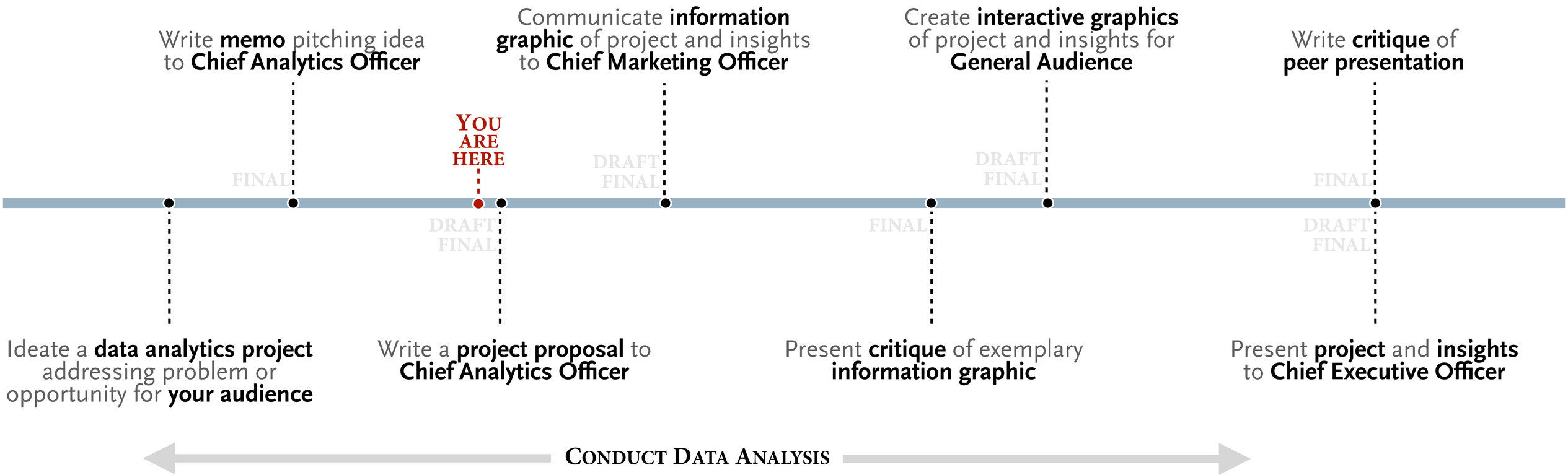
Storytelling with data

06 | grammar of graphics, Doumont applied to data encoding, color, typologies

course overview | main course deliverables



grammar of graphics

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*)

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

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

ELEMENTS : graphs (*e.g., points, lines*) and their aesthetic attributes (*e.g., color, opacity, shape, size, orientation*)

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

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

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

p <-

  # functions for data ink

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

  # functions for non-data ink

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

DATA TRANSFORMATIONS

ELEMENTS
SCALES & GUIDES
COORDINATES

GUIDES

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

Doumont's *three laws of communication* applied to data encoding

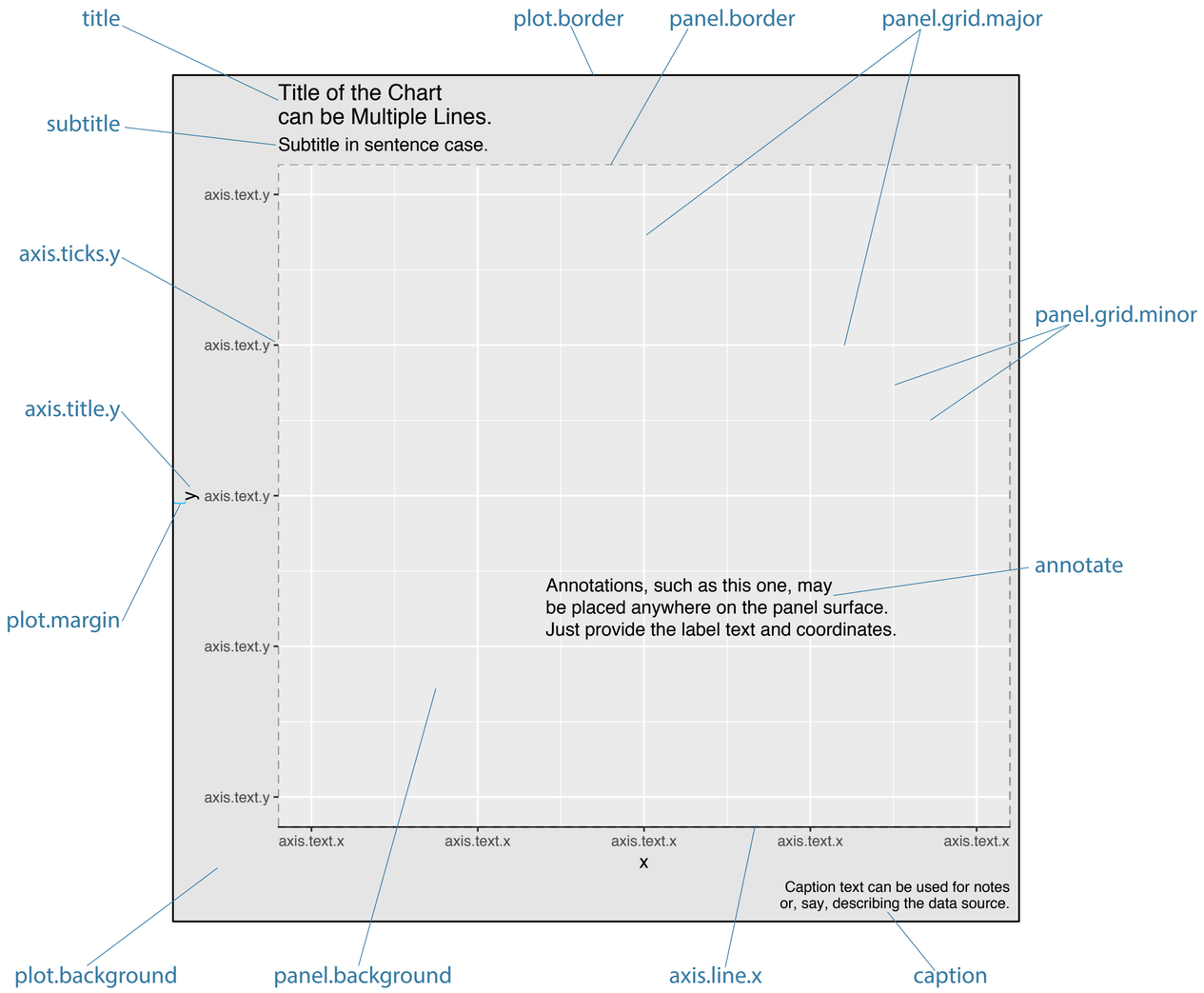
Doumont applied to data encoding, recall Doumont's three laws of communication

Adapt to your audience

Maximize the signal-to-noise ratio

Use effective redundancy

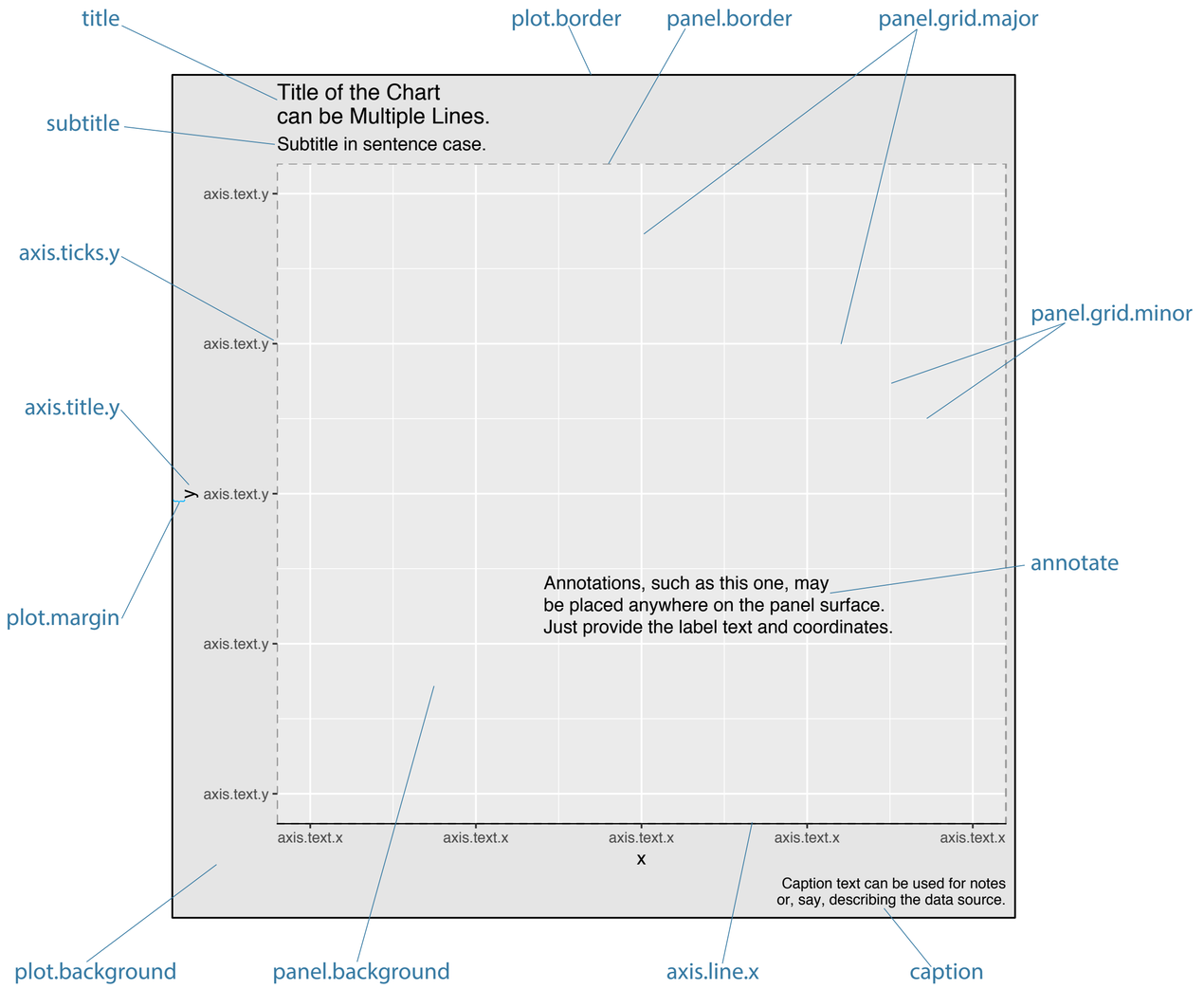
Doumont applied to data encoding, **non-“data ink”**



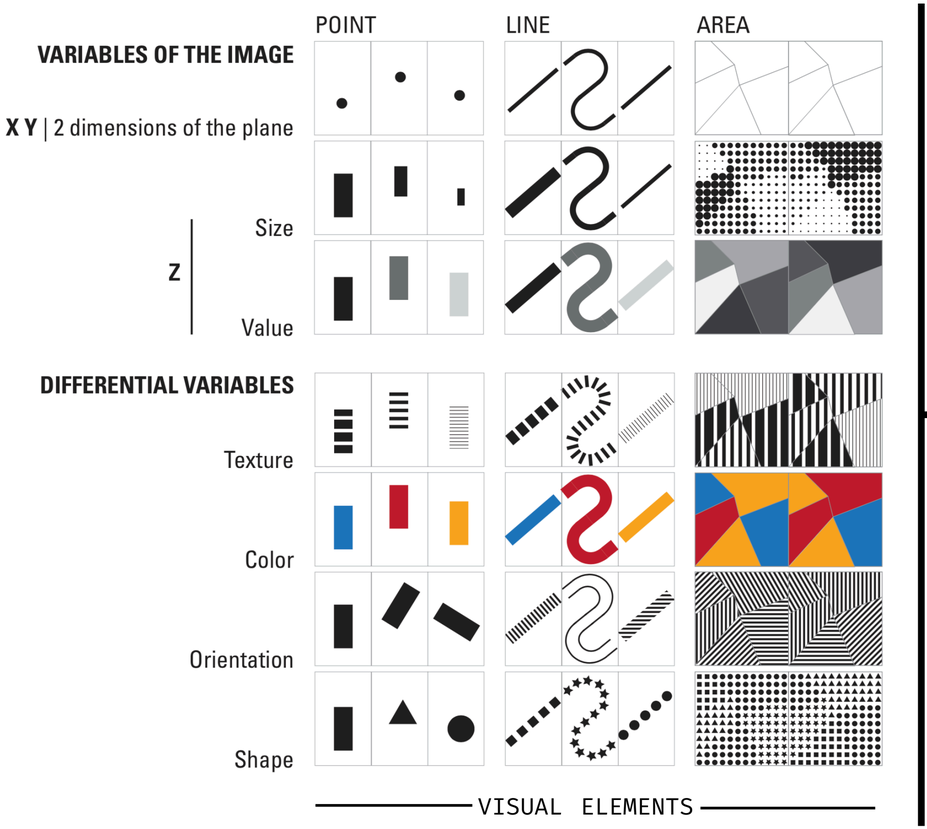
Doumont applied to data encoding, non-data ink — example functions to draw non-data ink in ggplot2

