Streams Processing

Exploratory Data Analysis

What it is

	Incident Type	Location	Borough	Creation Date	Closed Date	Latitude	Longitude
1595	HazMat-Chemical	300 Western Avenue	Staten Island	07/02/2013 11:30:49 AM	07/02/2013 12:48:04 PM	40.633754480426916	-74.18251802332459
1596	Fire-1st Alarm	300 Western Ave	Staten Island	05/20/2011 02:00:23 PM	07/02/2013 12:53:27 PM	40.633754480426916	-74.18251802332459
1597	Utility-Gas Service Line	1047 Amsterdam Avenue	Manhattan	07/02/2013 01:08:11 PM	07/02/2013 02:13:39 PM	40.80395045070998	-73.96313868034538
1598	Fire-10-76 (Commercial High Rise)	22 Cortlandt St	Manhattan	07/02/2013 02:58:13 PM	07/02/2013 04:11:12 PM	40.71022280163124	-74.01089676246752
1599	Fire-2nd Alarm	511 Lexington Ave	Manhattan	06/05/2013 08:51:40 AM	07/03/2013 10:28:37 AM	40.75510242312456	-73.97320355186292
1600	Utility-Water Main	55 East Houston Street	Manhattan	07/02/2013 03:34:51 PM	07/03/2013 07:06:48 PM	40.72472374319593	-73.99429247627018
1601	LawEnforcement-White Powder	900 Fteley AVe	Bronx	07/03/2013 09:53:09 PM	07/03/2013 11:23:37 PM	40.82290402603169	-73.87003989233642
1602	LawEnforcement-Suspicious Package	28-34 49th Street	Queens	07/04/2013 01:19:58 AM	07/04/2013 01:47:37 AM	40.76128774263084	-73.90718556366119
1603	HazMat-Liquid	23rd Street & 3rd Avenue	Brooklyn	07/02/2013 11:22:12 AM	07/04/2013 10:40:59 AM	40.662790925443	-73.99886460131124
1604	Structural-Sidewalk Shed	West 165th Street & Broadway	Manhattan	07/04/2013 08:50:13 AM	07/04/2013 10:50:31 AM	40.8391780038646	-73.94113521565507
1605	Fire-Metro North Train on Fire	East Tremont Ave & Park Ave	Bronx	07/04/2013 12:55:53 PM	07/04/2013 02:27:27 PM		
1606	Fire-3rd Alarm	125 Lake Avenue	Staten Island	07/04/2013 11:03:36 AM	07/04/2013 02:40:58 PM	40.63351755393437	-74.15094186010192
1607	Fire-10-77 (Residential High Rise)	1535 University Avenue	Bronx	07/04/2013 11:20:02 PM	07/05/2013 12:19:30 AM	40.84588291295465	-73.92194063355016
1608	Rescue-Technical		Manhattan	07/05/2013 08:33:33 AM	07/05/2013 11:04:31 AM		
1609	Structural-Partial Collapse	120 Riverside Drive	Manhattan	07/05/2013 12:25:53 AM	07/05/2013 01:26:28 PM	40.78854036460794	-73.98089288622866
1610	Utility-Gas Service Line	218 West 147 Street	Manhattan	07/05/2013 03:37:12 PM	07/05/2013 05:14:17 PM	40.823309727773825	-73.93904279472251
1611	Utility-Power Outage		Bronx	07/05/2013 05:30:49 PM	07/05/2013 08:30:26 PM	40.894557751747016	-73.86105620593477
1612	Utility-Power Outage		Staten Island	07/06/2013 01:53:45 AM	07/06/2013 10:53:46 AM		
1613	Utility-Water Main	26 Madison Street	Manhattan	07/06/2013 12:07:09 AM	07/06/2013 08:14:04 PM	40.71177959709003	-73.99963929106451
1614	Utility-Power Outage	Ralph Avenue & Fulton Street	Brooklyn (NYCHA-Brevoort	07/06/2013 01:12:33 PM	07/06/2013 08:16:17 PM	40.6788705990481	-73.92164580117112

