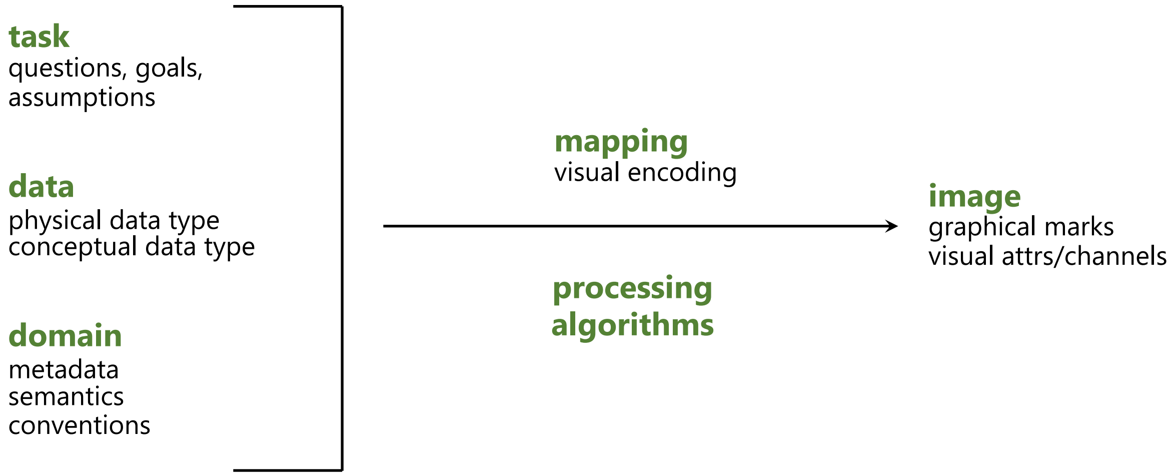
DATA & IMAGE MODELS

CS 448B | Fall 2025

MANEESH AGRAWALA

1

The big picture



task
questions, goals, assumptions

data
physical data type
conceptual data type

domain
metadata
semantics
conventions

mapping
visual encoding

processing algorithms

image
graphical marks
visual attrs/channels

7

TODAY

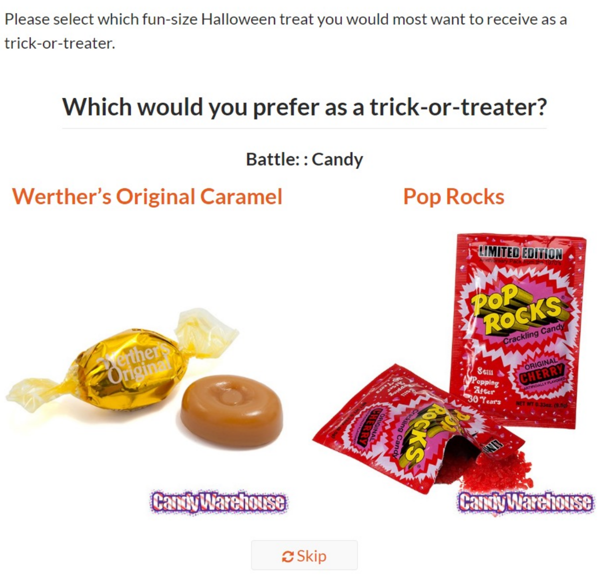
Learning Objectives

1. Identify *properties* of data and images
2. Decide how to *encode data using visual attributes/channels*
3. Define concepts of *expressiveness* and *effectiveness*
4. Develop *automated chart design* algorithm

8

DATA

9



Please select which fun-size Halloween treat you would most want to receive as a trick-or-treater.

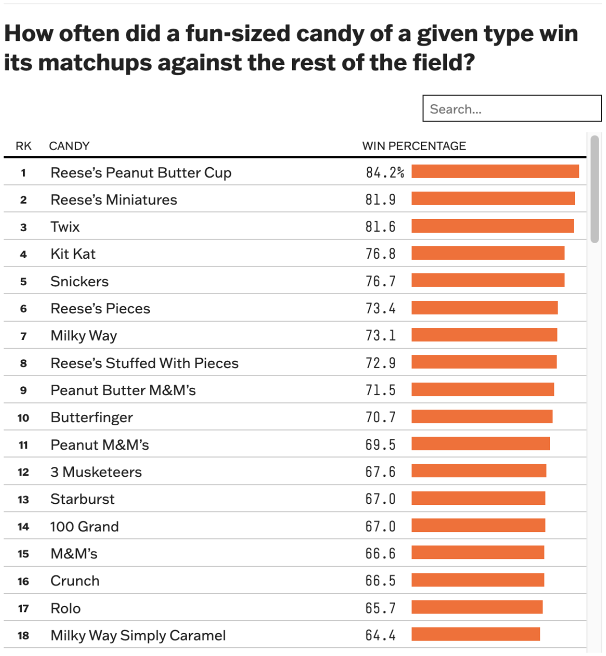
Which would you prefer as a trick-or-treater?

Battle: : Candy

Werther's Original Caramel **Pop Rocks**

http://walthickey.com/2017/10/18/whats-the-best-halloween-candy/

10



How often did a fun-sized candy of a given type win its matchups against the rest of the field?

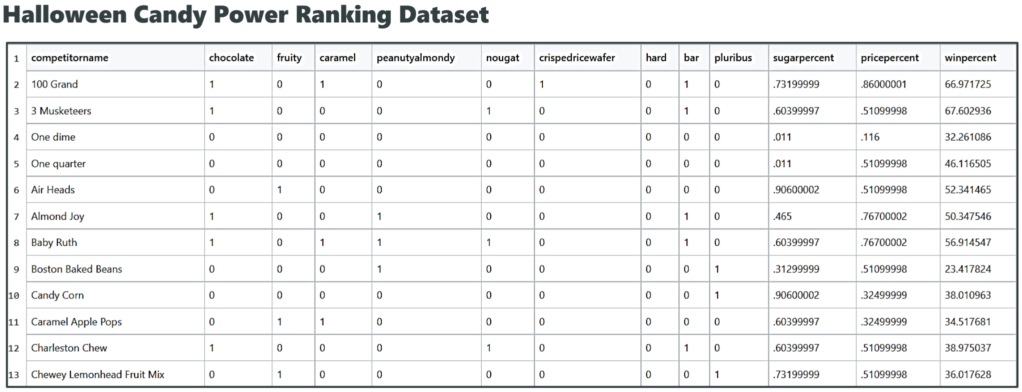
Search...

RK	CANDY	WIN PERCENTAGE
1	Reese's Peanut Butter Cup	84.2%
2	Reese's Miniatures	81.9
3	Twix	81.6
4	Kit Kat	76.8
5	Snickers	76.7
6	Reese's Pieces	73.4
7	Milky Way	73.1
8	Reese's Stuffed With Pieces	72.9
9	Peanut Butter M&M's	71.5
10	Butterfinger	70.7
11	Peanut M&M's	69.5
12	3 Musketeers	67.6
13	Starburst	67.0
14	100 Grand	67.0
15	M&M's	66.6
16	Crunch	66.5
17	Rolo	65.7
18	Milky Way Simply Caramel	64.4

http://walthickey.com/2017/10/18/whats-the-best-halloween-candy/

11

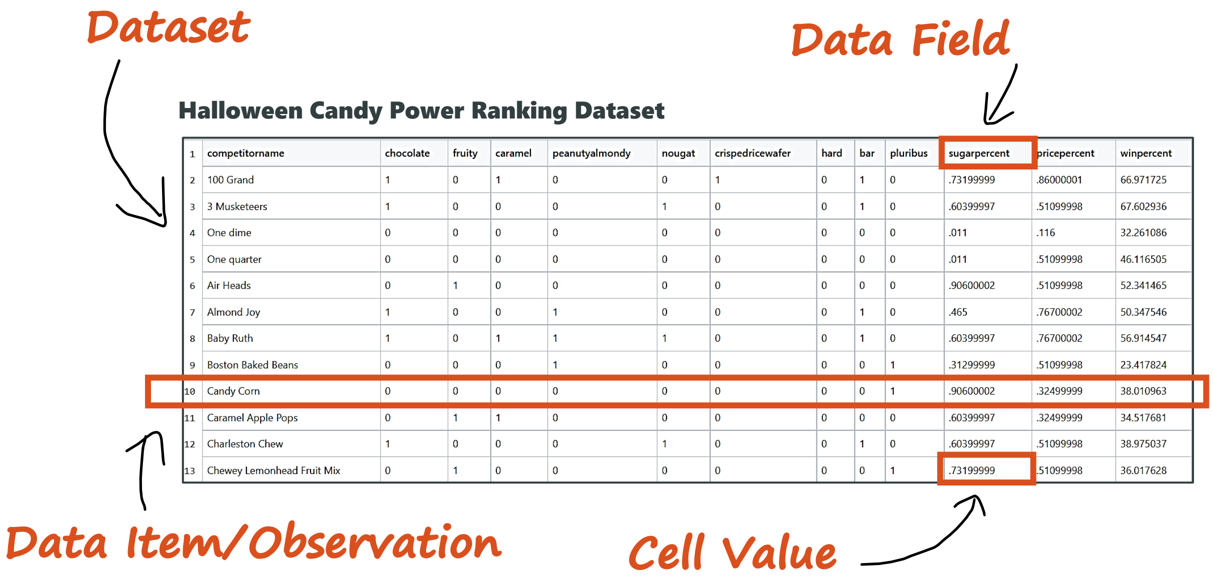
DATA TABLE



Halloween Candy Power Ranking Dataset

1	competitorname	chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus	sugarpercent	pricepercent	winpercent
2	100 Grand	1	0	1	0	0	1	0	1	0	.73199999	.86000001	66.971725
3	3 Musketeers	1	0	0	0	1	0	0	1	0	.60399997	.51099998	67.602936
4	One dime	0	0	0	0	0	0	0	0	0	.011	.116	32.261086
5	One quarter	0	0	0	0	0	0	0	0	0	.011	.51099998	46.116505
6	Air Heads	0	1	0	0	0	0	0	0	0	.90600002	.51099998	52.341465
7	Almond Joy	1	0	0	1	0	0	0	1	0	.465	.76700002	50.347546
8	Baby Ruth	1	0	1	1	1	0	0	1	0	.60399997	.76700002	56.914547
9	Boston Baked Beans	0	0	0	1	0	0	0	0	1	.31299999	.51099998	23.417824
10	Candy Corn	0	0	0	0	0	0	0	0	1	.90600002	.32499999	38.010963
11	Caramel Apple Pops	0	1	1	0	0	0	0	0	0	.60399997	.32499999	34.517681
12	Charleston Chew	1	0	0	0	1	0	0	1	0	.60399997	.51099998	38.975037
13	Chewey Lemonhead Fruit Mix	0	1	0	0	0	0	0	0	1	.73199999	.51099998	36.017628

https://fivethirtyeight.com/videos/the-ultimate-halloween-candy-power-ranking/



Dataset

Data Field

Data Item/Observation

Cell Value

Halloween Candy Power Ranking Dataset

1	competitorname	chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus	sugarpercent	pricepercent	winpercent
2	100 Grand	1	0	1	0	0	1	0	1	0	.73199999	.86000001	66.971725
3	3 Musketeers	1	0	0	0	1	0	0	1	0	.60399997	.51099998	67.602936
4	One dime	0	0	0	0	0	0	0	0	0	.011	.116	32.261086
5	One quarter	0	0	0	0	0	0	0	0	0	.011	.51099998	46.116505
6	Air Heads	0	1	0	0	0	0	0	0	0	.90600002	.51099998	52.341465
7	Almond Joy	1	0	0	1	0	0	0	1	0	.465	.76700002	50.347546
8	Baby Ruth	1	0	1	1	1	0	0	1	0	.60399997	.76700002	56.914547
9	Boston Baked Beans	0	0	0	1	0	0	0	0	1	.31299999	.51099998	23.417824
10	Candy Corn	0	0	0	0	0	0	0	0	1	.90600002	.32499999	38.010963
11	Caramel Apple Pops	0	1	1	0	0	0	0	0	0	.60399997	.32499999	34.517681
12	Charleston Chew	1	0	0	0	1	0	0	1	0	.60399997	.51099998	38.975037
13	Chewey Lemonhead Fruit Mix	0	1	0	0	0	0	0	0	1	.73199999	.51099998	36.017628

https://fivethirtyeight.com/videos/the-ultimate-halloween-candy-power-ranking/

DATA MODELS & CONCEPTUAL MODELS

Data models are formal descriptions

Math: Sets with operations on them

Examples: integers with +, - and × operators
reals/floats with +, -, × and ÷

Conceptual models are mental constructions

Include semantics and support reasoning

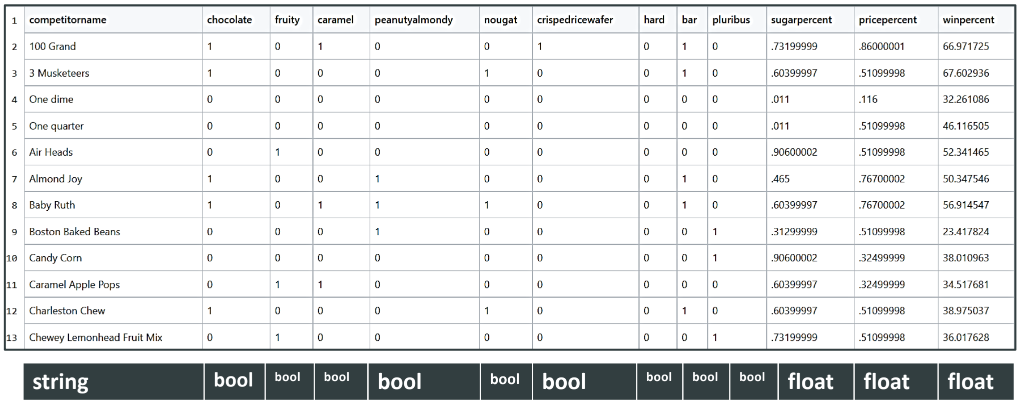
Examples (data vs. conceptual)

