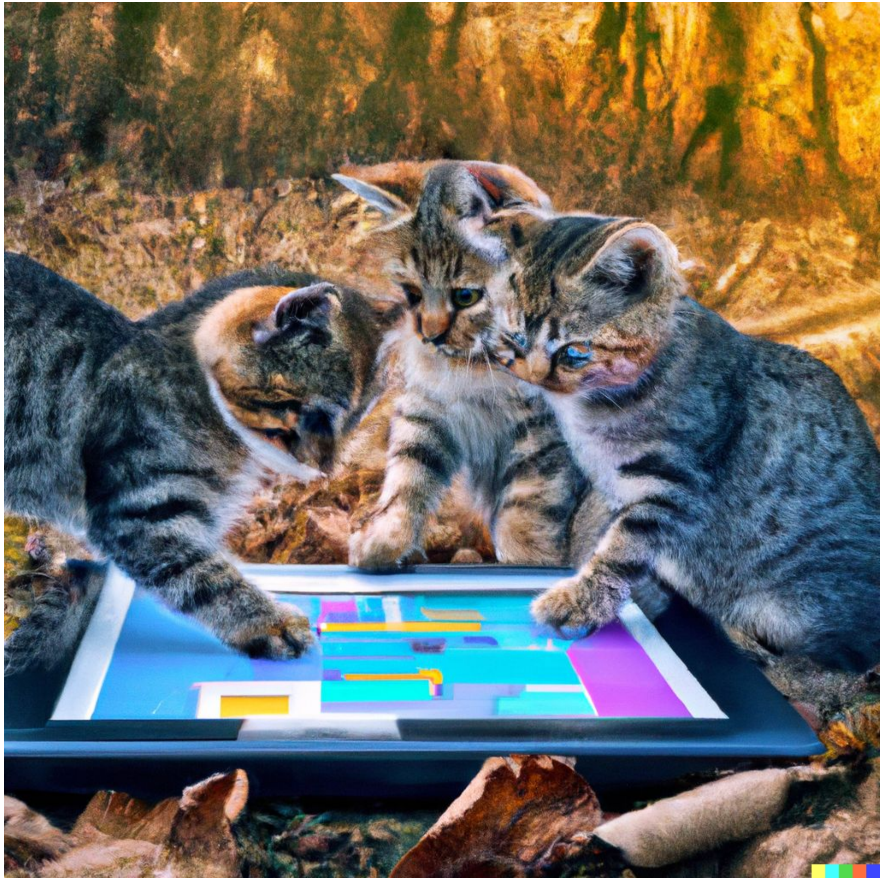
FALL 2023
DSE 12700
VISUAL ANALYTICS

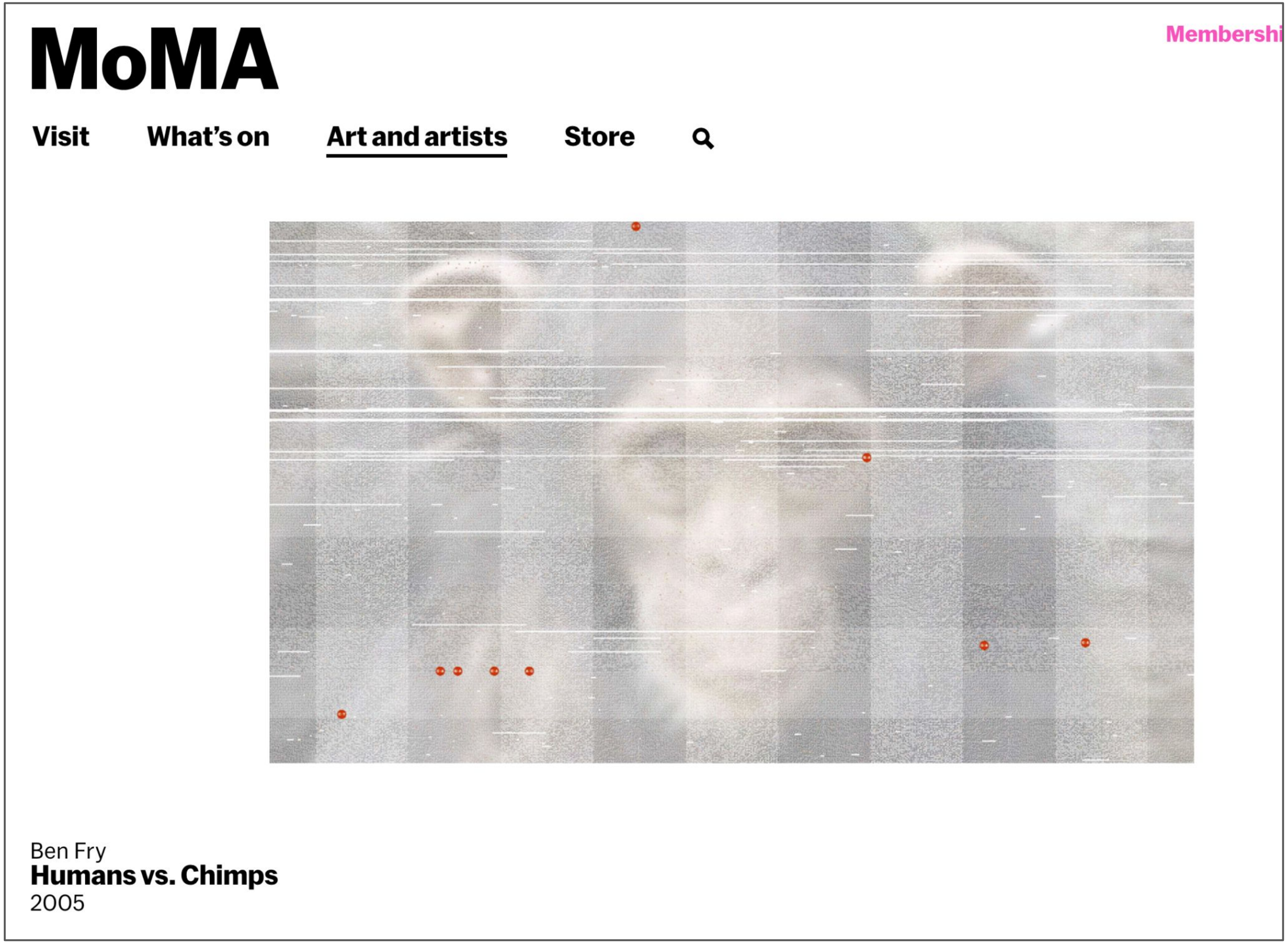
Professor
Madeline Blount
she/her

Week 2



Dall-E2, tabby kittens creating colorful digital charts in a forest, photorealistic style

Ben Fry
MIT, Fathom



Ben Fry
Humans vs. Chimps
2005

https://www.moma.org/collection/works/110354

acquire

live or changing data sources

parse