NYC OD: Emergency Response Incidents

Any method of **looking at data** that does not include formal statistical modeling and inference

Why it matters

Confirmatory statistical analyses are based on models.

$$y = Ax + \mathcal{N}(0, \sigma^2)$$

Structural component

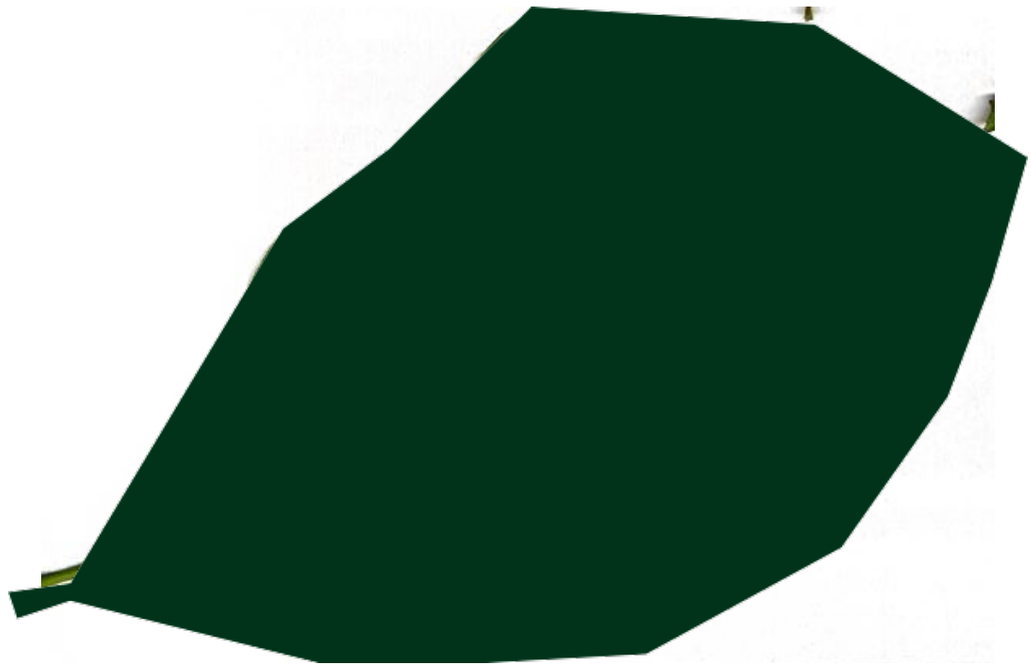
Random component

Signal

Noise

Why it matters

Models are not perfect representations of the real world.

