Tidying: The Preliminary First/Last Step(s) for Data Modeling?

Ravichandran Sarangan
ABCC, FNCLR

Announcement

Drs. Randy Johnson and Ravichandran Sarangan will be running a Data Tidying R-workshop soon. Hope you all can join us and Please look out for the announcement.

Agenda

- What is Tidy data?
  - Definition
  - Examples

- How to tidy up the untidy data?
  - Many examples

Acknowledgement(s)

- Dr. Randy Johnson

- Statistics for Lunch team

- Ideas for the talk came from several places
  - Key resources are
    - Hadley Wickham (**RStudio**)
    - Karl Broaman **(Univ Wisconsin**)
    - Roger Peng, Brian Caffo, Jeff Leek (**Johns Hopkins**)
    - Rafael Irizarry (**Harvard**)

Tidy Data Not For Us?

We don’t model big data

Excel Data as shared with us

Conc(mM)	VER						C29						Z53						K21					
	R1	R2	R3	R4	R5	R6	R1	R2	R3	R4	R5	R6	R1	R2	R3	R4	R5	R6	R1	R2	R3	R4	R5	R6
0.2	77.0538	33.0005	94.3916	74.3856	76.9227	64.9896	27.2454	70.8575	95.0289	66.9970	38.9787		23.8547	27.0286	16.2678	76.9873	50.0295		86.1360	40.1581	64.0368	76.7385	89.3777	44.1586
2	64.6542	56.2507	39.3620	75.2164	7.6627	9.4396	16.8141	8.8717	49.4840	17.5830	43.0331		108.6977	41.7782	60.28067	61.60561	91.26872		24.65882	21.29545	21.29487	22.29555	22.38884	23.48461
20	100.5052	77.0049	90.2609	113.7513	66.0418	128.0548	61.3101	117.6954	75.0741	102.5791	119.8568		60.45198	99.30211	76.58083	34.81724	88.2234		43.99294	55.40685	88.4444	44.47773	52.47771	64.21042
200	22.7551	67.6204	37.4174	85.3549	44.9639	14.5867	7.4392	90.1611	85.5179	16.6576	66.8401		42.41031	23.80522	19.32308	3.62231	61.70047		91.78032	91.77401	100.0076	100.3718	97.39394	93.57512
2000	67.4748	45.0833	77.9093	64.6854	68.9216	23.2428	55.4230	32.8498	69.6015	42.3984	22.4975		11.19209	69.68825	2.876452	75.0157	23.32492		90.49027	98.52109	113.5611	107.5951	91.99885	92.37767
5000													22.85579	9.23904		87.16471	12.6984			88.55632	112.9554			

R1 to R6 are Replicates

Cell Growth Data (two more drug data not shown)

Test the significance of difference between the drugs

R1 to R6 are Replicates

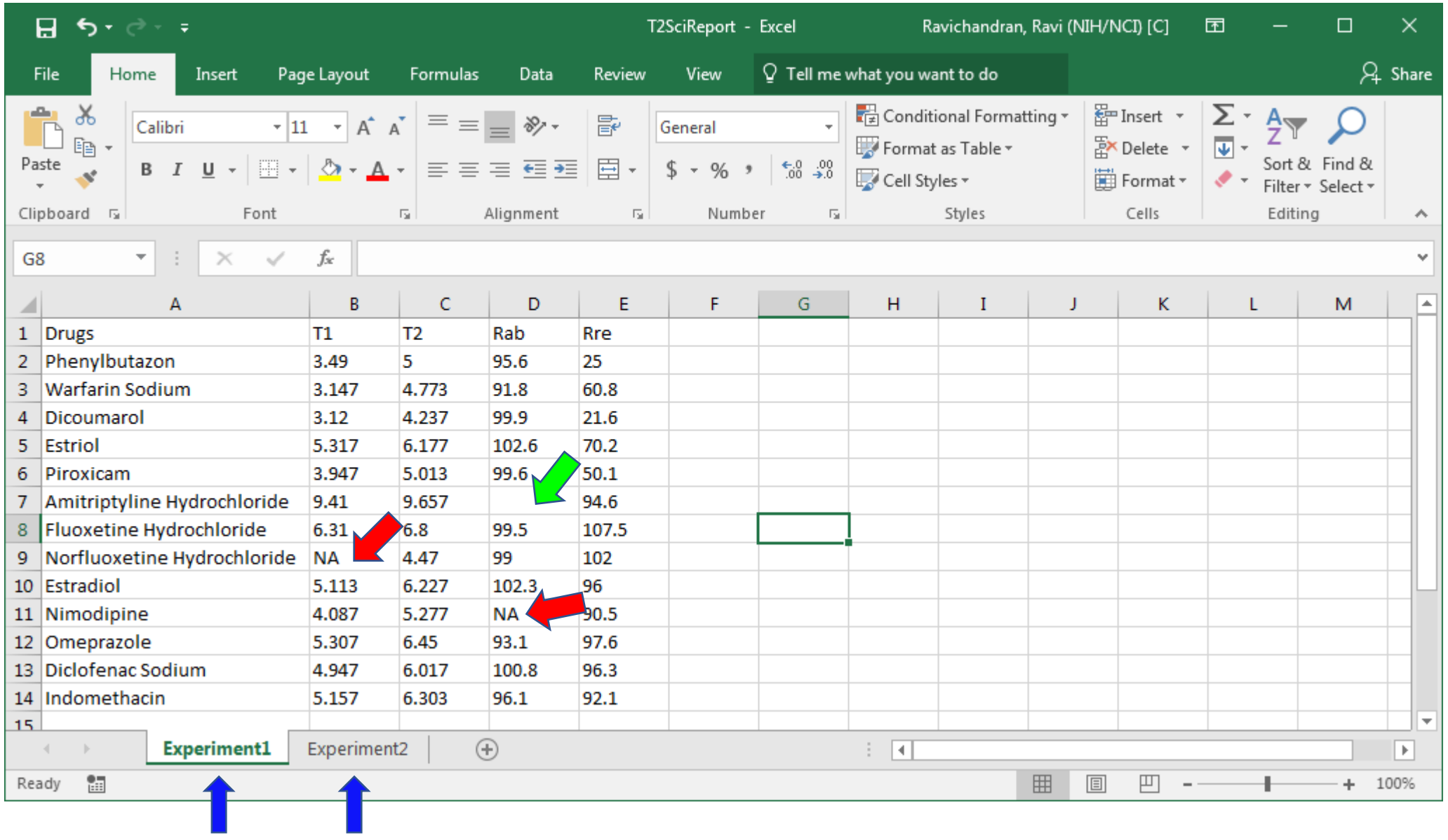
Concentration(mM)	VER					
	R1	R2	R3	R4	R5	R6
0.2	63.2915	68.6234	70.0896	38.9158	61.2060	80.7211
2	30.8031	59.2299	37.4290	79.4987	58.8359	28.4906
20	26.5052	108.7251	57.8343	51.9349	92.3215	114.4090
200	38.0781	64.5364	28.5504	54.0257	55.7642	17.1096
2000	37.1207	49.2985	30.0768	63.0939	28.1204	51.4099
5000						

Concentration(mM)	C29					
	R1	R2	R3	R4	R5	R6
0.2	81.9251	25.9428	88.3891	46.5541	27.8573	
2	58.1008	78.8404	8.7778	5.7540	75.8824	
20	96.3247	57.1182	102.3849	101.1493	78.0180	
200	32.5744	36.5940	81.5093	14.7338	63.7707	
2000	73.5428	74.1092	74.0179	41.6010	48.7619	
5000						

Concentration(mM)	K21					
	R1	R2	R3	R4	R5	R6
0.2	77.2862	22.9933	26.2564	49.7822	73.4912	25.5989
2	24.65441	22.76161	22.14644	21.66432	21.87527	22.8385
20	62.09718	66.93179	88.4444	17.09395	70.17626	20.70322
200	91.78032	91.77401	100.0076	100.3718	97.39394	93.57512
2000	90.49027	98.52109	113.5611	107.5951	91.99885	92.37767
5000		88.55632	112.9554			

Concentration(mM)	Z53					
	R1	R2	R3	R4	R5	R6
0.2	25.1243	71.3255	79.4123	20.7826	68.1928	
2	99.3483	28.27143	52.67069	108.1993	71.78687	
20	100.345	94.43867	97.50873	70.42895	88.2234	
200	27.76815	49.31593	16.73357	14.3842	68.24119	
2000	14.208	68.5947	15.57453	16.63922	53.24721	
5000	34.22645	84.47106		80.83111	21.71818	

