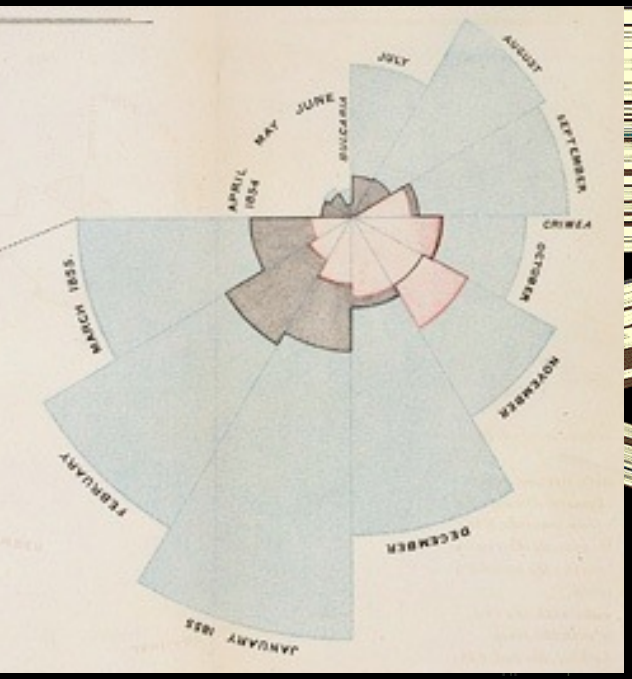
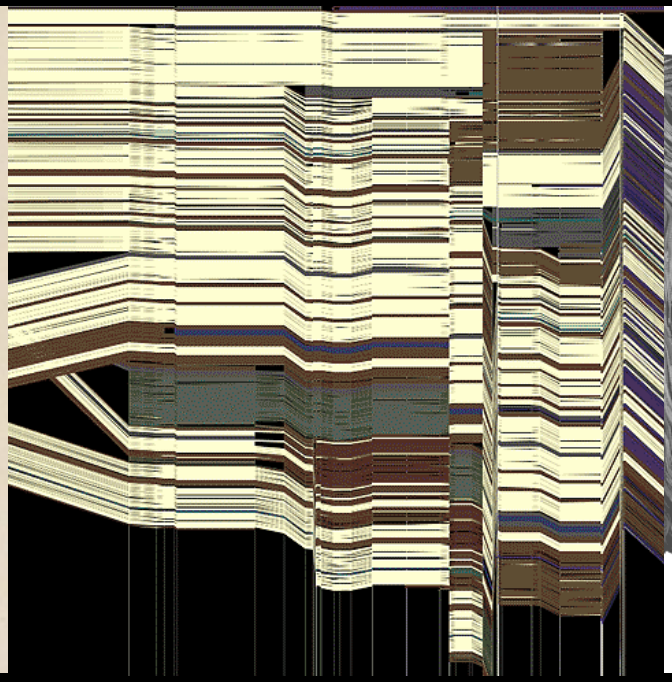
CSE 412 - Intro to Data Visualization

