What’s the Deal with Data?

AN INTRODUCTION TO DATA FOR THE UNCERTAIN AND INTERESTED

MARTIN CHANDLER

MARTIN.CHANDLER@MCGILL.CA

Agenda

- What is Data?
- Why is Data?
- Where is Data?
- How is Data?
- Questions?

What is data?

- Data is facts, in different forms, about a particular subject, used to answer a question or to make an argument.
- This can include numbers (numerical data), information about locations (geospatial data), or words about something (descriptive or otherwise)
- Eg: The number of couples introduced through the pineapple export business is a point of data (being 1, as far as I know)
- The number of dogs living in a particular household is another data point

Some terminology

- Data: facts, or factual information, used for reasoning or analysis
- Dataset: A collection of facts, often gathered together to be manipulated
- Variable: A characteristic that is observed (eg: age, marital status, etc). May be numeric ("33") or descriptive ("common law").
- Case: The unit of analysis (ie each submission: if 1000 surveys were collected, the microdata file will have 1000 "cases").

Variable

Case

Home ID	Postal Code	Dog	Breed	Temperament
001	M6R 1N6	0	N/A	N/A
002	L2S 2A7	1	Beagle	Friendly
003	B2R 1S2	2	Dachshund; Terrier	Cute; Derpy
004	H4H 1V1	0	N/A	N/A

Statistics

- Calculated figures produced using methods developed through modes of inquiry – counts, totals, averages, means, etc.
- Ex: Approval rates for governing party
- Number of households with dogs, in a given area

Microdata

- A dataset with individual pieces of recorded information
- Ex: Individual responses to a survey
- Student test scores by schools
- Exact breakdown of spending for an event

Statistics vs Microdata

Respondents had an average of 0.75 dogs, and a derp factor of 1, or light

Home ID	Postal Code	Dog	Breed	Temperment
001	M6R 1N6	0	N/A	N/A
002	L2S 2A7	1	Beagle	Friendly
003	B2R 1S2	2	Dachshund; Malamute	Cute; Derpy
004	H4H 1V1	0	N/A	N/A

Which should I use, statistics or microdata?

- “How much/many?” vs “Why?”
- Eg: “I need data on the number of Spanish speakers in Montréal.” You want statistics!
- Or: “I’m exploring the relationship between radon and lung cancer in Montréal”. You want microdata!

Statistics vs Microdata

- Why does all of this matter?
- Different platforms and tools for finding
- Different restrictions and requirements for use (e.g. privacy)
- Different methods and software for working with the information

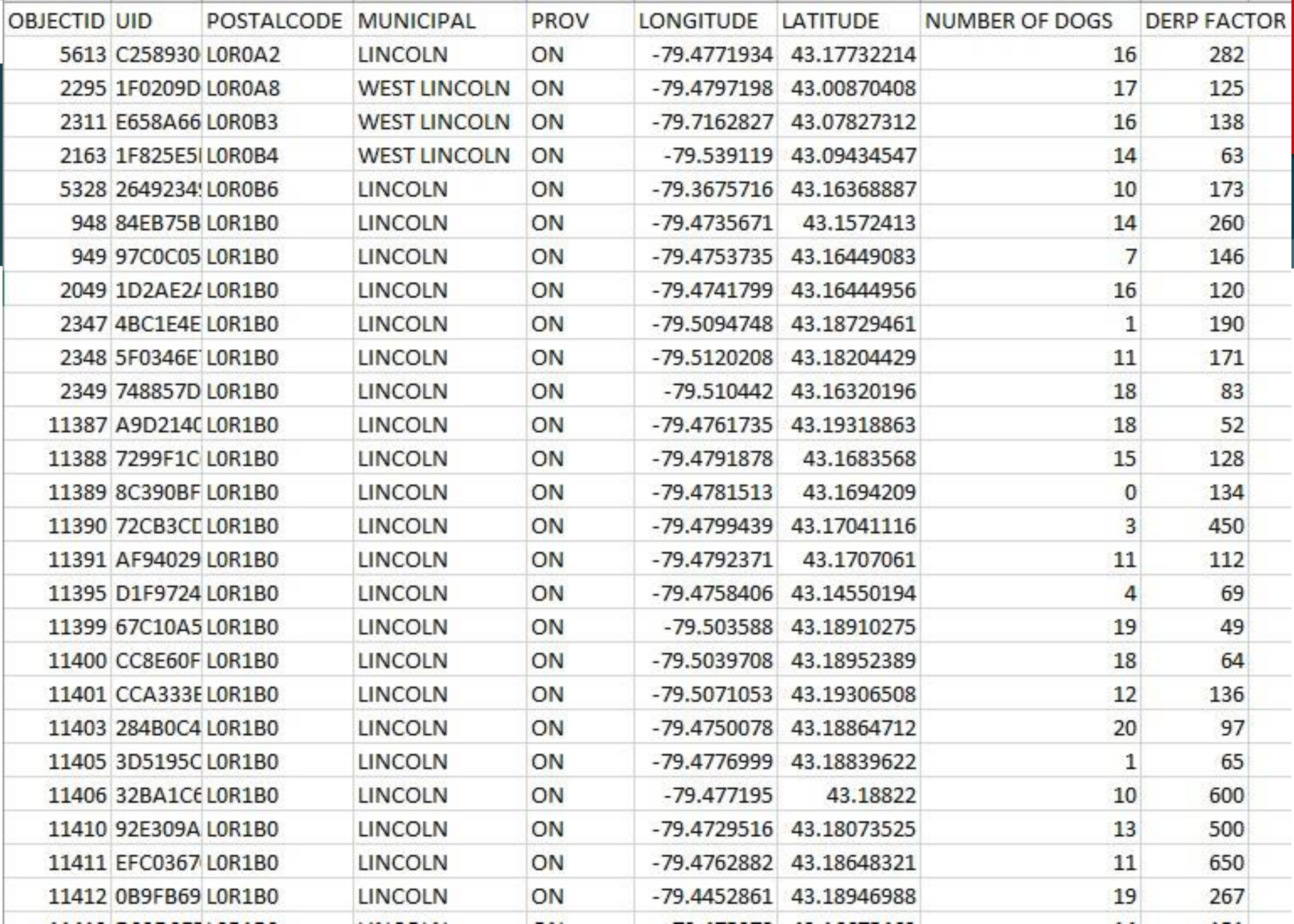
Kinds of Data

- Qualitative - data that is expressed in natural language.
- Quantitative - data that relates to a measurable number, or a certain quantity.