Conc(mM)	Drug	Rep	Cell-Growth(%)
0.2	VER	R1	63.2915
0.2	C29	R1	81.2951



	A	B	C	D	E
1	Drugs	T1	T2	Rab	Rre
2	Phenylbutazon	3.49	5	95.6	25
3	Warfarin Sodium	3.147	4.773	91.8	60.8
4	Dicoumarol	3.12	4.237	99.9	21.6
5	Estriol	5.317	6.177	102.6	70.2
6	Piroxicam	3.947	5.013	99.6	50.1
7	Amitriptyline Hydrochloride	9.41	9.657		94.6
8	Fluoxetine Hydrochloride	6.31	6.8	99.5	107.5
9	Norfluoxetine Hydrochloride	NA	4.47	99	102
10	Estradiol	5.113	6.227	102.3	96
11	Nimodipine	4.087	5.277	NA	90.5
12	Omeprazole	5.307	6.45	93.1	97.6
13	Diclofenac Sodium	4.947	6.017	100.8	96.3
14	Indomethacin	5.157	6.303	96.1	92.1
15					

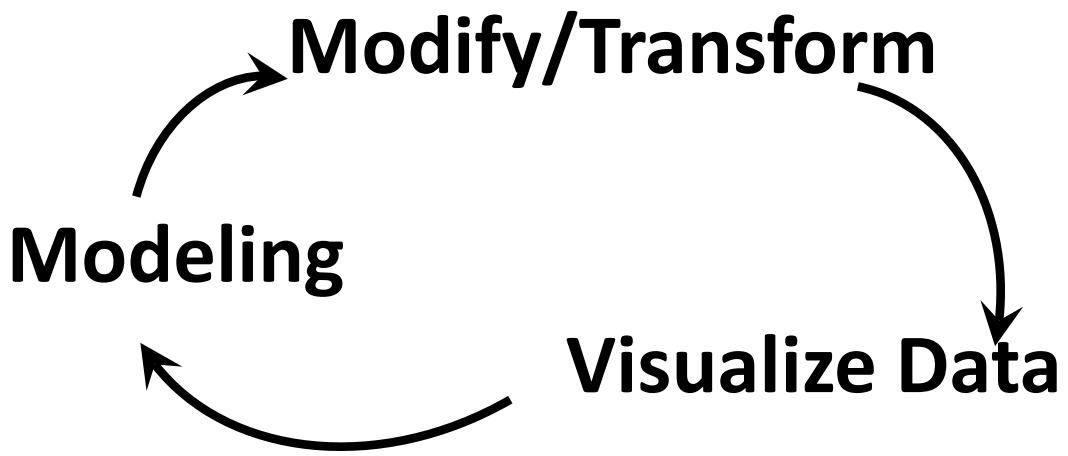
No manipulations had been done; full dataset

Excel sheet with 10 Worksheets someone sent you ☺

Raw

No clear procedures

Clean up (mostly missing data)



Raw Data: Hospital Compare Data

Rectangular Data

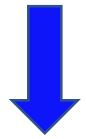
1	City	State	ZIP Code	County	Measure Name	Measure ID	VHA National Rate	Compare to National	Denominator	Score	Lower Estimate	Higher Estimate	Footnotes	Measure Start Date
1055	MILWAUKEE	WI	53295	MILWAUKEE	Rate of readmission for chi	READM-30-COPD	16.28	No different than the VHA	425	18.49	15.34	22.06		10/1/2012
1056	MILWAUKEE	WI	53295	MILWAUKEE	Heart failure (HF) 30-Day R	READM-30-HF	19.34	No different than the VHA	551	21.51	18.72	24.72		10/1/2012
1057	MILWAUKEE	WI	53295	MILWAUKEE	Pneumonia (PN) 30-Day Re	READM-30-PN	14.74	No different than the VHA	235	14.67	12.19	17.57		10/1/2012
1058	MADISON	WI	53705	DANE	Acute Myocardial Infarctio	MORT-30-AMI	9.21	No different than the VHA	142	7.8	5.56	10.28		10/1/2012
1059	MADISON	WI	53705	DANE	Death rate for chronic obst	MORT-30-COPD	5.99	No different than the VHA	170	4.65	2.81	7.15		10/1/2012
1060	MADISON	WI	53705	DANE	Heart failure (HF) 30-Day M	MORT-30-HF	7.72	No different than the VHA	408	7.15	5.52	9.25		10/1/2012
1061	MADISON	WI	53705	DANE	Pneumonia (PN) 30-Day M	MORT-30-PN	8.64	No different than the VHA	210	7.12	5.2	9.33		10/1/2012
1062	MADISON	WI	53705	DANE	Acute Myocardial Infarctio	READM-30-AMI	15.57	No different than the VHA	156	14.76	12.14	17.34		10/1/2012
1063	MADISON	WI	53705	DANE	Rate of readmission for chi	READM-30-COPD	16.28	No different than the VHA	185	17.9	14.27	22.65		10/1/2012
1064	MADISON	WI	53705	DANE	Heart failure (HF) 30-Day R	READM-30-HF	19.34	No different than the VHA	460	19.78	17.27	22.99		10/1/2012
1065	MADISON	WI	53705	DANE	Pneumonia (PN) 30-Day Re	READM-30-PN	14.74	No different than the VHA	217	16.17	13.47	19.51		10/1/2012
1066	CHEYENNE	WY	82001	LARAMIE	Acute Myocardial Infarctio	MORT-30-AMI	9.21	Number of Cases Too Smal	Not Available	Not Available	Not Available	Not Available	1 - The number of cases/pa	10/1/2012
1067	CHEYENNE	WY	82001	LARAMIE	Death rate for chronic obst	MORT-30-COPD	5.99	No different than the VHA	102	6.08	3.78	10.21		10/1/2012
1068	CHEYENNE	WY	82001	LARAMIE	Heart failure (HF) 30-Day M	MORT-30-HF	7.72	No different than the VHA	77	7.52	5.18	10.42		10/1/2012
1069	CHEYENNE	WY	82001	LARAMIE	Pneumonia (PN) 30-Day M	MORT-30-PN	8.64	No different than the VHA	143	8.1	6.01	10.66		10/1/2012
1070	CHEYENNE	WY	82001	LARAMIE	Acute Myocardial Infarctio	READM-30-AMI	15.57	Number of Cases Too Smal	Not Available	Not Available	Not Available	Not Available	1 - The number of cases/pa	10/1/2012
1071	CHEYENNE	WY	82001	LARAMIE	Rate of readmission for chi	READM-30-COPD	16.28	No different than the VHA	115	15.4	11.66	19.94		10/1/2012
1072	CHEYENNE	WY	82001	LARAMIE	Heart failure (HF) 30-Day R	READM-30-HF	19.34	No different than the VHA	77	18.74	15.12	22.63		10/1/2012
1073	CHEYENNE	WY	82001	LARAMIE	Pneumonia (PN) 30-Day Re	READM-30-PN	14.74	No different than the VHA	147	15.01	12.52	18.32		10/1/2012
1074	SHERIDAN	WY	82801	SHERIDAN	Acute Myocardial Infarctio	MORT-30-AMI	9.21	Number of Cases Too Smal	Not Available	Not Available	Not Available	Not Available	1 - The number of cases/pa	10/1/2012
1075	SHERIDAN	WY	82801	SHERIDAN	Death rate for chronic obst	MORT-30-COPD	5.99	Number of Cases Too Smal	Not Available	Not Available	Not Available	Not Available	1 - The number of cases/pa	10/1/2012
1076	SHERIDAN	WY	82801	SHERIDAN	Heart failure (HF) 30-Day M	MORT-30-HF	7.72	Number of Cases Too Smal	Not Available	Not Available	Not Available	Not Available	1 - The number of cases/pa	10/1/2012
1077	SHERIDAN	WY	82801	SHERIDAN	Pneumonia (PN) 30-Day M	MORT-30-PN	8.64	No different than the VHA	34	8.83	6.36	12.3		10/1/2012
1078	SHERIDAN	WY	82801	SHERIDAN	Acute Myocardial Infarctio	READM-30-AMI	15.57	Number of Cases Too Smal	Not Available	Not Available	Not Available	Not Available	1 - The number of cases/pa	10/1/2012
1079	SHERIDAN	WY	82801	SHERIDAN	Rate of readmission for chi	READM-30-COPD	16.28	Number of Cases Too Smal	Not Available	Not Available	Not Available	Not Available	1 - The number of cases/pa	10/1/2012
1080	SHERIDAN	WY	82801	SHERIDAN	Heart failure (HF) 30-Day R	READM-30-HF	19.34	Number of Cases Too Smal	Not Available	Not Available	Not Available	Not Available	1 - The number of cases/pa	10/1/2012
1081	SHERIDAN	WY	82801	SHERIDAN	Pneumonia (PN) 30-Day Re	READM-30-PN	14.74	No different than the VHA	32	14.28	11.04	17.55		10/1/2012
1082	ORLANDO	FL	32827	ORANGE	Acute Myocardial Infarctio	MORT-30-AMI	9.21	Not Available	Not Available	Not Available	Not Available	Not Available	5 - Results are not available	10/1/2012
1083	ORLANDO	FL	32827	ORANGE	Death rate for chronic obst	MORT-30-COPD	5.99	Not Available	Not Available	Not Available	Not Available	Not Available	5 - Results are not available	10/1/2012
1084	ORLANDO	FL	32827	ORANGE	Heart failure (HF) 30-Day M	MORT-30-HF	7.72	Not Available	Not Available	Not Available	Not Available	Not Available	5 - Results are not available	10/1/2012
1085	ORLANDO	FL	32827	ORANGE	Pneumonia (PN) 30-Day M	MORT-30-PN	8.64	Not Available	Not Available	Not Available	Not Available	Not Available	5 - Results are not available	10/1/2012
1086	ORLANDO	FL	32827	ORANGE	Acute Myocardial Infarctio	READM-30-AMI	15.57	Not Available	Not Available	Not Available	Not Available	Not Available	5 - Results are not available	10/1/2012
1087	ORLANDO	FL	32827	ORANGE	Rate of readmission for chi	READM-30-COPD	16.28	Not Available	Not Available	Not Available	Not Available	Not Available	5 - Results are not available	10/1/2012
1088	ORLANDO	FL	32827	ORANGE	Heart failure (HF) 30-Day R	READM-30-HF	19.34	Not Available	Not Available	Not Available	Not Available	Not Available	5 - Results are not available	10/1/2012
1089	ORLANDO	FL	32827	ORANGE	Pneumonia (PN) 30-Day Re	READM-30-PN	14.74	Not Available	Not Available	Not Available	Not Available	Not Available	5 - Results are not available	10/1/2012