1D floats vs. temperature

3D tuple of floats vs. spatial location in 3D

15

DATA MODEL

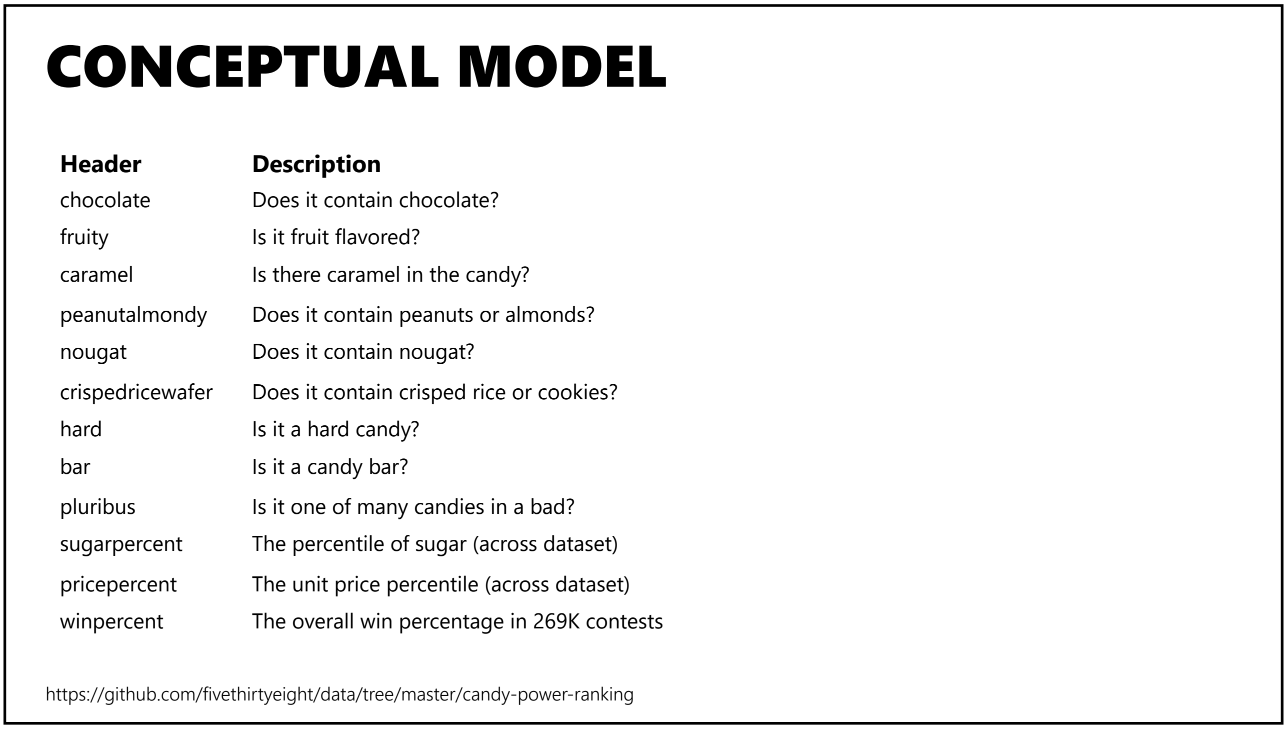


1	competitorname	chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus	sugarpercent	pricepercent	winpercent
2	100 Grand	1	0	1	0	0	1	0	1	0	.73199999	.86000001	66.971725
3	3 Musketeers	1	0	0	0	1	0	0	1	0	.60399997	.51099998	67.602936
4	One dime	0	0	0	0	0	0	0	0	0	.011	.116	32.261086
5	One quarter	0	0	0	0	0	0	0	0	0	.011	.51099998	46.116505
6	Air Heads	0	1	0	0	0	0	0	0	0	.90600002	.51099998	52.341465
7	Almond Joy	1	0	0	1	0	0	0	1	0	.465	.76700002	50.347546
8	Baby Ruth	1	0	1	1	1	0	0	1	0	.60399997	.76700002	56.914547
9	Boston Baked Beans	0	0	0	1	0	0	0	0	1	.31299999	.51099998	23.417824
10	Candy Corn	0	0	0	0	0	0	0	0	1	.90600002	.32499999	38.010963
11	Caramel Apple Pops	0	1	1	0	0	0	0	0	0	.60399997	.32499999	34.517681
12	Charleston Chew	1	0	0	0	1	0	0	1	0	.60399997	.51099998	38.975037
13	Chewey Lemonhead Fruit Mix	0	1	0	0	0	0	0	0	1	.73199999	.51099998	36.017628
	string	bool	bool	bool	bool	bool	bool	bool	bool	bool	float	float	float

How is data stored in the database?

https://github.com/fivethirtyeight/data/tree/master/candy-power-ranking

16

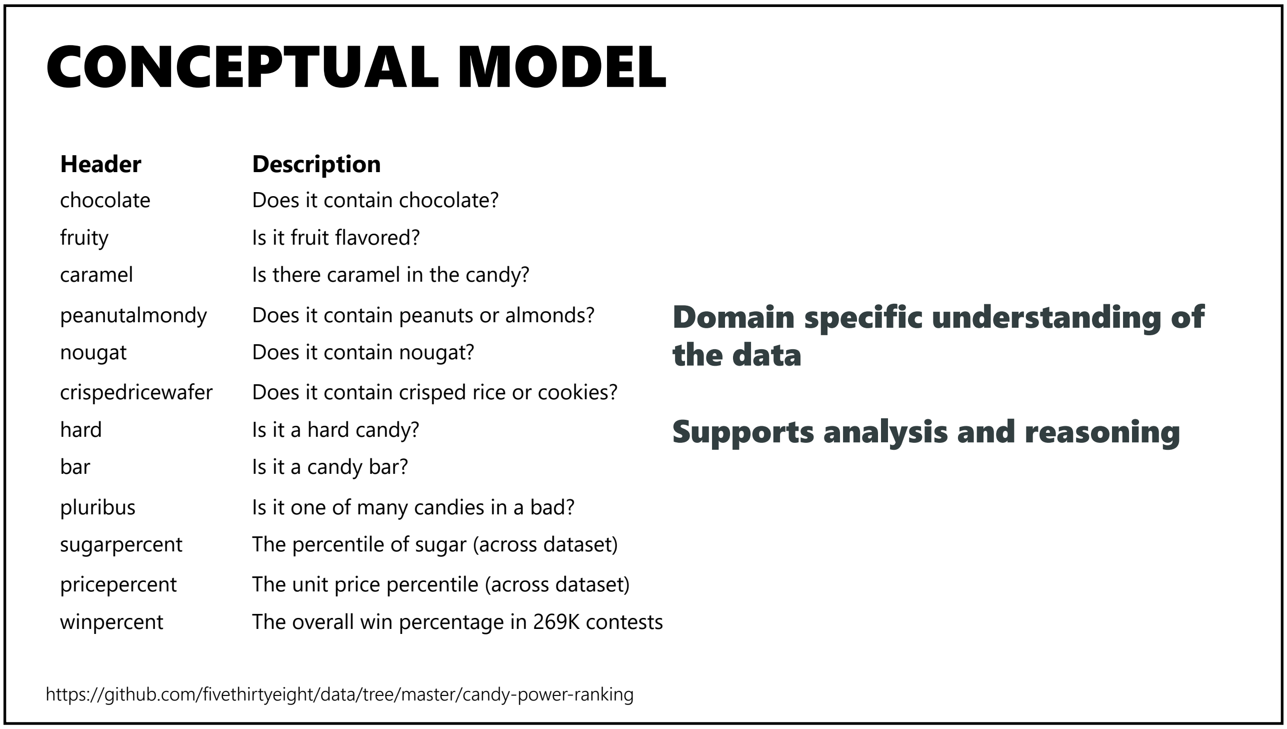


CONCEPTUAL MODEL

Header	Description
chocolate	Does it contain chocolate?
fruity	Is it fruit flavored?
caramel	Is there caramel in the candy?
peanutalmondy	Does it contain peanuts or almonds?
nougat	Does it contain nougat?
crispedricewafer	Does it contain crisped rice or cookies?
hard	Is it a hard candy?
bar	Is it a candy bar?
pluribus	Is it one of many candies in a bad?
sugarpercent	The percentile of sugar (across dataset)
pricepercent	The unit price percentile (across dataset)
winpercent	The overall win percentage in 269K contests

https://github.com/fivethirtyeight/data/tree/master/candy-power-ranking

17



CONCEPTUAL MODEL

Header	Description
chocolate	Does it contain chocolate?
fruity	Is it fruit flavored?
caramel	Is there caramel in the candy?
peanutalmondy	Does it contain peanuts or almonds?
nougat	Does it contain nougat?
crispedricewafer	Does it contain crisped rice or cookies?
hard	Is it a hard candy?
bar	Is it a candy bar?
pluribus	Is it one of many candies in a bad?
sugarpercent	The percentile of sugar (across dataset)
pricepercent	The unit price percentile (across dataset)
winpercent	The overall win percentage in 269K contests

Domain specific understanding of the data

Supports analysis and reasoning

https://github.com/fivethirtyeight/data/tree/master/candy-power-ranking

18

On the theory of
scales of measurements
S. S. Stevens, 1946

DATA TYPES

N - Nominal (labels)

Fruits: Apples, oranges, …
Operations: =, ≠

O - Ordered

Quality of eggs: Grade AA, A, B
Operations: =, ≠, <, >

Q - Interval (location of zero arbitrary)

Dates: Jan, 19, 2016; Loc.: (LAT 33.98, LON -118.45)
Like a geometric point. Cannot compare directly
Only differences (i.e. intervals) may be compared
Operations =, ≠, <, >, -

Q - Ratio (location of zero fixed)

Physical measurement: Length, Mass, …
Counts and amounts
Like a geometric vector, origin is meaningful
Operations: =, ≠, <, >, -, ÷

19

NOMINAL, ORDINAL, QUANTITATIVE

Header	Description	
competitorname	Name of candy	N
chocolate	Does it contain chocolate?	N (maybe O)
fruity	Is it fruit flavored?	N (maybe O)
caramel	Is there caramel in the candy?	N (maybe O)
peanutalmondy	Does it contain peanuts or almonds?	N (maybe O)
nougat	Does it contain nougat?	N (maybe O)
crispedricewafer	Does it contain crisped rice or cookies?	N (maybe O)
hard	Is it a hard candy?	N (maybe O)
bar	Is it a candy bar?	N (maybe O)
pluribus	Is it one of many candies in a bad?	N (maybe O)
sugarpercent	The percentile of sugar (across dataset)	Q-Ratio
pricepercent	The unit price percentile (across dataset)	Q-Ratio
winpercent	The overall win percentage in 269K contests	Q-Ratio

https://github.com/fivethirtyeight/data/tree/master/candy-power-ranking

20

DATA TYPES

DIMENSIONS

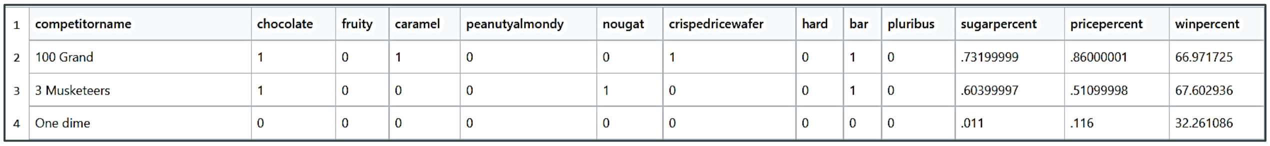
Dimensions are often the **independent** variables

Dimensions contain **qualitative values that describe the data item** (such as names, dates, or geographical data)

MEASURES

Measures are often the **dependent** variables

Measures contain numeric, **quantitative values that you can measure *in the experiment***. Measures can be aggregated (sum, count, average, std. deviation).