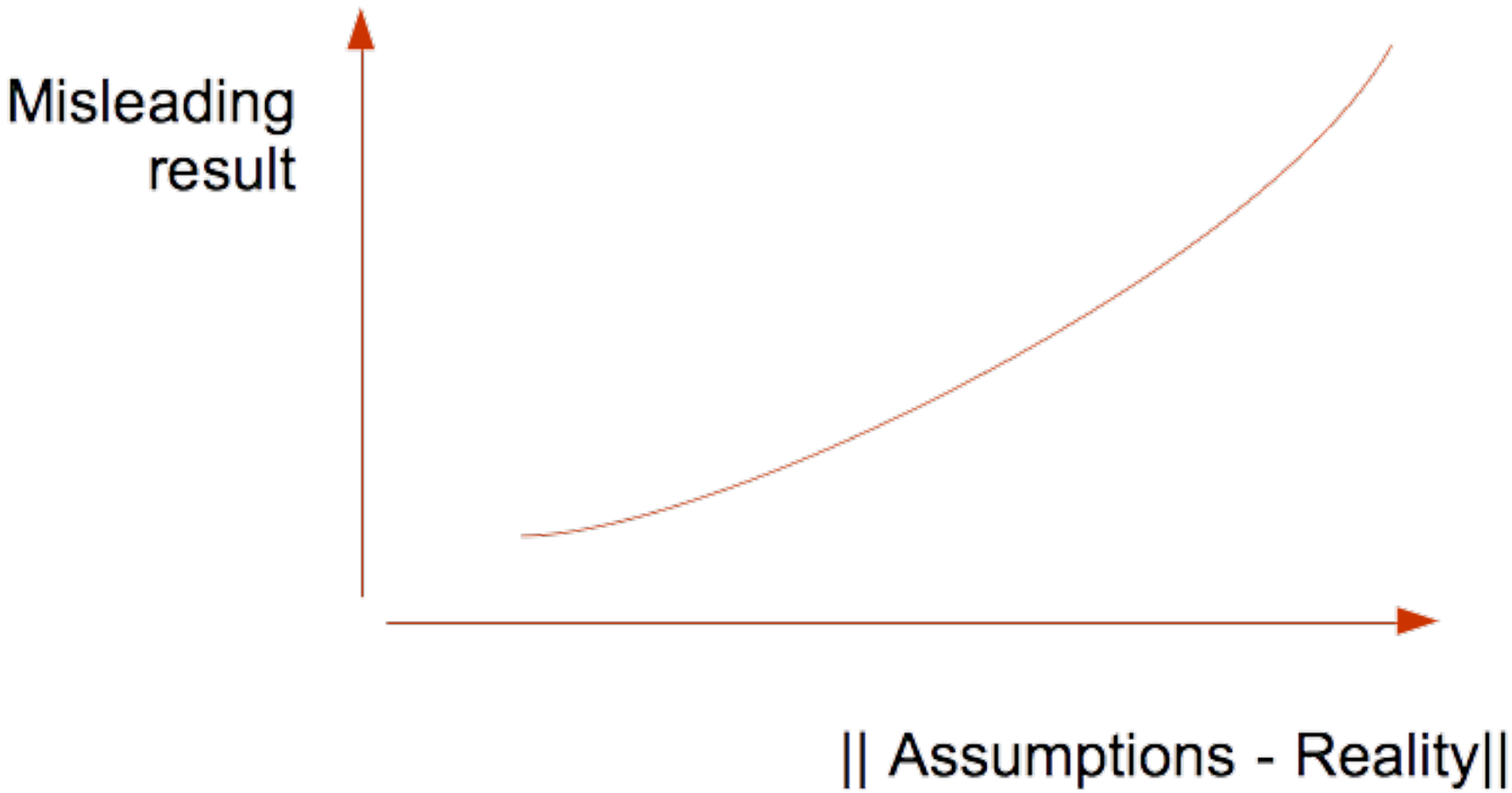


https://commons.wikimedia.org/w/index.php?curid=521370

But some are **close enough** to be useful!

Why it matters

What is close enough to reality?



Statistical inference always depends on **model assumptions** about the data.

Use EDA for:

- Detecting **data noise**
- Checking **assumptions**
- Selecting data **models**
- Determining **relationships** between the **explanatory** variables
- Determining **relationships** between **explanatory** and **outcome** variables

Techniques

Look at the raw data

- Look at the top and bottom of your data.
- How much missing data?
- How noisy is the data?

Compute summary statistics

- What values the variables take?
- How often variables take those values?

