DSSP
Table Wrangling & Data Visualization

A. Bichat - E. Le Pennec

ÉCOLE POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY

Spring 2019

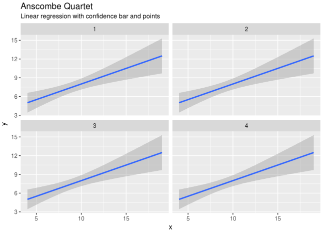
Outline

Anscombe Quartet

From

	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

to



From Table to Graph

- Need to manipulate tables.
- Need to visualize tables.

Outline

Anatomy of a Table

Tidy Table

a	b	c	d
5			A
1			B
4			A
5			B
2			B

Anatomy of a Table

Tidy Table

a	b	c	d
5			A
1			B
4			A
5			B
2			B

- Each observation forms a row.

Anatomy of a Table

Tidy Table

a	b	c	d
5			A
1			B
4			A
5			B
2			B

- Each observation forms a row.
- Each variable forms a column.

Anatomy of a Table

Tidy Table

a	b	c	d
5			A
1			B
4			A
5			B
2			B

- Each observation forms a row.
- Each variable forms a column.

- Columns are named, rows are not.
- Columns made of values of the same type.

Anatomy of a Table

Tidy Table

a	b	c	d
5			A
1			B
4			A
5			B
2			B

- Each observation forms a row.
- Each variable forms a column.
- Each type of observational unit forms a table.

- Columns are named, rows are not.
- Columns made of values of the same type.
- Codd's 3rd normal form...

Anatomy of a Table

Tidy Table

a	b	c	d
5			A
1			B
4			A
5			B
2			B

- Each observation forms a row.
- Each variable forms a column.
- Each type of observational unit forms a table.

- Columns are named, rows are not.
- Columns made of values of the same type.
- Codd's 3rd normal form...

In practice

- Definition of observation may depend on the task.
- Tidying data is a real work!

Table Actions

a	b	c	d
5			A
1			B
4			A
5			B
2			B

Verbs

Table Actions

a	b	c	d
5			A
1			B
4			A
5			B
2			B

a	b	c	d
5			A
4			A

Filter

Verbs

- Rows: Filter,

Table Actions

a	b	c	d
5			A
1			B
4			A
5			B
2			B

a	d
5	A
1	B
4	A
5	B
2	B

Select

Verbs

- Rows: Filter,
- Columns: Remove,

Table Actions

a	b	c	d
5			A
1			B
4			A
5			B
2			B
3			A

Insert

a	b	c	d
5			A
1			B
4			A
5			B
2			B

Verbs

- Rows: Filter, Insert
- Columns: Remove,

Table Actions

a	b	c	d
5			A
1			B
4			A
5			B
2			B

a	b	c	d	e
5			A	10
1			B	2
4			A	8
5			B	10
2			B	4

Add

Verbs

- Rows: Filter, Insert
- Columns: Remove, Add

Table Actions

a	b	c	d
5			A
1			B
4			A
5			B
2			B

a	b	c	d	e
5			A	10
1			B	2
4			A	8
5			B	10
2			B	4

Transform

Verbs

- Rows: Filter, Insert
- Columns: Remove, Add
- Rows/Columns: Transform, ,

Table Actions

a'
17

Summarize

a	b	c	d
5			A
1			B
4			A
5			B
2			B

Verbs

- Rows: Filter, Insert
- Columns: Remove, Add
- Rows/Columns: Transform, Summarize,

Table Actions

a	b	c	d
5			A
1			B
4			A
5			B
2			B

a	b	c	d
1			B
2			B
4			A
5			A
5			B

Sort

Verbs

- Rows: Filter, Insert
- Columns: Remove, Add
- Rows/Columns: Transform, Summarize, Sort

Table Actions

a	b	c	d
5			A
1			B
4			A
5			B
2			B

a	b	c	d
5			A
4			A

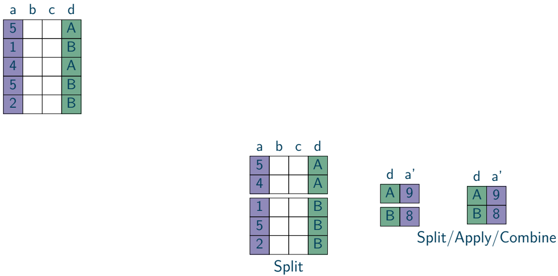
1			B
5			B
2			B

Split

Verbs

- Rows: Filter, Insert
- Columns: Remove, Add
- Rows/Columns: Transform, Summarize, Sort
- Split

Table Actions



Verbs

- Rows: Filter, Insert
- Columns: Remove, Add
- Rows/Columns: Transform, Summarize, Sort
- Split/Apply/Combine (Summarize)

Table Actions

a'
17

Summarize

a	b	c	d
5			A
1			B
4			A
5			B
2			B
3			A

Insert

a	b	c	d
5			A
1			B
4			A
5			B
2			B

a	b	c	d
5			A
4			A

Filter

a	b	c	d	e
5			A	10
1			B	2
4			A	8
5			B	10
2			B	4

Add/Transform

a	d
5	A
1	B
4	A
5	B
2	B

Select

a	b	c	d
1			B
2			B
4			A
5			A
5			B

Sort

a	b	c	d
5			A
4			A

1			B
5			B
2			B

Split

d	a'
A	9
B	8

Split/Apply/Combine

Verbs

- Rows: Filter, Insert
- Columns: Remove, Add
- Rows/Columns: Transform, Summarize, Sort
- Split/Apply/Combine (Summarize)

Table Actions

a
5
1
4
5
2
3

a	b	c	d
5			A
1			B
4			A
5			B
2			B
3			A

Insert
`INSERT/UNION`

a'
17

Summarize
`SELECT`

a	b	c	d
5			A
1			B
4			A
5			B
2			B

a	b	c	d
5			A
4			A

Filter
`WHERE`

a	b	c	d	e
5			A	10
1			B	2
4			A	8
5			B	10
2			B	4

Add/Transform
`SELECT`

a	d
5	A
1	B
4	A
5	B
2	B

Select
`SELECT`

a	b	c	d
1			B
2			B
4			A
5			A
5			B

Sort
`ORDER BY`

a	b	c	d
5			A
4			A

1			B
5			B
2			B

Split
`GROUP BY`

d	a'
A	9
B	8

Split/Apply/Combine
`SELECT + GROUP BY`

Verbs

- Rows: Filter, Insert
- Columns: Remove, Add
- Rows/Columns: Transform, Summarize, Sort
- Split/Apply/Combine (Summarize)

Table Actions

a'
17

Summarize
`summarize`

a	b	c	d
5			A
1			B
4			A
5			B
2			B

a	b	c	d
5			A
1			B
4			A
5			B
2			B
3			A

Insert
`bind_rows`

a	b	c	d
5			A
4			A

Filter
`filter`

a	b	c	d	e
5			A	10
1			B	2
4			A	8
5			B	10
2			B	4

Add/Transform
`bind_cols/mutate`

a	d
5	A
1	B
4	A
5	B
2	B

Select
`select`

a	b	c	d
1			B
2			B
4			A
5			A
5			B

Sort
`arrange`

a	b	c	d
5			A
4			A

1			B
5			B
2			B

Split
`group_by`

d	a'
A	9
B	8

Split/Apply/Combine
`group_by` + `summarize`

Verbs

- Rows: Filter, Insert
- Columns: Remove, Add
- Rows/Columns: Transform, Summarize, Sort
- Split/Apply/Combine (Summarize)

Table Actions

a'
17

Summarize
`agg`

a	b	c	d
5			A
1			B
4			A
5			B
2			B

a	b	c	d
5			A
1			B
4			A
5			B
2			B
3			A

Insert
`concat`

a	b	c	d
5			A
4			A

Filter
`query`

a	b	c	d	e
5			A	10
1			B	2
4			A	8
5			B	10
2			B	4

Add/Transform
`assign`

a	d
5	A
1	B
4	A
5	B
2	B

Select
`loc`

a	b	c	d
1			B
2			B
4			A
5			A
5			B

Sort
`sort_values`

a	b	c	d
5			A
4			A

1			B
5			B
2			B

Split
`groupby`

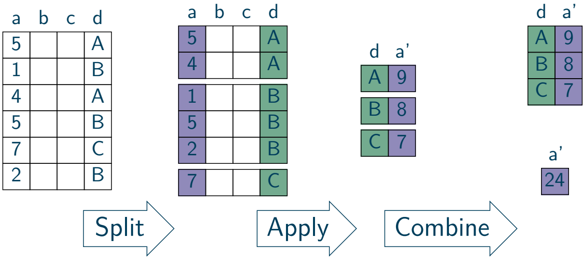
d	a'
A	9
B	8

Split/Apply/Combine
`groupby` + `agg`

Verbs

- Rows: Filter, Insert
- Columns: Remove, Add
- Rows/Columns: Transform, Summarize, Sort
- Split/Apply/Combine (Summarize)

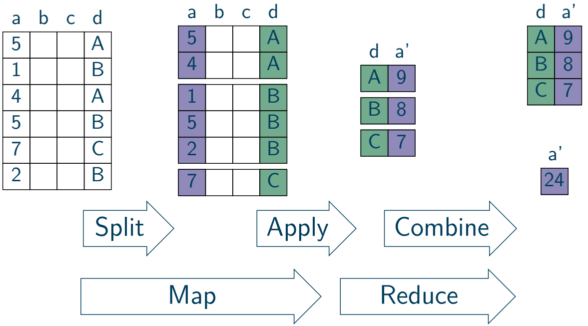
Split/Apply/Combine



Split/Apply/Combine

- **Split** the data by some **grouping** variable
- **Apply** some function to each group independently
- **Combine** the data into some **output**.

Split/Apply/Combine



Split/Apply/Combine

- **Split** the data by some **grouping** variable
- **Apply** some function to each group independently
- **Combine** the data into some **output**.

- **Map/Reduce** $\sim$ **Split/Apply/ Combine**.

Joining Two Tables

d	k	e
1	A	
2	B	
3	A	
4	D	

a	b	c	k
1			A
2			B
3			A
4			C
5			B

Join between two tables along a key

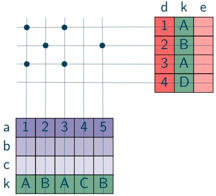
Joining Two Tables

d	k	e
1	A	
2	B	
3	A	
4	D	

a	1	2	3	4	5
b					
c					
k	A	B	A	C	B

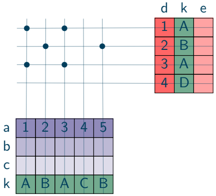
Join between two tables along a key

Joining Two Tables



Join between two tables along a key

Joining Two Tables



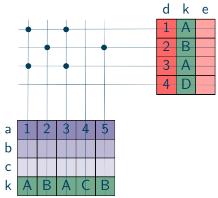
a	b	c	k	d	e
1			A	1	
1			A	3	
2			B	2	
3			A	1	
3			A	3	
5			B	2	

Inner

Join between two tables along a key

- Inner

Joining Two Tables



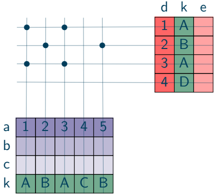
a	b	c	k	d	e
1			A	1	
1			A	3	
2			B	2	
3			A	1	
3			A	3	
4			C	?	?
5			B	2	

Left

Join between two tables along a key

- Inner
- Outer: left

Joining Two Tables



a	b	c	k	d	e
1			A	1	
3			A	1	
2			B	2	
5			B	2	
1			A	3	
3			A	3	
?	?	?	D	4	

Right

Join between two tables along a key

- Inner
- Outer: left, right

Joining Two Tables

d	k	e
1	A	
2	B	
3	A	
4	D	

a	1	2	3	4	5
b					
c					
k	A	B	A	C	B

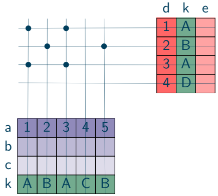
a	b	c	k	d	e
1			A	1	
1			A	3	
2			B	2	
3			A	1	
3			A	3	
4			C	?	?
5			B	2	
?	?	?	D	4	

Full

Join between two tables along a key

- Inner
- Outer: left, right, full

Joining Two Tables



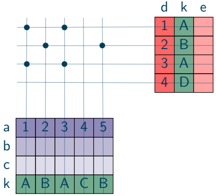
a	b	c	k
1			A
2			B
3			A
5			B

Semi

Join between two tables along a key

- Inner
- Outer: left, right, full
- Other: semi

Joining Two Tables



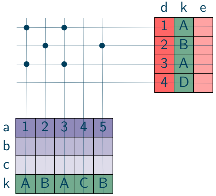
a	b	c	k
4			C

Anti

Join between two tables along a key

- Inner
- Outer: left, right, full
- Other: semi, anti

Joining Two Tables



a	b	c	k	d	e
1			A	1	
1			A	3	
2			B	2	
3			A	1	
3			A	3	
5			B	2	

Inner

a	b	c	k	d	e
1			A	1	
1			A	3	
2			B	2	
3			A	1	
3			A	3	
4			C	?	?
5			B	2	

Left

a	b	c	k	d	e
1			A	1	
3			A	1	
2			B	2	
5			B	2	
1			A	3	
3			A	3	
?	?	?	D	4	

Right

a	b	c	k	d	e
1			A	1	
1			A	3	
2			B	2	
3			A	1	
3			A	3	
4			C	?	?
5			B	2	
?	?	?	D	4	