- Primary - data that was collected for the purpose it is used (the data was collected specifically for this study).
- Secondary - data that was collected for another purpose, but is then re-used for another study (ie the data was collected for one study, but can be used in another one).

(Other) Kinds of Data

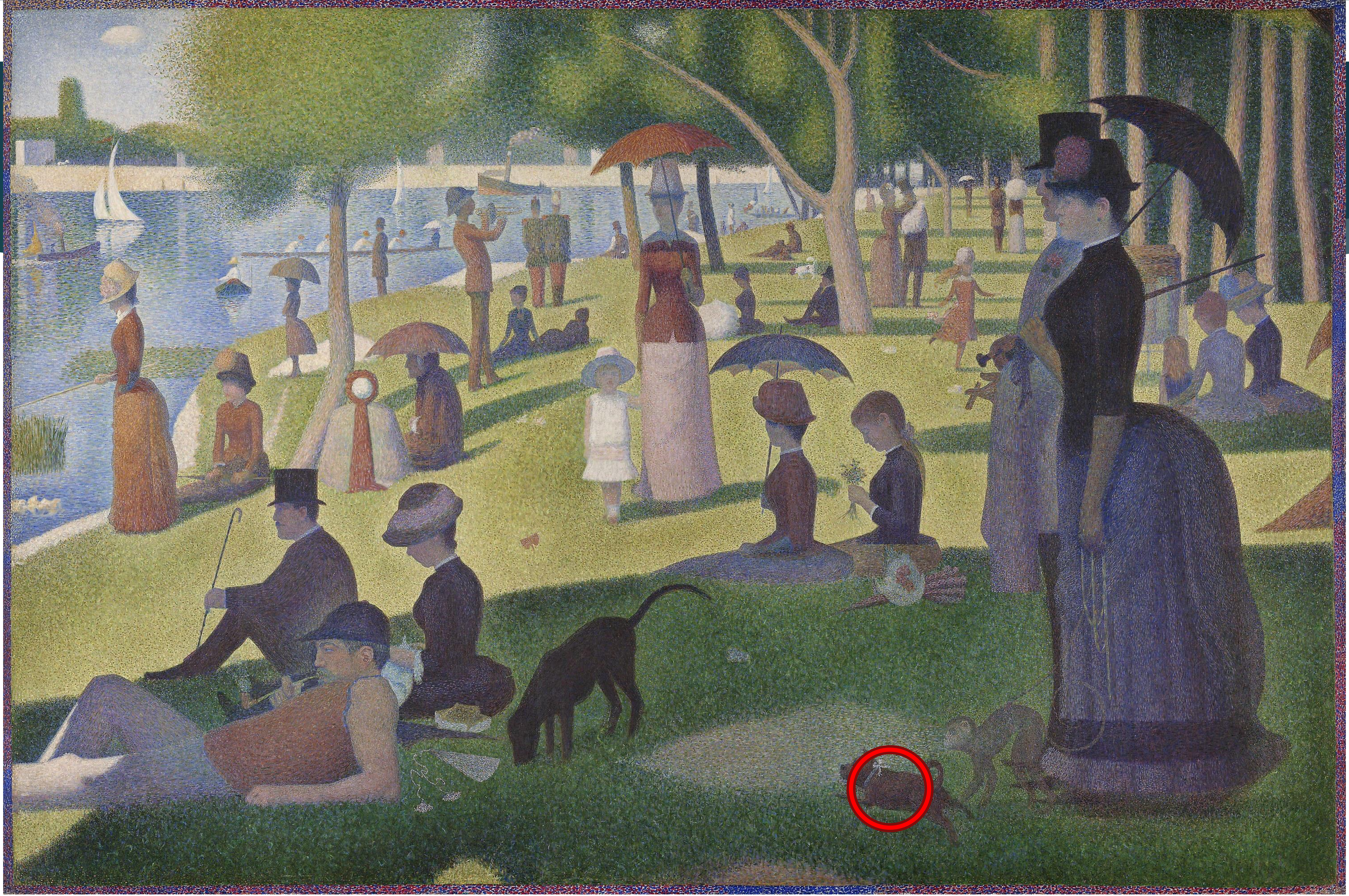
- Numeric data – data based on numbers
- Geospatial data – data that involves a location
- Textual data – you can analyze texts for meaning, style, even (perhaps) psychological states.
- Visual data – aspects about images, such as subjects, lighting, colours, etc.
- Acoustic data – sound levels, pitches, etc.



OBJECTID	UID	POSTALCODE	MUNICIPAL	PROV	LONGITUDE	LATITUDE	NUMBER OF DOGS	DERP FACTOR
5613	C258930	L0R0A2	LINCOLN	ON	-79.4771934	43.17732214	16	282
2295	1F0209D	L0R0A8	WEST LINCOLN	ON	-79.4797198	43.00870408	17	125
2311	E658A66	L0R0B3	WEST LINCOLN	ON	-79.7162827	43.07827312	16	138
2163	1F825E5	L0R0B4	WEST LINCOLN	ON	-79.539119	43.09434547	14	63
5328	2649234	L0R0B6	LINCOLN	ON	-79.3675716	43.16368887	10	173
948	84EB75B	L0R1B0	LINCOLN	ON	-79.4735671	43.1572413	14	260
949	97C0C05	L0R1B0	LINCOLN	ON	-79.4753735	43.16449083	7	146
2049	1D2AE2	L0R1B0	LINCOLN	ON	-79.4741799	43.16444956	16	120
2347	4BC1E4E	L0R1B0	LINCOLN	ON	-79.5094748	43.18729461	1	190
2348	5F0346E	L0R1B0	LINCOLN	ON	-79.5120208	43.18204429	11	171
2349	748857D	L0R1B0	LINCOLN	ON	-79.510442	43.16320196	18	83
11387	A9D214	L0R1B0	LINCOLN	ON	-79.4761735	43.19318863	18	52
11388	7299F1C	L0R1B0	LINCOLN	ON	-79.4791878	43.1683568	15	128
11389	8C390BF	L0R1B0	LINCOLN	ON	-79.4781513	43.1694209	0	134
11390	72CB3C	L0R1B0	LINCOLN	ON	-79.4799439	43.17041116	3	450
11391	AF94029	L0R1B0	LINCOLN	ON	-79.4792371	43.1707061	11	112
11395	D1F9724	L0R1B0	LINCOLN	ON	-79.4758406	43.14550194	4	69
11399	67C10A5	L0R1B0	LINCOLN	ON	-79.503588	43.18910275	19	49
11400	CC8E60F	L0R1B0	LINCOLN	ON	-79.5039708	43.18952389	18	64
11401	CCA333E	L0R1B0	LINCOLN	ON	-79.5071053	43.19306508	12	136
11403	284B0C4	L0R1B0	LINCOLN	ON	-79.4750078	43.18864712	20	97
11405	3D5195C	L0R1B0	LINCOLN	ON	-79.4776999	43.18839622	1	65
11406	32BA1C6	L0R1B0	LINCOLN	ON	-79.477195	43.18822	10	600
11410	92E309A	L0R1B0	LINCOLN	ON	-79.4729516	43.18073525	13	500
11411	EFC0367	L0R1B0	LINCOLN	ON	-79.4762882	43.18648321	11	650
11412	0B9FB69	L0R1B0	LINCOLN	ON	-79.4452861	43.18946988	19	267

What is Data?

