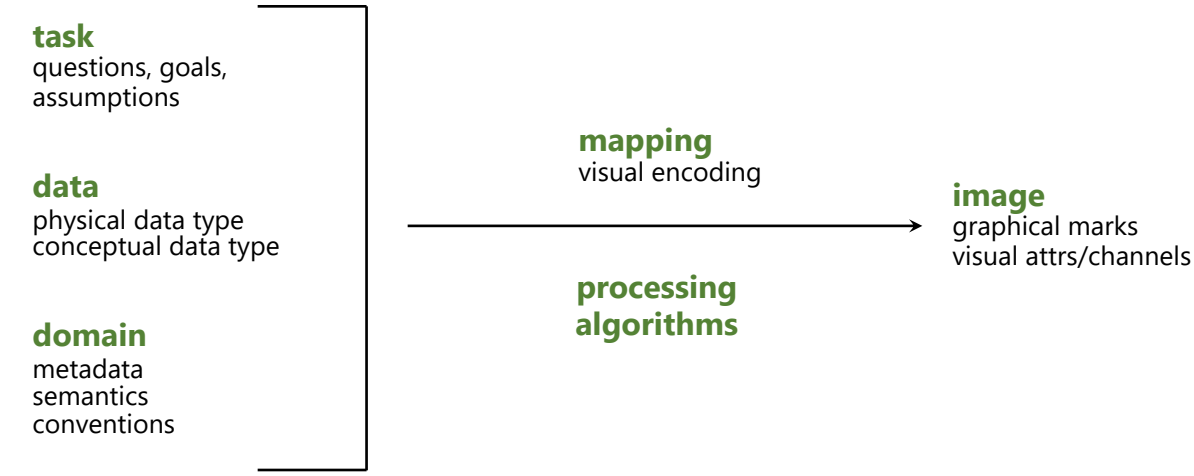
# DATA & IMAGE MODELS

CS 448B | Fall 2024

MANEESH AGRAWALA

1

## The big picture




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


Skip

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**

| | 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/

12

Dataset

Data Field

**Halloween Candy Power Ranking Dataset**

| | 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 |

Data Item/Observation

Cell Value

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

13

# 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

| | 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


# 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).

| | 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

23

# 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

24

# 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 |

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

# 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 |

27

# DATA TABLES & TRANSFORMATIONS

28

# 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

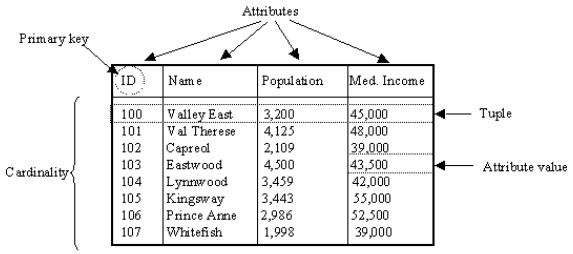
Sorting (ORDER BY) – order rows

Aggregation (GROUP BY, SUM, MIN, …)

partition rows into groups and summarize

Combination (JOIN, UNION, …)