NOT all Not Available data are same

https://catalog.data.gov/dataset/hospital-compare-data-17295

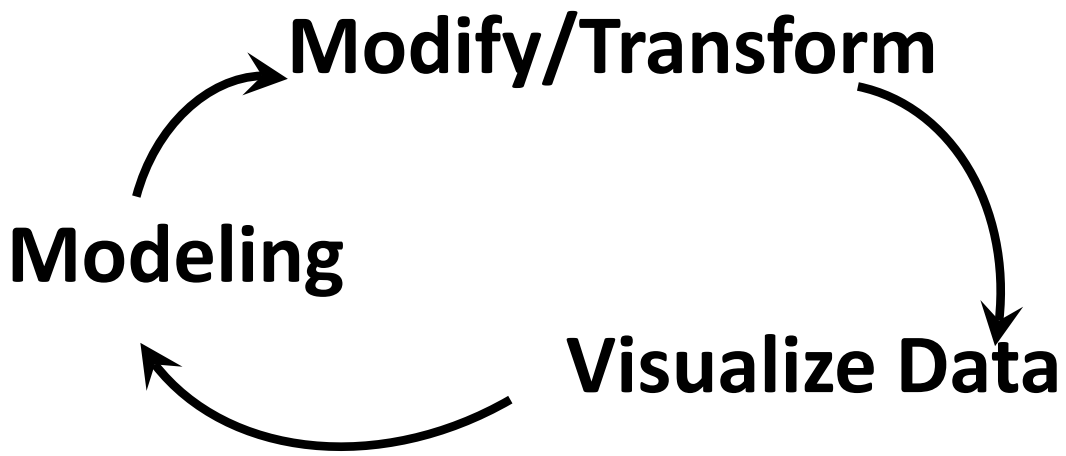
Less time munching the data from one format to another

Raw



Tidy

Series of clearly defined steps, procedures



Raw Data: Starting from the basics

- **Examples of Raw Data**
  - Census data; output from your equipment

- **Why are most data untidy?**
  - People collecting the data enter the data in a convenient way
  - People collecting the data don’t worry about variables and analysis

Raw Data: Starting from the basics

- **What is wrong with Raw Data?**
  - Almost 80% of data analysis is spent in cleaning the data (tidying the data)
  - Non-Messy, cleaned data is rare

- **OK, let us clean it?**
  - No proper procedure exits

Common Data Structure

- Dataset contains Rows and Columns

- Columns are labelled and Rows are labeled sometimes

For clarity, let us look at a simple dataset

Common Data Representations

NA:
Not Available Data

Various other formats exist for missing data

	TreatmentA	TreatmentB
John Smith	NA	2
Jane Doe	16	22
Mary Johnson	3	1

OR

	John Smith	Jane Doe	Mary Johnson
TreatmentA	NA	16	3
TreatmentB	2	11	2

H. Wickam, Journal of Statistical Software, 4 (59), 1-23 (2014)

Data Semantics

Dataset?

- Collection of **values** (quantitative: numbers; qualitative: strings)

- Each value belongs to a **variable** and an **observation**

  - **Variable:** contains all values that measure the same attribute (treatmentA etc.) across units.

    Units: People or objects

  - An **observation** contains all values measured on the same unit (person, John Smith etc.)

Unstructured (or Untidy) to Tidy

Two different presentation of a table

Completely Crossed experiment

	TreatmentA	TreatmentB
John Smith	NA	2
Jane Doe	16	22
Mary Johnson	3	1

NA: Not Available

Two different presentation of a table

Variables

Name	TreatmentA	TreatmentB
John Smith	NA	2
Jane Doe	16	22
Mary Johnson	3	1

Values

Observation 1

TreatmentA/TreatmentB are not headers?!!
They are variables

One treatment/person

Tidy

Name	Treatment	Result
John Doe	A	4
Jane Doe	B	1
John Smith	A	NA
John Smith	B	18
Mary Johnson	A	6
Mary Johnson	B	7

Untidy → Tidy

NA: Not Available

- Dataset contains 18 values representing 3 variables and 6 observations.
  - Name: **3** possible **values**
    - John Smith, Jane Doe and Mary Johnson
  - Treatment with **2** possible **values**
    - A or B
  - Result with **6** possible **values**
    - NA, 16, 3, 2, 11 and 1

Name	Treatment	Result
John Doe	A	4
Jane Doe	B	1
John Smith	A	NA
John Smith	B	18
Mary Johnson	A	6
Mary Johnson	B	7

Issues

- **What are observations/variables?**
  - Sometimes difficult to identify

- **Missing variables (we call NAs here)**
  - Really missing or not possible

Rules Tidy Data

Treatment
A
B
A
B
A
B

1. Each column should represent a single variable (i.e., field)
2. Each row should represent a single observation (i.e., record)

Jane Doe	B	1

3. Each cell should contain only one (single) value
   - ❑ 20/120000

Interrelated relationship—Not possible to satisfy only two without satisfying the other one.

Tidy Rules

- Closely follows Codd's Third norm (3NF)
  - https://en.wikipedia.org/wiki/Third_normal_form#Definition_of_third_normal_form

Code Book or Data documentation/dictionary

- Information about the variables
  - Units etc
- Study design information
- Additional assumptions
  - Created new groups (low-income, mid-income, high-income groups etc.)

Code Book or Data documentation/dictionary

- **What kind of study was this (prospective study)?**
- **Where was this carried out?**
- **When did the study began?**
- **Where was the study carried out?**
- **How many participants and other details.**

Variable	Description	Units	Range or Count
SEX	Participants sex	1 = Male 2 = Female	n = 3500 n = 3800
Age	Number of days since last exam		0-2300

Untidy → Tidy

✓Each cell contains single value

✓Each variable forms a column

X Each observation forms a row

	TreatmentA	TreatmentB
John Smith	NA	2
Jane Doe	16	22
Mary Johnson	3	1

Each cell contains single value

Each variable forms a column

Each observation forms a row

Name	Treatment	Result
Jane Doe	A	4
Jane Doe	B	1
John Smith	A	NA
John Smith	B	18
Mary Johnson	A	6
Mary Johnson	B	7

More Examples of Tidying

Your Turn: How will you tidy this data set?

Unstructured or Untidy

	Habitat		
Species	X	Y	Z
A	0	3	0
B	1	0	2

Structured or Tidy

Species	Habitat	Abundance
A	X	0
A	Y	3
A	Z	0
B	X	1
B	Y	0
B	Z	2

Based on Figure 1 from Baldridge et al, DOI: 10.4033/iee.2013.6b.6.f

Yes, all these are the same data

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

Which one is tidy?