1	competitorname	chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus	sugarpercent	pricepercent	winpercent
2	100 Grand	1	0	1	0	0	1	0	1	0	.73199999	.86000001	66.971725
3	3 Musketeers	1	0	0	0	1	0	0	1	0	.60399997	.51099998	67.602936
4	One dime	0	0	0	0	0	0	0	0	0	.011	.116	32.261086

NOTE: Distinction is not strict. The same variable may be treated either way depending on the task

21

DIMENSION OR MEASURE

Header	**Description**
competitorname	Name of candy
chocolate	Does it contain chocolate?
fruity	Is it fruit flavored?
caramel	Is there caramel in the candy?
peanutalmondy	Does it contain peanuts or almonds?
nougat	Does it contain nougat?
crispedricewafer	Does it contain crisped rice or cookies?
hard	Is it a hard candy?
bar	Is it a candy bar?
pluribus	Is it one of many candies in a bad?
sugarpercent	The percentile of sugar (across dataset)
pricepercent	The unit price percentile (across dataset)
winpercent	The overall win percentage in 269K contests

https://github.com/fivethirtyeight/data/tree/master/candy-power-ranking

22

DIMENSION OR MEASURE

Header	Description
competitorname	Name of candy
chocolate	Does it contain chocolate?
fruity	Is it fruit flavored?
caramel	Is there caramel in the candy?
peanutalmondy	Does it contain peanuts or almonds?
nougat	Does it contain nougat?
crispedricewafer	Does it contain crisped rice or cookies?
hard	Is it a hard candy?
bar	Is it a candy bar?
pluribus	Is it one of many candies in a bad?
sugarpercent	The percentile of sugar (across dataset)
pricepercent	The unit price percentile (across dataset)
winpercent	The overall win percentage in 269K contests

https://github.com/fivethirtyeight/data/tree/master/candy-power-ranking

DIMENSION OR MEASURE

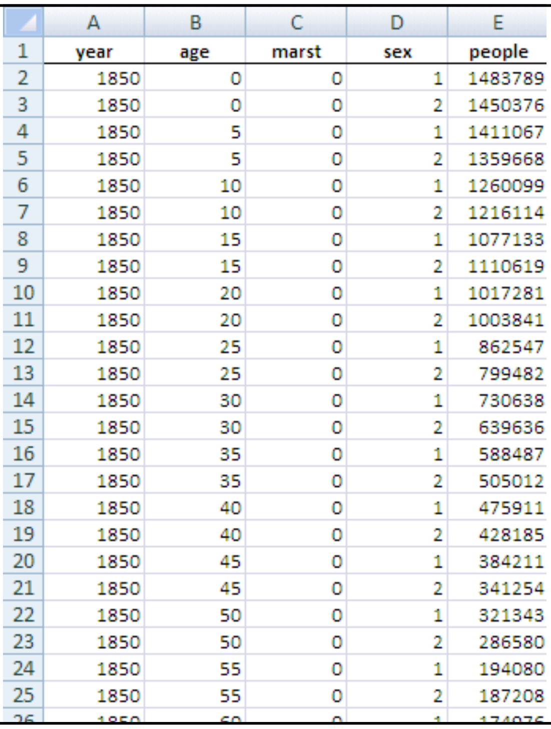
Header	Description
competitorname	Name of candy
chocolate	Does it contain chocolate?
fruity	Is it fruit flavored?
caramel	Is there caramel in the candy?
peanutalmondy	Does it contain peanuts or almonds?
nougat	Does it contain nougat?
crispedricewafer	Does it contain crisped rice or cookies?
hard	Is it a hard candy?
bar	Is it a candy bar?
pluribus	Is it one of many candies in a bad?
sugarpercent	The percentile of sugar (across dataset)
pricepercent	The unit price percentile (across dataset)
winpercent	The overall win percentage in 269K contests

https://github.com/fivethirtyeight/data/tree/master/candy-power-ranking

U.S. CENSUS DATA

People Count: # of people in subgroup
Year: 1850 – 2000 (every decade)
Age: 0 – 90+
Sex: Male, Female
Marital Status: Single, Married, Divorced, …

2348 data points

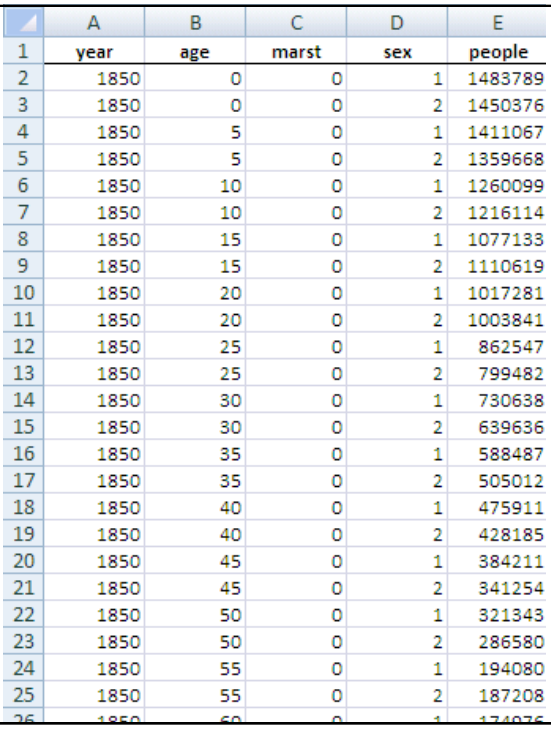


	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080
25	1850	55	0	2	187208
26	1850	60	0	1	174076

25

CENSUS N, O, Q

People Count: **Q-Ratio**
Year: **Q-Interval**
Age: **Q-Ratio**
Sex: **N**
Marital Status: **N**

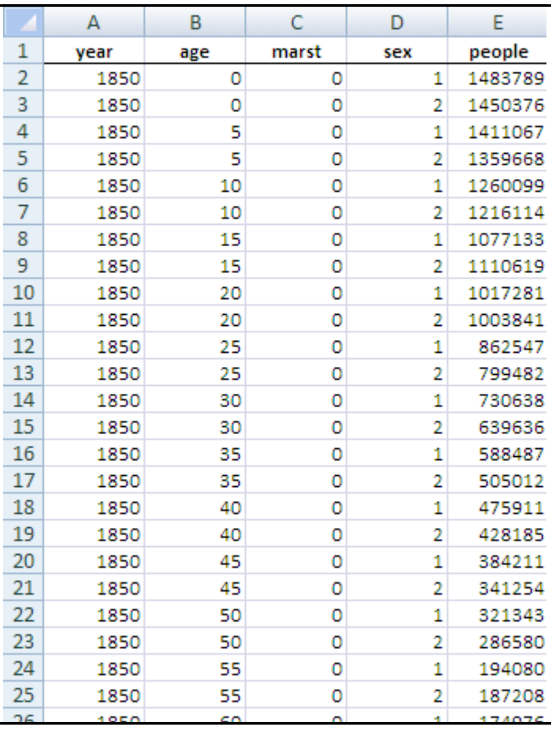


	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080
25	1850	55	0	2	187208
26	1850	60	0	1	174076

26

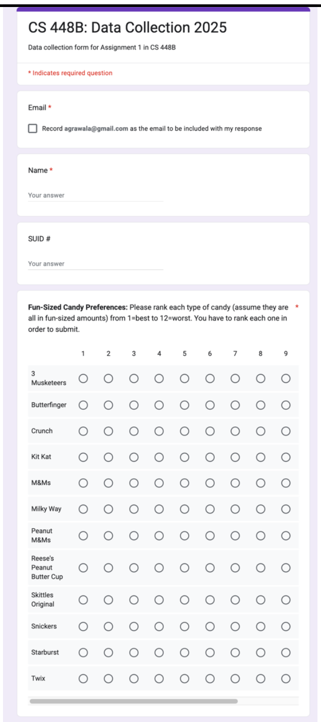
CENSUS DIM., MEAS.

People Count:	Measure
Year:	Dimension
Age:	Measure
Sex:	Measure
Marital Status:	Measure



	A	B	C	D	E
1	year	age	marst	sex	people
2	1850	0	0	1	1483789
3	1850	0	0	2	1450376
4	1850	5	0	1	1411067
5	1850	5	0	2	1359668
6	1850	10	0	1	1260099
7	1850	10	0	2	1216114
8	1850	15	0	1	1077133
9	1850	15	0	2	1110619
10	1850	20	0	1	1017281
11	1850	20	0	2	1003841
12	1850	25	0	1	862547
13	1850	25	0	2	799482
14	1850	30	0	1	730638
15	1850	30	0	2	639636
16	1850	35	0	1	588487
17	1850	35	0	2	505012
18	1850	40	0	1	475911
19	1850	40	0	2	428185
20	1850	45	0	1	384211
21	1850	45	0	2	341254
22	1850	50	0	1	321343
23	1850	50	0	2	286580
24	1850	55	0	1	194080
25	1850	55	0	2	187208
26	1850	60	0	1	174976

27



CS 448B: Data Collection 2025

Data collection form for Assignment 1 in CS 448B

* Indicates required question

Email *

□ Record agrawala@gmail.com as the email to be included with my response

Name *

Your answer

SUID #

Your answer

Fun-Sized Candy Preferences: Please rank each type of candy (assume they are all in fun-sized amounts) from 1=best to 12=worst. You have to rank each one in order to submit. *

	1	2	3	4	5	6	7	8	9
3 Musketeers	○	○	○	○	○	○	○	○	○
Butterfinger	○	○	○	○	○	○	○	○	○
Crunch	○	○	○	○	○	○	○	○	○
Kit Kat	○	○	○	○	○	○	○	○	○
M&Ms	○	○	○	○	○	○	○	○	○
Milky Way	○	○	○	○	○	○	○	○	○
Peanut M&Ms	○	○	○	○	○	○	○	○	○
Reese's Peanut Butter Cup	○	○	○	○	○	○	○	○	○
Skittles Original	○	○	○	○	○	○	○	○	○
Snickers	○	○	○	○	○	○	○	○	○
Starburst	○	○	○	○	○	○	○	○	○
Twix	○	○	○	○	○	○	○	○	○

Data: Halloween Candy Class Ranking

As part of the first lecture in this class you have been asked to fill out a form asking you to rank 12 types of candy commonly passed out on Halloween (fun-sized portions) from 1=best to 12=worst. We have aggregated and wrangled the data to produce a data table with the following information.

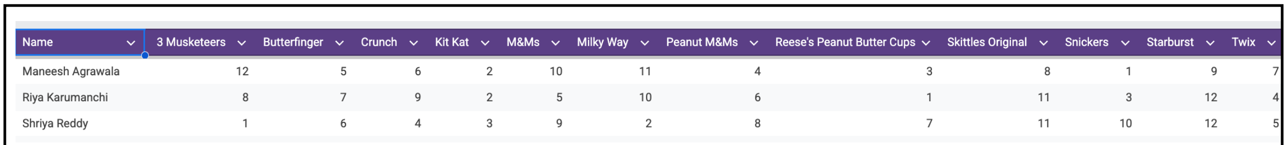
Number of records: 12
Data fields:

- **candy:** Name of candy.
- **chocolate:** Does it contain chocolate?
- **fruity:** Is it fruit flavored?
- **caramel:** Is there caramel in the candy?
- **peanutyalmondy:** Does it contain peanuts, peanut butter or almonds?
- **nougat:** Does it contain nougat?
- **crispedricewafer:** Does it contain crisped rice, wafers or a cookie component?
- **hard:** Is it hard candy?
- **bar:** Is it a candy bar?
- **pluribus:** Is it one of many candies in a bag or box?
- **sugarpercent:** The percentile of sugar it fall under within a larger data set of 86 candies.
- **pricepercent:** The unit price percentile compared with a larger data set of 86 candies.
- **classwinpercent:** The win percentage based on all the pairwise ranking match-ups in our class.

The data is available in csv format at (we've processed and wrangled it for you to convert ranks into classwinpercent for each candy) : classHalloweenCandy2025.csv.

Note that all the fields other than **classwinpercent** are from a subset of Walt Hickey's Halloween Candy Article as available from Kaggle.

28



Name	3 Musketeers	Butterfinger	Crunch	Kit Kat	M&Ms	Milky Way	Peanut M&Ms	Reese's Peanut Butter Cups	Skittles Original	Snickers	Starburst	Twix
Maneesh Agrawala	12	5	6	2	10	11	4	3	8	1	9	7
Riya Karumanchi	8	7	9	2	5	10	6	1	11	3	12	4
Shriya Reddy	1	6	4	3	9	2	8	7	11	10	12	5

Computing classwinpercent

a. numrows: 50

b. numcandies: 12

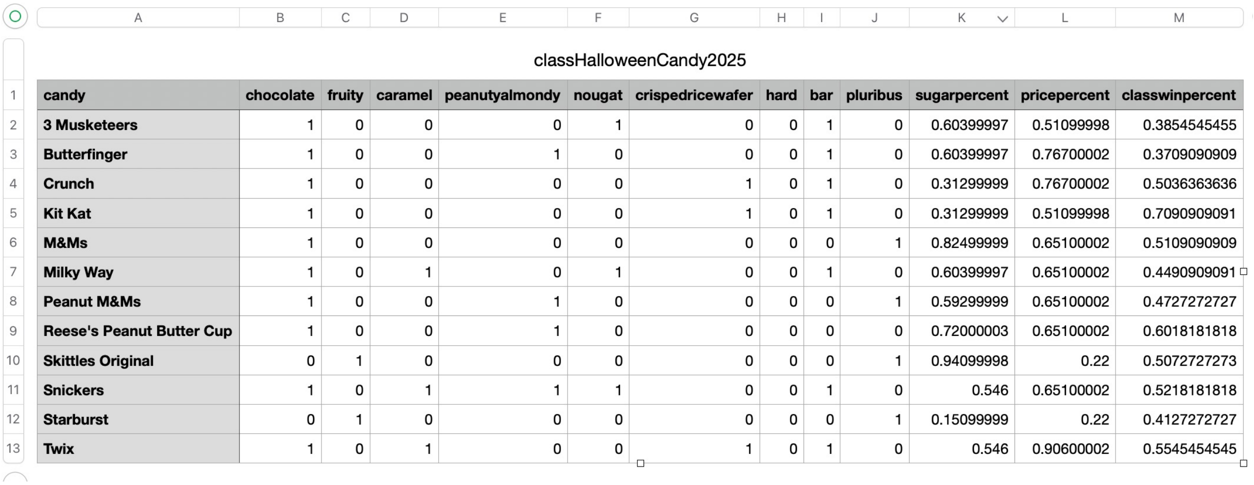
c. num pairwise wins per candy, per row: numcandies – candyrank

d. num pairwise wins per candy: sum over rows (numcandies-candyrank)

e. total num pairwise test: numrows (numcandies – 1) = 50*11

f. classwinpercent for each candy: **d/e**

29



classHalloweenCandy2025

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	candy	chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus	sugarpercent	pricepercent	classwinpercent
2	3 Musketeers	1	0	0	0	1	0	0	1	0	0.60399997	0.51099998	0.3854545455
3	Butterfinger	1	0	0	1	0	0	0	1	0	0.60399997	0.76700002	0.3709090909
4	Crunch	1	0	0	0	0	1	0	1	0	0.31299999	0.76700002	0.5036363636
5	Kit Kat	1	0	0	0	0	1	0	1	0	0.31299999	0.51099998	0.7090909091
6	M&Ms	1	0	0	0	0	0	0	0	1	0.82499999	0.65100002	0.5109090909
7	Milky Way	1	0	1	0	1	0	0	1	0	0.60399997	0.65100002	0.4490909091
8	Peanut M&Ms	1	0	0	1	0	0	0	0	1	0.59299999	0.65100002	0.4727272727
9	Reese's Peanut Butter Cup	1	0	0	1	0	0	0	0	0	0.72000003	0.65100002	0.6018181818
10	Skittles Original	0	1	0	0	0	0	0	0	1	0.94099998	0.22	0.5072727273
11	Snickers	1	0	1	1	1	0	0	1	0	0.546	0.65100002	0.5218181818
12	Starburst	0	1	0	0	0	0	0	0	1	0.15099999	0.22	0.4127272727
13	Twix	1	0	1	0	0	1	0	1	0	0.546	0.90600002	0.5545454545