modular parsers for new data sources

filter

automation of tedious manual processes

modify filter in real-time

mine

modify parameters of statistical methods in real-time

represent

rapid prototyping and iteration

juxtapose large amounts of data

try multiple representations

refine

change design rules without manual redesign

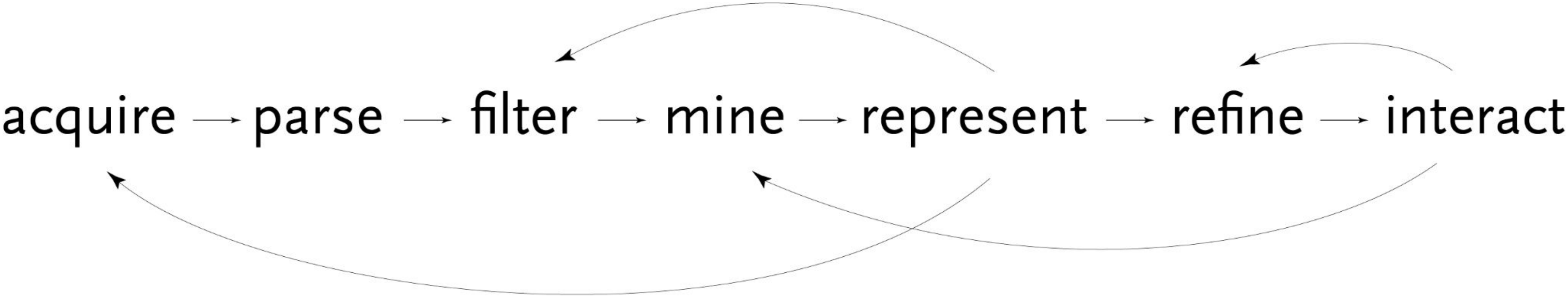
computation as its own “medium”