Visualize

- Show comparisons
- Show structure
- Show multivariate data

Summary stats

Range, max, min

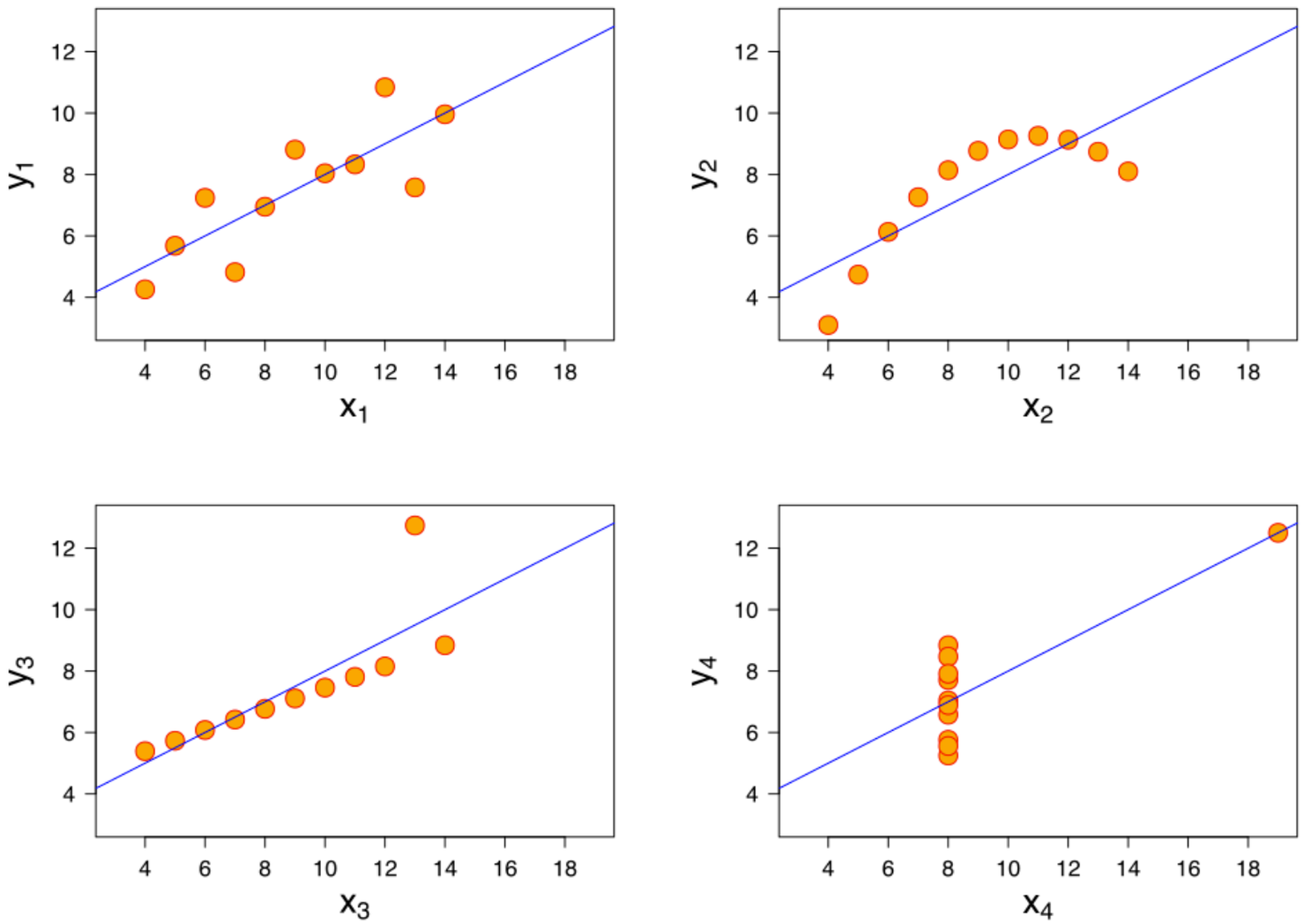
Mean, mode, median

Variance

Correlation

…

Beware of summary stats



Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician. 27 (1): 17–21.
From wikipedia

