

EXPLORATORY DATA ANALYSIS & ELICITATION

PETRA ISENBERG

VISUAL ANALYTICS

ANALYSIS COMPONENTS

Remember: not necessarily in this order or linear

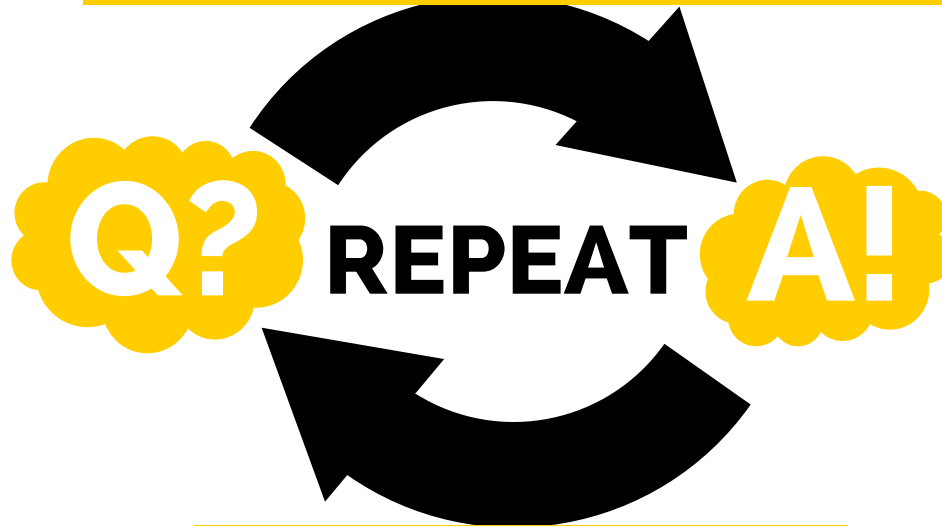


WHY DO YOU NEED DATA?

(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)

John W. Tukey

EXPLORATORY 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

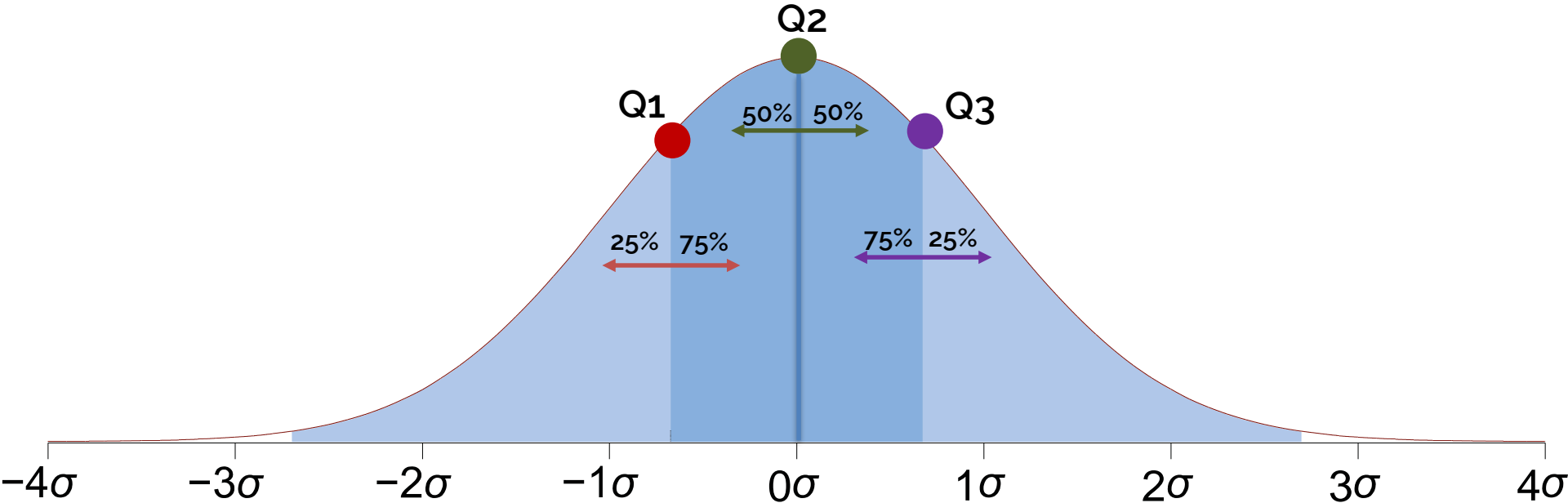
The sample maximum (largest observation)

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 | 9998844331929433361107
5 | 97666666554422210097731
6 | 898665441077761065