30

DATA TABLES & TRANSFORMATIONS

31

RELATIONAL ALGEBRA [Codd 1970] / SQL

Operations on data tables: table(s) in, table out

Projection (SELECT) – choose a set of columns

Selection (WHERE) – filter rows

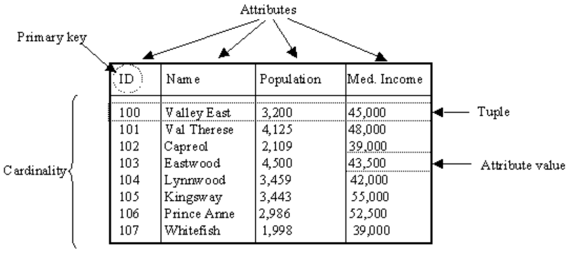
Sorting (ORDER BY) – order rows

Aggregation (GROUP BY, SUM, MIN, …)

partition rows into groups and summarize

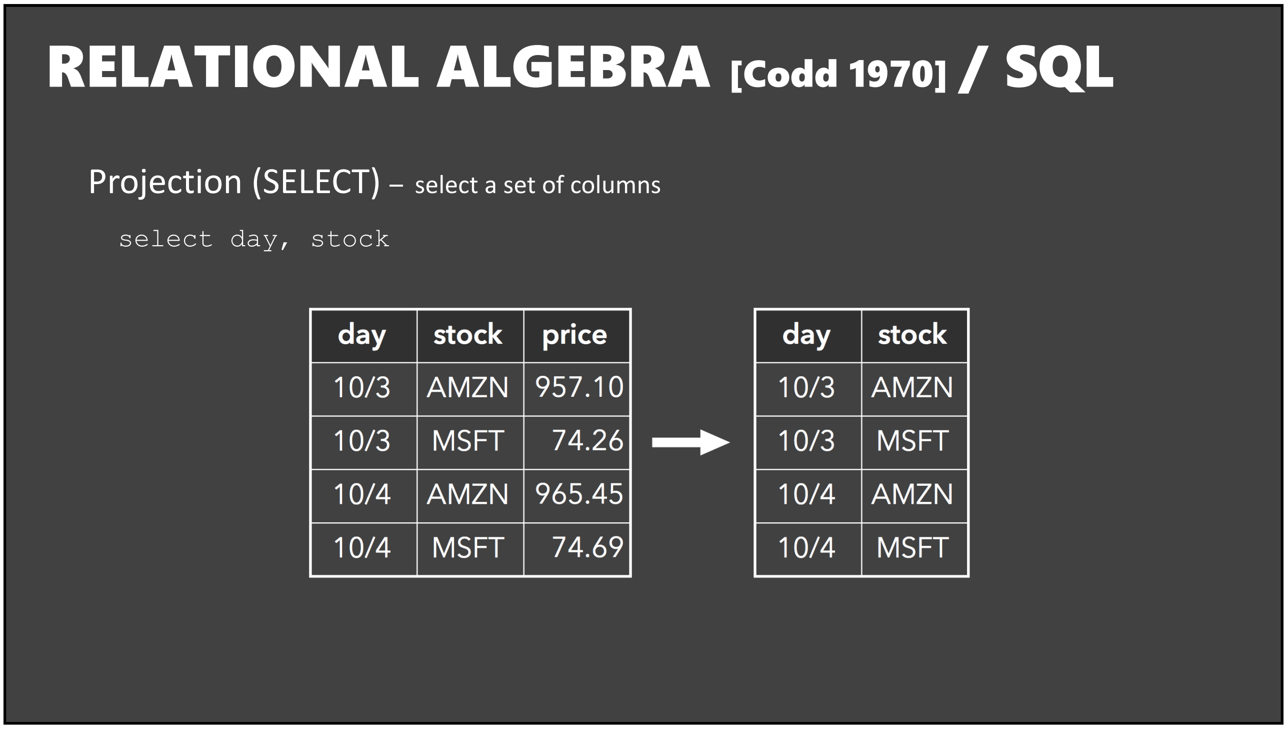
Combination (JOIN, UNION, …)

integrate data from multiple tables



ID	Name	Population	Med. Income
100	Valley East	3,200	45,000
101	Val Therese	4,125	48,000
102	Capreol	2,109	39,000
103	Eastwood	4,500	43,500
104	Lynnwood	3,459	42,000
105	Kingsway	3,443	55,000
106	Prince Anne	2,986	52,500
107	Whitefish	1,998	39,000

32



RELATIONAL ALGEBRA [Codd 1970] / SQL

Projection (SELECT) – select a set of columns

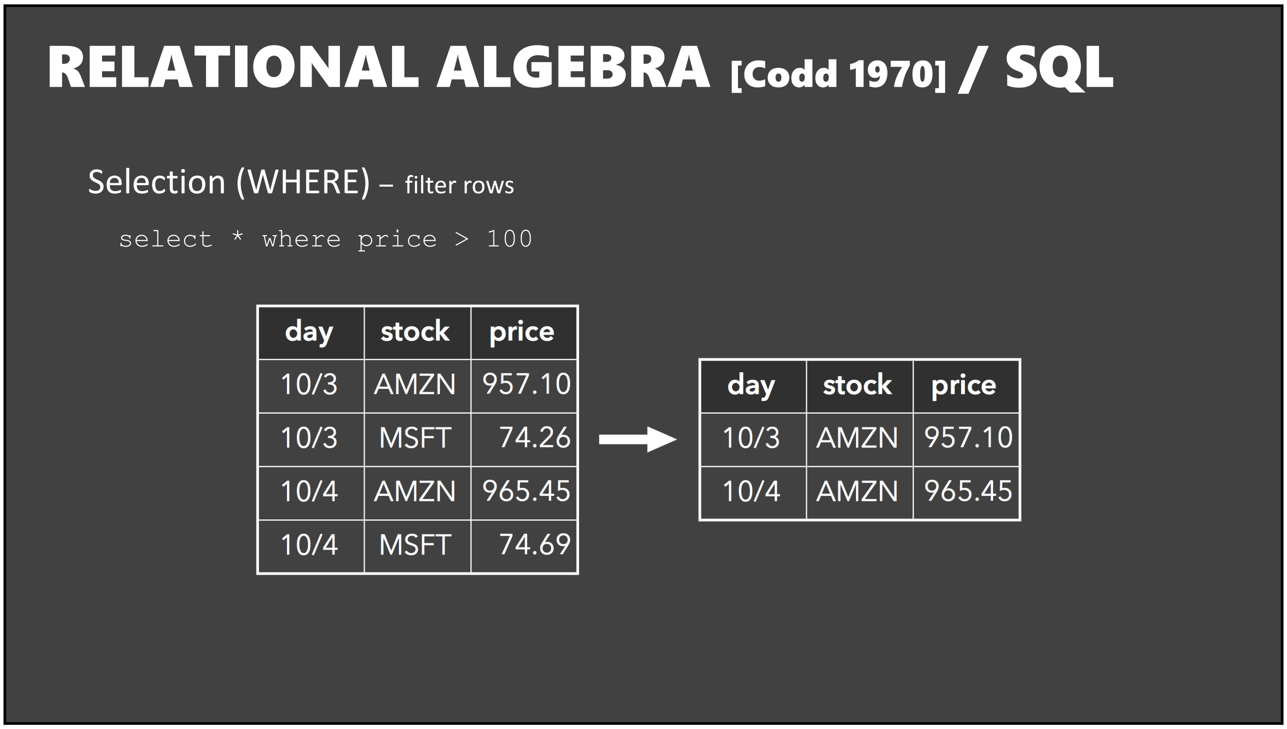
```
select day, stock
```

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69

→

day	stock
10/3	AMZN
10/3	MSFT
10/4	AMZN
10/4	MSFT

33



RELATIONAL ALGEBRA [Codd 1970] / SQL

Selection (WHERE) – filter rows

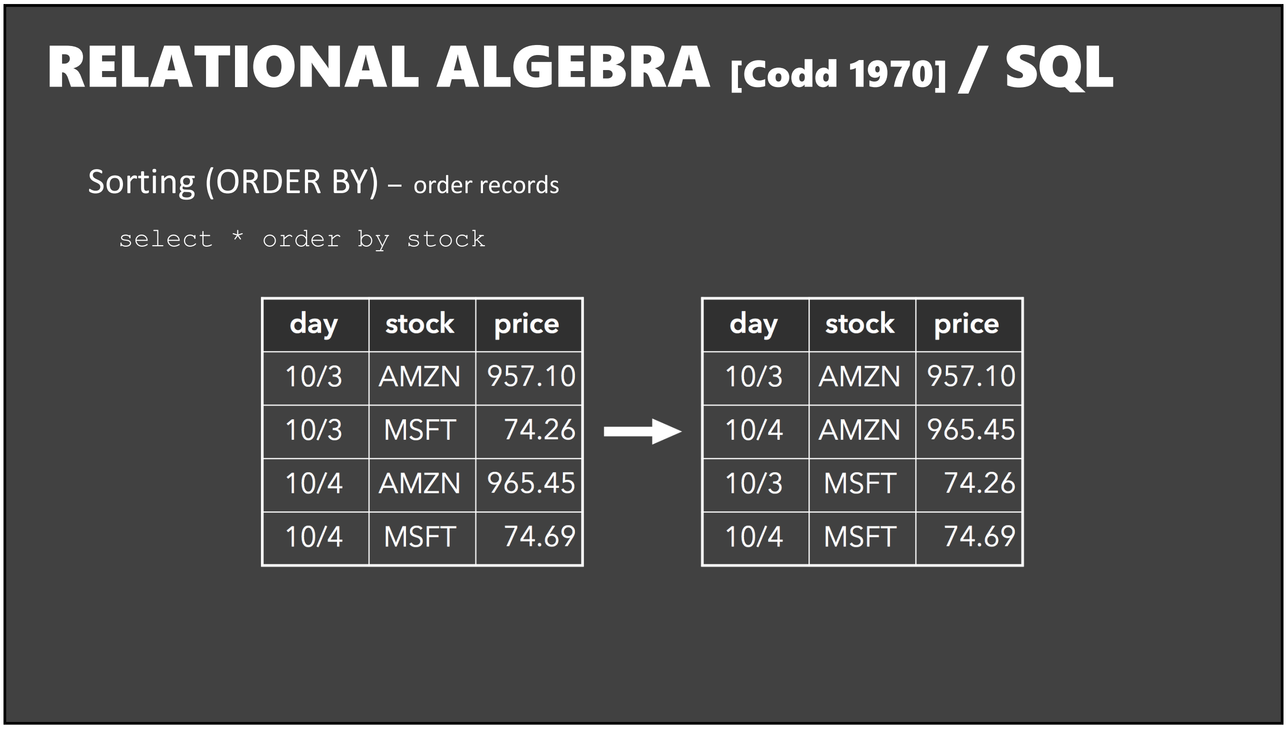
```
select * where price > 100
```

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69

→

day	stock	price
10/3	AMZN	957.10
10/4	AMZN	965.45

34



RELATIONAL ALGEBRA [Codd 1970] / SQL

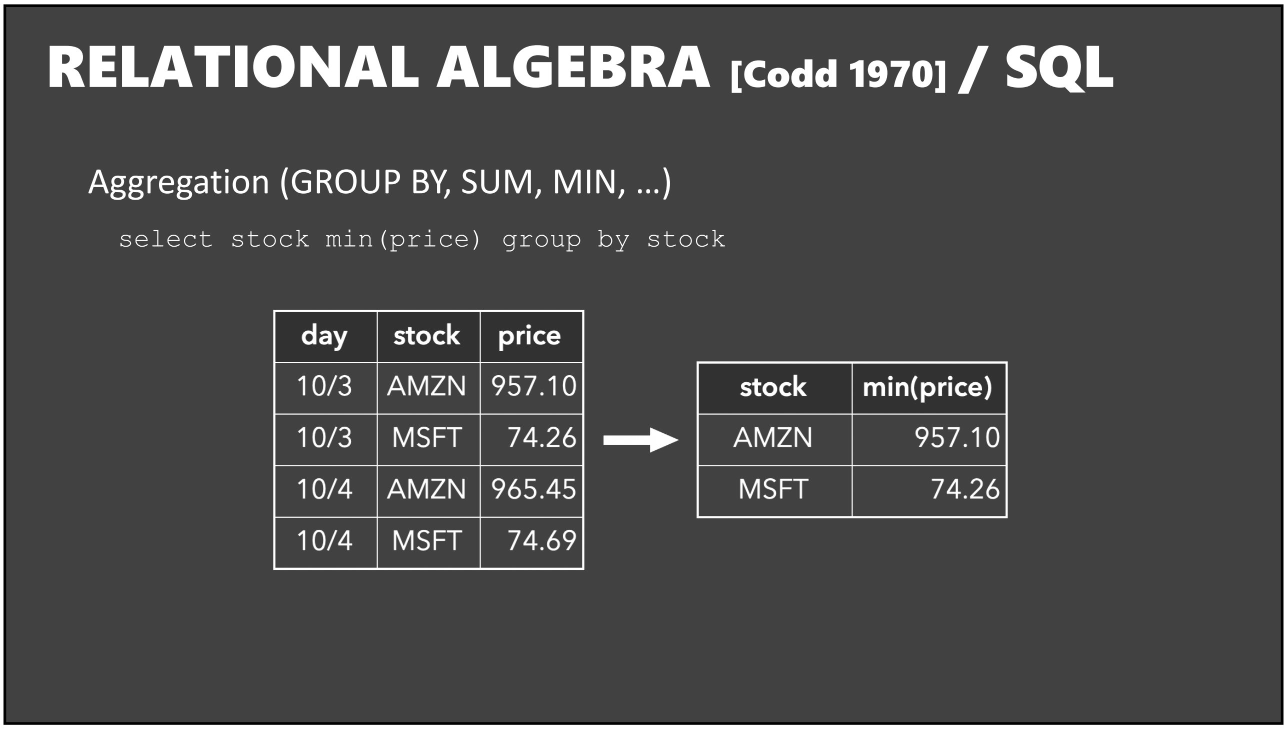
Sorting (ORDER BY) – order records

```
select * order by stock
```