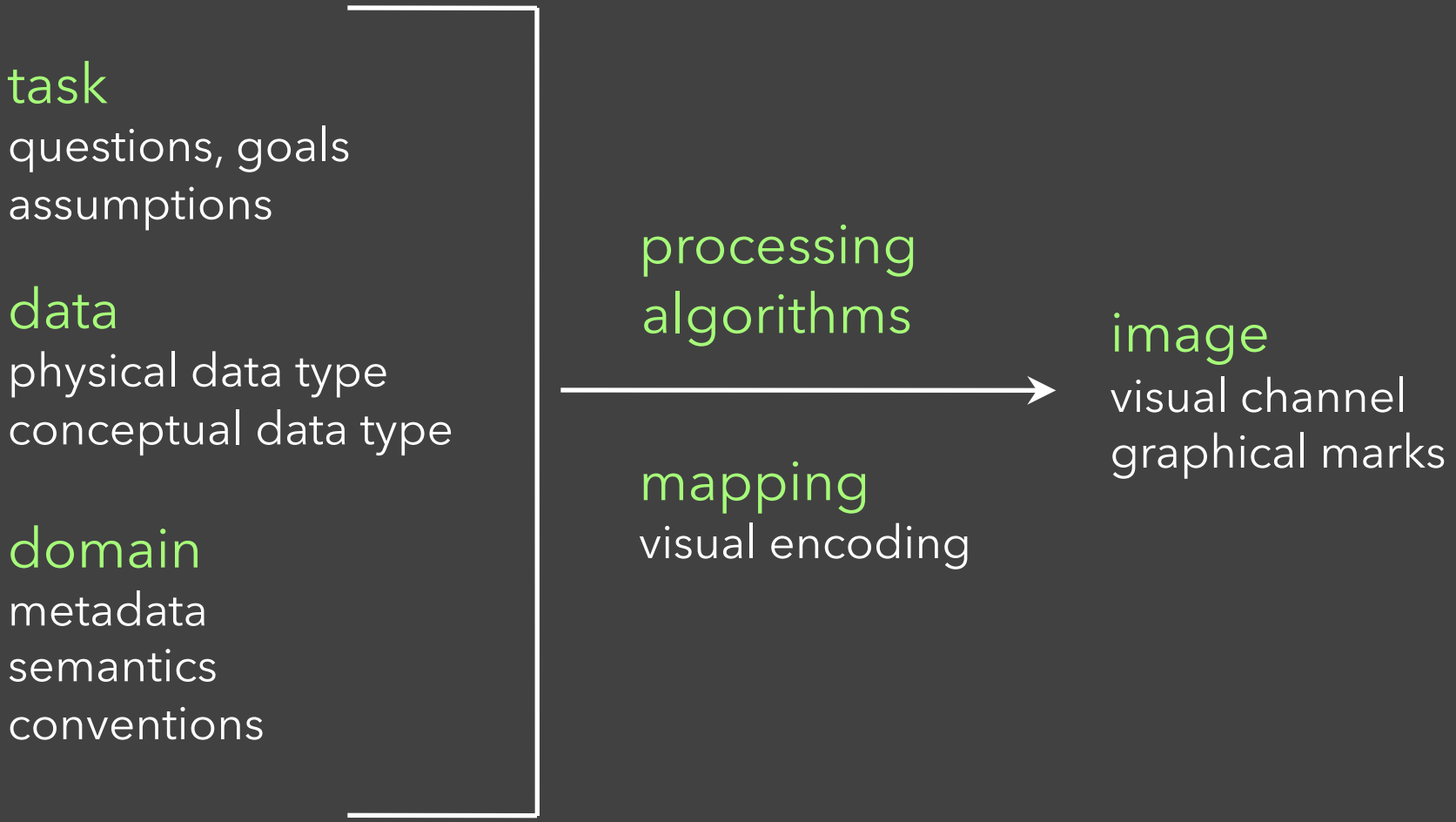
Data Models





Jane Hoffswell University of Washington

The Big Picture



Topics

Today Properties of Data

Friday Properties of Images

Friday Mapping Data to Images

Data Models / Conceptual Models

Data models are formal descriptions
Math: sets with operations on them
Example: integers with + and x operators

Conceptual models are mental constructions
Include semantics and support reasoning

Examples (data vs. conceptual)
1D floats vs. temperatures
3D vector of floats vs. spatial location

Taxonomy of Data Types (?)

1D (sets and sequences)
Temporal
2D (maps)
3D (shapes)
nD (relational)
Trees (hierarchies)
Networks (graphs)

Are there others?