integrate data from multiple tables




29

# 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 |

30

# 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 |

31

# 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 |

32

# 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 |

33

# 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
```

34

# ANNOUNCEMENTS

41

# 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-fa24/

42

# READING/NOTEBOOK/LECTURE RESPONSES

**Good responses typically exhibit one or more**

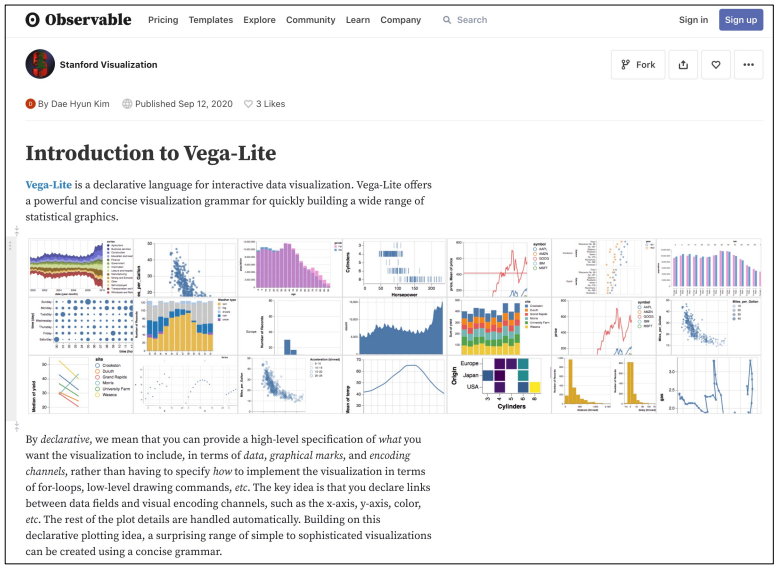
- **Critiques** of arguments made in the papers/lectures
- **Analysis** of implications or future directions for ideas in readings/lectures
- **Insightful questions** about the readings/lectures

**Responses should not be summaries**

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

43

# OBSERVABLE NOTEBOOKS / VEGA-LITE




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

**Do the exercises** (fork notebook)

**This Fri 9/27 at 9am** Mhar will run a Zoom session talking about the basics of Observable and how to do Data Wrangling using Tools in Observable.

44

# ASSIGNMENT 1: VISUALIZATION DESIGN

**Due Mon 9/30 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?

## Updated 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: Stanford Olympic Medals**

Stanford Athletics recently published a dataset on the Olympic Medal History of Stanford students and alumni. We have extracted and wrangled a data table containing information Olympic medals won by Stanford students since 1912. Our data contains the following information:

**Number of records: 335**
**Data fields:**

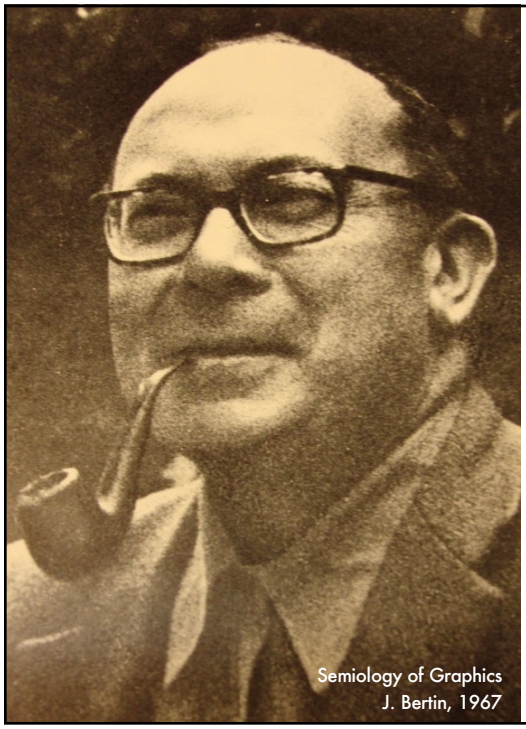
- **Athlete Name:** Name of athlete.
- **Host City:** Name of city that hosted the Olympics.
- **Host Country:** Name of country that hosted the Olympics.
- **Year:** Year the summer Olympics took place.
- **Athlete Team Country:** Country whose team the athlete represented.
- **Sport:** Name of sport athlete competed in.
- **Event:** Name of event athlete competed in ("–" indicates event name is same as name of sport).
- **Medal:** Type of medal won (ties are indicated in parentheses).

The extracted dataset is available in csv format: olympic_athletes-wrangled-2024.csv

45

# IMAGE

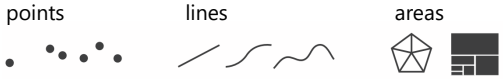
47




# MARKS & VISUAL ATTRs

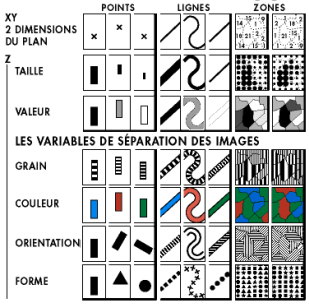
**Marks:** geometric primitives

points lines areas



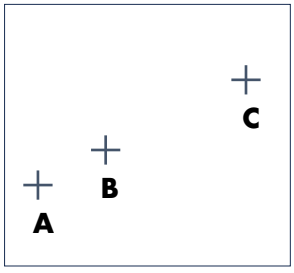
**Visual Attributes:** control mark appearance

Position (2x)
Size
Value
Texture
Color
Orientation
Shape




48

# 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**

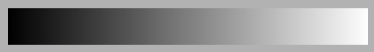
50

# 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

51

# 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

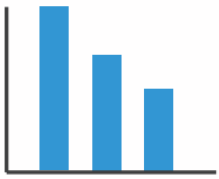
**Note: Q ⊂ O ⊂ N**

52

# VISUAL ENCODING

53

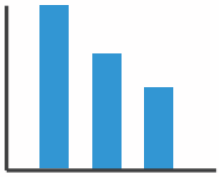
## ENCODINGS: MAP DATA to MARK ATTRIBUTES



mark: rect
data → size (height)

54