7 | 98855431100652108073
8 | 653322122937
9 | 377655421000493
10 | 0984433165212
11 | 4963201631
12 | 45421164
13 | 47830
14 | 00
15 | 676
16 | 52
17 | 92
18 | 5
19 | 39730
```

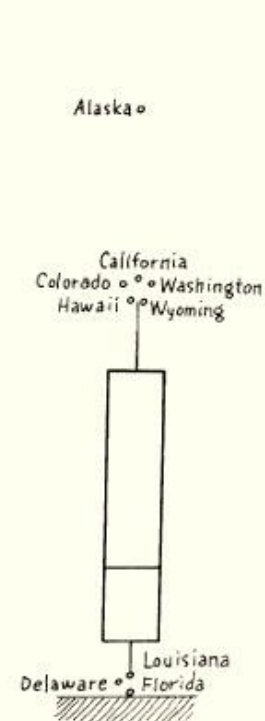


BOX PLOTS

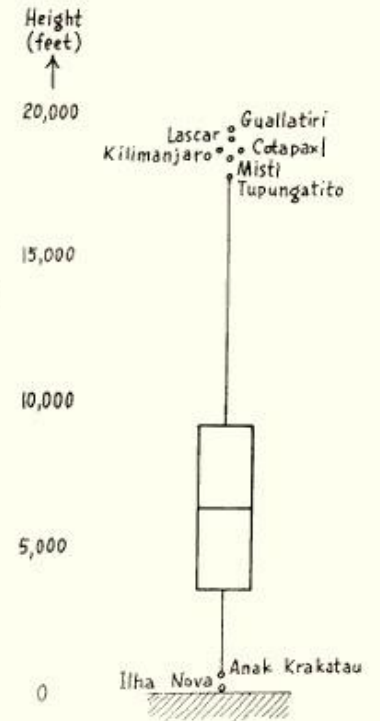
exhibit 6 of chapter 2: various heights

Box-and-whisker plots with end values identified

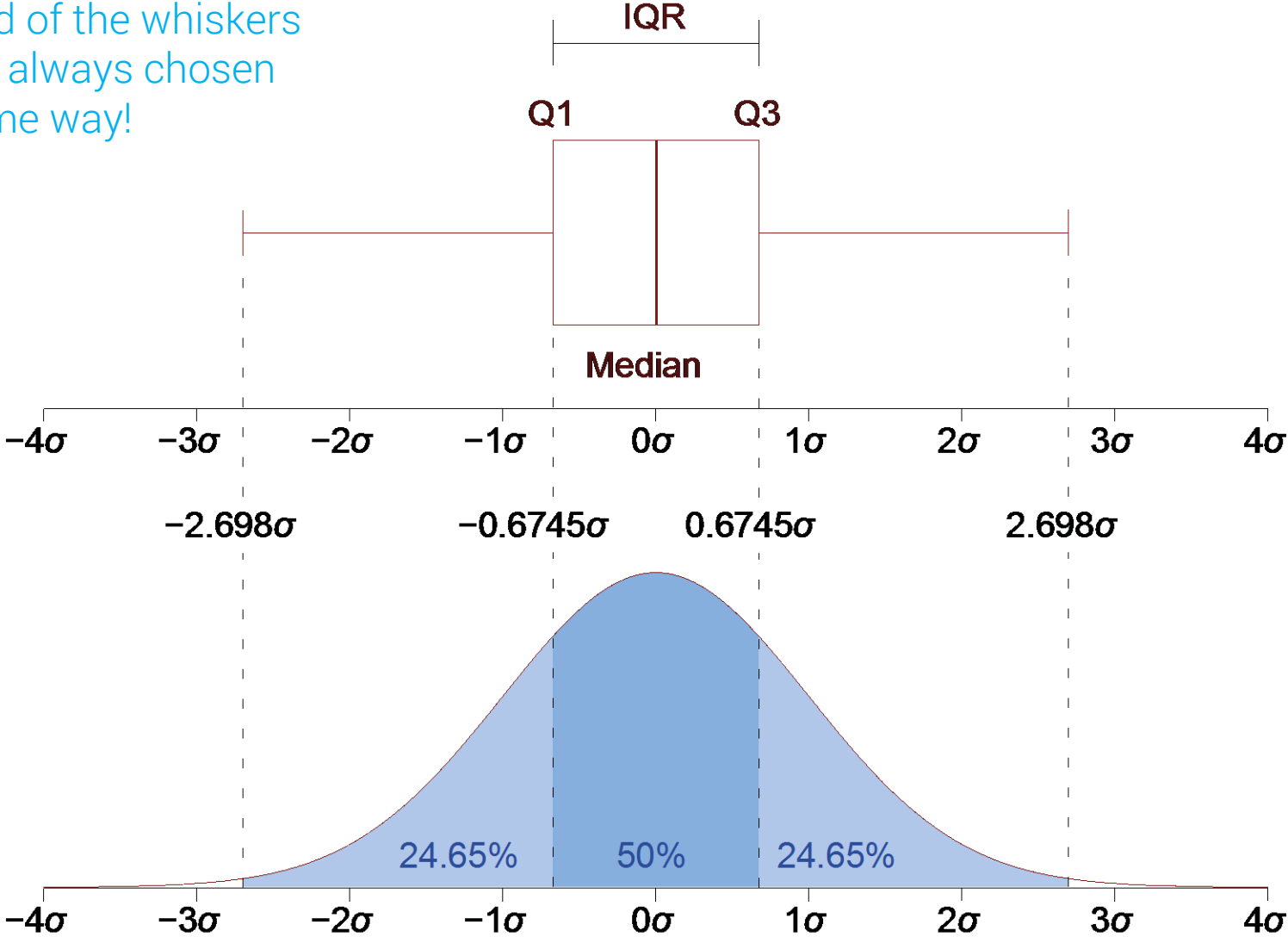
A) HEIGHTS of 50 STATES



B) HEIGHTS of 219 VOLCANOS



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



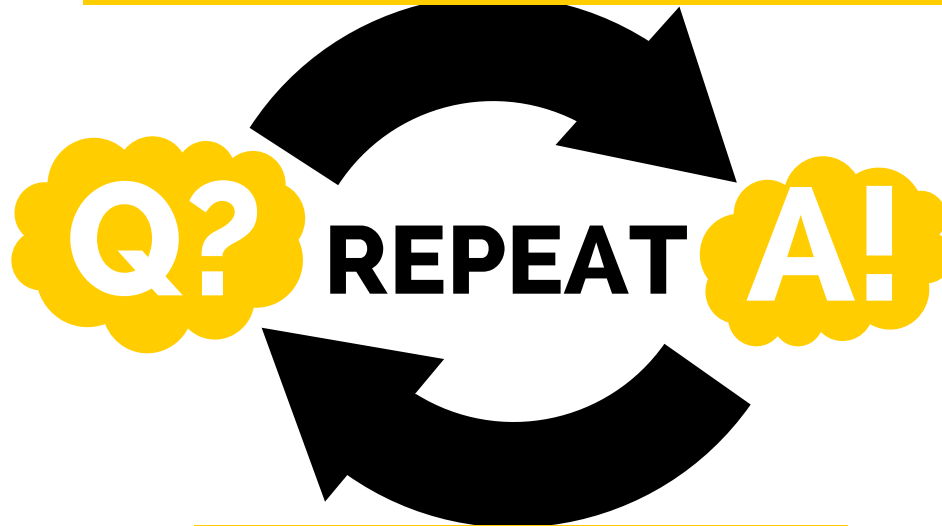