interact

smooth transition between states to maintain context

additional information as viewpoint shifts

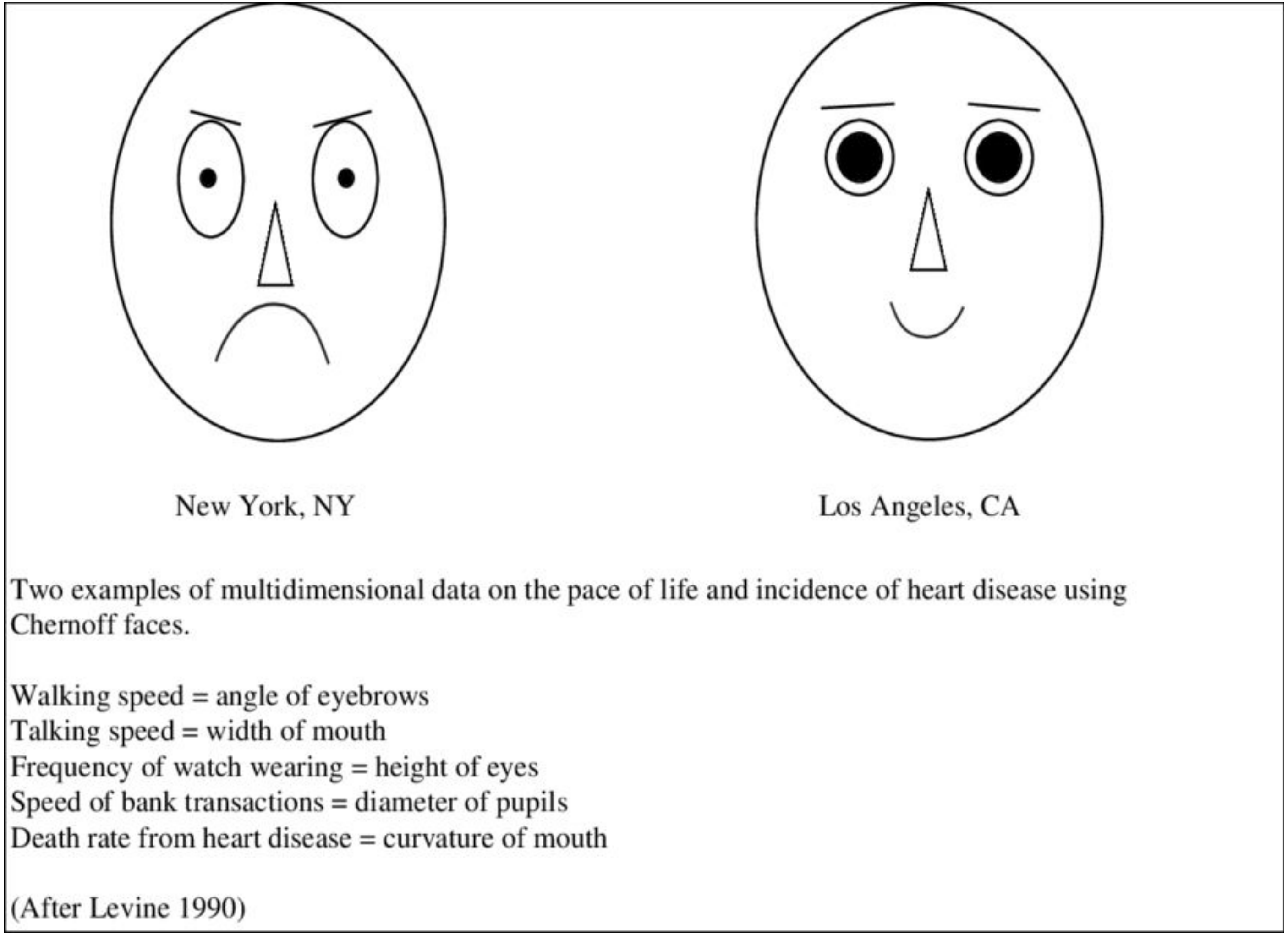
COMPUTER SCIENCE		MATHEMATICS, STATISTICS, AND DATA MINING		GRAPHIC DESIGN		INFOVIS AND HCI
acquire	parse	filter	mine	represent	refine	interact



acquire → parse → filter → mine → represent → refine → interact

“More data is not implicitly better, and often serves to simply confuse the situation … A focus on the question helps define what that minimum requirements are.” - Fry, Ch. 5

“Knowledge of the audience is essential, for knowing what might be appropriate at each step.” - Fry, Ch. 5



Two examples of multidimensional data on the pace of life and incidence of heart disease using Chernoff faces.

Walking speed = angle of eyebrows
Talking speed = width of mouth
Frequency of watch wearing = height of eyes
Speed of bank transactions = diameter of pupils
Death rate from heart disease = curvature of mouth

(After Levine 1990)

https://www.researchgate.net/figure/Chernoff-faces-showing-the-pace-of-life-and-heart-disease_fig3_242246033