The eyes have it: A task by data type taxonomy for information visualization [Shneiderman 96]

Nominal, Ordinal & Quantitative

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, …

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, …

O - Ordered

- Quality of meat: Grade A, AA, AAA

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, …

O - Ordered

- Quality of meat: Grade A, AA, AAA

Q - Interval (location of zero arbitrary)

- Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)
- Only differences (i.e., intervals) may be compared

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Fruits: apples, oranges, …

O - Ordered

- Quality of meat: Grade A, AA, AAA

Q - Interval (location of zero arbitrary)

- Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)
- Only differences (i.e., intervals) may be compared

Q - Ratio (zero fixed)

- Physical measurement: Length, Mass, Time duration, …
- Counts and amounts

Nominal, Ordinal & Quantitative

N - Nominal (labels or categories)

- Operations: =, ≠

O - Ordered

- Operations: =, ≠, <, >

Q - Interval (location of zero arbitrary)

- Operations: =, ≠, <, >, -
- Can measure distances or spans

Q - Ratio (zero fixed)

- Operations: =, ≠, <, >, -, %
- Can measure ratios or proportions

From Data Model to N, O, Q

Data Model

32.5, 54.0, -17.3, …

Floating point numbers

Conceptual Model

Temperature (°C)

Data Type

Burned vs. Not-Burned (N)

Hot, Warm, Cold (O)

Temperature Value (Q-interval)

Dimensions & Measures

Dimensions (~ independent variables)
Often discrete variables describing data (N, O)
Categories, dates, binned quantities

Measures (~ dependent variables)
Data values that can be aggregated (Q)
Numbers to be analyzed
Aggregate as sum, count, avg, std. dev…

Not a strict distinction. The same variable may be treated either way depending on the task.

Example: U.S. Census Data

Example: U.S. Census Data

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

Example: U.S. Census

People Count
Year
Age
Sex
Marital Status

2,348 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	174976
27	1850	60	0	2	162236
28	1850	65	0	1	106827
29	1850	65	0	2	105534
30	1850	70	0	1	73677
31	1850	70	0	2	71762
32	1850	75	0	1	40834
33	1850	75	0	2	40229
34	1850	80	0	1	23449
35	1850	80	0	2	22949
36	1850	85	0	1	8186
37	1850	85	0	2	10511
38	1850	90	0	1	5259
39	1850	90	0	2	6569
40	1860	0	0	1	2120846
41	1860	0	0	2	2092162

Census: N, O, Q-Interval, Q-Ratio?

People Count	Q-Ratio
Year	Q-Interval *(O)*
Age	Q-Ratio *(O)*
Sex	N
Marital Status	N

Census: Dimension or Measure?

People Count	Measure
Year	Dimension
Age	Depends!
Sex	Dimension
Marital Status	Dimension

Census Data Demo

Administrivia

A1: Visualization Design

Design a static visualization for a data set.

The climate of a place can have a tremendous impact on people's lived experience. You will examine average monthly climate measurements for six major U.S. cities, roughly covering the edges of the continental United States.

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

A1: Visualization Design

Pick a **guiding question**, use it to title your vis.
Design a **static visualization** for that question.
You are free to **use any tools** (inc. pen & paper).

Deliverables (upload via Canvas; see A1 page)
Image of your visualization (PNG or JPG format)
Short description + design rationale (≤ 4 paragraphs)

Due by **11:59 pm PT, next Monday April 5th**.

Course Participation

Quiz & discussion comments on class forum (Ed).

Both are due each Monday, by 11:59pm up through week 8 of the quarter.

You have 1 “pass” (quiz + comment) for the quarter.

First discussion and quiz are now posted on Ed
Due by **11:59 pm PT, next Monday April 5th**.