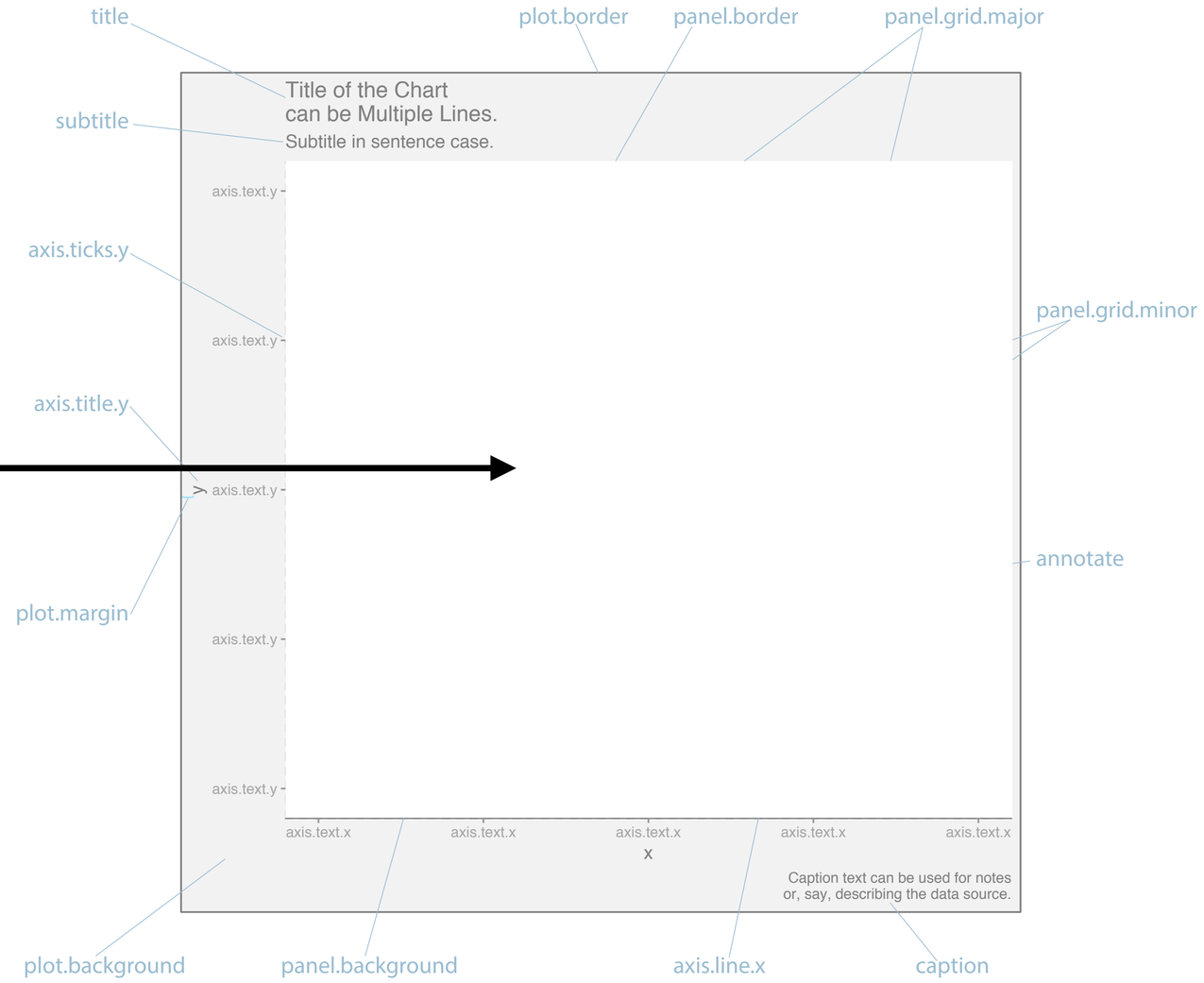
```
# load grammar of graphics
library(ggplot2)

p <-

  # functions for data ink

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

  # functions for non-data ink

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

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



Doumont applied to data encoding, "data ink" — Jacques Bertin's visual channels for encoding data

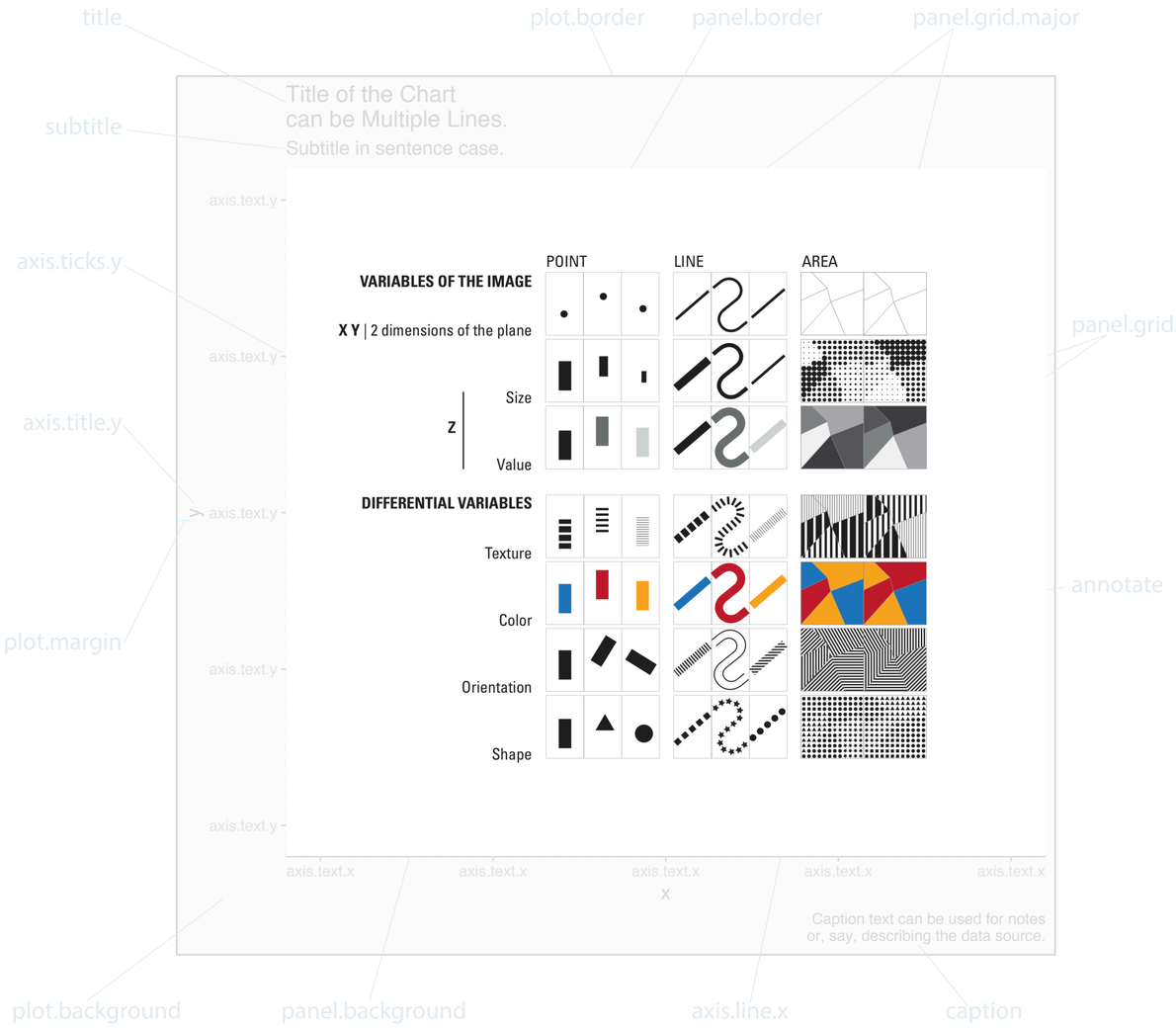




Doumont applied to data encoding, the data ink — example functions to draw encoded data in ggplot2

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

p <-

  # functions for data ink

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

  # functions for non-data ink

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

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



Doumont applied to data encoding, Tufte — data-ink maximization, *within reason*

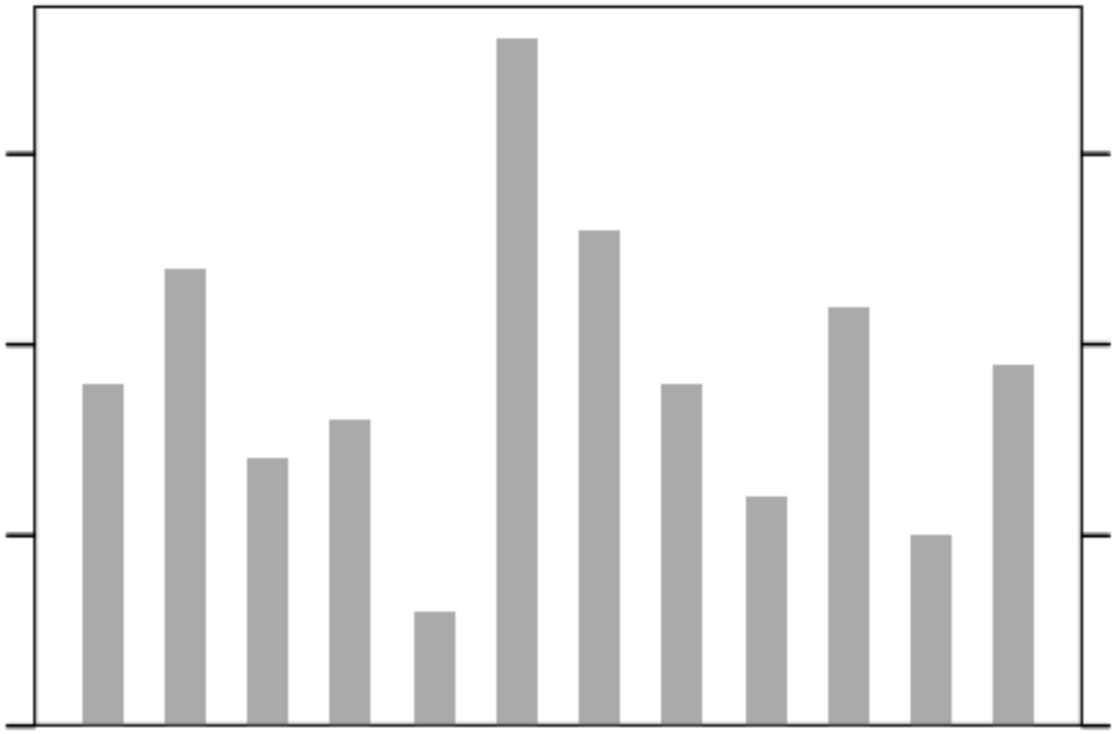
$$\text{data-ink ratio} = \frac{\text{data-ink}}{\text{total ink used to print the graphic}}$$

$$= \text{proportion of a graphic's ink devoted to the non-redundant display of data-information}$$

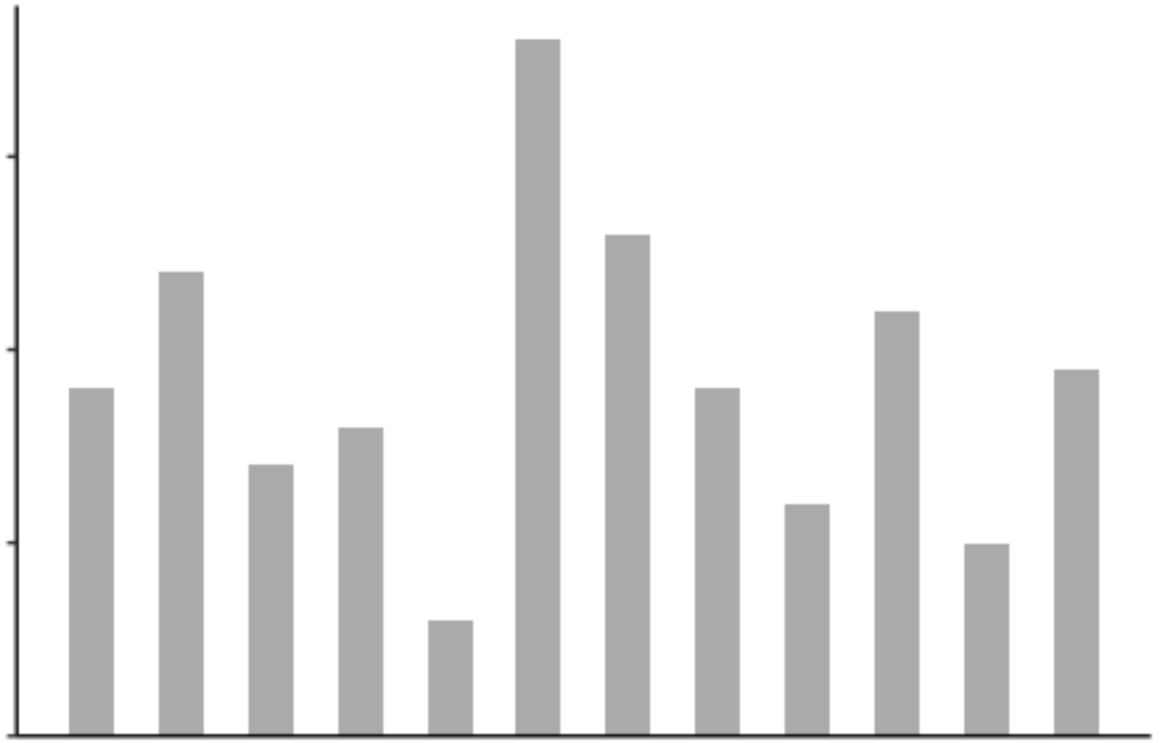
$$= \text{1.0 – proportion of a graphic that can be erased without loss of data-information}$$

Doumont — "maximize the signal-to-noise ratio"

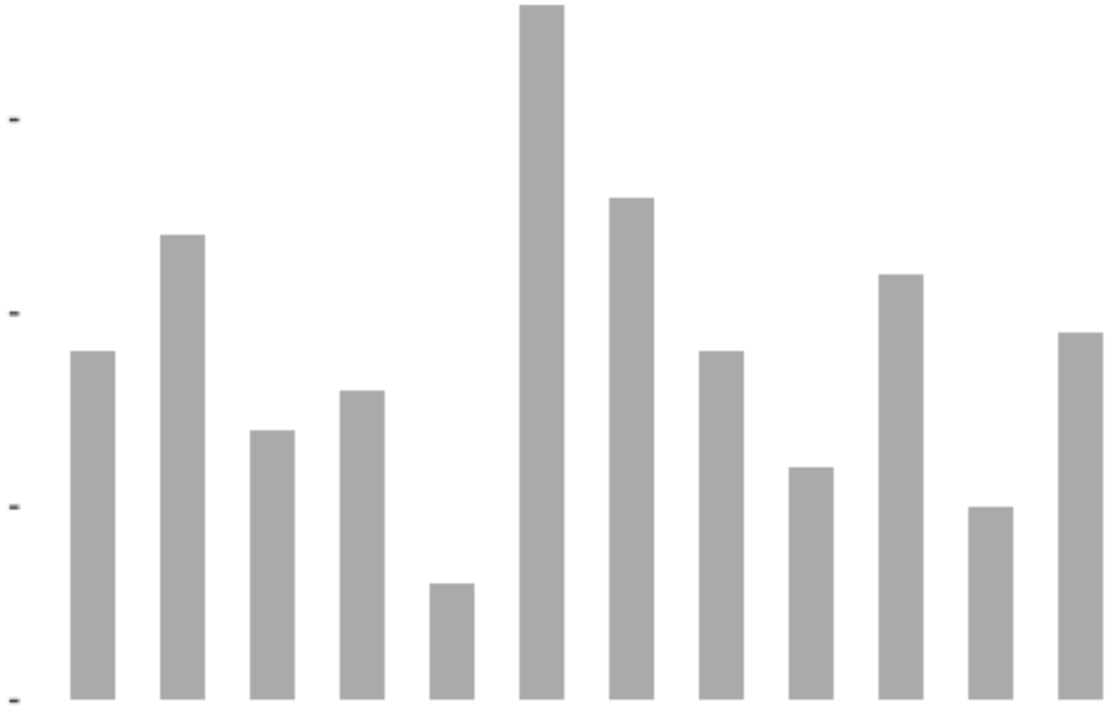
Doumont applied to data encoding, experimentation example — redesigning a bar chart



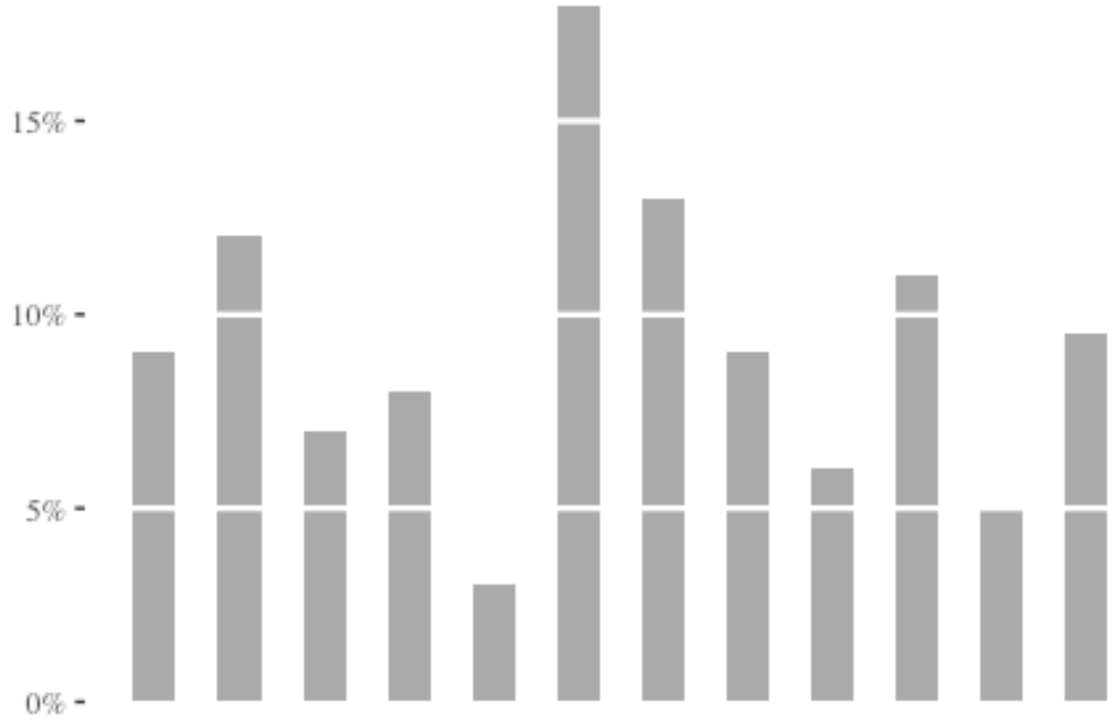
Doumont applied to data encoding, experimentation example — redesigning a bar chart



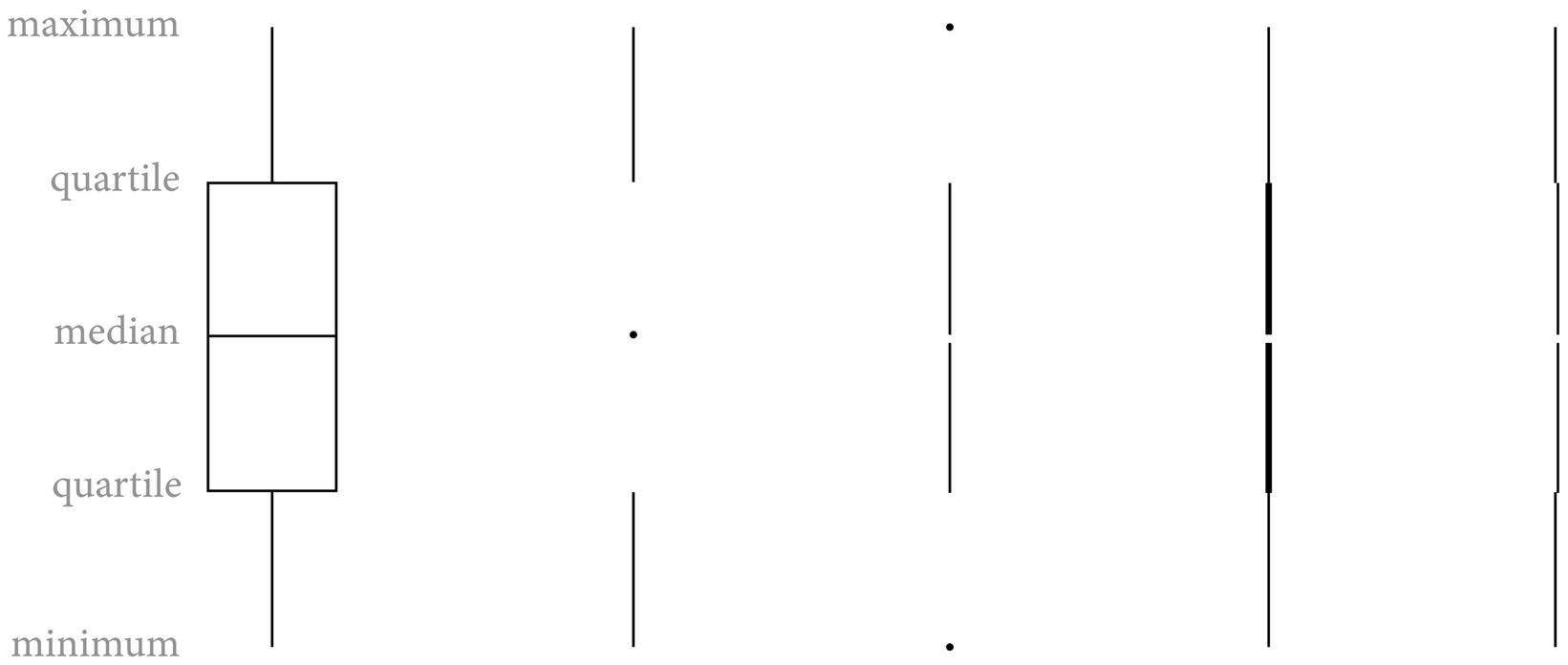
Doumont applied to data encoding, experimentation example — redesigning a bar chart



Doumont applied to data encoding, experimentation example — redesigning a bar chart



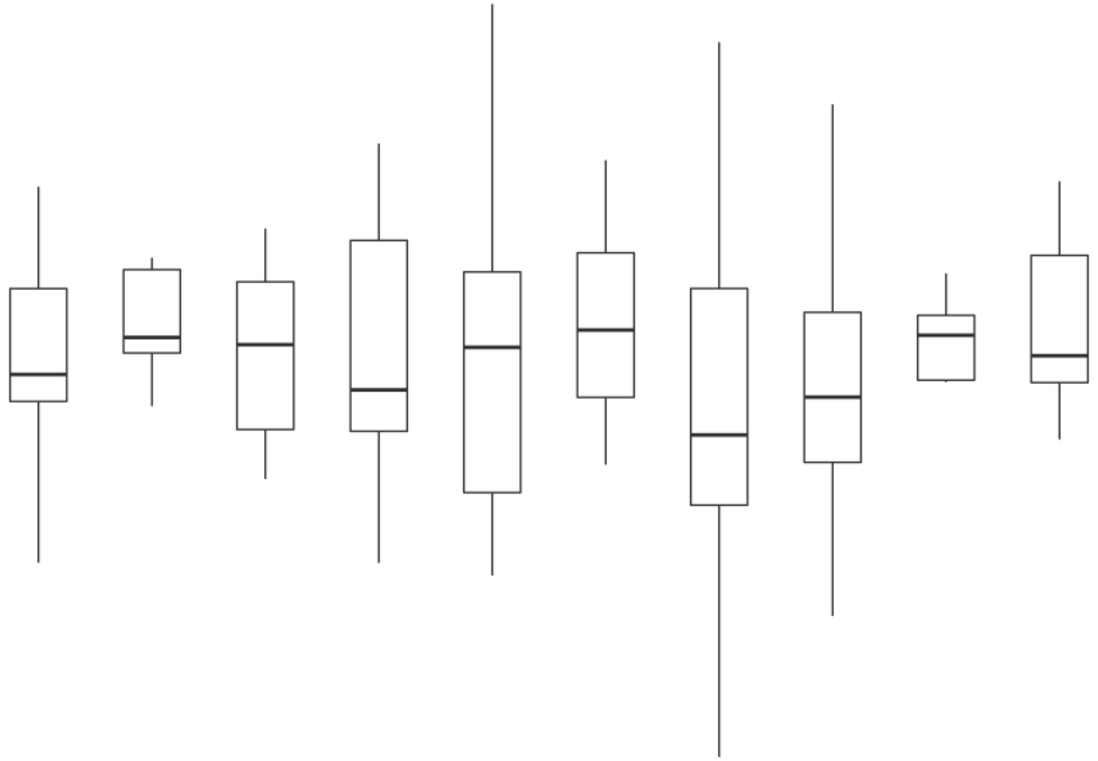
Doumont applied to data encoding, experimentation example — redesigning John Tukey's box plot



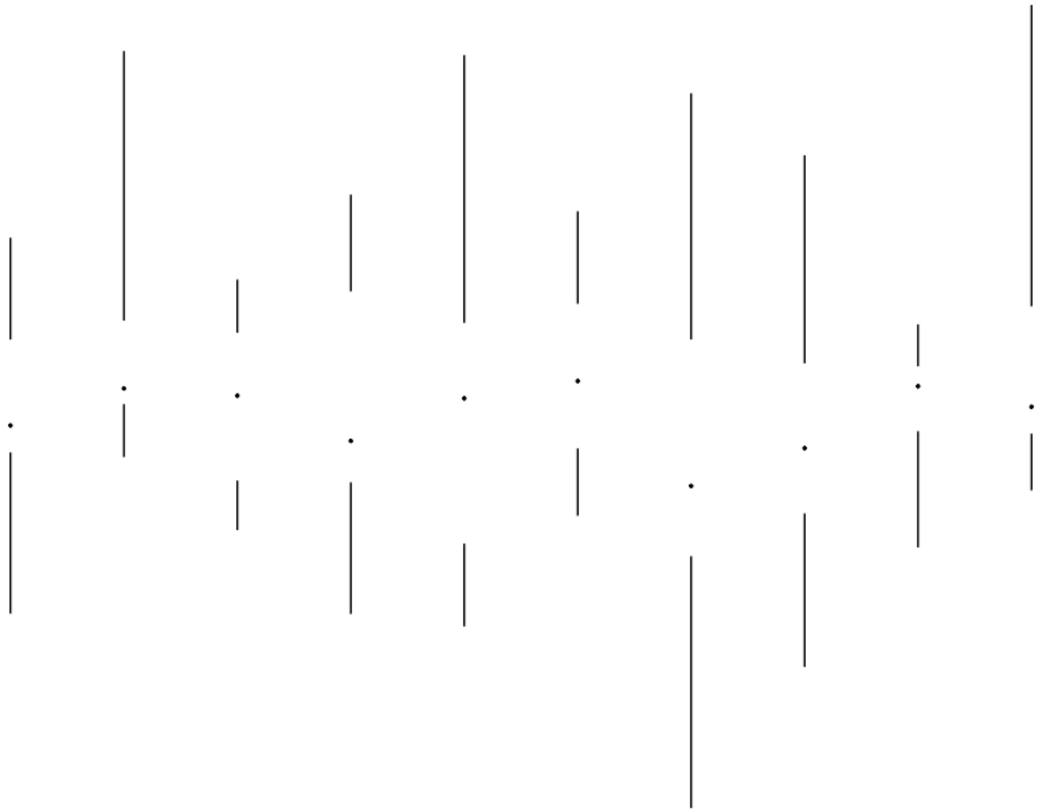
"In these revisions of the box plot, . . . the best overall arrangement naturally also rests on statistical and aesthetic criteria — in other words, the procedure is one of *reasonable* data-ink maximizing."

— Tufte, Edward, *The Visual Display of Quantitative Information*

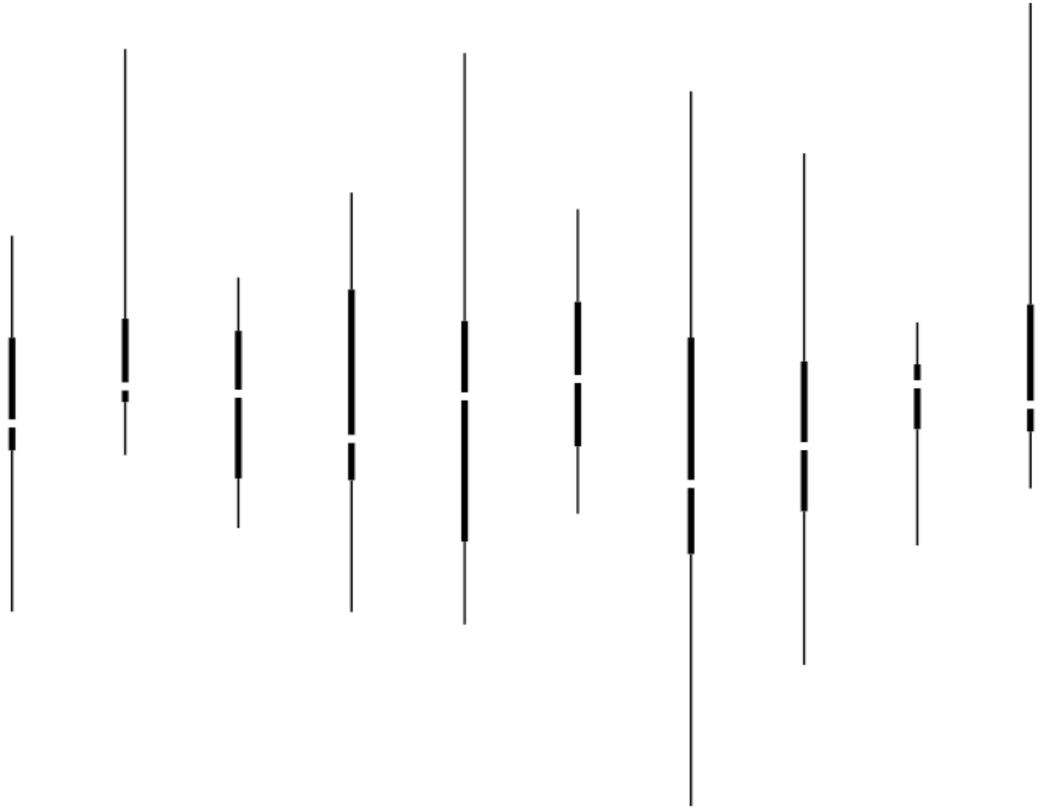
Doumont applied to data encoding, experimentation example — redesigning John Tukey's box plot



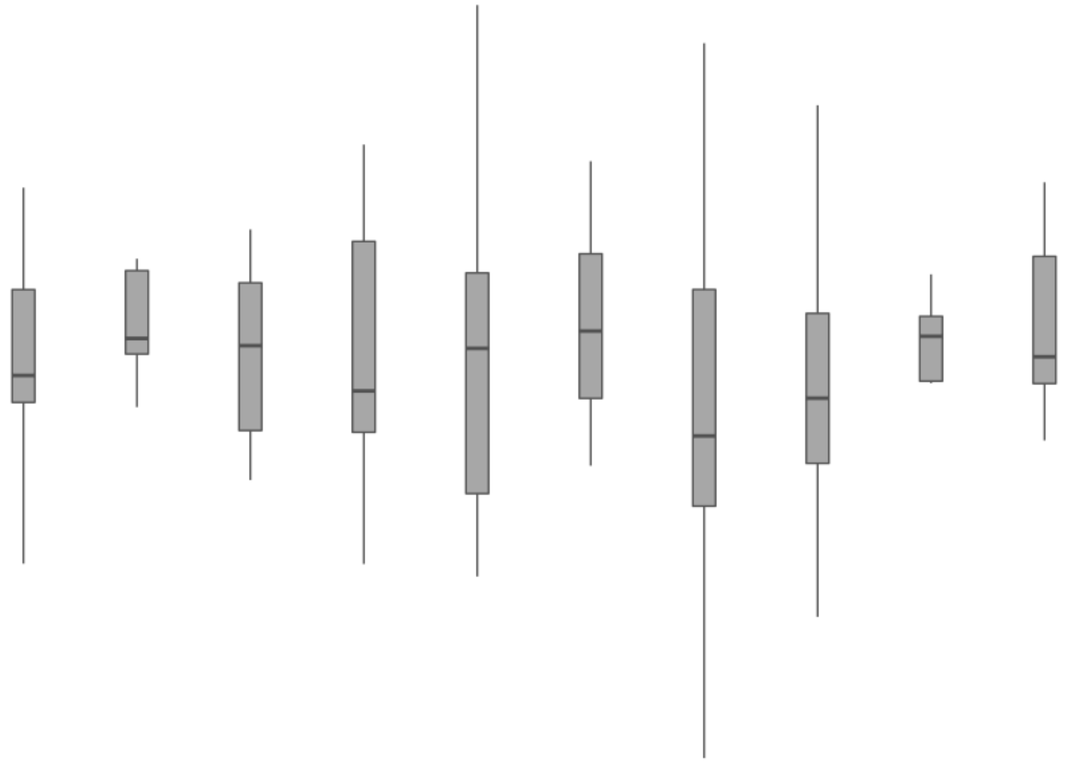
Doumont applied to data encoding, experimentation example — redesigning John Tukey's box plot



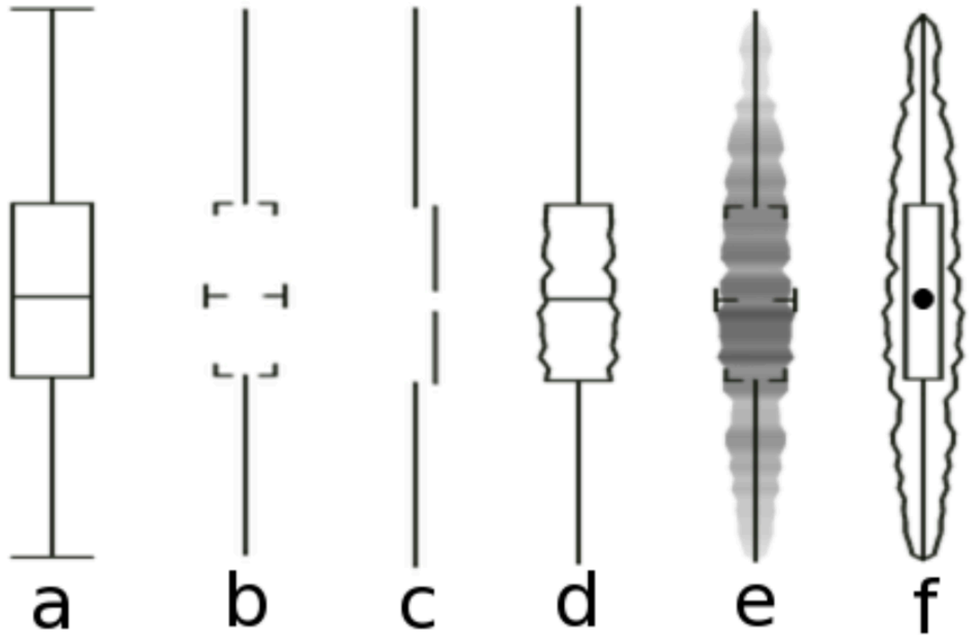
Doumont applied to data encoding, experimentation example — redesigning John Tukey's box plot



Doumont applied to data encoding, experimentation example — redesigning John Tukey's box plot



Doumont applied to data encoding, **which works best? —** iterative process of *creating*, *questioning*, *testing*!



Anderson et al. (2011) found that, of the four kinds of boxplot shown in Figure 1.7, the minimalist version from Tufte's own work (option C) proved to be the most cognitively difficult for viewers to interpret.