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69

day	stock	price
10/3	AMZN	957.10
10/4	AMZN	965.45
10/3	MSFT	74.26
10/4	MSFT	74.69

35



RELATIONAL ALGEBRA [Codd 1970] / SQL

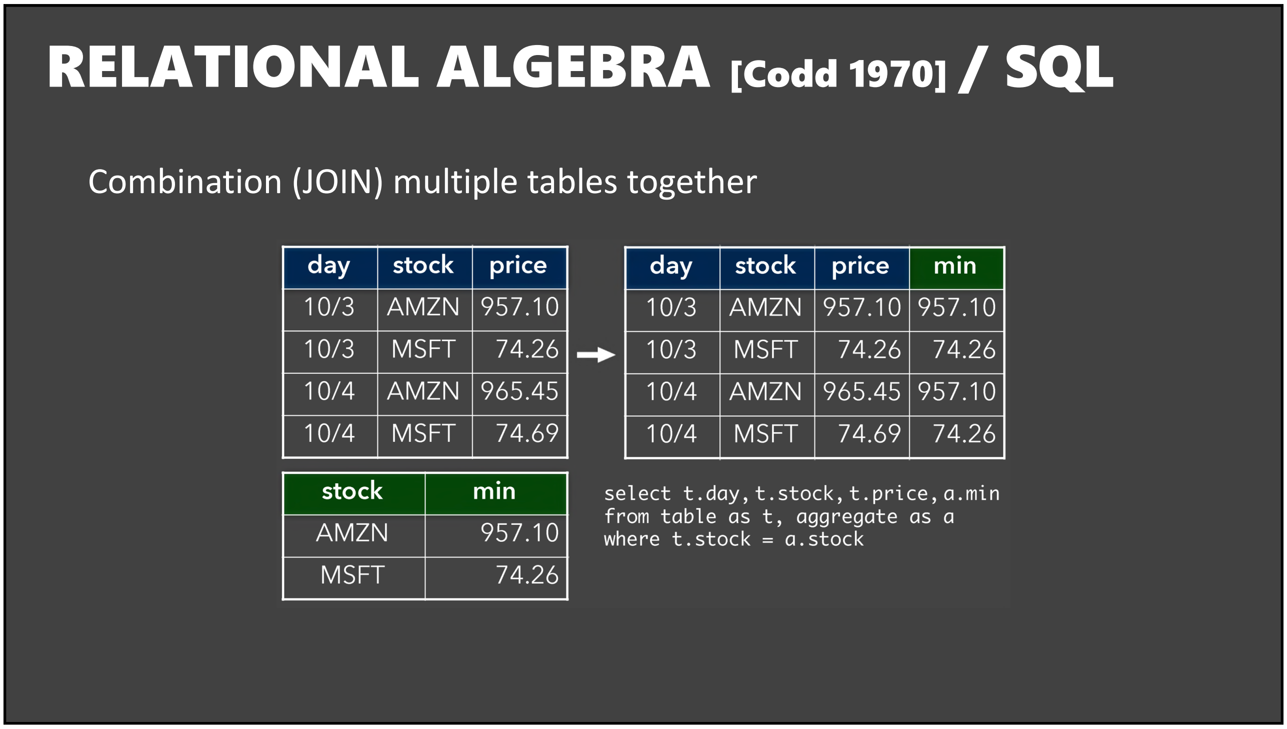
Aggregation (GROUP BY, SUM, MIN, …)

```
select stock min(price) group by stock
```

day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69

stock	min(price)
AMZN	957.10
MSFT	74.26

36



RELATIONAL ALGEBRA [Codd 1970] / SQL

Combination (JOIN) multiple tables together

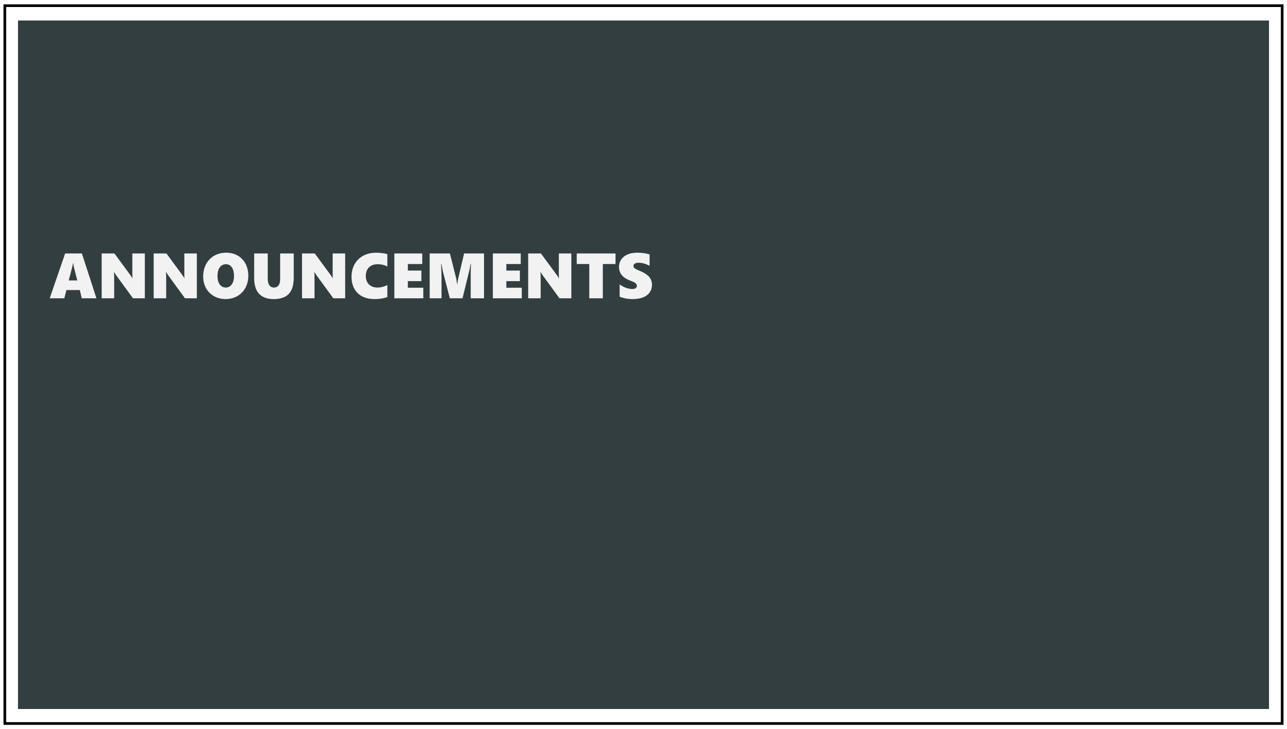
day	stock	price
10/3	AMZN	957.10
10/3	MSFT	74.26
10/4	AMZN	965.45
10/4	MSFT	74.69

stock	min
AMZN	957.10
MSFT	74.26

day	stock	price	min
10/3	AMZN	957.10	957.10
10/3	MSFT	74.26	74.26
10/4	AMZN	965.45	957.10
10/4	MSFT	74.69	74.26

```
select t.day, t.stock, t.price, a.min
from table as t, aggregate as a
where t.stock = a.stock
```

37



ANNOUNCEMENTS

44

CLASS PARTICIPATION REQUIREMENTS

Complete required **readings** and **notebooks** before class

Attend class and be a part of the in-class discussion

Post at least 1 discussion substantive comment/question per week

- Due by 8pm the following Sunday
- 1 free pass for the quarter

Class home page
https://magrawala.github.io/cs448b-fa25/

45

READING/NOTEBOOK/LECTURE RESPONSES

Good responses typically exhibit one or more

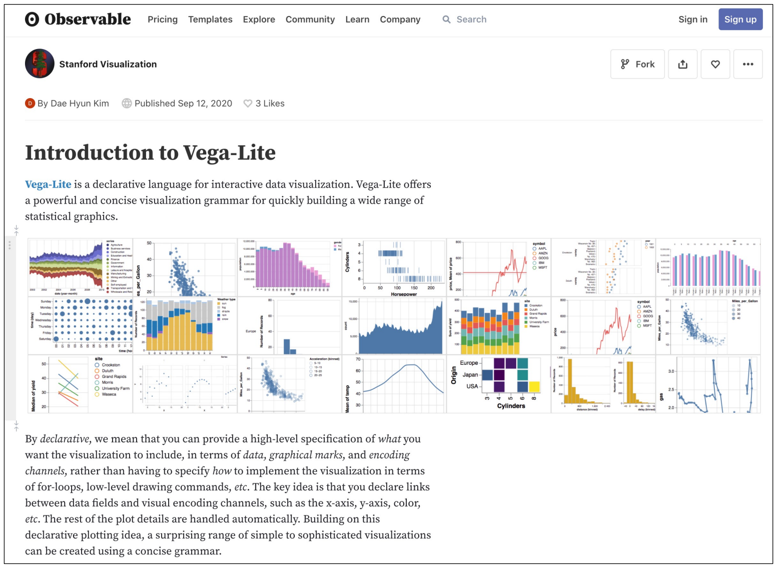
- **Critiques** of arguments made in the papers/lectures
- **Analysis** of implications or future directions for ideas in readings/lectures
- **Insightful questions** about the readings/lectures

Responses should not be summaries

- Should be substantive (1-2 paragraphs is typical)

46

OBSERVABLE NOTEBOOKS / VEGA-LITE



By Dae Hyun Kim · Published Sep 12, 2020 · 3 Likes

Introduction to Vega-Lite

Vega-Lite is a declarative language for interactive data visualization. Vega-Lite offers a powerful and concise visualization grammar for quickly building a wide range of statistical graphics.

By *declarative*, we mean that you can provide a high-level specification of *what* you want the visualization to include, in terms of *data*, *graphical marks*, and *encoding channels*, rather than having to specify *how* to implement the visualization in terms of for-loops, low-level drawing commands, *etc*. The key idea is that you declare links between data fields and visual encoding channels, such as the x-axis, y-axis, color, *etc*. The rest of the plot details are handled automatically. Building on this declarative plotting idea, a surprising range of simple to sophisticated visualizations can be created using a concise grammar.

Vega-Lite is a *declarative* API for programming visualizations

Do the exercises (fork notebook)

This Fri 9/26 10:30-11:30 We will run a Zoom session talking about the basics of Observable and how to do Data Wrangling using Tools in Observable.

47

ASSIGNMENT 1: VISUALIZATION DESIGN

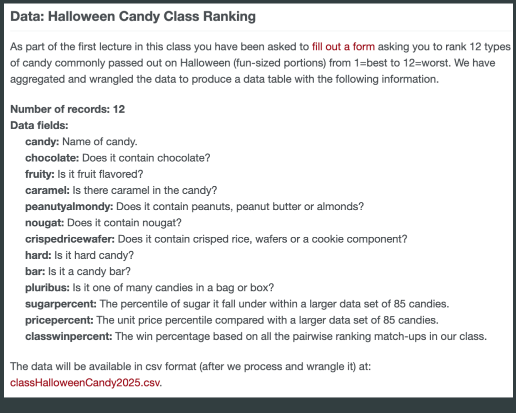
Due Mon 9/29 6am!

Design a static visualization for a data set

You must choose the message you want to convey. What question(s) do you want to answer? What insight do you want to communicate?

Submission guidelines

Make public Google Slide with your visualization
Put your name, suid and write-up in notes
Submit public link to slide on Canvas



Data: Halloween Candy Class Ranking

As part of the first lecture in this class you have been asked to fill out a form asking you to rank 12 types of candy commonly passed out on Halloween (fun-sized portions) from 1=best to 12=worst. We have aggregated and wrangled the data to produce a data table with the following information.

Number of records: 12
Data fields:

- **candy:** Name of candy.
- **chocolate:** Does it contain chocolate?
- **fruity:** Is it fruit flavored?
- **caramel:** Is there caramel in the candy?
- **peanutyalmondy:** Does it contain peanuts, peanut butter or almonds?
- **nougat:** Does it contain nougat?
- **crispedricewafer:** Does it contain crisped rice, wafers or a cookie component?
- **hard:** Is it hard candy?
- **bar:** Is it a candy bar?
- **pluribus:** Is it one of many candies in a bag or box?
- **sugarpercent:** The percentile of sugar it fall under within a larger data set of 85 candies.
- **pricepercent:** The unit price percentile compared with a larger data set of 85 candies.
- **classwinpercent:** The win percentage based on all the pairwise ranking match-ups in our class.

