Data Models and Representations

Sungahn Ko

HAiV

0

Outline

HAiV

- Foundational topics for course
- Data types
- Data representations
- Data models
- Tables

1

1

Data

HAiV

- So far we have looked at many examples of visualization
- Ignored the fundamentals
- It all starts with data

- Good definition?
  - Collected from the world
  - Represents the world (somehow)

2

2

Data

HAiV

- **Data is just an abstraction of a real phenomenon**

- **Corollaries:**
  - Visualization is only as good as the data
  - Visualizations can be misleading
  - Good data is important!

3

3

Data and Datasets

- **Data is everywhere!**
- **Almost all of it is unstructured (95%)**
  - Images
  - Video
  - Sound
  - Log files
  - Text
  - Web pages

Data and Datasets

- Need regular and structured datasets to analyze and visualize this data

- Often we must do this ourselves!

Existing Structured Data

- **Resources exist that collect data on the Web**
- **Data.gov**
  - US Federal government dataset collection
- **UCB Library:**
  - https://guides.lib.berkeley.edu/c.php?g=1257448&p=9237051
- **UCI ML Data**
  - https://archive.ics.uci.edu/datasets
- **Public Data Portal Korea:** https://www.data.go.kr/index.do
- **AI data in Korea:** **https://aihub.or.kr/aihubdata/data/list.do**

6

6

Deriving Structured Data: Wrangler (CHI 2011)

- **http://vis.stanford.edu/wrangler/**
- **Demo!**
  - https://vimeo.com/19185801

7

7

Data Models

- **How to capture and structure our data?**
- **Often use three types of entities:**
  - Attributes
    - Characteristics of objects and relations
    - Property of an entity
    - **Example**: age, gender, color of object

8

8

Relational Data Model

- Records in a data table
- Structured from amenable to analysis and visualization
- Fixed-length tuples (attributes)
- Each column (attribute) has a domain (type)
- Relational databases also allow relations between cases (often through related tables) – not our focus today

9

9

Relational Algebra

- Manipulating relation data models
- Formalized in the standardized SQL language
  - Standard Query Language
- Selection (SELECT)
- Projection (WHERE)
- Sorting (ORDER BY)
- Aggregation (GROUP BY, SUM, MIN, …)

10

10

Relational Algebra

- **Set operations (UNION, INTERSECT…)**
- **Join (INNER JOIN, …)**

11

11

Example: Relational Data

Students

ID#	Name	Phone	DOB
500	Matt	555-4141	06/03/70
501	Jenny	867-5309	3/15/81
502	Sean	876-9123	10/31/82

ID#	ClassID	Sem
500	1001	Fall02
501	1002	Fall02
501	1002	Spr03
502	1003	S203

Takes_Course

ClassID	Title	ClassNum
1001	Intro to Informatics	I101
1002	Data Mining	I400
1003	Internet and Society	I400

Courses

Variable Types

- **Nominal data (labels) *[Category data]***
  - Supports only equality (same or different)
  - Examples: gender, car brand, fruit, bus number
- **Ordinal data (ordered) *[Integer data]***
  - Obeys the < relation, ordered set
  - Examples: days of week, Fresh/Soph./Junior/Senior

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

Variable Types

- **Quantitative data *[Real-number data]***
  - Supports arithmetic operations
  - Interval (zero arbitrary)
    - Example: Dates, location
  - Ratio (zero fixed)
    - Example: age, temperature, stock value

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

Mathematical Operations

- **N - Nominal (labels)**
  - Operations: =, ≠
- **O – Ordered**
  - Operations: =, ≠, <, >
- **Q - Interval (Location of zero arbitrary)**
  - Operations: =, ≠, <, >, -
  - Can measure distances or spans

Mathematical Operations

- **Q - Ratio (zero fixed)**
  - Operations: =, ≠, <, >, -, ÷
  - Can measure ratios or proportions

Metadata

- **Data about data (derived data)**
- **Describes:**
  - Definition
  - Structure
  - Administration

Metadata

HAiV

- **Examples:**
  - Types of variables in data table
  - Language of a particular text
  - Dimensions, bit depth, timestamp for a photograph
- **Metadata is often useful when treating data, and sometimes also for visualization!**

18

Data Dimensions

HAiV

- **Common dimensions: 1, 2, 3**
  - 1 dimension – univariate
    - Temperature readings
  - 2 dimensions – bivariate
    - Positions on map (lat/long)
  - 3 dimensions – trivariate
    - Positions in space (3D)
- **For more than 3 dimensions**
  - Multivariate
  - Hypervariate

19

Example: US Census Data

HAiV

- People: # of people in group
- 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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
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