Cues like labels and gridlines, together with some strictly superfluous embellishment of data points or other design elements, may often be an aid rather than an impediment to interpretation.

Doumont — "adapt to your audience"

Doumont applied to data encoding, **data ink** — one of many design considerations

"Maximizing data ink (within reason) is but a single dimension of a complex and multivariate design task.

The principle helps conduct
experiments in graphical design.

Some of those
experiments will succeed.

There remain, however, many **other considerations** in the design of statistical graphics — not only of efficiency, but also of **complexity**, **structure**, **density**, and even **beauty**."

— Tufte, Edward, *The Visual Display of Quantitative Information*

Doumont applied to data encoding, which works best? — iterative process of *creating*, *questioning*, *testing*!

Design is a search problem

Prototypes should emphasize speed over polish.

Get fresh eyes frequently. Invite criticism.

Move from exploring to refining.

Doumont applied to data encoding, **which works best?** — iterative process of *creating, questioning, testing*!

SECTIONS HOME SEARCH

The New York Times

Account

Joe Biden, in Video, Says He Will Be 'More Mindful' of Personal Space

Senate Republicans Go 'Nuclear' to Speed Trump Confirmations

Subpoena for Mueller Report and Documents Approved by House Judiciary Committee

House Intelligence Committee Seeks Documents From Trump's Inaugural

In Rare S
Leader S

POLITICS | ONE-PARTY RULE

SHARE

Taking the Battle to the States

By **HAEYOUN PARK**, **JEREMY ASHKENAS** and MIKE BOSTOCK JAN. 11, 2014

Republicans or Democrats have single-party control of both the legislature and the governor's office in 36 states, the most in six decades. Lawmakers in these states have been seeking to reshape government policy in recent years, from legalizing same-sex marriage to restricting labor unions. Some of these laws were passed after the rapid rise of single-party control in 2010; others have been in place for years. Below is a look at where states stand on some key issues.

88 million people live in states with **Democratic** control of both legislatures and the governor's office.

69 million people live in states with **mixed** control.

157 million people live in states with **Republican** control.

Did not pass new restrictions in 2013

Abortion

Twenty-two states — all but six controlled by Republicans — enacted 70 abortion restrictions in 2013, according to the Guttmacher Institute, a research group. Laws ranged from bans on abortions 20 weeks after fertilization to limitations on insurance coverage of abortions.

Passed new restrictions in 2013

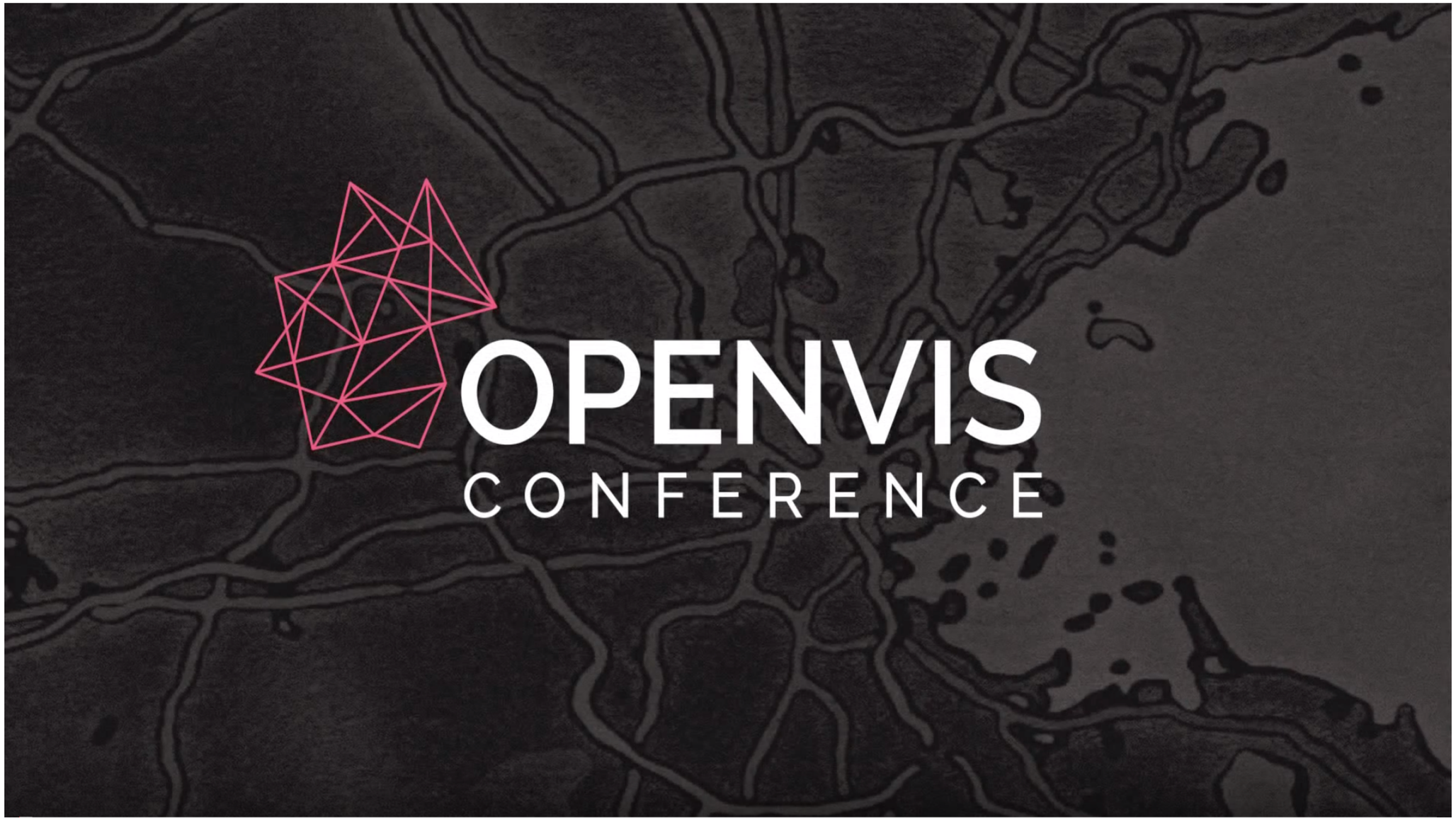
Expanded Medicaid

Medicaid

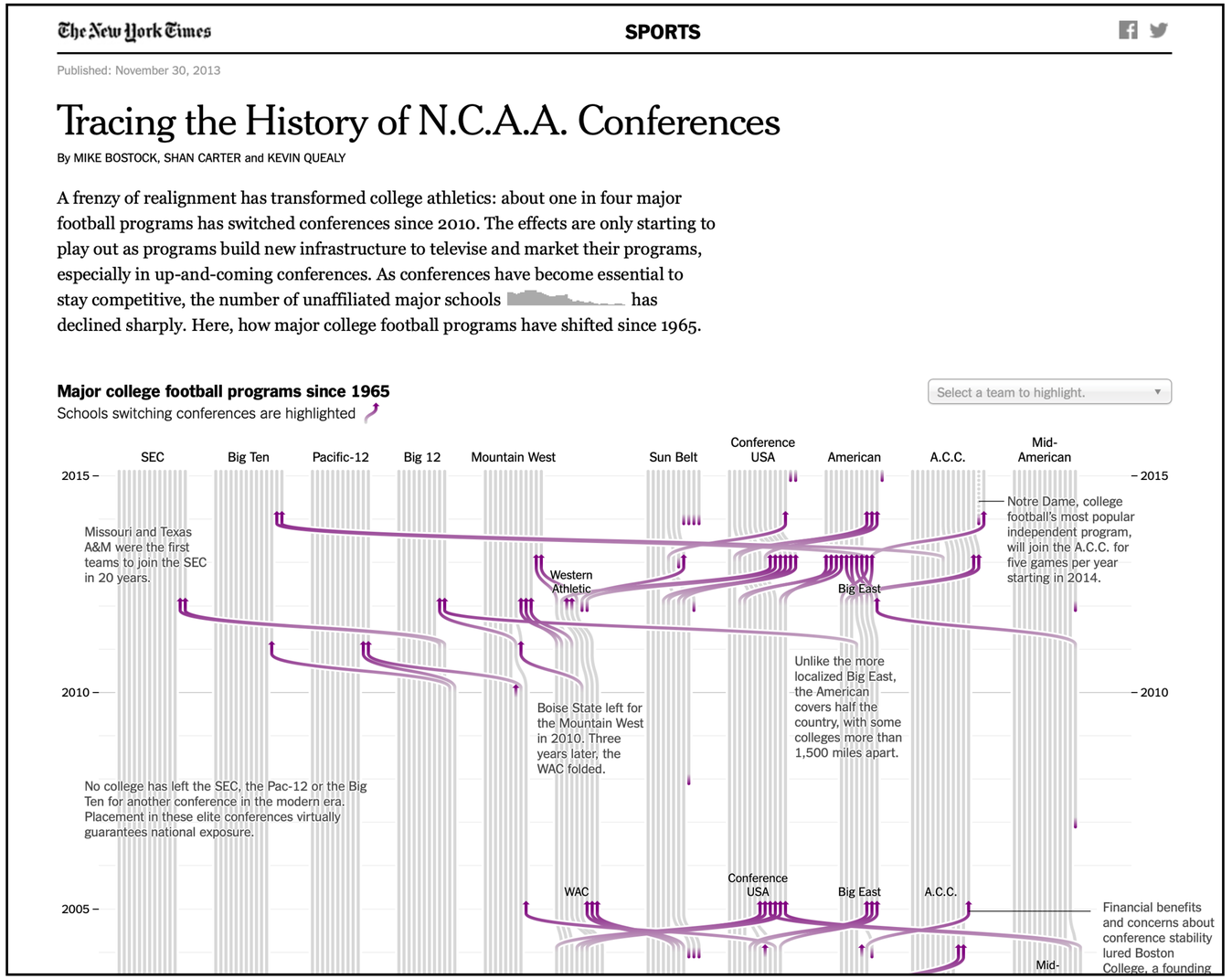
A 2012 Supreme Court decision allowed states to decide whether to expand Medicaid to more low-income adults under the Affordable Care Act. All 13 Democratic states have expanded the program. Most Republican states did not. Pennsylvania, Indiana and Tennessee have not expanded it but may do so.