The data will be available in csv format (after we process and wrangle it) at: classHalloweenCandy2025.csv.

48

IMAGE

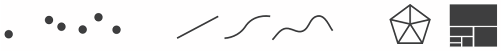
50

Semiology of Graphics
J. Bertin, 1967

MARKS & VISUAL ATTRs

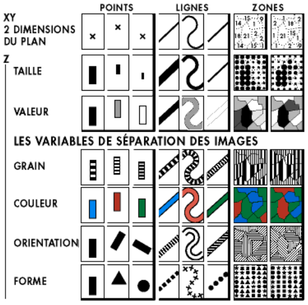
Marks: geometric primitives

points lines areas



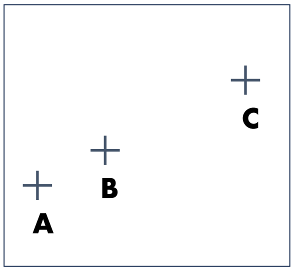
Visual Attributes: control mark appearance

- Position (2x)
- Size
- Value
- Texture
- Color
- Orientation
- Shape



51

CODING INFORMATION IN POSITION



1. A, B, C are distinguishable
2. Three points are colinear: B between A and C
3. BC is twice as long as AB

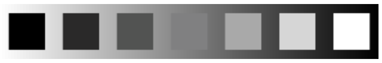
∴ Encode quantitative variables

"Resemblance, order and proportional are the three signfields in graphics." - Bertin

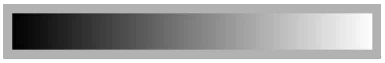
CODING INFORMATION IN COLOR

Value is perceived as ordered

∴ Encode ordinal variables (O)

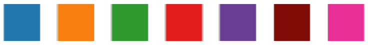


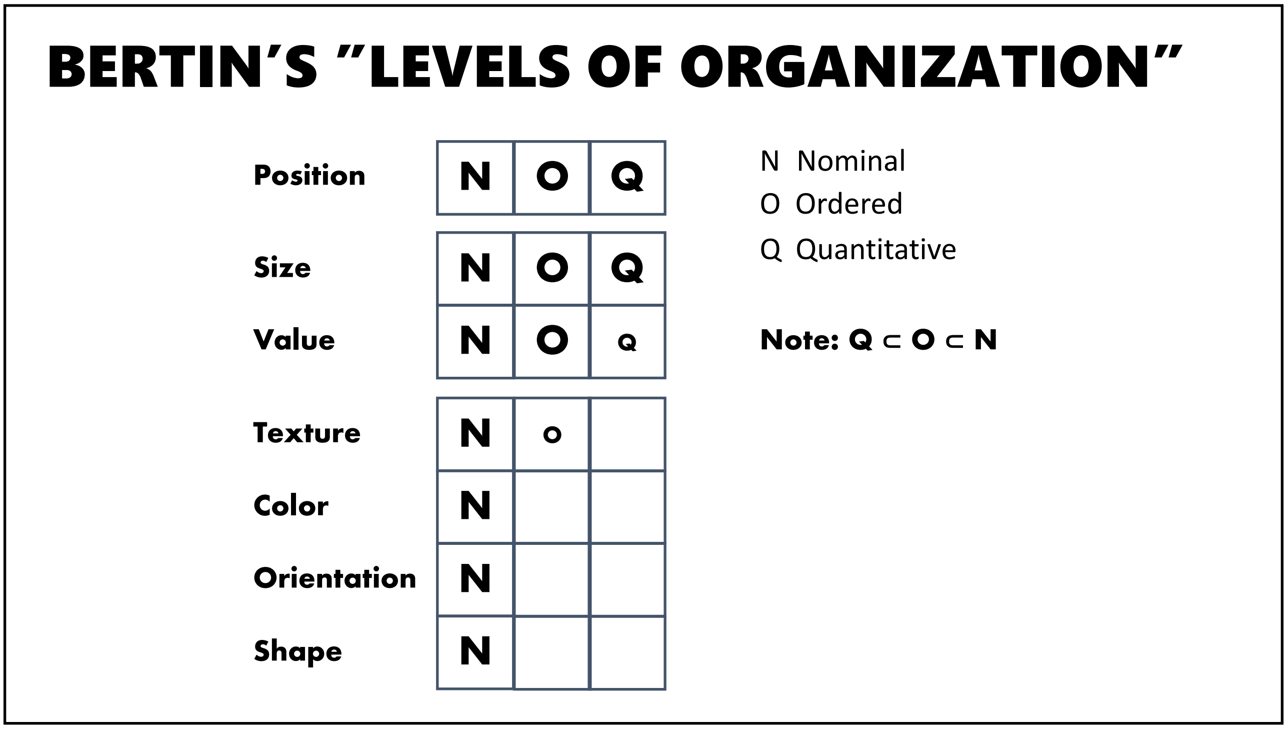
∴ Encode continuous variables (Q) [not as well]



Hue is normally perceived as unordered

∴ Encode nominal variables (N) using color





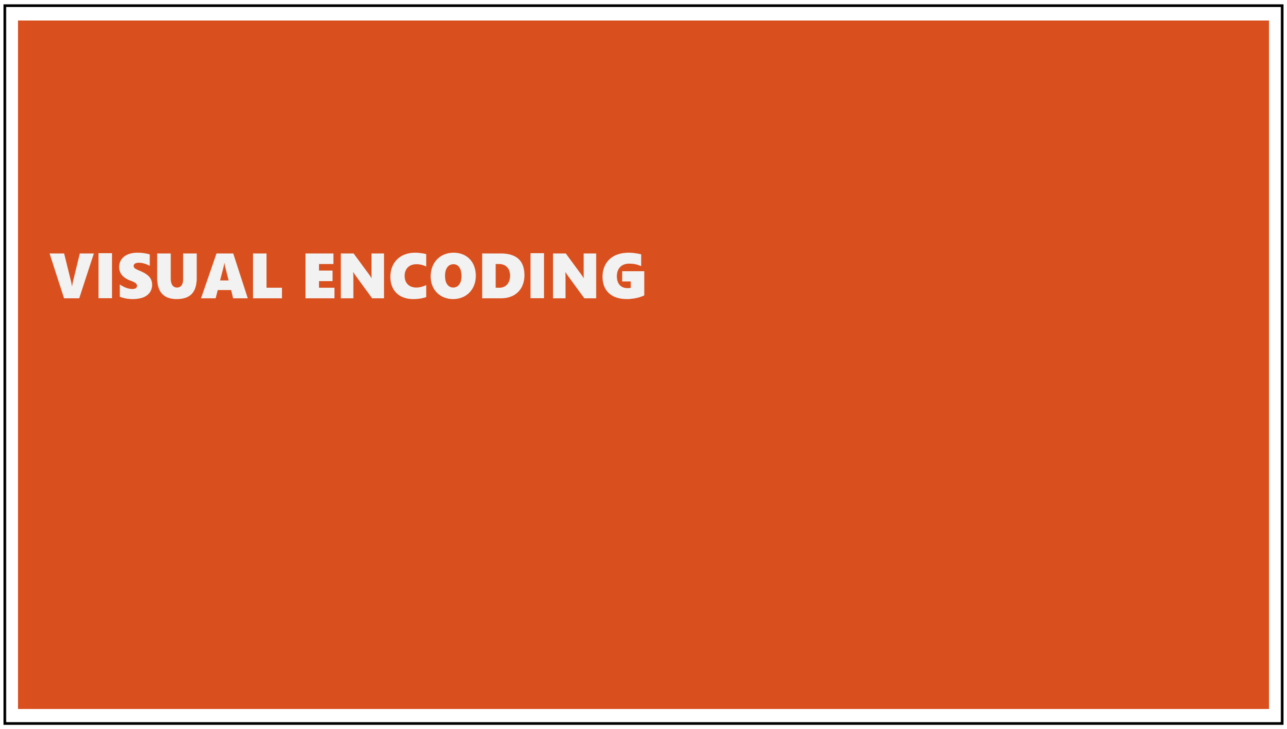
BERTIN'S "LEVELS OF ORGANIZATION"

Position	**N**	**O**	**Q**
Size	**N**	**O**	**Q**
Value	**N**	**O**	**Q**
Texture	**N**	**O**	
Color	**N**		
Orientation	**N**		
Shape	**N**		

N Nominal
O Ordered
Q Quantitative

Note: Q ⊂ O ⊂ N

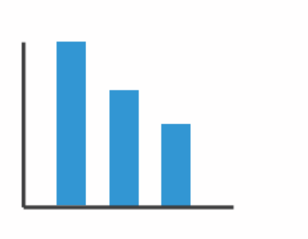
55



VISUAL ENCODING

56

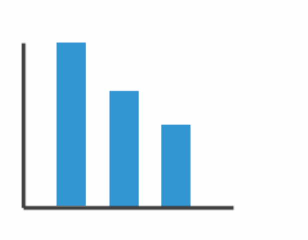
ENCODINGS: MAP DATA to MARK ATTRIBUTES



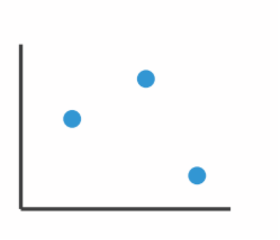
mark: rect
data → size (height)

57

ENCODINGS: MAP DATA to MARK ATTRIBUTES



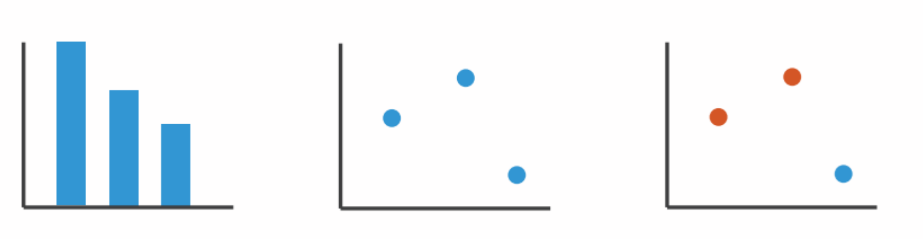
mark: rect
data → size (height)



mark: points
$data_1$ → x-pos
$data_2$ → y-pos

58

ENCODINGS: MAP DATA to MARK ATTRIBUTES



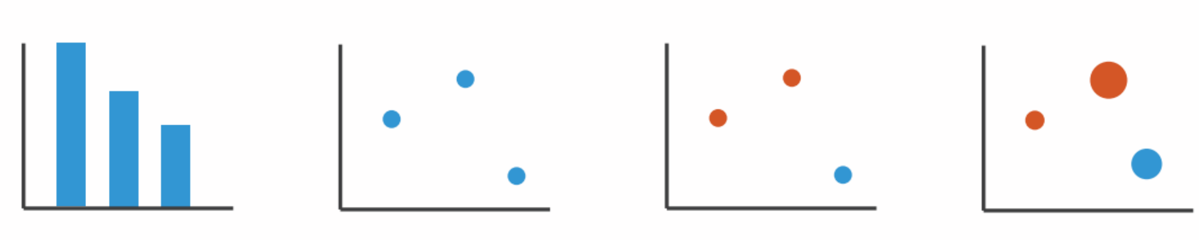
mark: rect
data → size (height)

mark: points
data$_1$ → x-pos
data$_2$ → y-pos

mark: points
data$_1$ → x-pos
data$_2$ → y-pos
data$_3$ → color

59

ENCODINGS: MAP DATA to MARK ATTRIBUTES



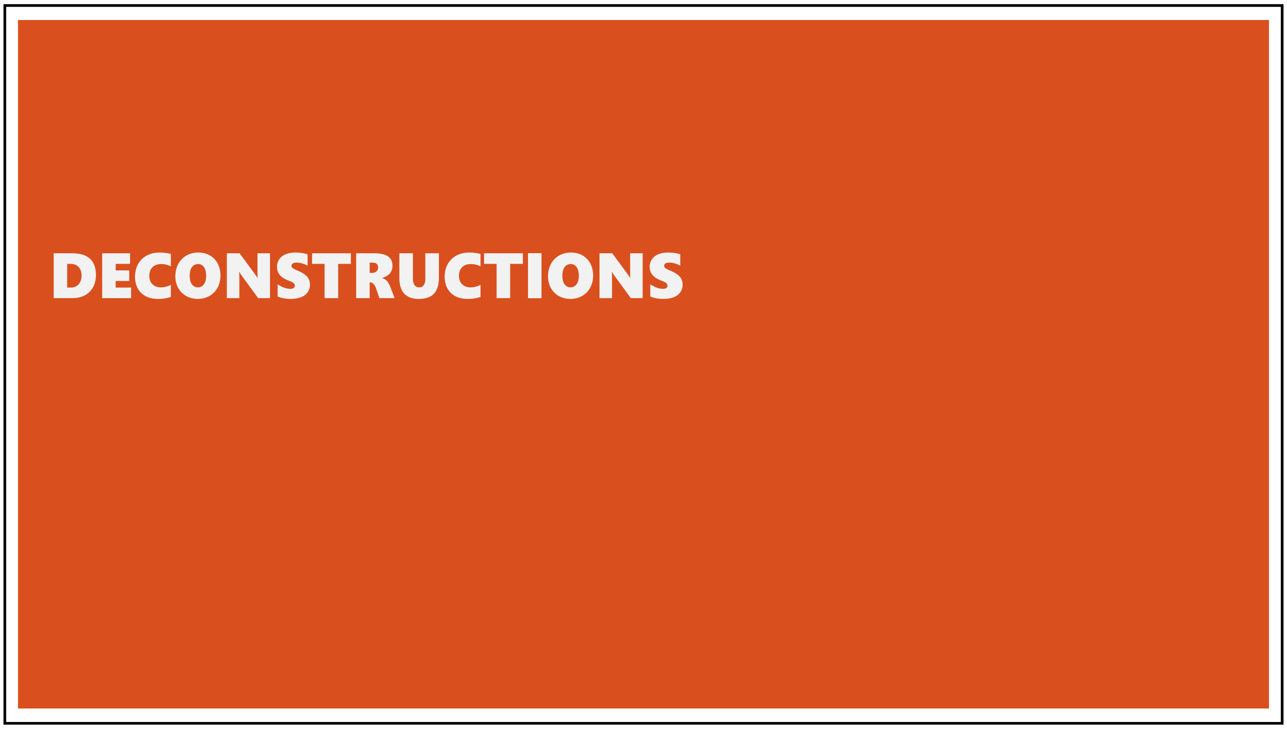
mark: rect
data → size (height)