Full

a	b	c	k
1			A
2			B
3			A
5			B

Semi

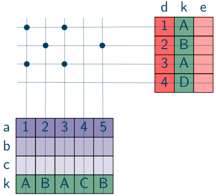
a	b	c	k
4			C

Anti

Join between two tables along a key

- Inner
- Outer: left, right, full
- Other: semi, anti

Joining Two Tables



a	b	c	k	d	e
1			A	1	
1			A	3	
2			B	2	
3			A	1	
3			A	3	
5			B	2	

Inner
`JOIN`

a	b	c	k	d	e
1			A	1	
1			A	3	
2			B	2	
3			A	1	
3			A	3	
4			C	?	?
5			B	2	

Left
`LEFT JOIN`

a	b	c	k	d	e
1			A	1	
3			A	1	
2			B	2	
5			B	2	
1			A	3	
3			A	3	
?	?	?	D	4	

Right
`RIGHT JOIN`

a	b	c	k	d	e
1			A	1	
1			A	3	
2			B	2	
3			A	1	
3			A	3	
4			C	?	?
5			B	2	
?	?	?	D	4	

Full
`FULL JOIN`

a	b	c	k
1			A
2			B
3			A
5			B

Semi
`SEMI JOIN`

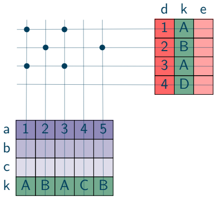
a	b	c	k
4			C

Anti
`ANTI JOIN`

Join between two tables along a key

- Inner
- Outer: left, right, full
- Other: semi, anti

Joining Two Tables



a	b	c	k	d	e
1			A	1	
1			A	3	
2			B	2	
3			A	1	
3			A	3	
5			B	2	

Inner
`inner_join`

a	b	c	k	d	e
1			A	1	
1			A	3	
2			B	2	
3			A	1	
3			A	3	
4			C	?	?
5			B	2	

Left
`left_join`

a	b	c	k	d	e
1			A	1	
3			A	1	
2			B	2	
5			B	2	
1			A	3	
3			A	3	
?	?	?	D	4	

Right
`right_join`

a	b	c	k	d	e
1			A	1	
1			A	3	
2			B	2	
3			A	1	
3			A	3	
4			C	?	?
5			B	2	
?	?	?	D	4	

Full
`full_join`

a	b	c	k
1			A
2			B
3			A
5			B

Semi
`semi_join`

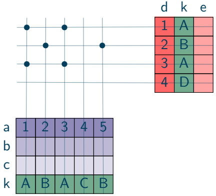
a	b	c	k
4			C

Anti
`anti_join`

Join between two tables along a key

- Inner
- Outer: left, right, full
- Other: semi, anti

Joining Two Tables



a	b	c	k	d	e
1			A	1	
1			A	3	
2			B	2	
3			A	1	
3			A	3	
5			B	2	

Inner
`merge` (inner)

a	b	c	k	d	e
1			A	1	
1			A	3	
2			B	2	
3			A	1	
3			A	3	
4			C	?	?
5			B	2	

Left
`merge` (left)

a	b	c	k	d	e
1			A	1	
3			A	1	
2			B	2	
5			B	2	
1			A	3	
3			A	3	
?	?	?	D	4	

Right
`merge` (right)

a	b	c	k	d	e
1			A	1	
1			A	3	
2			B	2	
3			A	1	
3			A	3	
4			C	?	?
5			B	2	
?	?	?	D	4	

Full
`merge` (full)

a	b	c	k
1			A
2			B
3			A
5			B

Semi
`merge` (inner) + `loc`

a	b	c	k
4			C

Anti
`merge` (inner) + `loc`

Join between two tables along a key

- Inner
- Outer: left, right, full
- Other: semi, anti

Pivot

a	b	c	d	e
E	A	5		
E	B	1		
F	A	4		
F	B	5		
G	C	7		
G	B	2		

- Several ways to organize information.

Table Reshaping

Pivot

a	b	c	d	e
E	A	5		
E	B	1		
F	A	4		
F	B	5		
G	C	7		
G	B	2		

Pivot

a	A	B	C
E	5	1	?
F	5	5	?
G	?	2	7

- Several ways to organize information.

Table Reshaping

- Pivot: Categories to columns

Pivot

a	b	c	d	e
E	A	5		
E	B	1		
F	A	4		
F	B	5		
G	C	7		
G	B	2		

Pivot

a	A	B	C
E	5	1	?
F	5	5	?
G	?	2	7

Pivot/Unpivot

a	b	c
E	A	5
E	B	1
F	A	4
F	B	5
G	C	7
G	B	2

- Several ways to organize information.

Table Reshaping

- Pivot: Categories to columns
- Unpivot: Columns to categories

Pivot

a	b	c	d	e
E	A	5		
E	B	1		
F	A	4		
F	B	5		
G	C	7		
G	B	2		

Pivot

`PIVOT`

a	A	B	C
E	5	1	?
F	5	5	?
G	?	2	7

Pivot/Unpivot

`PIVOT/UNPIVOT`

a	b	c
E	A	5
E	B	1
F	A	4
F	B	5
G	C	7
G	B	2

- Several ways to organize information.

Table Reshaping

- Pivot: Categories to columns
- Unpivot: Columns to categories

Pivot

a	b	c	d	e
E	A	5		
E	B	1		
F	A	4		
F	B	5		
G	C	7		
G	B	2		

Pivot

spread

a	A	B	C
E	5	1	?
F	5	5	?
G	?	2	7

Pivot/Unpivot

spread/gather

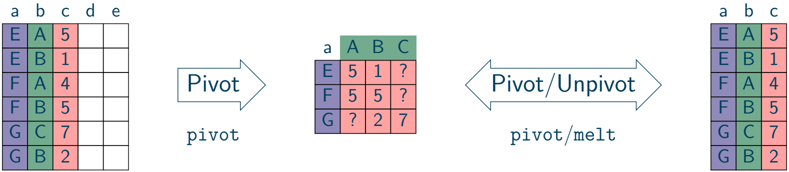
a	b	c
E	A	5
E	B	1
F	A	4
F	B	5
G	C	7
G	B	2

- Several ways to organize information.

Table Reshaping

- Pivot: Categories to columns
- Unpivot: Columns to categories

Pivot

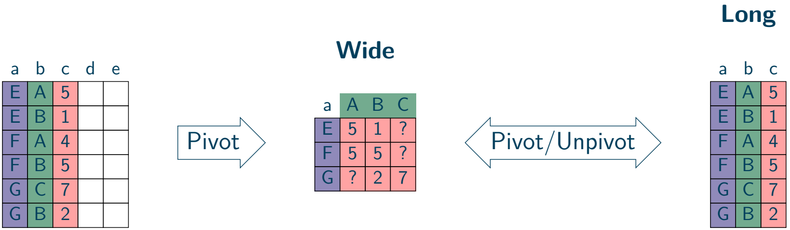


- Several ways to organize information.

Table Reshaping

- Pivot: Categories to columns
- Unpivot: Columns to categories

Pivot



- Several ways to organize information.

Table Reshaping

- Pivot: Categories to columns
- Unpivot: Columns to categories

- Wide / Long format

Pivot

a	b	c	d	e
E	A	5		
E	B	1		
F	A	4		
F	B	5		
G	C	7		
G	B	2		

Pivot

a	A	B	C
E	5	1	?
F	5	5	?
G	?	2	7

Pivot/Unpivot

a	b	c
E	A	5
E	B	1
F	A	4
F	B	5
G	C	7
G	B	2

- Several ways to organize information.

Table Reshaping

- Pivot: Categories to columns
- Unpivot: Columns to categories

- Best format depends on the task

Table Sources

Local Files: `CSV`, `JSON`, `xls`,...

- Dedicated import(/export) libraries
- Often weekly typed.
- Cleaning often required.

- Typed variants exists: `arrow`, `feather`...

Database: `SQL`, `NoSQL`,...

- Dedicated DB connectors.
- `SQL` as a common language.
- Often already cleaned.

- Local files may be accessed through a DB interface...

Web: `HTML`

- Web scraping library
- Often a lot of cleaning

- Web API $\sim$ Local files...

Outline

Outline

Why Data Visualization?

Data visualization can:

- provide a clear understanding of patterns in data
- detect hidden structures in data
- condense information

Why Data Visualization?

Data visualization can:

- provide a clear understanding of patterns in data
- detect hidden structures in data
- condense information

- Anscombe's quartet example:

Anscombe Quartet

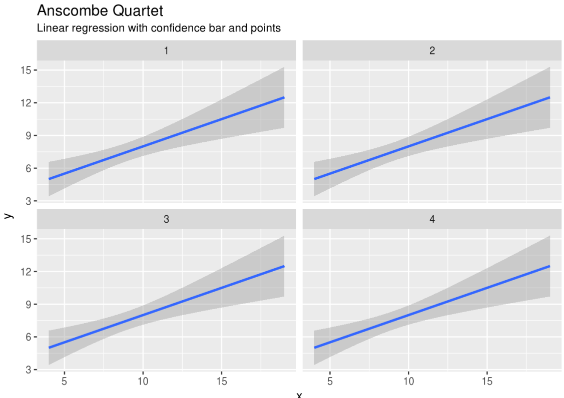
	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

Why Data Visualization?

Data visualization can:

- provide a clear understanding of patterns in data
- detect hidden structures in data
- condense information

- Anscombe's quartet example:

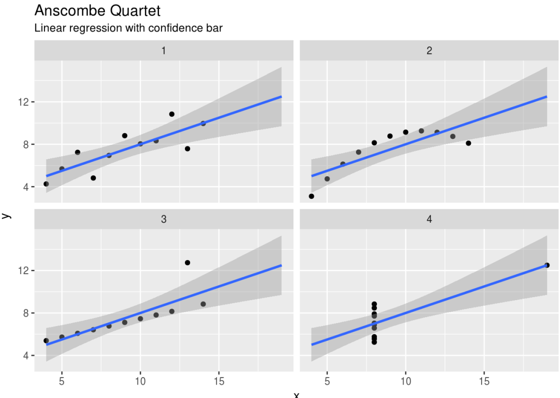


Why Data Visualization?

Data visualization can:

- provide a clear understanding of patterns in data
- detect hidden structures in data
- condense information

- Anscombe's quartet example:



Data Visualization

Focus of today

- Standard data visualization techniques,
- Review of various graphical techniques,
- Principle of good data presentation,
- Example of implementation with R.

Not the focus of this lecture

- *Infographics*
- Cognitive aspect of data perception...

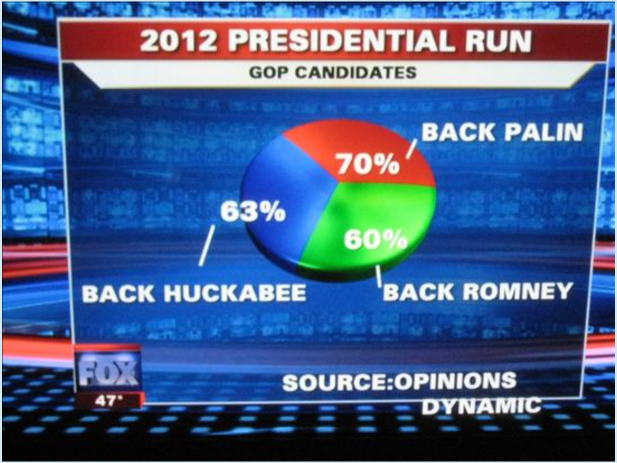
Goal

- Exposure to various plotting techniques.
- Proof of concept with R.
- *Visualize* the power of appropriate data graphics techniques

Outline

Bad Data Visualization

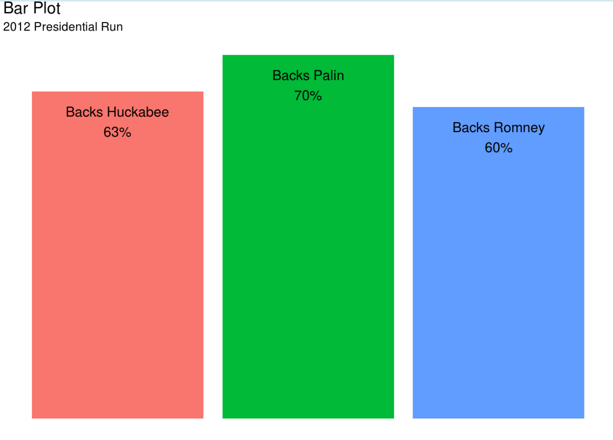
No Comment!