Have not expanded at this time

Doumont applied to data encoding, **which works best?** — iterative process of *creating*, *questioning*, *testing*!



Doumont applied to data encoding, **which works best?** — iterative process of *creating, questioning, testing*!



The New York Times **SPORTS**

Published: November 30, 2013

Tracing the History of N.C.A.A. Conferences

By MIKE BOSTOCK, SHAN CARTER and KEVIN QUEALY

A frenzy of realignment has transformed college athletics: about one in four major football programs has switched conferences since 2010. The effects are only starting to play out as programs build new infrastructure to televise and market their programs, especially in up-and-coming conferences. As conferences have become essential to stay competitive, the number of unaffiliated major schools has declined sharply. Here, how major college football programs have shifted since 1965.

Major college football programs since 1965

Schools switching conferences are highlighted

Select a team to highlight.

SEC | Big Ten | Pacific-12 | Big 12 | Mountain West | Sun Belt | Conference USA | American | A.C.C. | Mid-American

2015 – 2015

2010 – 2010

2005

Missouri and Texas A&M were the first teams to join the SEC in 20 years.

Notre Dame, college football's most popular independent program, will join the A.C.C. for five games per year starting in 2014.

Western Athletic

Big East

Boise State left for the Mountain West in 2010. Three years later, the WAC folded.

Unlike the more localized Big East, the American covers half the country, with some colleges more than 1,500 miles apart.

No college has left the SEC, the Pac-12 or the Big Ten for another conference in the modern era. Placement in these elite conferences virtually guarantees national exposure.

WAC

Conference USA

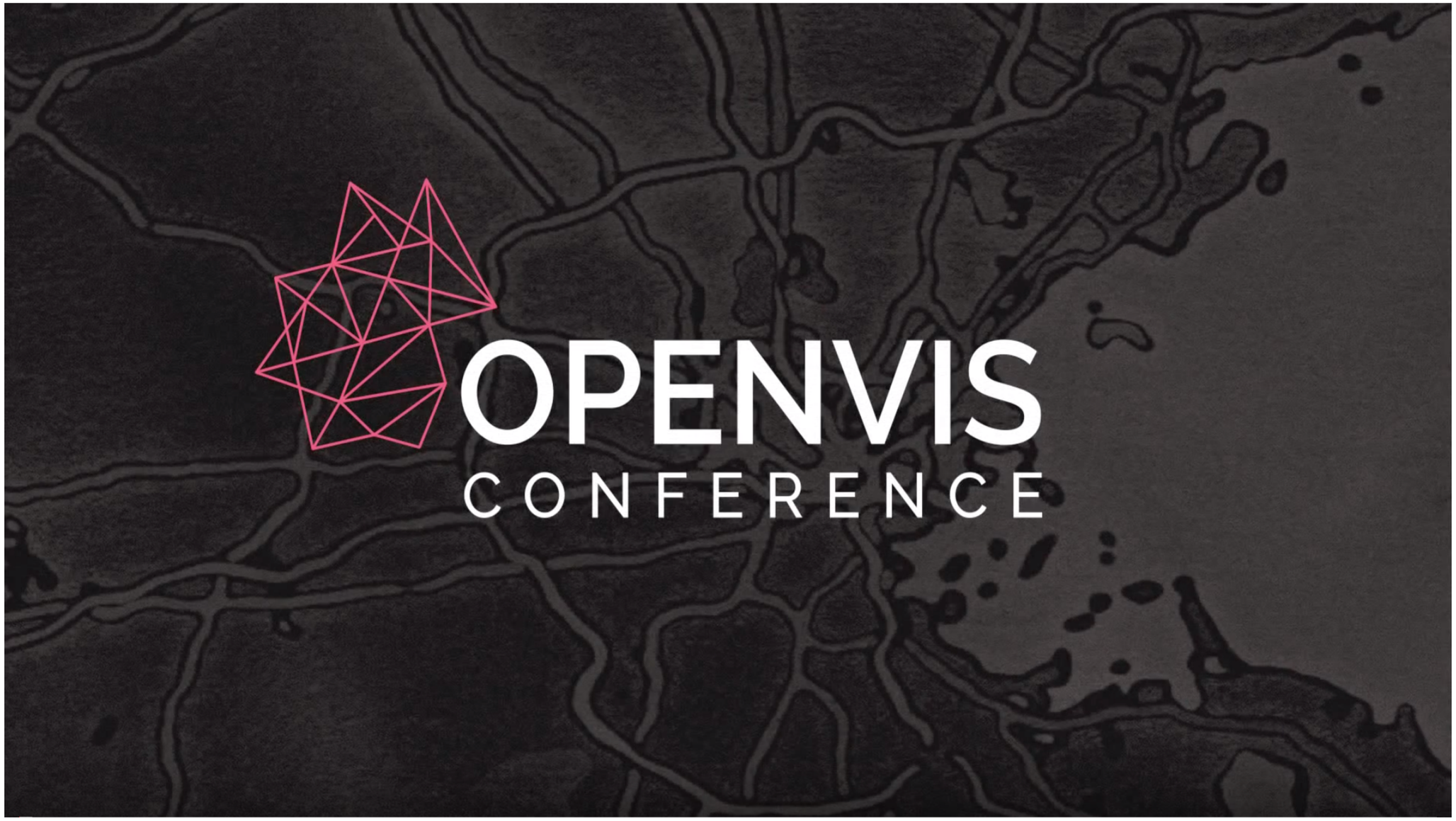
Big East

A.C.C.

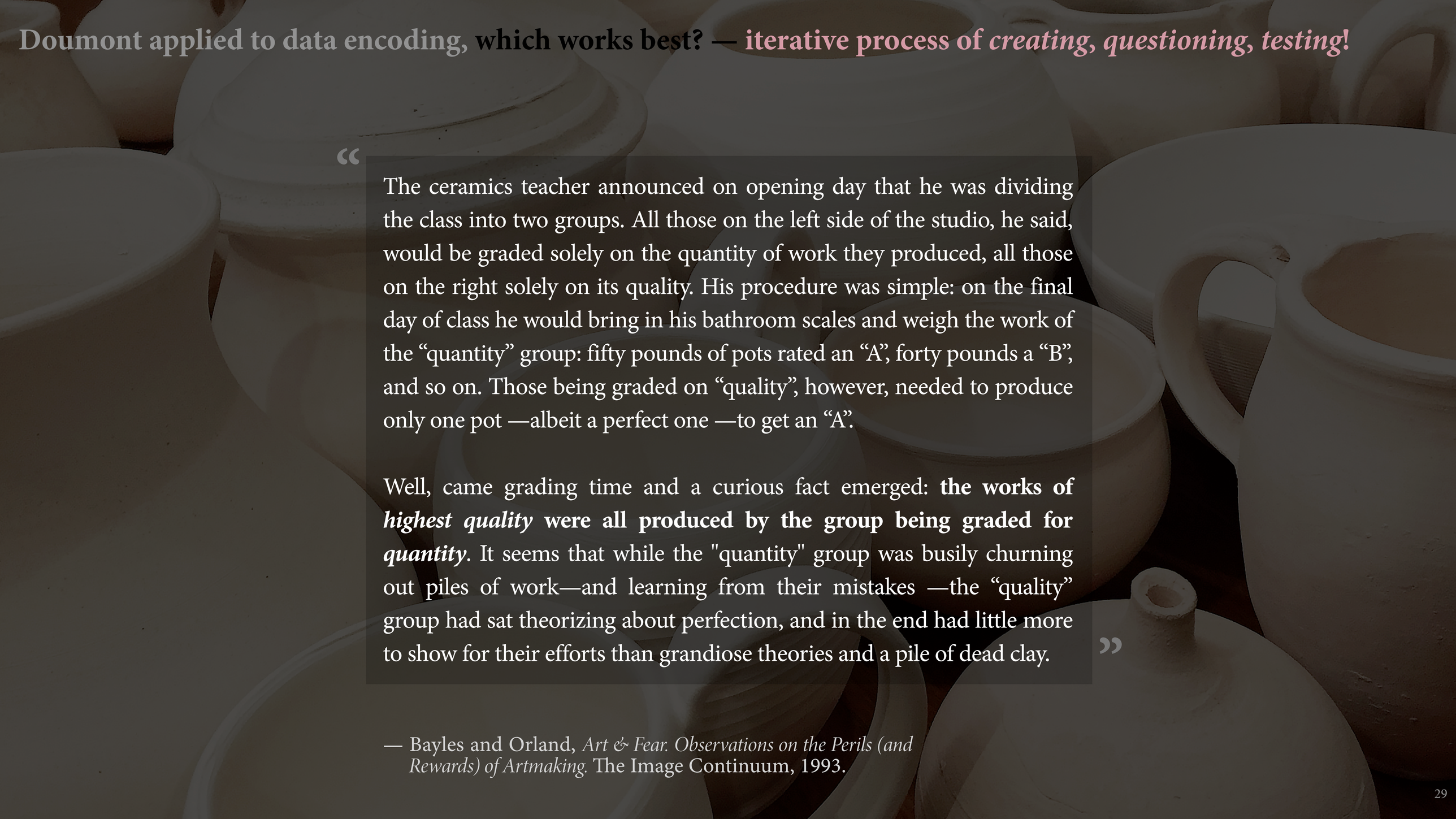
Financial benefits and concerns about conference stability lured Boston College, a founding

Mid-

Doumont applied to data encoding, which works best? — iterative process of *creating*, *questioning*, *testing*!



Doumont applied to data encoding, which works best? — iterative process of *creating*, *questioning*, *testing*!

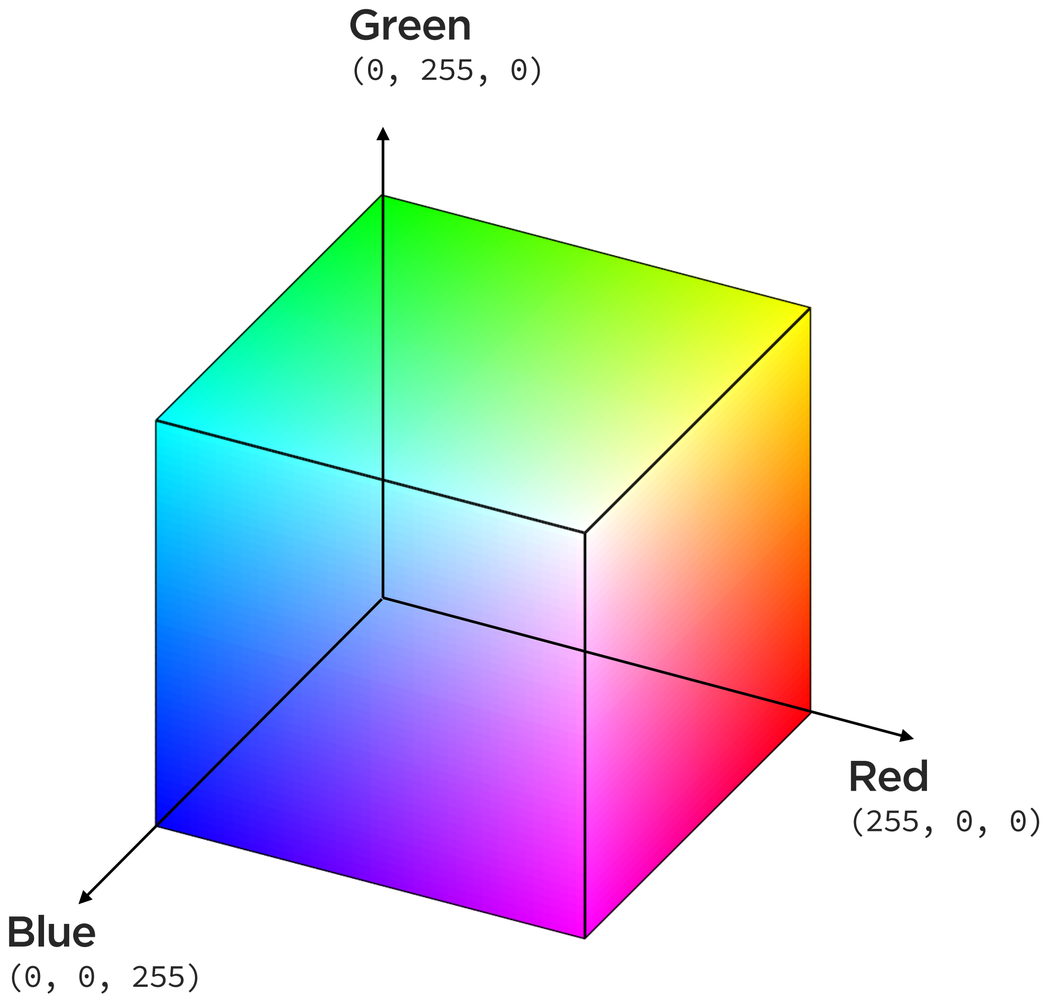


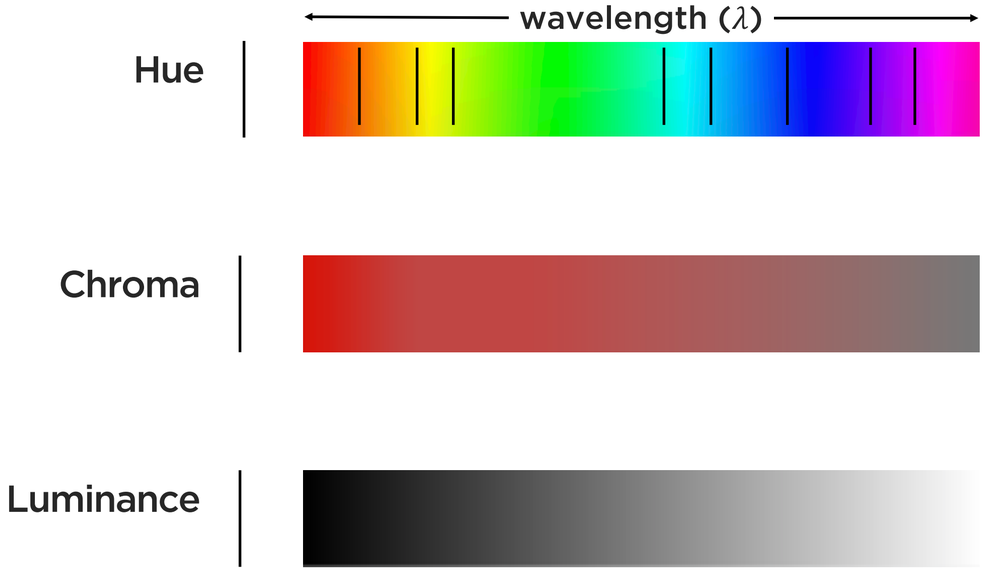
> The ceramics teacher announced on opening day that he was dividing the class into two groups. All those on the left side of the studio, he said, would be graded solely on the quantity of work they produced, all those on the right solely on its quality. His procedure was simple: on the final day of class he would bring in his bathroom scales and weigh the work of the "quantity" group: fifty pounds of pots rated an "A", forty pounds a "B", and so on. Those being graded on "quality", however, needed to produce only one pot —albeit a perfect one —to get an "A".
>
> Well, came grading time and a curious fact emerged: **the works of *highest quality* were all produced by the group being graded for *quantity***. It seems that while the "quantity" group was busily churning out piles of work—and learning from their mistakes —the "quality" group had sat theorizing about perfection, and in the end had little more to show for their efforts than grandiose theories and a pile of dead clay.