EXPLORATORY ANALYSIS IS ABOUT **UNDERSTANDING DATA** AND **CHECKING ASSUMPTIONS**

- 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

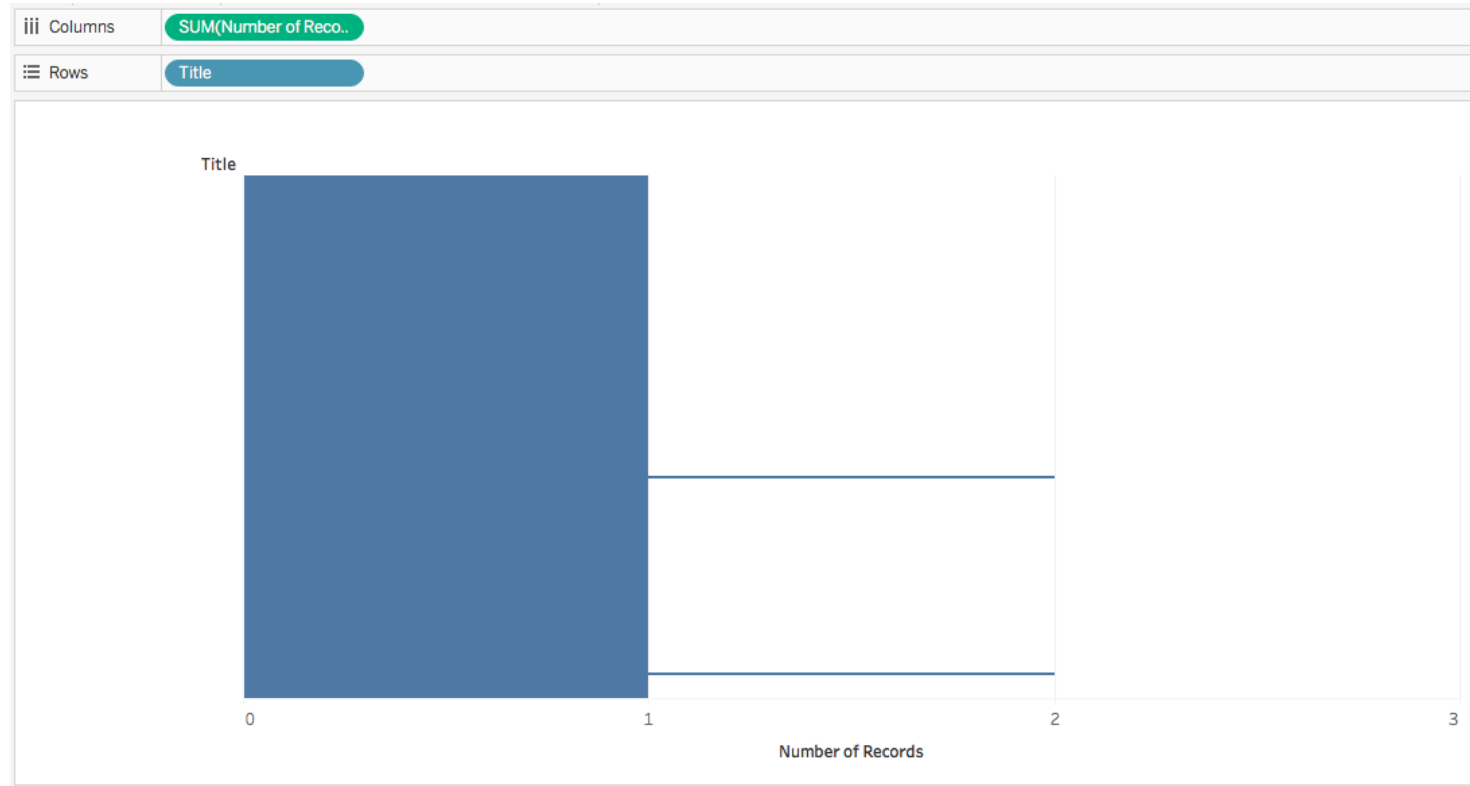
ARE THE FIELDS CORRECT?

# 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
			2	1	5.00000	859,046,895	1995.00
			16	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	156				95.00
9	Sudden Death (1995)	Action	16				95.00
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

More Accurate

↑

↓

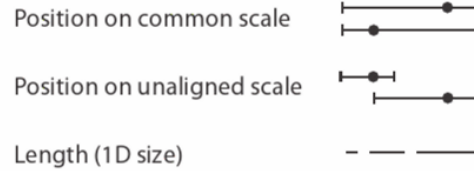
Less Accurate

	Quantitative	Ordinal	Nominal
	Position 	Position 	Position 
	Length 	Density 	Hue 
	Angle 	Saturation 	Density 
	Slope 	Hue 	Saturation 
	Area 	Length 	Shape 
	Density 	Angle 	Length 
	Saturation 	Slope 	Angle 
	Hue 	Area 	Slope 
	Shape 	Shape 	Area 

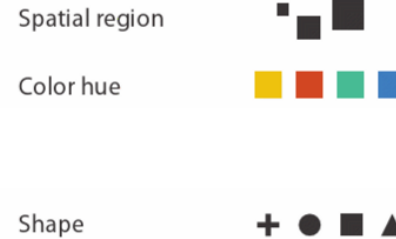
[JACQUES BERTIN REFINED BY CLEVELAND & MCGILL THEN BY CARD & MACKINLAY]

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

➔ Magnitude Channels: Ordered Attributes



➔ Identity Channels: Categorical Attributes



▲ Most
Effectiveness
Least ▼

EASY SOLUTION
ONLY USE THESE!