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

country	year	rate
Afghanistan	1999	745/19987071
Afghanistan	2000	2666/20595360
Brazil	1999	37737/172006362
Brazil	2000	80488/174504898
China	1999	212258/1272915272
China	2000	213766/1280428583

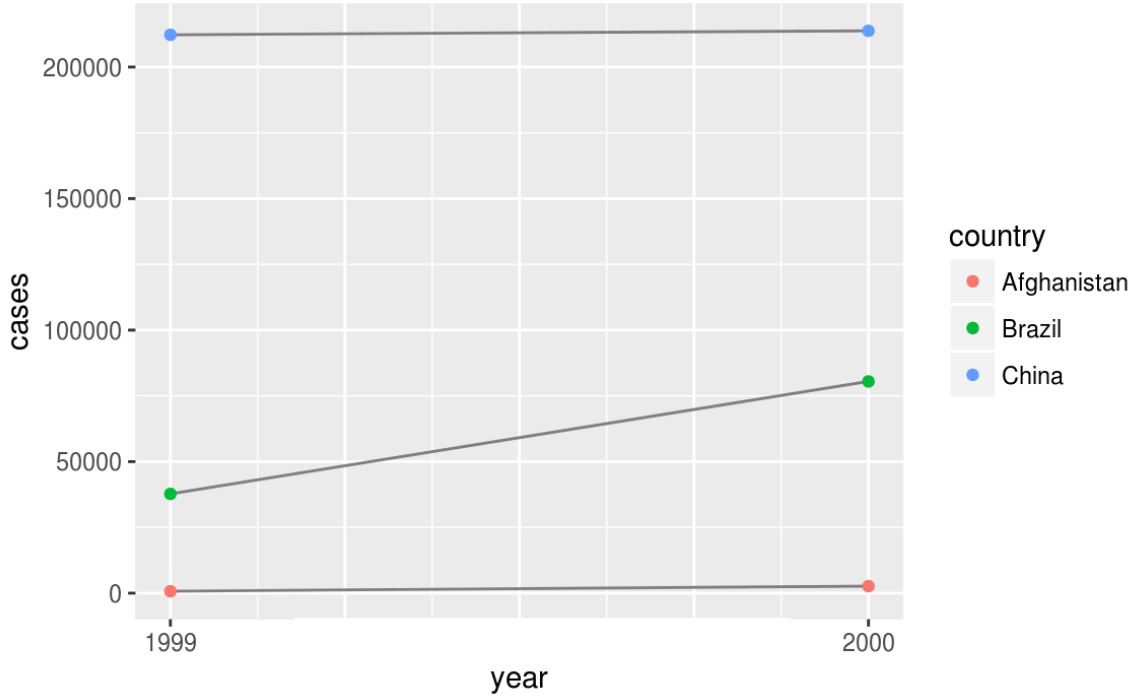
H. Wickam, Journal of Statistical Software, 4 (59), 1-23 (2014)

Easy to Compute/Plot

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

rate = (cases/population) * 10000

country	year	cases	population	rate
Afghanistan	1999	745	19987071	0.3727
Afghanistan	2000	2666	20595360	1.2945
Brazil	1999	37737	172006362	2.1939
Brazil	2000	80488	174504898	4.6124
China	1999	212258	1272915272	1.6675
China	2000	213766	1280428583	1.6695



n = # of cases

year	n
1999	250740
2000	296920

How can we tidy this table?

Retention times of the Conventional HPLC column (T1)
Online SPE system (T2)

Each row should represent a single observation (i.e., record)

Drugs	T1 (min)	T2 (min)
Phenylbutazon	3.490	5.000
Warfarin Sodium	3.147	4.773
Dicoumarol	3.120	4.237
Estriol	5.317	6.177
Piroxicam	3.947	5.013
Amitriptyline Hydrochloride	9.410	9.657
Fluoxetine Hydrochloride	6.310	6.800
Norfluoxetine Hydrochloride	3.400	4.470
Estradiol	5.113	6.227
Nimodipine	4.087	5.277
Omeprazole	5.307	6.450
Diclofenac Sodium	4.947	6.017
Indomethacin	5.157	6.303

How many variables are in this table?
More than one answer!!

Modified Table 2 from Sci Rep, 2015; 5, 14866, 10.1038/srep14866

Retention times of the Conventional HPLC column (T1) Online SPE system (T2)

Drugs	T1 (min)	T2 (min)
Phenylbutazon	3.490	5.000
Warfarin Sodium	3.147	4.773
Dicoumarol	3.120	4.237
Estriol	5.317	6.177
Piroxicam	3.947	5.013
Amitriptyline Hydrochloride	9.410	9.657
Fluoxetine Hydrochloride	6.310	6.800
Norfluoxetine Hydrochloride	3.400	4.470
Estradiol	5.113	6.227
Nimodipine	4.087	5.277
Omeprazole	5.307	6.450
Diclofenac Sodium	4.947	6.017
Indomethacin	5.157	6.303

Drugs	Time(min)	values
Phenylbutazon	T1	3.49
Warfarin Sodium	T1	3.15
Dicoumarol	T1	3.12
Estriol	T1	5.32
Piroxicam	T1	3.95
Amitriptyline Hydrochloride	T1	9.41
Fluoxetine Hydrochloride	T1	6.31
Norfluoxetine Hydrochloride	T1	3.40
Estradiol	T1	5.11
Nimodipine	T1	4.09

... with 42 more rows

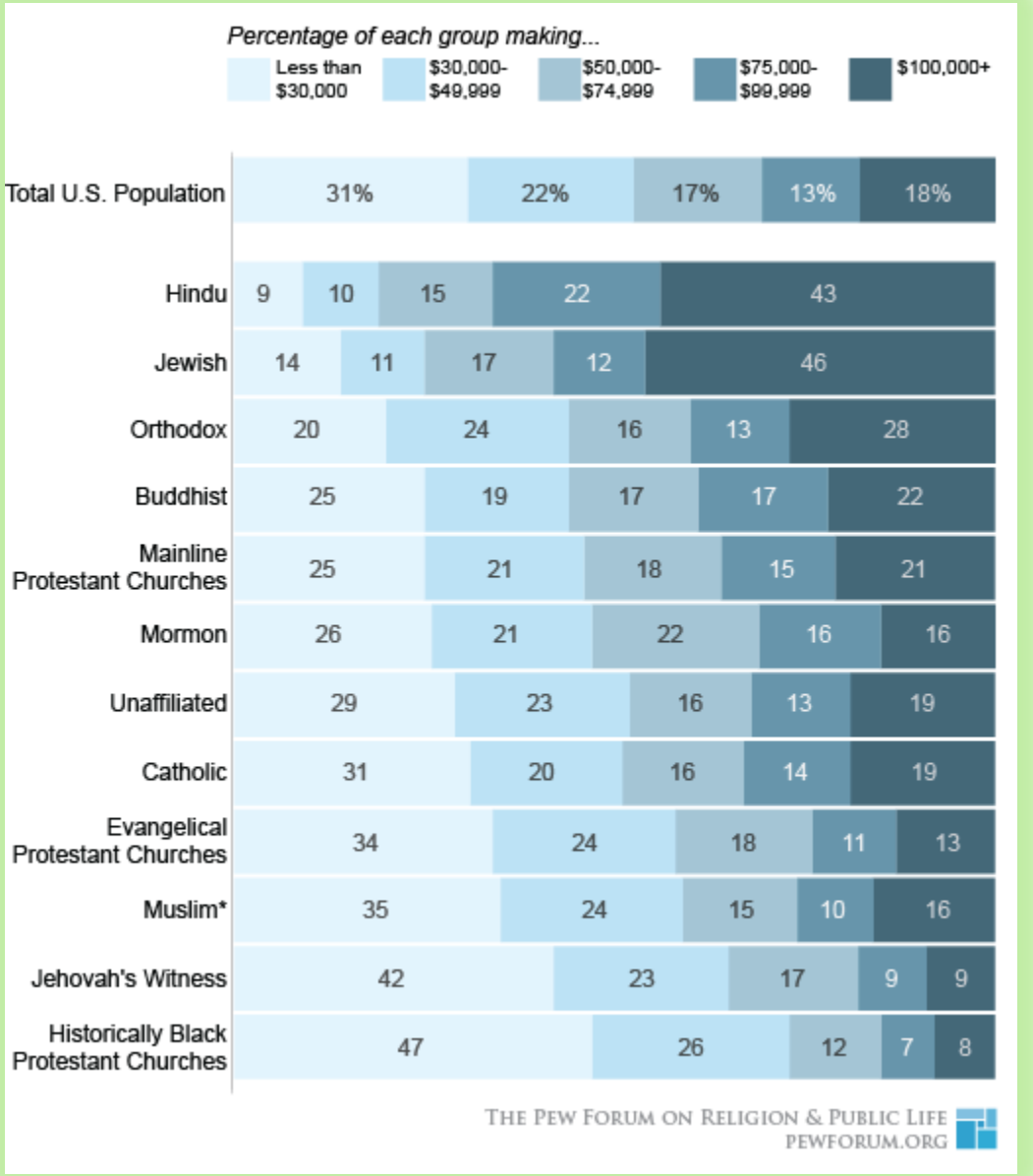
Tidying Unstructured Datasets

Specific issues and how to tidy them

http://www.pewforum.org/2009/01/30/income-distribution-within-us-religious-groups/

Column headers are values, not variable names

- Income distributions within U.S. Religious groups



Example based on H. Wickam, Journal of Statistical Software, 4 (59), 1-23 (2014)

Column headers are values, not variable names

Only first three lines of the dataset shown

Religion	<$10k	<10-20k	$20-30k	$30-40k	$40-50k	$50-75k	$75-100k	$100-150k	>$150k	Refused/Don't know
Agnostic	27	34	60	81	76	137	122	109	84	96
Atheist	12	27	37	52	35	70	73	59	74	76
Buddist	27	21	30	34	33	58	62	39	53	54

- Type of data representation is good for presentations but not good for analysis

How many variables are in this table?