— Bayles and Orland, *Art & Fear. Observations on the Perils (and Rewards) of Artmaking.* The Image Continuum, 1993.

encoding data as color

encoding data as color, encode data using color spaces, which are mathematical models





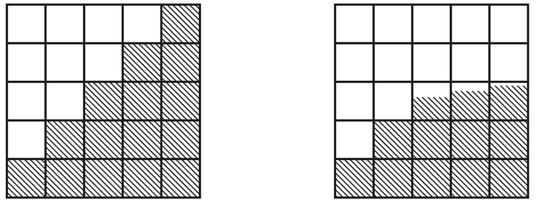
encoding data as color, how can we map data to light, whether using its hue, chroma, or luminance?

encoding data as color, perceived brightness is nonlinear function of luminance

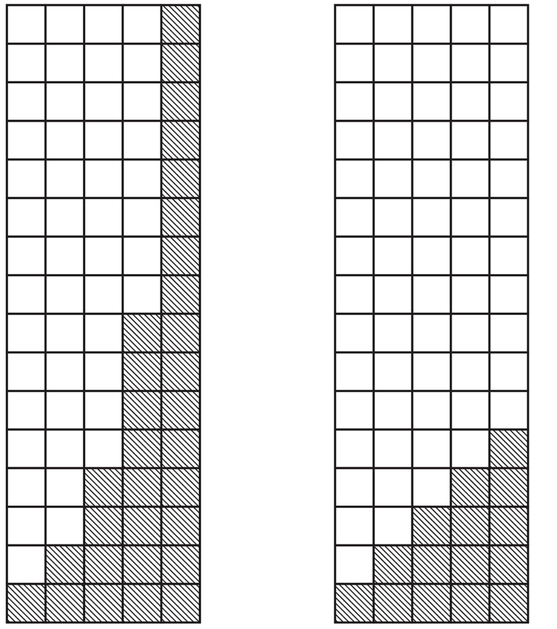
LUMINANCE : the *measured* amount of light coming from some region of space.

BRIGHTNESS : the *perceived* amount of light coming from that region of space.

encoding data as color, visual perception of arithmetical progression depends on physical geometric progression

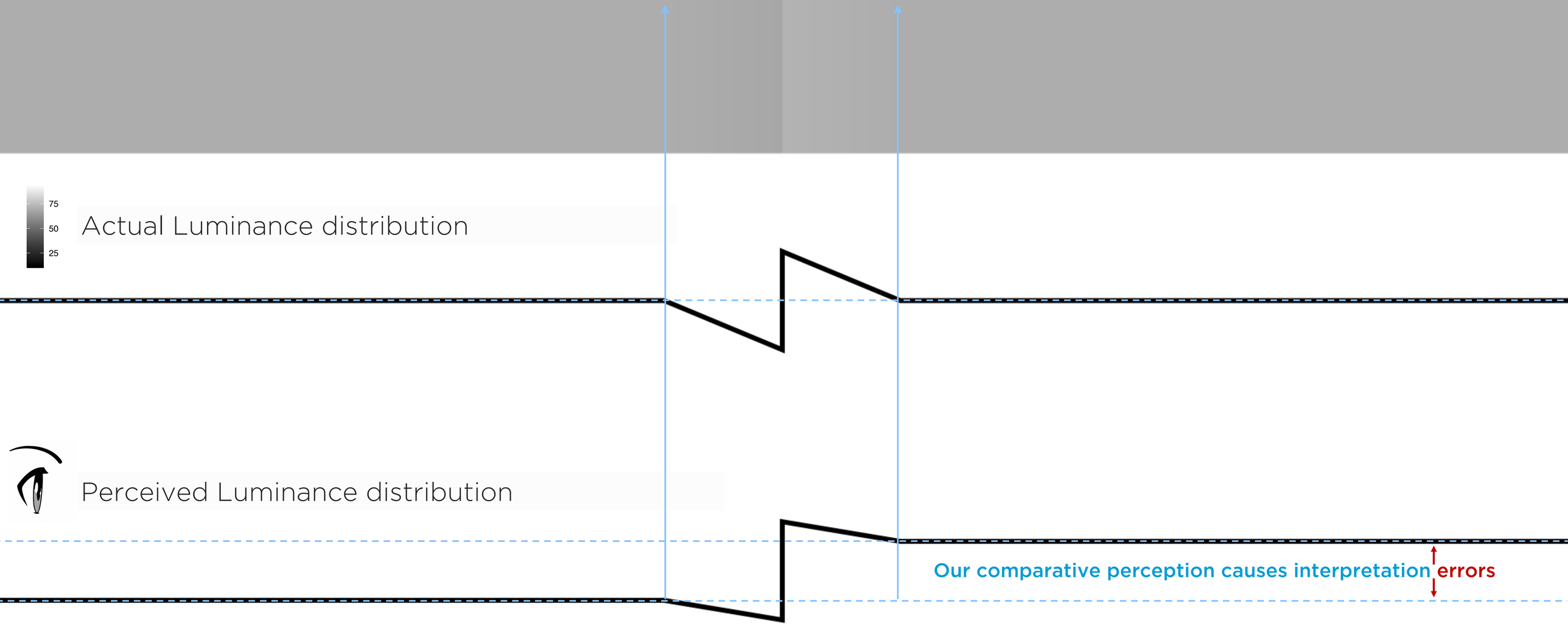


THIS PHYSICAL FACT REDUCES TO THIS PSYCHOLOGICAL EFFECT

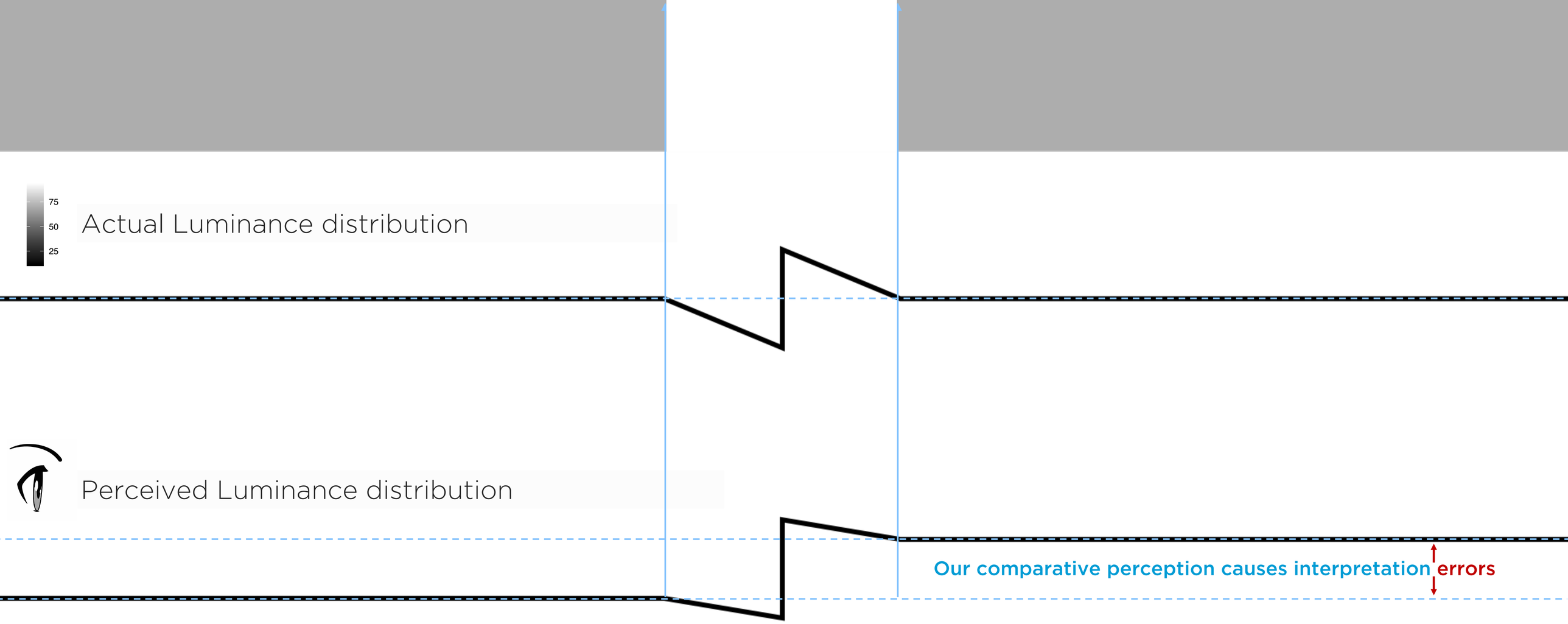


THIS PHYSICAL FACT PRODUCES THIS PSYCHOLOGICAL EFFECT

color, **humans have evolved to see edge contrasts. We see comparative — not absolute — luminance value.**

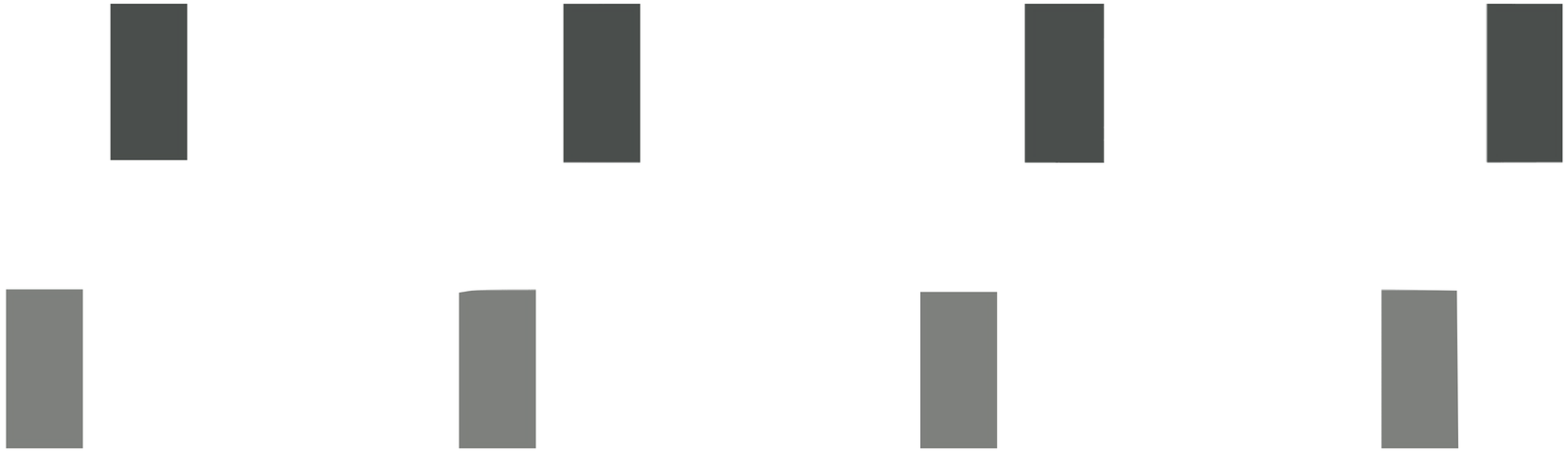


color, **humans have evolved to see edge contrasts. We see comparative — not absolute — luminance value.**

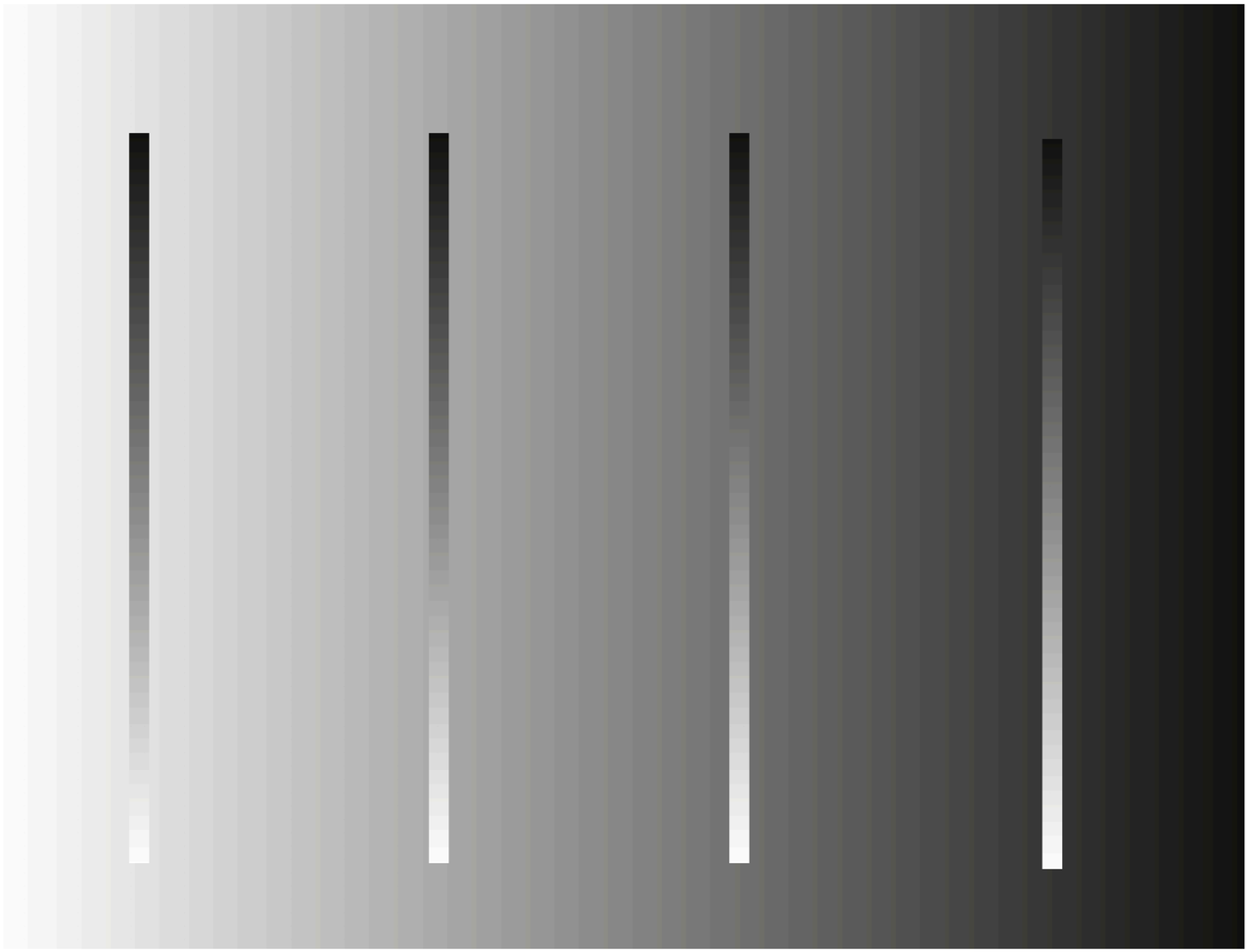


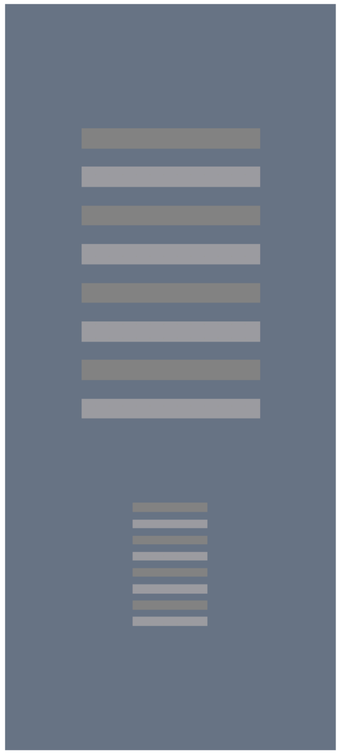
color, background and adjacent luminance can interfere with our perception

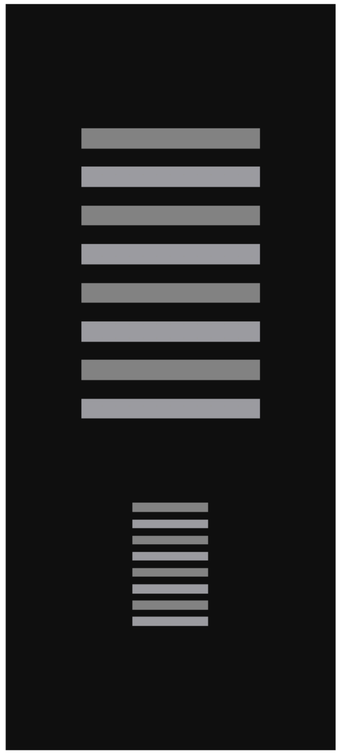
color, background and adjacent luminance can interfere with our perception



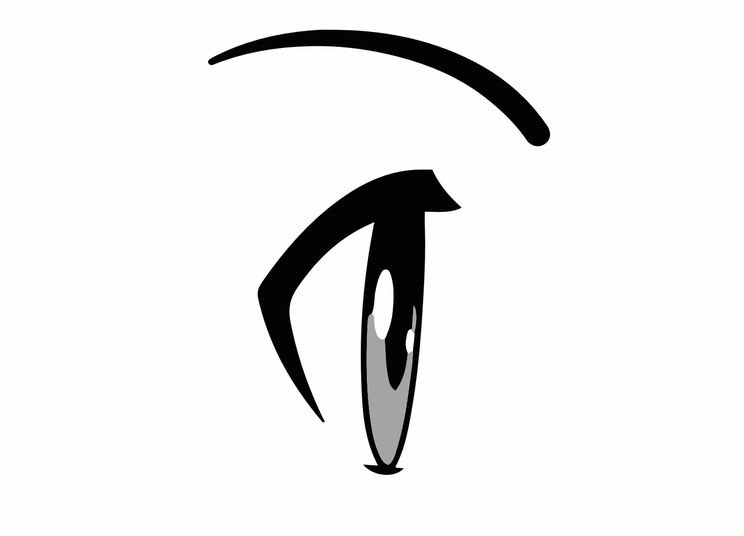
color, high foreground to background luminance contrast enhances shape, lower contrast enhances grayscale

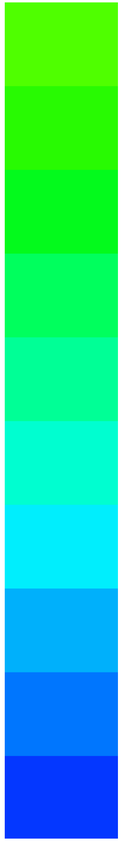






color, as with luminance, hue values in the RGB color space fail to uniformly scale across values





color, HSL colorspace is intuitive, but not perceptually uniform in each attribute

Same luminance or lightness?

HSL(250, 100, 100)

HSL(250, 100, 100)

HSL(60, 100, 100)

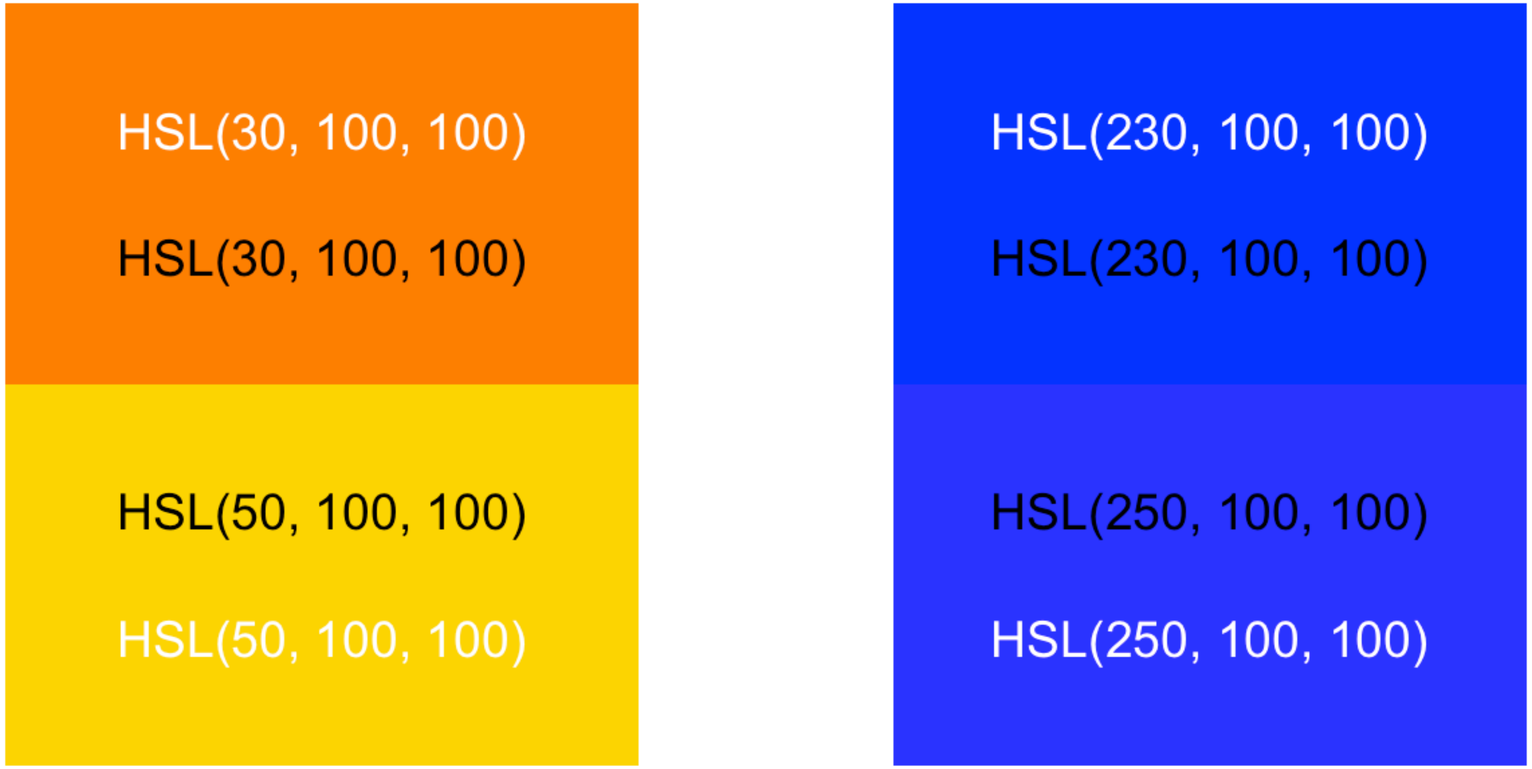
HSL(60, 100, 100)

color, HSL colorspace is intuitive, but not perceptually uniform in each attribute

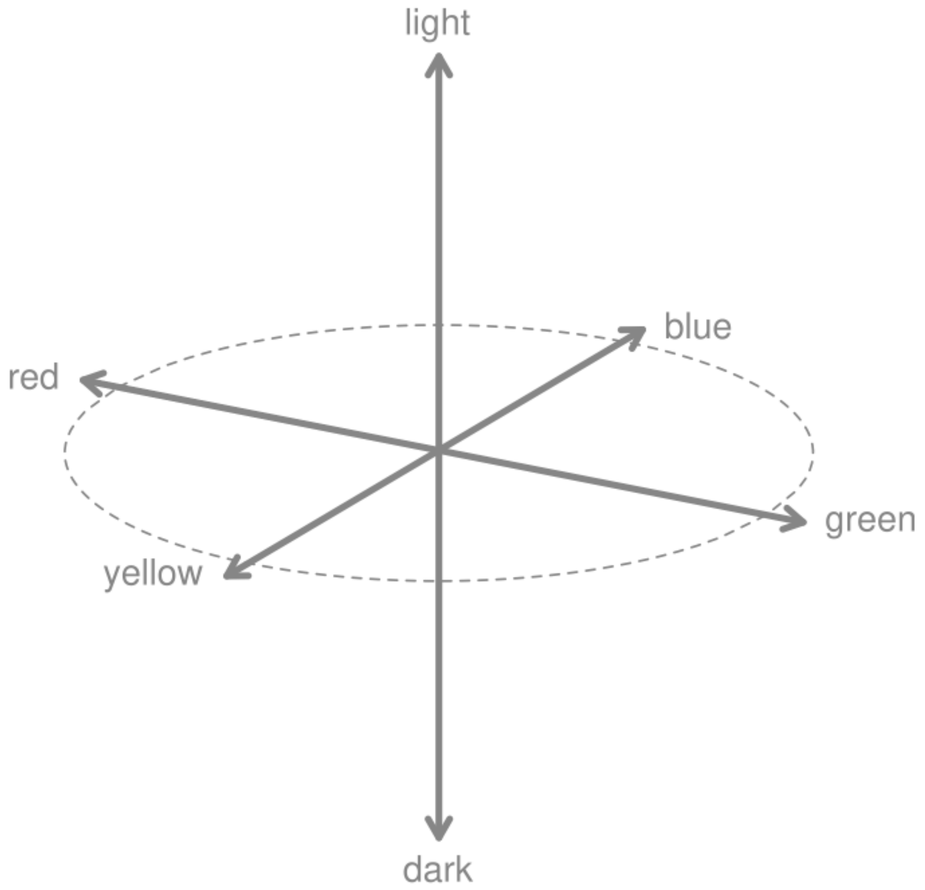
Same saturation?

color, HSL colorspace is intuitive, but not perceptually uniform in each attribute

Equal difference between hues?



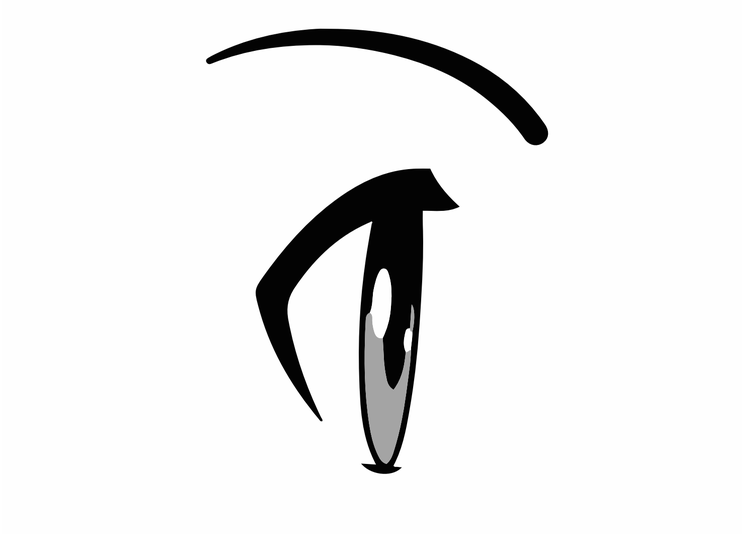
color, HSL colorspace is intuitive, but not perceptually uniform in each attribute



The International Commission on Illumination (CIE) studied human perception and re-mapped color into a space where we perceive color changes uniformly.

Their **CIELuv** color model has two dimensions — u and v — that represent color scales from red to green and yellow to blue.

color, perceptually uniform color spaces better represent quantity



color, example encoding data as perceptually uniform color attributes: R · ggplot2 · HSLuv

Load functions for mapping data to perceptually-uniform color
https://github.com/ssp3nc3r/hsluv-rcpp

