

# The Shape of Data

ME314: Introduction to Data Science and Machine Learning

---

Ken Benoit

11 July 2024

# Plan today

- The nature of data
- Units of data
- Data types
- More about git
- Representing text data: Unicode
- Representing dates
- Data and datasets
  - “tidy” data
  - reshaping data
  - normalization forms
- Lab preview

# Data is Fundamental

*“You can have data without information, but you cannot have information without data.” – Daniel Keys Moran, an American computer programmer and science fiction writer.*

*“In God we trust. All others must bring data.” – W. Edwards Deming, statistician, professor, author, lecturer, and consultant.*

## Structured data: Index cards

- Origins in the 19th century, with botanist [Carl Linnaeus](#), who needed to record species that he was studying
- This was a form of *database*
  - each piece of information about a species formed a *field*
  - each species' entry in the system formed a *record*
  - the records were *indexed* using some reference system

# Heyday: Use in libraries to catalog books



a record looked like this

Bil

STC  
3056

Bigges, Walter, d. 1585?

A summarie and true discourse of Sir Frances Drake's West Indian voyage. Wherein were taken, the townes of Saint Iago, Sancto Domingo, Cartagena & Saint Augustine. Imprinted at London by Richard Field, 1589.

A<sup>2</sup>, B-G<sup>4</sup>, H<sup>2</sup>. 4to.

Begun by Captain Bigges, continued by Lieutenant Croftes. Edited by Thomas Cates.

STC number arbitrarily assigned. Apparently

(Continued

on next card)

27/14

## Dewey decimal system

- a proprietary library classification system first published in the United States by Melvil Dewey in 1876
- scheme is made up of ten classes, each divided into ten divisions, each having ten sections
- the system's notation uses Arabic numbers, with three whole numbers making up the main classes and sub-classes and decimals creating further divisions
- Example:

500 Natural sciences and mathematics

    510 Mathematics

        516 Geometry

            516.3 Analytic geometries

                516.37 Metric differential geometries

                    516.375 Finsler Geometry

## How to index?

- Problem: Could only sort the cards one way
- Re-referencing was literally a manual operation
- Contrast with the idea of electronic indexes, where assets are stored once, and many indexing and reference systems can be applied

# Modern database manager

- “Normalizes” data into relational tables, linked by “keys” (more on this on Day 3)

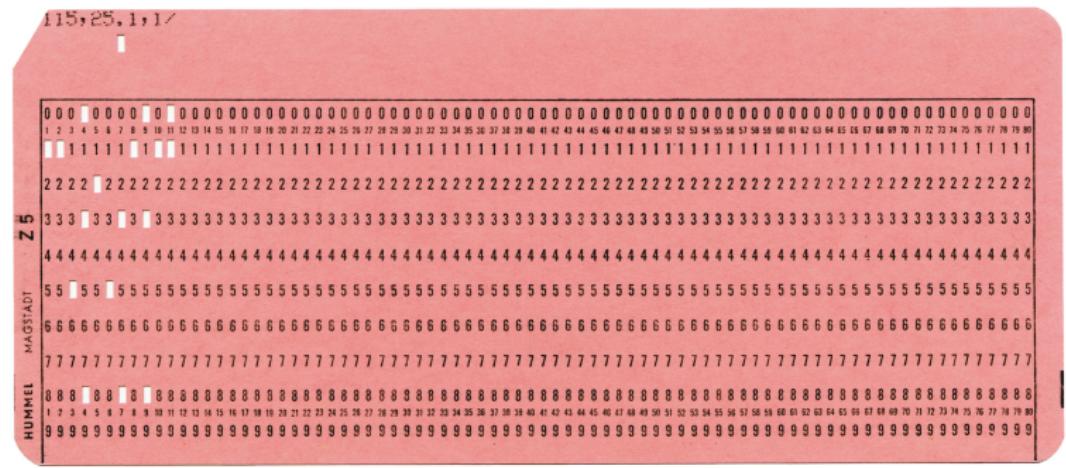
School Table	
ID	Name
S001	University of Technology
S002	University of Applied Science

Student Table			
School ID	ID	Name	DOB
S001	UT-1000	Tommy	05/06/1995
S001	UT-1000	Better	16/04/1995
S002	UAS-1000	Linda	02/09/1995
S002	UAS-1000	Jonathan	22/06/1995