- Is this data?
- Yes!



Can I create data?

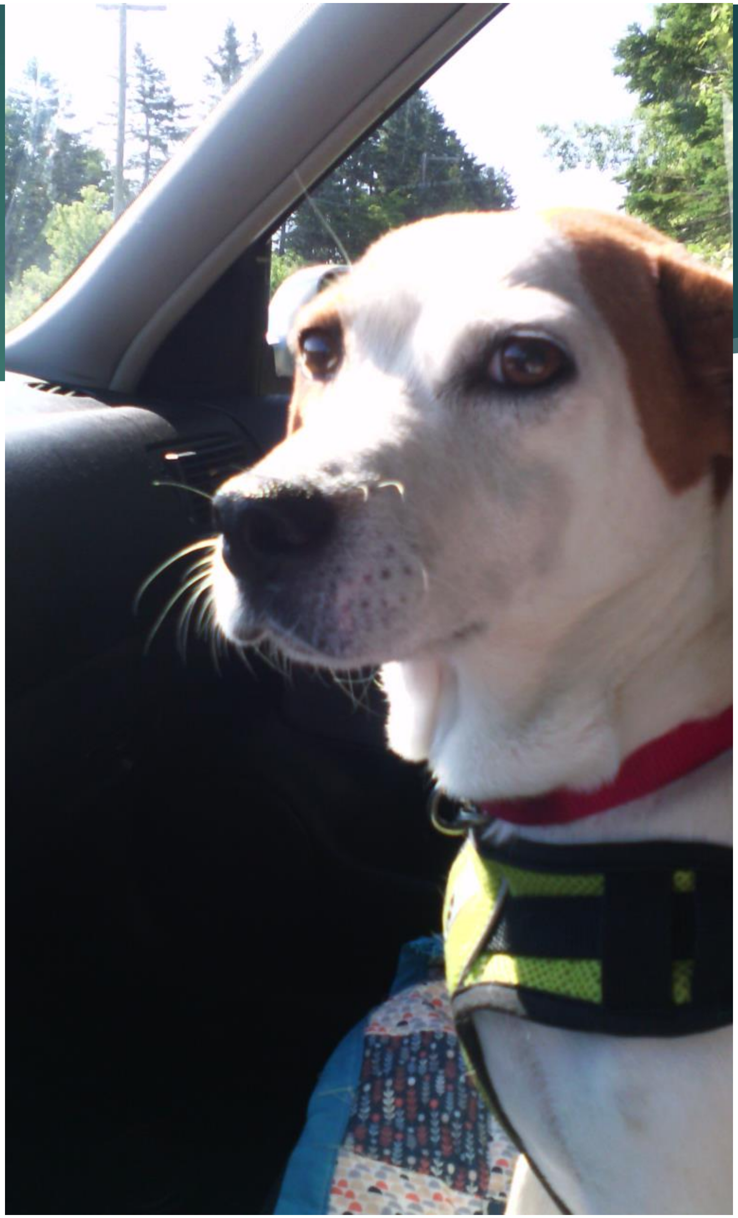
- Yes!
- You’re doing so right now – you’re a part of McGill University (for X amount of time); you’re doing X at McGill; you’re at this talk; your phone is in a particular place; you’re using a computer, or paper and pen/pencil.
- How do you make your own data set? Depends on the question
- If it’s about people – you can survey people, or collect it from institutions that are willing and able to share it

Why is Data?

- What’s the point of data?
- To answer a question!
- Specifically, to answer a question on a broader scale
- Example: Do most people at McGill like dogs, or cats?

Example

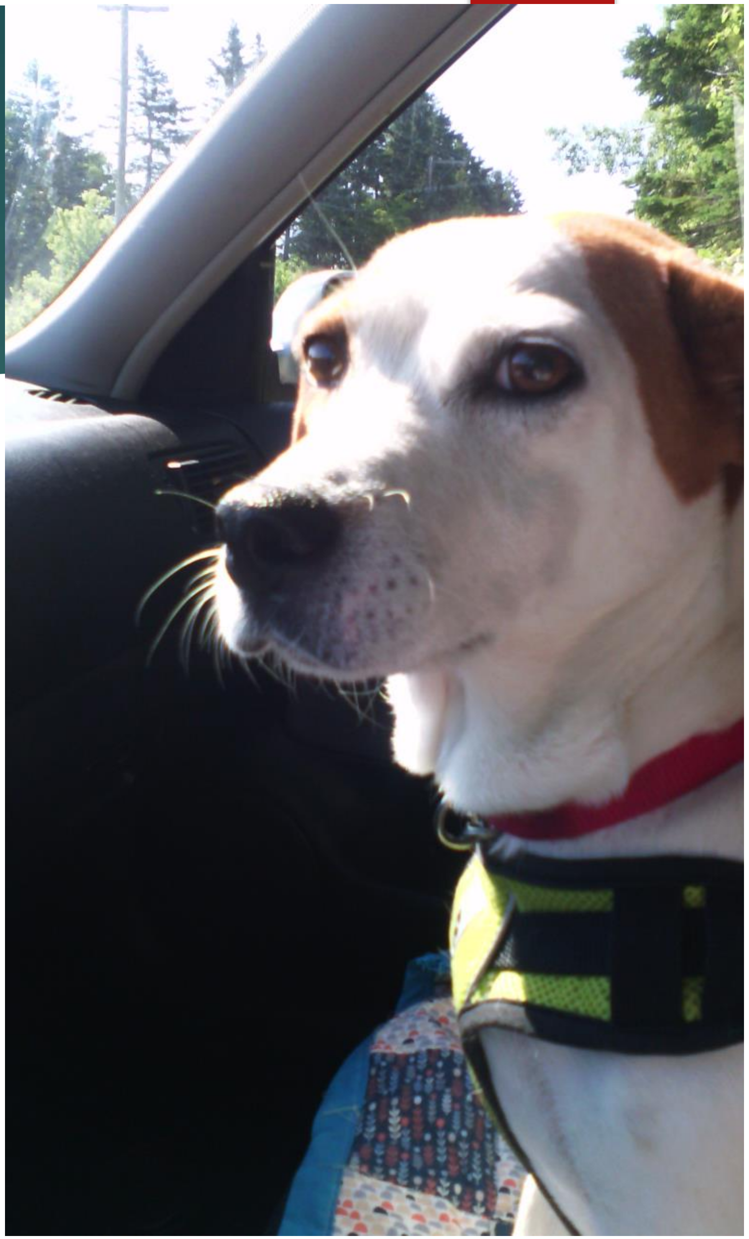
- I like dogs



Example

- Some people like cats

Quick Poll
Cat or Dog?