```
library(HSLuv)
```

Create sample data encoded as hue, saturation, luminance

```
df <- expand.grid(H = c(20, 290),
                  S = seq(0, 100, by = 10),
                  L = seq(0, 100, by = 10))
```

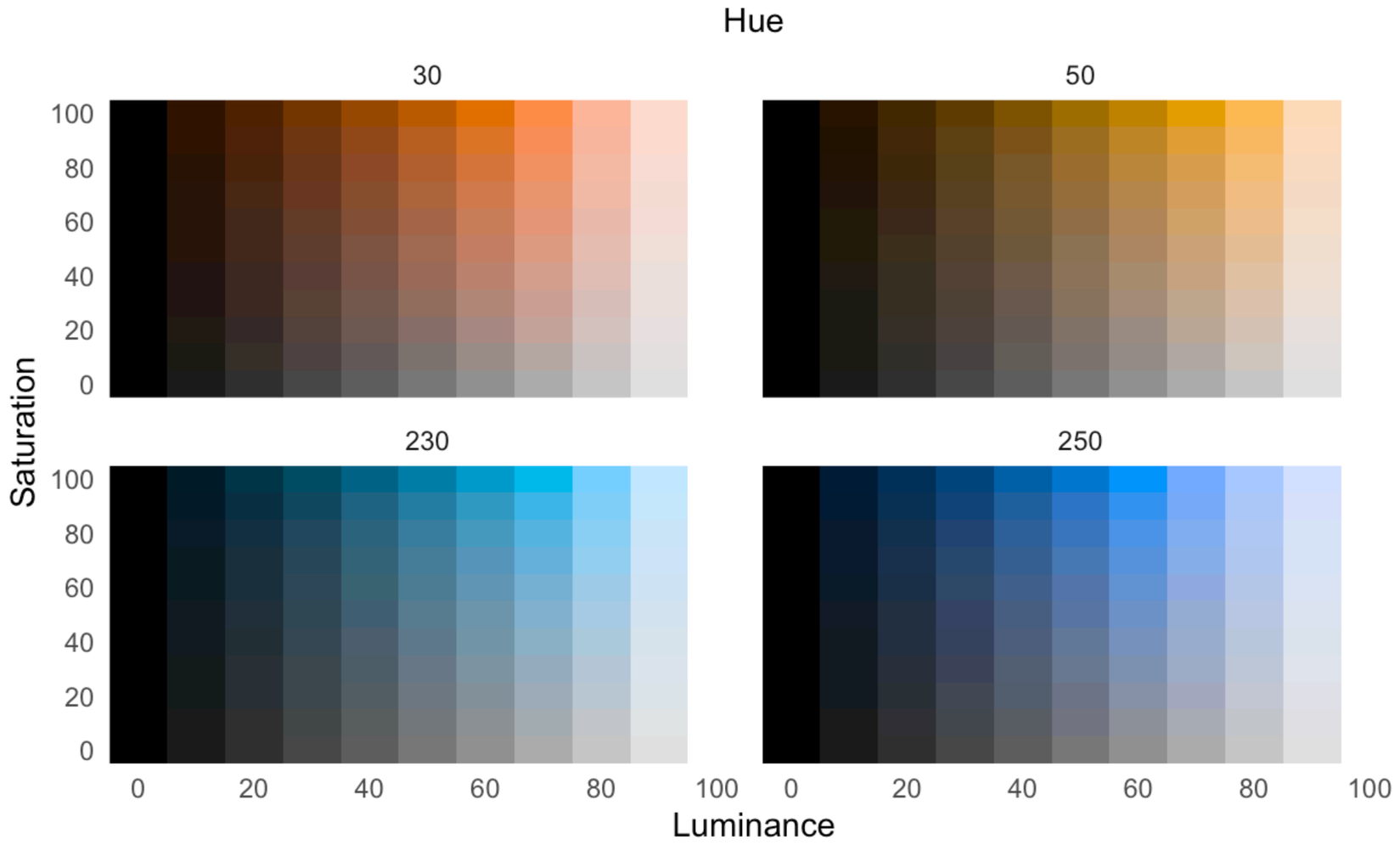
Convert HSLuv scaled values to RGB
color space as hex code #RRGGBB

```
df$colors <- with(df, hsluv_hex(H, S, L) )
```

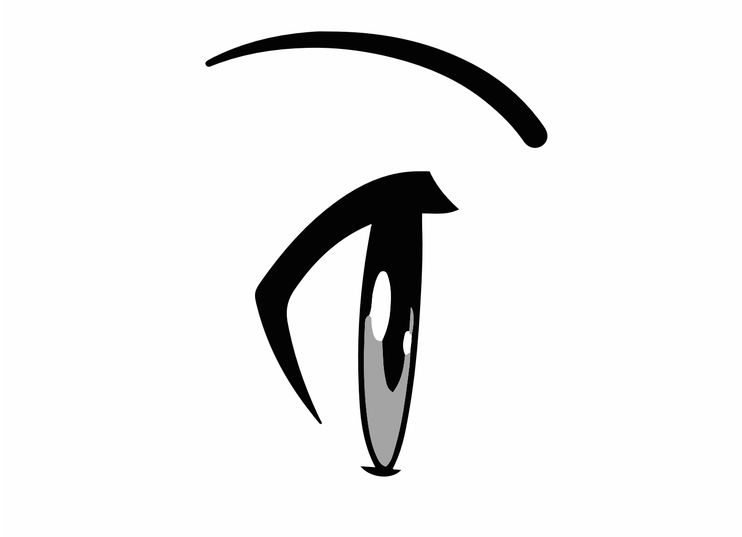
Plot data encoded as colors