Source: Fox News

Bad Data Visualization

A possible fix



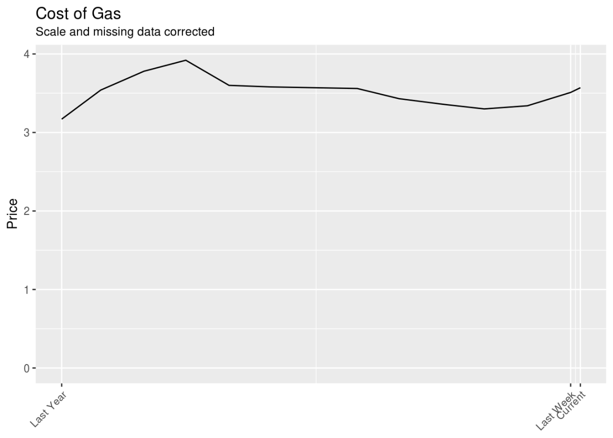
Bad Data Visualization

Scale Issue + Missing Data

Source: Fox News

Bad Data Visualization

Scale Issue Corrected + Missing Data Corrected



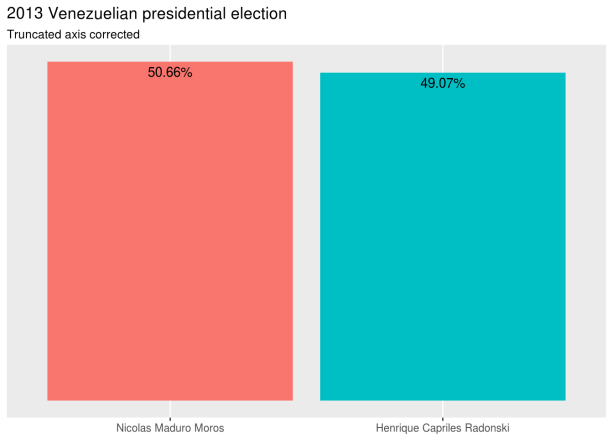
Bad Data Visualization

Truncated Axis Issue

Source: Venezolana de Television

Bad Data Visualization

Truncated Axis Issue Corrected



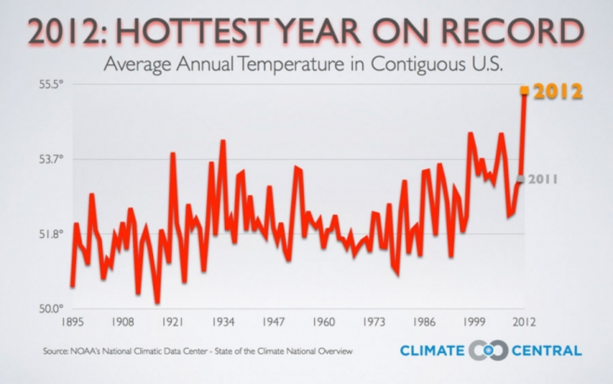
Bad Data Visualization

Scale Issue + Selection Issue

Source: Climate Central

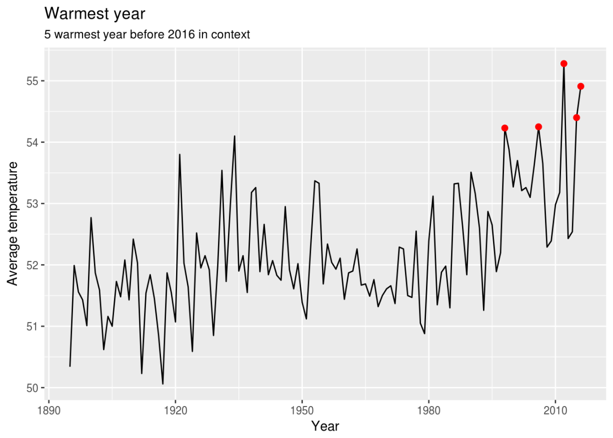
Bad Data Visualization

Scale Issue Corrected + Selection Issue Corrected



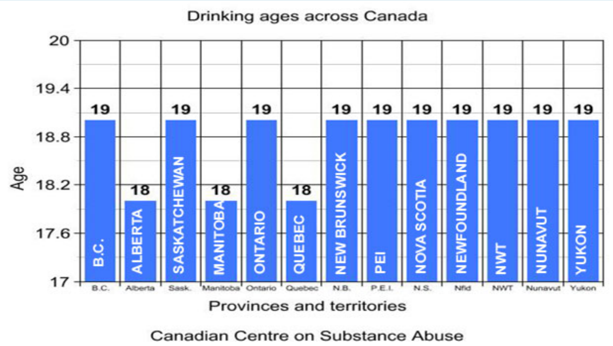
Source: Climate Central

Scale Issue Corrected + Selection Issue Corrected (2016)



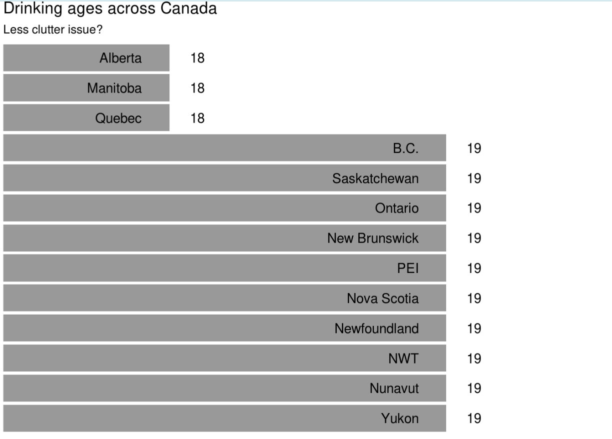
Bad Data Visualization

Truncated Axis + Clutter Issue

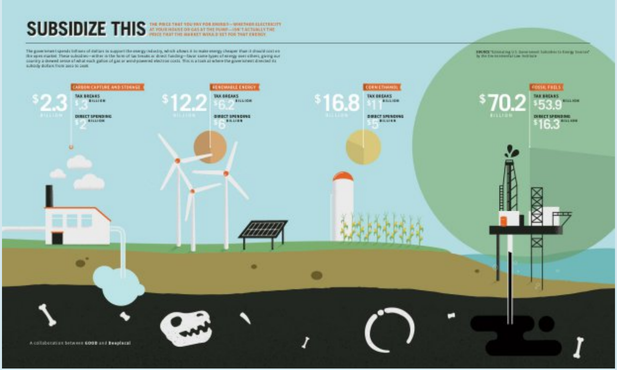


Source: Canadian Center on Substance Abuse

Truncated Axis + Clutter Issue Corrected?

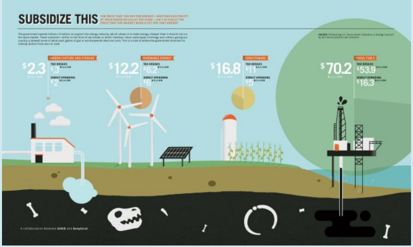


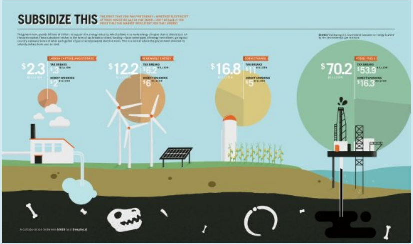
Area Issue



Source: Good Magazine

Area Issue Corrected





Source: Good Magazine

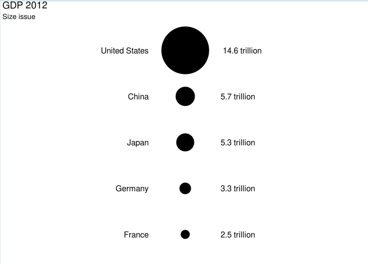
Bad Data Visualization

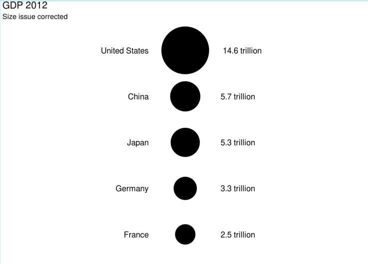
Area Issue

Source: Contrast

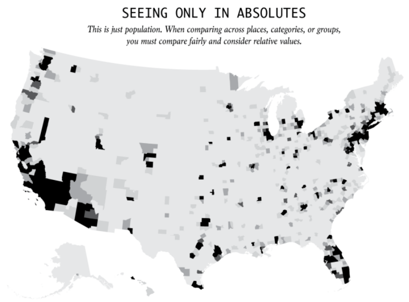
Bad Data Visualization

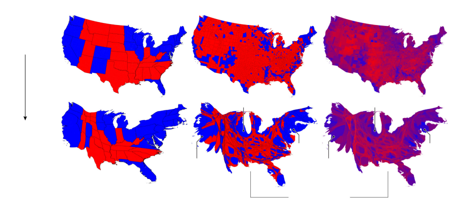
Area Issue Corrected





Map Issue

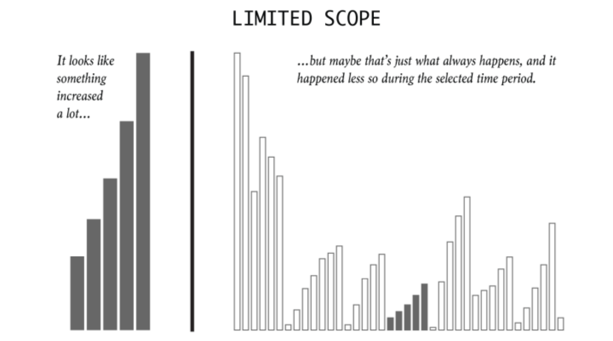




Sources: Flowing Data / M. Newman

Bad Data Visualization

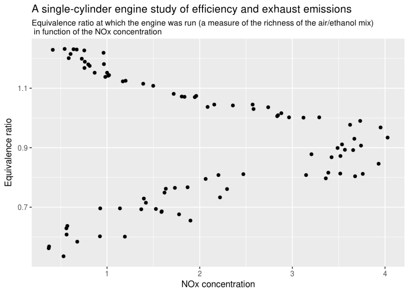
Limited Scope Issue (Corrected)



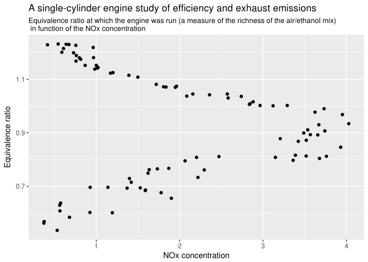
Source: Flowing Data

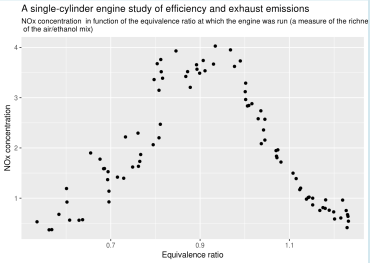
Bad Data Visualization

Unusual Axis Issue

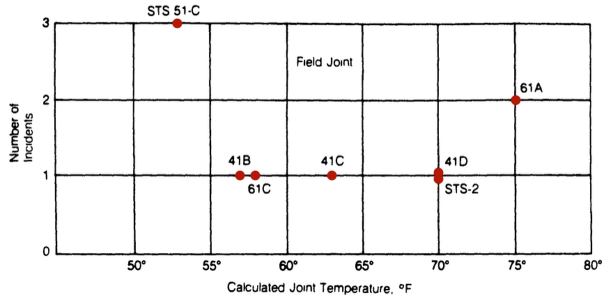


Unusual Axis Issue Corrected



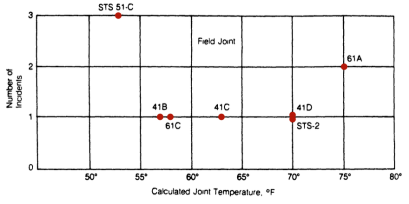


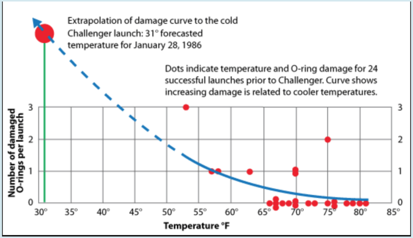
Catastrophic Issue



Source: NASA

Catastrophic Issue Corrected





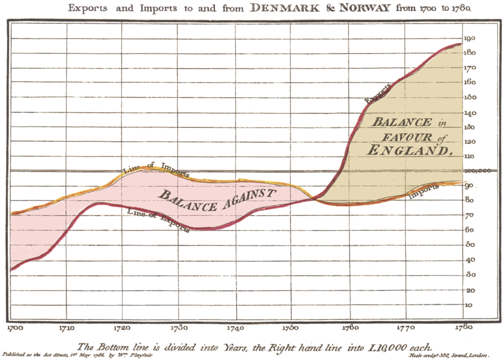
Sources: NASA / E. Tufte

Outline

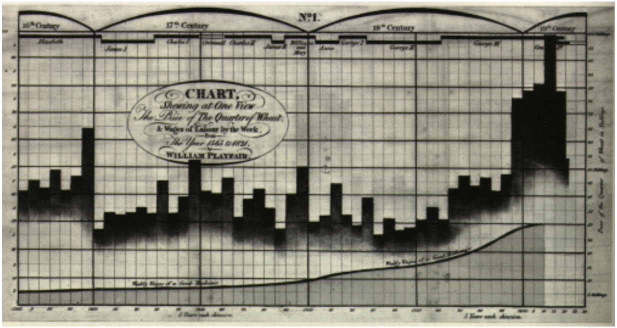
Playfair

William Playfair (1759-1823)

- Generally viewed as the inventor of most of the common graphical forms used to display data: line plots, bar chart and pie chart



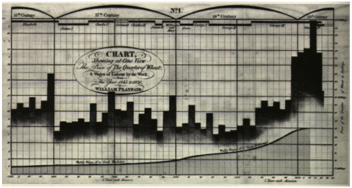
Playfair

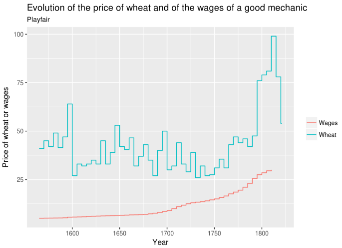


- Unfortunately often flawed...