We just collected some data!

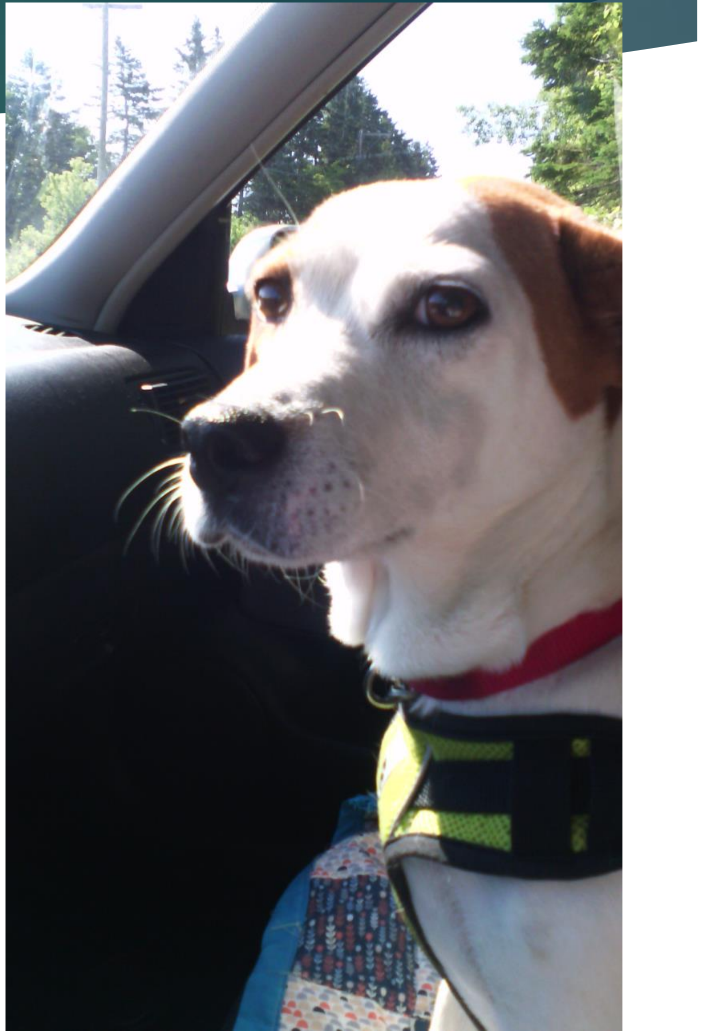
- Yes, it can be that easy!
- Though collecting data that way won't hold up as part of academic research
- Also, if your study involves people, you'll need approval from the Research Ethics Board (https://www.mcgill.ca/research/research/compliance/human/reb-i-ii-iii)