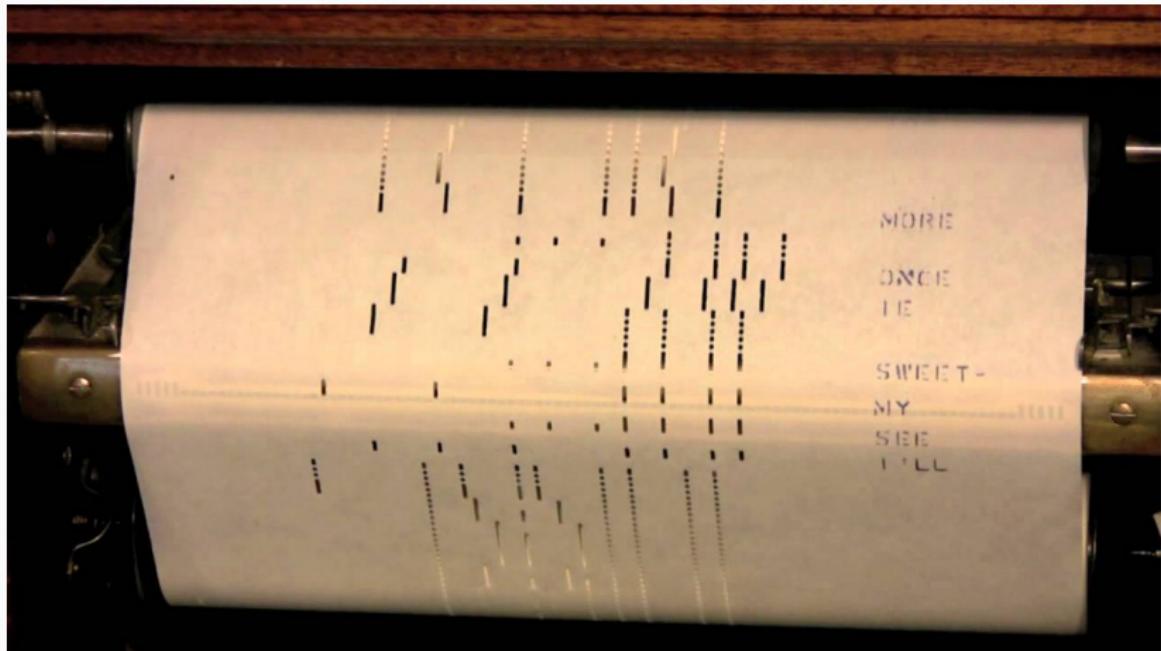
# Punch cards (and legacy systems)

- How we used to enter programs into the computer



- Pre-computer origins: 18-th century use in textile looms
  - Responsible for the 80-character legacy

## Who knows what this is?



That is a [piano roll](#), where music to be played by a [player piano](#) is encoded.

# The point: Information is more than data

## Data

- raw, unorganized facts that need to be processed
- unusable until it is organized

## Information

- created when data is processed, organized, structured
- needs to be situated in an appropriate *context* in order to become useful

**There are important differences in how humans and computers treat data as information**

## Basic units of data

- Bits
  - smallest unit of storage, a 0 or 1
  - anything that can store two states - now “transistors”, used to be vacuum tubes
  - with  $n$  bits, can store  $2^n$  patterns - so one byte can store 256 patterns

## Bytes

- eight *bits* = one *byte*
- “eight bit encoding” - represented characters, such as A represented as 65

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0 000	NUL	(null)	32	20 040	6#32;	Space		64	40 100	6#64;	Ø	96	60 140	6#96;	`		
1	1 001	SOH	(start of heading)	33	21 041	6#33;	!	!	65	41 101	6#65;	A	97	61 141	6#97;	a		
2	2 002	STX	(start of text)	34	22 042	6#34;	"	"	66	42 102	6#66;	B	98	62 142	6#98;	b		
3	3 003	ETX	(end of text)	35	23 043	6#35;	#	#	67	43 103	6#67;	C	99	63 143	6#99;	c		
4	4 004	EOT	(end of transmission)	36	24 044	6#36;	\$	\$	68	44 104	6#68;	D	100	64 144	6#100;	d		
5	5 005	ENQ	(enquiry)	37	25 045	6#37;	%	%	69	45 105	6#69;	E	101	65 145	6#101;	e		
6	6 006	ACK	(acknowledge)	38	26 046	6#38;	&	&	70	46 106	6#70;	F	102	66 146	6#102;	f		
7	7 007	BEL	(bell)	39	27 047	6#39;	'	'	71	47 107	6#71;	G	103	67 147	6#103;	g		
8	8 010	BS	(backspace)	40	28 050	6#40;	(	(	72	48 110	6#72;	H	104	68 150	6#104;	h		
9	9 011	TAB	(horizontal tab)	41	29 051	6#41;	)	)	73	49 111	6#73;	I	105	69 151	6#105;	i		
10	A 012	LF	(NL line feed, new line)	42	2A 052	6#42;	*	*	74	4A 112	6#74;	J	106	6A 152	6#106;	j		
11	B 013	VT	(vertical tab)	43	2B 053	6#43;	+	+	75	4B 113	6#75;	K	107	6B 153	6#107;	k		
12	C 014	FF	(NP form feed, new page)	44	2C 054	6#44;	,	,	76	4C 114	6#76;	L	108	6C 154	6#108;	l		
13	D 015	CR	(carriage return)	45	2D 055	6#45;	-	-	77	4D 115	6#77;	M	109	6D 155	6#109;	m		
14	E 016	SO	(shift out)	46	2E 056	6#46;	.	.	78	4E 116	6#78;	N	110	6E 156	6#110;	n		
15	F 017	SI	(shift in)	47	2F 057	6#47;	/	/	79	4F 117	6#79;	O	111	6F 157	6#111;	o		
16	10 020	DLE	(data link escape)	48	30 060	6#48;	0	0	80	50 120	6#80;	P	112	70 160	6#112;	p		
17	11 021	DC1	(device control 1)	49	31 061	6#49;	1	1	81	51 121	6#81;	Q	113	71 161	6#113;	q		
18	12 022	DC2	(device control 2)	50	32 062	6#50;	2	2	82	52 122	6#82;	R	114	72 162	6#114;	r		
19	13 023	DC3	(device control 3)	51	33 063	6#51;	3	3	83	53 123	6#83;	S	115	73 163	6#115;	s		
20	14 024	DC4	(device control 4)	52	34 064	6#52;	4	4	84	54 124	6#84;	T	116	74 164	6#116;	t		
21	15 025	NAK	(negative acknowledge)	53	35 065	6#53;	5	5	85	55 125	6#85;	U	117	75 165	6#117;	u		
22	16 026	SYN	(synchronous idle)	54	36 066	6#54;	6	6	86	56 126	6#86;	V	118	76 166	6#118;	v		
23	17 027	ETB	(end of trans. block)	55	37 067	6#55;	7	7	87	57 127	6#87;	W	119	77 167	6#119;	w		
24	18 030	CAN	(cancel)	56	38 070	6#56;	8	8	88	58 130	6#88;	X	120	78 170	6#120;	x		
25	19 031	EM	(end of medium)	57	39 071	6#57;	9	9	89	59 131	6#89;	Y	121	79 171	6#121;	y		
26	1A 032	SUB	(substitute)	58	3A 072	6#58;	:	:	90	5A 132	6#90;	Z	122	7A 172	6#122;	z		
27	1B 033	ESC	(escape)	59	3B 073	6#59;	:	:	91	5B 133	6#91;	[	123	7B 173	6#123;	{		
28	1C 034	FS	(file separator)	60	3C 074	6#60;	<	<	92	5C 134	6#92;	\	124	7C 174	6#124;			
29	1D 035	GS	(group separator)	61	3D 075	6#61;	=	=	93	5D 135	6#93;	]	125	7D 175	6#125;	}		
30	1E 036	RS	(record separator)	62	3E 076	6#62;	>	>	94	5E 136	6#94;	^	126	7E 176	6#126;	~		
31	1F 037	US	(unit separator)	63	3F 077	6#63;	?	?	95	5F 137	6#95;	_	127	7F 177	6#127;	DEL		

