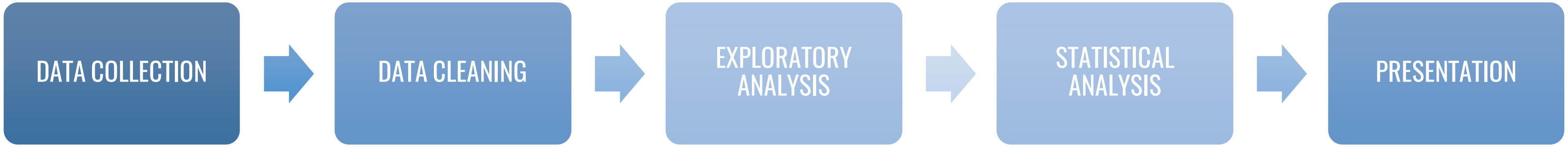
EXPLORATORY DATA ANALYSIS & ELICITATION

PETRA ISENBERG

VISUAL ANALYTICS

ANALYSIS COMPONENTS

Remember: not necessarily in this order or linear

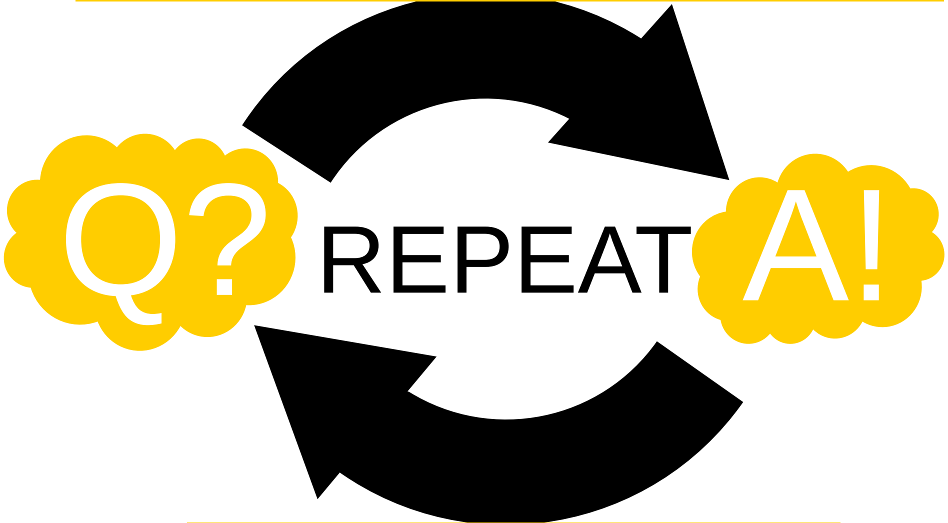


WHY DO YOU NEED DATA?

(HINT: Usually, because you have a question you need to answer!)

ANALYSIS CIRCLE

GATHERING DATA,
APPLYING STATISTICAL TOOLS,
AND CONSTRUCTING GRAPHICS
TO ADDRESS QUESTIONS



INSPECT “ANSWERS” AND
ASSESS NEW QUESTIONS

DATA IS ONLY AS GOOD AS THE QUESTIONS YOU ASK

Some people say....

WHERE DO QUESTIONS COME FROM?

WHERE DO QUESTIONS COME FROM?

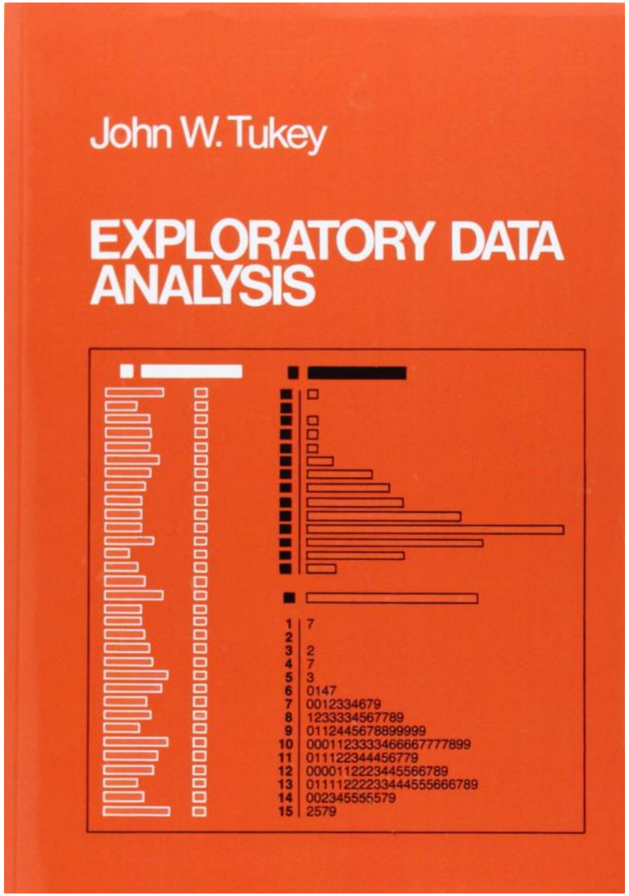
STAKEHOLDERS

EXPLORATORY ANALYSIS

“EXPLORATORY DATA ANALYSIS”

JOHN TUKEY

(IN CONTRAST TO “**CONFIRMATORY**” DATA ANALYSIS)



Based on insights developed at **Bell Labs** in the 60's

Introduced a number of novel techniques for visualizing and summarizing data:

- 5-number summary
- Box plots
- Stem and leaf diagrams

EXPLORATORY ANALYSIS IS ABOUT UNDERSTANDING DATA AND CHECKING ASSUMPTIONS

- IS THE DATA **CORRECT**?
- DOES IT **MATCH OUR PREVIOUS EXPECTATIONS**?
- IS THERE **A RELATIONSHIP**?
  A CORRELATION?
  A TREND?
  ETC.?

E.D.A. CIRCA ~1970

- Mostly done by hand
  (computation is expensive and inaccessible)

- Simple statistical summaries and charts

TUKEY’S 5-NUMBER SUMMARY

The sample minimum (smallest observation)

The lower quartile

The median (middle value)

The upper quartile

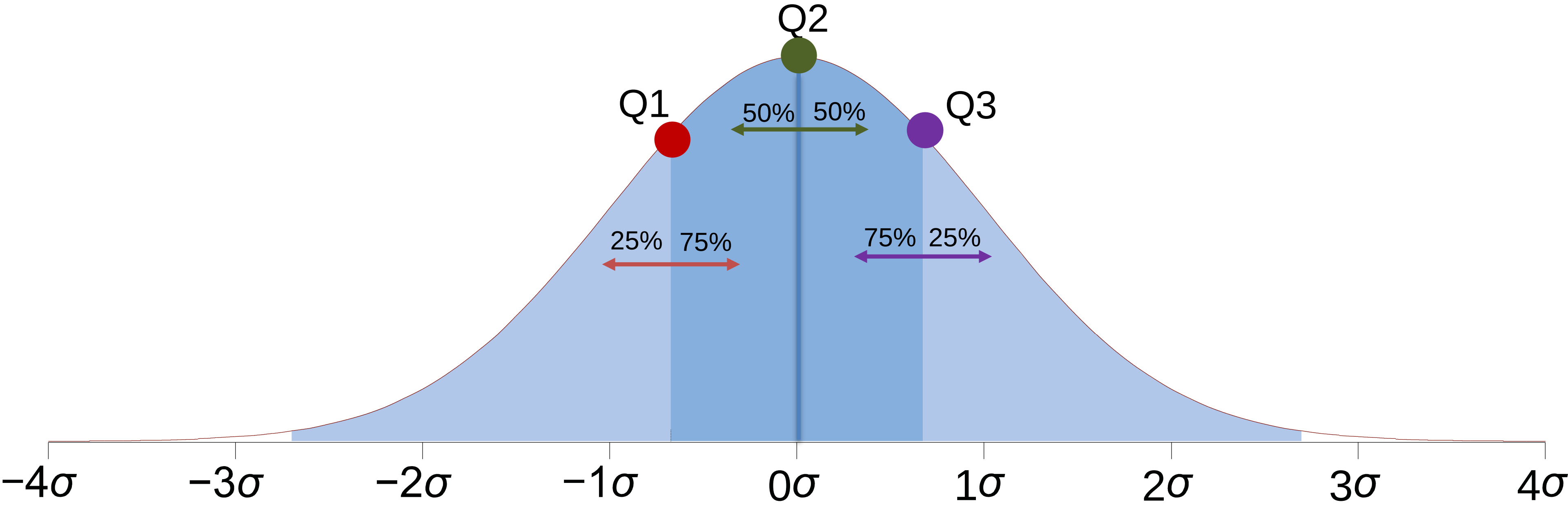
The sample maximum (largest observation)

WHAT'S A QUARTILE?

Q1 = lower quartile / first quartile / 25th percentile

Q2 = median / second quartile / 50th percentile

Q3 = upper quartile / third quartile / 75th percentile



5 NUMBER SUMMARY IN R

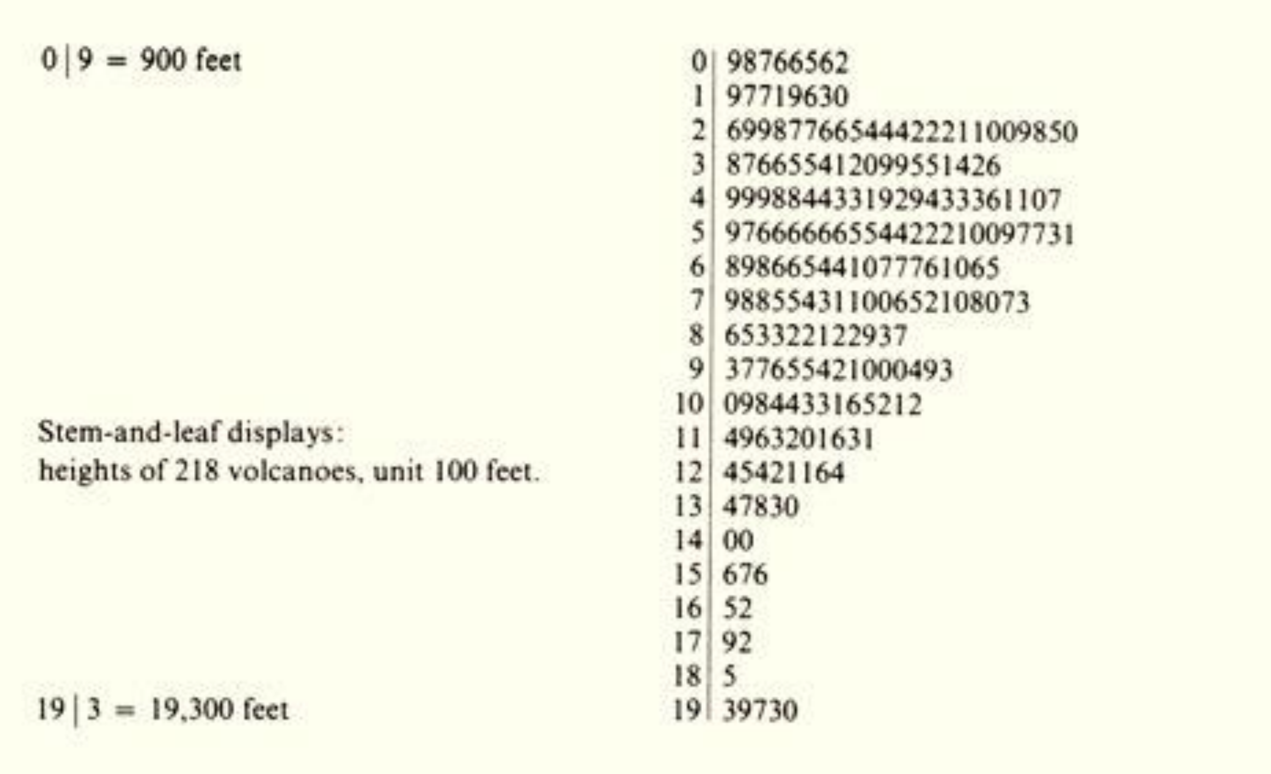
```
➢ moons <- c(0, 0, 1, 2, 63, 61, 27, 13)
➢ fivenum(moons)

[1] 0.0   0.5   7.5   44.0   63.0

➢ summary(moons)

Min. 1st Qu. Median Mean 3rd Qu. Max.
0.0    0.5          7.5      20.88   44.0     63
```

← Note: mean added

STEM-AND-LEAF PLOTS

Volcano heights:

900 feet
1957 feet
823 feet
2620 feet
19300 feet
730 feet
1753 feet
603 feet
2930 feet
12400 feet
650 feet
3663 feet



0 | 9 = 900 feet

Stem-and-leaf displays:
heights of 218 volcanoes, unit 100 feet.