```
library(ggplot2)

ggplot(df) +

  theme_minimal() +

  theme(panel.grid = element_blank(),
        axis.text.x.top = element_blank()) +

  geom_point(aes(L, S),
             color = '#eeeeee',
             fill = df$colors,
             size = 10,
             shape = 22) +

  scale_x_continuous(breaks = seq(0, 100, by = 20),
                     sec.axis = sec_axis(~., name = 'Hue')) +

  scale_y_continuous(breaks = seq(0, 100, by = 20)) +

  facet_wrap(~H) +

  labs(x = 'Luminance',
       y = 'Saturation')
```

color, example encoding data as perceptually uniform color attributes: R · ggplot2 · HSLuv

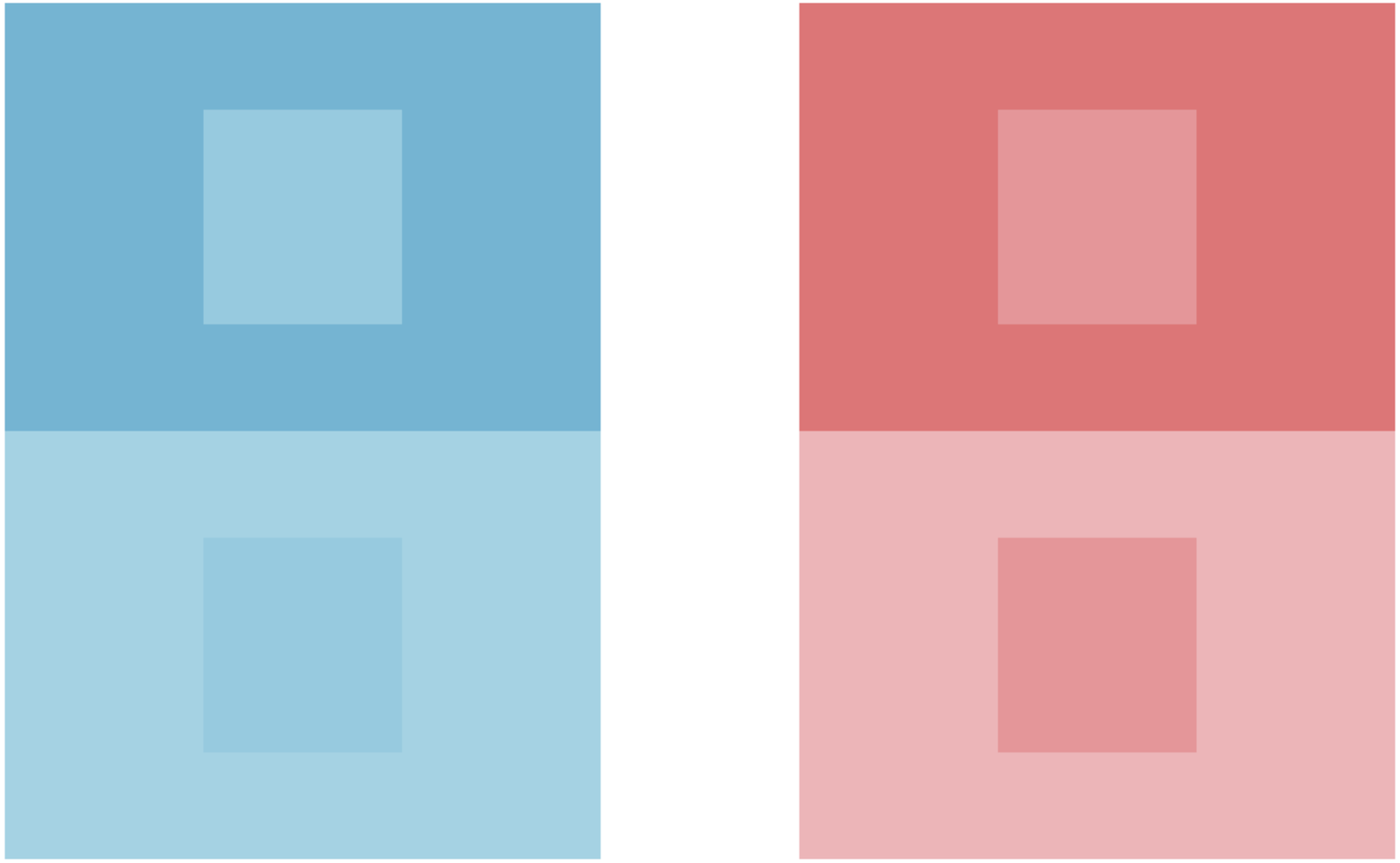


color, perceptually uniform color spaces also help in distinguishing categorical data

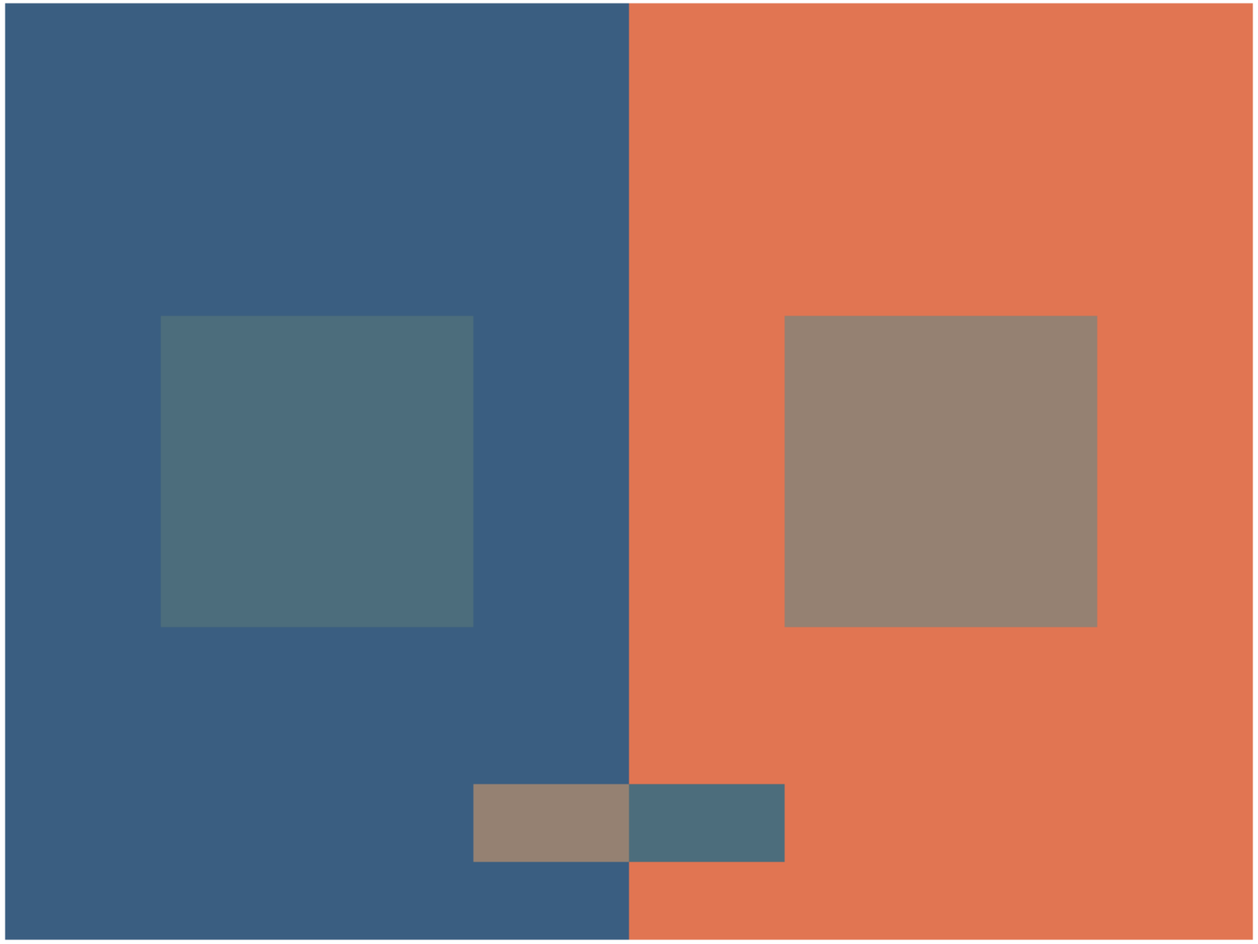


interaction of color

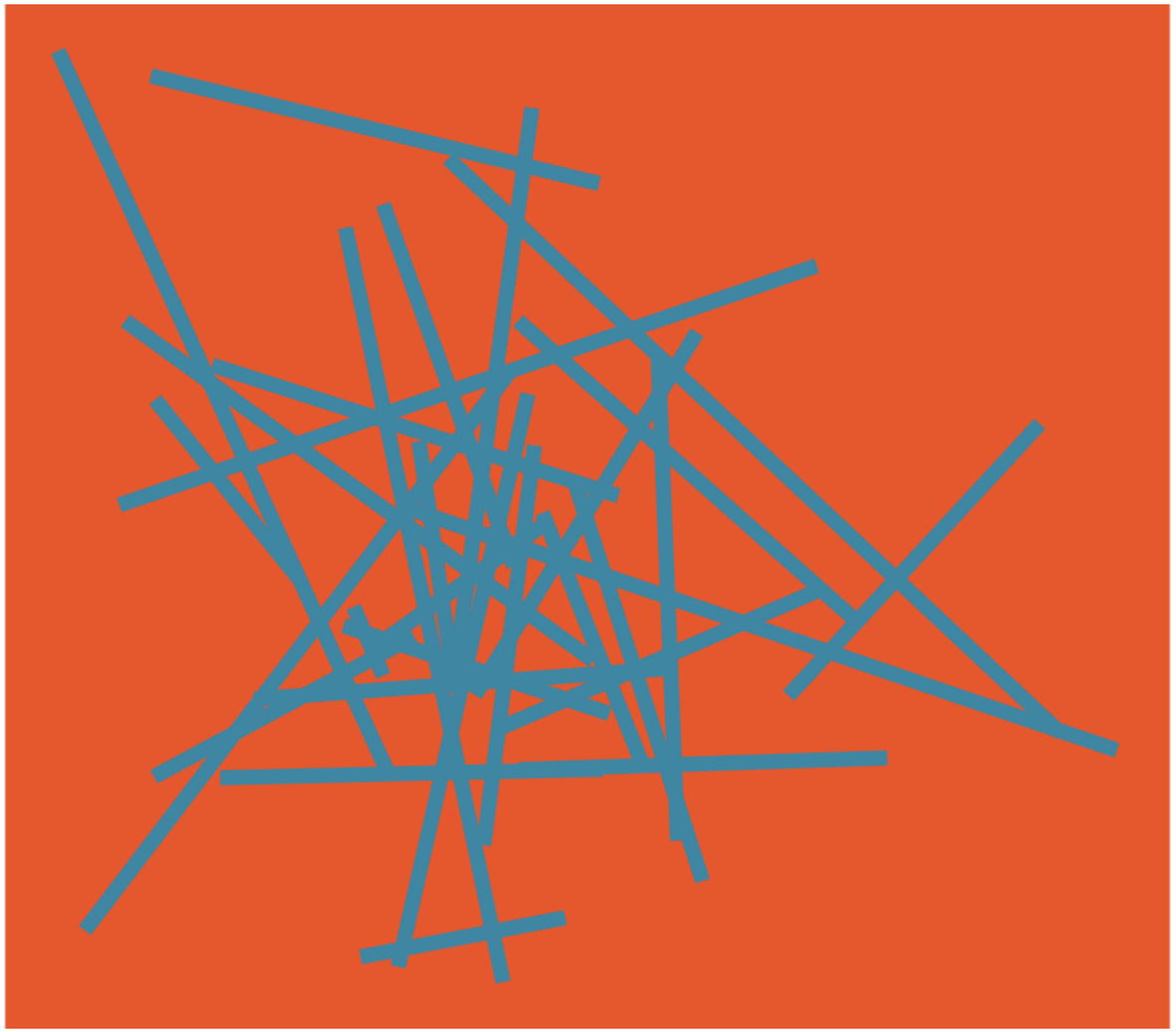
interaction of color, one color appearing as two



interaction of color, two different colors look alike



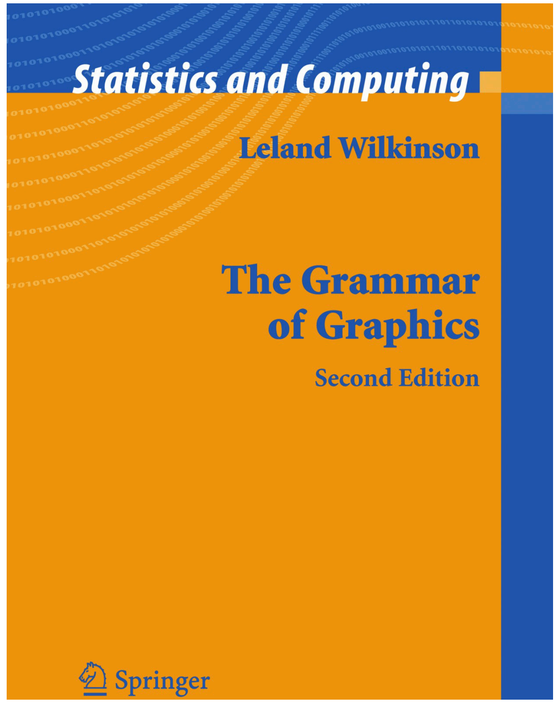
interaction of color, vibrating boundaries, occurs with contrasting hues of similar luminance



charts are *mere typologies* of graphics — don’t limit yourself

think data encodings, *not* charts …

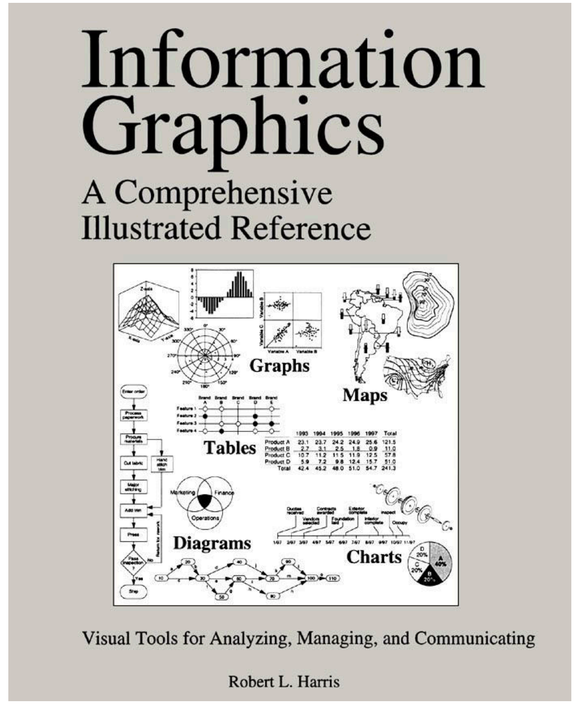
but chart typologies *can* help us learn and discuss encodings



"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.*



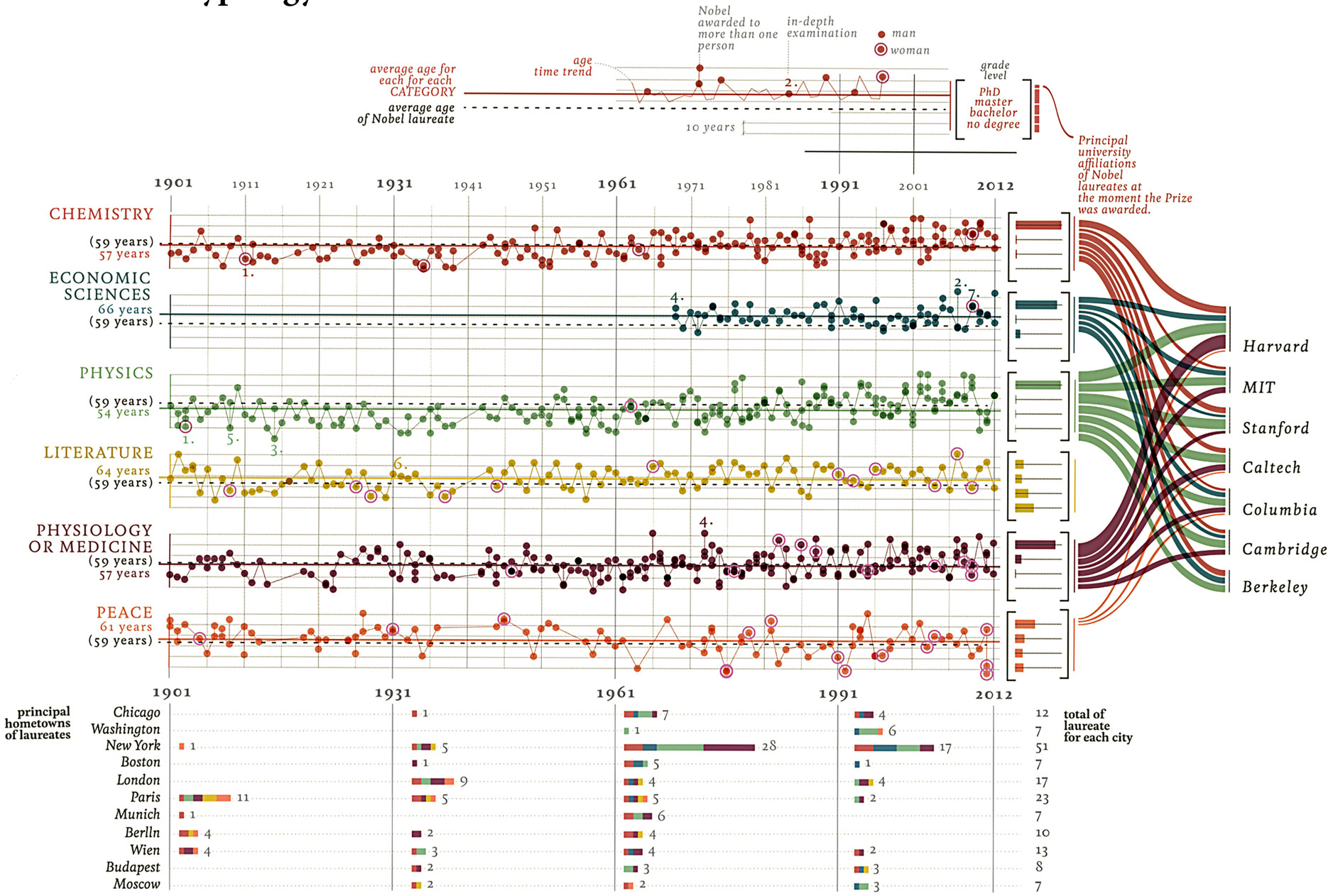
a graphics study — deconstructing Lupi's *Nobels, no degrees*, identifying typologies

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.

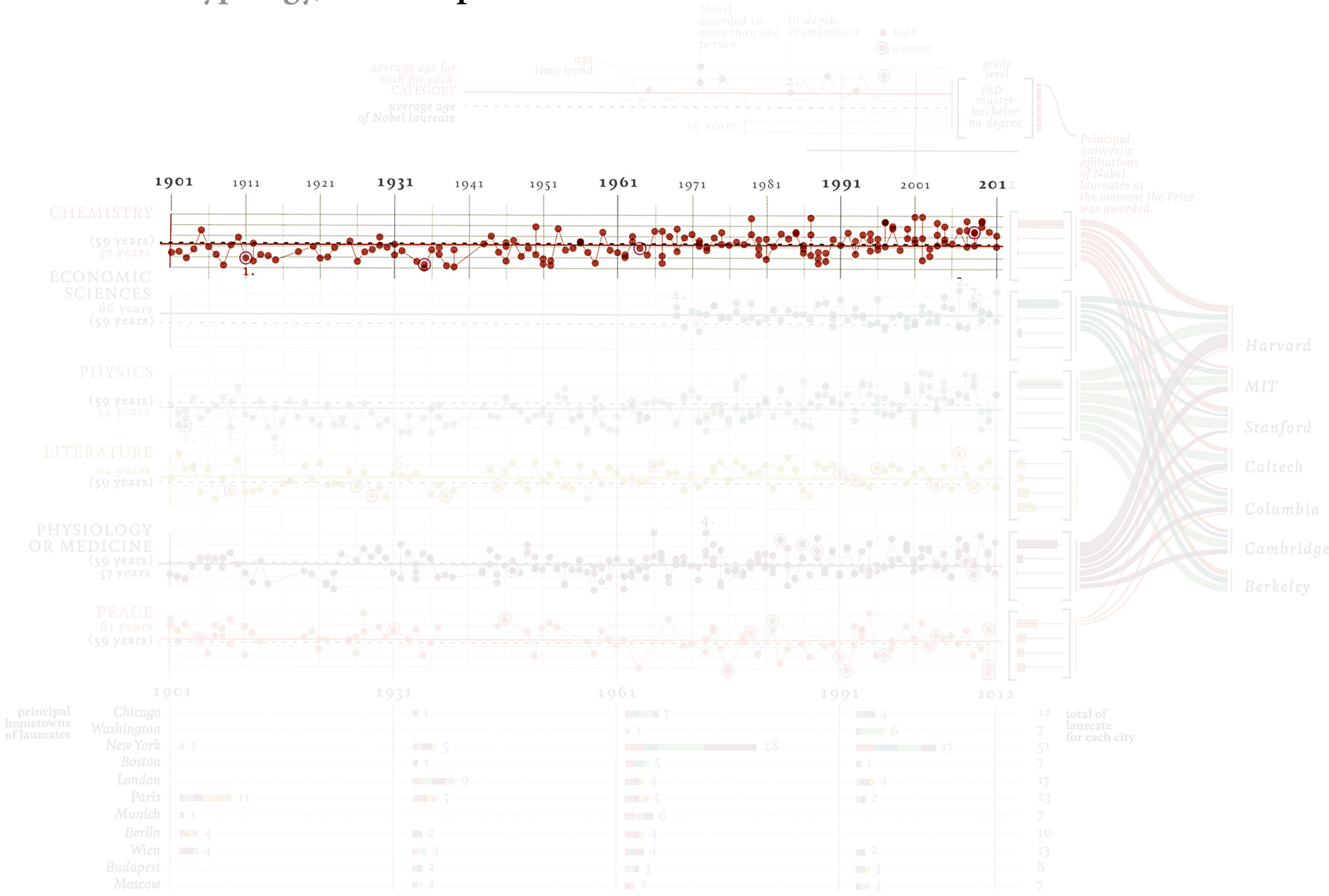
Don't be intimidated! — Just methodically experiment with encodings for each data type, then organize them.

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



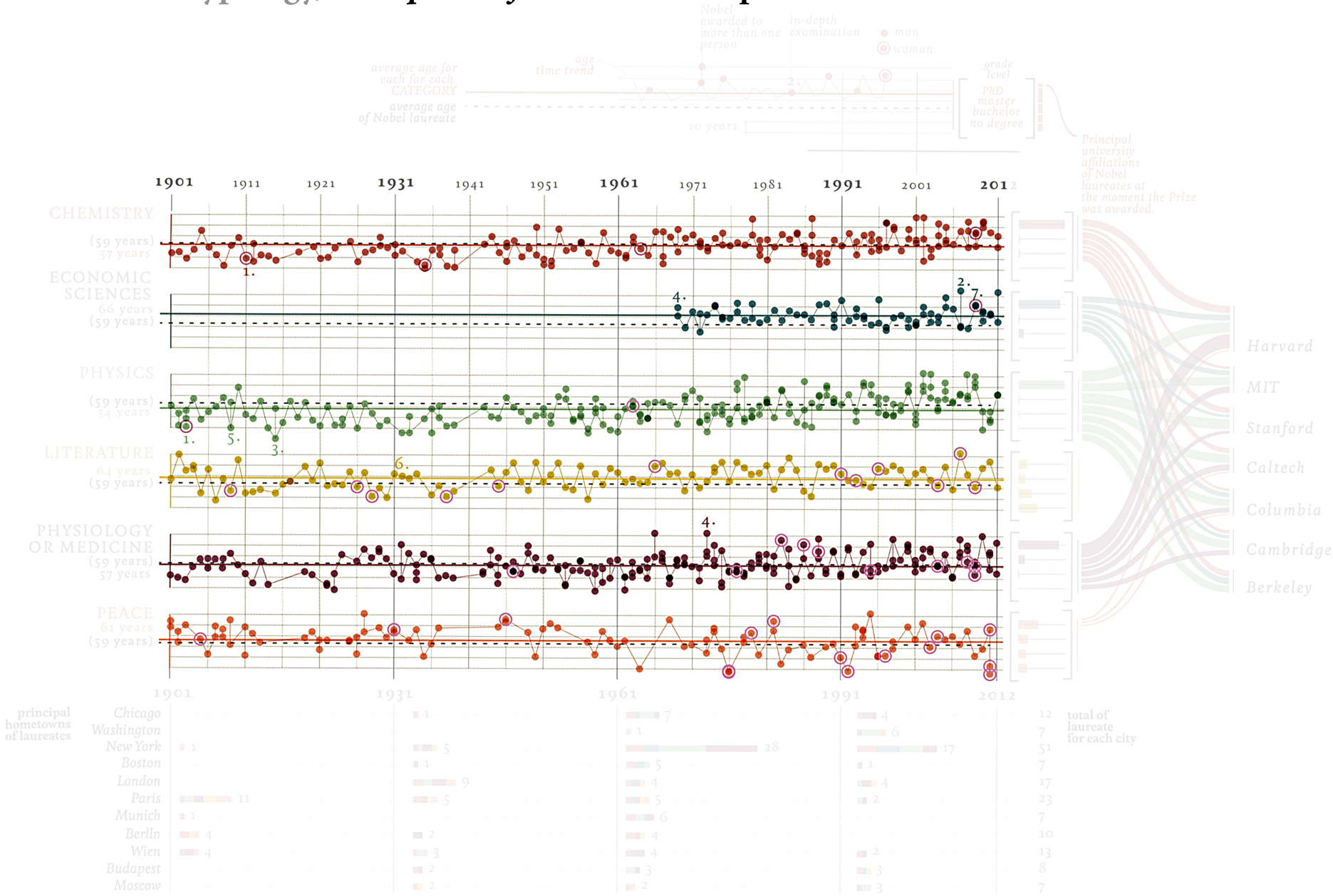
a graphics study through deconstruction and typology, a scatter plot and line charts

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


  scale_color_manual(
    values = c(category_colors, sex_colors),
    breaks = c(category_names, sex_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)
```