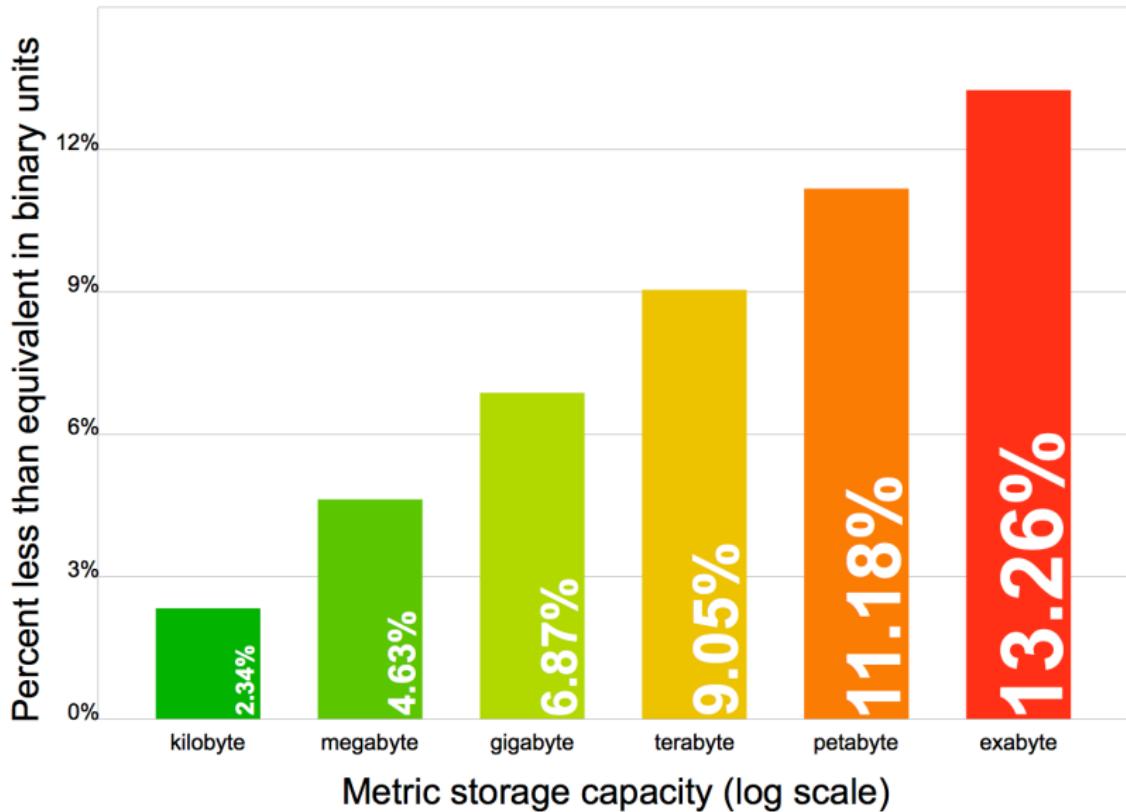
Source: [www.LookupTables.com](http://www.LookupTables.com)