Tamara Munzner
University of British Columbia

Domain-independent vocabulary for data visualization

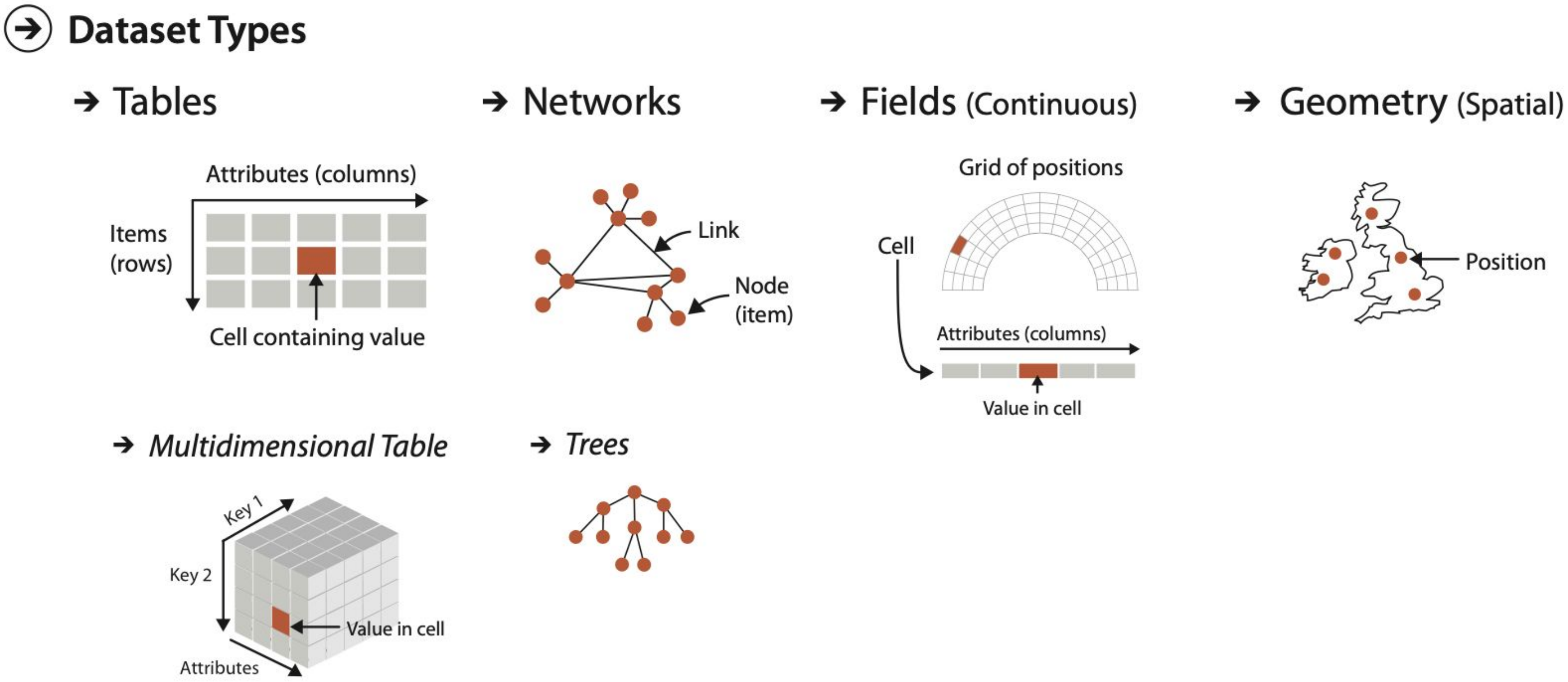
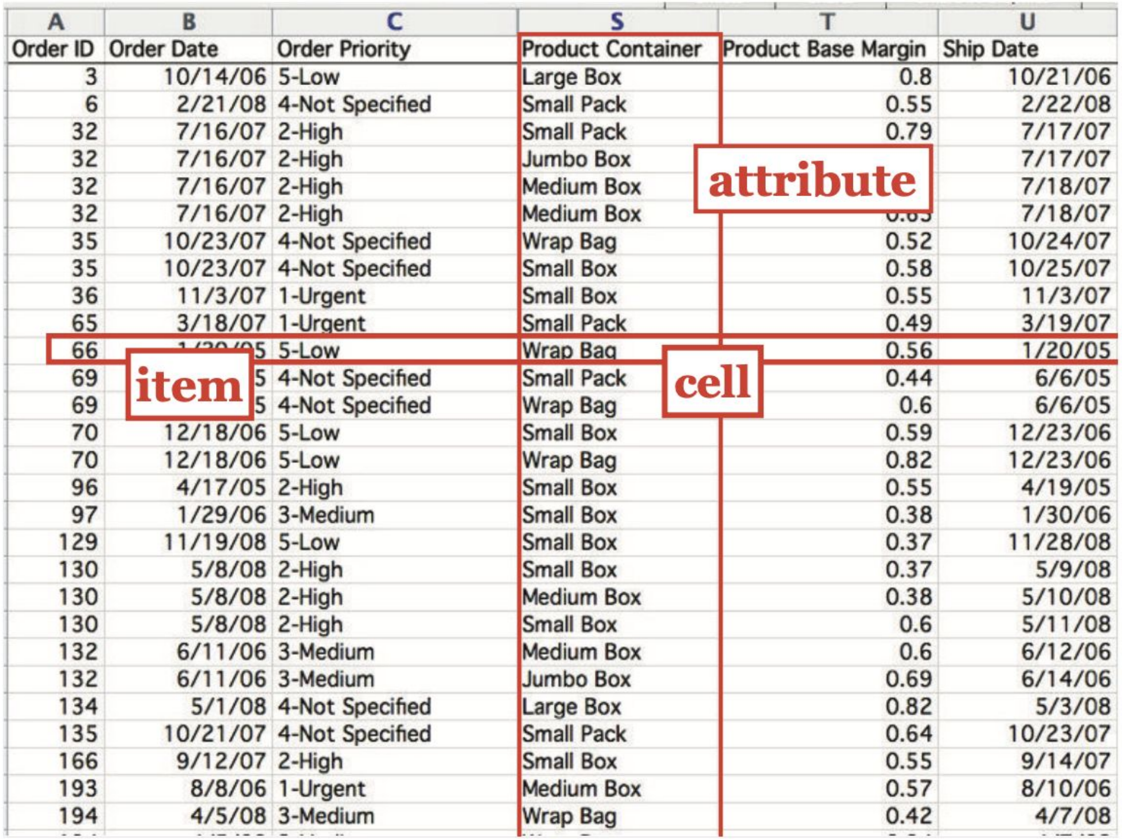


Figure 2.4. The detailed structure of the four basic dataset types.



A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	[illegible]	7/17/07
32	7/16/07	2-High	Medium Box	[illegible]	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	[illegible]	4-Not Specified	Small Pack	0.44	6/6/05
69	[illegible]	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