19 | 3 = 19,300 feet

```
 0 | 98766562
 1 | 97719630
 2 | 69987766544422211009850
 3 | 876655412099551426
 4 | 999884433192943336 1107
 5 | 97666666554422210097731
 6 | 898665441077761065
 7 | 988554311006521080 73
 8 | 653322122937
 9 | 377655421000493
10 | 0984433165212
11 | 4963201631
12 | 45421164
13 | 47830
14 | 00
15 | 676
16 | 52
17 | 92
18 | 5
19 | 39730
```

BOX PLOTS

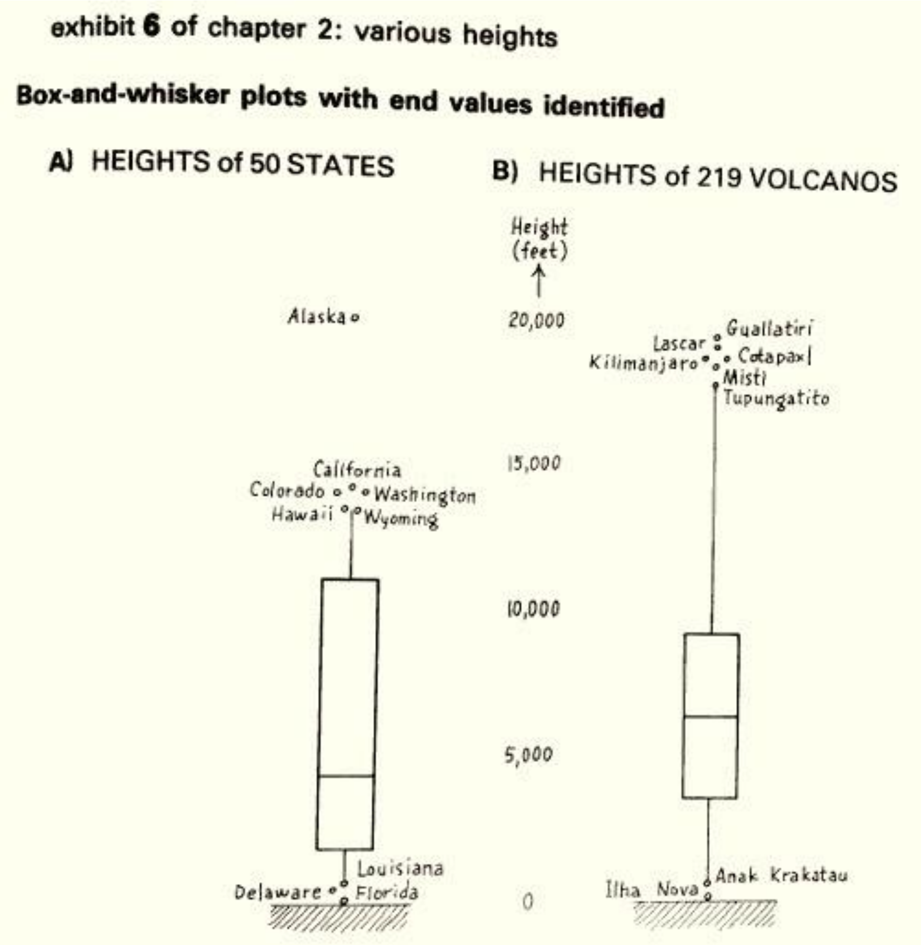


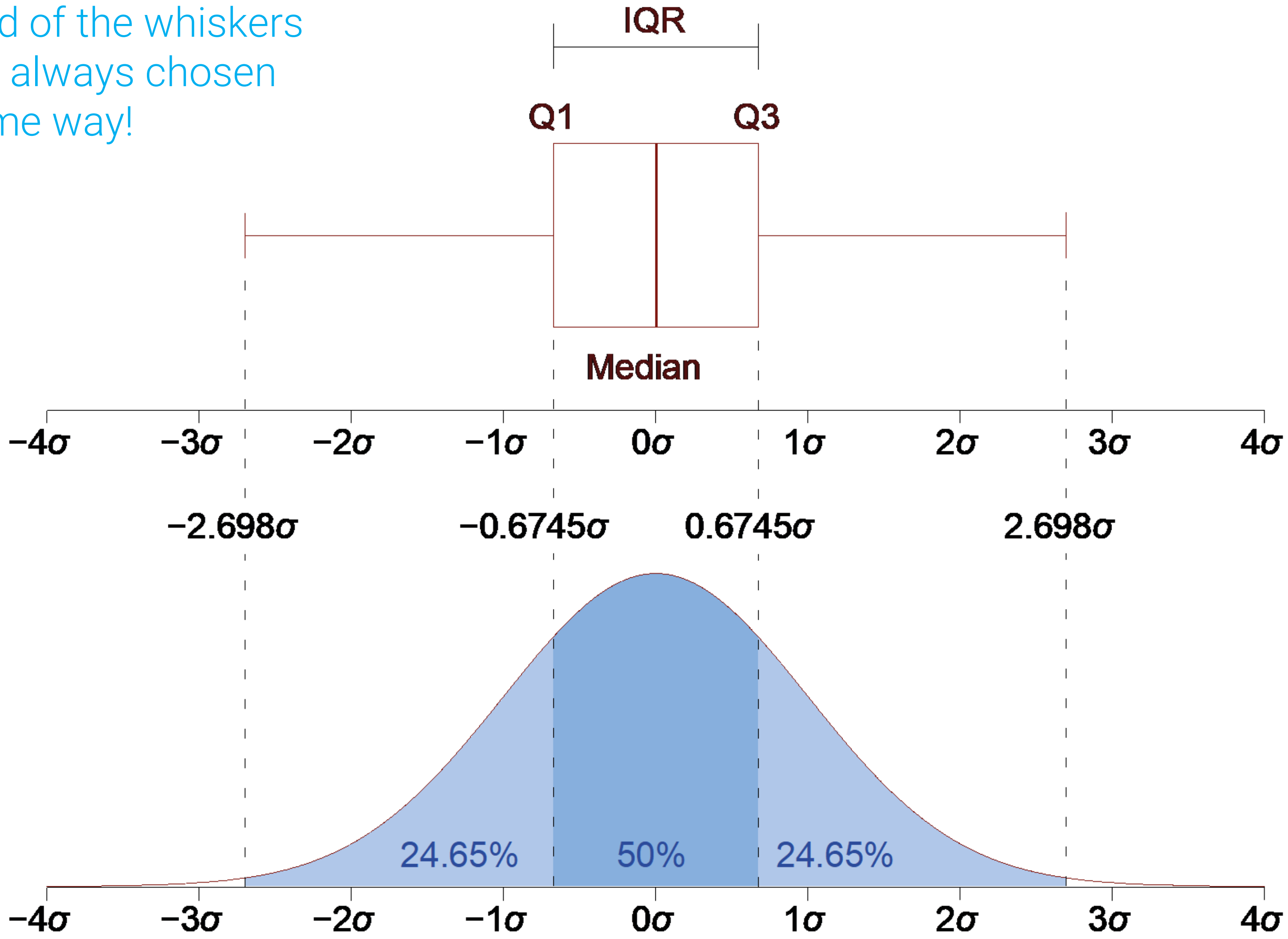
exhibit 6 of chapter 2: various heights

Box-and-whisker plots with end values identified

A) HEIGHTS of 50 STATES

B) HEIGHTS of 219 VOLCANOS

The end of the whiskers are not always chosen the same way!



EXPLORATORY ANALYSIS IS ABOUT UNDERSTANDING DATA AND CHECKING ASSUMPTIONS

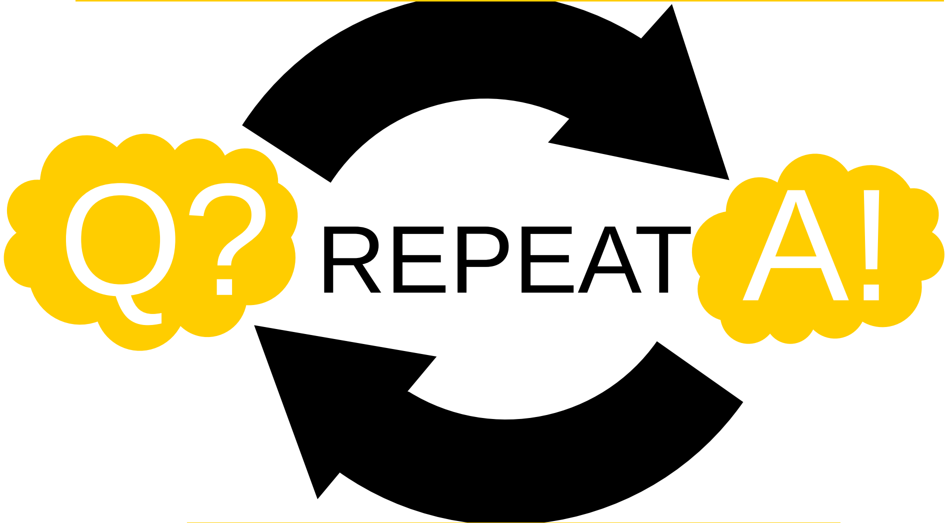
- IS THE DATA **CORRECT**?
- DOES IT **MATCH OUR PREVIOUS EXPECTATIONS**?
- IS THERE **A RELATIONSHIP**?

  A CORRELATION?
  A TREND?
  ETC.?

BUT, HOW SHOULD WE GO ABOUT DOING THIS?

ANALYSIS CIRCLE

GATHERING DATA,
APPLYING STATISTICAL TOOLS,
AND CONSTRUCTING GRAPHICS
TO ADDRESS QUESTIONS



INSPECT “ANSWERS” AND
ASSESS NEW QUESTIONS

START SIMPLE

IT'S EASY TO GET SIDETRACKED TRYING TO DO COMPLICATED ANALYSES AND MISS THE BASIC STUFF

SOME FIRST STEPS TO START WITH

1. Plot the raw data
2. Plot simple statistics
3. Look at plots together

DON'T TRY TO CREATE A WHOLE NEW CHART ALL AT ONCE!
CHECK YOUR LOGIC AT EVERY STEP.