## multi-byte units

unit	abbreviation	total bytes	nearest decimal equivalent
kilobyte	KB	$1,024^1$	$1000^1$
megabyte	MB	$1,024^2$	$1000^2$
gigabyte	GB	$1,024^3$	$1000^3$
terabyte	TB	$1,024^4$	$1000^4$
petabyte	PB	$1,024^5$	$1000^5$
exabyte	EB	$1,024^6$	$1000^6$
zettabyte	ZB	$1,024^7$	$1000^7$
yottabyte	YB	$1,024^8$	$1000^8$

- this is why 1GB is greater than 1 billion bytes

# Comparison of decimal and binary units



## Data types: Generically

- objects are *bound* to an identifier, e.g.

```
temperature = 98.6  
print(temperature)
```

98.6

- here, `temperature` is a variable name assigned to the literal floating-point object with the value of 98.6
- in Python, this is an instance of the `float` class
- identifiers in R and Python are *case-sensitive*
- some identifiers are typically reserved, e.g.

```
False, True, None, or, and # Python  
FALSE, TRUE, NA           # R
```

- All programming languages use comments, for humans to read
- this is anything that follows the # character in both Python and R

*“Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do.” – Donald Knuth, Literate Programming (1984)*

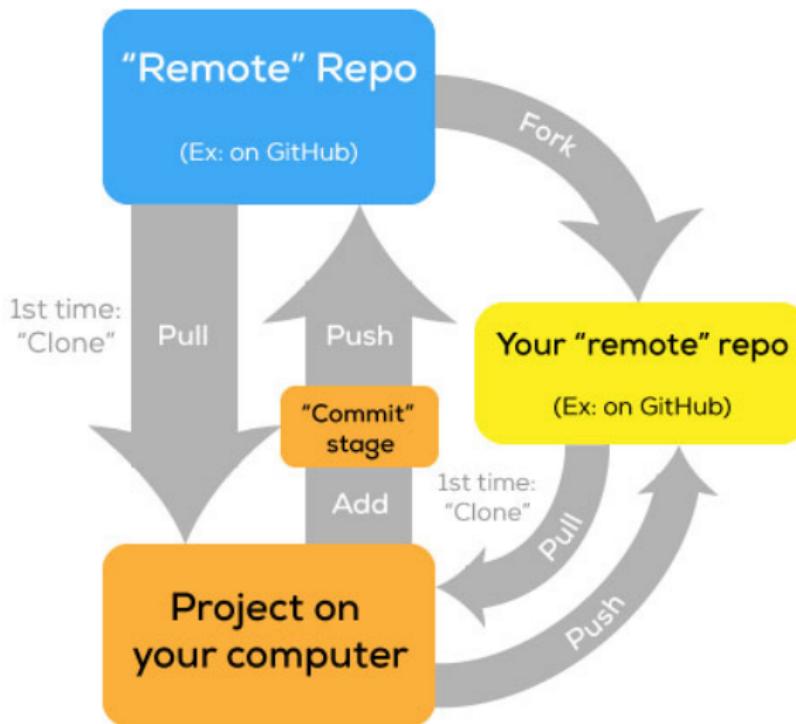
# git

- git: a revision control system
- Allows for complete history of changes, branching, staging areas, and flexible and distributed workflows

- simplified workflow (from [Anita Cheng's excellent blog post](#))

## Git for non-developers

(in a tiny nutshell)



## different ways to use git

through the command line, or through GUI clients, or through most IDEs and (good) editors



lectures — bash — 94x25

```
[KB-Office-iMac:lectures kbenoit$ cd ~/Dropbox/GitHub/lse-st445/lectures/
[KB-Office-iMac:lectures kbenoit$ git add *
[KB-Office-iMac:lectures kbenoit$ git commit -m "Update week 1 lectures"
[master ca47e74] Update week 1 lectures
 10 files changed, 1402 insertions(+)
 create mode 100644 week01/R_example.ipynb
 create mode 100644 week01/ST445_wk1_admin.ipynb
 create mode 100644 week01/ST445_wk1_class.ipynb
 create mode 100644 week01/ST445_wk1_lecture.ipynb
 create mode 100644 week01/figs/cardcatalog.jpg
 create mode 100644 week01/figs/cardcatalog2.jpg
 create mode 100644 week01/figs/photos.png
 create mode 100644 week01/figs/pianoroll.png
 create mode 100644 week01/figs/punchard.jpg
 create mode 100644 week01/figs/relational_data.png
[KB-Office-iMac:lectures kbenoit$ git push
warning: redirecting to https://github.com/lse-st445/lectures/
Counting objects: 14, done.
Delta compression using up to 8 threads.
Compressing objects: 100% (14/14), done.
Writing objects: 100% (14/14), 5.57 MiB | 3.14 MiB/s, done.
Total 14 (delta 0), reused 0 (delta 0)
To http://github.com/lse-st445/lectures
 2d26aa9..ca47e74  master -> master
KB-Office-iMac:lectures kbenoit$
```

## different ways to use git

or, RStudio

RStudio: Review Changes

Changes History master | Stage | Revert | Ignore | Pull | Push

Staged	Status	Path
<input checked="" type="checkbox"/>	A	day02/ME314_day2.Rmd
<input checked="" type="checkbox"/>	A	day02/ME314_day2.html
<input checked="" type="checkbox"/>	M	day02/ME314_day2.ipynb
<input checked="" type="checkbox"/>	A	day02/libs/header-attrs/header-attrs.js
<input checked="" type="checkbox"/>	A	day02/libs/remark-css/default-fonts.css
<input checked="" type="checkbox"/>	A	day02/libs/remark-css/default.css