# ENCODINGS: MAP DATA to MARK ATTRIBUTES

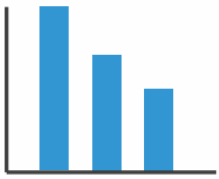


mark: rect
data → size (height)

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

55

# ENCODINGS: MAP DATA to MARK ATTRIBUTES



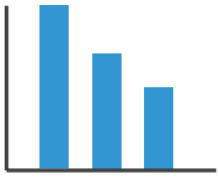
mark: rect
data → size (height)

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

mark: points
$data_1$ → x-pos
$data_2$ → y-pos
$data_3$ → color

56

# ENCODINGS: MAP DATA to MARK ATTRIBUTES



mark: rect
data → size (height)

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

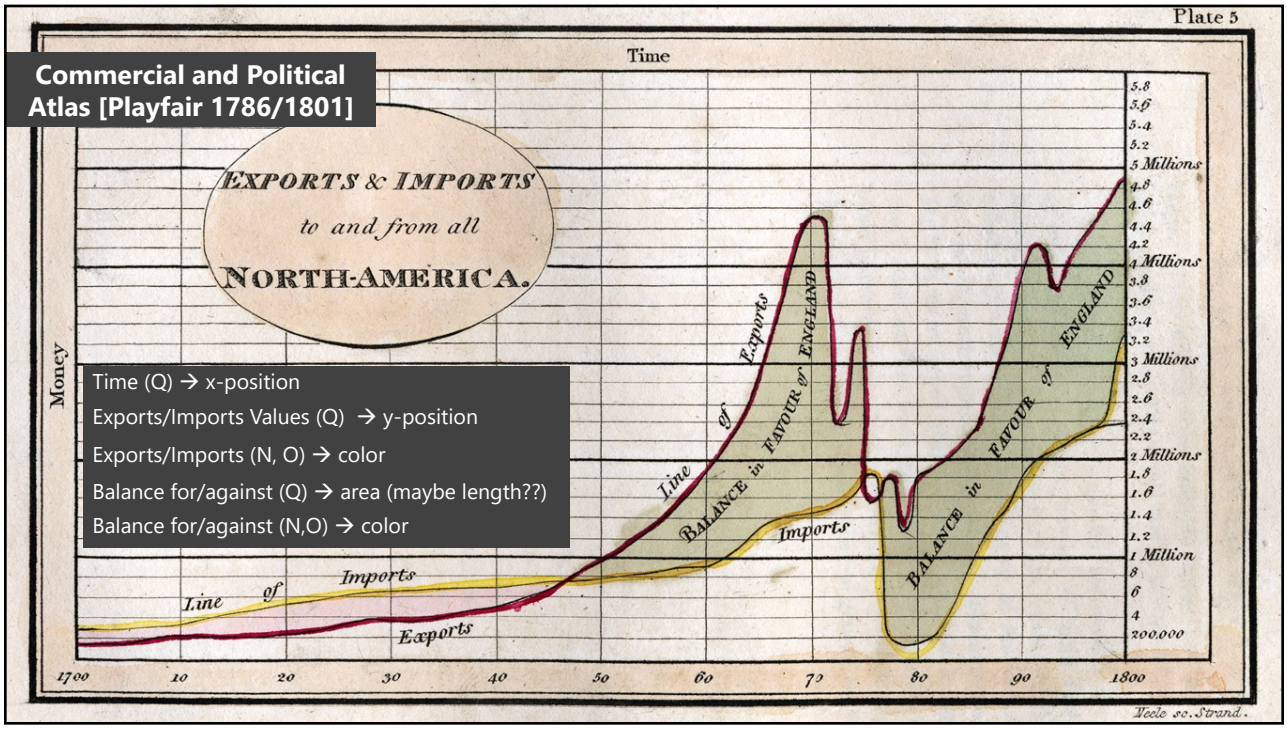
mark: points
$data_1$ → x-pos
$data_2$ → y-pos
$data_3$ → color

mark: points
$data_1$ → x-pos
$data_2$ → y-pos
$data_3$ → color
$data_4$ → size

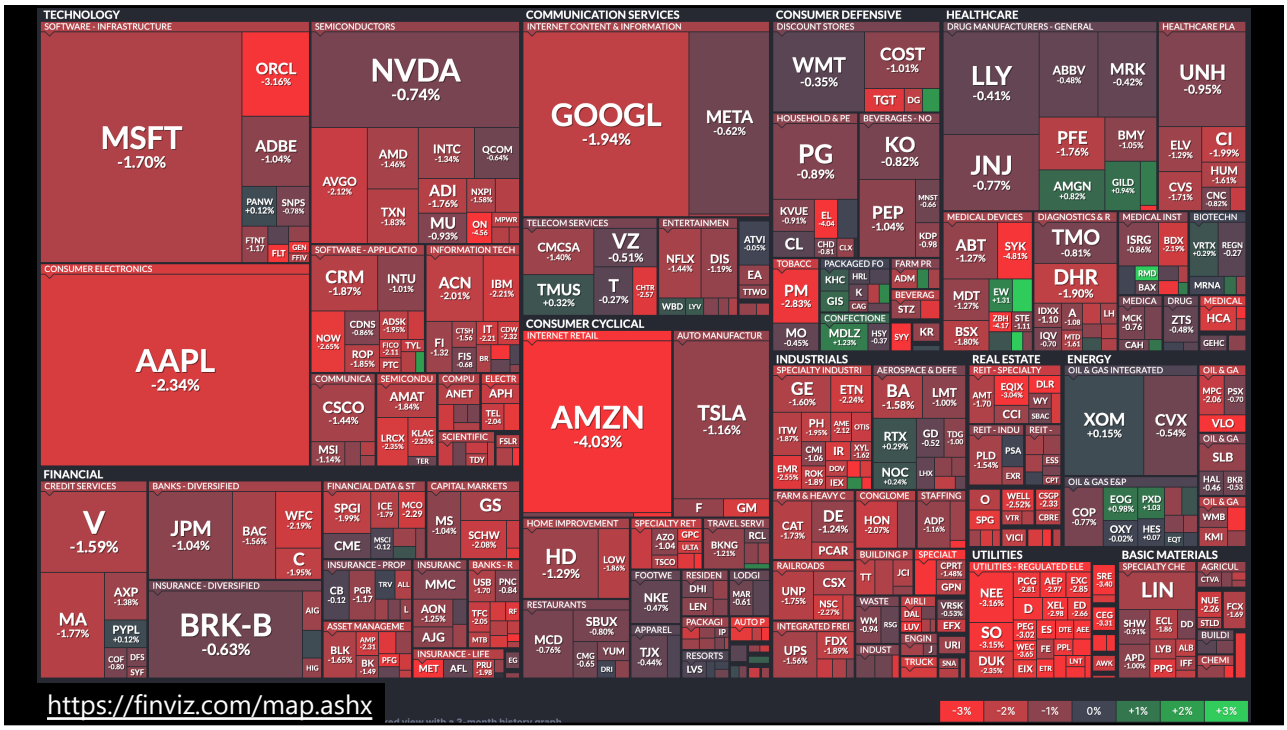
57

# DECONSTRUCTIONS

58


Commercial and Political Atlas [Playfair 1786/1801]
Plate 5
Time
EXPORTS & IMPORTS
to and from all
NORTH-AMERICA.
Money
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
Line of Imports
Exports
BALANCE in FAVOUR of ENGLAND
Line of Exports
Imports
BALANCE in FAVOUR of ENGLAND
5.8
5.6
5.4
5.2
5 Millions
4.8
4.6
4.4
4.2
4 Millions
3.8
3.6
3.4
3.2
3 Millions
2.8
2.6
2.4
2.2
2 Millions
1.8
1.6
1.4
1.2
1 Million
8
6
4
200,000
1700
10
20
30
40
50
60
70
80
90
1800
Veele sc. Strand.

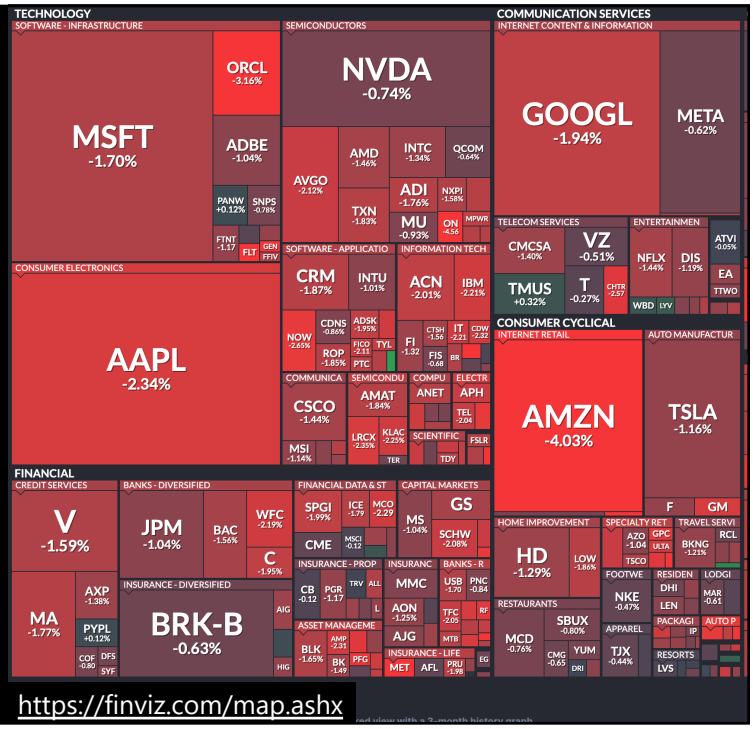
61


TECHNOLOGY
SOFTWARE - INFRASTRUCTURE
MSFT -1.70%
ORCL -3.16%
ADBE -1.04%
PANW +0.12%
SNPS -0.78%
FTNT -1.17%
GEN
FLT
FFIV
CONSUMER ELECTRONICS
AAPL -2.34%
SEMICONDUCTORS
NVDA -0.74%
AVGO -2.12%
AMD -1.46%
INTC -1.34%
QCOM -0.64%
TXN -1.83%
