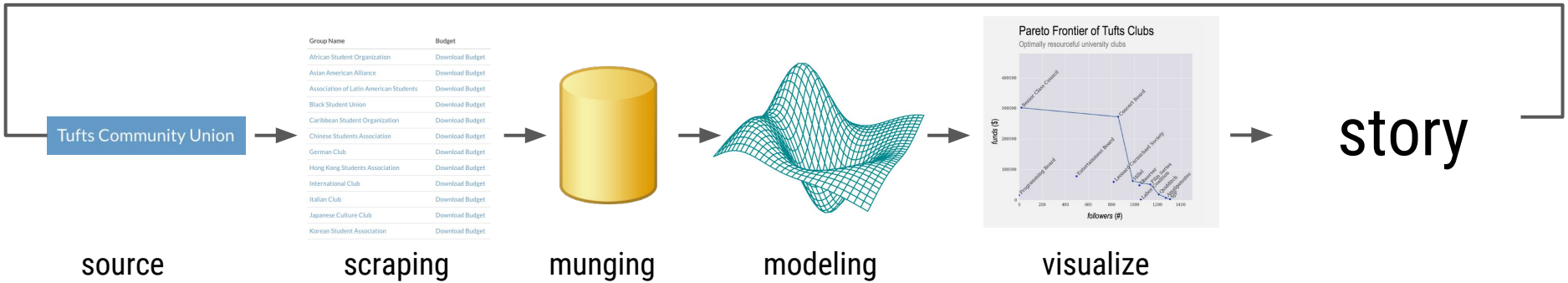
Data Science and Journalism

A Practical Guide

Tufts Independent Data Journal

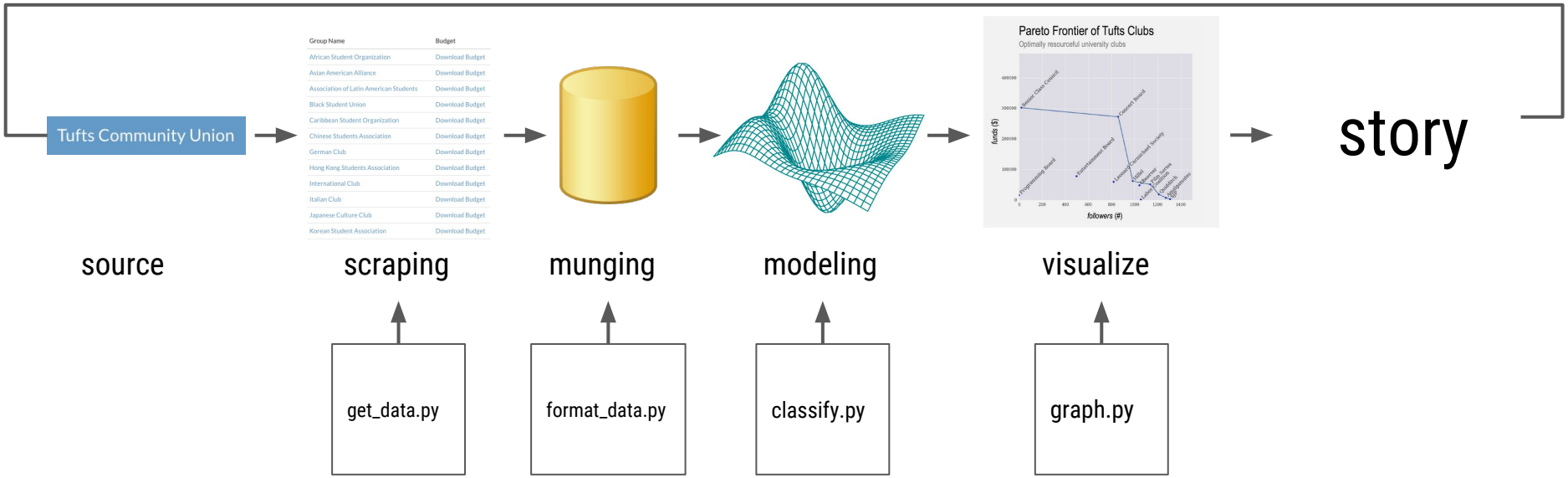
Recap

“Journalism as a storytelling pipeline.”



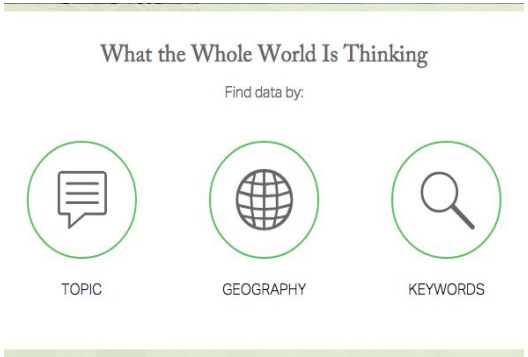
Recap

"Python as an easy scripting tool to use at every step."