Data Tables & Transformations

Relational Data Model

Represent data as a **table** (or *relation*)

Each **row** (or *tuple*) represents a record
    Each record is a fixed-length tuple

Each **column** (or *field*) represents a variable
    Each field has a *name* and a *data type*

A table's **schema** is the set of names and types

A **database** is a collection of tables (relations)

Relational Algebra [Codd '70] / SQL

Operations on Data Tables: table(s) in, table out

Relational Algebra [Codd '70] / SQL

Operations on Data Tables: table(s) in, table out

Project (`select`): select a set of columns

Filter (`where`): remove unwanted rows

Sort (`order by`): order records

Aggregate (`group by`, `sum`, `min`, `max`, ...):

partition rows into groups + summarize

Combine (`join`, `union`, ...):

integrate data from multiple tables

Relational Algebra [Codd '70] / SQL

Project (`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

Relational Algebra [Codd '70] / SQL

Filter (`where`): remove unwanted 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

Relational Algebra [Codd '70] / SQL

Sort (`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

Relational Algebra [Codd '70] / SQL

Aggregate (`group by`, `sum`, `min`, `max`, ...):

```
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

Relational Algebra [Codd '70] / SQL

Join (`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
```

Roll-Up and Drill-Down

Want to examine population by year and age?

Roll-up the data along the desired dimensions

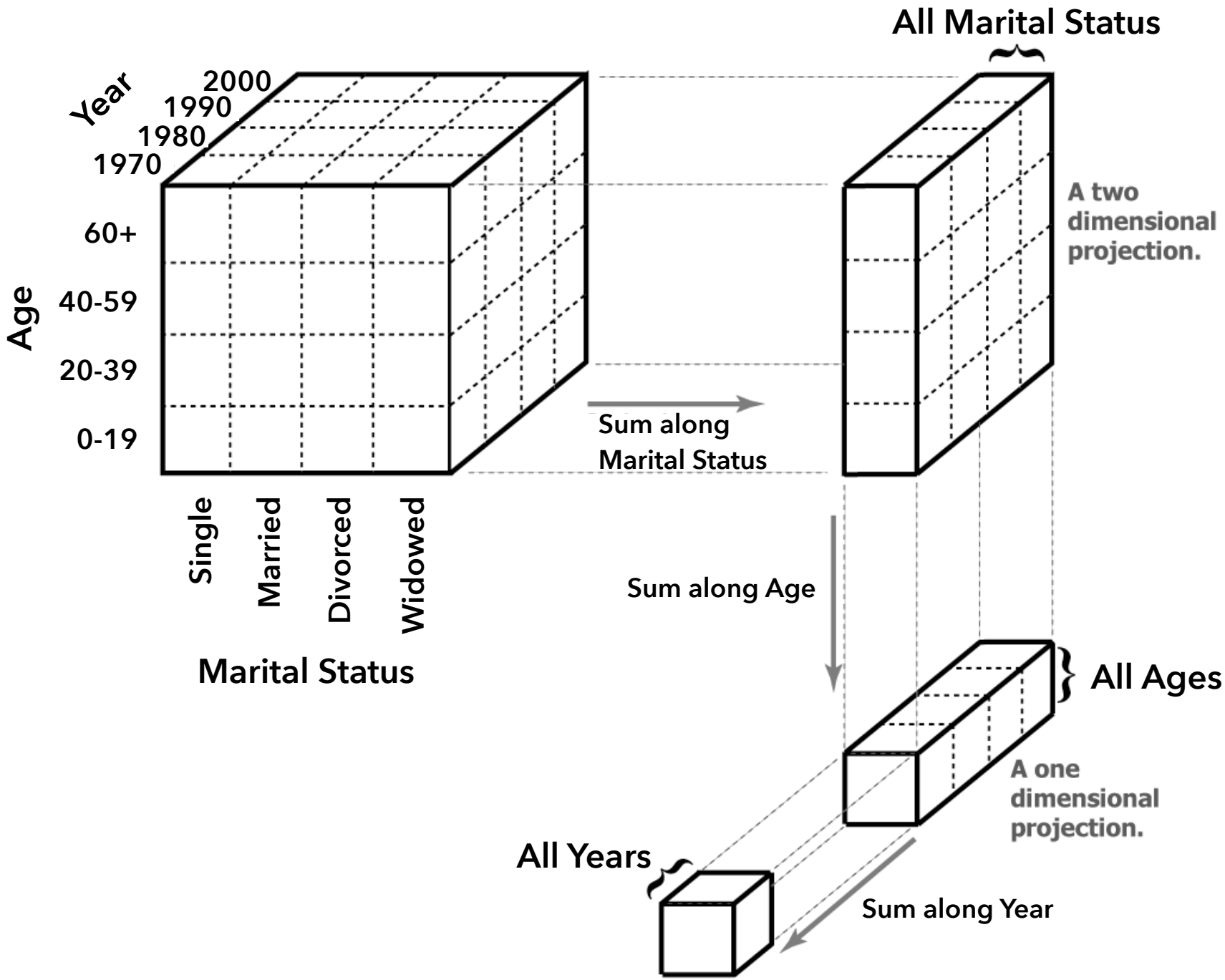

```
SELECT year, age, sum(people)
FROM census
GROUP BY year, age
```

