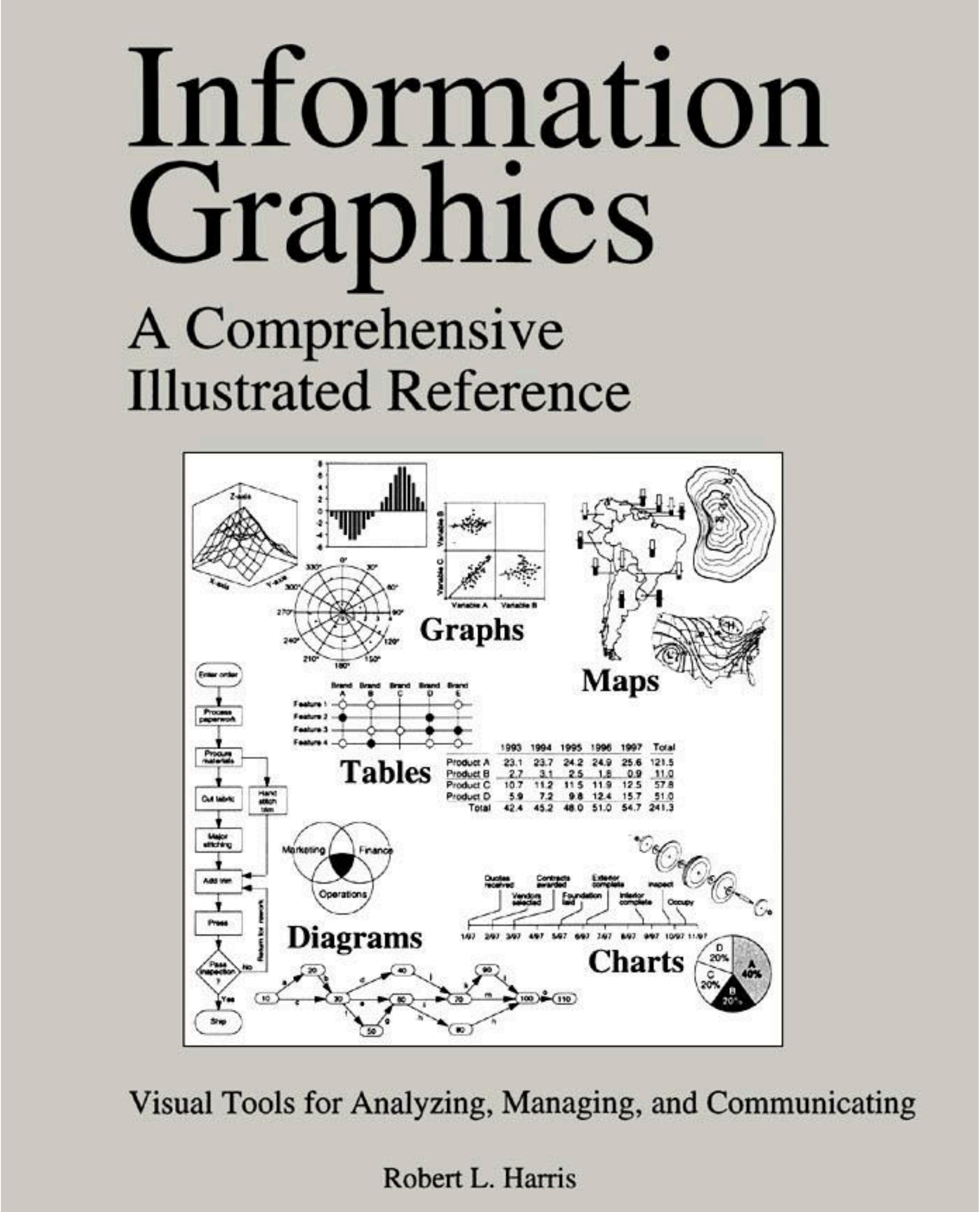


“We often call graphics charts. There are pie charts, bar charts, line charts, and so on. [We should] **shun chart typologies. Charts are usually instances of much more general objects.**

Once we understand that a pie is a divided bar in polar coordinates, we can construct other polar graphics that are less well known. We will also come to realize why a histogram is not a bar chart and why many other graphics that look similar nevertheless have different grammars…. **Elegant design requires us to think about a theory of graphics, not charts.**”

— Leland Wilkinson, *The Grammar of Graphics, Second.*

but chart typologies *can* help us learn and discuss encodings



resources

References

Spencer, Scott."Visual, Sec. 2-2.1.3" In *Data in Wonderland.* 2021. https://ssp3nc3r.github.io/data_in_wonderland.

Anscombe, F J. "*Graphs in Statistical Analysis*." The American Statistician 27, no. 1 (February 1973): 17–21.

Bertin, Jacques. *Semiology of Graphics: Diagrams Networks Maps*. Redlands: ESRI Press, 2010.

Harris, Robert L. *Information Graphics: A Comprehensive Illustrated Reference*. New York: Oxford University Press, 1999.

Koponen, Juuso, and Jonatan Hildén. *Data Visualization Handbook*. First. Finland: Aalto Art Books, 2019.

Leborg, Christian. *Visual Grammar*. Princeton Architectural Press, 2004.

Spencer, Scott. "*Approximating the Components of Lupi's Nobels, No Degrees*." Blog, March 15, 2019. https://ssp3nc3r.github.io/post/approximating-the-components-of-lupi-s-nobel-no-degrees/.

Spencer, Scott. "*Demonstration of Layers in Graphics*." Publications, March 6, 2020. https://ssp3nc3r.github.io/publications/Spencer-2020-Demonstration-of-layers-in-graphics.pdf .

Tufte, Edward R. *The Visual Display of Quantitative Information*. Second. Graphics Press, 2001.

———. *Visual Explanations. Images and Quantities, Evidence and Narrative*. Graphics Press, 1997.

Wickham, Hadley. "*A Layered Grammar of Graphics*." Journal of Computational and Graphical Statistics 19, no. 1 (January 2010): 3–28.

Wickham, Hadley, Danielle Navarro, and Thomas Lin. *Ggplot2: Elegant Graphics for Data Analysis*. Third. Springer, 2021. https://ggplot2-book.org/.

Wilke, C. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. First edition. Sebastopol, CA: O'Reilly Media, 2019.

Wilkinson, Leland. *The Grammar of Graphics*. Second. Springer, 2005.