LOOKING AT DATA WITH
“THE PAINTER’S EYE”

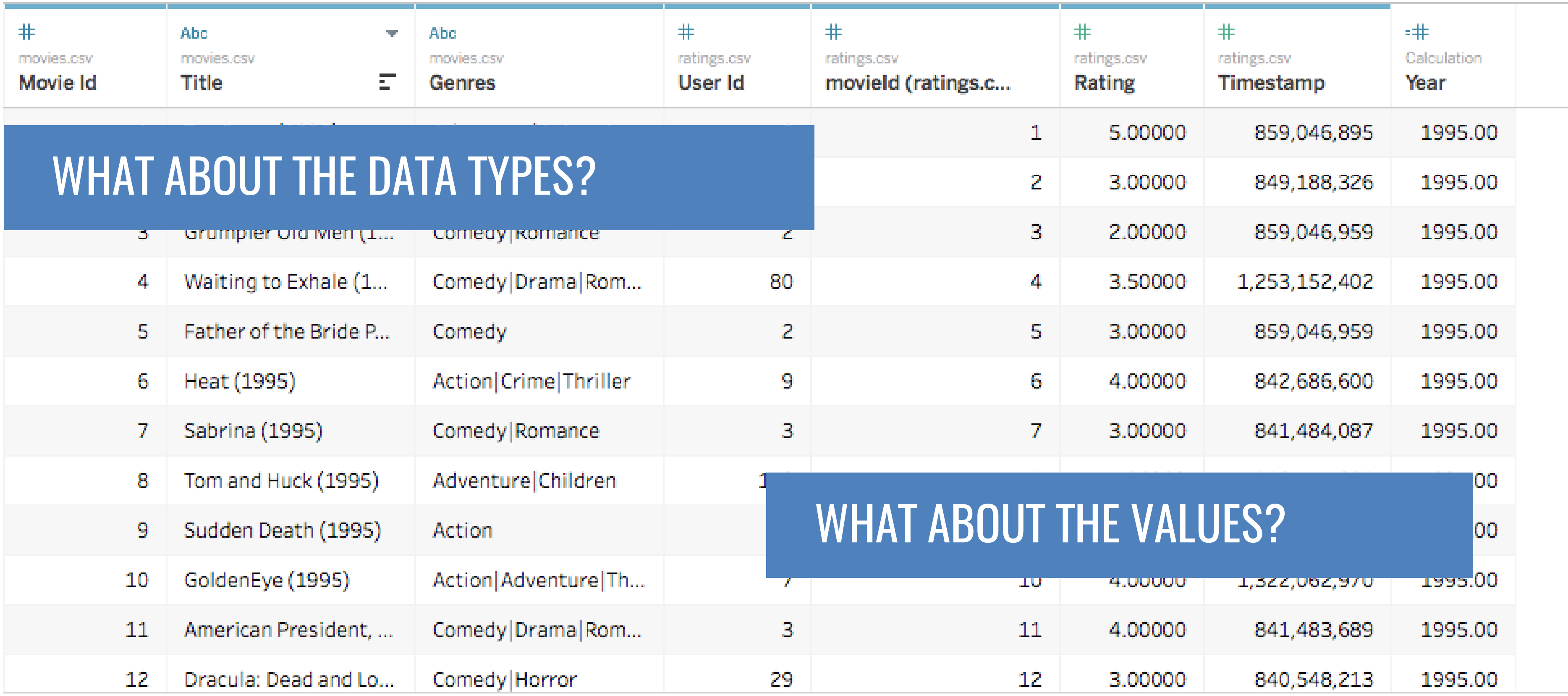
J. BERTIN

EMBRACING
“SLOW DATA”

STEPHEN FEW

PLOT THE RAW DATA

ARE THE FIELDS CORRECT?

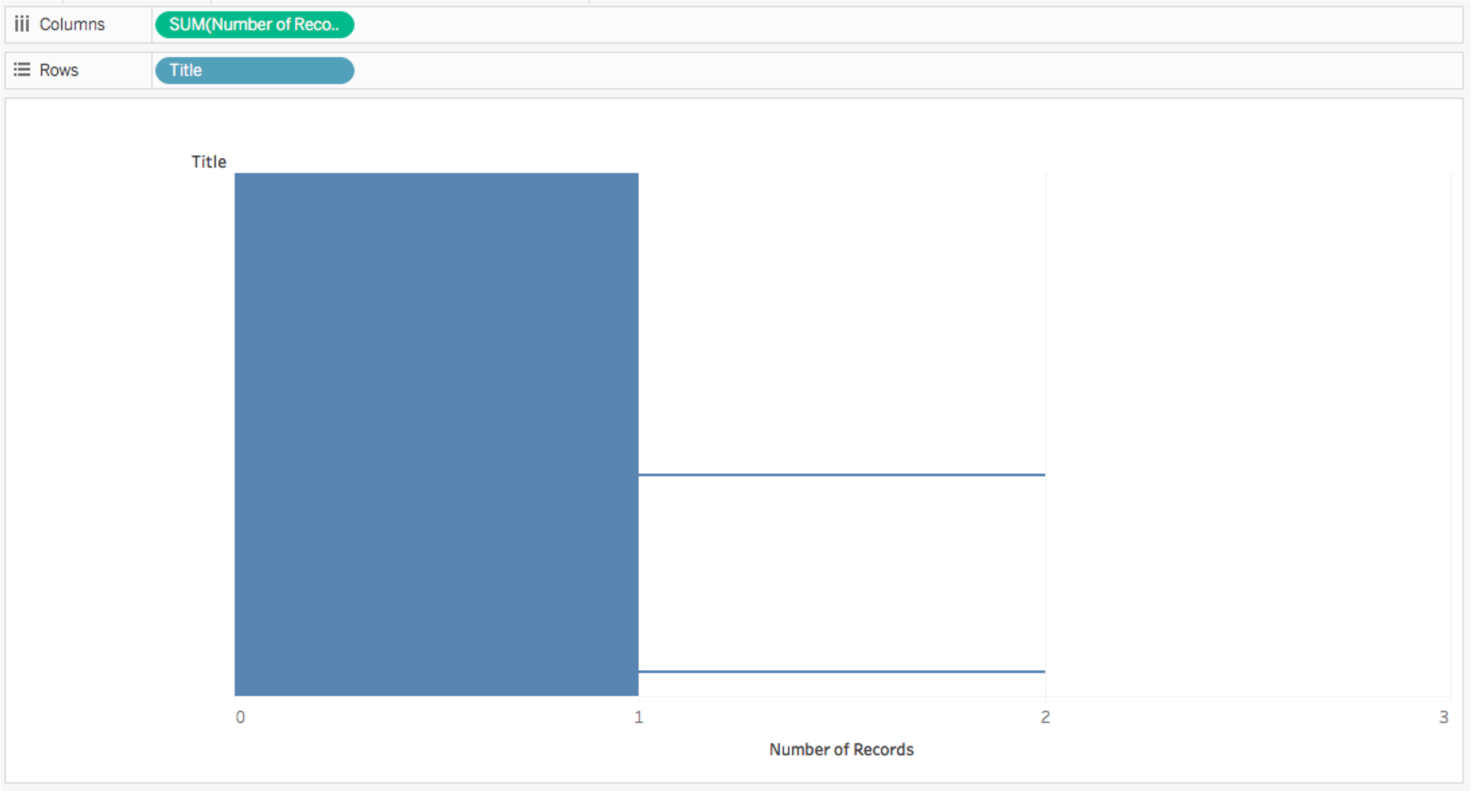


# movies.csv Movie Id	Abc movies.csv Title	Abc movies.csv Genres	# ratings.csv User Id	# ratings.csv movieId (ratings.c...	# ratings.csv Rating	# ratings.csv Timestamp	=# Calculation Year
				1	5.00000	859,046,895	1995.00
				2	3.00000	849,188,326	1995.00
3	Grumpier Old Men (1...	Comedy|Romance	2	3	2.00000	859,046,959	1995.00
4	Waiting to Exhale (1...	Comedy|Drama|Rom...	80	4	3.50000	1,253,152,402	1995.00
5	Father of the Bride P...	Comedy	2	5	3.00000	859,046,959	1995.00
6	Heat (1995)	Action|Crime|Thriller	9	6	4.00000	842,686,600	1995.00
7	Sabrina (1995)	Comedy|Romance	3	7	3.00000	841,484,087	1995.00
8	Tom and Huck (1995)	Adventure|Children	1				00
9	Sudden Death (1995)	Action					00
10	GoldenEye (1995)	Action|Adventure|Th...	7	10	4.00000	1,322,062,970	1995.00
11	American President, ...	Comedy|Drama|Rom...	3	11	4.00000	841,483,689	1995.00
12	Dracula: Dead and Lo...	Comedy|Horror	29	12	3.00000	840,548,213	1995.00

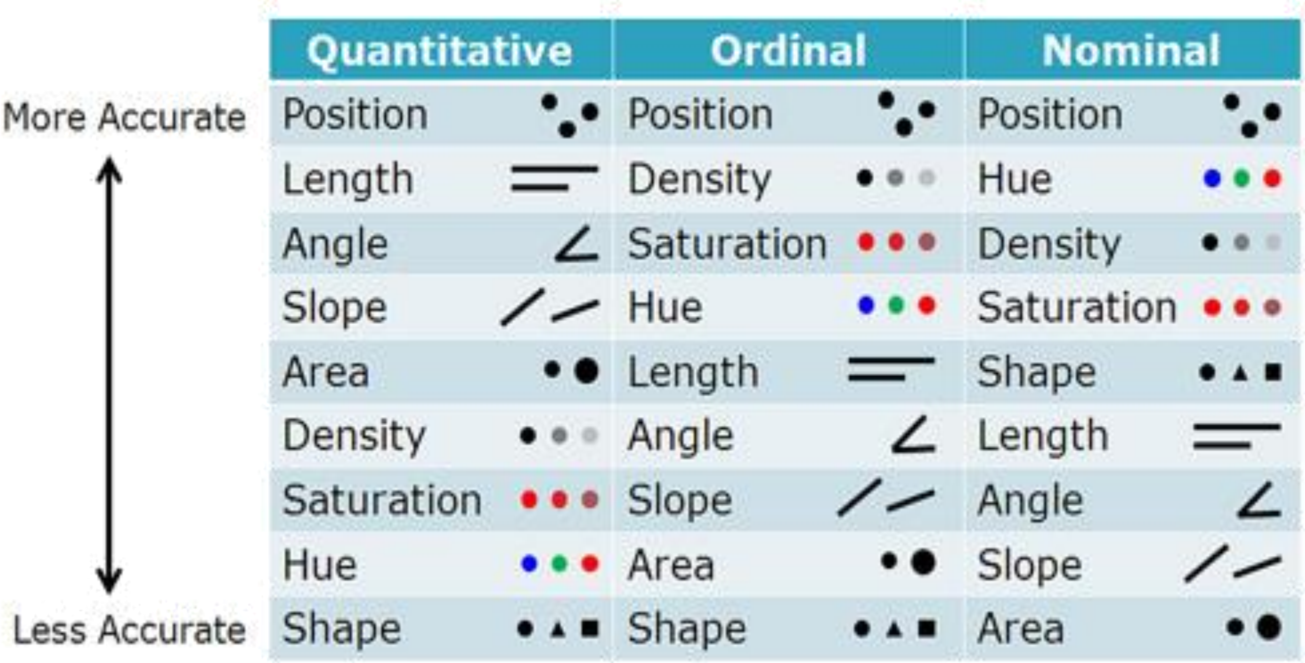
WHAT ABOUT THE DATA TYPES?

WHAT ABOUT THE VALUES?