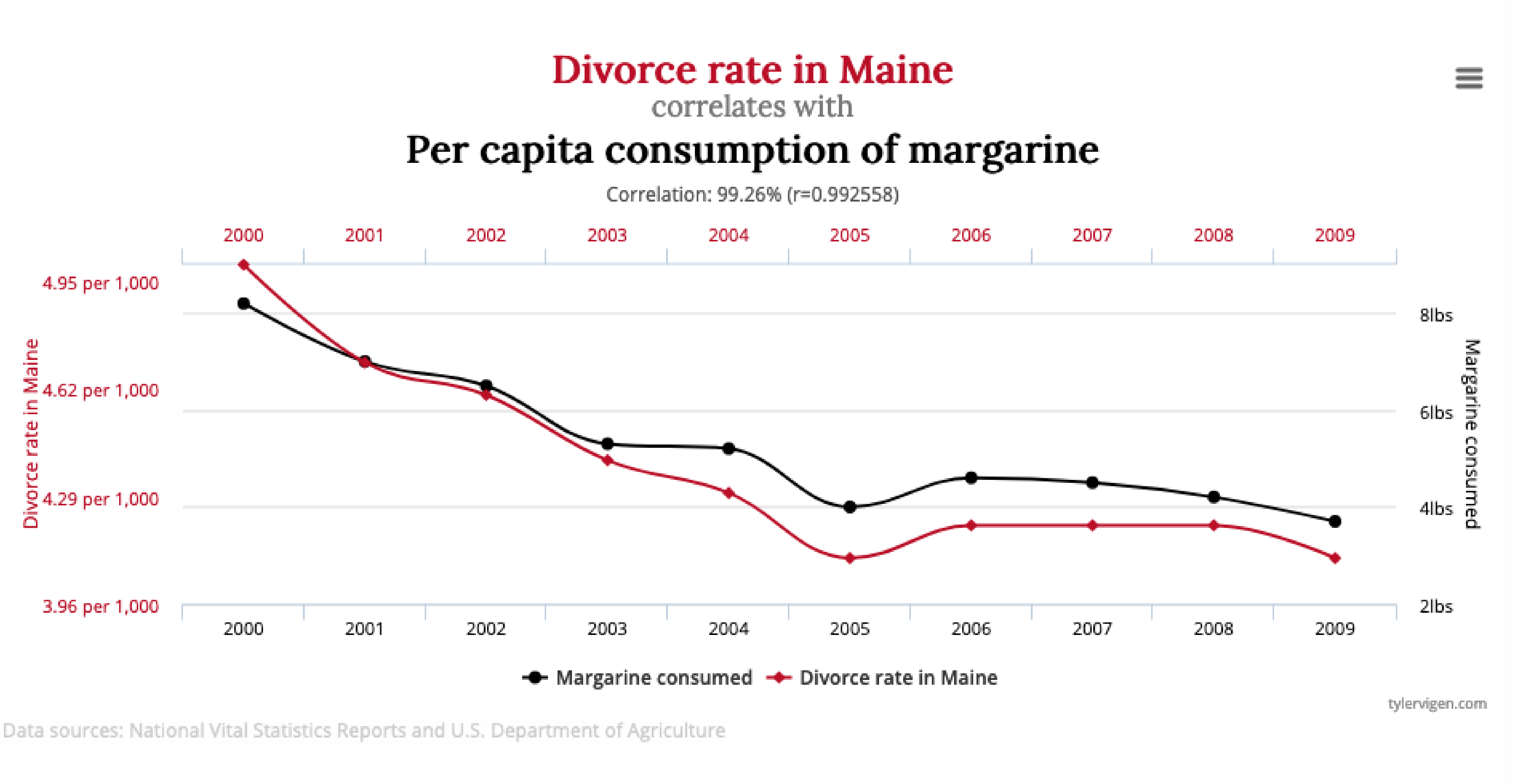
Some important points about data

- Data is not neutral
- Data collection is not neutral
- Some things to think about with data:
  - How was it collected? Eg survey? By phone, in person, online?
  - Why was it collected? What question was the researcher trying to answer?
  - Where was it collected Different places have different people!
  - How much data was collected? A smaller dataset may work for for some questions, whereas a larger one is needed for others
    - Eg A paper exploring information sources for a group of people can use a smaller dataset. But, the research drawn will not be definitive for all members of that group!
    - Example/shameless self-promotion: https://journals.library.ualberta.ca/eblip/index.php/EBLIP/article/view/29515

Looking at our example:

- One of these two looks inherently friendlier
- Even the placement on a screen can impact how people see things (Toepel, Das & Soest)





Divorce rate in Maine
correlates with
Per capita consumption of margarine

Correlation: 99.26% (r=0.992558)

Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

Data Management

- Always good to manage your data!
- Plan for your data – how you will
  - create it
  - structure it
  - and store it
- Some tools and information are available at https://www.mcgill.ca/library/services/data-services/organizing
- Contact Alisa Rod, our Research Data Management Specialist

Where is Data?

- A few good sources for data:
  - Statistics Canada (access through Statistics Canada website or the Census Analyzer) – includes the Census and CANSIM tables.
  - Many governments (federal, provincial, and municipal) now have Open Data Portals.
  - Various non-profits, NGOs, and other organizations also have datasets for use: sometimes open, sometimes licensed.

- You can also create your own data, using tools like surveys or Collector for ArcGIS
- But, it all depends on the question!

How is Data?

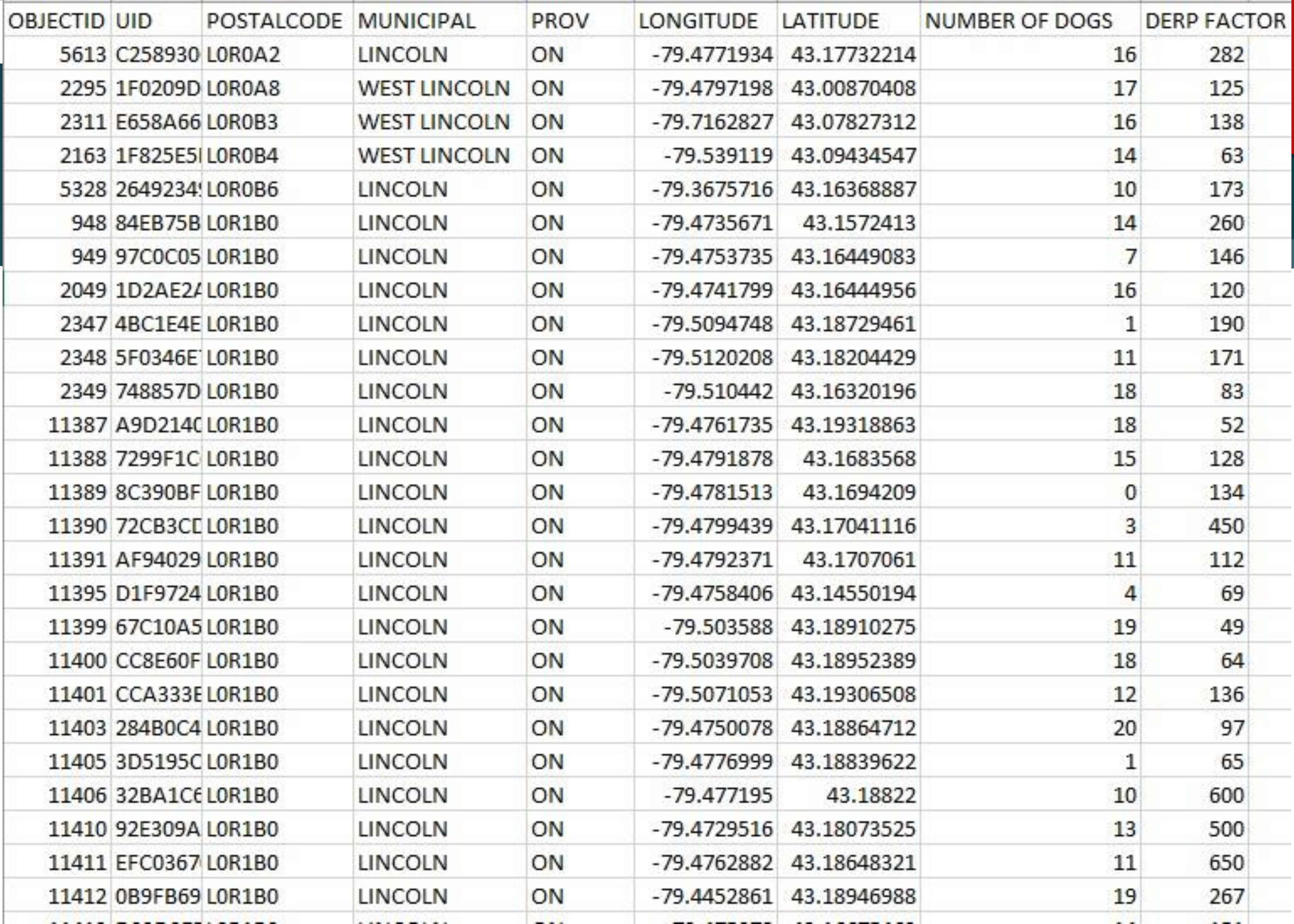
- Data can be stored in many formats, both digital and analog
- Digital: .csv/spreadsheet, word document, a table, .xml, .tiff, .shp, .kml, survey responses, .mp3, etc.
- Analog: Handwritten table, notebook, picture, painting, etc.

Open data

- Open data is (generally) free of cost and (generally) is free of licensing restrictions.
- (Costs for reproduction may be involved, though these will be reasonable)
- (Licensing allows for re-use and re-distribution; attribution is still required)

Presenting: Data!