Recap : Source / Scraping

GALLUP Analytics



polls

YAHOO! Finance

time series

United States Census Bureau

STATISTICS FROM 2000 CENSUS

Planning Area	Central	Northeast	West	South	Outside City*	Citywide
Population	33,550	31,121	22,254	27,099		114,024
Number of dwelling units	11,237	12,971	10,050	12,446		46,704
Owner occupied	2,588	5,807	6,682	5,608		20,685
Renter	8,402	6,692	3,076	6,838		17,008
Average household size	2.21	2.27	2.26	2.15		2.22
Median income	30,627	57,898	59,939	51,447		46,299
Person density per acre	22	5	6	6		9.75
% of households with children	9%	30%	30%	27%		24%
Disability status - ages 5-64	8%	8%	10%	13%		9%
Disability status - ages 65 & up	37%	30%	32%	39%		34%
Minority comp. African American	6%	9%	6%	13%		9%
Minority composition - Asian	11%	21%	3.5%	9%		12%
Number of parks	23	54	36	38	6	157
Acreage of parks	125.67	885.40	605.63	341.43	130.41	2088.54
Acreage parkland/1000 residents	3.7	28.45	27.21	12.6		18.32
Percent student population						33%

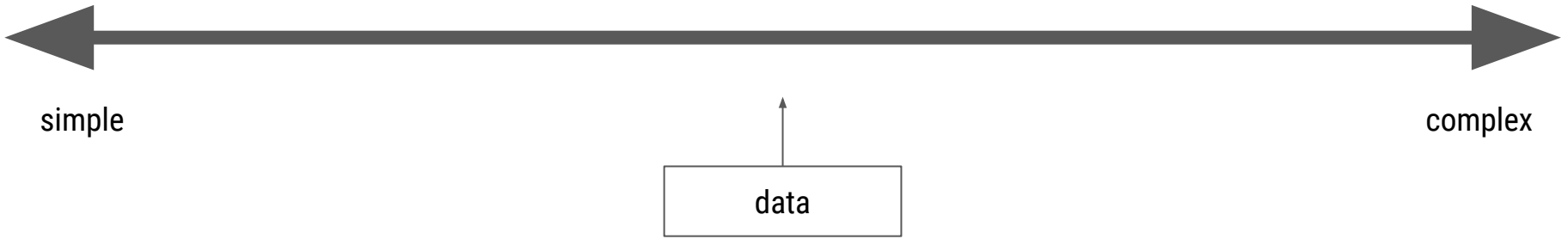
*Outside City refers to parks such as Marshall and Dolph which are outside of the official City limits of Ann Arbor, but are still part of the park system.

census

Munging / Modelling

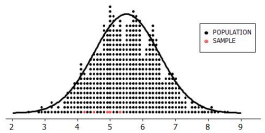
(the data science part)

“Your model should be a function of:
1. the complexity of your data / domain.
2. the complexity of your question.”



Munging / Modelling

(the data science part)



summary
statistics

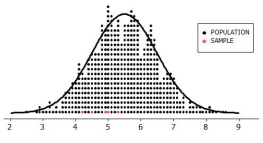
simple

complex

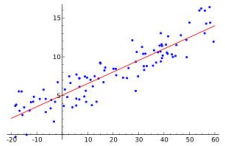
data

Munging / Modelling

(the data science part)



summary statistics



regression

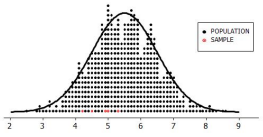
simple

complex

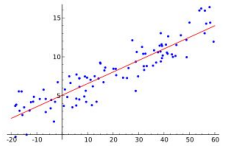
data

Munging / Modelling

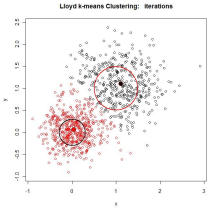
(the data science part)



summary statistics



regression



clustering

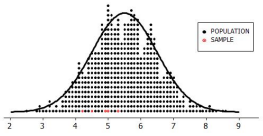
simple

complex

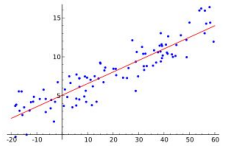
data

Munging / Modelling

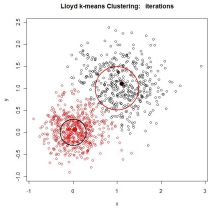
(the data science part)