USE THE SIMPLEST REPRESENTATION YOU CAN TO EVALUATE ALL OF THE DATA



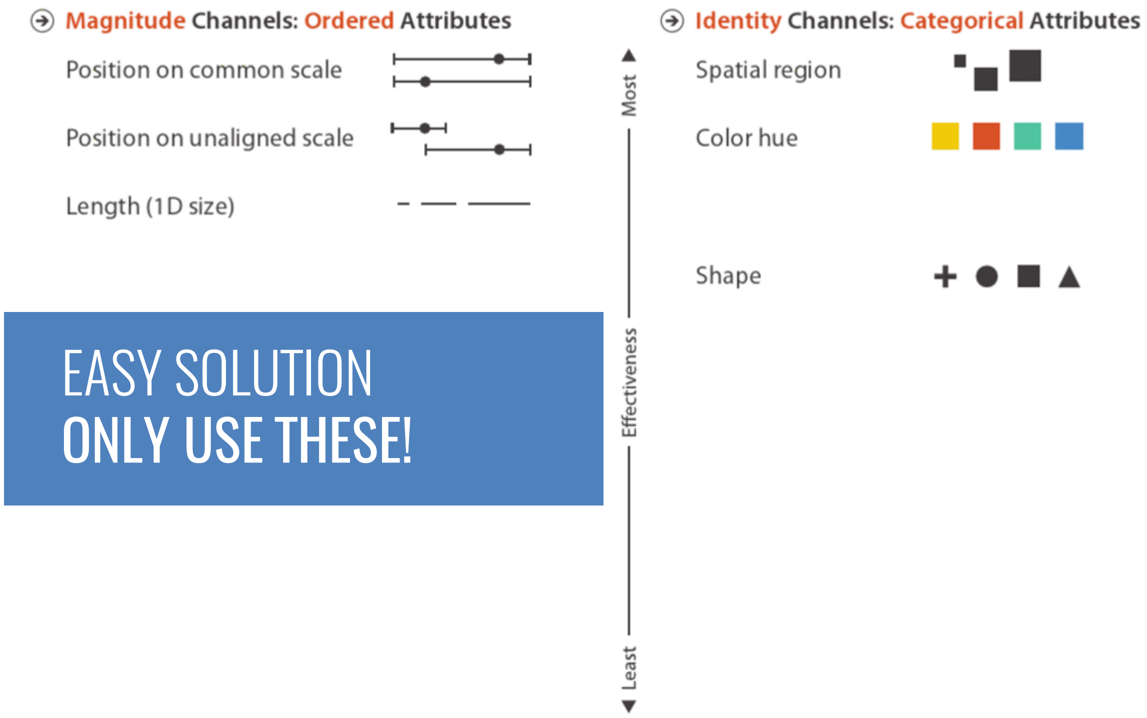
CHOOSE REPRESENTATIONS THAT MAKE IT EASY TO COMPARE DIFFERENCES AND SEE PATTERNS



	Quantitative	Ordinal	Nominal
More Accurate	Position	Position	Position
↑	Length	Density	Hue
	Angle	Saturation	Density
	Slope	Hue	Saturation
	Area	Length	Shape
	Density	Angle	Length
	Saturation	Slope	Angle
↓	Hue	Area	Slope
Less Accurate	Shape	Shape	Area

[JACQUES BERTIN REFINED BY CLEVELAND & MCGILL THEN BY CARD & MACKINLAY]

CHOOSE REPRESENTATIONS THAT MAKE IT EASY TO COMPARE DIFFERENCES AND SEE PATTERNS



Magnitude Channels: Ordered Attributes

Position on common scale

Position on unaligned scale

Length (1D size)

Identity Channels: Categorical Attributes

Spatial region

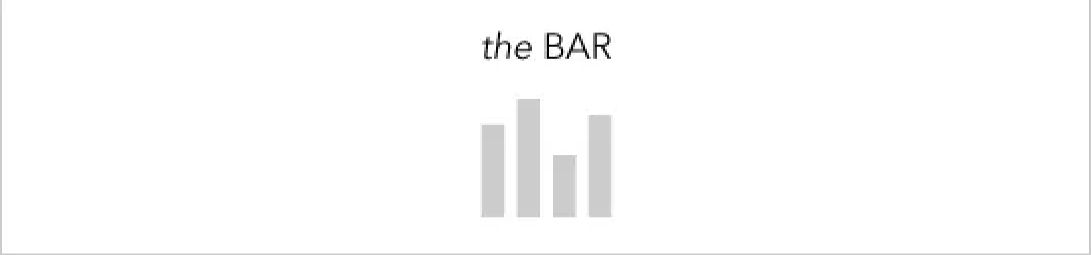
Color hue

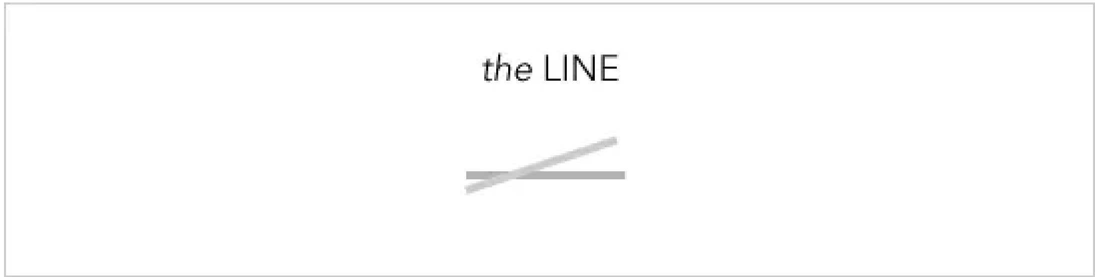
Shape

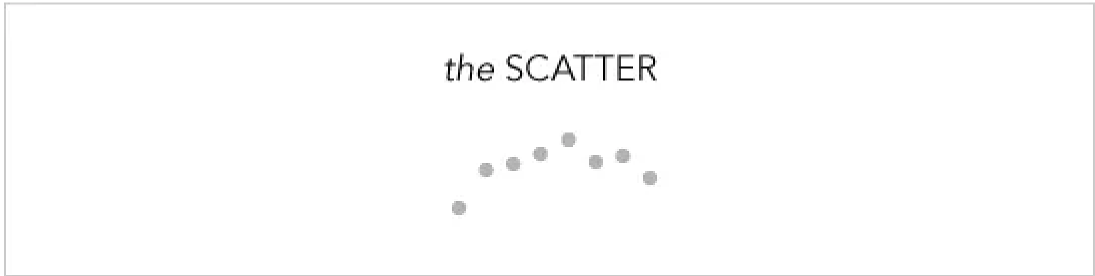
EASY SOLUTION
ONLY USE THESE!