ADI -1.76%
NXPI -1.58%
MU -0.93%
ON -4.56%
MPWR
SOFTWARE - APPLICATIO
INFORMATION TECH
CRM -1.87%
INTU -1.01%
ACN -2.01%
IBM -2.21%
CDNS -0.86%
ADSK -1.95%
NOW -2.65%
ROP -1.85%
FICO -2.11%
TYL
CTSH -1.56
IT -2.21
CDW -2.32
FI -1.32
FIS -0.68
BR
COMMUNICA
SEMICONDU
COMPU
ELECTR
CSCO -1.44%
AMAT -1.84%
ANET
APH
TEL -2.04
MSI -1.14%
LRCX -2.35%
KLAC -2.22%
TER
SCIENTIFIC
TDY
FSLR
FINANCIAL
CREDIT SERVICES
BANKS - DIVERSIFIED
V -1.59%
JPM -1.04%
BAC -1.56%
WFC -2.19%
C -1.95%
AXP -1.38%
MA -1.77%
PYPL +0.12%
COF -0.80
DFS
SYF
INSURANCE - DIVERSIFIED
BRK-B -0.63%
AIG
HIG
FINANCIAL DATA & ST
CAPITAL MARKETS
SPGI -1.99%
ICE -1.79
MCO -2.29
CME
MSCI -0.12
MS -1.04%
GS
SCHW -2.08%
INSURANCE - PROP
INSURANC
BANKS - R
CB -0.12
PGR -1.17
TRV
ALL
L
MMC
USB -1.70
PNC -0.84
AON -1.25%
TFC -2.05
RF
ASSET MANAGEME
AMP -2.31
AJG
MTB
BLK -1.65%
BK -1.49
PFG
INSURANCE - LIFE
EG
MET
AFL
PRU -1.98
COMMUNICATION SERVICES
INTERNET CONTENT & INFORMATION
GOOGL -1.94%
META -0.62%
TELECOM SERVICES
ENTERTAINMEN
CMCSA -1.40%
VZ -0.51%
TMUS +0.32%
T -0.27%
CHTR -2.57
NFLX -1.44%
DIS -1.19%
WBD
LYV
ATVI -0.05%
EA
TTWO
CONSUMER CYCLICAL
INTERNET RETAIL
AUTO MANUFACTUR
AMZN -4.03%
TSLA -1.16%
F
GM
HOME IMPROVEMENT
SPECIALTY RET
TRAVEL SERVI
HD -1.29%
LOW -1.86%
AZO -1.04
GPC
ULTA
TSCO
BKNG -1.21%
RCL
FOOTWE
RESIDEN
LODGI
NKE -0.47%
DHI
LEN
MAR -0.61
RESTAURANTS
APPAREL
PACKAGI
IP
AUTO P
MCD -0.76%
SBUX -0.80%
CMG -0.65
YUM
DRI
TJX -0.44%
RESORTS
LVS
CONSUMER DEFENSIVE
DISCOUNT STORES
WMT -0.35%
COST -1.01%
TGT
DG
HOUSEHOLD & PE
BEVERAGES - NO
PG -0.89%
KO -0.82%