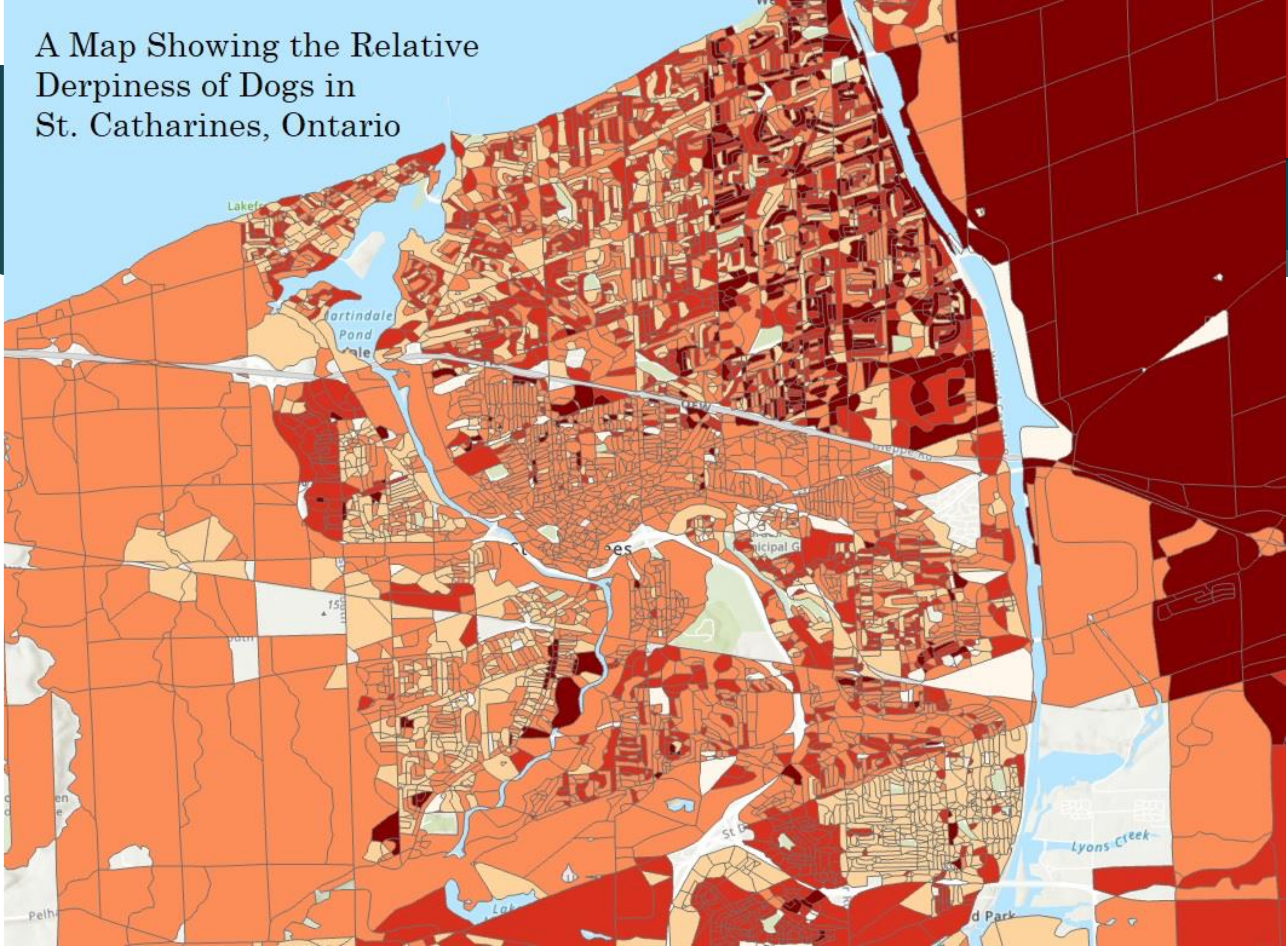
- Data visualizations are a good way to communicate information in an easy-to-understand manner
- They can reveal new information, or help with analysis
- They help your audience to understand your message
- Are an aesthetically pleasing form of presentation



OBJECTID	UID	POSTALCODE	MUNICIPAL	PROV	LONGITUDE	LATITUDE	NUMBER OF DOGS	DERP FACTOR
5613	C258930	L0R0A2	LINCOLN	ON	-79.4771934	43.17732214	16	282
2295	1F0209D	L0R0A8	WEST LINCOLN	ON	-79.4797198	43.00870408	17	125
2311	E658A66	L0R0B3	WEST LINCOLN	ON	-79.7162827	43.07827312	16	138
2163	1F825E5	L0R0B4	WEST LINCOLN	ON	-79.539119	43.09434547	14	63
5328	2649234	L0R0B6	LINCOLN	ON	-79.3675716	43.16368887	10	173
948	84EB75B	L0R1B0	LINCOLN	ON	-79.4735671	43.1572413	14	260
949	97C0C05	L0R1B0	LINCOLN	ON	-79.4753735	43.16449083	7	146
2049	1D2AE2A	L0R1B0	LINCOLN	ON	-79.4741799	43.16444956	16	120
2347	4BC1E4E	L0R1B0	LINCOLN	ON	-79.5094748	43.18729461	1	190
2348	5F0346E	L0R1B0	LINCOLN	ON	-79.5120208	43.18204429	11	171
2349	748857D	L0R1B0	LINCOLN	ON	-79.510442	43.16320196	18	83
11387	A9D214C	L0R1B0	LINCOLN	ON	-79.4761735	43.19318863	18	52
11388	7299F1C	L0R1B0	LINCOLN	ON	-79.4791878	43.1683568	15	128
11389	8C390BF	L0R1B0	LINCOLN	ON	-79.4781513	43.1694209	0	134
11390	72CB3CD	L0R1B0	LINCOLN	ON	-79.4799439	43.17041116	3	450
11391	AF94029	L0R1B0	LINCOLN	ON	-79.4792371	43.1707061	11	112
11395	D1F9724	L0R1B0	LINCOLN	ON	-79.4758406	43.14550194	4	69
11399	67C10A5	L0R1B0	LINCOLN	ON	-79.503588	43.18910275	19	49
11400	CC8E60F	L0R1B0	LINCOLN	ON	-79.5039708	43.18952389	18	64
11401	CCA333E	L0R1B0	LINCOLN	ON	-79.5071053	43.19306508	12	136
11403	284B0C4	L0R1B0	LINCOLN	ON	-79.4750078	43.18864712	20	97
11405	3D5195C	L0R1B0	LINCOLN	ON	-79.4776999	43.18839622	1	65
11406	32BA1C6	L0R1B0	LINCOLN	ON	-79.477195	43.18822	10	600
11410	92E309A	L0R1B0	LINCOLN	ON	-79.4729516	43.18073525	13	500
11411	EFC0367	L0R1B0	LINCOLN	ON	-79.4762882	43.18648321	11	650
11412	0B9FB69	L0R1B0	LINCOLN	ON	-79.4452861	43.18946988	19	267