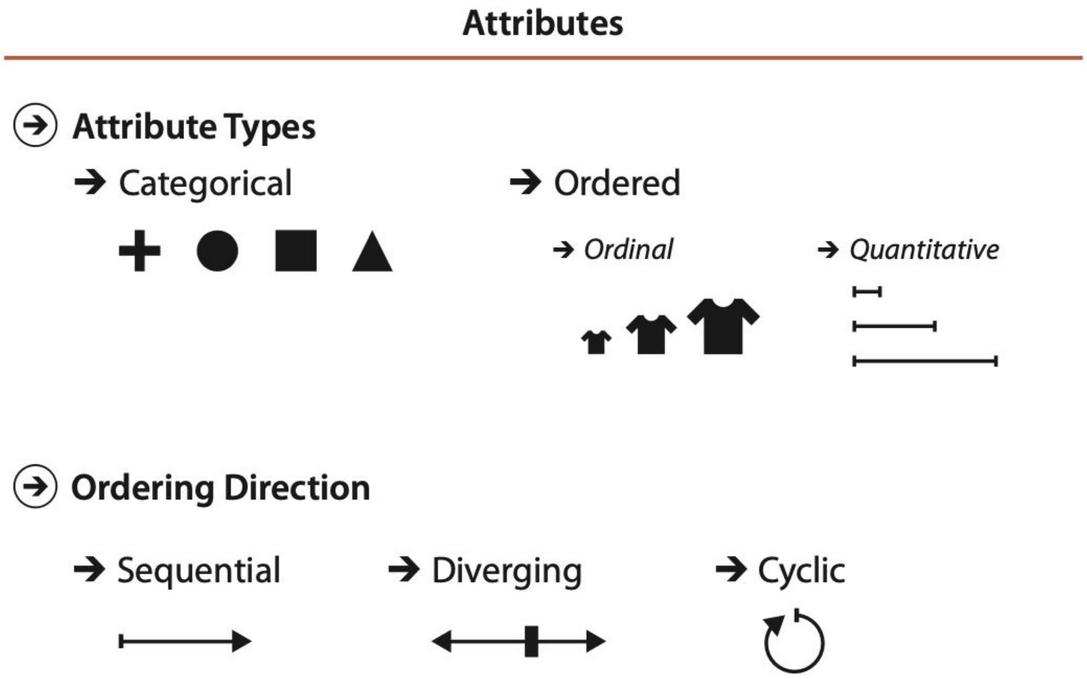
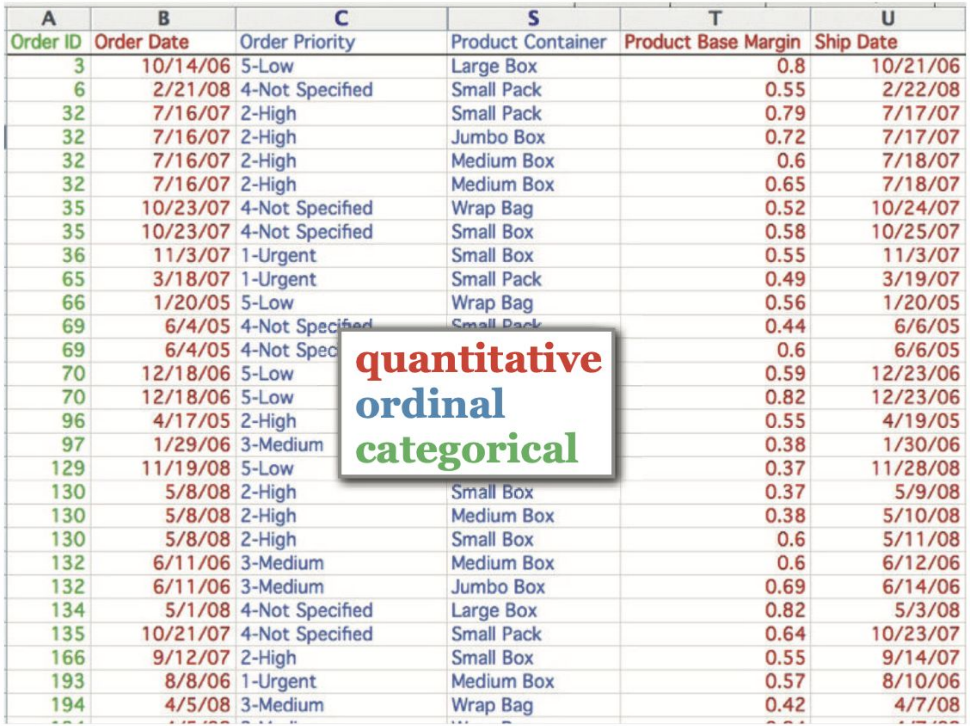