KVUE -0.91%
EL -4.04
CHD -0.81
CLX
PEP -1.04%
MNST -0.66
KDP -0.98
TOBACC
PACKAGED FO
FARM PR
CL
KHC
HRL
K
ADM
BEVERAG
PM -2.83%
GIS
CAG
STZ
CONFECTIONE
MDLZ +1.23%
HSY -0.37
SYY
KR
MO -0.45%
BSX -1.80%
HEALTHCARE
DRUG MANUFACTURERS - GENERAL
HEALTHCARE PLA
LLY -0.41%
ABBV -0.48%
MRK -0.42%
UNH -0.95%
JNJ -0.77%
PFE -1.76%
BMY -1.05%
AMGN +0.82%
GILD +0.96%
ELV -1.29%
CI -1.99%
CVS -1.71%
HUM -1.61%
CNC -0.82%
MEDICAL DEVICES
DIAGNOSTICS & R
MEDICAL INST
BIOTECHN
ABT -1.27%
SYK -4.81%
TMO -0.81%
DHR -1.90%
ISRG -0.86%
BDX -2.19%
VRTX +0.29%
REGN -0.27
MDT -1.27%
EW +1.31
RMD
BAX
MRNA
IDXX -1.10
A -1.08
LH
MEDICA
DRUG
MEDICAL
MCK -0.76
ZTS -0.48%
HCA
ZBH -4.17
STE -1.11
IQV -0.70
MTD -1.61
CAH
GEHC
INDUSTRIALS
SPECIALTY INDUSTRI
AEROSPACE & DEFE
GE -1.60%
ETN -2.24%
BA -1.58%
LMT 1.00%
ITW -1.87%
PH -1.95%
AME -2.12
OTIS
RTX +0.29%
GD -0.32
TDG 1.00
CMI -1.06
IR
XYL -1.62
EMR -2.55%
ROK -1.89
DOV
IEX
NOC +0.24%
LHX
FARM & HEAVY C
CONGLOME
STAFFING
CAT -1.73%
DE 1.24%
HON -2.07%
ADP -1.16%
PCAR
RAILROADS
BUILDING P
SPECIALT
UNP -1.75%
CSX
NSC -2.27%
TT
JCI
CPRT -1.48%
GPN
INTEGRATED FREI
WASTE
AIRLI
VRSK -0.53%
EFX
UPS -1.56%
FDX -1.89%
WM -0.94
RSG
DAL
LUV
ENGIN
J
INDUST
TRUCK
URI
SNA
REAL ESTATE
REIT - SPECIALTY
EQIX -2.04%
DLR
AMT -1.70
WY
CCI
SBAC
REIT - INDU
REIT -
PLD -1.54%
PSA
ESS
EXR
CPT
ENERGY
OIL & GAS INTEGRATED
OIL & GA
XOM +0.15%
CVX -0.54%
MPC -2.06
PSX -0.70
VLO
OIL & GA
SLB
OIL & GAS E&P
HAL -0.46
BKR -0.53
O
WELL -2.52%
CSGP -2.33
CBRE
COP -0.77%
EOG +0.98%
PXD +1.03
OIL & GA
SPG
VTR
OXY -0.02%
HES +0.07
EQT