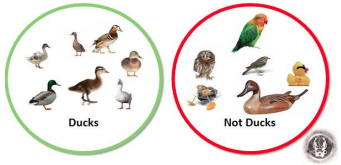
summary statistics



regression



clustering



classification

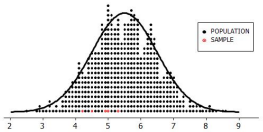
simple

complex

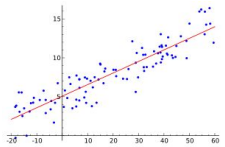
data

Munging / Modelling

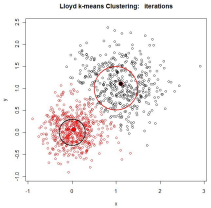
(the data science part)



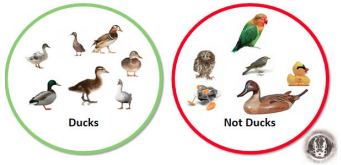
summary statistics



regression



clustering



classification

natural language processing

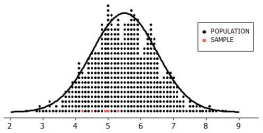
simple

complex

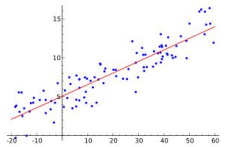
data

Munging / Modelling

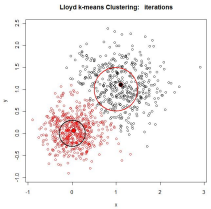
(the data science part)



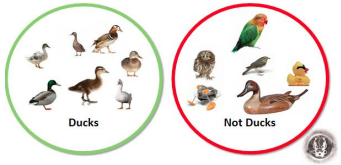
summary statistics



regression



clustering



classification

natural language processing

prediction

simple

complex

data

A Simple Rule for Data Science:

A Simple Rule for Data Science:

Ask the right question.

A Simple Rule for Data Science:

Ask the right question.
Find the right model.

A Simple Rule for Data Science:

Ask the right question.
Find the right model.
Learn how to use it.

Case Study : Presidential Candidates

Ideas on data sources, polls, etc?

What are some relevant questions to ask?

How could we answer those questions using ‘data science’?

Case Study : University Trends

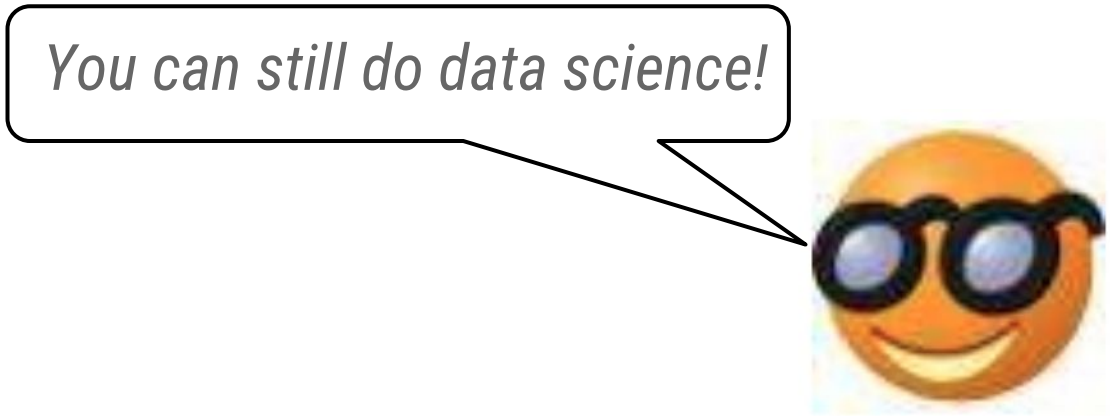
Ideas on data sources, polls, etc?

What are some relevant questions to ask?

How could we answer those questions using ‘data science’?

So what if I don’t really know Python?

So what if I don’t really know Python?



Demo

Python Modules

- SciPy: http://www.scipy.org/
  - high level mathematics/advanced data analysis capabilities
  - widely used
- Pandas: http://pandas.pydata.org/
  - open source, easy to use data analysis libraries
  - allows you to perform analysis in Python, rather than only do munging/preparation
  - makes Python a viable alternative to R
- Scikit-learn: https://github.com/scikit-learn/scikit-learn
  - open source machine learning module
  - interfaces well with matplotlib for creating visuals
  - built on top of SciPy