Only first three lines of the dataset shown

Religion	<$10k	<10-20k	$20-30k	$30-40k	$40-50k	$50-75k	$75-100k	$100-150k	>$150k	Refused/Don't know
Agnostic	27	34	60	81	76	137	122	109	84	96
Atheist	12	27	37	52	35	70	73	59	74	76
Buddist	27	21	30	34	33	58	62	39	53	54

Tidying

Religion	Income	freq
Agnostic	<$10k	27
Agnostic	$10-20k	34
Agnostic	$20-30k	60
Agnostic	$30-40k	81
Agnostic	$40-50k	76
Agnostic	$50-75k	137
Agnostic	$75-100k	122
Agnostic	$100-150k	109
Agnostic	>$150k	84
Agnostic	Don't know/refused	96

To tidy this table, we have to stack (melt) the data.
This means turning the **columns** into **rows.**
Making WIDE datasets → Long or Tall

Multiple variables are stored in one column

country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014	f1524	f2534
AD	2000	0	0	1	0	0	0	0	NA	f014	NA	NA
AE	2000	2	4	4	6	5	12	10	NA	f014	16	1
AF	2000	52	228	183	149	129	94	80	NA	f014	414	565
AG	2000	0	0	0	0	0	0	1	NA	f014	1	1
AL	2000	2	19	21	14	24	19	16	NA	f014	11	10
AM	2000	2	152	130	131	63	26	21	NA	f014	24	27
AN	2000	0	0	1	2	0	0	0	NA	f014	0	1
AO	2000	186	999	1003	912	482	312	194	NA	f014	1142	1091
AR	2000	97	278	594	402	419	368	330	NA	f014	544	479
AS	2000	NA	NA	NA	NA	1	1	NA	NA	f014	NA	NA
AT	2000	1	17	30	59	42	23	41	NA	f014	11	22
AU	2000	3	16	35	25	24	19	49	NA	f014	15	19
AZ	2000	0	9	24	33	42	30	0	NA	f014	3	3
BA	2000	4	56	82	99	66	58	77	NA	f014	30	46
BB	2000	0	0	0	2	0	0	0	NA	f014	0	1
BD	2000	256	3640	5643	5750	4718	3667	2837	NA	f014	3029	3238

TB dataset (modified by Hadley); originally from WHO

m014: Male years 0-14
m1524: Male Years 15-24
Similar entries for Female.
The dataset is restricted for Year 2000

Note the zeros and NA
Sometimes they mean different things in the dataset

How many variables are in this table?

sex and age were under one column (also called column)

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

(a) Molten data

country	year	sex	age	cases
AD	2000	m	0–14	0
AD	2000	m	15–24	0
AD	2000	m	25–34	1
AD	2000	m	35–44	0
AD	2000	m	45–54	0
AD	2000	m	55–64	0
AD	2000	m	65+	0
AE	2000	m	0–14	2
AE	2000	m	15–24	4
AE	2000	m	25–34	4
AE	2000	m	35–44	6
AE	2000	m	45–54	5
AE	2000	m	55–64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

(b) Tidy data

country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014	f1524	f2534
AD	2000	0	0	1	0	0	0	0	NA	f014	NA	NA
AE	2000	2	4	4	6	5	12	10	NA	f014	16	1
AF	2000	52	228	183	149	129	94	80	NA	f014	414	565
AG	2000	0	0	0	0	0	0	1	NA	f014	1	1
AL	2000	2	19	21	14	24	19	16	NA	f014	11	10
AM	2000	2	152	130	131	63	26	21	NA	f014	24	27
AN	2000	0	0	1	2	0	0	0	NA	f014	0	1
AO	2000	186	999	1003	912	482	312	194	NA	f014	1142	1091
AR	2000	97	278	594	402	419	368	330	NA	f014	544	479
AS	2000	NA	NA	NA	NA	1	1	NA	NA	f014	NA	NA
AT	2000	1	17	30	59	42	23	41	NA	f014	11	22
AU	2000	3	16	35	25	24	19	49	NA	f014	15	19
AZ	2000	0	9	24	33	42	30	0	NA	f014	3	3
BA	2000	4	56	82	99	66	58	77	NA	f014	30	46
BB	2000	0	0	0	2	0	0	0	NA	f014	0	1
BD	2000	256	3640	5643	5750	4718	3667	2837	NA	f014	3029	3238

Tidy

country	year	sex	age	cases
AD	2000	m	0–14	0
AD	2000	m	15–24	0
AD	2000	m	25–34	1
AD	2000	m	35–44	0
AD	2000	m	45–54	0

Note the final table is more meaningful

Country, Year, Sex, Age are all fixed

Cases are unknown

Variables are stored in both rows and columns

- Daily temperatures in Cuernavaca, Mexico (2010)

Column Names: id, year,, month, element, d1-d31 (days 1-31)

TMIN/TMAX: Minimum/Maximum Temperature; ID: Weather station Identifier

NOT ALL NAs are the same:
Not all months have 31 days;
Can these impossible entries be removed?

id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12	d13	d14	d15	d16	d17	d18	d19	d20	d21	d22	d23	d24	d25	d26	d27	d28	d29	d30	d31
MX000017004	2010	1	TMAX	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	278	NA	
MX000017004	2010	1	TMIN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	145	NA	
MX000017004	2010	2	TMAX	NA	273	241	NA	NA	NA	NA	NA	NA	NA	297	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	299	NA	NA	NA	NA	NA	NA	NA	
MX000017004	2010	2	TMIN	NA	144	144	NA	NA	NA	NA	NA	NA	NA	134	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	107	NA	NA	NA	NA	NA	NA	NA	
MX000017004	2010	3	TMAX	NA	NA	NA	NA	321	NA	NA	NA	NA	345	NA	NA	NA	NA	NA	311	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
MX000017004	2010	3	TMIN	NA	NA	NA	NA	142	NA	NA	NA	NA	168	NA	NA	NA	NA	NA	176	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
MX000017004	2010	4	TMAX	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	363	NA	NA	NA	
MX000017004	2010	4	TMIN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	167	NA	NA	NA	
MX000017004	2010	5	TMAX	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	332	NA	NA	NA	
MX000017004	2010	5	TMIN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	182	NA	NA	NA	
MX000017004	2010	6	TMAX	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	280	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	301	NA	
MX000017004	2010	6	TMIN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	175	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	180	NA	
MX000017004	2010	7	TMAX	NA	NA	286	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	299	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
MX000017004	2010	7	TMIN	NA	NA	175	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	165	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
MX000017004	2010	8	TMAX	NA	NA	NA	NA	296	NA	NA	290	NA	NA	NA	NA	298	NA	NA	NA	NA	NA	NA	NA	NA	NA	264	NA	297	NA	NA	NA	280	NA	254
MX000017004	2010	8	TMIN	NA	NA	NA	NA	158	NA	NA	173	NA	NA	NA	NA	165	NA	NA	NA	NA	NA	NA	NA	NA	NA	150	NA	156	NA	NA	NA	153	NA	154
MX000017004	2010	10	TMAX	NA	NA	NA	NA	270	NA	281	NA	NA	NA	NA	NA	NA	295	287	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	312	NA	NA	NA	
MX000017004	2010	10	TMIN	NA	NA	NA	NA	140	NA	129	NA	NA	NA	NA	NA	NA	130	105	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	150	NA	NA	NA	
MX000017004	2010	11	TMAX	NA	313	NA	272	263	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	281	277	NA	NA	NA	
MX000017004	2010	11	TMIN	NA	163	NA	120	79	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	121	142	NA	NA	NA	
MX000017004	2010	12	TMAX	299	NA	NA	NA	NA	278	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
MX000017004	2010	12	TMIN	138	NA	NA	NA	NA	105	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	