WMB
VICI
KMI
UTILITIES
UTILITIES - REGULATED ELE
BASIC MATERIALS
SPECIALTY CHE
NEE -3.16%
PCG -2.81
AEP -2.97
EXC -2.85
SRE -1.40
LIN
AGRICUL
CTVA
D
XEL -2.99
ED -2.66
CEG -3.31
NUE -2.26
FCX -1.69
SO -3.15%
ES
DTE
AEE
SHW -0.91%
ECL -1.86
DD
STLD
PEG -3.02
WEC -3.60
FE
PPL
LYB
ALB
BUILDI
DUK -2.35%
EIX
ETR
LNT
AWK
APD -1.00%
PPG
IFF
CHEMI
https://finviz.com/map.ashx
-3%
-2%
-1%
0%
+1%
+2%
+3%

62

# ENCODINGS

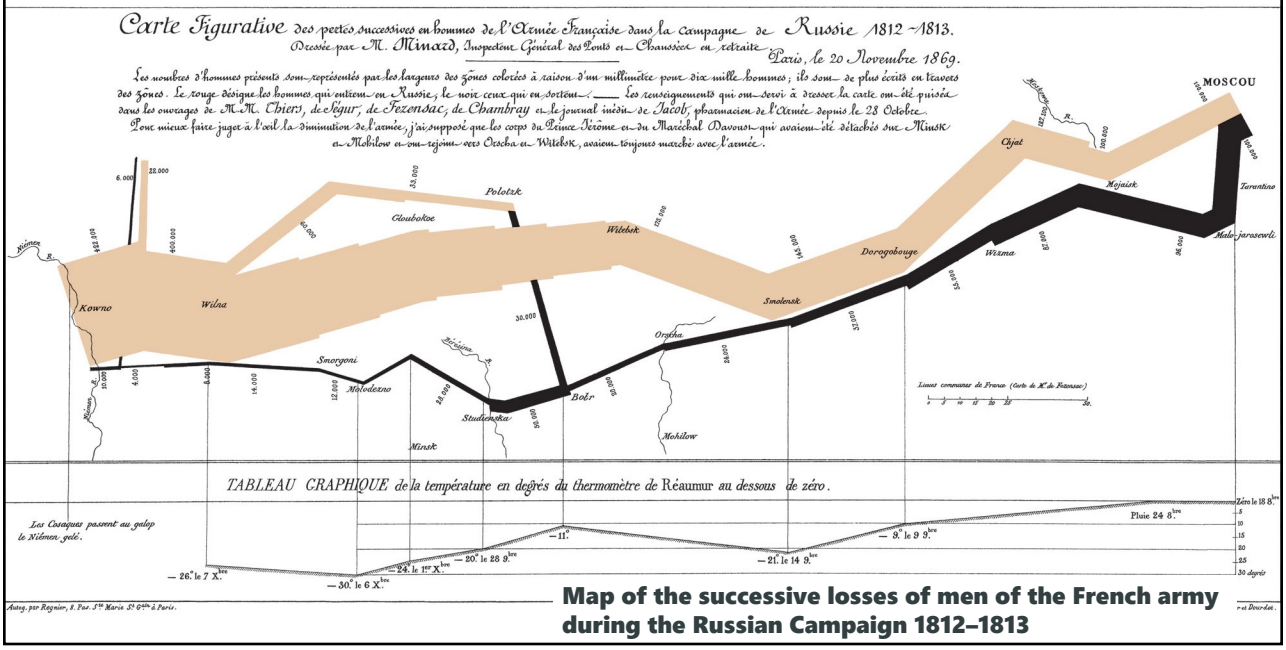
market cap (Q) → rectangle size
mkt sector (N), mkt cap (Q) → rect. pos.
loss vs. gain (N, O) → color hue
magnitude of loss or gain (Q) → color value




https://finviz.com/map.ashx

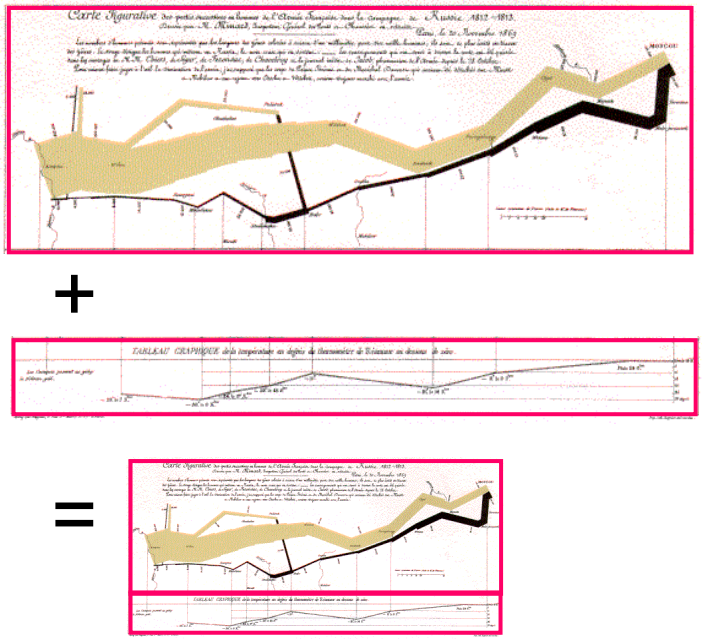
63

# MINARD's MARCH on MOSCOW




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

64




65

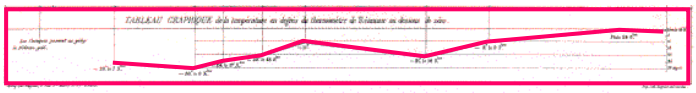
# MARK COMPOSITION

temperature (Q) → y-position

+

longitude (Q), time (O) → x-position

=




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

66

# MARK COMPOSITION

latitude (Q) → y-position

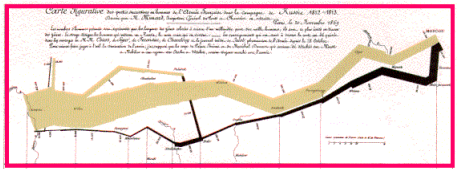
\+

longitude (Q) → x-position

\+

army size (Q) → width

=



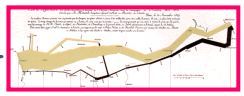
army position (Q x Q) and army size (Q)

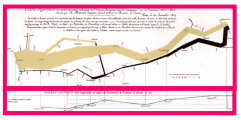
67

latitude (Q)

longitude (Q)

army size (Q)



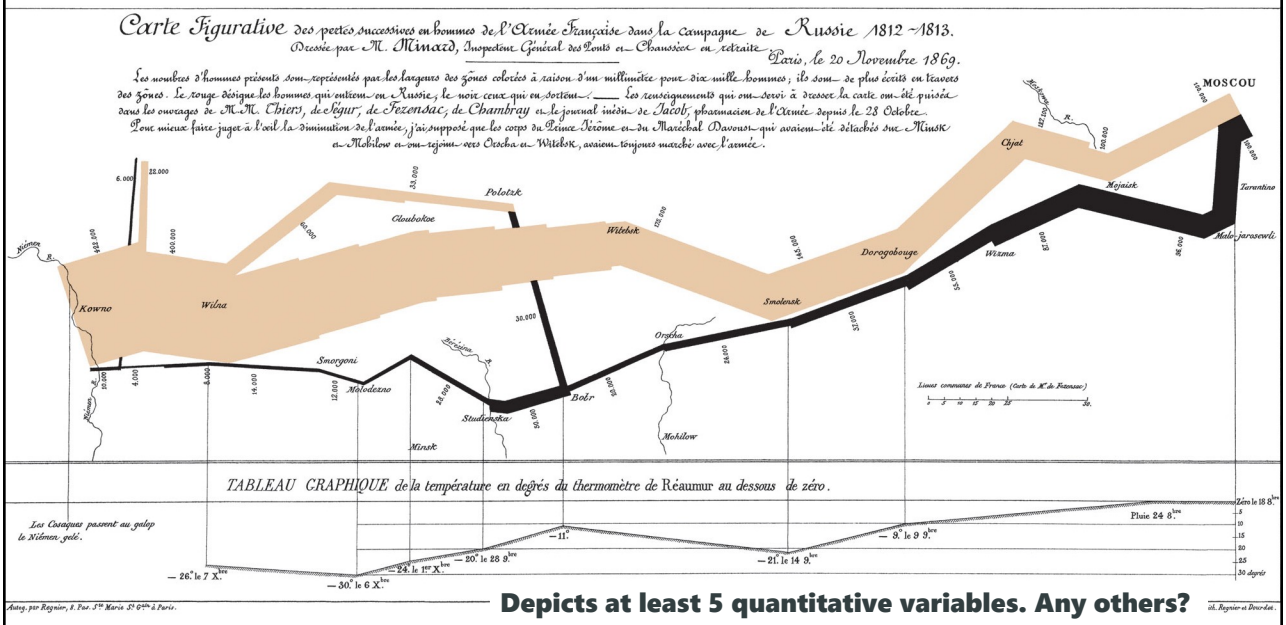


temperature (Q)

longitude (Q), time(O)

68

# MINARD's MARCH on MOSCOW




69

# FORMALIZING DESIGN

70

# COMBINATORICS OF ENCODINGS

## Challenge:

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