mark: points
data$_1$ → x-pos
data$_2$ → y-pos

mark: points
data$_1$ → x-pos
data$_2$ → y-pos
data$_3$ → color

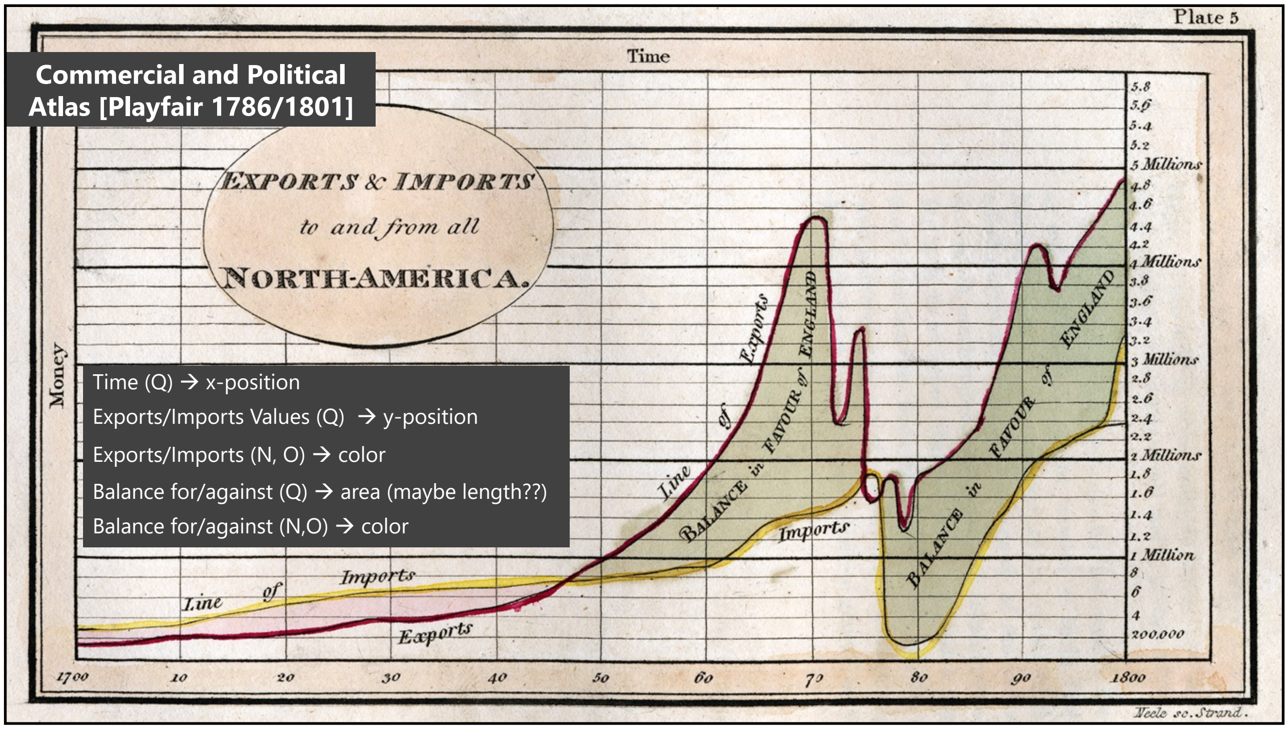
mark: points
data$_1$ → x-pos
data$_2$ → y-pos
data$_3$ → color
data$_4$ → size

60



DECONSTRUCTIONS

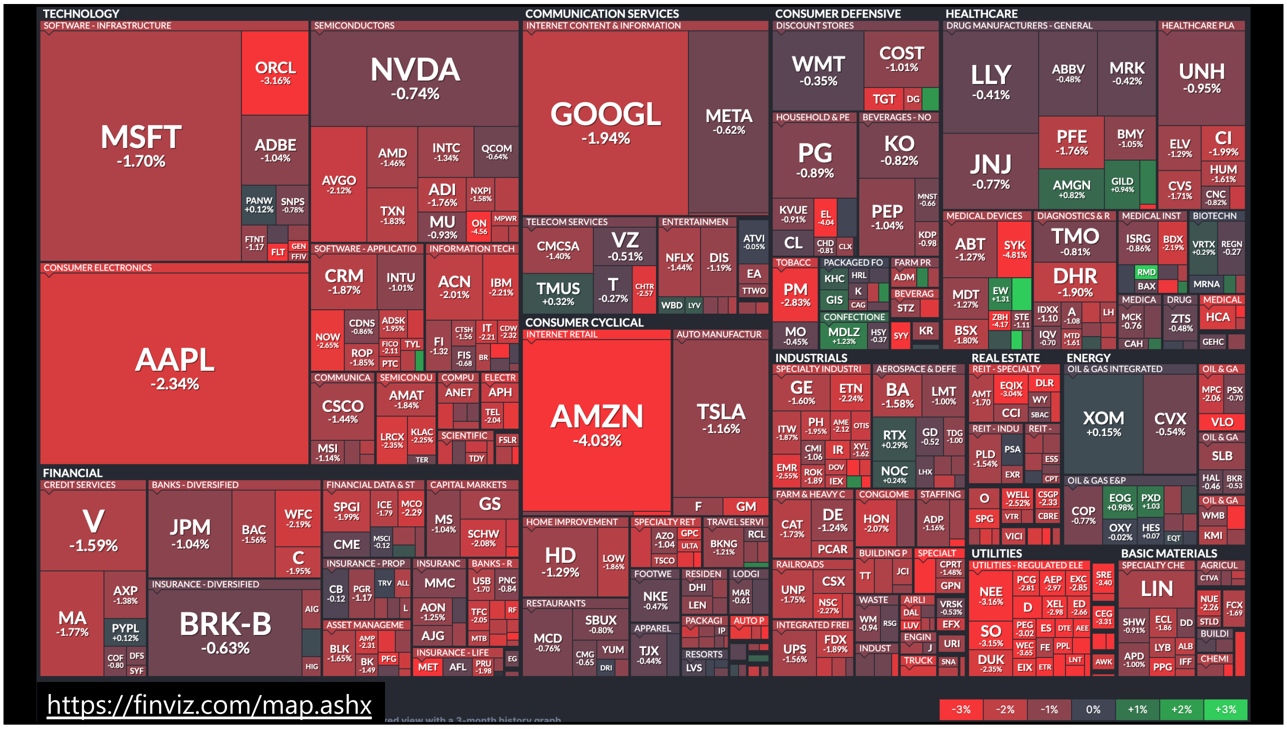
61



Commercial and Political Atlas [Playfair 1786/1801]

- Time (Q) → x-position
- Exports/Imports Values (Q) → y-position
- Exports/Imports (N, O) → color
- Balance for/against (Q) → area (maybe length??)
- Balance for/against (N,O) → color

64



https://finviz.com/map.ashx

65

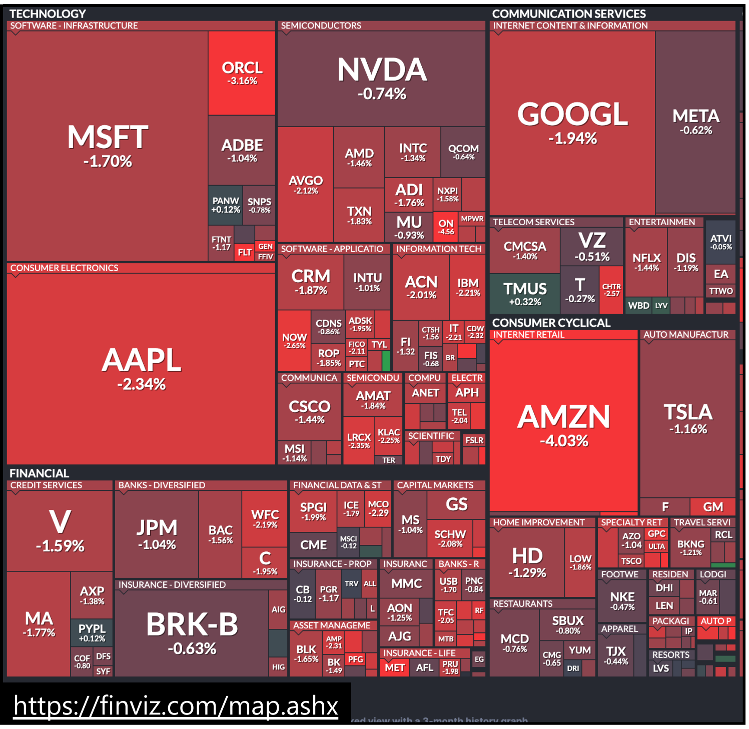
ENCODINGS

market cap (Q) → rectangle size

mkt sector (N), mkt cap (Q) → rect. pos.

loss vs. gain (N, O) → color hue

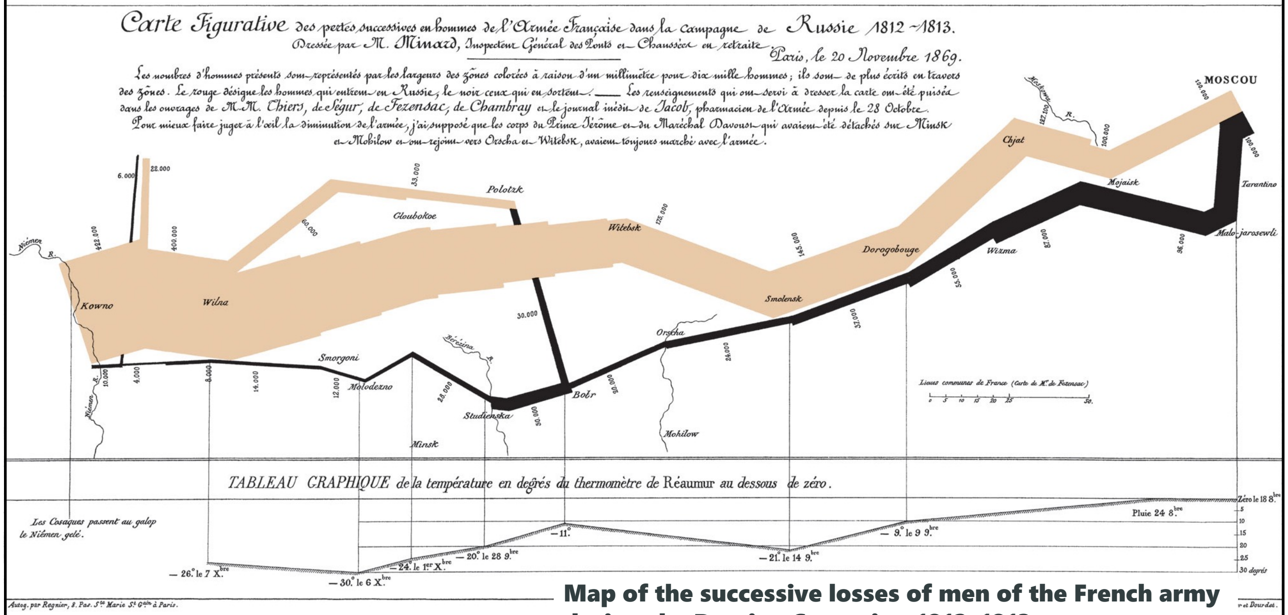
magnitude of loss or gain (Q) → color value



https://finviz.com/map.ashx

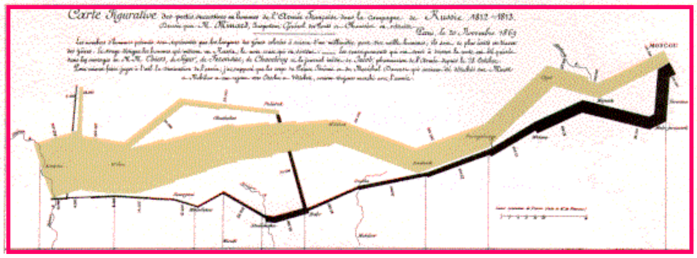
66

MINARD's MARCH on MOSCOW

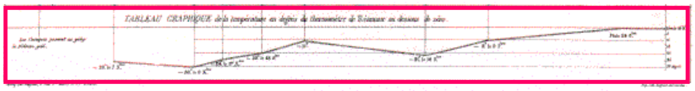


Map of the successive losses of men of the French army during the Russian Campaign 1812–1813

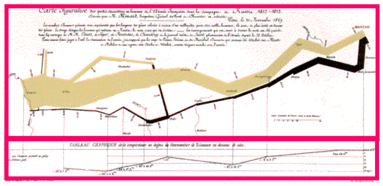
67



+



=



68

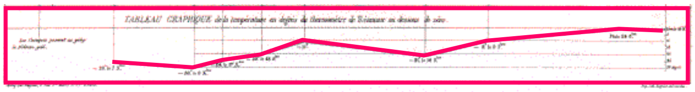
MARK COMPOSITION

temperature (Q) → y-position

+

longitude (Q), time (O) → x-position

=



temp across space & time (Q x Q,O)