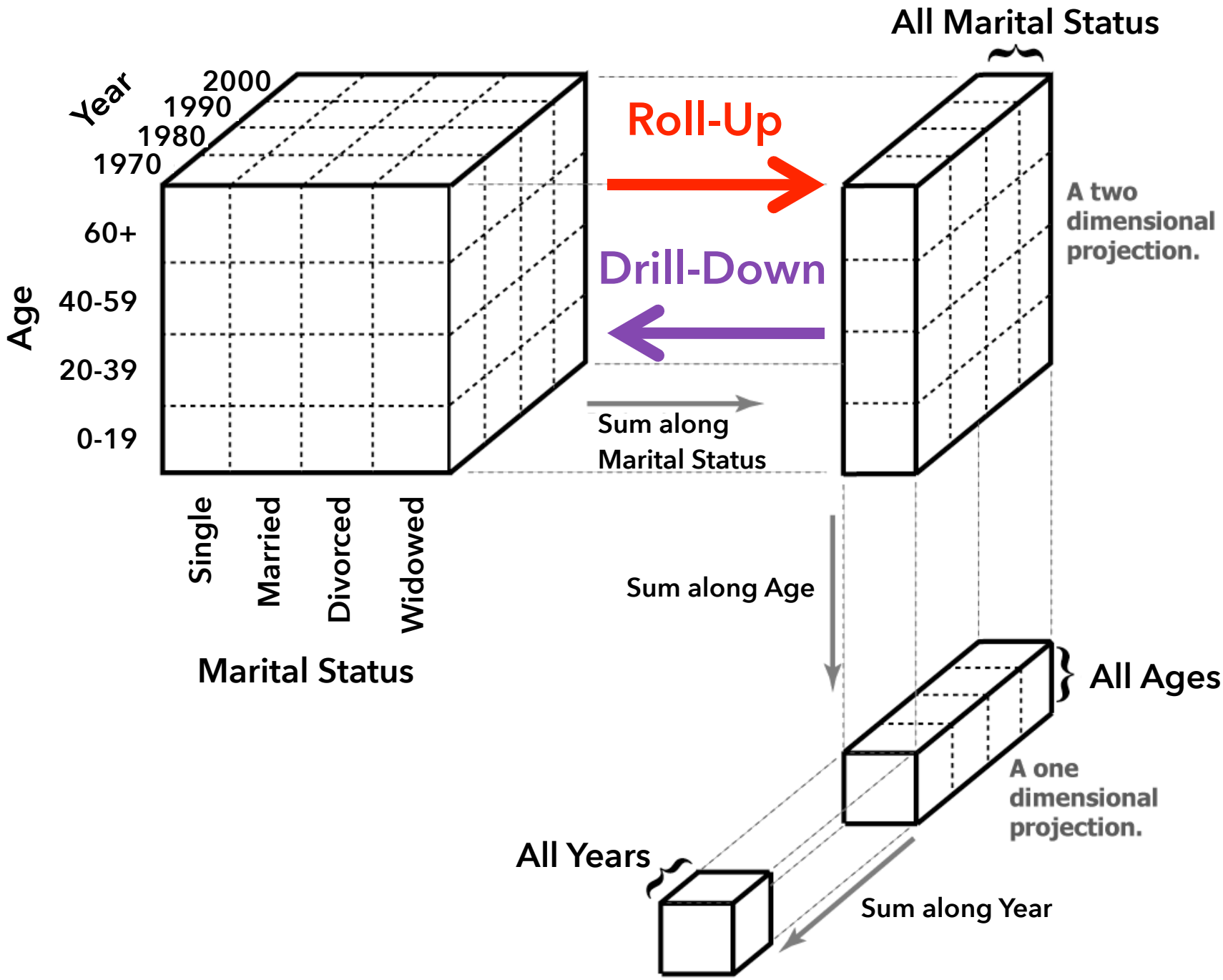
Roll-Up and Drill-Down

Want to see the breakdown by marital status?

Drill-down into additional dimensions

```
SELECT year, age, marst, sum(people)
FROM census
GROUP BY year, age, marst
```


All Marital Status
Year
2000
1990
1980
1970
Age
60+
40-59
20-39
0-19
Single
Married
Divorced
Widowed
Marital Status
A two dimensional projection.
Sum along Marital Status
Sum along Age
All Ages
A one dimensional projection.
All Years
Sum along Year


All Marital Status
Year
2000
1990
1980
1970
Roll-Up
Drill-Down
A two dimensional projection.
Age
60+
40-59
20-39
0-19
Sum along Marital Status
Single
Married
Divorced
Widowed
Marital Status
Sum along Age
All Ages
A one dimensional projection.
All Years
Sum along Year

ORIGINAL

YEAR	AGE	MARST	SEX	PEOPLE
1850	0	0	1	1,483,789
1850	5	0	1	1,411,067
1860	0	0	1	2,120,846
1860	5	0	1	1,804,467
. . .				

PIVOTED (or CROSS-TABULATION)

AGE	MARST	SEX	1850	1860	. . .
0	0	1	1,483,789	2,120,846	. . .
5	0	1	1,411,067	1,804,467	. . .
. . .					

Which format might we prefer? Why?

Tidy Data [Wickham 2014]

How do rows, columns, and tables match up with observations, variables, and types? In “tidy” data:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

The advantage is that this provides a flexible starting point for analysis, transformation, and visualization.