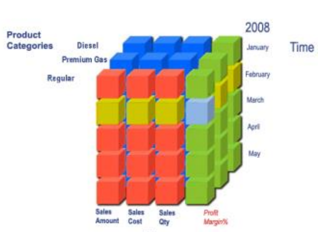
20

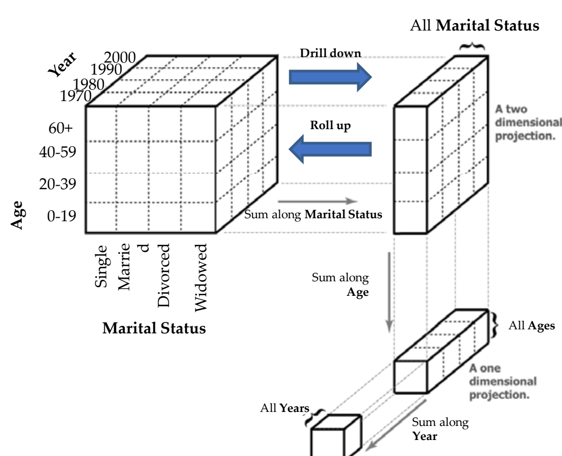
20

OLAP Cube – US Census Data

HAiV

- **OLAP – Online Analytical Processing**
  - Manipulation
  - Analysis
- **…of data from multiple perspectives**





[J. Heer]

21

21

How to Represent Tabular Data?

- **Standard answer in this course: graphs!**
  - Statistical data graphics
  - Bar charts, line charts, pie charts, etc.
- **There is a simpler way: tables**
  - Also a graphical representation!
  - Textual representation
  - Useful for direct lookups

22

22

How to Represent Tabular Data?

- **When to use which format?**
  - Tables: looking up individual values, precise data
  - Graphs: relationships, comparisons

23

23

Side Note

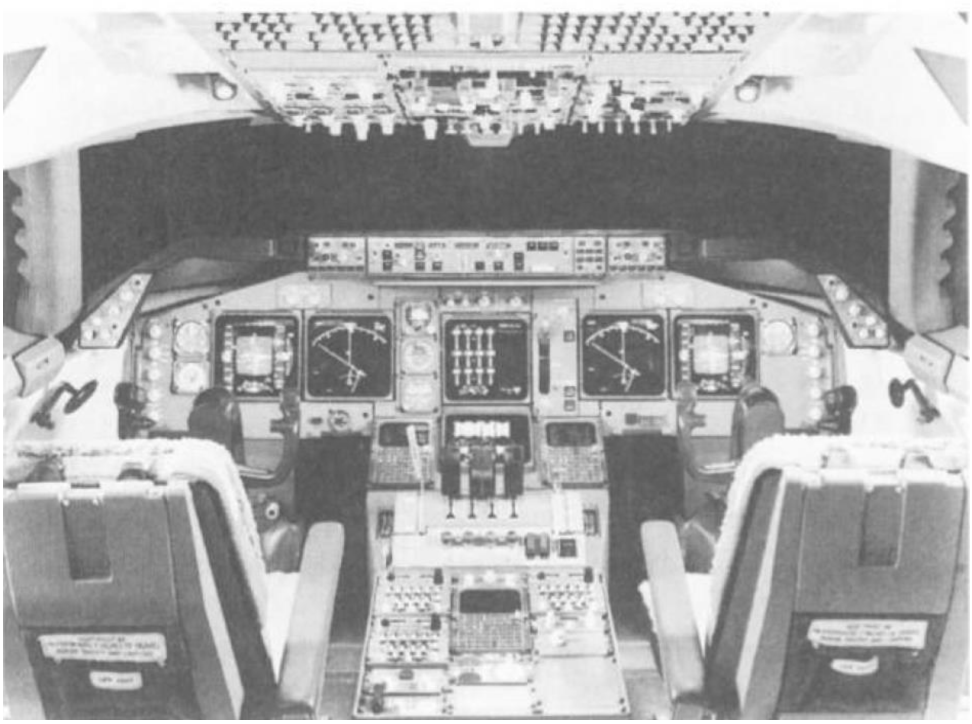
HAiV

A: Which number is larger?
284 912

B: Which number is larger?
284 312

People answer A faster than B. Why?

"The form of representation most appropriate for an artifact depends on the task to be performed" – D. A. Norman, 1993