How many variables are in this table?

id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12	d13	d14	d15	d16	d17	d18	d19	d20	d21	d22	d23	d24	d25	d26	d27	d28	d29	d30	d31
MX000017004	2010	1	TMAX	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	278	NA
MX000017004	2010	1	TMIN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	145	NA
MX000017004	2010	2	TMAX	NA	273	241	NA	NA	NA	NA	NA	NA	NA	297	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	299	NA	NA	NA	NA	NA	NA	NA	NA

Messy Data

Tidy Data

	id	year	month	day	tmax	tmin
1	MX000017004	2010	1	30	278	145
2	MX000017004	2010	2	2	273	144
3	MX000017004	2010	2	3	241	144
4	MX000017004	2010	2	11	297	134
5	MX000017004	2010	2	23	299	107
6	MX000017004	2010	3	5	321	142
7	MX000017004	2010	3	10	345	168
8	MX000017004	2010	3	16	311	176
9	MX000017004	2010	4	27	363	167
10	MX000017004	2010	5	27	332	182
11	MX000017004	2010	6	17	280	175
12	MX000017004	2010	6	29	301	180
13	MX000017004	2010	7	3	286	175
14	MX000017004	2010	7	14	299	165
15	MX000017004	2010	8	5	296	158
16	MX000017004	2010	8	8	290	173
17	MX000017004	2010	8	13	298	165
18	MX000017004	2010	8	23	264	150
19	MX000017004	2010	8	25	297	156
20	MX000017004	2010	8	29	280	153
21	MX000017004	2010	8	31	254	154
22	MX000017004	2010	10	5	270	140
23	MX000017004	2010	10	7	281	129
24	MX000017004	2010	10	14	295	130
25	MX000017004	2010	10	15	287	105
26	MX000017004	2010	10	28	312	150
27	MX000017004	2010	11	2	313	163
28	MX000017004	2010	11	4	272	120
29	MX000017004	2010	11	5	263	79
30	MX000017004	2010	11	26	281	121
31	MX000017004	2010	11	27	277	142
32	MX000017004	2010	12	1	299	138
33	MX000017004	2010	12	6	278	105

Real-life example based on a question from a NCI investigator

Anonymized data

Maximal response for effect A 1000 nM (Receptor1)	Maximal response for effect A at lower dose 100 nM (Receptor1)	Maximal response for effect B 1000 nM (Receptor2)	Maximal response for effect B at lower dose 100 nM (Receptor2)
2.4459	0.9793	1.7922	1.7534
1.5752	0.0608	0.8562	-1.2930
2.6959	0.8674	2.0225	1.0184
1.2401	0.6337	3.7147	0.5151
4.5276	2.0354	2.9920	-0.8623
2.4625		0.9328	-0.7610
3.5255		1.6204	
2.3753			

Maximal responses were determined with an average +/- SEM

SEM: Standard deviation of the mean

- Two different biological responses each with two step dose response curve (Drug concentrations: 100nM and 1000nM)
- The 1000 nM drug concentration gives the maximum dose response for both effects
  - Note the absolute response is different

- Question?
  - If the 100 nM response for effect A is different from that of effect B

Maximal response for effect A 1000 nM (Receptor1)	Maximal response for effect A at lower dose 100 nM (Receptor1)	Maximal response for effect B 1000 nM (Receptor2)	Maximal response for effect B at lower dose 100 nM (Receptor2)
2.4459	0.9793	1.7922	1.7534
1.5752	0.0608	0.8562	-1.2930
2.6959	0.8674	2.0225	1.0184
1.2401	0.6337	3.7147	0.5151
4.5276	2.0354	2.9920	-0.8623
2.4625		0.9328	-0.7610
3.5255		1.6204	
2.3753			

Untidy

Tidy

Group	Drug Concentration (nM)	Response
A	1000	2.4459
A	1000	1.5752
A	1000	2.6959
A	1000	1.2401
A	1000	4.5276
A	1000	2.4625
A	1000	3.5255
A	1000	2.3753
A	100	0.9793
A	100	0.0608
A	100	0.8674
A	100	0.6337
A	100	2.0354
B	1000	1.7922
B	1000	0.8562
B	1000	2.0225
B	1000	3.7147
B	1000	2.9920
B	1000	0.9328
B	1000	1.6204
B	100	1.7534
B	100	-1.2930
B	100	1.0184
B	100	0.5151
B	100	-0.8623
B	100	-0.7610

Another fun example

Total number of words spoken by each Race and Gender

Movie 1

The Fellowship Of The Ring

Race	Female	Male
Elf	1229	971
Hobbit	14	3644
Man	0	1995

Movie 2

The Two Towers

Race	Female	Male
Elf	331	513
Hobbit	0	2463
Man	401	3589

Movie 3

The Return of the King

Race	Female	Male
Elf	183	510
Hobbit	2	2673
Man	268	2459

Example from Dr. Jenny Bryan, Univ of British Columbia

https://www.stat.ubc.ca/~jenny/

Questions on Untidy dataset

Movie 1

The Fellowship Of The Ring

Race	Female	Male
Elf	1229	971
Hobbit	14	3644
Man	0	1995

- Good Format for report??
- Violates most of the Tidy rules, Column values are not really # of Females but the number of words spoken by Females
- If we have to compute the total number of words spoken by Male Elf (for example) from all the movies, we have to spend some time extracting data from tables. Do you think it is easy?
- Want to add more data, how easy is it?

Let us tidy the dataset

Analysis becomes very easy

All three tables were merged into one table

We are ready to answer questions like:

What race spoke most words in a movie?

How different is the dominant race in "The Two Towers" compared to the "The Return of the King" ?