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

71

# 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

72

# EXPRESSIVENESS CRITERIA [Mackinlay 1986]

## Expressiveness

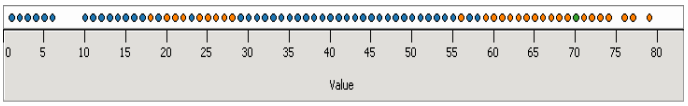
A set of facts is expressible in a visual language if the sentences (i.e., the visualizations) in the language express ***all*** the facts in the set of data, and ***only*** the facts in the data.

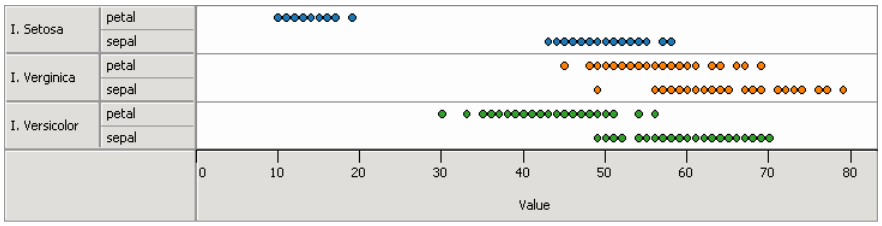
73

# 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




74

# EXPRESSES FACTS NOT IN THE DATA

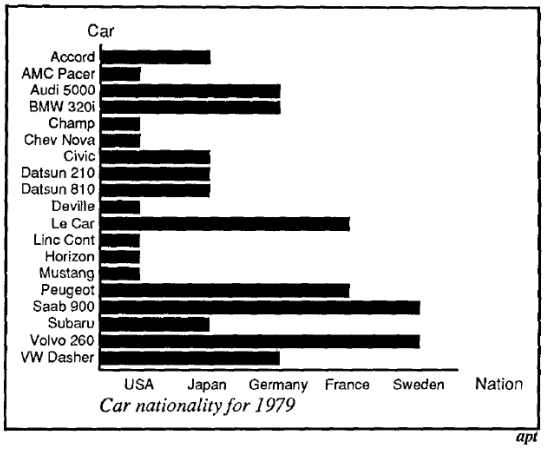



Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

Length is interpreted as encoding a quantitative value

75

# EFFECTIVENESS CRITERIA [Mackinlay 1986]

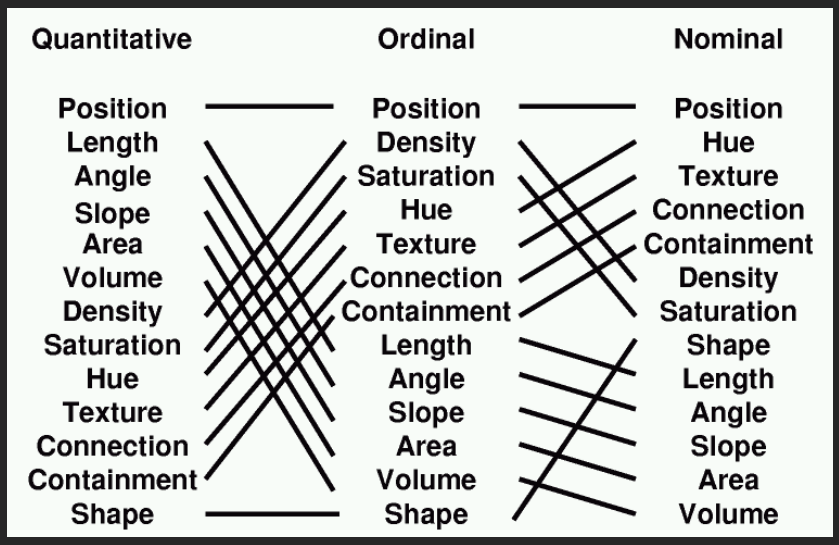
## Effectiveness

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

76

# MACKINLAY'S RANKING




Conjectured *effectiveness* of encodings by data type

77

# 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**

78

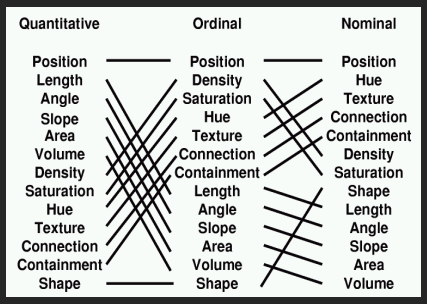

# APT [Mackinlay 1986]

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)




→

**mark: lines**

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

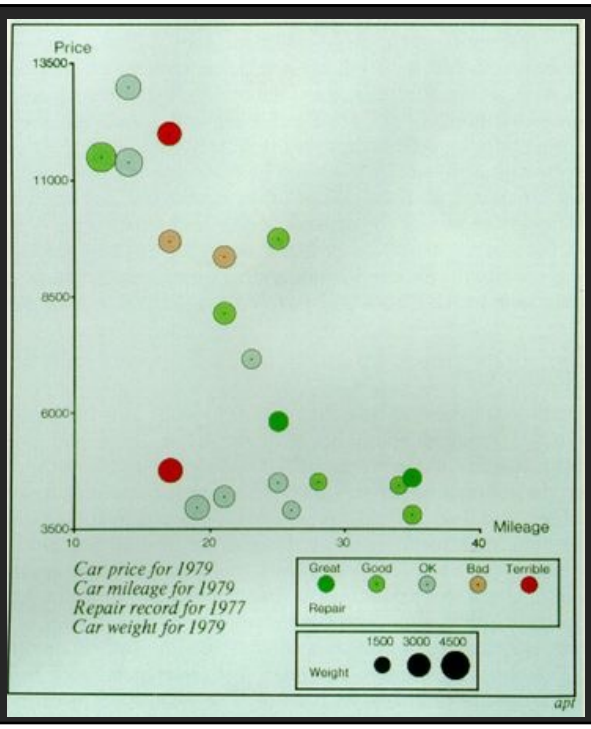
80

# APT [Mackinlay 1986]

Automatically generated chart for cars data

**Cars Data**

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




81

# 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**

82

# 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

83