MARK COMPOSITION

latitude (Q) → y-position

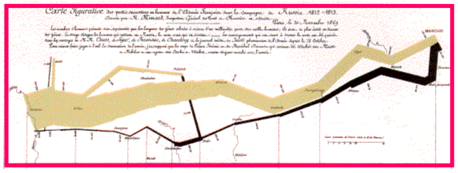
+

longitude (Q) → x-position

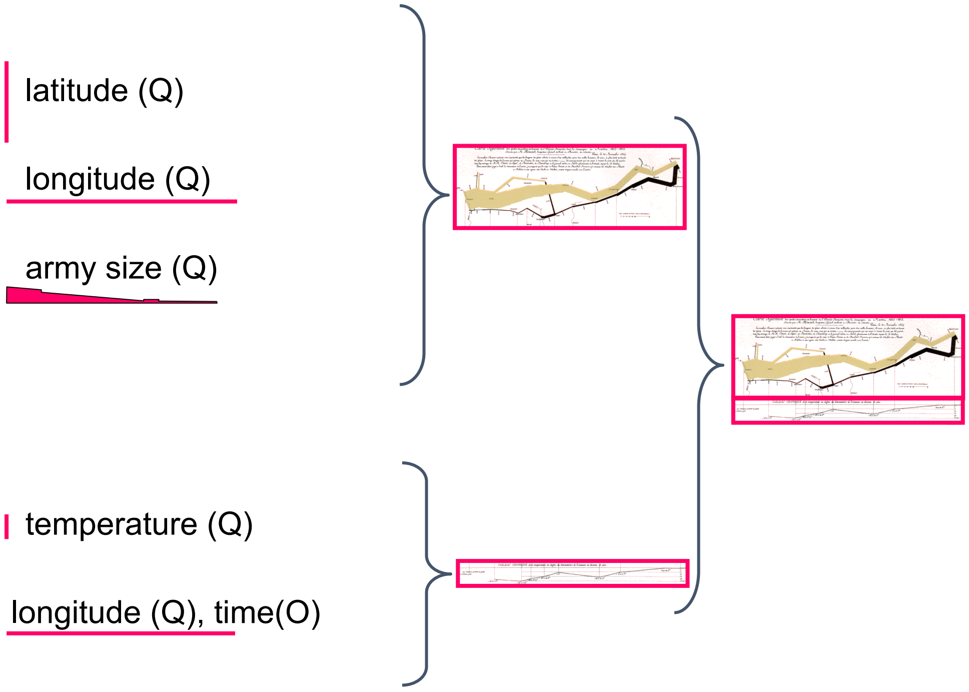
+

army size (Q) → width

=



army position (Q x Q) and army size (Q)



latitude (Q)

longitude (Q)

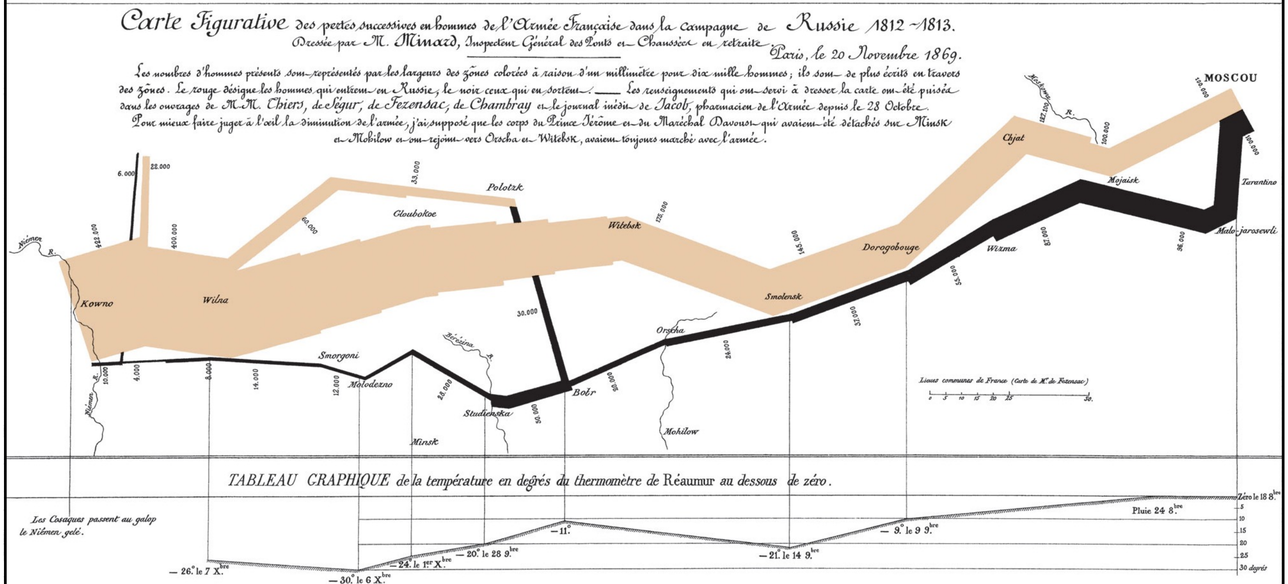
army size (Q)

temperature (Q)

longitude (Q), time(O)

71

MINARD's MARCH on MOSCOW



Depicts at least 5 quantitative variables. Any others?

72

FORMALIZING DESIGN

73

COMBINATORICS OF ENCODINGS

Challenge:

Assume **k** visual attributes/channels and **n** data fields

Pick the best encoding from the exponential number of possibilities **$(n+1)^k$**

74

PRINCIPLES

Challenge

Assume **k** visual attributes/channels and **n** data fields

Pick the best encoding from the exponential number of possibilities $\mathbf{(n+1)^k}$

Principle of Consistency

Properties of image (visual variables) should match properties of data

Principle of Importance Ordering

Encode most important information in the most effective way

75

EXPRESSIVENESS CRITERIA [Mackinlay 1986]

Expressiveness

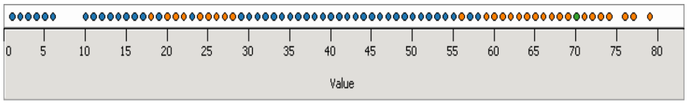
A set of facts is expressible in a visual language if the sentences (i.e., the visualizations) in the language express ***all*** the facts in the set of data, and ***only*** the facts in the data.

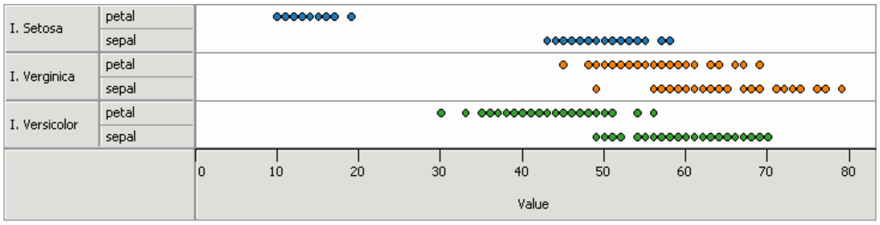
76

CANNOT EXPRESS ALL THE FACTS

Horizontal dot plot

A one-to-many (1 → N) relation cannot be expressed in a single horizontal dot plot because multiple tuples are mapped to the same position





77

EXPRESSES FACTS NOT IN THE DATA

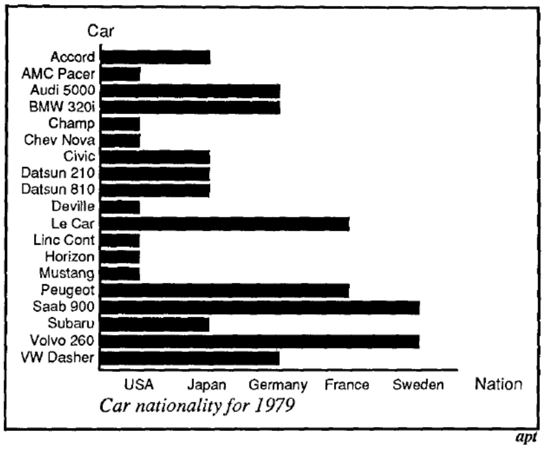


Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

Length is interpreted as encoding a quantitative value

78

EFFECTIVENESS CRITERIA [Mackinlay 1986]

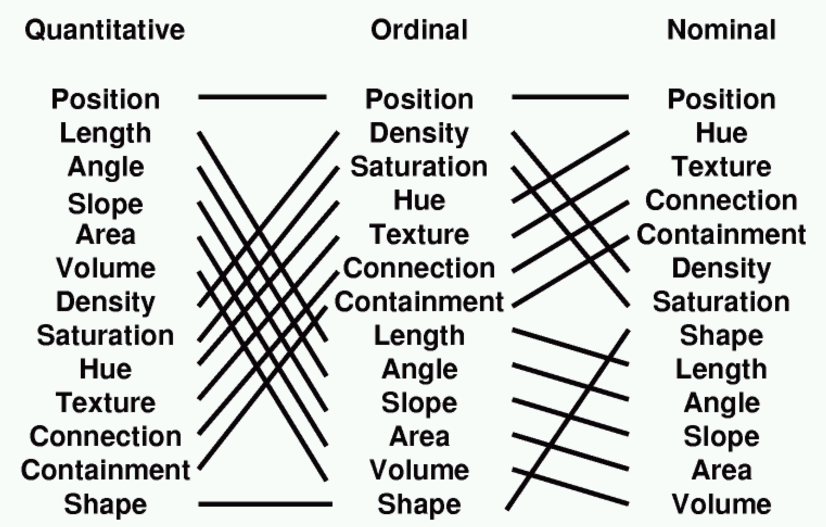
Effectiveness

A visualization is more effective than another visualization if the information conveyed by one visualization is more readily ***perceived*** than the information in the other visualization.

Subject of the Perception Lecture

79

MACKINLAY’S RANKING



Quantitative	Ordinal	Nominal
Position	Position	Position
Length	Density	Hue
Angle	Saturation	Texture
Slope	Hue	Connection
Area	Texture	Containment
Volume	Connection	Density
Density	Containment	Saturation
Saturation	Length	Shape
Hue	Angle	Length
Texture	Slope	Angle
Connection	Area	Slope
Containment	Volume	Area
Shape	Shape	Volume

Conjectured *effectiveness* of encodings by data type

80

AUTOMATIC CHART DESIGN [Mackinlay 1986]

APT – "A Presentation Tool"

User formally specifies data model and type
Input: list of data variables ordered by importance

APT searches over the design space
Tests expressiveness of each visual encoding (rule-based)
Generates encodings that pass test
Rank by perceptual effectiveness criteria

Outputs *most effective* visualization

81

Automating the design of graphical presentation of relational information
J. Mackinlay, 1986

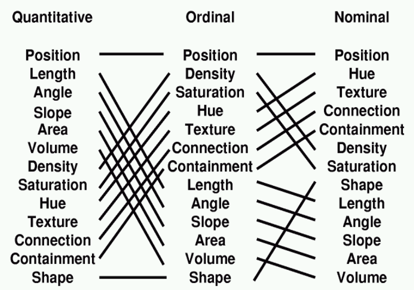
APT [Mackinlay 1986]

Encode most important data using highest ranking visual variable for the data type

Price	Mileage	Weight	Repair
13,500	22	3000	great
7,200	31	1500	ok
11,300	12	4200	terrible
...	...	...	...

→

1. Price (Q)
2. Mileage (Q)
3. Weight (Q)
4. Repair (N)



Quantitative	Ordinal	Nominal
Position	Position	Position
Length	Density	Hue
Angle	Saturation	Texture
Slope	Hue	Connection
Area	Texture	Containment
Volume	Connection	Density
Density	Containment	Saturation
Saturation	Length	Shape
Hue	Angle	Length
Texture	Slope	Angle
Connection	Area	Slope
Containment	Volume	Area
Shape	Shape	Volume

→

mark: lines

Price (Q) → y-pos
Mileage (Q) → x-pos
Weight (Q) → size
Repair (N) → color

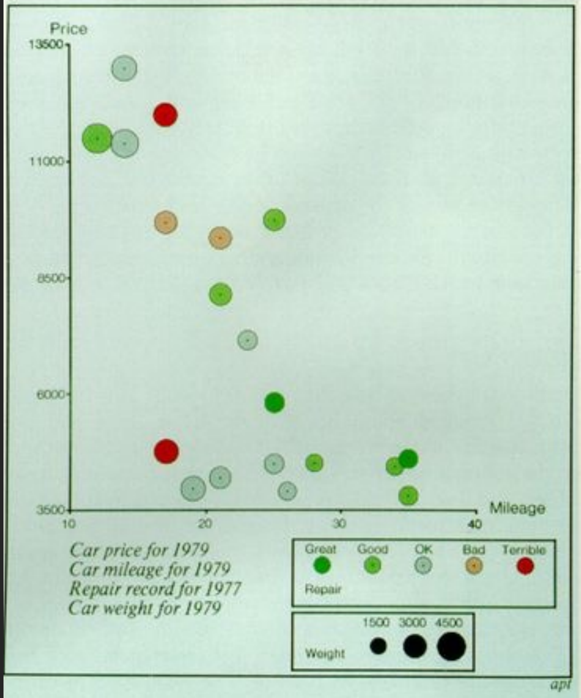
83

APT [Mackinlay 1986]

Automatically generated chart for cars data

Cars Data

1. Price (Q)
2. Mileage (Q)
3. Weight (Q)
4. Repair (Q)



LIMITATIONS

Does not cover many visualization techniques

Networks, maps, diagrams

Also, 3D, animation, illustration, …

Does not consider interaction

Does not consider semantics or conventions

Assumes single visualization as output

SUMMARY

Formal specification

Data model: tidy data, N,O,Q types

Image model: marks, visual attributes/channels

Encodings map data to mark attributes/channels

Choose *expressive* and *effective* encodings

Rule-based test of expressiveness

Perceptual effectiveness rankings

86

87