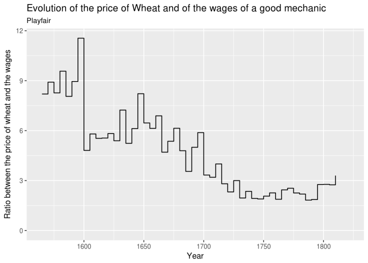
Source: W. Playfair

Playfair





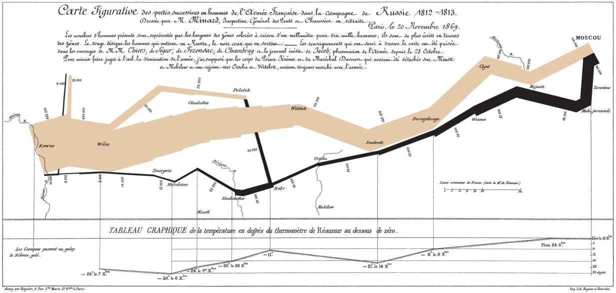
Issue Resolved?



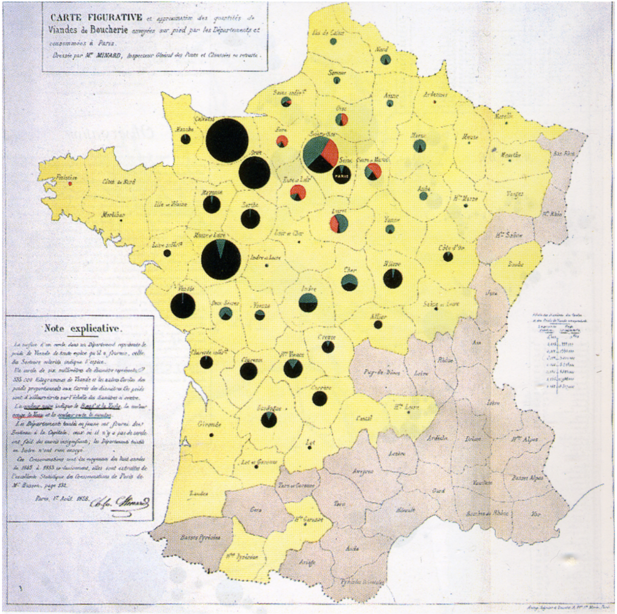
Source: W. Playfair

Charles Minard (1781-1870)

- contributed significantly in the field of information graphics in civil engineering and statistics and in particular in geographic maps.



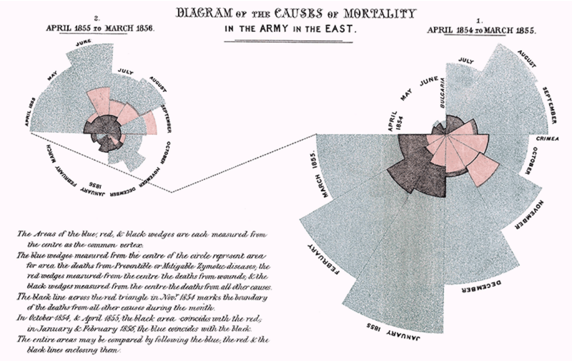
Minard



Source: Ch. Minard

Florence Nightingale (1820-1910)

- Mostly famous as a the mother of modern nursing. She also contributed to the use of graphical representation.



Source: F. Nightingale

Snow

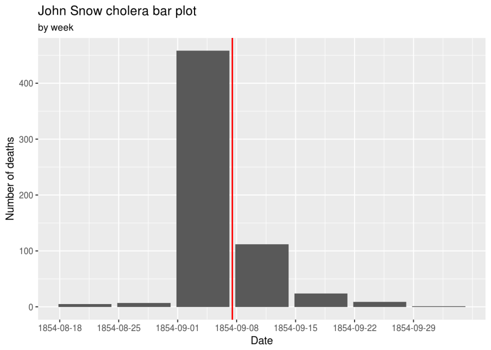
John Snow (1813-1858)

- An English physician famous for tracing the source of a cholera outbreak in London.

Source: J. Snow

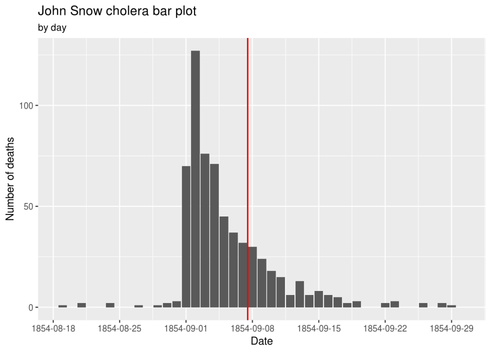
Snow

Story


John Snow cholera bar plot
by week
Number of deaths
400
300
200
100
0
1854-08-18
1854-08-25
1854-09-01
1854-09-08
1854-09-15
1854-09-22
1854-09-29
Date

Snow

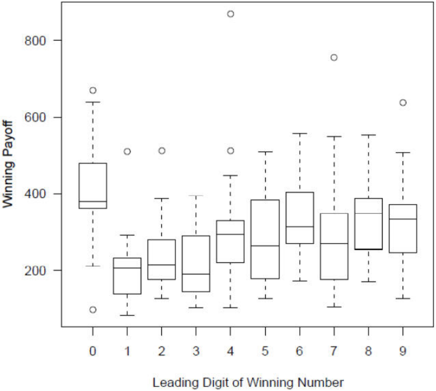
Reality



Fisher and Tukey

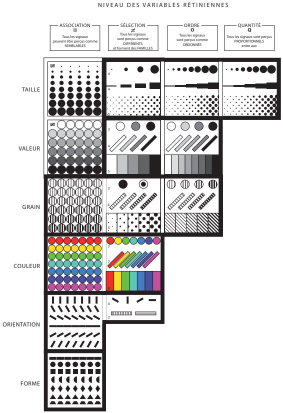
Ronald Fisher (1890-1962) and John Tukey (1915-2000)

- Advance graphical methods for the analysis of data.
- Fisher: plot the data to understand relationships.
- Tukey promoted Exploratory Data Analysis!
- Tukey created the box plot and the stem and leaf plot.



Source: J. Tukey

Bertin



Jacques Bertin, "Sémiologie Graphique", 1973.

Jacques Bertin (1918-2010)

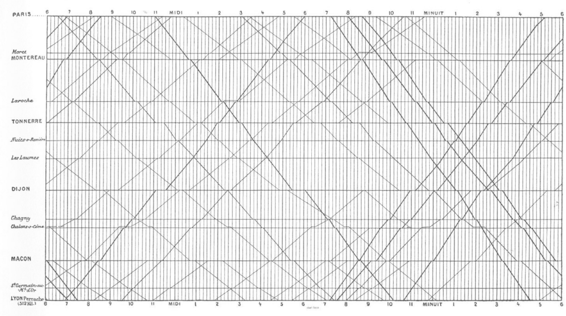
- *sémiologie graphique*!
- Systematic system of sign for information transmission.

Source: J. Bertin

Tufte

Edward Tufte (1942-)

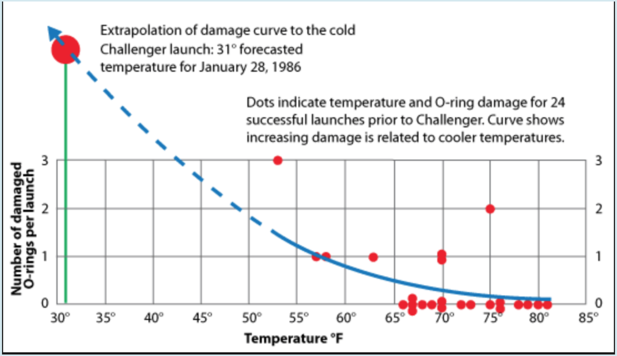
- Probably the most widely know works on data visualization.
- Highly compressed, elegant, and informative data, as expressed in dense printed graphics.
- Importance of *beauty* aspect...



E.J. Marey (1885)

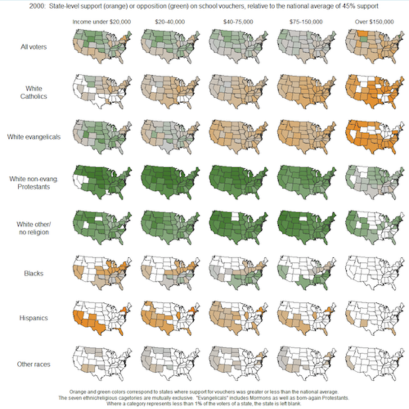
Source: E. Tufte

Challenger corrected!



Source: E. Tufte

Small Multiples



Source: A. Gelman

Tufte

Sparklines

Example sparklines in small multiple

Index	Day	Value	Change
Dow Jones		10765.45	–32.82 (–0.30%)
S&P 500		1256.92	–8.10 (–0.64%)

Example sparklines of the Dow Jones Industrial Average and S&P 500 on February 7, 2006

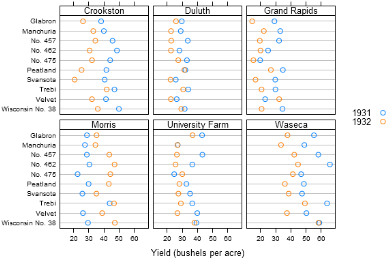
Source: E. Tufte

Tufte

Tufter (now popular) Principles and Terminology

- **Graphics reveal data** - show the data without distorting it - “above all else show the data”
- **Small multiples** - understanding one slice makes understanding others easier
- **Lie factor** - effect shown/effect in reality
- **Graphical Integrity** - no lies, let data vary, not design
- **Data density** - maximize data/ink ratio
- **Sparklines** - seems they haven’t caught on
- **Chartjunk** - self-explanatory
- **Powerpoint is responsible for most of the world’s sorrows** [The Cognitive Style of Powerpoint]

Cleveland

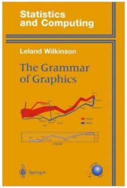


William Cleveland (1943-)

- His Elements of Graphing Data and Visualizing Data pioneered systematic considerations of legibility
- Cleveland is particularly known for promoting the dot plot as a alternative to bars and pies.
- The dot plot provides clarity and easy comparison of data.
- Cleveland also pioneered Trellis graphics that emphasizes comparison of multiple panels of data.

Source: W. Cleveland

Wilkinson

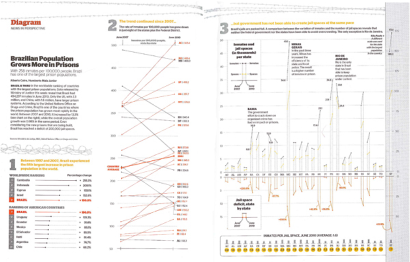


Leland Wilkinson (1945-)

- Its Grammar of Graphics was extremely influential in thinking about graphics:
  - Grammar means "rules for art and science"
  - Specifies rules both mathematical and aesthetic
  - Earlier graph producers focused on aesthetics of static content
  - Dynamic graphics and scientific visualization, by contrast, require sophisticated designs to enable brushing, drill-down, zooming, linking
  - The Grammar of Graphics is easily adapted to this approach
- `ggplot2` (Hadley Wickham) is inspired by this formalism!

Source: L. Wilkinson

Cairo



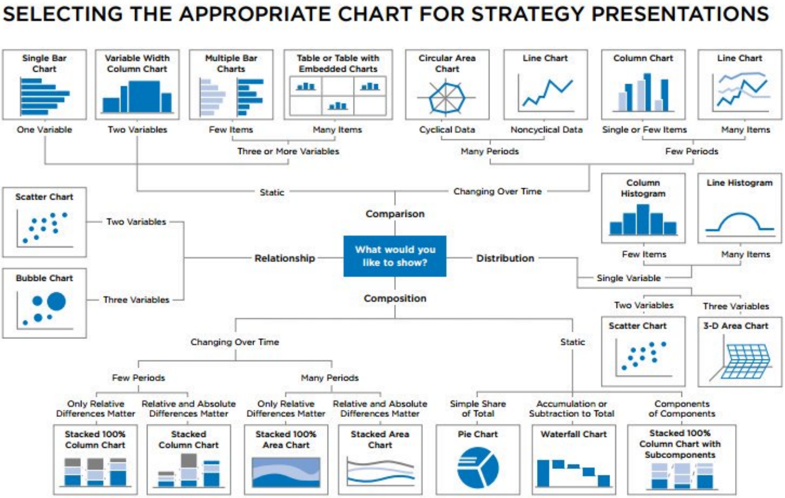
Alberto Cairo

- Data Journalism / Importance of storytelling!
- *The functional art* : An introduction to Information Graphics and Visualization, the communication of facts and data by means of charts, graphs, maps, and diagrams.
- *The truthful art* : Explains how to transform elementary principles of data and scientific reasoning into tools that you can use in daily life to interpret data sets and extract stories from them.

Source: A. Cairo

Outline

How to Pick the Right Chart?



Source: Unknown

How to Build a Graph?

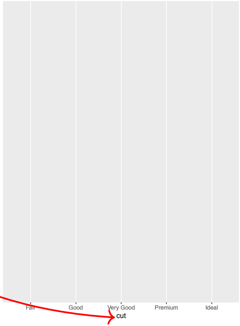
cut	n
Ideal	21551
Premium	13791
Very Good	12082
Good	4906
Fair	1610

Bar Plot Example

- Start from the **data** and a blank canvas.

How to Build a Graph?

cut	n
Ideal	21551
Premium	13791
Very Good	12082
Good	4906
Fair	1610

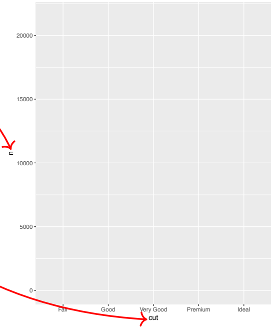


Bar Plot Example

- Start from the **data** and a blank canvas.
- Map **variables** to **axis**.