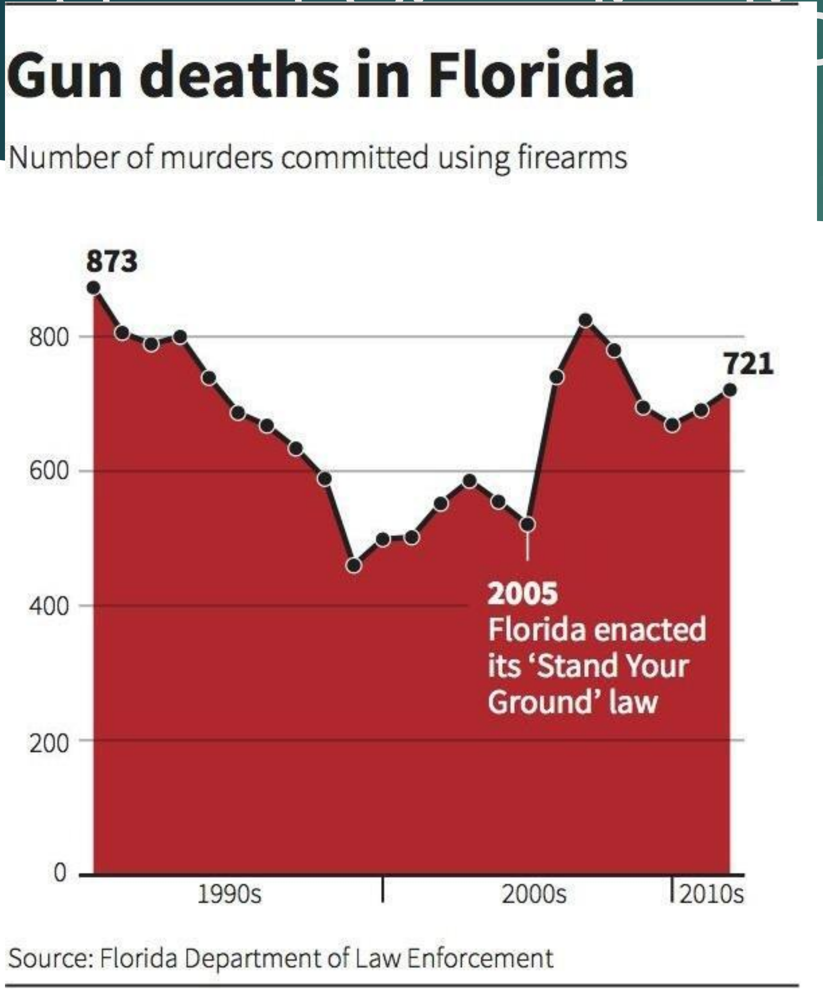
Visualization principles

- Remember your audience
  - Chemistry professors and professional actors have different contexts
- Prepare and explore your data
  - Make sure it is clean able to be visualized
- Select your visualization form
  - Check with different types of visualization – trials can reveal new things
- Test it on a friend
  - Always get feedback to make sure it’s communicating what you want it to