Commit message  
Update Day 2  
 Amend previous commit

Show Staged Unstaged Context 5 lines ▲  Ignore Whitespace  Unstage All

```
@@ -0,0 +1,1055 @@
1 ---
2 title: "The Shape of Data"
3 author: "Ken Benoit"
4 date: "ME314 Lecture 2, 22 June 2021"
5 output:
6   xaringan::moon_reader:
7     css: [default, default-fonts]
8     lib_dir: libs
9     toc: true
10    toc_depth: 2
11    number_sections: true
12    nature:
```

# different ways to use git

or through editors (here, the excellent [Visual Studio Code](#) editor)

The screenshot shows the Visual Studio Code interface with the following details:

- SOURCE CONTROL** panel on the left: Shows a list of changes made to files like `ME314_day1.2.html`, `ME314_day2.Rmd`, and `DS202_Week_9.pdf`.
- Message** input field: "Enter to commit on ..."
- Commit** button: "✓ Commit" (disabled)
- ME314\_day2.Rmd — lectures** tab: Contains Rmd code for a presentation.
- PROBLEMS** and **TERMINAL** tabs: The TERMINAL tab shows the following output:

```
day02 > ME314_day2.Rmd
1 ---
2   title: "The Shape of Data"
3   subtitle: "ME314: Introduction to Data Science and Machine
Learning"
4   author: "Ken Benoit"
5   date: "11 July 2023"
6   output:
7     beamer_presentation:
8       keep_tex: no
9       theme: metropolis
10      latex_engine: xelatex
11      slide_level: 2
=====
downloaded 18 KB
trying URL 'https://cran.r-project.org/bin/macosx/big-sur-arm64/contrib/4.3/
languageserver_0.3.15.tgz'
Content type 'application/x-gzip' length 818528 bytes (799 KB)
=====
downloaded 799 KB

The downloaded binary packages are in
/var/folders/rc/y2gv0_dn0z19yggq271w7f9dr0000gs/T//RtmpI9wPaM/downloa
ded_packages
* Terminal will be reused by tasks, press any key to close it.
```

# GitHub

- a website and hosting platform for git repositories
- and so much more...
  - publishing websites: <http://kenbenoit.net>, whose source code is at <https://github.com/kbenoit/kbenoit.github.io>
  - “continuous integration” hooks: <https://github.com/kbenoit/spacyr> (for instance - see the badges)
  - Issue tracking: <https://github.com/kbenoit/spacyr/issues>
  - Inspecting code: (e.g.) <https://github.com/kbenoit/spacyr/blob/master/R/python-functions.R>
- GitHub classroom
- Free stuff for students! <https://education.github.com/pack>

## More resources

### More great resources for using git/GitHub

- An easy git Cheatsheet, by Nina Jaeschke and Roger Dudler
- git - the simple guide by Roger Dudler

Some people have entire, open-source, user-commented books online, such as:

- *R for Data Science*
  - with [source code here](#)
  - with [GitHub issues here](#)
  - and pull requests - [examples here](#)

# Markdown (and other markup languages)

- Idea of a “markup” language: HTML, XML, LaTeX
- “Markdown”
  - Created by John Gruber as a simple way for non-programming types to write in an easy-to-read format that could be converted directly into HTML
  - No opening or closing tags
  - Plain text, and can be read when not rendered
- Markdown has [many “flavours”](#)

## Markdown example

This is a markdown example.

\* bullet list 1  
\* bullet list 2  
*"I love deadlines. I like the whooshing sound they make as they fly by."*

– Douglas Adams

```
## Markdown example
```

This is a markdown example

- \* bullet list 1
- \* bullet list 2

```
> "[I love deadlines. I like the whooshing sound they make as they fly by.]"
-- _Douglas Adams_
```

A good reference for Markdown:

<https://ia.net/writer/support/general/markdown-guide/>.

## A note on “meta-data”

- Data that provides information about other (primary) data
- Usually not meant to be analyzed as data itself
- Example: HTML