Our pivoted table variant was not “tidy”!

(This is a variant of normalized forms in DB theory)

Common Data Formats

CSV: Comma-Separated Values (d3.csv)

```
year,age,marst,sex,people
1850,0,0,1,1483789
1850,5,0,1,1411067
...
```

Common Data Formats

CSV: Comma-Separated Values (d3.csv)

```
year,age,marst,sex,people
1850,0,0,1,1483789
1850,5,0,1,1411067
...
```

JSON: JavaScript Object Notation (d3.json)

```
[
 {"year":1850,"age":0,"marst":0,"sex":1,"people":1483789},
 {"year":1850,"age":5,"marst":0,"sex":1,"people":1411067},
 ...
]
```

Required Readings for Fri 4/2

Chapter 3. Technology Fundamentals

Solid familiarity with the following concepts will make your time with D3 a lot less frustrating and a lot more rewarding. Consider this a brief refresher course on Web-Making 101.

WARNING

Beware! This is a pretty dense chapter, packed with years' worth of web development knowledge, and nothing in here is specific to D3. I recommend skimming just the information that is new to you, and skipping the rest. You can always reference this chapter later as questions arise.

Chapter 3 in Interactive Data Visualization for the Web, 2nd Edition. Scott Murray.

Quiz Section: Data Wrangling

First quiz section will be tomorrow, Thur 4/1

Introduction to Observable

Discussion of data formats and transformation

Up Next: Jane's Office Hour (link on Canvas)