How to Build a Graph?

cut	n
Ideal	21551
Premium	13791
Very Good	12082
Good	4906
Fair	1610

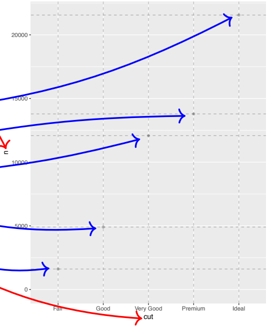


Bar Plot Example

- Start from the **data** and a blank canvas.
- Map **variables** to **axis**.

How to Build a Graph?

cut	n
Ideal	21551
Premium	13791
Very Good	12082
Good	4906
Fair	1610

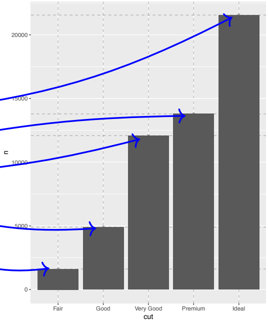


Bar Plot Example

- Start from the **data** and a blank canvas.
- Map **variables** to **axis**.
- Place the data with a **meaningful** mapping.

How to Build a Graph?

cut	n
Ideal	21551
Premium	13791
Very Good	12082
Good	4906
Fair	1610

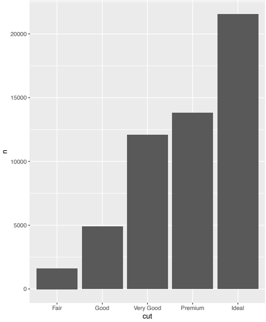


Bar Plot Example

- Start from the **data** and a blank canvas.
- Map **variables** to **axis**.
- Place the data with a **meaningful** mapping.
- Draw the graph with a certain **geometry**.

How to Build a Graph?

cut	n
Ideal	21551
Premium	13791
Very Good	12082
Good	4906
Fair	1610

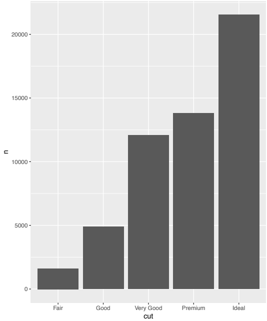


Bar Plot Example

- Start from the **data** and a blank canvas.
- Map **variables** to **axis**.
- Place the data with a **meaningful** mapping.
- Draw the graph with a certain **geometry**.

How to Build a Graph?

cut	n
Ideal	21551
Premium	13791
Very Good	12082
Good	4906
Fair	1610



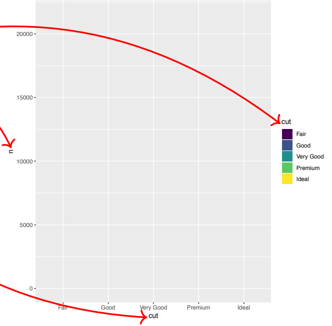
Bar Plot Example

- Start from the **data** and a blank canvas.
- Map **variables** to **axis**.
- Place the data with a **meaningful** mapping.
- Draw the graph with a certain **geometry**.

- Systematic way of describing a graph.

How to Build a Graph?

cut	n
Ideal	21551
Premium	13791
Very Good	12082
Good	4906
Fair	1610



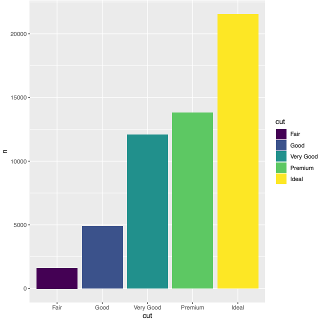
Bar Plot Example

- Start from the **data** and a blank canvas.
- Map **variables** to **axis**.
- Place the data with a **meaningful** mapping.
- Draw the graph with a certain **geometry**.

- Systematic way of describing a graph.

How to Build a Graph?

cut	n
Ideal	21551
Premium	13791
Very Good	12082
Good	4906
Fair	1610

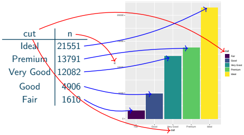


Bar Plot Example

- Start from the **data** and a blank canvas.
- Map **variables** to **axis**.
- Place the data with a **meaningful** mapping.
- Draw the graph with a certain **geometry**.

- Systematic way of describing a graph.

Grammar of Graphics

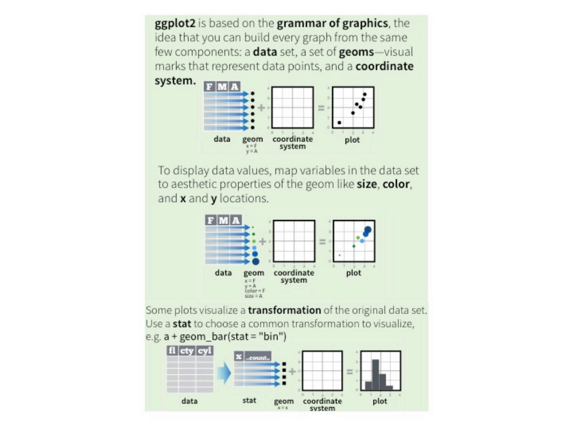


Wilkinson

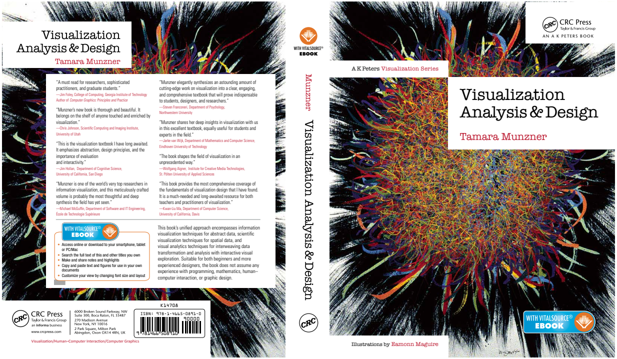
- **DATA** - weighting, reshaping, counting, bootstrapping
- **VARIABLES** - transform, sort, log, rank., resid., quant.
- **STATISTICS** - statistic preprocessing of data
- **AESTHETICS** - mapping between position/color/size/... and variable
- **SCALES** - nominal, ordinal, interval, ratio...
- **GEOMETRY** - line, area, etc., along with modifiers like jitter and dodge
- **COORDINATES** - refers to the coordinate system of the graph (cartesian, polar, etc.)
- **FACETS** - subgroups, multiway tables
- **GUIDES** - legends, axes, color scales, keys

Grammar of Graphics

ggplot2 cheatsheet



- ggplot2 (Hadley Wickham) is inspired by this formalism!



Visualization Analysis and Design

- Book from Tamara Munzner
- Published by CRC Press in 2014

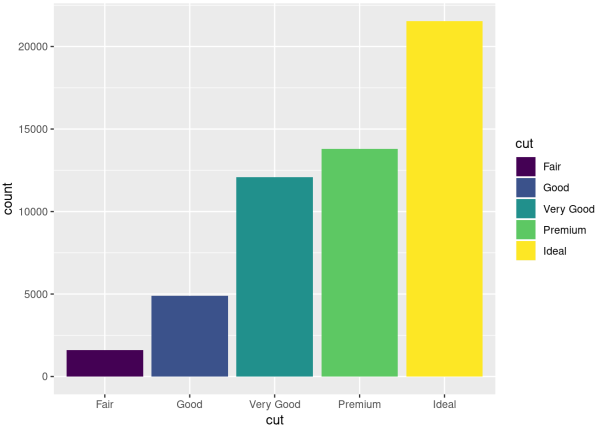
- Supplementary slides *taken* from her slidedesk!

Source: T. Munzner

Outline

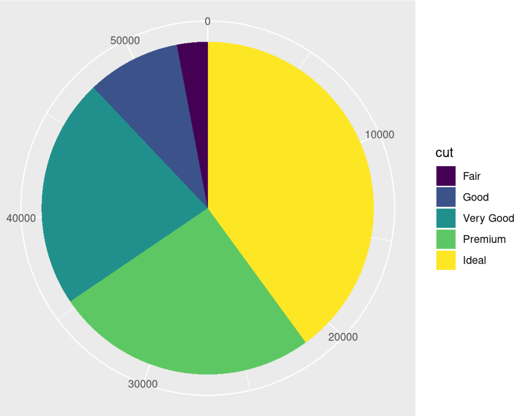
Outline

Quantities - Bar



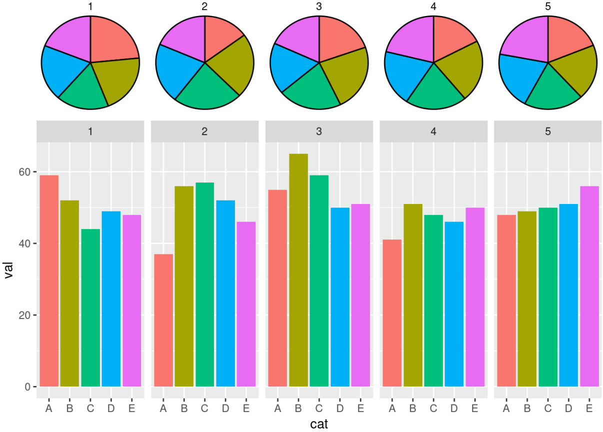
- Adapted to counts and quantities.

Proportion - Pie



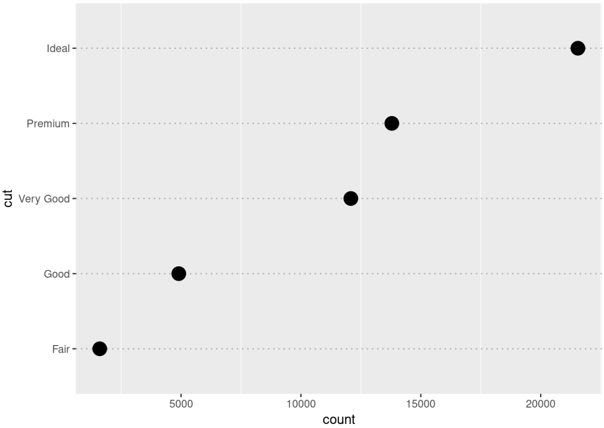
- Should not be used for comparison...

Proportion - Pie



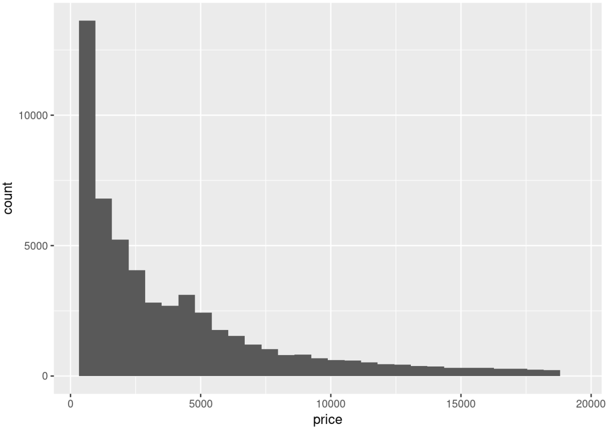
- Bar plots more efficient for comparison.
- Pie plots more efficient for global proportions!

Quantities - Cleveland Dot Plots



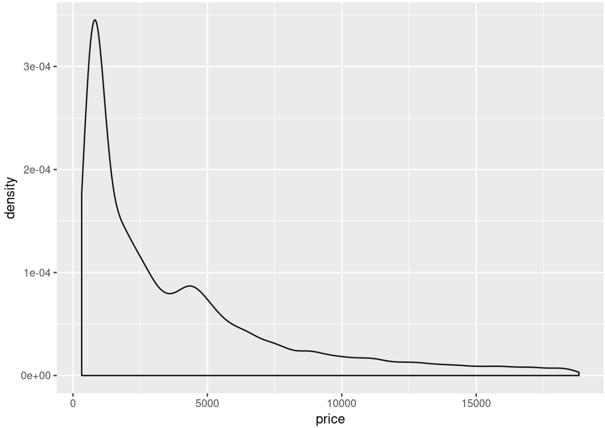
- Less *ink*, more pleasant...

Distribution - Histogram and density



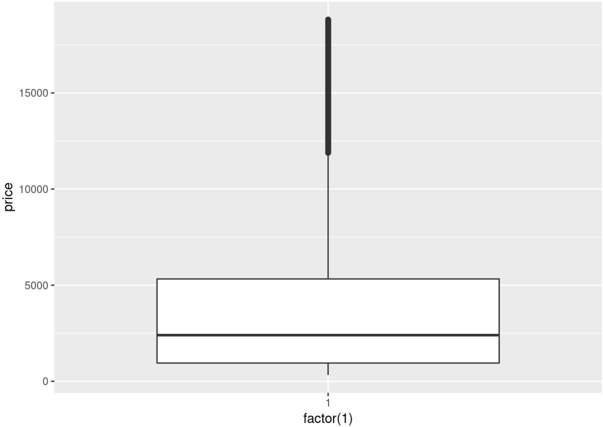
- Easily interpretable
- Adapted to continuous variable.

Distribution - Histogram and density



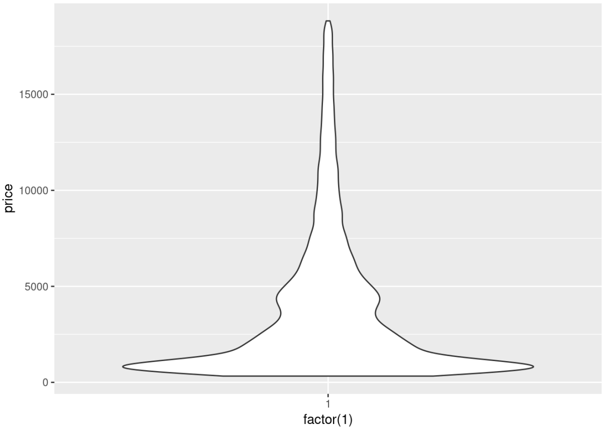
- Regularized view...