TAMARA MUNZNER

DEFAULT TO SIMPLE AND EFFECTIVE CHART TYPES







+ COLOUR & SHAPE
TO SHOW CATEGORIES

Images from Nathan Yau

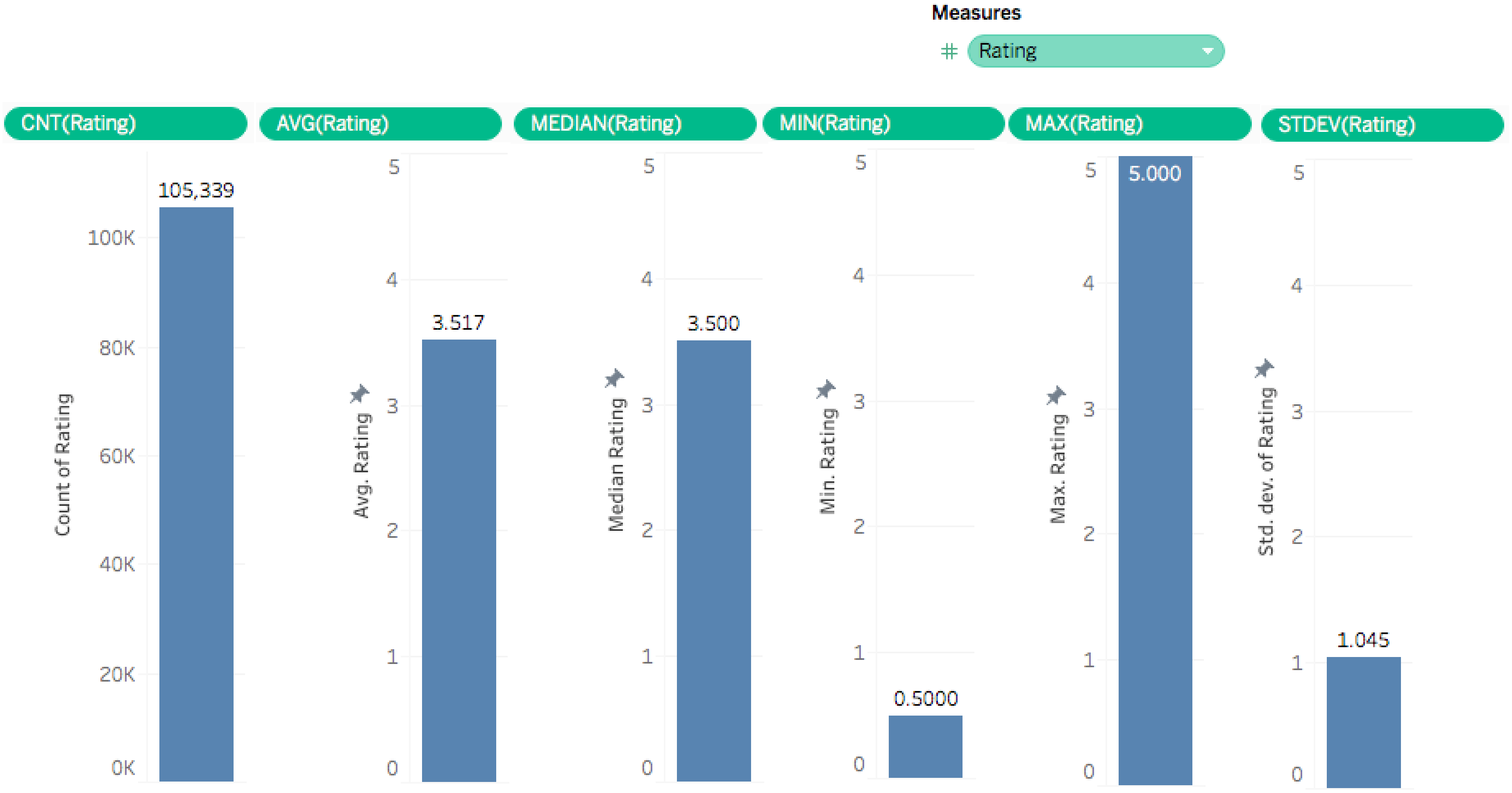
SOME FIRST STEPS TO START WITH

1. Plot the raw data

2. **Plot simple statistics**

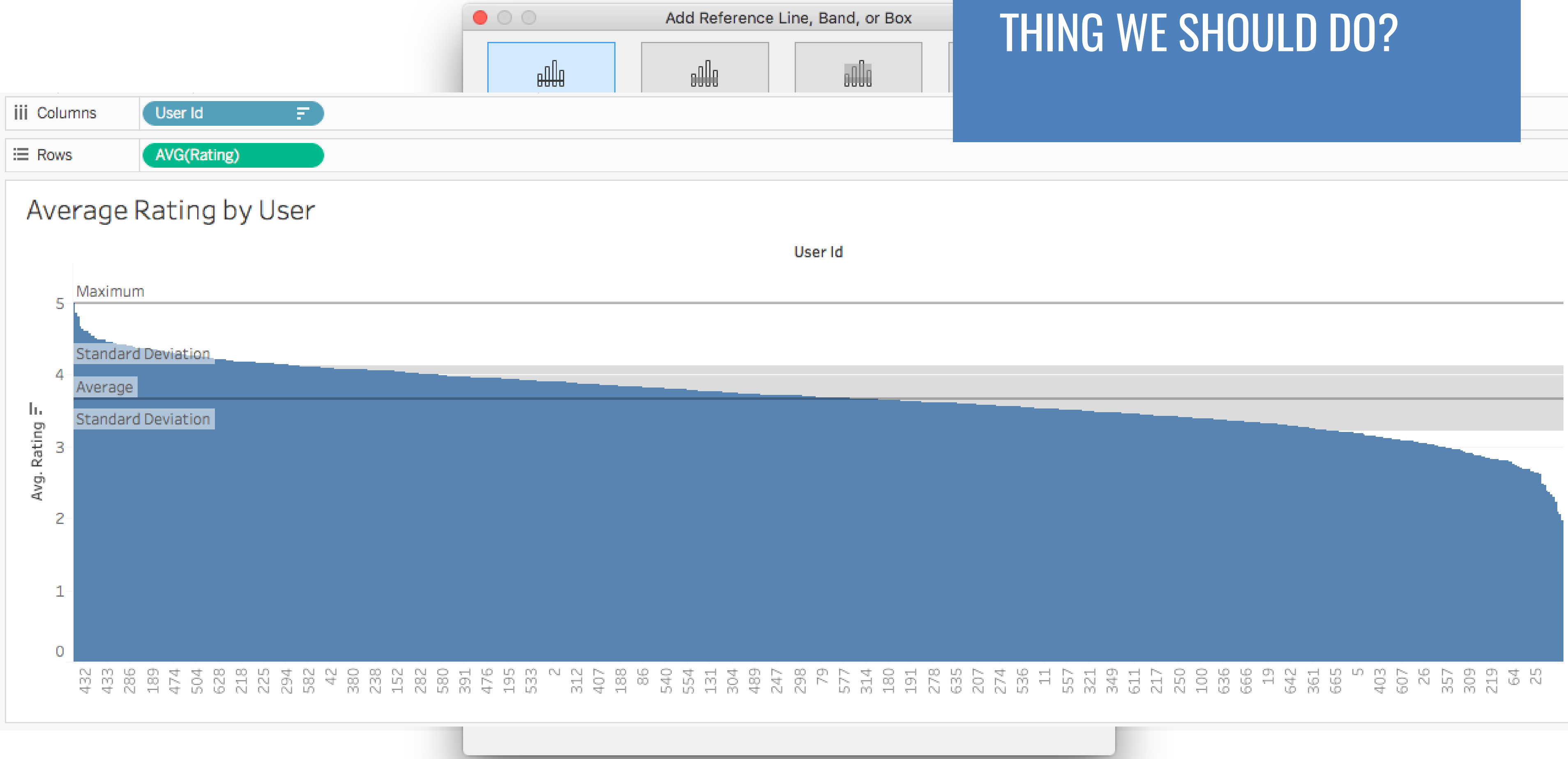
3. Look at plots together

CHECK SIMPLE STATISTICS



CHECK SIMPLE STATISTICS

WHAT'S ONE MORE EASY THING WE SHOULD DO?



Add Reference Line, Band, or Box

Columns: User Id

Rows: AVG(Rating)

Average Rating by User

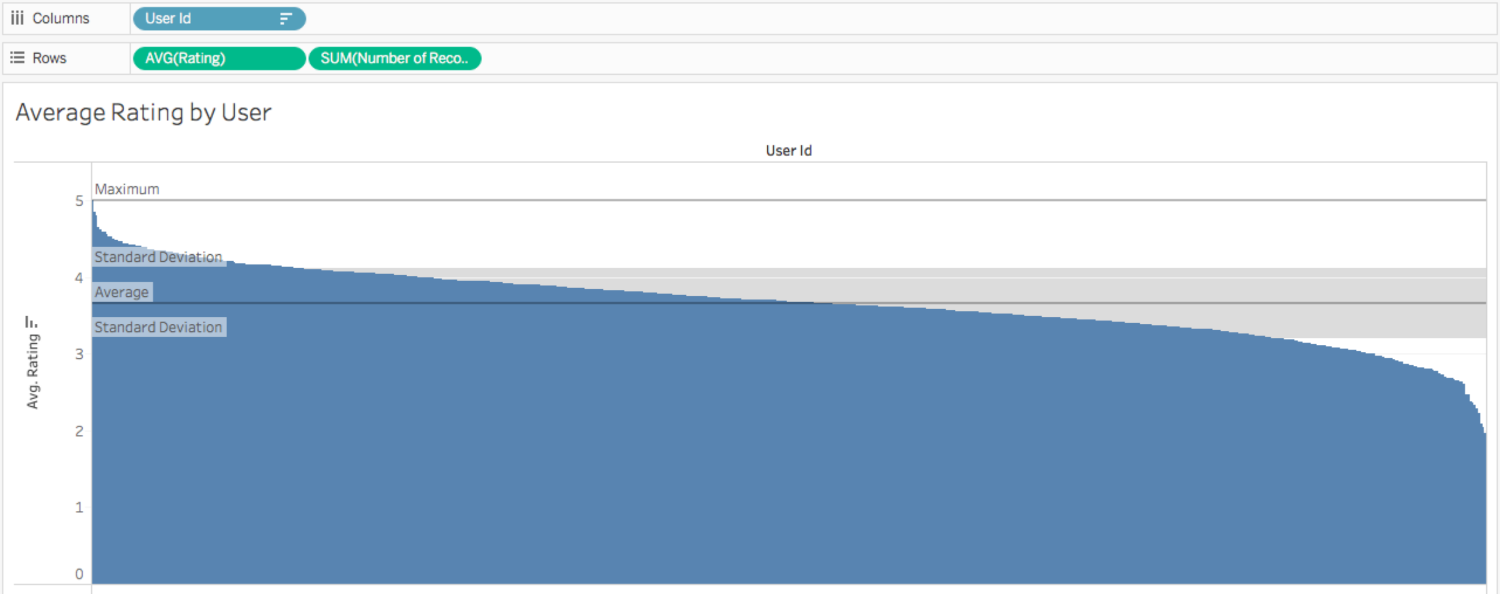
SOME FIRST STEPS TO START WITH

1. Plot the raw data

2. Plot simple statistics

3. **Look at plots together**

COMPARE MULTIPLE PLOTS



UNDERSTANDING DISTRIBUTIONS

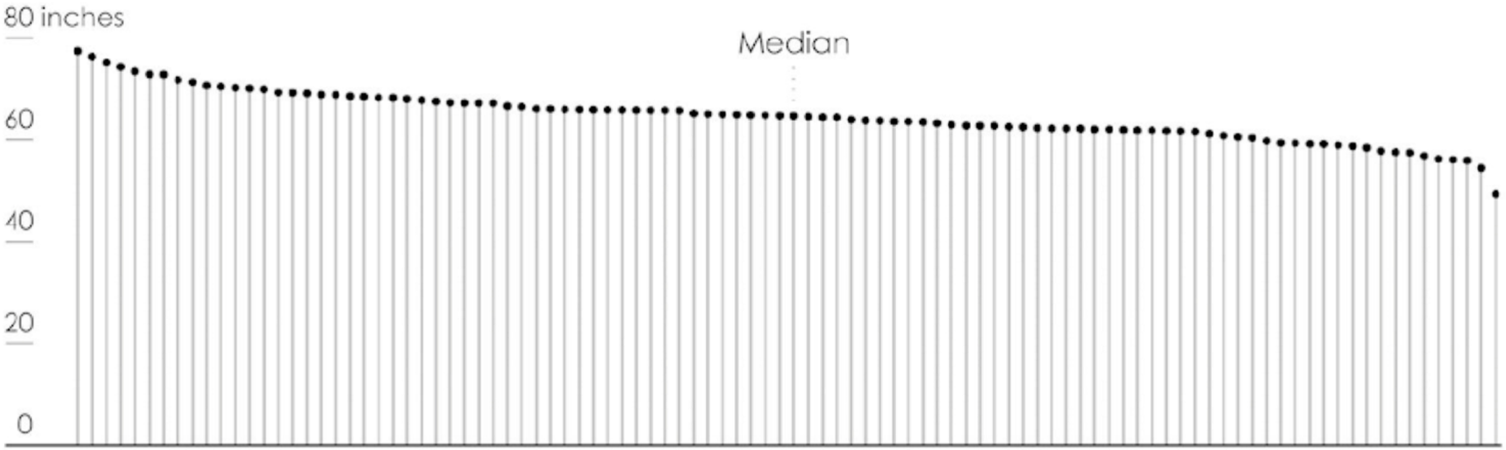


FIGURE 4-52 *Heights of imaginary people, sorted from shortest to tallest*

YAU 2013

ASKING PEOPLE

Requirement: we have stakeholders, not necessarily data

QUESTIONS FROM STAKEHOLDERS

ELICITATION

ELICITATION

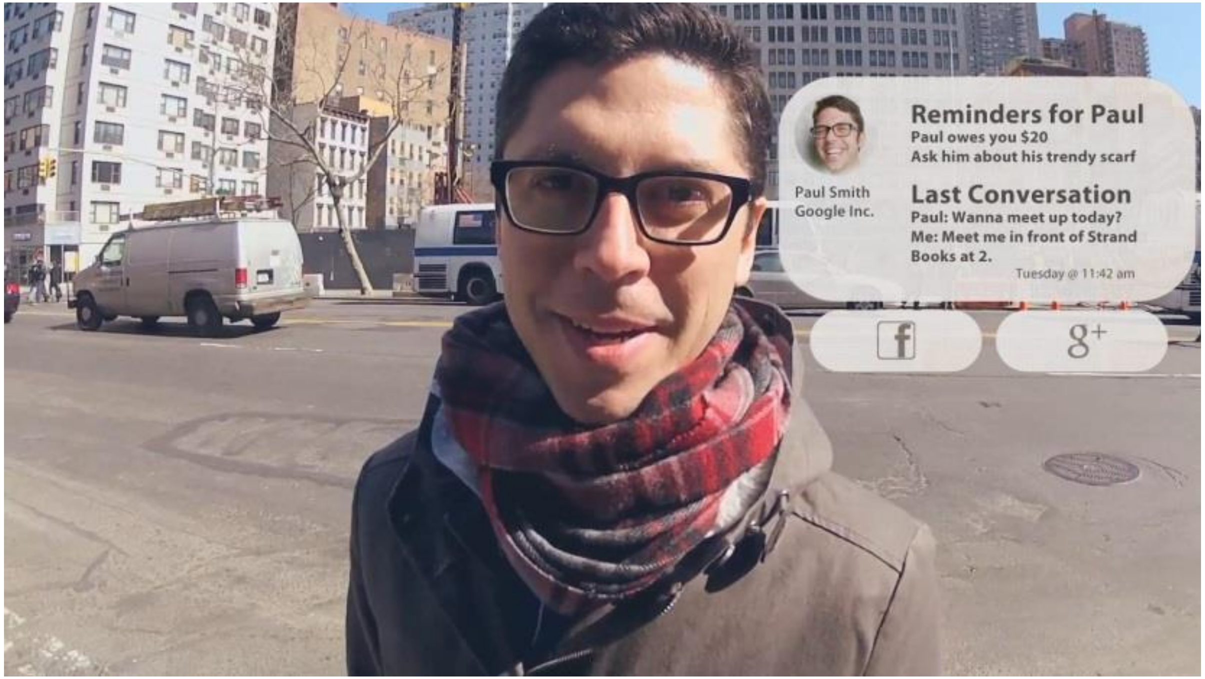
= GATHERING INFORMATION DIRECTLY FROM PEOPLE

ELICITATION IN RELATED FIELDS

In Human-Computer Interaction

WHY IS UI DESIGN HARD?

We’ve never “seen” it before



Paul Smith
Google Inc.
Reminders for Paul
Paul owes you $20
Ask him about his trendy scarf
Last Conversation
Paul: Wanna meet up today?
Me: Meet me in front of Strand Books at 2.
Tuesday @ 11:42 am
f
g+

WHY IS UI DESIGN HARD?

- We've never "seen" it before
- We aren't the people using it

WHY IS UI DESIGN HARD?