DEFAULT TO SIMPLE AND EFFECTIVE CHART TYPES

the BAR



the LINE



the SCATTER



**+ COLOUR & SHAPE
TO SHOW CATEGORIES**

SOME FIRST STEPS TO START WITH

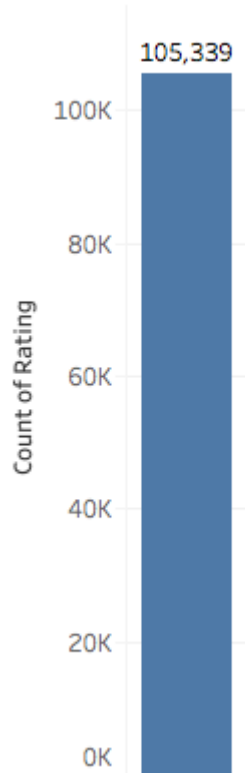
1. Plot the raw data
2. Plot simple statistics
3. Look at plots together

CHECK SIMPLE STATISTICS

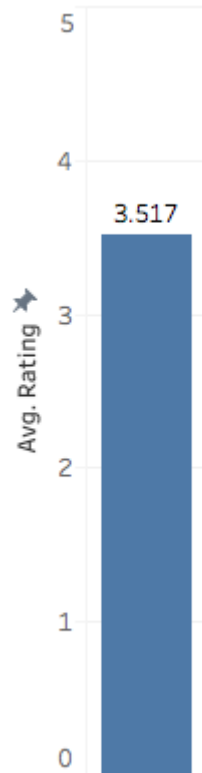
Measures

Rating

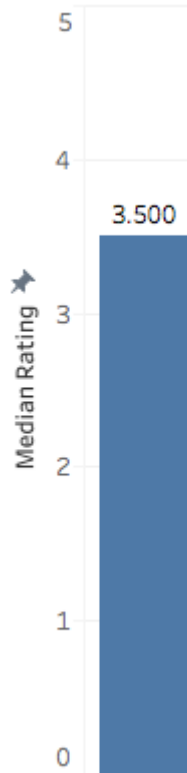
CNT(Rating)



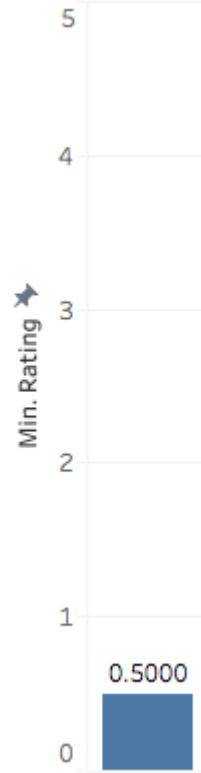
AVG(Rating)



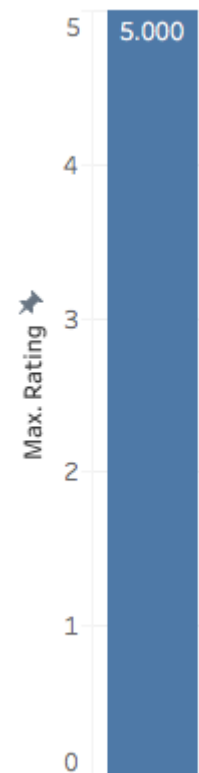
MEDIAN(Rating)



MIN(Rating)



MAX(Rating)

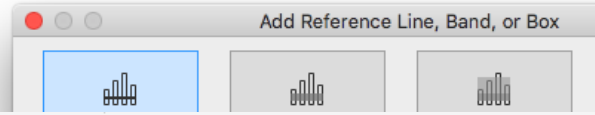


STDEV(Rating)



CHECK SIMPLE STATISTICS

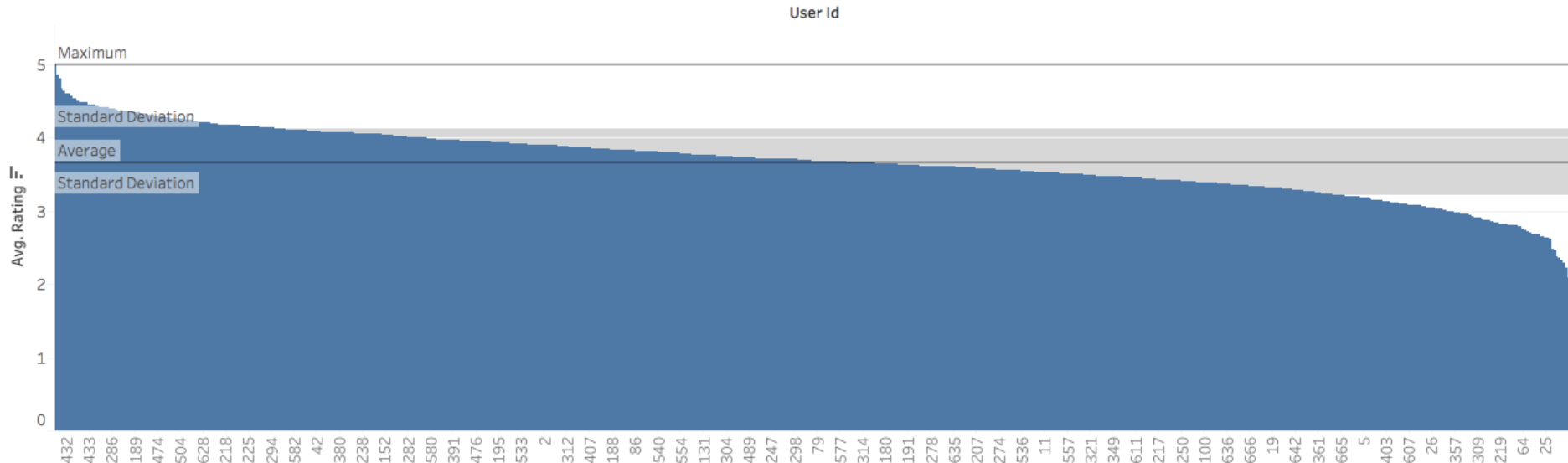
WHAT'S ONE MORE EASY THING WE SHOULD DO?



