Data

C. Andrews

2014-02-18

Data models and Conceptual models

Data model		Conceptual model
2.3	float	Temperature Olympic score
“Iris”	String	Body part Flower part Proper name
(5,7)	tuple / vector	location in space dimensions
syntax		**semantics**

Types of data

Interval

a distance can be computed between two values

Ratio

there is a fixed origin, or absolute smallest value on the scale

Nominal

labels without ordering (e.g., genders, types of fruit)

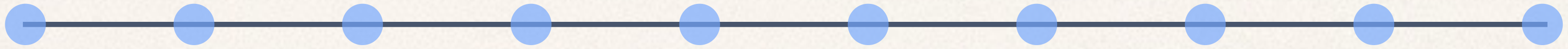
Ordinal

a well ordered set of values (e.g., grades, military rank)

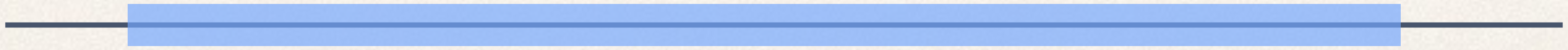
Qualitative measures

Binary (0 or 1, True or False)

Discrete (e.g., number of eyes, number of birds in a flock)



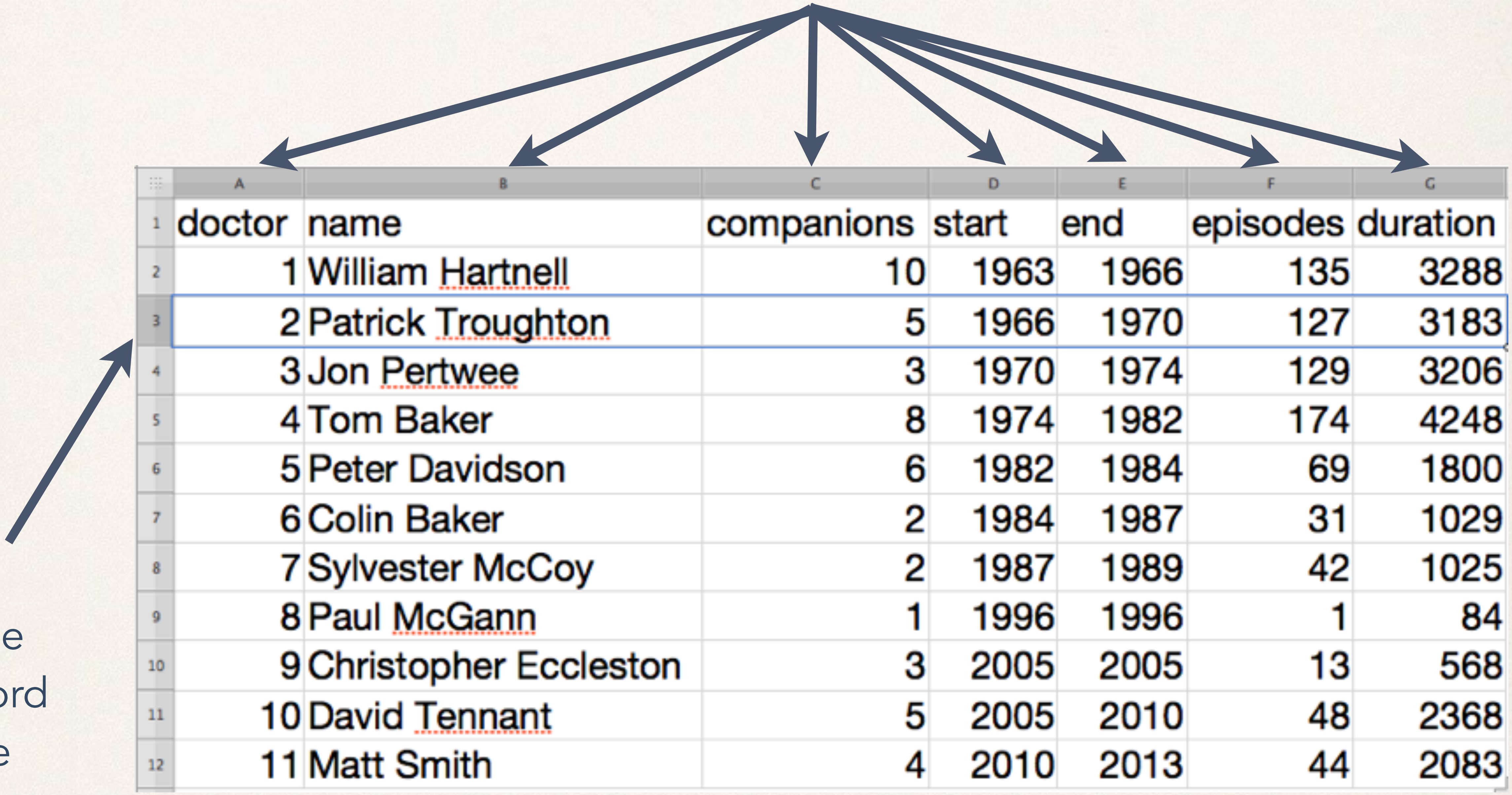
Continuous (e.g., temperature, length)