- We’ve never “seen” it before
- We aren’t the people using it
- We can’t anticipate how people will use it

WHY IS UI DESIGN HARD?

- We’ve never “seen” it before
- We aren’t the people using it
- We can’t anticipate how people will use it

WHY IS ANALYSIS HARD?

ARE THERE PROCESSES THAT CAN BE FOLLOWED?

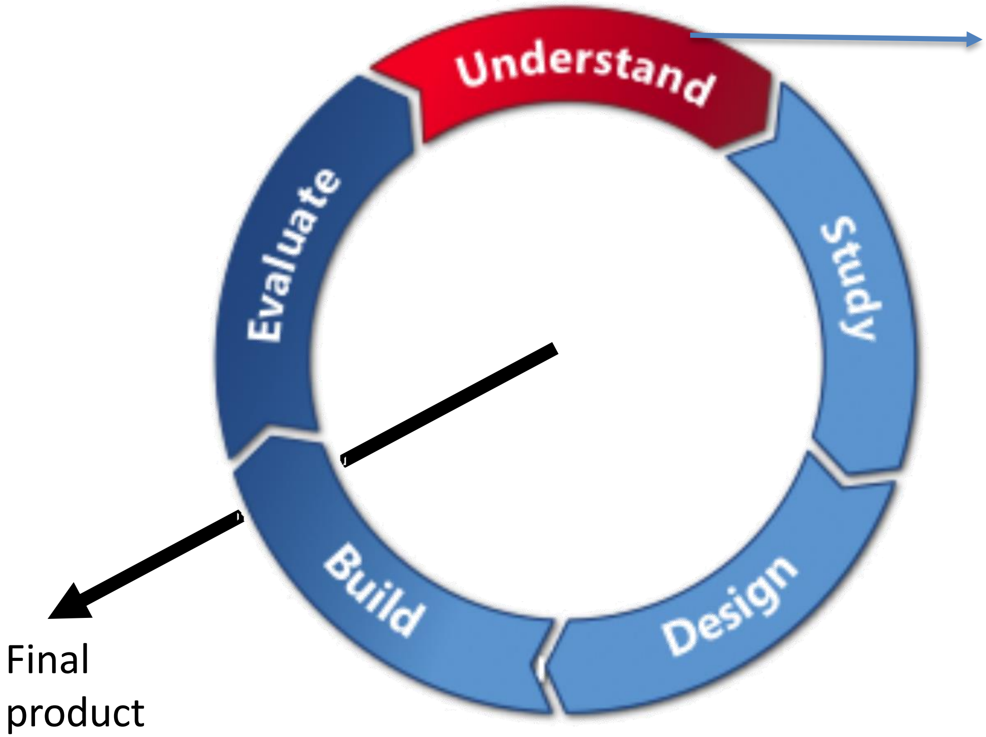
THE USER-CENTERED APPROACH

- early focus on users and tasks
- empirical measurement
- iterative design

FOUR BASIC ACTIVITIES

1. establishing requirements
2. designing alternatives
3. prototyping
4. evaluating

THE DESIGN LIFECYCLE



- what human values do we wish to design for?
- what are the various morale, personal, and social impacts of the proposed system?

Harper et al., 2008

HOW DOES THIS AFFECT ME?

YOU ARE AN ANALYTIC TOOL DESIGNER / DEV?

→You will go through this cycle

YOU ARE THE ANALYST

→ You will go through a version of this cycle

For you to think about:
How does the design life cycle relate to the analysis cycle we looked at earlier?

BACK TO: ELICITATION

Or .. Establishing requirements

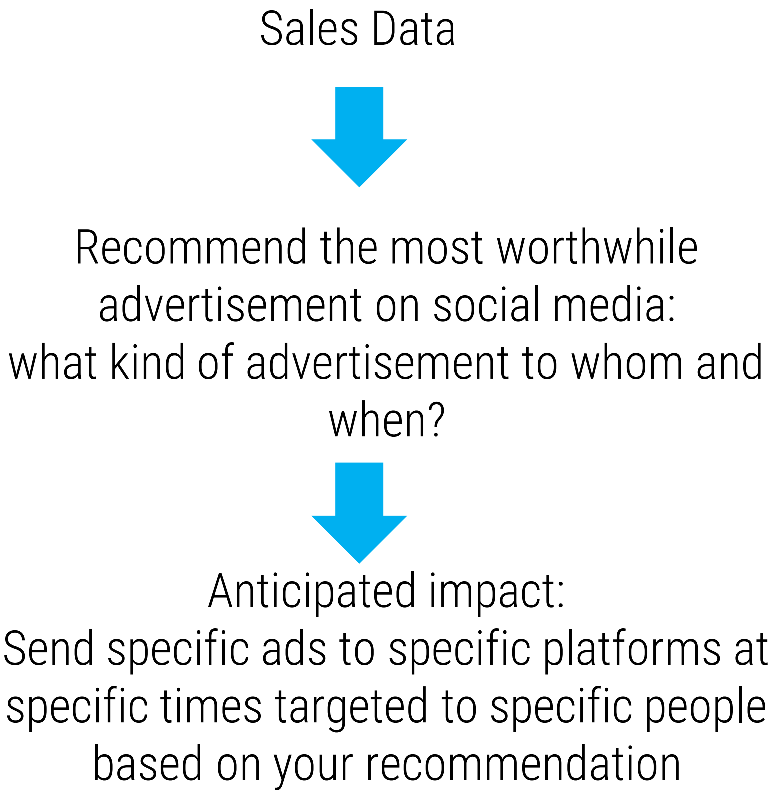
1) IDENTIFY STAKEHOLDERS

STAKEHOLDERS

Anyone who is affected by your data analysis project or might have a strong interest in it

Owners
Deciders
Doers
Consumers

EXAMPLE



Sales Data

Recommend the most worthwhile advertisement on social media: what kind of advertisement to whom and when?

Anticipated impact:
Send specific ads to specific platforms at specific times targeted to specific people based on your recommendation

Who are potential stakeholders?

- The person who hired you
- The person who is responsible for ads in the company
- The people who have to implement you recommendations
- The database people delivering data to you
- Other departments who might want to use your recommendations
- Governments, e.g. if you might invade someone’s privacy

IDENTIFY THE MOST IMPORTANT STAKEHOLDERS

The list can get very large

Which people will most affect your project or benefit from your project

QUESTIONS TO IDENTIFY KEY STAKEHOLDERS

1) Is the stakeholder importantly impacted by your work or strongly impacts your work or performance?
2) Can you identify what you want from the stakeholder?
3) Do you want a dynamic relationship with the stakeholder?
4) Can you exist without or easily replace the stakeholder?
5) Have you already included the stakeholders in another group of people?

https://hbr.org/2014/03/five-questions-to-identify-key-stakeholders

2) ELICIT INFORMATION

FROM STAKEHOLDERS

LEARN MOTIVATIONS & EXPECTATION FOR YOUR ANALYSIS

Goal

STEPS

- Articulate concrete descriptions of stakeholders (roles in analysis, interests, …)
- Use these descriptions to determine which types of questions you need to ask them

RESEARCH METHODS

observing and/or interviewing stakeholders of your analysis

- find out what current analysis methods they use, what data they have, what they really need (depending on their role)
- go from abstract stakeholders → real people with real needs

example:
if you are doing an analysis to aid the sales department target their sales, observe them in how they currently do this

IF YOU CAN'T MEET STAKEHOLDERS

- carefully select and interview their representatives
- MUST be people with direct contact with stakeholders and intimate knowledge and experience of their needs and what they do
- people who work with them are the best

Example:
talk to front-line sales staff about their customers if you cannot observe or talk to customers directly. Better: interview/observe front-line staff as they deal with customers

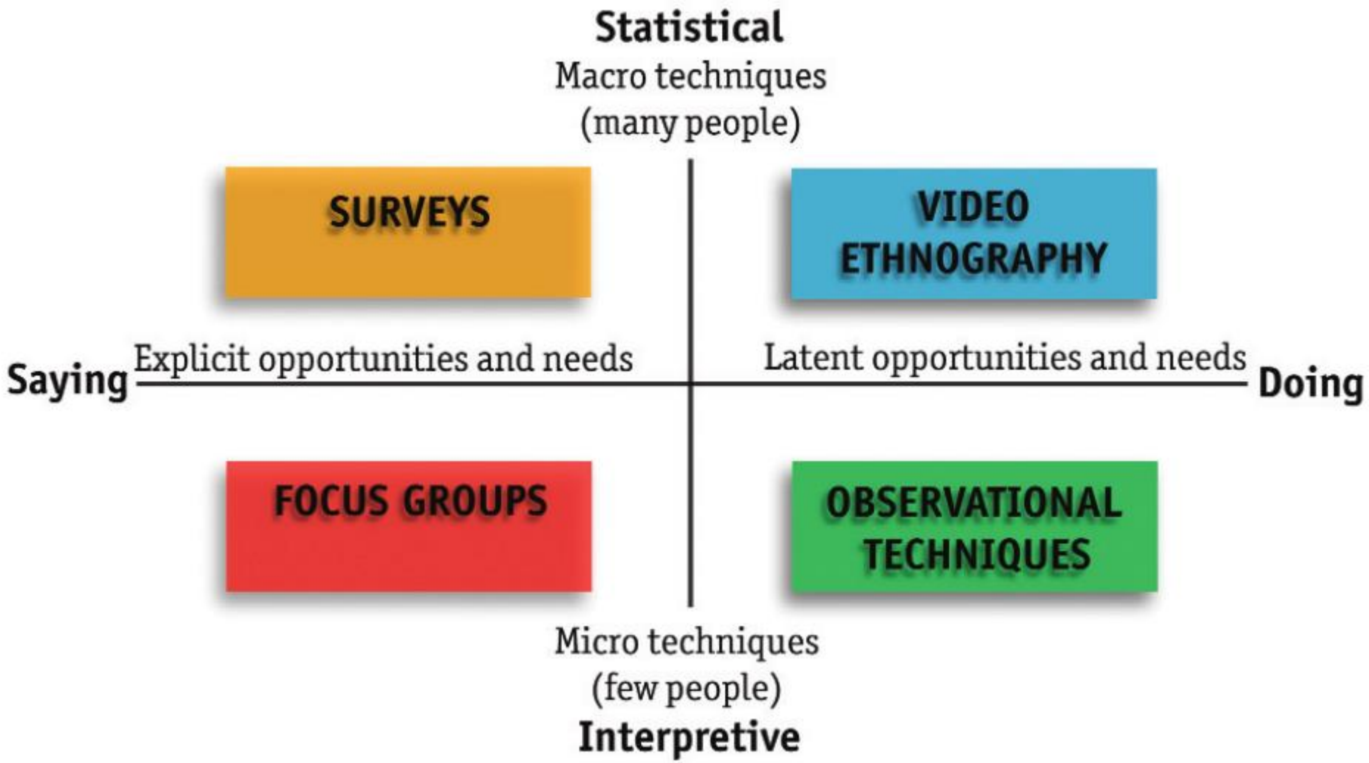
IF ALL ELSE FAILS

make your beliefs about the stakeholders and their needs explicit

- if you cannot get in touch with stakeholders or their representatives
- use your team to articulate their assumptions about stakeholders and their needs/tasks
- risk: resulting descriptions do not resemble reality → only use as last resort

RESEARCH METHODS

categories and examples (there are more methods than just these)



From: Moggridge – Designing Interactions

RESEARCH METHODS

from the analyst's perspective:

- **observe**: stakeholders and their behavior in context
- **engage**: interact with and interview stakeholders
- **immerse**: experience what stakeholders experience

OBSERVATION METHODS

Look

(SOME) OBSERVATION METHODS

- A Day in the Life
- Behavioral Archaeology
- Behavioral Mapping
- **Fly on the Wall**
- Guided Tours
- Personal Inventory
- Rapid Ethnography
- **Shadowing**
- Social Network Mapping
- Still-Photo Survey
- Time-Lapse Video

GENERAL OBSERVATION METHODS

- natural
  - no interference from the investigator
- controlled
  - the investigator sets a task and observes it being carried out
- participatory
  - the investigator actively joins in the activity being observed to gain a firsthand activity

ASK THEM TO HELP

Ask

WHEN LOOKING IS NOT ENOUGH…

- LOOKing gives you great insight into the state of the world
- but it doesn't tell you <u>why</u> people are acting the way they do, or what their goals, needs, or feelings are

katieb50 on flickr

PROBLEMS WITH ASKING

- people can be unduly influenced by cultural context (hype), and what they think you expect them to say (this rocks!) (remember the iphone 5 video I showed you)
- people may lie—deliberately to save face (embarrassment, cultural / polite)
- people may lie—their boss is around

WAIT, ARE PEOPLE COMPLETELY USELESS?

people are really good at telling us a few things:

- what they are doing right now.
- how they are feeling right now.
- what their goal is right now.