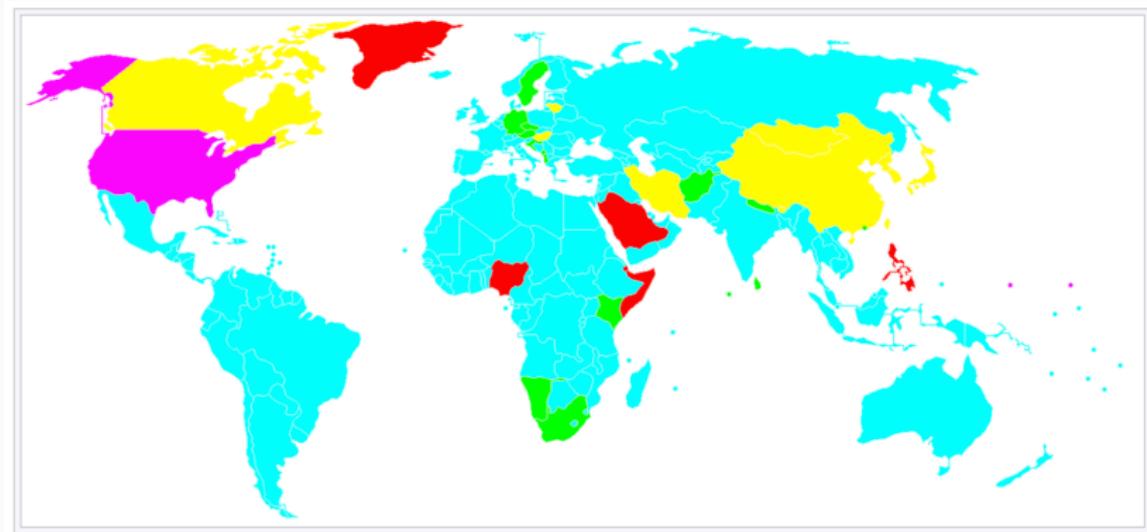
```
<!DOCTYPE html>  
<html class="client-nojs" lang="en" dir="ltr">  
<head>  
<meta charset="UTF-8"/>  
<title>Metadata - Wikipedia</title>
```

- Example of a standard attempting to address this need: [Dublin Core Metadata Initiative](#)

# Representing dates: Different formats?

Color	Order styles	Main regions and countries (approximate population of each region in millions)	Approximate population in millions
Cyan	DMY	Asia (Central, SE, West), Australia (25), New Zealand (5), parts of Europe (c. 640), Latin America (625), North Africa (195), India (1315), Indonesia (265), Bangladesh (165), Russia (145)	3565
Yellow	YMD	Bhutan, Canada (35), China (1385), Koreas (75), Taiwan (24), Hungary (10), Iran (80), Japan (125), Lithuania (5), Mongolia (5). Known in other countries due to ISO 8601.	1745
Magenta	MDY	United States (325), Federated States of Micronesia, Marshall Islands	325
Red	DMY, MDY	Malaysia (35), Nigeria (190), Philippines (105), Saudi Arabia (35), Somalia (10)	380
Green	DMY, YMD	Afghanistan (28), Albania (3), Austria (9), Czech Republic (11), Kenya (49), Macau (1), Maldives, Montenegro, Namibia (2), Nepal (29), Singapore (6), South Africa (56), Sri Lanka (21), Sweden (10) <sup>[1]</sup>	225

## Representing dates: Different formats?



## Representing dates

Description	Format	Examples
American month and day	mm/dd	"5/12", "10/27"
American month, day and year	mm/dd/y	"1/17/2006"
Four digit year, month and day with slashes	YY/mm/dd	"2008/6/30", "1978/12/22"
Four digit year and month (GNU)	YY-mm	"2008-6", "2008-06", "1978-12"
Year, month and day with dashes	y-mm-dd	"2008-6-30", "78-12-22", "8-6-21"
Day, month and four digit year	dd-mm-YY	"30-6-2008"
Day, month and two digit year	dd.mm.yy	"30.6.08"
Day, textual month and year	dd-m y	"30-June 2008"
Textual month and four	m YY	"June 2008",

## ISO8601: Imposing common standards

- Purpose: to provide unambiguous and well-defined method of representing dates and times
- Goal: to avoid misinterpretation of numeric representations of dates and times, particularly when data are transferred between countries with different conventions for writing numeric dates and time
- First published in 1988
- Introduces a common notation, and a common order (most-to-least-significant order [YYYY]-[MM]-[DD])
  - matches lexicographical order with chronological order
- Uses codes for date and time elements, to represent dates (and times) in either a basic format (no separators) or in an extended format with added separators (to enhance human readability)

# ISO8601 formatting components

Symbol	Meaning	Example	Notes
YYYY	4-digit year	2017	Avoids the "Y2K" problem
MM	2-digit day of the month	10	
DD	2-digit day of the month	03	
Www	Week number	52	
D	Weekday number	2	Starts on Monday!
hh	hour (0-24)	10	24 is only used to denote midnight at the end of a

# Coordinated Universal Time (UTC)

- World standard for time
- Does not include Daylight Savings Time
- Interchangeable with Greenwich Mean Time (GMT), but GMT is no longer precisely defined by the scientific community
- **Time zones around the world** are expressed using positive or negative offsets from UTC
- French v. English
  - English speakers originally proposed *CUT* (for “coordinated universal time”)
  - French speakers proposed *TUC* (for “temps universel coordonné”)
  - Compromise: *UTC*

## POSIX Time

- a system for describing a point in time, defined as the number of seconds that have elapsed since 00:00:00 UTC, Thursday, 1 January 1970
- Also known as “Unix time”, or “**epoch time**”, because it represents elapsed time from a defined “epoch”
- Problem: How much elapsed time can you store?