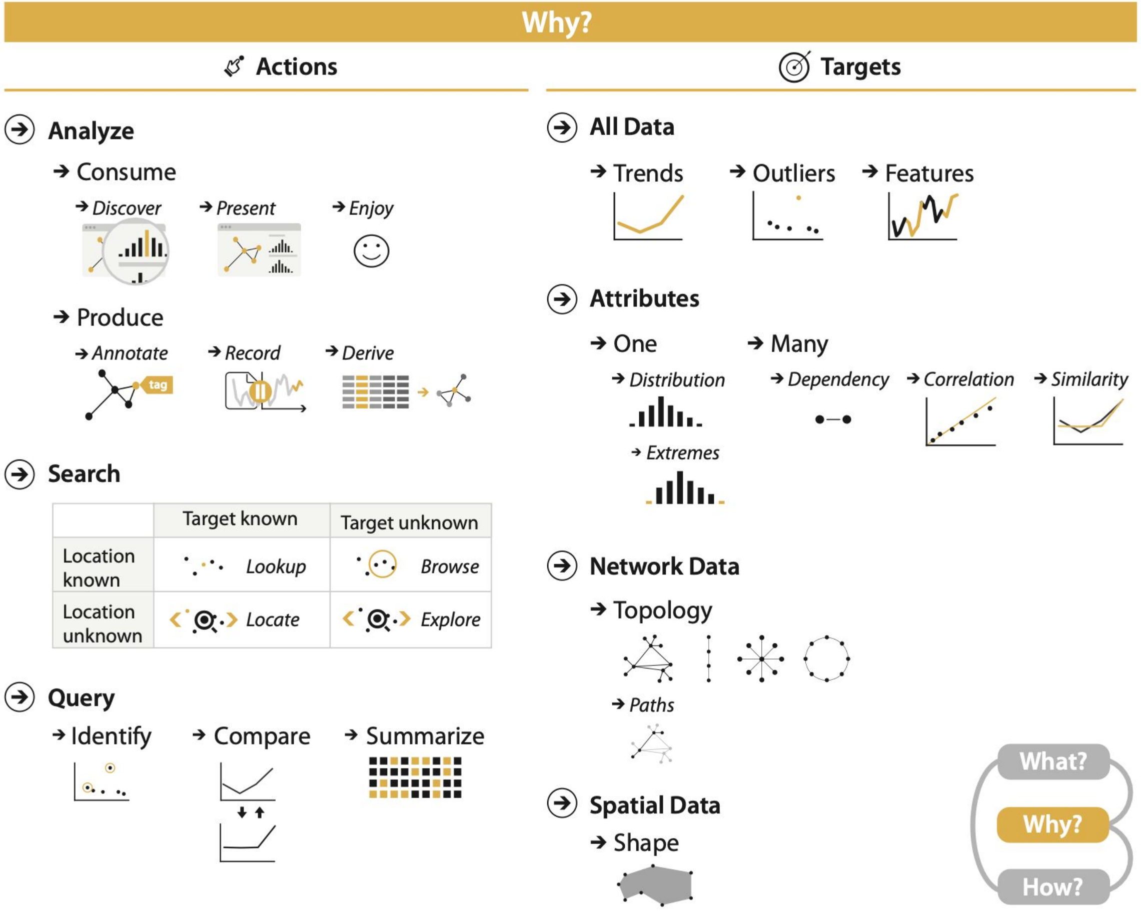