IDEALLY, COMBINE INTERVIEW WITH OBSERVATION

- watch people in their own environment
- watch people do everyday tasks

- opportunities for new questions arise from:
  - workarounds
  - breakdowns
  - unexpected uses of existing tools/methods

(SOME) ASKING METHODS

- Camera Journal
- Card Sort
- Cognitive Maps
- Collage
- Conceptual Landscape
- Draw the Experience
- Extreme User **Interviews**
- Five Whys?
- Foreign Correspondents
- Narration
- **Surveys & Questionnaires**
- Unfocus Group
- Word-Concept Association

METHOD: INTERVIEWS

Types:

- Unstructured - exploratory and in-depth
- Structured - are scripted with pre-written questions
- Semi-structured - guided by a script but can become more open as it progresses
- Group (focus groups) - allows diversity and more views/issues to be raised and reflected on

METHOD: INTERVIEWS

Two question types

- ‘closed questions’ have a predetermined answer format, e.g., ‘yes’ or ‘no’
- ‘open questions’ - no predetermined format

TYPES OF QUESTIONS

- What has been tried before?
- How did it turn out?
- What do you think needs to be done?
- …

METHOD: SURVEYS & QUESTIONNAIRES

- ask a series of targeted questions in order to ascertain particular characteristics and perception of users
- this is a quick way to elicit answers from a large number of people

example:
developing a new gift-wrap packaging concept the IDEO team conducted web-based surveys to collect consumer perspectives from many people around the world

SURVEYS & QUESTIONNAIRES

very popular method

- good for finding out about attitudes, values, opinions, likes and dislikes
- can be administered to large populations, web-based, paper or email
- sampling can be a problem when size of population is unknown
- can be offputting to people if appears too long
- 40% response rate is high, 20% is often acceptable

QUESTIONNAIRE CONTENT

- be clear on the goal
- open and closed questions
  - What do you think about X?
  - Which of the following are things you might use?
    - a, b, c, d, e
- rating scales
  - I think X is a good idea
    - 1 strongly disagree to 5 strongly agree
- be sure to pilot your questionnaire

QUESTIONNAIRE DESIGN

how it is structured is key

- impact of a question can be influenced by its order
- strike a balance between using white space and keeping the questionnaire compact
- decide whether phrases will all be positive, all negative or mixed
- providing check boxes and drop down menus to choose from - makes it easier to fill in
- open-ended questions allow for more interview-like comments

ASK & LOOK

Often observations and asking are combined

METHODOLOGY: ETHNOGRAPHY

- collection of methods
- includes field work done in natural settings
  - Spend as much time as you can with people relevant to the design topic.
  - Establish their trust in order to visit and/or participate in their natural habitat and witness specific activities
- study of the large picture
  - get more complete context of activities
  - get objective perspective with rich description of people, environments, and interactions
  - use a “wide-angle research lens”
- goal: elicit user requirements that would be hard for a typical user to articulate
- very (!) time intensive

ETHNOGRAPHIC METHOD: CONTEXTUAL INQUIRY

- combining “looking” and “asking” by immersing oneself into a particular context/culture: *understand mental models and work practices*

- “the core premise of Contextual Inquiry is very simple:
  - go where the customer works,
  - observe the customer as he or she works, and
  - talk to the customer about the work.

  do that, and you can’t help but gain a better understanding of your customer.”

AFTER HAVING DONE ALL THIS…

What’s next?

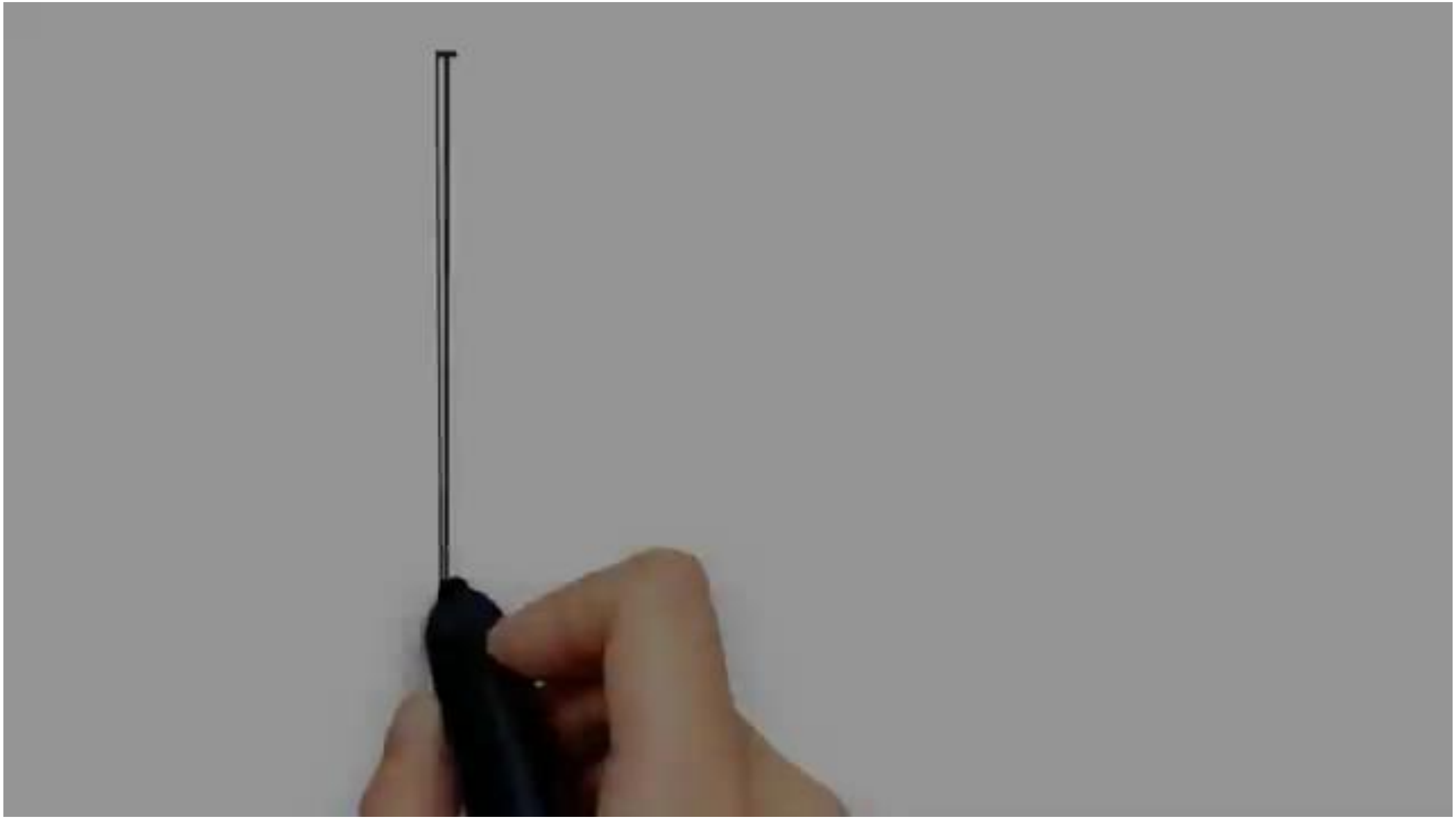
www.projectcartoon.com

What the customer really needed

http://www.projectcartoon.com/cartoon/2

IDENTIFY DATA & VARIABLES FOR YOUR ANALYSIS

FIND OUT IF STAKEHOLDERS AGREE ABOUT THE PROBLEM YOU WILL TRY TO ADDRESS



From: © Coursera

TYPES OF RESEARCH QUESTIONS

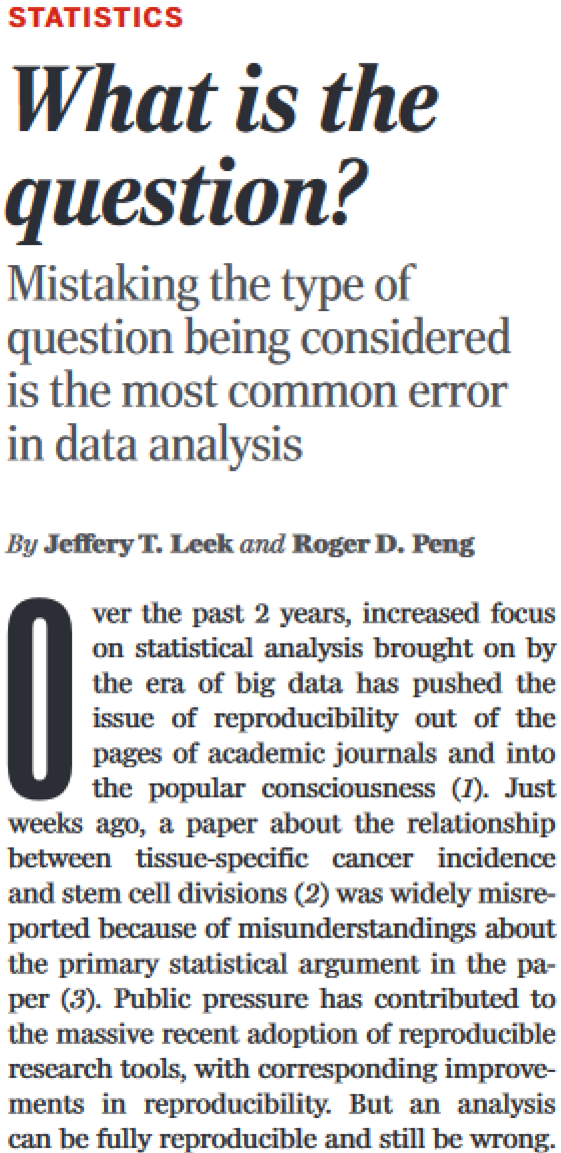
TOPIC: VISUALIZATION RESEARCH

Imagine you would like to communicate data about visualization research

RESEARCH QUESTIONS

- Simple & boring
  - Numbers of papers at IEEE VIS 2015
- Boring
  - Numbers of papers by P. Isenberg in 2015
- Interesting (unfortunately not simple)
  - In the domain of visual analytics growing or shrinking?
  - Are visual analytics and visualization the same community?
  - Are research interests of specific researchers changing?
  - What are new research trends in visual analytics?
  - To which university should I go to do a PhD in visual analytics?
  - Who are good reviewers for a certain topic?
  - Who should be in the program committee of VAST / VIS 2017?
  - How does a change in affiliation impact a researcher's interests?
  - I there a relation between affiliation and citations?
  - Are there gender biases in the domains of visualization? How do they compare to computer science in general?