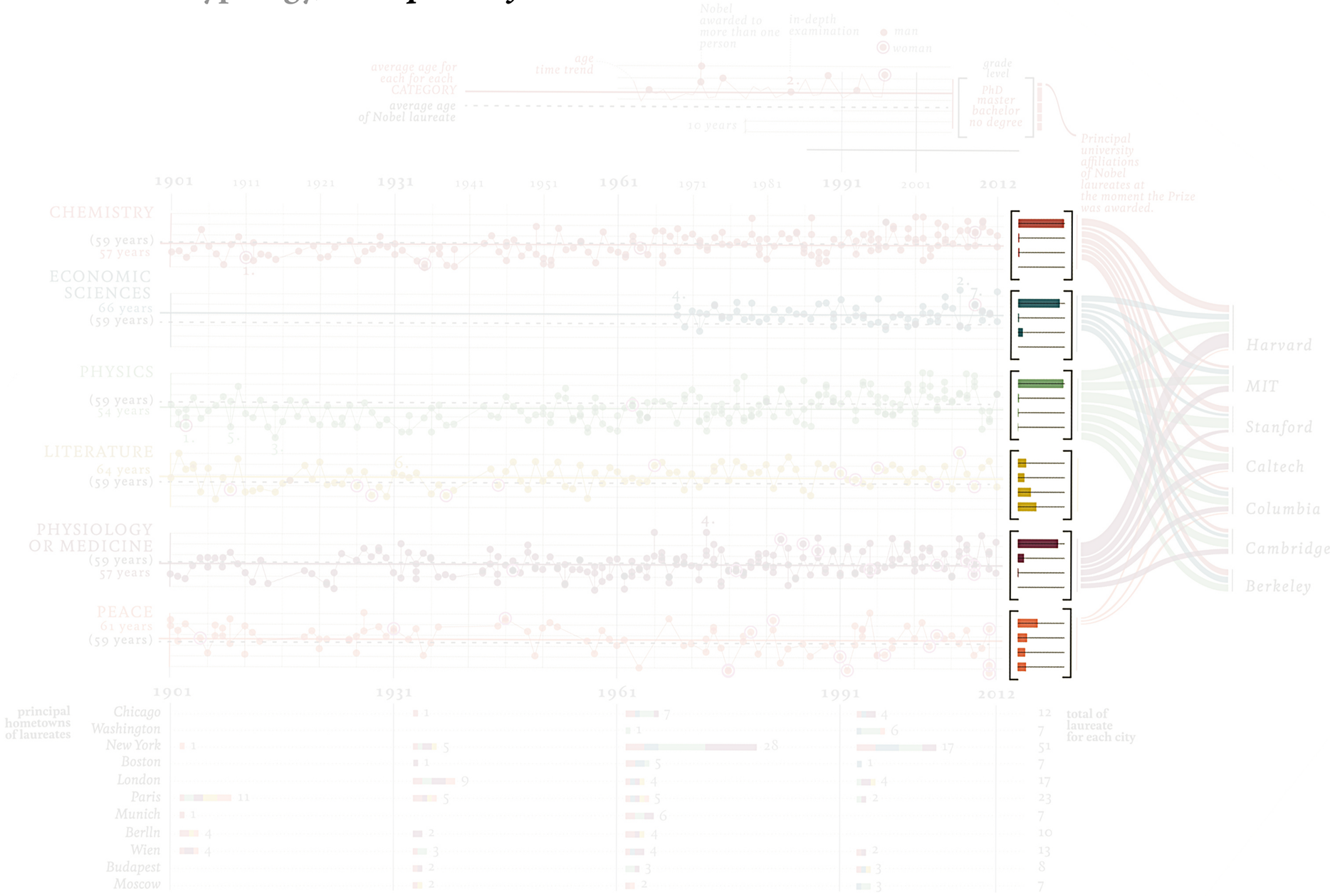
a graphics study through deconstruction and typology, *multiples or facets* of scatter plots and line charts

```
ggplot(df) +

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

  scale_color_manual(
    values = c(category_colors, sex_colors),
    breaks = c(category_names, sex_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)
```



a graphics study through deconstruction and typology, *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")
```

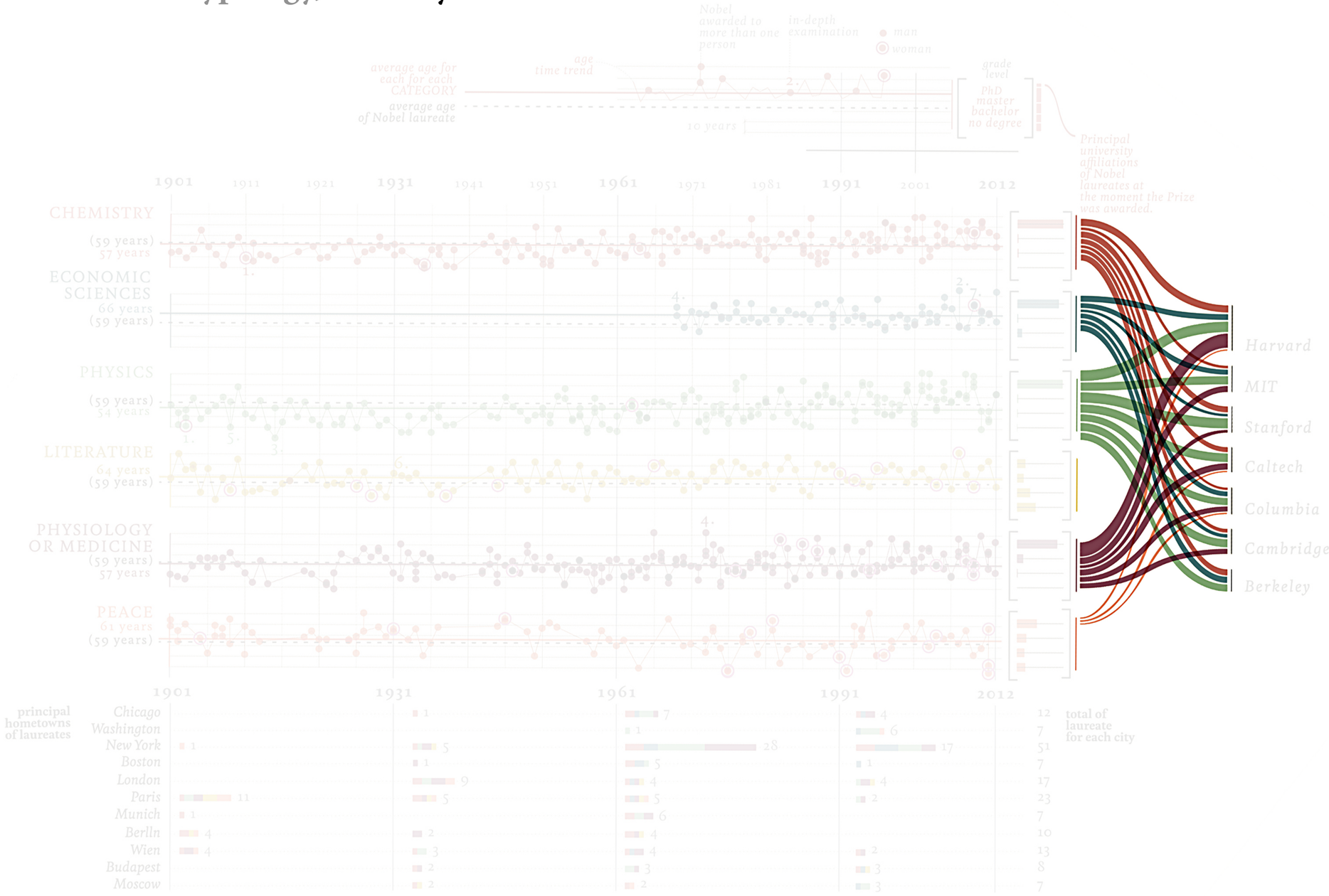


a graphics study through deconstruction and typology, *multiples or facets* of stacked bar charts

```
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)
```

a graphics study through deconstruction and typology, **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)
```



resources

References

Spencer, Scott. "Color," "Layers and separation," "Maximize information in visual displays," "Layering and opacity," Sec. 2.1.3.1-2.2.3 In *Data in Wonderland.* 2021. https://ssp3nc3r.github.io/data_in_wonderland.

Albers, Josef. *Interaction of Color*. Yale University Press, 2006, and interactive app on iPad: http://interactionofcolor.com

Boronine, Alexei. "*Color Spaces for Human Beings*." HSLuv.org (blog), March 26, 2012. https://www.hsluv.org and https://www.boronine.com/2012/03/26/Color-Spaces-for-Human-Beings/

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.

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/.

Tufte, Edward R. "Layers and Separation" and "Color and Information." In *Envisioning Information*. Graphics Press, 1990.

———. "Data-Ink and Graphical Redesign" and "Data-Ink Maximization and Graphical Design." In *The Visual Display of Quantitative Information*. Second. Graphics Press, 2001.

Ware, Colin. *Information Visualization: Perception for Design*. Fourth. Philadelphia: Elsevier, Inc, 2020.

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

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.