A Map Showing the Relative Derpiness of Dogs in St. Catharines, Ontario

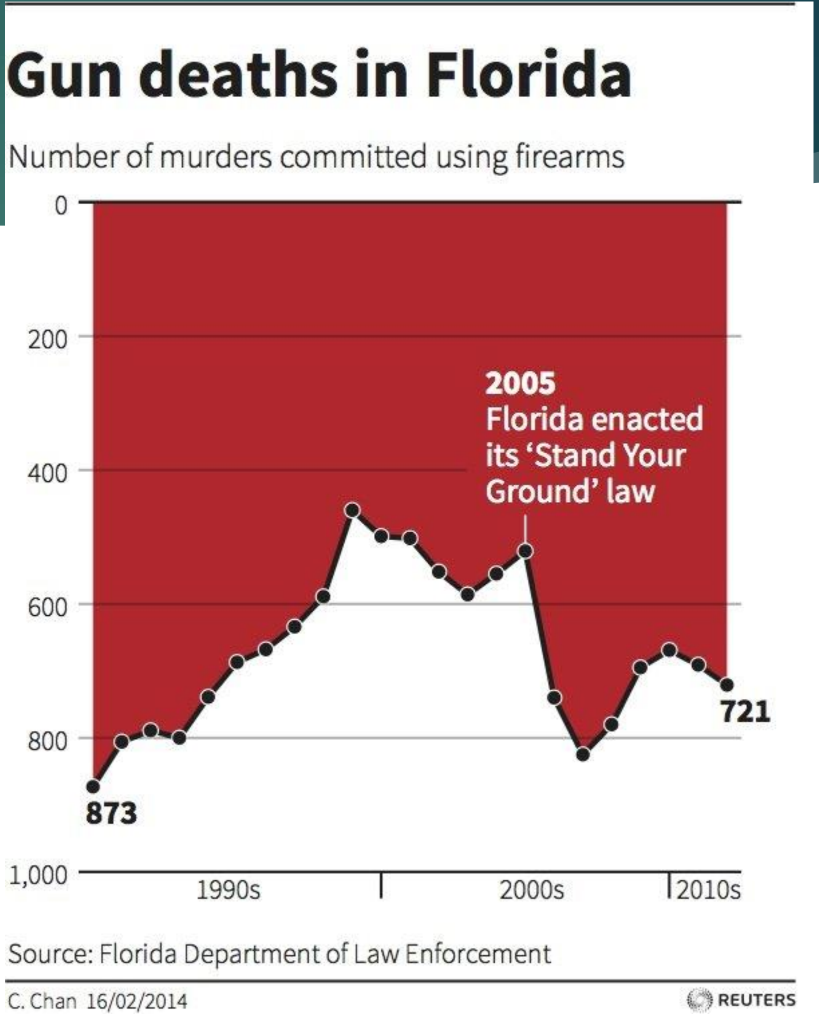
Abused Visualization



Gun deaths in Florida

Number of murders committed using firearms

Source: Florida Department of Law Enforcement



Gun deaths in Florida

Number of murders committed using firearms

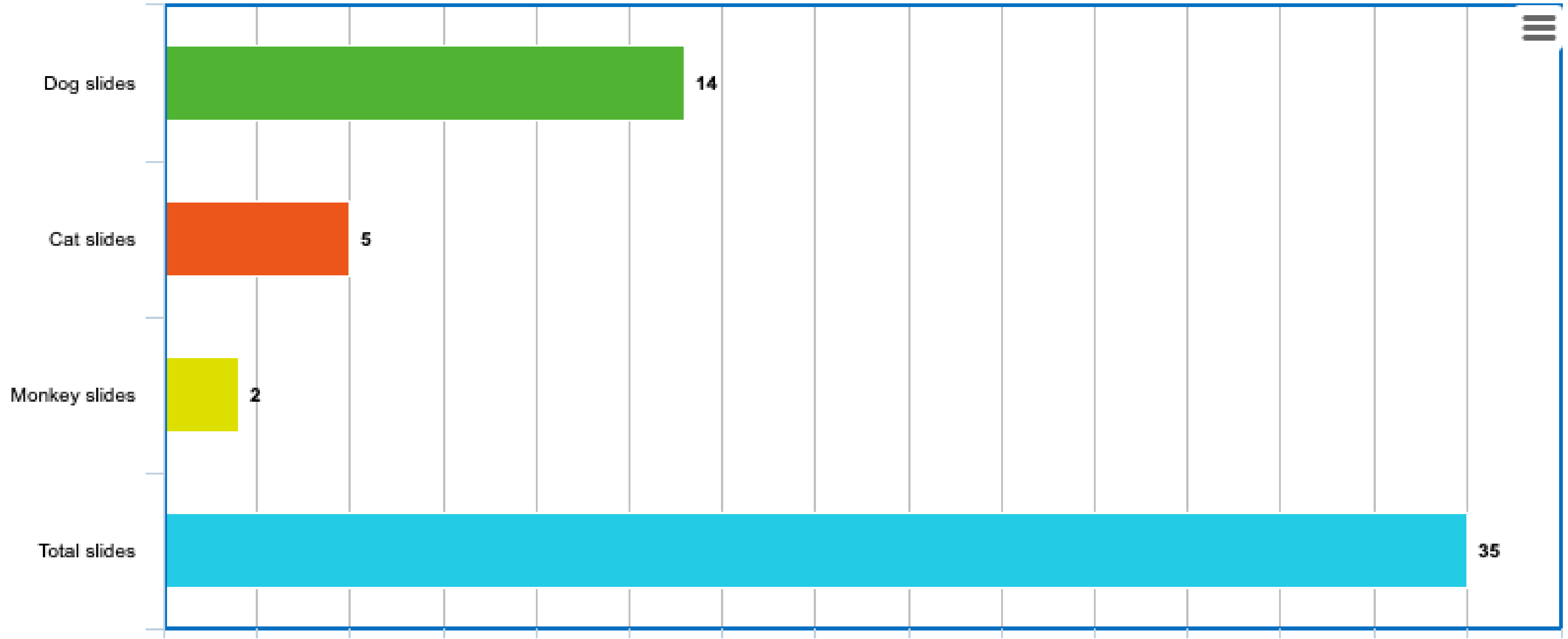
Source: Florida Department of Law Enforcement

C. Chan 16/02/2014

Citing Data

- Yes, data must be cited!
- Numeric data citation libguide: http://libraryguides.mcgill.ca/datacitation
- Geospatial data citation guidelines: https://acmla-acacc.ca/docs/ACMLA_BestPracticesCitations.pdf

In Conclusion



Sources

- Chan, C. (2014). Gun deaths in Florida. *ThomsonReuters.com*. Available: http://graphics.thomsonreuters.com/14/02/US-FLORIDA0214.gif (accessed June 12, 2019).
- Chandler, M. (2019). A Map Showing the Relative Depriness of Dogs in St. Catharines. Data layers: OpenStreetMap contributers; DMTI Spatial, Inc Multuple Enhanced Postal Codes (MEP). Brock University, St. Catharines, ON: Generated by Martin Chandler, April 24, 2019. Using: ArcGIS Pro [GIS]. Version 2.3.2. Redlands, CA: Esri, 2018.
- Chandler, M. (2019). The Information Searching Behaviour of Music Directors. *Evidence Based Library and Information Practice*, *14*(2), 85-99. https://doi.org/10.18438/eblip29515
- DMTI Spatial, Inc. (2014). Multiple Enhanced Postal Codes (MEP)[computer file]. Available from Scholars GeoPortal: http://geo.scholarsportal.info.proxy.library.brocku.ca/#r/details/_uri@=643323315. (accessed June 6, 2019).
- Engel, P. (2014, Feb 18). This Chart Shows An Alarming Rise In Florida Gun Deaths After 'Stand Your Ground' Was Enacted. *BusinessInsider.com*. Available: https://www.businessinsider.com/gun-deaths-in-florida-increased-with-stand-your-ground-2014-2 (accessed June 12, 2019).
- Merriam Webster. (2019). "Data". Available: https://www.merriam-webster.com/dictionary/data (accessed June 9, 2019).
- *Multiple Enhanced Postal Codes (MEP) – 2014* [computer file]. Markham, Ontario: DMTI Spatial Inc., 2014. Available: Scholars GeoPortal http://geo.scholarsportal.info.proxy.library.brocku.ca/#r/details/_uri@=643323315$DMTI_2014_CanMapPS_MEP_CAN (accessed June 10, 2019).
- Seurat, G. (1884). Un dimanche après-midi à l'Île de la Grande Jatte [painting]. Available: https://g.co/arts/Ke9bVTioqsWYQmqP6 (accessed June 12, 2019).
- Vigen, T. (n.d.). Divorce rate in Maine correlates with Per capita consumption of margarine. *Spurious Correlations*. Available : http://www.tylervigen.com/spurious-correlations (accessed June 11, 2019).
- Winkler, E. (2019, June). Was Shakespeare a Woman? *The Atlantic, June 2019, 86-95.*
- Toepoel, V., Das, J.W.M. & van Soest, A. (2009). Design of Web Questionnaires: The Effect of Layout in Rating Scales. *Journal of Official Statistics 25*(4), 509-528. Available: https://core.ac.uk/download/pdf/6315116.pdf (accessed June 10, 2019).
- Trimble, L. (2017). Finding Canadian Statistics & Data. Presentation for graduated student library assistants, Map & Data Library, University of Toronto.

Questions?

Martin Chandler

martin.chandler@mcgill.ca