## The “Year 2038 Problem”

- Many Unix-like operating systems which keep time as seconds elapsed from the epoch date of January 1, 1970
- For signed 32-bit integers, this means that cannot encode times after 03:14:07 UTC on 19 January 2038
- Times beyond that will wrap around and be stored internally as a negative number, which these systems will interpret as having occurred on 13 December 1901
- A solution: 64-bit signed integers allow a new wraparound date that is 20x greater than the estimated age of the universe: approximately 290 billion years in the future

# Dataset manipulation

## What is a “Dataset”?

- A dataset *is* a “rectangular” formatted table of data in which all the values of the same variable must be in a single column
- A dataset *is not*:
  - the results of tabulating a dataset
  - any set of summary statistics on a dataset
  - a series of relational tables
- Many of the datasets we use have been artificially reshaped in order to fulfill this criterion of rectangularity
  - This means “non-normalized” data
  - Often confounds variables with their values

# The difference between a dataset and table

- This is a *table*:

		Lost	Won
Challenger	266	60	
Incumbent	32	106	

- This is a (partial) dataset:

		district	incumbf	wonseatf
1	Carlow Kilkenny	Challenger		Lost
2	Carlow Kilkenny	Challenger		Lost
5	Carlow Kilkenny	Incumbent		Won
100	Donegal South West	Challenger		Lost
459		Wicklow	Incumbent	Won
464		Wicklow	Challenger	Lost

# Hadley Wickham's three rules for “tidy” datasets

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.

country	year	cases	population
Afghanistan	2000	735	152071
Afghanistan	2000	566	20395360
Brazil	1999	31737	172006362
Brazil	2000	81488	17404898
China	1999	21258	127201272
China	2000	21066	12802583

variables

country	year	cases	population
Afghanistan	1999	735	152071
Afghanistan	2000	566	20395360
Brazil	1999	31737	172006362
Brazil	2000	81488	17404898
China	1999	21258	127201272
China	2000	21066	12802583

observations

country	year	cases	population
Afghanistan	1999	735	152071
Afghanistan	2000	566	20395360
Brazil	1999	31737	172006362
Brazil	2000	81488	17404898
China	1999	21258	127201272
China	2000	21066	12802583

values

## Example from *R for Data Science*

- Example of non-tidy data

```
suppressPackageStartupMessages(library("tidyverse"))
print(table4a)
```

```
## # A tibble: 3 x 3
##   country     `1999` `2000`
##   <chr>       <dbl>   <dbl>
## 1 Afghanistan    745    2666
## 2 Brazil        37737   80488
## 3 China         212258  213766
```

## What we need to do: make this into “long” format

country	year	cases	country	1999	2000
Afghanistan	1999	745	Afghanistan	745	2666
Afghanistan	2000	2666	Brazil	37737	80488
Brazil	1999	37737	China	212258	213766
Brazil	2000	80488			
China	1999	212258			
China	2000	213766			

table4

## Current recommended way: pivot\_longer()

```
pivot_longer(table4a, c(`1999`, `2000`),
            names_to = "year", values_to = "cases")

## # A tibble: 6 x 3
##   country     year   cases
##   <chr>       <chr>   <dbl>
## 1 Afghanistan 1999     745
## 2 Afghanistan 2000    2666
## 3 Brazil       1999   37737
## 4 Brazil       2000   80488
## 5 China        1999  212258
## 6 China        2000  213766
```

## Previously: “gathering”

```
gather(table4a, `1999`, `2000`, key = "year", value = "cases")

## # A tibble: 6 x 3
##   country     year   cases
##   <chr>       <chr>   <dbl>
## 1 Afghanistan 1999     745
## 2 Brazil       1999   37737
## 3 China        1999  212258
## 4 Afghanistan 2000    2666
## 5 Brazil       2000   80488
## 6 China        2000  213766
```

## Reshaping to wide format: pivot\_wider()

```
print(table2)
```

```
## # A tibble: 12 x 4
##   country     year type     count
##   <chr>       <dbl> <chr>     <dbl>
## 1 Afghanistan 1999 cases      745
## 2 Afghanistan 1999 population 19987071
## 3 Afghanistan 2000 cases     2666
## 4 Afghanistan 2000 population 20595360
## 5 Brazil       1999 cases     37737
## 6 Brazil       1999 population 172006362
## 7 Brazil       2000 cases     80488
## 8 Brazil       2000 population 174504898
## 9 China        1999 cases     212258
## 10 China       1999 population 1272915272
## 11 China       2000 cases     213766
## 12 China       2000 population 1280428583
```

# Objective: make long column into wide columns

country	year	key	value	country	year	cases	population
Afghanistan	1999	cases	745	Afghanistan	1999	745	19987071
Afghanistan	1999	population	19987071	Afghanistan	2000	266	20595360
Afghanistan	2000	cases	2666	Brazil	1999	37737	172006362
Afghanistan	2000	population	20595360	Brazil	2000	80488	174504898
Brazil	1999	cases	37737	China	1999	212258	1272915272
Brazil	1999	population	172006362	China	2000	213766	1280428583
Brazil	2000	cases	80488				
Brazil	2000	population	174504898				
China	1999	cases	212258				
China	1999	population	1272915272				
China	2000	cases	213766				
China	2000	population	1280428583				

table2

```
pivot_wider(table2, names_from = type, values_from = count)

## # A tibble: 6 x 4
##   country     year   cases population
##   <chr>       <dbl>   <dbl>      <dbl>
## 1 Afghanistan 1999    745 19987071
## 2 Afghanistan 2000   2666 20595360
## 3 Brazil      1999  37737 172006362
## 4 Brazil      2000  80488 174504898
## 5 China       1999 212258 1272915272
## 6 China       2000 213766 1280428583
```