Columns User Id

Rows AVG(Rating)

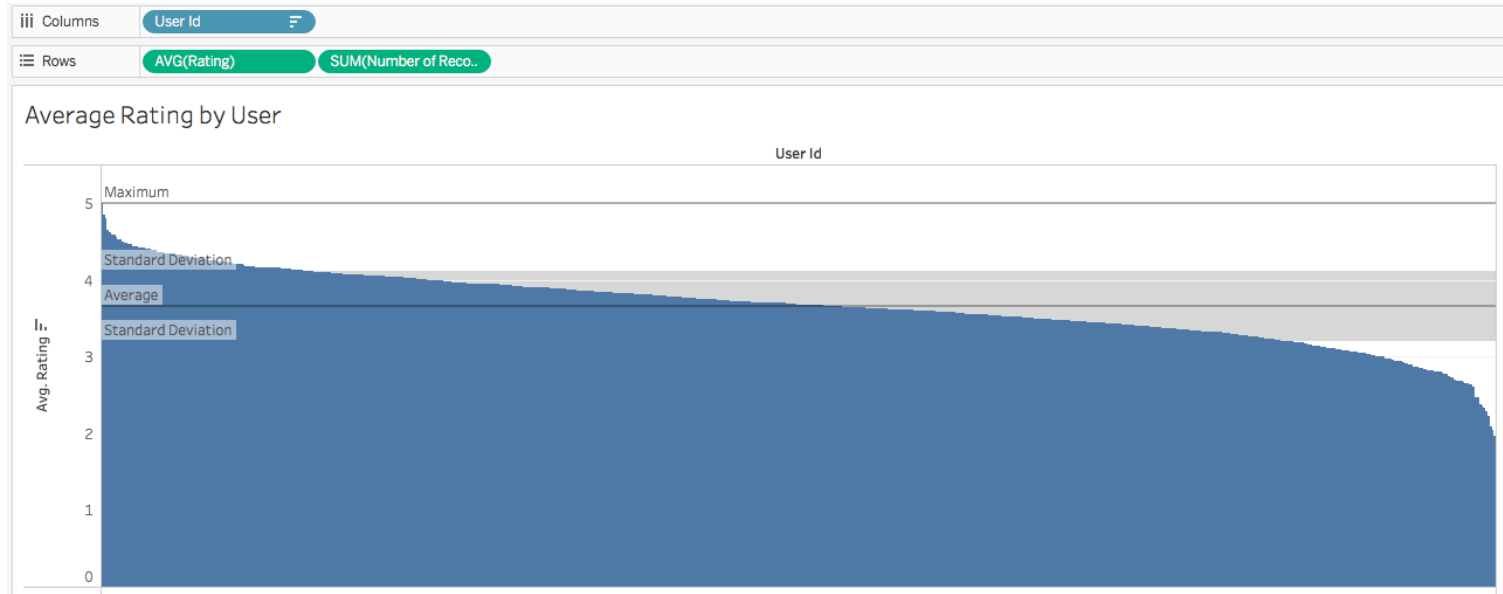
Average Rating by User



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

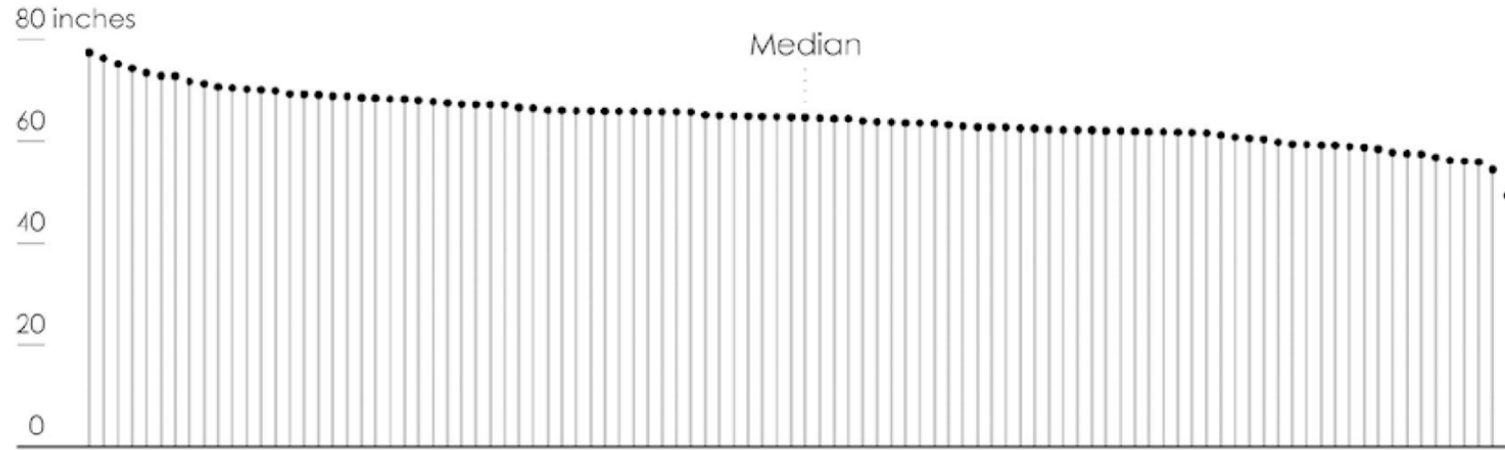
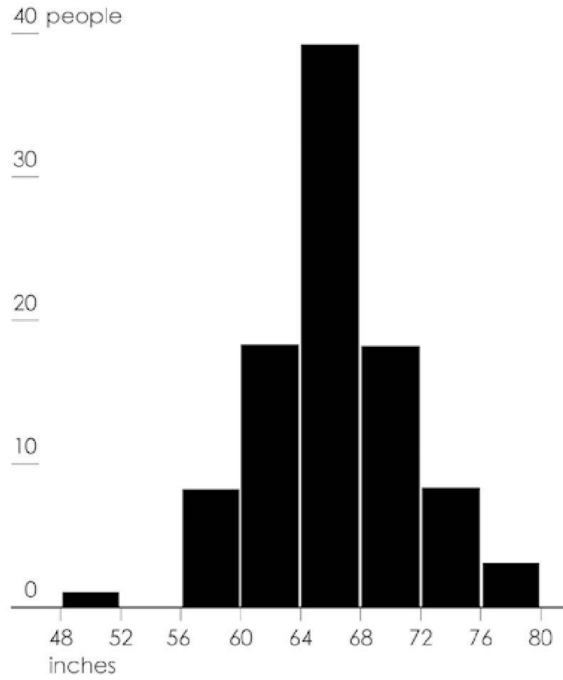
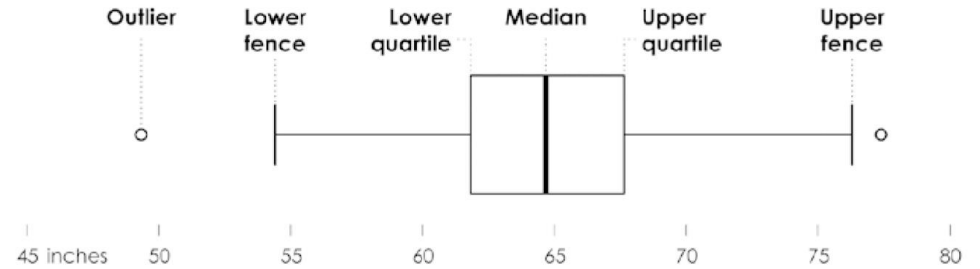