Distribution - Box and Whiskers



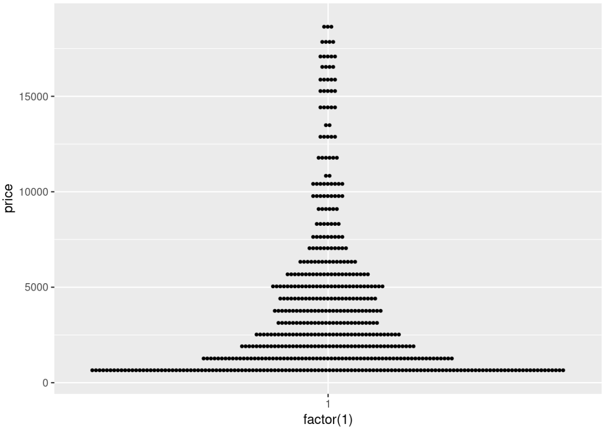
- Most classical representation after pie...

Distribution - Violin plot



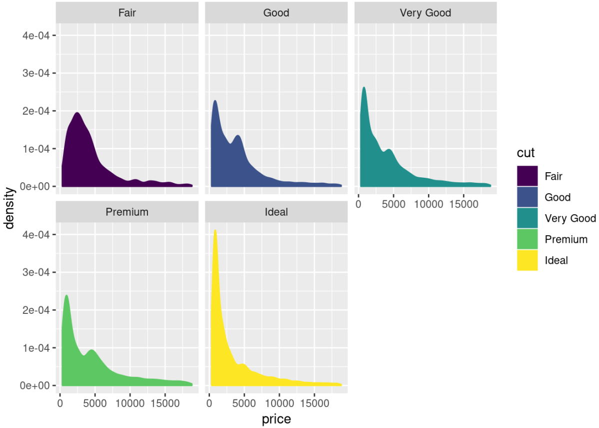
- Combined box plot and density estimation.

Distribution - Dot plots



- Combined binning and individuals...

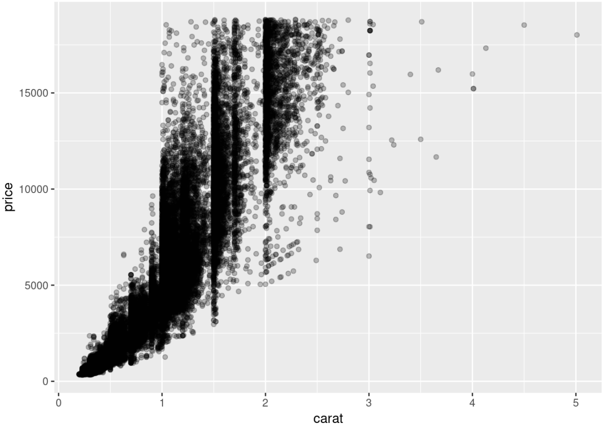
Grouping



- Key to construct complex representation.

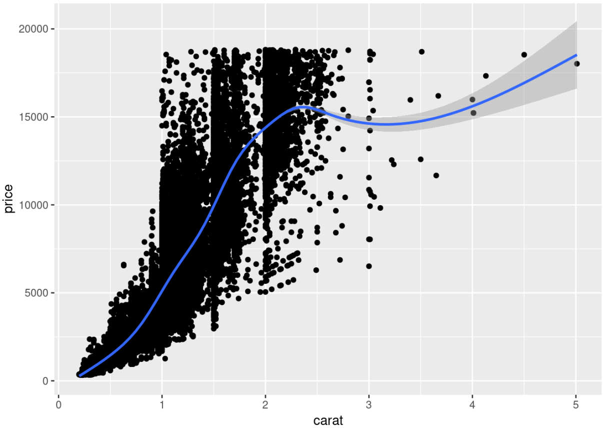
Outline

Relation - Scatter Plot



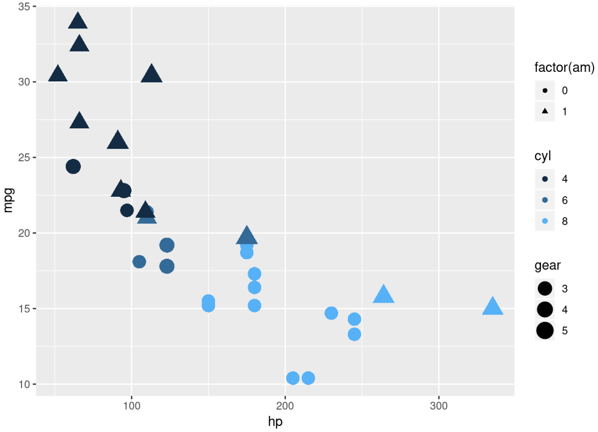
- Used to visualize the relationship between two variables.

Relation - Smoothing



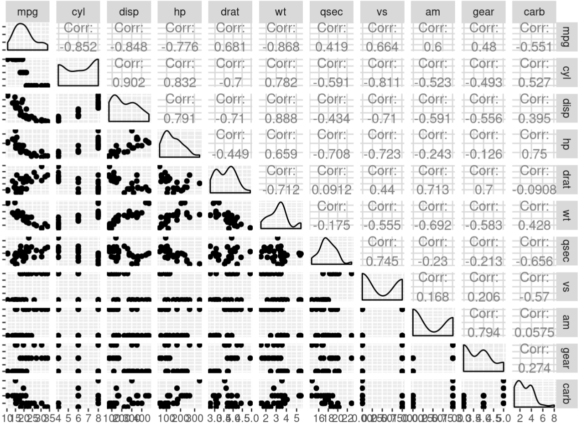
- Strong visual help.

Relation - Symbols



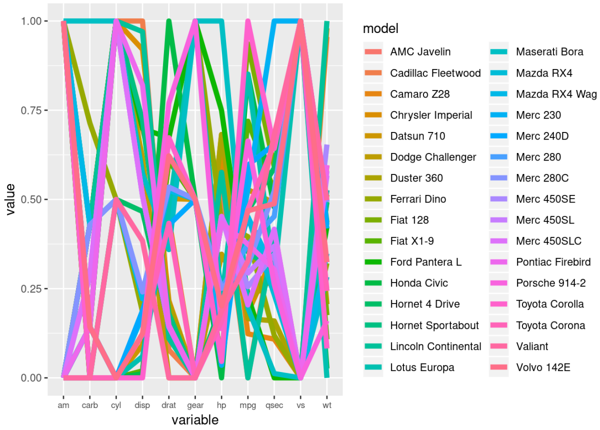
- Good idea to augment the information density...
- but can lead to too much complexity.

Grouping - Scatter Plot Matrix



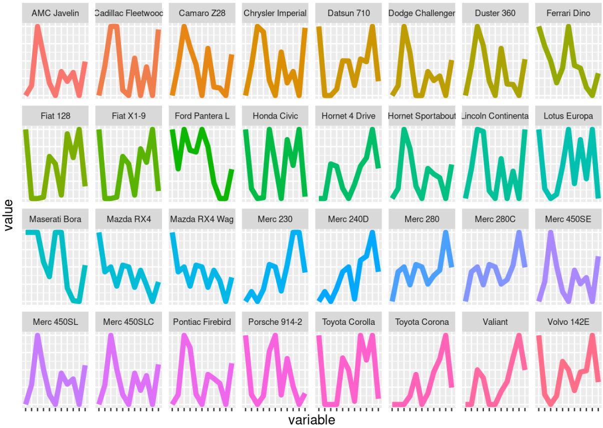
- Gather all the dependencies...

HighD - Parallel Coordinates / Radar Plot



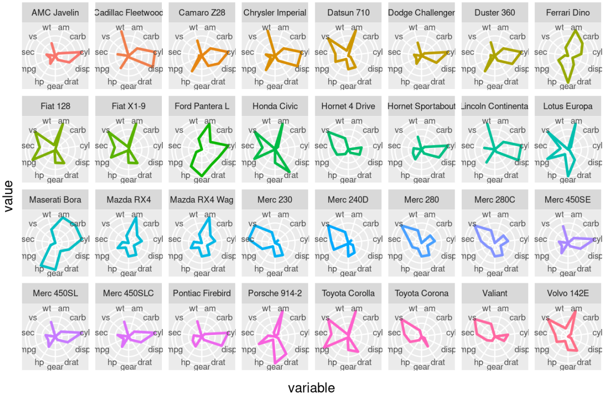
- Clever ideas to visualize groups.

HighD - Parallel Coordinates / Radar Plot



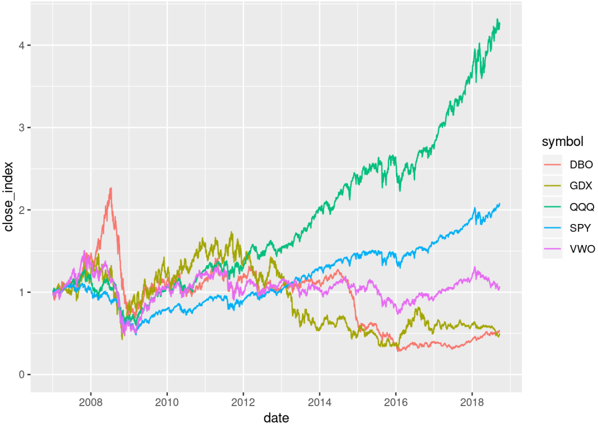
- Clever ideas to visualize groups.
- Example of small multiples

HighD - Parallel Coordinates / Radar Plot



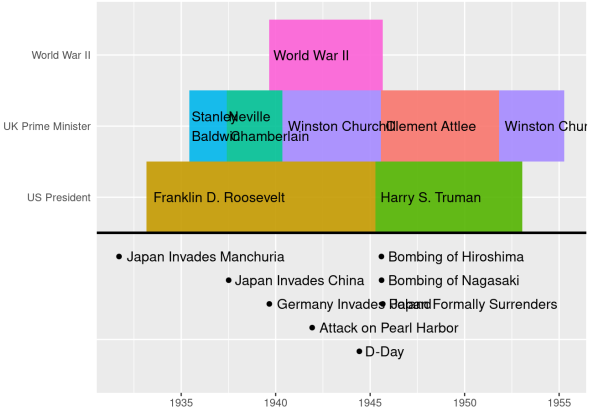
- Clever ideas to visualize groups.
- Example of small multiples

Evolution - Time series



- Order makes lines pertinent...
- Columns can also be used.

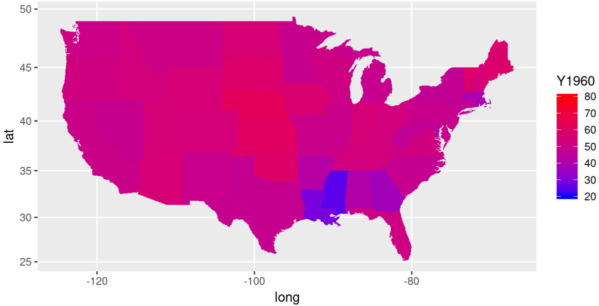
Evolution - Timeline



- Is this really a plot?

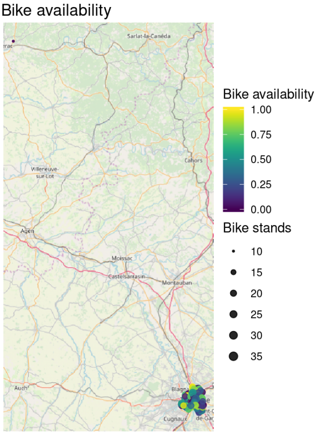
Outline

Map - Choroplets



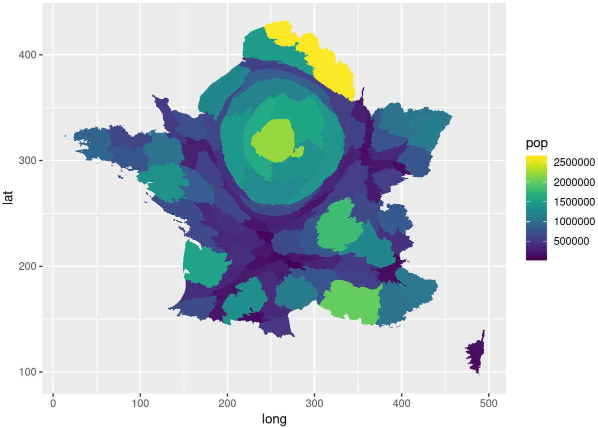
- Strong visual impact!

Map - Symbols



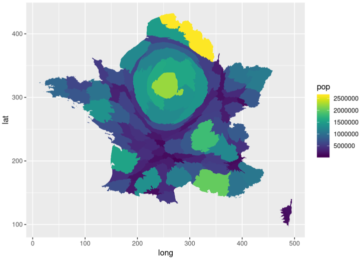
- Same ideas than decoration
- Could be extended to quite complex decorations...

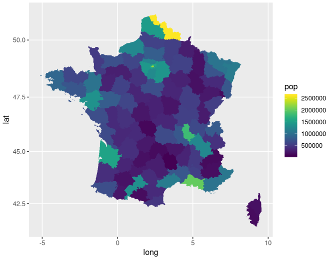
Map - Cartograms



- Mainly useful when the reference is known.