QUESTION TYPES



STATISTICS

What is the question?

Mistaking the type of question being considered is the most common error in data analysis

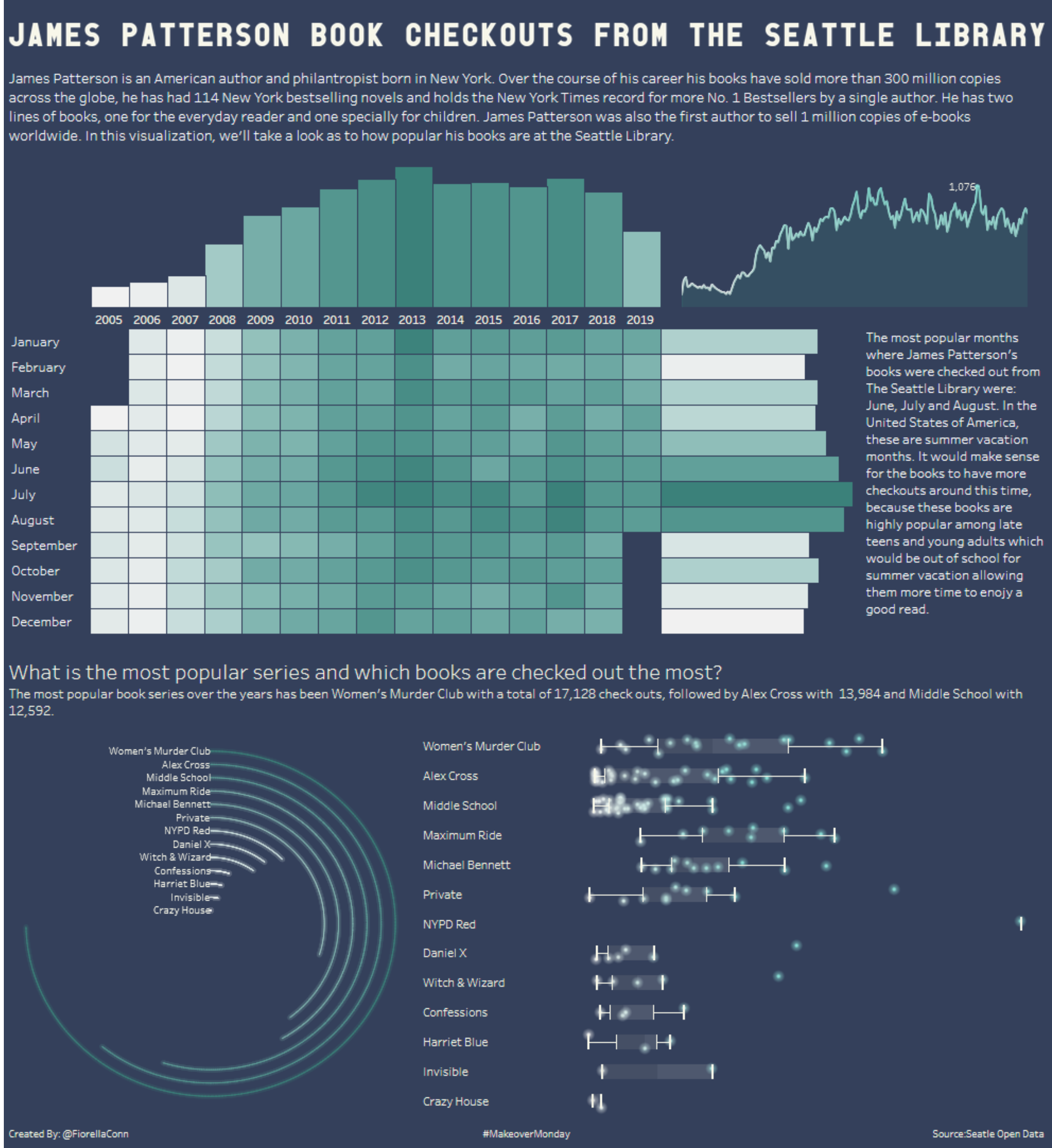
By **Jeffery T. Leek** *and* **Roger D. Peng**

Over the past 2 years, increased focus on statistical analysis brought on by the era of big data has pushed the issue of reproducibility out of the pages of academic journals and into the popular consciousness (*1*). Just weeks ago, a paper about the relationship between tissue-specific cancer incidence and stem cell divisions (*2*) was widely misreported because of misunderstandings about the primary statistical argument in the paper (*3*). Public pressure has contributed to the massive recent adoption of reproducible research tools, with corresponding improvements in reproducibility. But an analysis can be fully reproducible and still be wrong.

1. Descriptive
2. Exploratory
3. Inferential
4. Predictive
5. Causal
6. Mechanistic

DESCRIPTIVE

Describing something, mainly functions and characteristics



JAMES PATTERSON BOOK CHECKOUTS FROM THE SEATTLE LIBRARY

James Patterson is an American author and philantropist born in New York. Over the course of his career his books have sold more than 300 million copies across the globe, he has had 114 New York bestselling novels and holds the New York Times record for more No. 1 Bestsellers by a single author. He has two lines of books, one for the everyday reader and one specially for children. James Patterson was also the first author to sell 1 million copies of e-books worldwide. In this visualization, we'll take a look as to how popular his books are at the Seattle Library.

The most popular months where James Patterson's books were checked out from The Seattle Library were: June, July and August. In the United States of America, these are summer vacation months. It would make sense for the books to have more checkouts around this time, because these books are highly popular among late teens and young adults which would be out of school for summer vacation allowing them more time to enjoy a good read.

What is the most popular series and which books are checked out the most?

The most popular book series over the years has been Women's Murder Club with a total of 17,128 check outs, followed by Alex Cross with 13,984 and Middle School with 12,592.

Created By: @FiorellaConn

#MakeoverMonday

Source:Seatle Open Data

https://public.tableau.com/en-us/gallery/james-patterson-popularity-seattle-library?tab=viz-of-the-day&type=viz-of-the-day

EXPLORATORY

you analyze the data to see if there are;

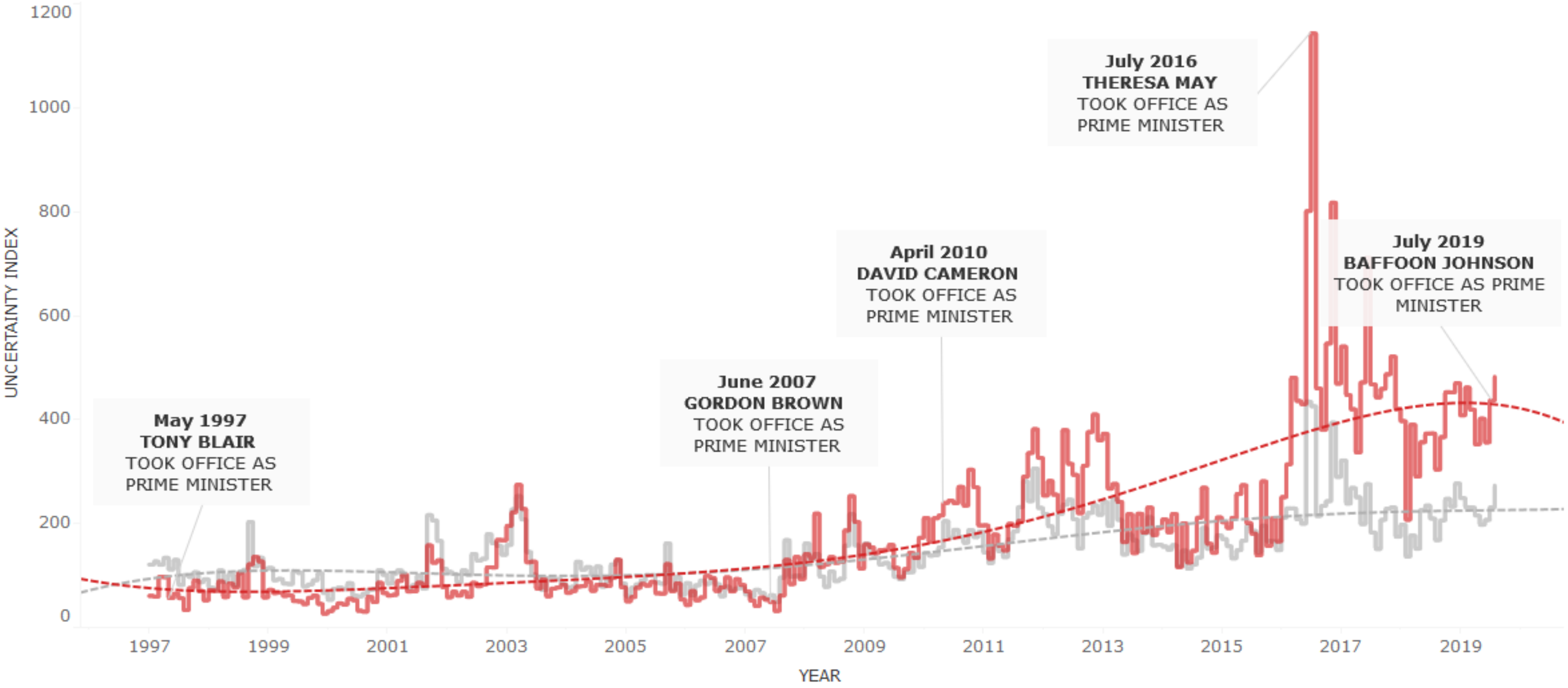
- *patterns*
- *trends*
- *or relationships between variables*

→ Generate hypotheses

HOW UNCERTAIN IS ECONOMIC POLICY IN THE UK AND EUROPE 1997-2019

EUROPEAN ECONOMIC POLICY UNCERTAINTY NEWS INDEX

The INDEX utilises the number of news articles containing the terms uncertain or uncertainty, economic or economy, as well as policy relevant terms (scaled by the smoothed total number of articles). Policy relevant terms include: 'policy', 'tax', 'spending', 'regulation', 'Bank of England', 'budget', and 'deficit'



May 1997
TONY BLAIR
TOOK OFFICE AS PRIME MINISTER

June 2007
GORDON BROWN
TOOK OFFICE AS PRIME MINISTER

April 2010
DAVID CAMERON
TOOK OFFICE AS PRIME MINISTER

July 2016
THERESA MAY
TOOK OFFICE AS PRIME MINISTER

July 2019
BAFFOON JOHNSON
TOOK OFFICE AS PRIME MINISTER

SOURCE: https://www.policyuncertainty.com/europe_monthly.html

SOURCE ACCREDITATION: Source: 'Measuring Economic Policy Uncertainty' by Scott R. Baker, Nicholas Bloom and Steven J. Davis at www.PolicyUncertainty.com.

DESIGN: @GREENYNORFOLK

INFERENTIAL

- Take a hypothesis
- Restate as a question
- Answer by testing on a different set of data

Hypothesis generated previously: among adults, eating at least 5 servings a day of fresh fruit and vegetables is associated with fewer viral illnesses per year.

→ *Study subset of French population*

PREDICTIVE

Find out what predicts something to occur

What will predict someone to eat a certain diet

CAUSAL

Find out what causes something to occur

What causes someone to eat a certain diet

MECHANISTIC

Find out *how* something causes something else

How does the diet lead to a reduction in viral ilnesses?

RESEARCH QUESTIONS

- Many data analyses answer multiple questions
- Questions are often influenced by the data you have

GOOD RESEARCH QUESTIONS

- Are of interest to your audience
- Have not already been answered
- Questions should stem from plausible framework
  - They have to possible make sense (can yoghurt sales predict pepperoni sales?)
- Questions should be answerable
- Should be specific enough to be answerable
  - Does x make you healthier? (what does healthier mean?)