Relational data model

Attributes

columns, variables, fields



	A	B	C	D	E	F	G
1	doctor	name	companions	start	end	episodes	duration
2	1	William Hartnell	10	1963	1966	135	3288
3	2	Patrick Troughton	5	1966	1970	127	3183
4	3	Jon Pertwee	3	1970	1974	129	3206
5	4	Tom Baker	8	1974	1982	174	4248
6	5	Peter Davidson	6	1982	1984	69	1800
7	6	Colin Baker	2	1984	1987	31	1029
8	7	Sylvester McCoy	2	1987	1989	42	1025
9	8	Paul McGann	1	1996	1996	1	84
10	9	Christopher Eccleston	3	2005	2005	13	568
11	10	David Tennant	5	2005	2010	48	2368
12	11	Matt Smith	4	2010	2013	44	2083

Item

row

tuple

record

case

Variable types

Dependent variables

measurements of the data

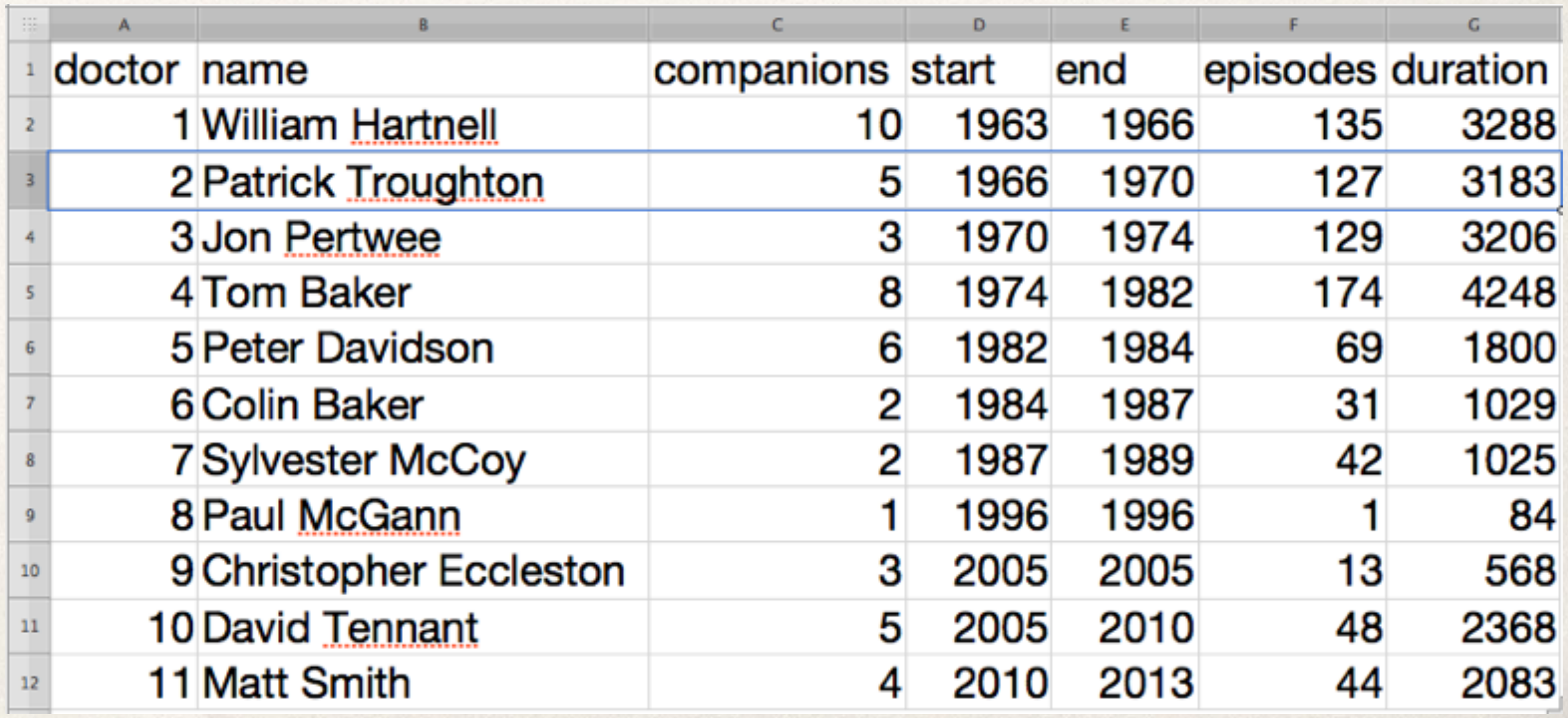
observations

can be analyzed and aggregated

Independent variables

values that describe the data

dates, categories, names



	A	B	C	D	E	F	G
1	doctor	name	companions	start	end	episodes	duration
2	1	William Hartnell	10	1963	1966	135	3288
3	2	Patrick Troughton	5	1966	1970	127	3183
4	3	Jon Pertwee	3	1970	1974	129	3206
5	4	Tom Baker	8	1974	1982	174	4248
6	5	Peter Davidson	6	1982	1984	69	1800
7	6	Colin Baker	2	1984	1987	31	1029
8	7	Sylvester McCoy	2	1987	1989	42	1025
9	8	Paul McGann	1	1996	1996	1	84
10	9	Christopher Eccleston	3	2005	2005	13	568
11	10	David Tennant	5	2005	2010	48	2368
12	11	Matt Smith	4	2010	2013	44	2083

Long vs wide data tables

Tree	age	circumference
1	118	30
1	484	58
1	664	87
1	1004	115
1	1231	120
1	1372	142
1	1582	145
2	118	33
2	484	69
2	664	111
2	1004	156
2	1231	172
2	1372	203
2	1582	203
3	118	30
3	484	51
3	664	75
3	1004	108
3	1231	115
3	1372	139
3	1582	140
4	118	32
4	484	62
4	664	112
4	1004	167

Tree	118	484	664	1004	1231	1372	1582
3	30	51	75	108	115	139	140
1	30	58	87	115	120	142	145
5	30	49	81	125	142	174	177
2	33	69	111	156	172	203	203
4	32	62	112	167	179	209	214

Data quality issues and conversions

Missing data