## Older method: spread()

```
spread(table2, key = type, value = count)
```

```
## # A tibble: 6 x 4
##   country     year   cases population
##   <chr>       <dbl>   <dbl>      <dbl>
## 1 Afghanistan 1999     745 19987071
## 2 Afghanistan 2000    2666 20595360
## 3 Brazil       1999  37737 172006362
## 4 Brazil       2000  80488 174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

But you should use the current `pivot_*`() functions in **tidyverse**.

## Example with Manifesto Data

```
load("cmpdata.Rdata")
glimpse(cmpdata)

## Rows: 709
## Columns: 83
## $ country      <dbl> 42, 42, 42, 42, 42, 42, 42, 42, 42, 42,
## $ countryname   <chr> "Austria", "Austria", "Austria", "Austr
## $ oecdmember    <int> 10, 10, 10, 10, 10, 10, 10, 10, 10,
## $ eumember       <int> 0, 0, 0, 0, 0, 0, 0, 0, 10, 10, 10,
## $ edate         <date> 1990-10-07, 1990-10-07, 1990-10-07, 19
## $ date          <int> 199010, 199010, 199010, 199010, 199410,
## $ party         <dbl> 42420, 42110, 42320, 42520, 42420, 4242
## $ partyname     <chr> "FPÖ Freedom Party", "GA Green Alternat
## $ parfam        <int> 40, 10, 30, 50, 40, 40, 10, 30, 50, 50,
## $ coderid       <int> 202, 202, 202, 202, 203, 203, 203, 203,
## $ coderyear     <int> 1991, 1991, 1991, 1991, 1998, 1998, 199
## $ manual        <int> 1, 1, 1, 1, NA, NA, NA, NA, NA, NA, NA,
## $ testresult    <dbl> 0.8160, 0.8160, 0.8160, 0.8160, 0.7650,
```

## select just ID variables and the codes

```
cmpdata <- cmpdata %>%
  select(countryname, date, partyname, starts_with("per")) %>%
  select(-pervote)
```

```
glimpse(cmpdata)
```

```
## Rows: 709
## Columns: 60
## $ countryname <chr> "Austria", "Austria", "Austria", "Austria"
## $ date         <int> 199010, 199010, 199010, 199010, 199410, 1
## $ partyname   <chr> "FPÖ Freedom Party", "GA Green Alternative"
## $ per101       <dbl> 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ per102       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ per103       <dbl> 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ per104       <dbl> 37, 0, 7, 6, 7, 0, 0, 13, 13, 0, 0, 0, 2,
## $ per105       <dbl> 1, 3, 2, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
## $ per106       <dbl> 1, 0, 4, 3, 0, 0, 3, 3, 9, 0, 1, 5, 0, 51
## $ per107       <dbl> 16, 17, 44, 22, 3, 0, 2, 26, 62, 3, 1, 11
```

```
cmpdata_long <-  
  pivot_longer(cmpdata, cols = starts_with("per"),  
               names_to = "code", values_to = "value")  
  
cmpdata_long  
  
## # A tibble: 40,413 x 5  
##   countryname    date partyname      code  value  
##   <chr>        <int> <chr>       <chr>  <dbl>  
## 1 Austria      199010 FPÖ Freedom Party per101     0  
## 2 Austria      199010 FPÖ Freedom Party per102     0  
## 3 Austria      199010 FPÖ Freedom Party per103     0  
## 4 Austria      199010 FPÖ Freedom Party per104    37  
## 5 Austria      199010 FPÖ Freedom Party per105     1  
## 6 Austria      199010 FPÖ Freedom Party per106     1  
## 7 Austria      199010 FPÖ Freedom Party per107    16  
## 8 Austria      199010 FPÖ Freedom Party per108     3  
## 9 Austria      199010 FPÖ Freedom Party per109     0  
## 10 Austria     199010 FPÖ Freedom Party per110     0  
## # i 40,403 more rows
```

## Compare pro- versus anti- EU sentiment

- per108: European Community/Union: Positive
- per110: European Community/Union: Negative

See details at the [Manifesto Project website](#)

```
cmpdata_long <- mutate(cmpdata_long, year = floor(date / 100))
cmpdata_long2 <- cmpdata_long %>%
  group_by(code, year) %>%
  summarize(total = sum(value)) %>%
  filter(code %in% c("per108", "per110"))

## `summarise()` has grouped output by 'code'. You can override
## `.groups` argument.
```

## cmpdata\_long2

```
## # A tibble: 44 x 3
## # Groups:   code [2]
##       code     year   total
##       <chr>   <dbl> <dbl>
## 1 per108  1990    170
## 2 per108  1991    650
## 3 per108  1992    232
## 4 per108  1993    921
## 5 per108  1994    659
## 6 per108  1995    498
## 7 per108  1996   1964
## 8 per108  1997    311
## 9 per108  1998    492
## 10 per108 1999   1470
## # i 34 more rows
```

```
ggplot(cmpdata_long2, aes(x = year, y = total,  
                           colour = code, group = code)) +  
  geom_line() +  
  scale_colour_manual(name = "EU Sentiment",  
                      labels = c("Positive", "Negative"),  
                      values =c("darkgreen", "red"))
```