Beware of summary stats

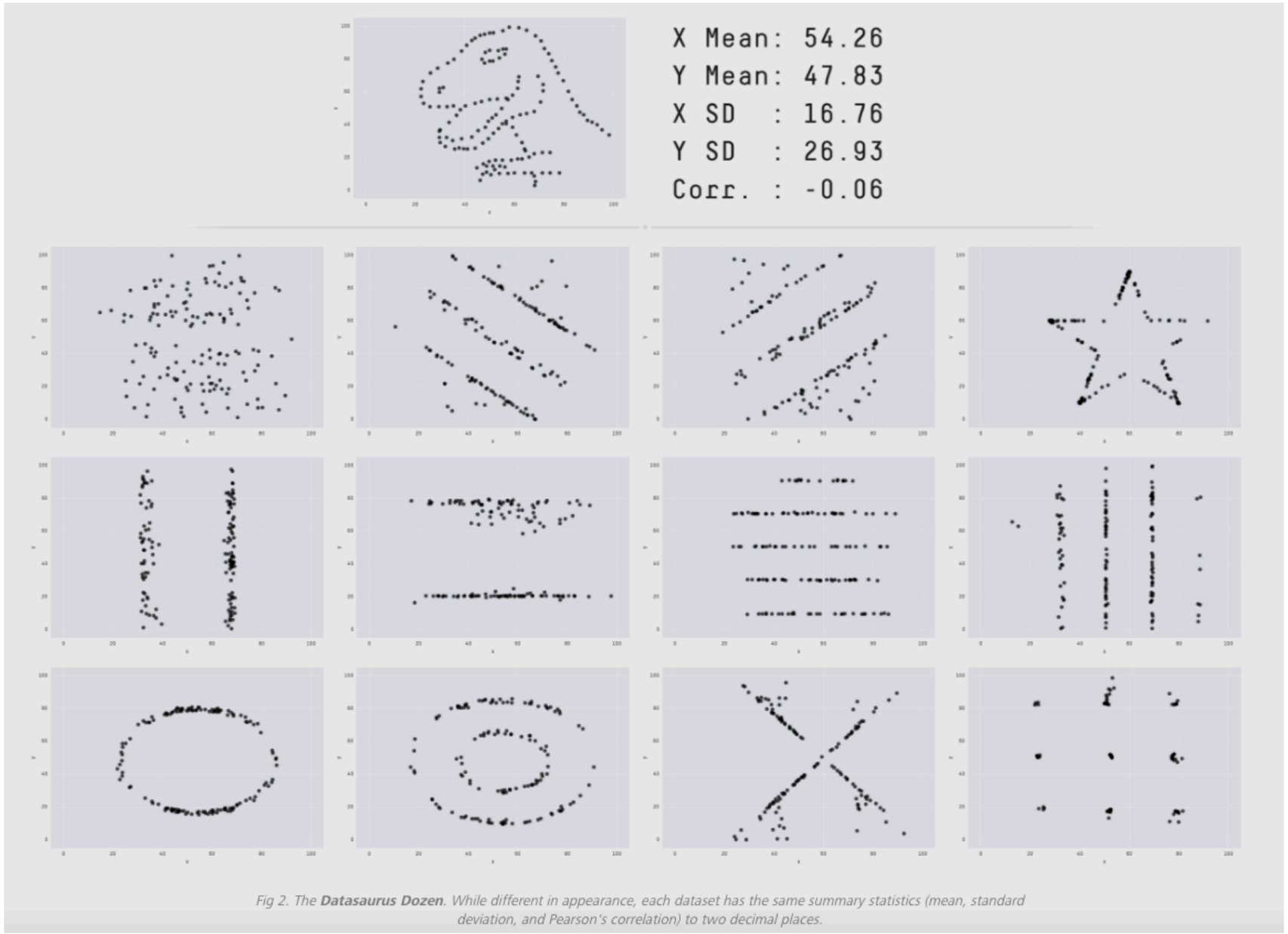


*Fig 2. The **Datasaurus Dozen**. While different in appearance, each dataset has the same summary statistics (mean, standard deviation, and Pearson's correlation) to two decimal places.*