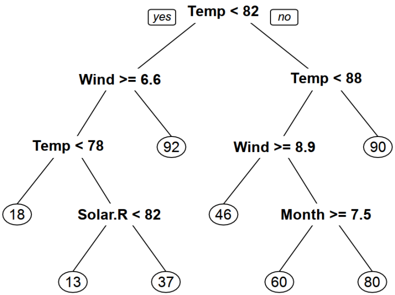
Map - Cartograms





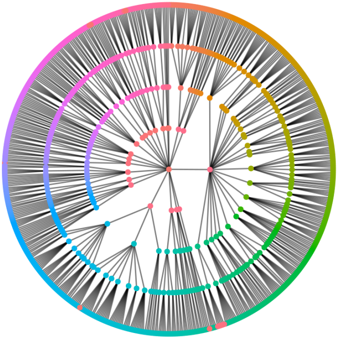
Outline

Hierarchy - Trees



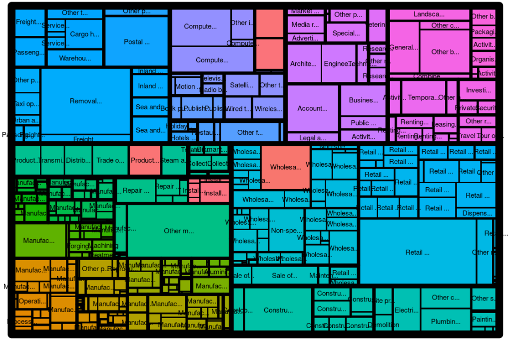
- Often use in classification...

Hierarchy - Tree Graph



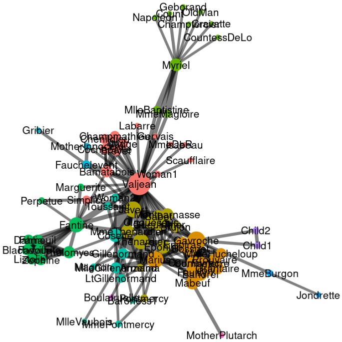
- Polar variant.

Hierarchy - Tree Map



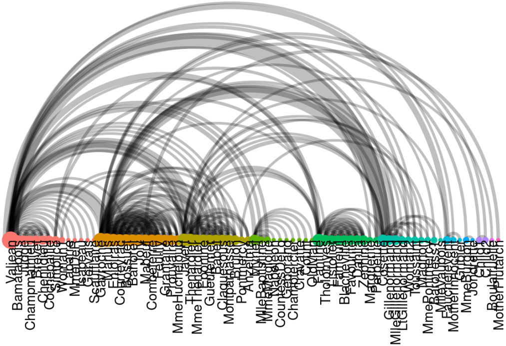
Outline

Network - Planar Layout



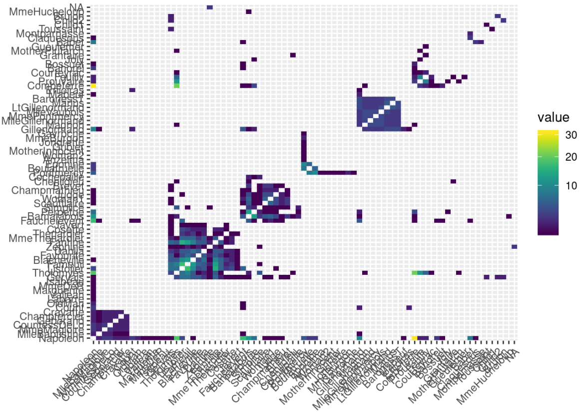
- Many possible layouts.

Network - Arc Diagram



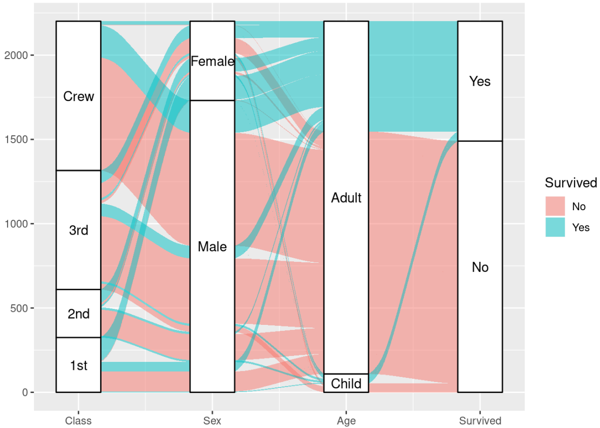
- Very different layout...

Network - Matrix View



- Adjacency matrix visualization.

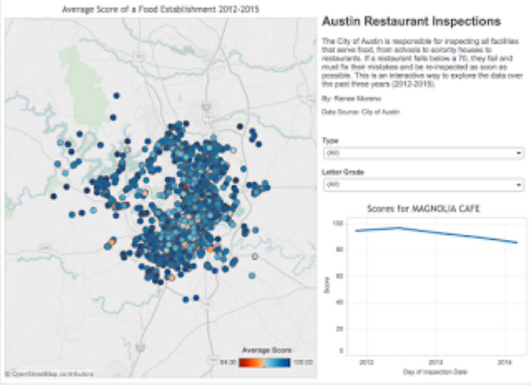
Network - Flow



- Vertice oriented visualization.

Outline

Interactivity And Dynamic Display



From static to dynamic

- Technology evolution: from paper to screen/mouse.
- Two directions:
  - Animation: use of time.
  - Interactivity: user interaction.

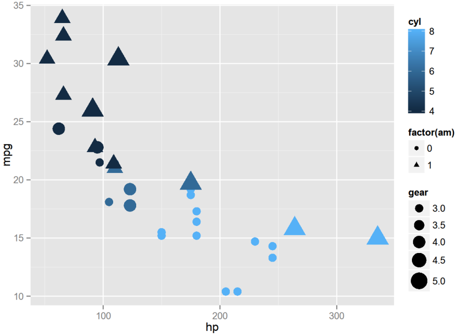
Source: RStudio

Animation

Animation

- Adapted to 1D mapping...
- Easily deployed (movie or animated picture)

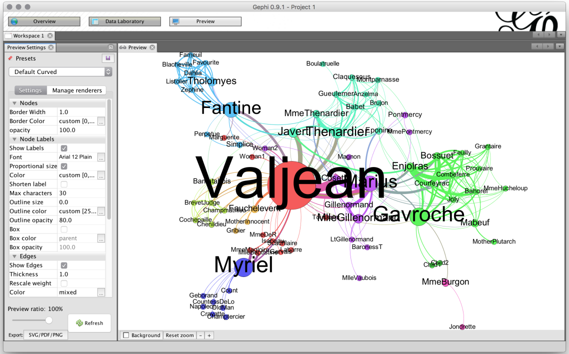
Interactivity



Interactivity: true dynamism

- More requirement.
- Solutions:
  - Standalone app (`Gephi`, `Tulip`...),
  - Javascript visualization libraries (`d3.js`, `bokeh`, `plotly`...)
  - Client/server infrastructure (`Shiny`, `flask`...)

Standalone App

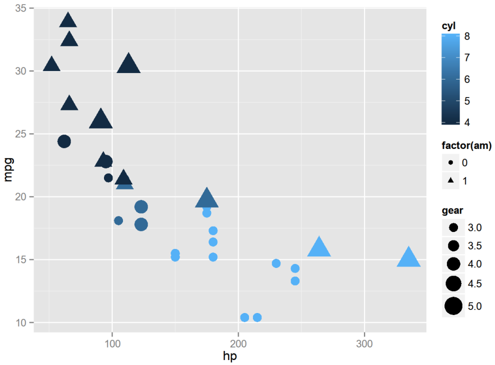


Dedicated software

- Gephi/Tulip for graphs
- Tableau
- ...

Source: Gephi

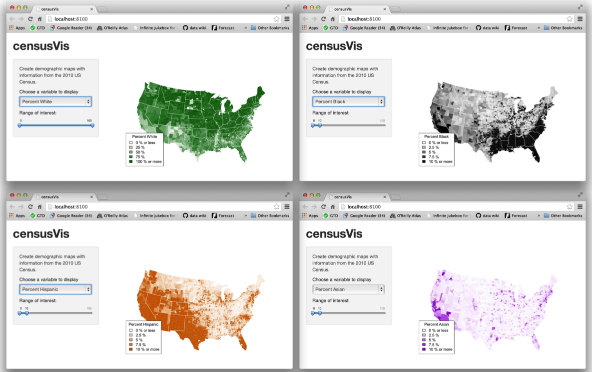
Javascript Based Interaction



Javascript based frameworks

- Local (lightweight) computation.
- Examples:
  - Tooltip, Zoom, Brushing...
  - Linked panels...

Client/Server



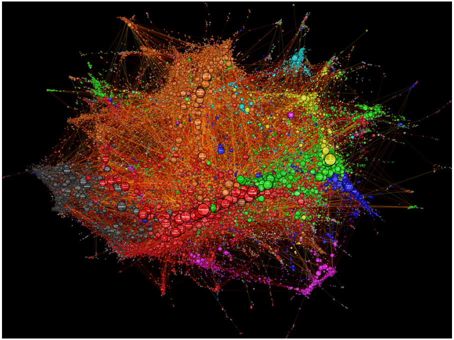
Client/Server Approach

- Shiny, Flask...
- Visualization on the client.
- Computation in the server.

Source: RStudio

Outline

Big Data Issues

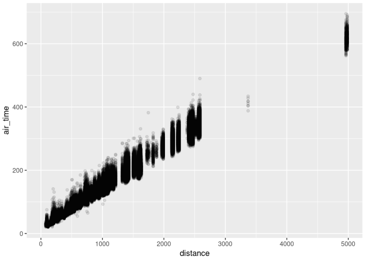


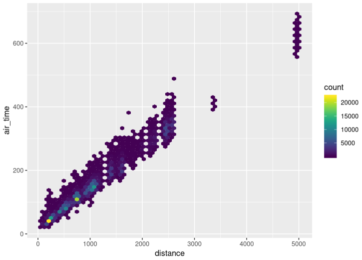
More data points than pixels!

- Even if the processing possible, it is almost impossible to visualize faithfully the data!
- Summarization/selection required:
    - Grouping by categories or binning,
    - Small multiples,
    - Interactive selection.

Source: LastFM

Binning

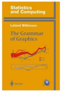




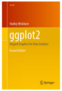
- Binning $\sim$ 2D histogram

Outline

References

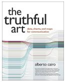


L. Wilkinson.
The Grammar of Graphics.
Springer, 2011



H. Wickham.
ggplot2: Elegant Graphics for Data Analysis.
Springer, 2016

A. Cairo.
The Functional Art: An introduction to information graphics and visualization.
New Riders, 2012



A. Cairo.
The Truthful Art: data, charts and maps for communication.
New Riders, 2016

T. Munzner.
Visualization Analysis and Design.
CRC Press, 2014

H. Wickham and G. Grolemund.
R for Data Science.
O'Reilly, 2017

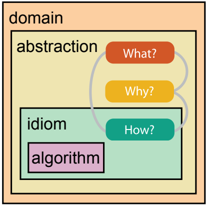
Outline

Outline

Nested model: Four levels of vis design

[A Nested Model of Visualization Design and Validation.
Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).]

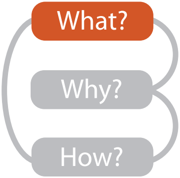
- *domain situation*
  - who are the target users?
- *abstraction*
  - translate from specifics of domain to vocabulary of vis
    - what is shown? **data** abstraction
    - why is the user looking at it? **task** abstraction
- *idiom*
  - how is it shown?
    - **visual encoding** idiom: how to draw
    - **interaction** idiom: how to manipulate
- *algorithm*
  - efficient computation



[A Multi-Level Typology of Abstract Visualization Tasks
Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013).]

Threats to validity differ at each level

[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).]



What?