FIGURE 4-52 *Heights of imaginary people, sorted from shortest to tallest*



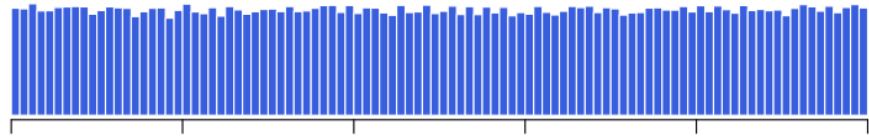
HISTOGRAMS



BOX PLOTS

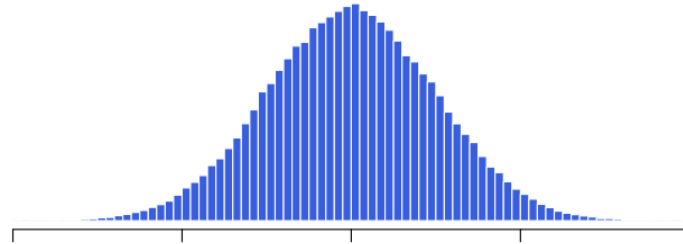
SOME SIMPLE DISTRIBUTIONS

UNIFORM



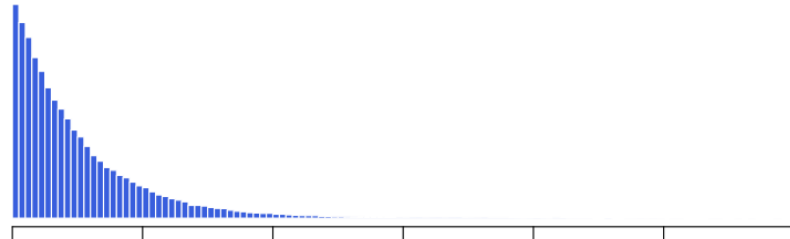
NORMAL