Example from Dr. Jenny Bryan, Univ of British Columbia

https://www.stat.ubc.ca/~jenny/

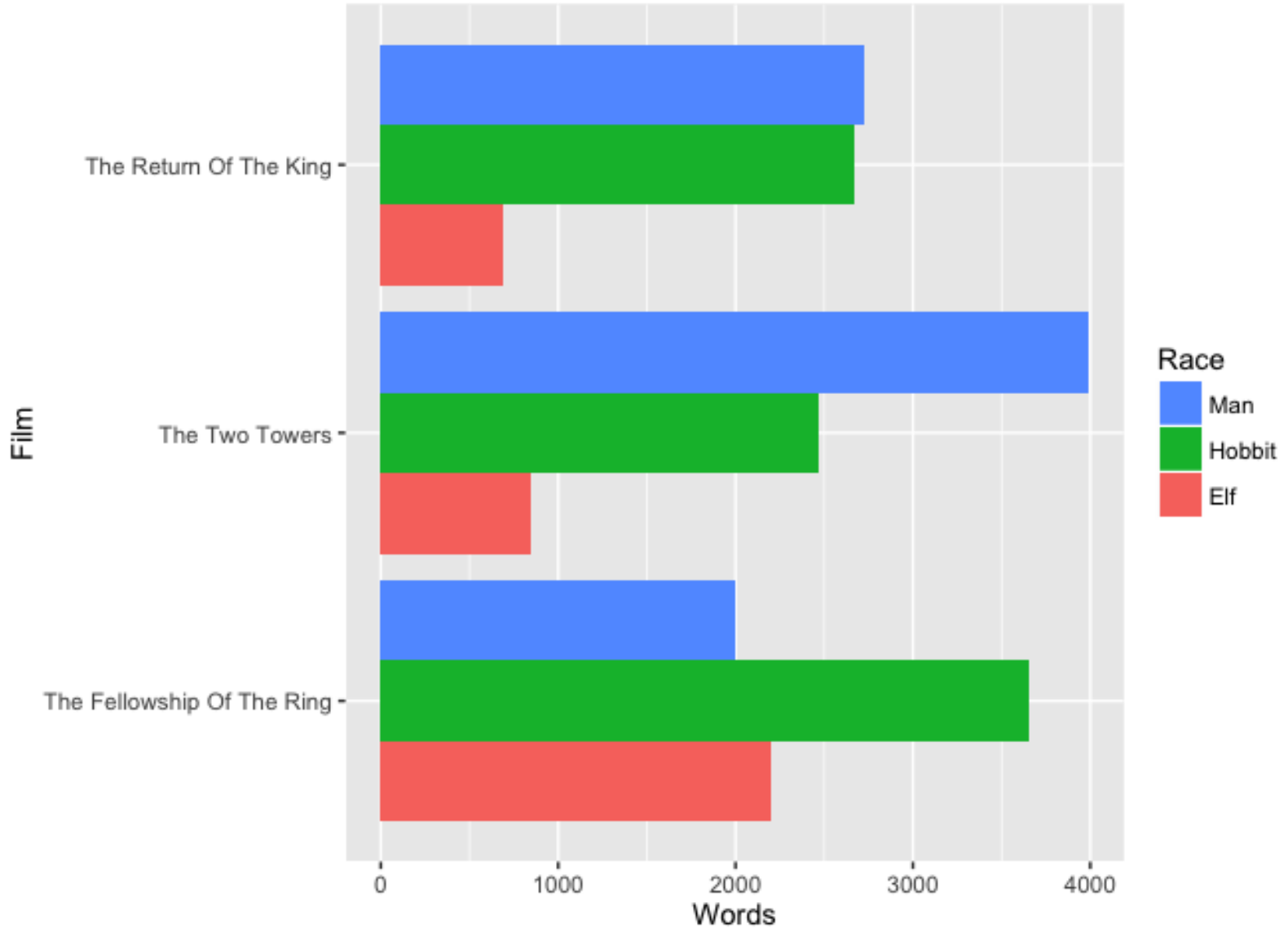
Table1 Table2 Table3

Tidying

Film	Gender	Race	Words
The Fellowship Of The Ring	Female	Elf	1229
The Fellowship Of The Ring	Male	Elf	971
The Fellowship Of The Ring	Female	Hobbit	14
The Fellowship Of The Ring	Male	Hobbit	3644
The Fellowship Of The Ring	Female	Man	0
The Fellowship Of The Ring	Male	Man	1995
The Two Towers	Female	Elf	331
The Two Towers	Male	Elf	513
The Two Towers	Female	Hobbit	0
The Two Towers	Male	Hobbit	2463
The Two Towers	Female	Man	401
The Two Towers	Male	Man	3589
The Return Of The King	Female	Elf	183

Only top 13 rows of the tidy dataset shown

Final benefit is the ease of analysis and creating figures



Example from Dr. Jenny Bryan, Univ of British Columbia

https://www.stat.ubc.ca/~jenny/

What about Biological (expression) data?

High dimensional RNASeq data 64102 x 8; GSE52778

SAMPLES: COL; Features: ROWS

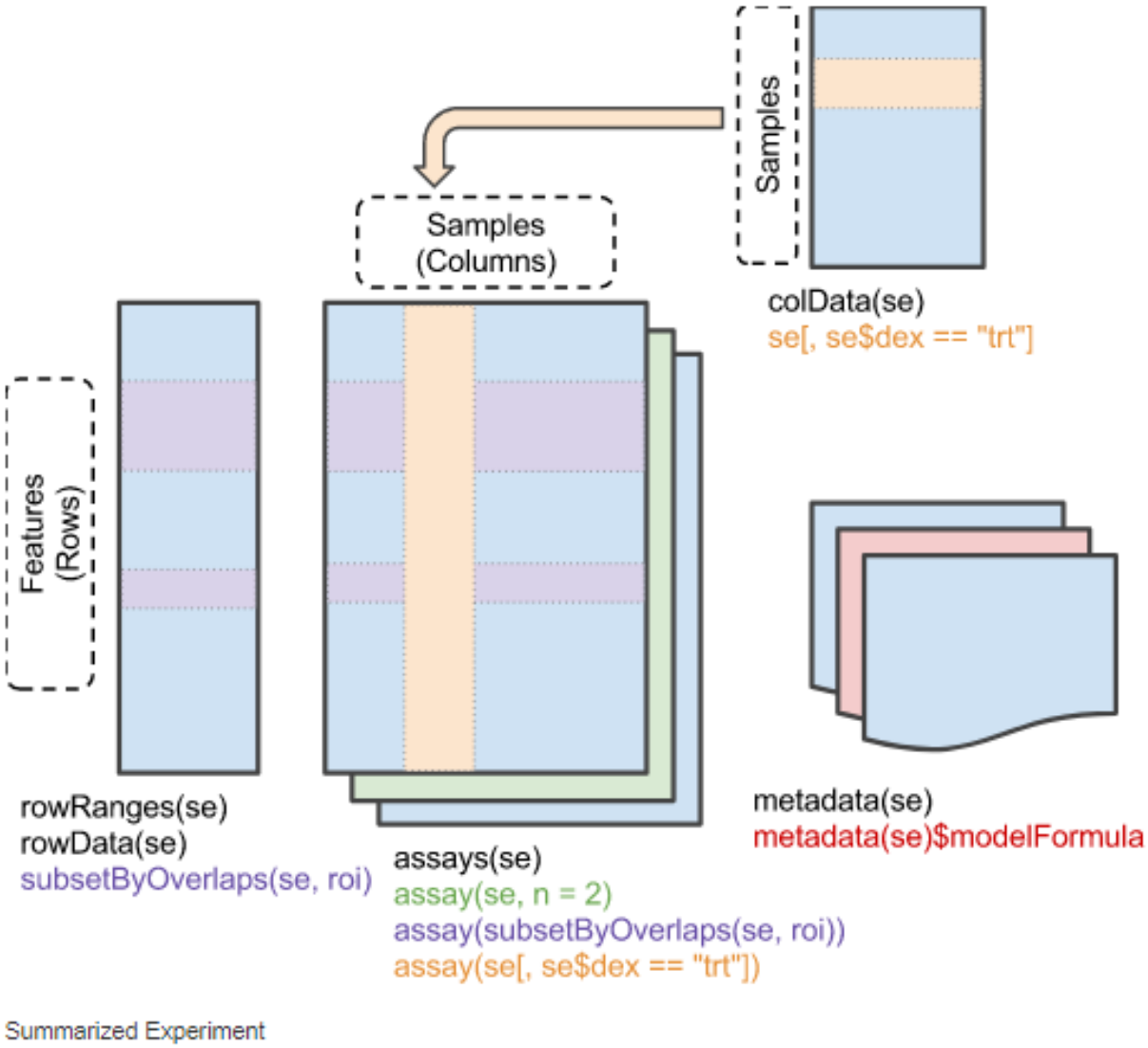
Genes	Assay1	Assay2	Assay3	Assay4	Assay5	Assay6	Assay7	Assay8
ENSG00000000003 TSPAN6	679	448	873	408	1138	1047	770	572
TNMD	0	0	0	0	0	0	0	0
ENSG00000000419 DPM1	467	515	621	365	587	799	417	508
SCYL3	260	211	263	164	245	331	233	229
ENSG00000000460 C1orf112	60	55	40	35	78	63	76	60
FGR	0	0	2	0	1	0	0	0
ENSG00000000971 CFH	3251	3679	6177	4252	6721	11027	5176	7995
FUCA2	1433	1062	1733	881	1424	1439	1359	1109
ENSG00000001084 GCLC	519	380	595	493	820	714	696	704
NFYA	394	236	464	175	658	584	360	269
ENSG00000001460 STPG1	172	168	264	118	241	210	155	177
NIPAL3	2112	1867	5137	2657	2735	2751	2467	2905
ENSG00000001497 LAS1L	524	488	638	357	676	806	493	475
ENPP4	71	51	211	156	23	38	134	172
ENSG00000001617 SEMA3F	555	394	905	415	727	697	618	599
CFTR	10	2	9	2	10	6	5	5
ENSG00000001629 ANKIB1	1660	1251	2259	1079	2462	2514	1888	1660

Summarized Experiment: RNA-Seq experiment of read counts per gene for airway smooth muscles

8 different experimental and assays 64,102 gene transcripts.