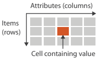
Datasets

Data Types
- Items
- Attributes
- Links
- Positions
- Grids

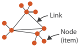
Data and Dataset Types

Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

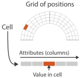
Dataset Types
- Tables



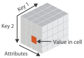
- Networks



- Fields (Continuous)



- *Multidimensional Table*



- *Trees*

- Geometry (Spatial)

Dataset Availability
- Static

- Dynamic

Attributes

Attribute Types
- Categorical

- Ordered
  - *Ordinal*

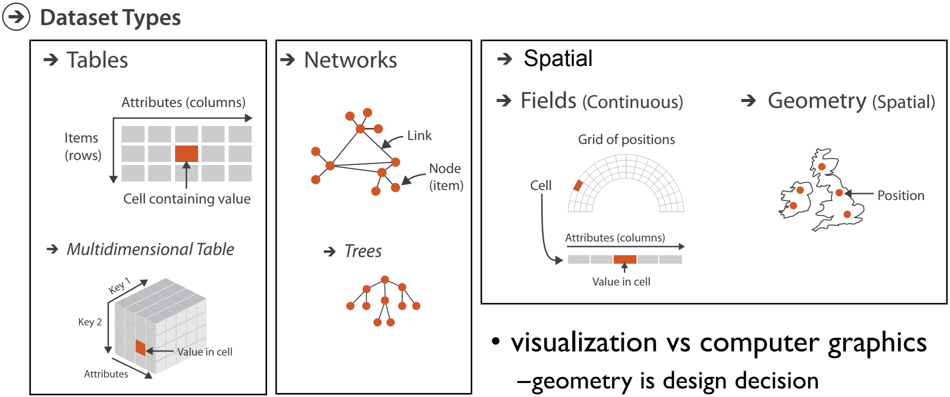
  - *Quantitative*

Ordering Direction
- Sequential

- Diverging

- Cyclic

Three major datatypes

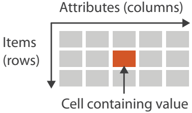


- visualization vs computer graphics
  - –geometry is design decision

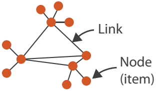
Types: Datasets and data

➔ Dataset Types

➔ Tables

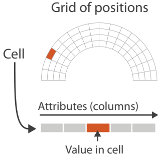


➔ Networks



➔ Spatial

➔ Fields (Continuous)



➔ Geometry (Spatial)

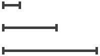
➔ Attribute Types

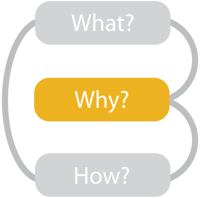
➔ Categorical

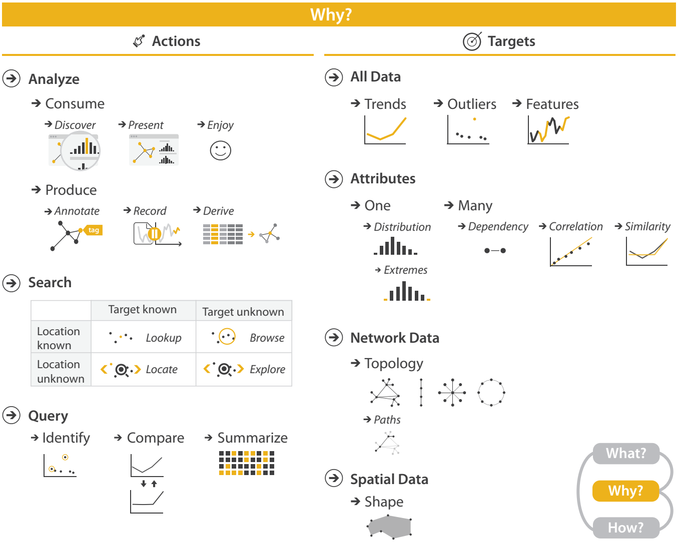
➔ Ordered

➔ *Ordinal*

➔ *Quantitative*







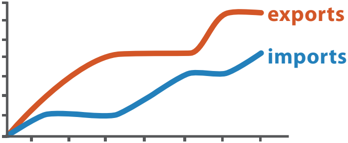
- {action, target} pairs
  - *discover distribution*
  - *compare trends*
  - *locate outliers*
  - *browse topology*

Actions: Analyze, Query

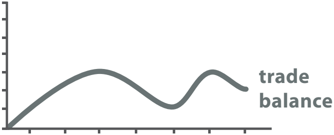
- analyze
  - consume
    - discover vs present
      - aka explore vs explain
    - enjoy
      - aka casual, social
  - produce
    - annotate, record, derive
- query
  - how much data matters?
    - one, some, all
- independent choices

Derive: Crucial Design Choice

- don't just draw what you're given!
  - decide what the right thing to show is
  - create it with a series of transformations from the original dataset
  - draw that
- one of the four major strategies for handling complexity



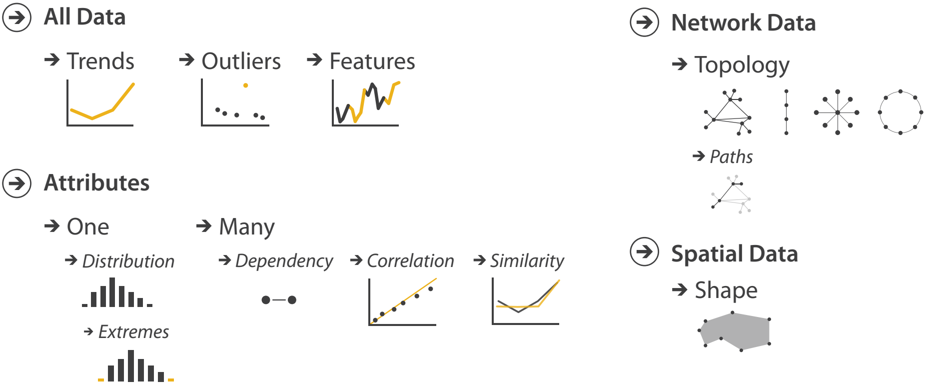
Original Data



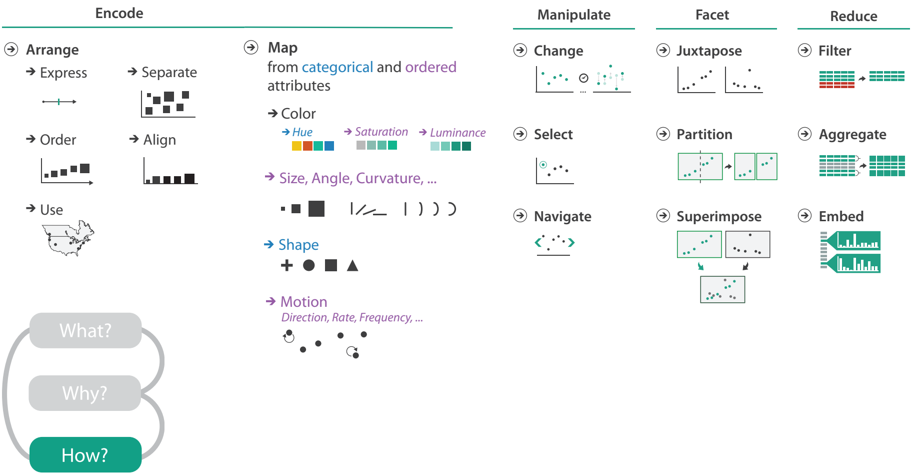
trade balance = *exports* − *imports*

Derived Data

Targets

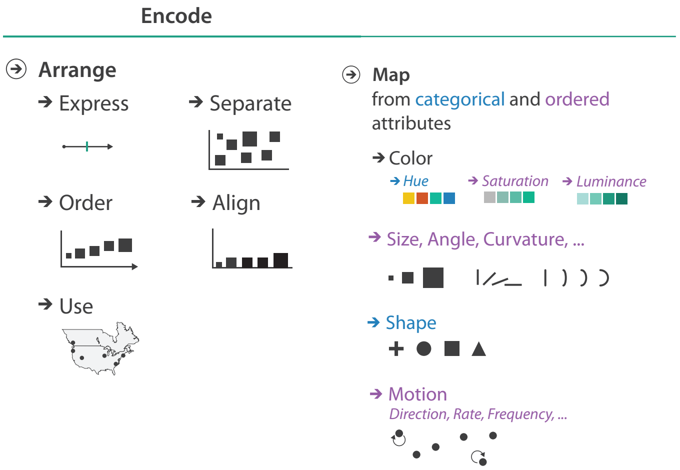

All Data
Trends
Outliers
Features
Attributes
One
Distribution
Extremes
Many
Dependency
Correlation
Similarity
Network Data
Topology
Paths
Spatial Data
Shape

How?



Encode
Arrange
Express
Separate
Order
Align
Use
Map
from categorical and ordered attributes
Color
Hue
Saturation
Luminance
Size, Angle, Curvature, ...
Shape
Motion
Direction, Rate, Frequency, ...
Manipulate
Change
Select
Navigate
Facet
Juxtapose
Partition
Superimpose
Reduce
Filter
Aggregate
Embed
What?
Why?
How?

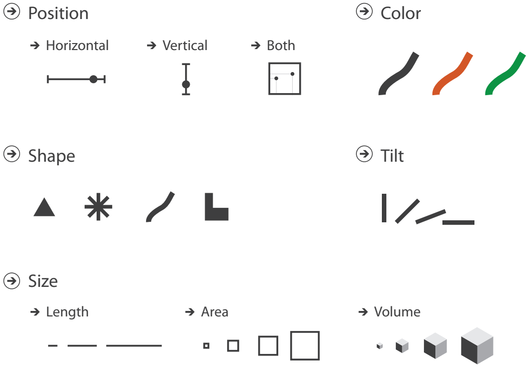
How to encode: Arrange space, map channels



Definitions: Marks and channels

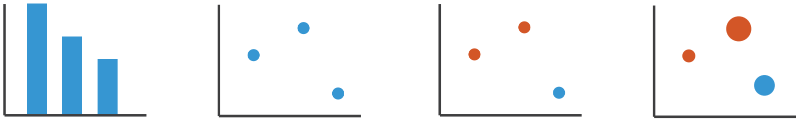
- marks
  - geometric primitives

- channels
  - control appearance of marks



Encoding visually with marks and channels

- analyze idiom structure
  - as combination of marks and channels



1:
vertical position

mark: line

2:
vertical position
horizontal position

mark: point

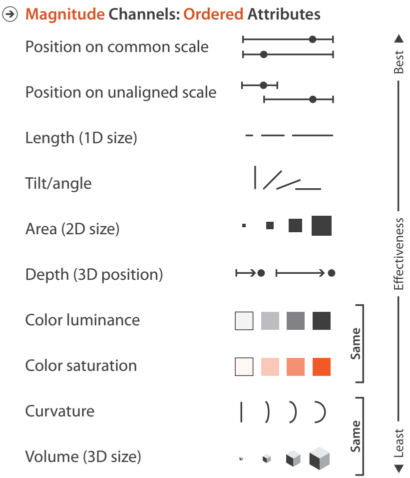
3:
vertical position
horizontal position
color hue

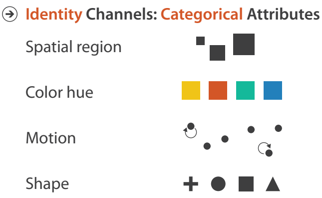
mark: point

4:
vertical position
horizontal position
color hue
size (area)

mark: point

Channels: Rankings

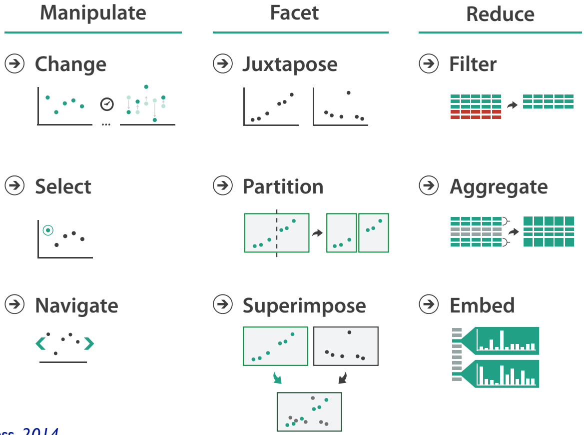




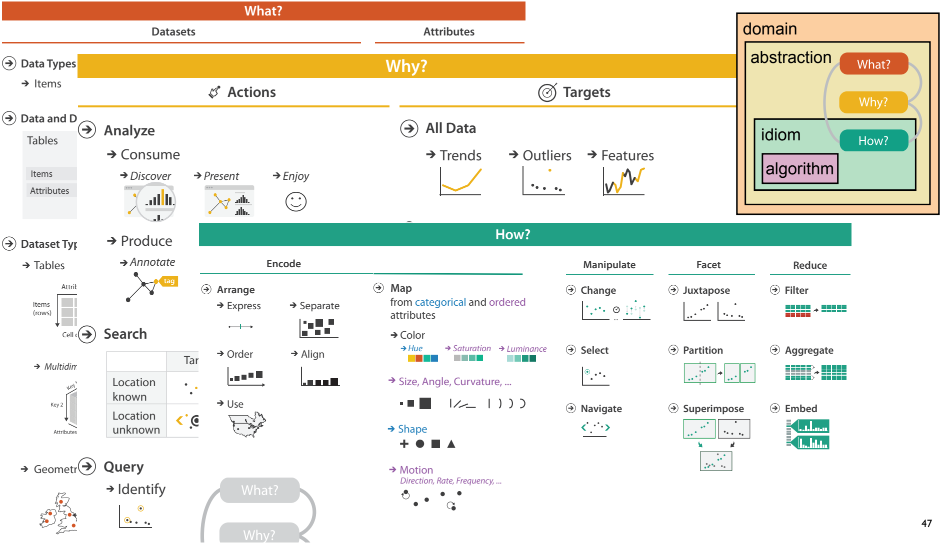
- expressiveness principle
  - match channel and data characteristics
- effectiveness principle
  - encode most important attributes with highest ranked channels

Four strategies to handle complexity

- derive new data to show within view
- change view over time
- facet across multiple views
- reduce items/attributes within single view



more at:
Visualization Analysis and Design.
Munzner. AK Peters Visualization Series, CRC Press, 2014.



What?
Datasets
Attributes
Data Types
Items
Data and D
Tables
Items
Attributes
Dataset Typ
Tables
Items (rows)
Cell
Multidim
Key 2
Attributes
Geometr
Why?
Actions
Targets
Analyze
Consume
Discover
Present
Enjoy
Produce
Annotate
tag
Search
Location known
Location unknown
Query
Identify
All Data
Trends
Outliers
Features
How?
Encode
Arrange
Express
Separate
Order
Align
Use
Map
from categorical and ordered attributes
Color
Hue
Saturation
Luminance
Size, Angle, Curvature, ...
Shape
Motion
Direction, Rate, Frequency, ...
Manipulate
Change
Select
Navigate
Facet
Juxtapose
Partition
Superimpose
Reduce
Filter
Aggregate
Embed
domain
abstraction
What?
Why?
idiom
How?
algorithm