no measurements, sensor problem, redacted

Erroneous data

typos, bad sensors, outliers

Data type conversions

fahrenheit to celsius, address to lat/lon

Entity resolution

different values for the same thing?

Reconciliation

combining multiple data sources

Dealing with missing and bad data

Discard the bad data

have to consider if the rest of the record is worth preserving

Assign a sentinel value

an obvious outlying value so we know where the insertion is

Assign and average value

simple substitution based on statistics

Create a value based on the nearest neighbors

a slightly better guess for some value types

Compute a substitute value

more advances statistical methods such as *imputation*

Data transformations

Normalizing

convert the data to fall within a common scale (usually 0-1)

$$d_{norm} = (d_{orig} - d_{min})/d_{max} - d_{min})$$

Convert to other scales

use log or power functions to change the scale

Aggregation

summing, averaging, etc...

binning, like we do for histograms

grouping by merging categorical data

File formats

- database tables
- spreadsheets
- delimited files: CSV, TSV
- XML
- JSON
- unstructured text

```
title,author,year
“Neuromancer”,”William Gibson”,1984
Cryptonomicon, Neal Stephenson,1999
“The Atrocity Archives”,”Charles Stross,2004
```

```
<book>
   <title>“Neuromancer”</title>
   <author>”William Gibson”</author>
   <year>1984</year>
</book>
<book>
   <title>Cryptonomicon</title>
   <author>Neal Stephenson</author>
   <year>1999</year>
</book>
<book>
    <title>“The Atrocity Archives”</title>
   <author>”Charles Stross</author>
   <year>2004</year>
</book>
```

```
[{"title":"“Neuromancer”","author":"”William
Gibson”","year":1984},
{"title":"Cryptonomicon","author":" Neal
Stephenson","year":1999},
{"title":"“The Atrocity
Archives”","author":"”Charles Stross","year":
2004}]
```