Figure 2.7. Attribute types are categorical, ordinal, or quantitative. The direction of attribute ordering can be sequential, diverging, or cyclic.



A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Spec		0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low		0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low		0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Figure 2.9. The order table with the attribute columns colored by their type; none of them is a key.



	Target known	Target unknown
Location known	Lookup	Browse
Location unknown	Locate	Explore

Figure 3.1. *Why* people are using vis in terms of actions and targets.

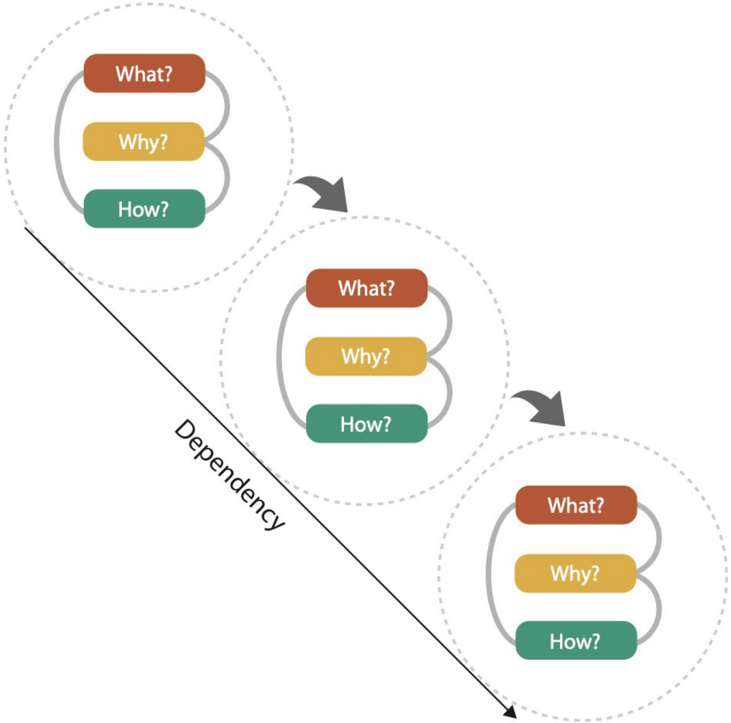
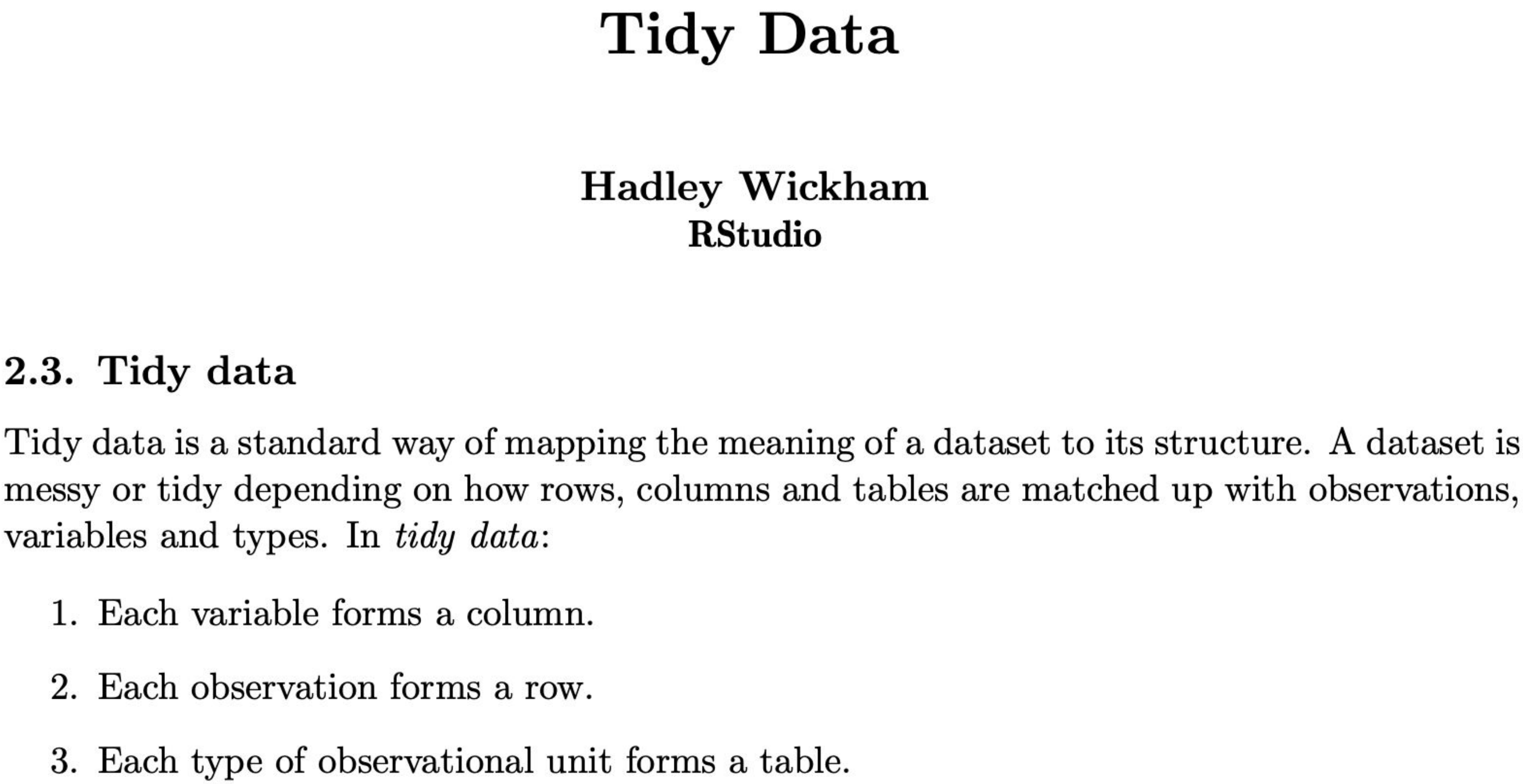