Example: Tables and Graphs

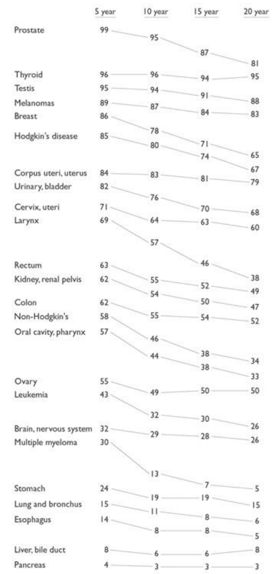
	Relative survival rate, % (SE)			
	5 years	10 years	15 years	20 years
Cancer site				
Oral cavity and pharynx	56·7 (1·3)	44·2 (1·4)	37·5 (1·6)	33·0 (1·8)
Oesophagus	14·2 (1·4)	7·9 (1·3)	7·7 (1·6)	5·4 (2·0)
Stomach	23·8 (1·3)	19·4 (1·4)	19·0 (1·7)	14·9 (1·9)
Colon	61·7 (0·8)	55·4 (1·0)	53·9 (1·2)	52·3 (1·6)
Rectum	62·6 (1·2)	55·2 (1·4)	51·8 (1·8)	49·2 (2·3)
Liver and intrahepatic bile duct	7·5 (1·1)	5·8 (1·2)	6·3 (1·5)	7·6 (2·0)
Pancreas	4·0 (0·5)	3·0 (0·5)	2·7 (0·6)	2·7 (0·8)
Larynx	68·8 (2·1)	56·7 (2·5)	45·8 (2·8)	37·8 (3·1)
Lung and bronchus	15·0 (0·4)	10·6 (0·4)	8·1 (0·4)	6·5 (0·4)
Melanomas	89·0 (0·8)	86·7 (1·1)	83·5 (1·5)	82·8 (1·9)
Breast	86·4 (0·4)	78·3 (0·6)	71·3 (0·7)	65·0 (1·0)
Cervix uteri	70·5 (1·6)	64·1 (1·8)	62·8 (2·1)	60·0 (2·4)
Corpus uteri and uterus, NOS	84·3 (1·0)	83·2 (1·3)	80·8 (1·7)	79·2 (2·0)
Ovary	55·0 (1·3)	49·3 (1·6)	49·9 (1·9)	49·6 (2·4)
Prostate	98·8 (0·4)	95·2 (0·9)	87·1 (1·7)	81·1 (3·0)
Testis	94·7 (1·1)	94·0 (1·3)	91·1 (1·8)	88·2 (2·3)
Urinary bladder	82·1 (1·0)	76·2 (1·4)	70·3 (1·9)	67·9 (2·4)
Kidney and renal pelvis	61·8 (1·3)	54·4 (1·6)	49·8 (2·0)	47·3 (2·6)
Brain and other nervous system	32·0 (1·4)	29·2 (1·5)	27·6 (1·6)	26·1 (1·9)
Thyroid	96·0 (0·8)	95·8 (1·2)	94·0 (1·6)	95·4 (2·1)
Hodgkin's disease	85·1 (1·7)	79·8 (2·0)	73·8 (2·4)	67·1 (2·8)
Non-Hodgkin lymphomas	57·8 (1·0)	46·3 (1·2)	38·3 (1·4)	34·3 (1·7)
Multiple myeloma	29·5 (1·6)	12·7 (1·5)	7·0 (1·3)	4·8 (1·5)
Leukaemias	42·5 (1·2)	32·4 (1·3)	29·7 (1·5)	26·2 (1·7)

Rates derived from SEER 1973–98 database (both sexes, all ethnic groups).[12]
NOS=not otherwise specified.

Table 4: **Most recent period estimates of relative survival rates, by cancer site**

Estimates of relative survival rates, by cancer site

	% survival rates and standard errors							
	5 year		10 year		15 year		20 year	
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
Melanomas	89.0	0.8	86.7	1.1	83.5	1.5	82.8	1.9
Breast	86.4	0.4	78.3	0.6	71.3	0.7	65.0	1.0
Hodgkin's disease	85.1	1.7	79.8	2.0	73.8	2.4	67.1	2.8
Corpus uteri, uterus	84.3	1.0	83.2	1.3	80.8	1.7	79.2	2.0
Urinary, bladder	82.1	1.0	76.2	1.4	70.3	1.9	67.9	2.4
Cervix, uteri	70.5	1.6	64.1	1.8	62.8	2.1	60.0	2.4
Larynx	68.8	2.1	56.7	2.5	45.8	2.8	37.8	3.1
Rectum	62.6	1.2	55.2	1.4	51.8	1.8	49.2	2.3
Kidney, renal pelvis	61.8	1.3	54.4	1.6	49.8	2.0	47.3	2.6
Colon	61.7	0.8	55.4	1.0	53.9	1.2	52.3	1.6
Non-Hodgkin's	57.8	1.0	46.3	1.2	38.3	1.4	34.3	1.7
Oral cavity, pharynx	56.7	1.3	44.2	1.4	37.5	1.6	33.0	1.8
Ovary	55.0	1.3	49.3	1.6	49.9	1.9	49.6	2.4
Leukemia	42.5	1.2	32.4	1.3	29.7	1.5	26.2	1.7
Brain, nervous system	32.0	1.4	29.2	1.5	27.6	1.6	26.1	1.9
Multiple myeloma	29.5	1.6	12.7	1.5	7.0	1.3	4.8	1.5
Stomach	23.8	1.3	19.4	1.4	19.0	1.7	14.9	1.9
Lung and bronchus	15.0	0.4	10.6	0.4	8.1	0.4	6.5	0.4
Esophagus	14.2	1.4	7.9	1.3	7.7	1.6	5.4	2.0
Liver, bile duct	7.5	1.1	5.8	1.2	6.3	1.5	7.6	2.0
Pancreas	4.0	0.5	3.0	1.5	2.7	0.6	2.7	0.8



"For […] small data sets, usually a simple table s hows the data more effectively than a graph, let al one a chartjunk graph." – E. R. Tufte, 2003

26

26

Questions?

28