High Dimensional Data (8 x 64,102)

Metadata



OPEN ACCESS Freely available online

PLOS ONE

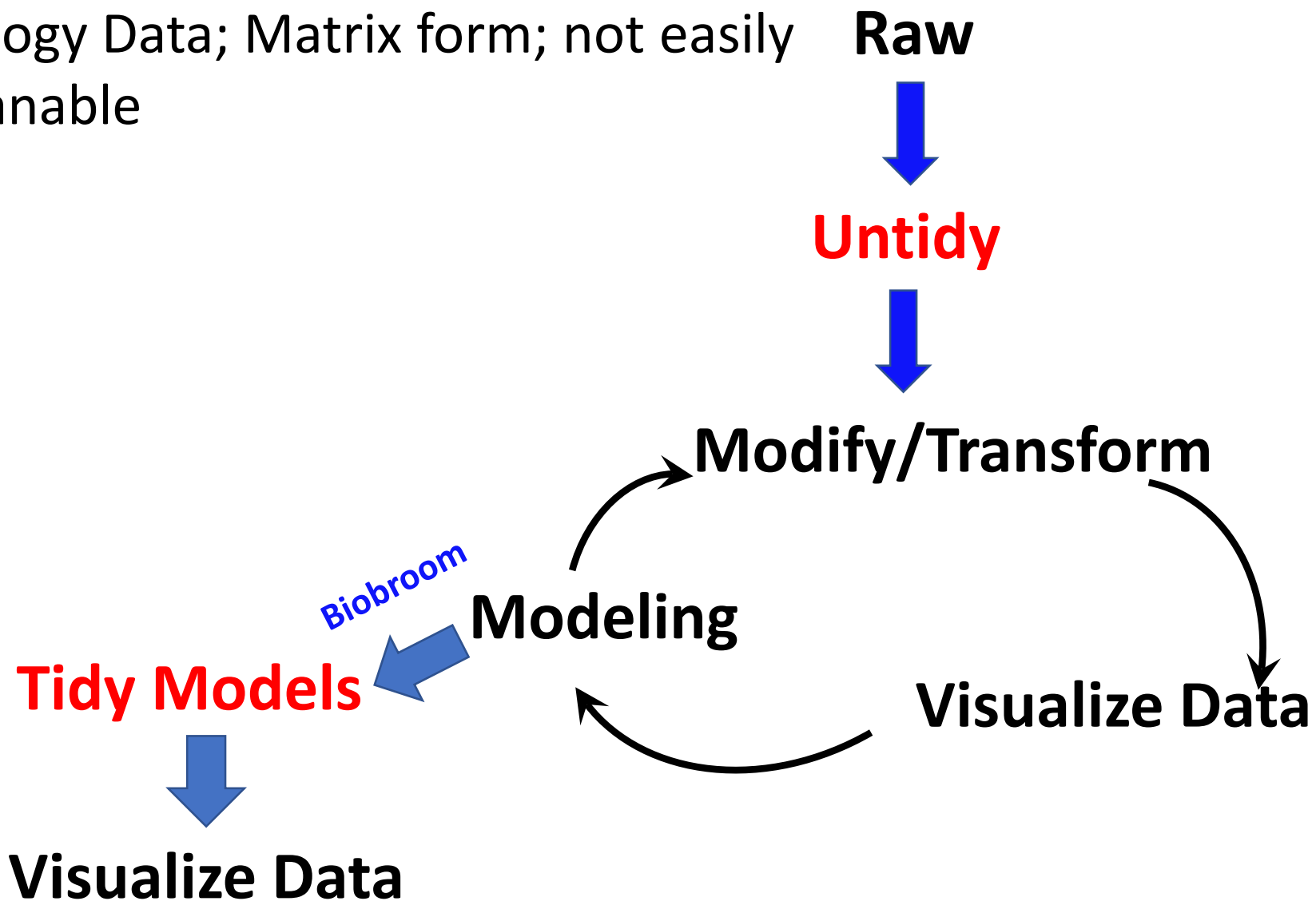
RNA-Seq Transcriptome Profiling Identifies *CRISPLD2* as a Glucocorticoid Responsive Gene that Modulates Cytokine Function in Airway Smooth Muscle Cells

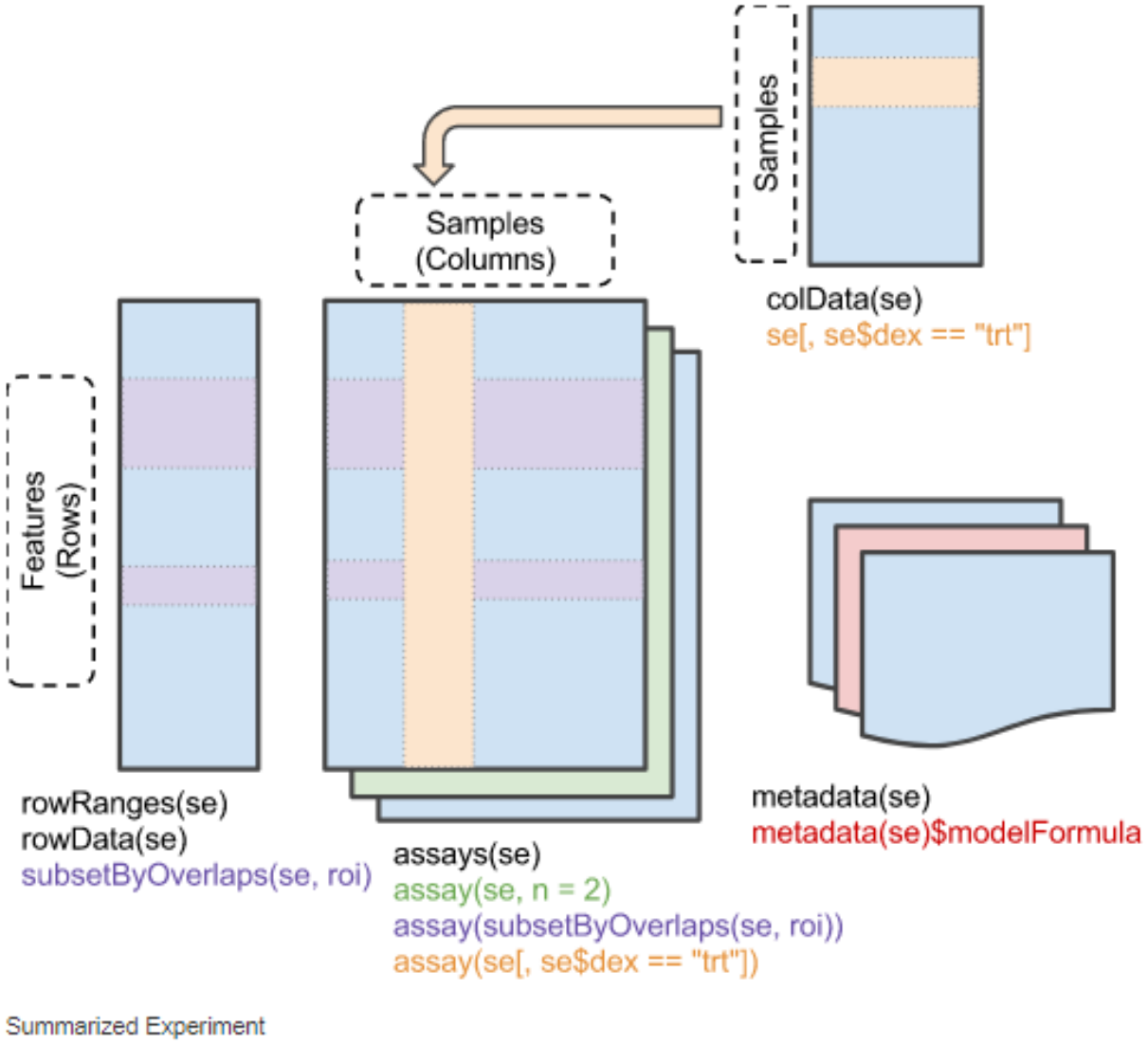
Blanca E. Himes[1,2,3*], Xiaofeng Jiang[4], Peter Wagner[4], Ruoxi Hu[4], Qiyu Wang[4], Barbara Klanderman[2], Reid M. Whitaker[1], Qingling Duan[1], Jessica Lasky-Su[1], Christina Nikolos[5], William Jester[5], Martin Johnson[5], Reynold A. Panettieri Jr.[5], Kelan G. Tantisira[1], Scott T. Weiss[1,2], Quan Lu[4*]

1 Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States of America, **2** Partners HealthCare Personalized Medicine, Boston, Massachusetts, United States of America, **3** Children's Hospital Informatics Program, Boston, Massachusetts, United States of America, **4** Program in Molecular and Integrative Physiological Sciences, Departments of Environmental Health, and Genetics and Complex Diseases, Harvard School of Public Health, Boston, Massachusetts, United States of America, **5** Pulmonary, Allergy and Critical Care Division, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Abstract

Biology Data; Matrix form; not easily cleanable





gene	sample	value
ENSG00000000003	SRR1039508	679
ENSG00000000005	SRR1039508	0
ENSG00000000419	SRR1039508	467
ENSG00000000457	SRR1039508	260
ENSG00000000460	SRR1039508	60
ENSG00000000938	SRR1039508	0
ENSG00000000971	SRR1039508	3251
ENSG00000001036	SRR1039508	1433
ENSG00000001084	SRR1039508	519
ENSG00000001167	SRR1039508	394

... with 512,806 more rows

Summary/Final Thoughts

- Sharing Data
  - Send CSV or text files along with CODE Book
  - Excel: One sheet only
- Tidying removes redundancy
- Easy for analysis
- Reduces chances of error
- May not be suitable for some datasets
  - Biological expression
    - For now use BioConductor ☺

Thanks

ravichandrans@mail.nih.gov