Figure 1.8. Analyzing vis usage as chained sequences of instances, where the output of one instance is the input to another.



Tidy Data

Hadley Wickham
RStudio

2.3. Tidy data

Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched 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.

religion	<$10k	$10–20k	$20–30k	$30–40k	$40–50k	$50–75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

Table 4: The first ten rows of data on income and religion from the Pew Forum. Three columns, $75–100k, $100–150k and >150k, have been omitted.

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

Table 6: The first ten rows of the tidied Pew survey dataset on income and religion. The `column` has been renamed to `income`, and `value` to `freq`.

pandas

- 🐍 **python**

- **Pa**nel **Da**ta, from econometrics data, 2008 (Wes McKinney)

- open source, now run by nonprofit

https://commons.wikimedia.org/wiki/File:6990634-panda-hug.jpg

Meet the Palmer penguins

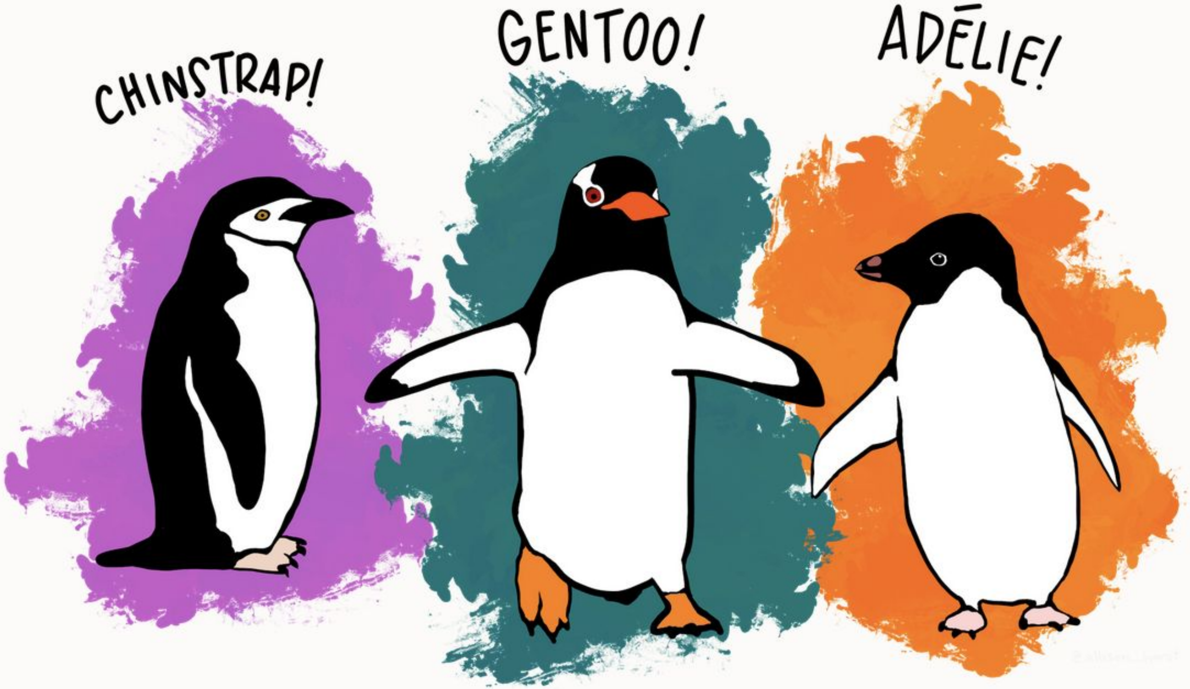
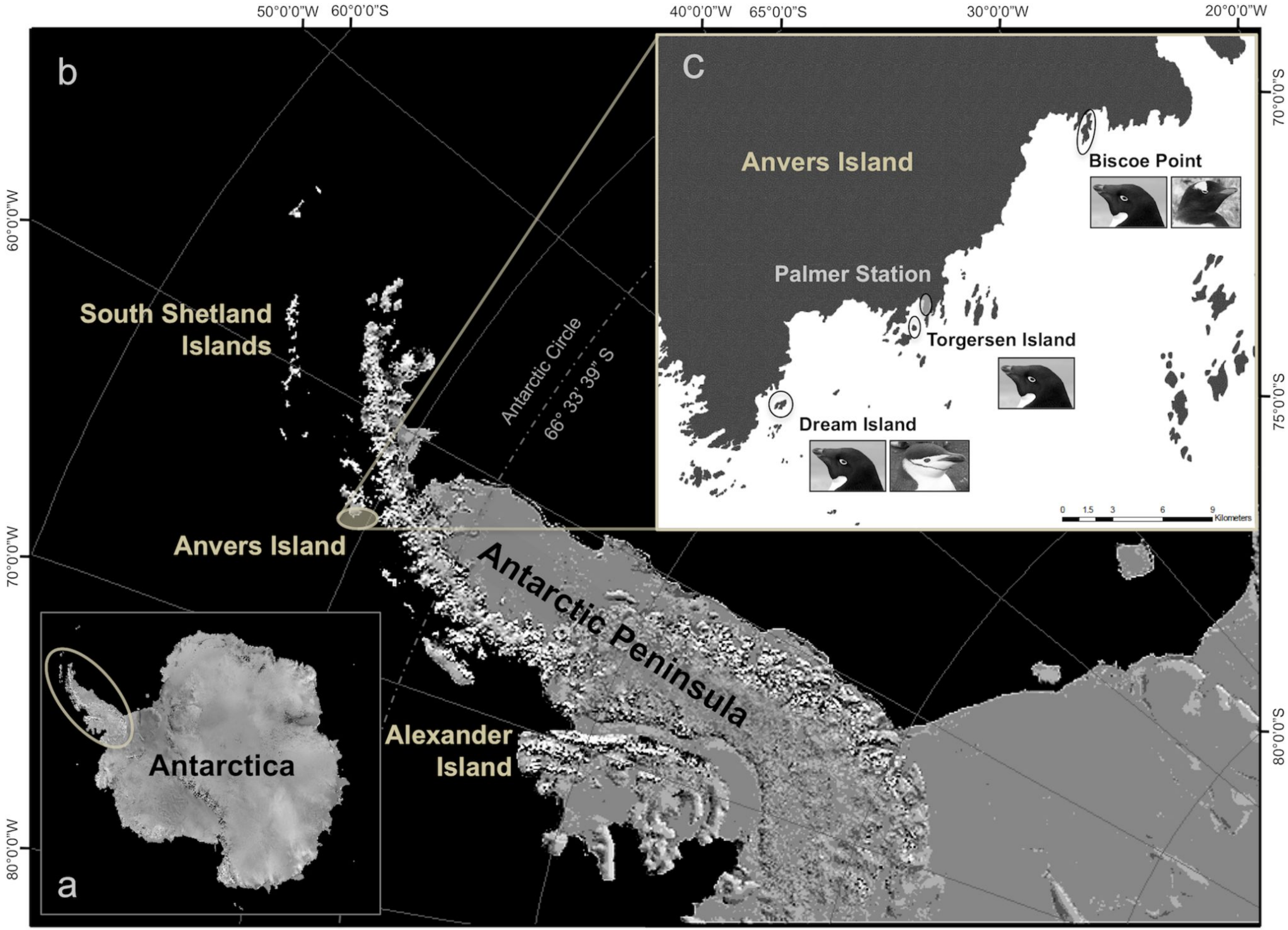


Image: S. Sternbach

Data were collected from 2007-2009 by **Dr. Kristen Gorman** with the [Palmer Station Long Term Ecological Research Program]()

Gorman et. al. made their data public - **Allison Horst and Allison Hill** turned into dataset package



https://www.researchgate.net/figure/Fieldwork-for-the-present-study-took-place-within-the-Palmer-Archipelago-near-Anvers_fig1_260557350

Chinstrap Penguins, Richard Sidley