https://www.autodeskresearch.com/publications/samestats

Data Visualization

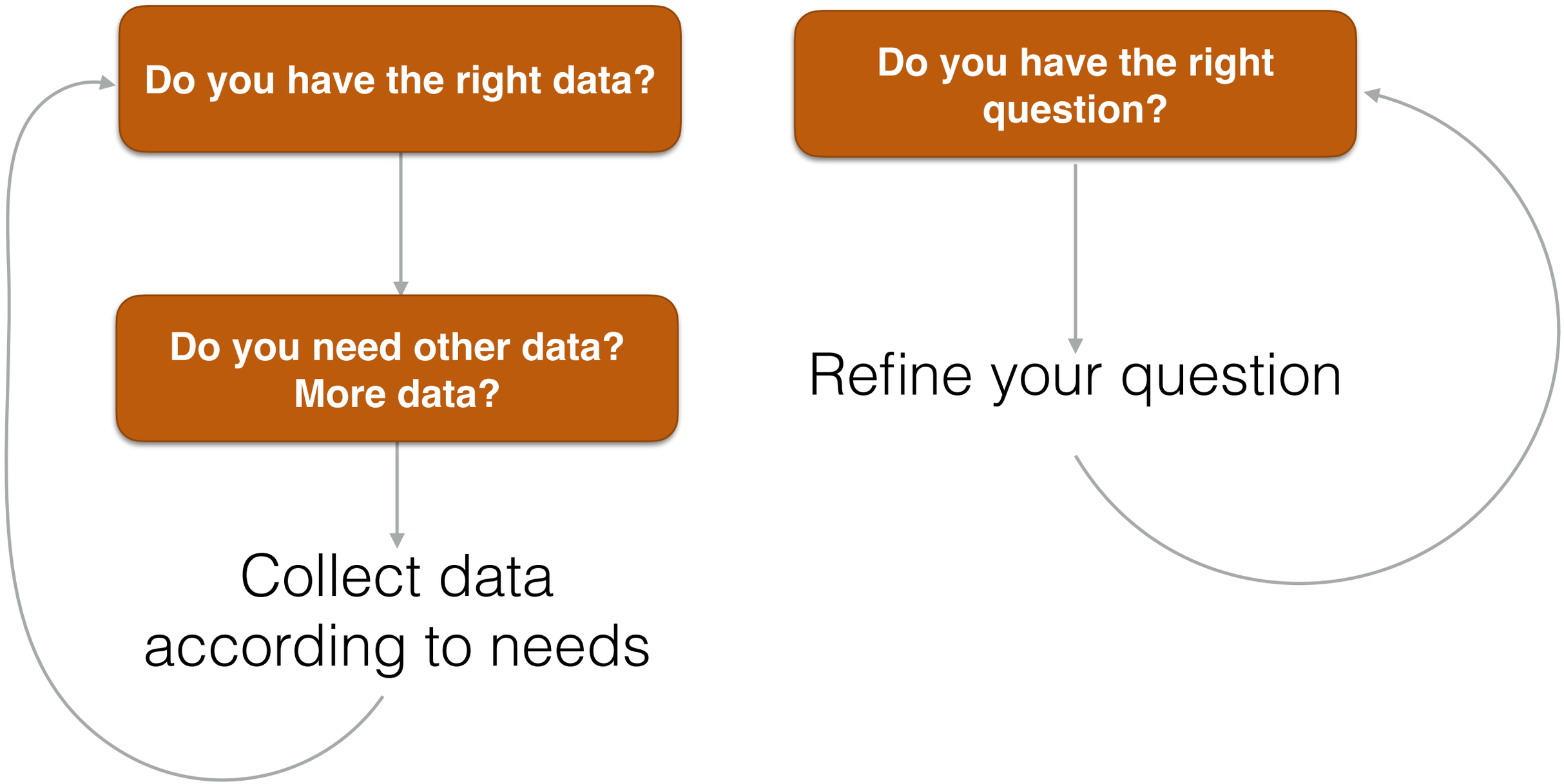
Data points across some features

Features across all data points

Histograms

…

Followup


Do you have the right data?
Do you need other data?
More data?
Collect data
according to needs
Do you have the right
question?
Refine your question