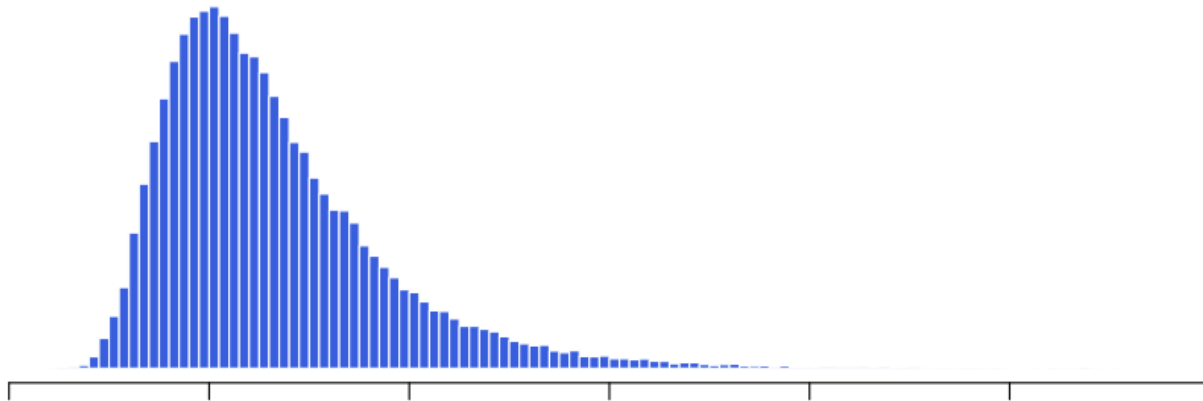
REPEATED MEASURES, VARIATION
IN POPULATIONS, ETC.



EXPONENTIAL

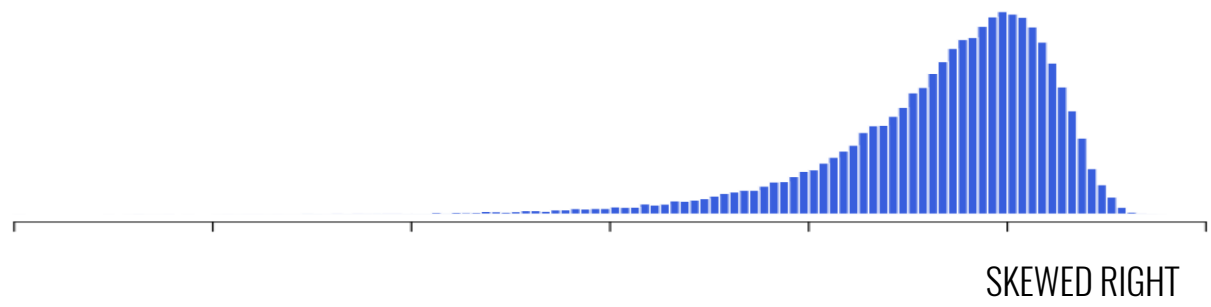
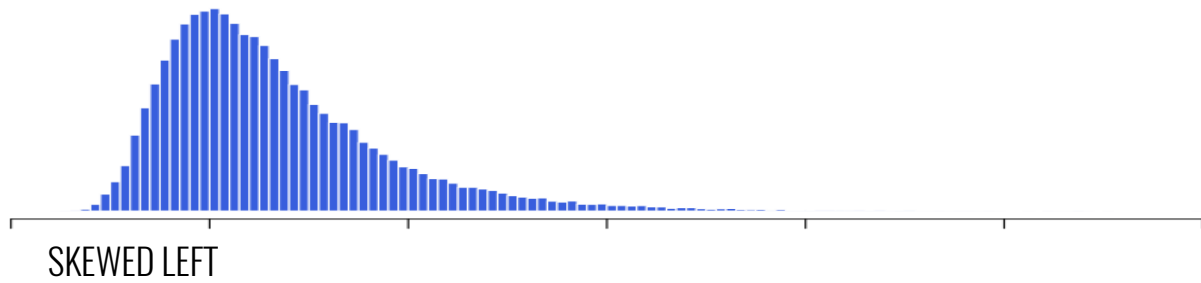
DURATIONS BETWEEN EVENTS, ETC.



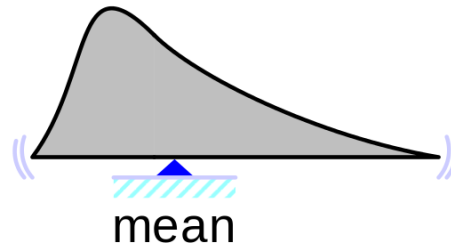
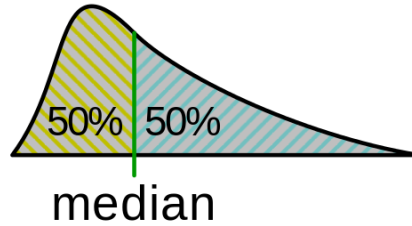
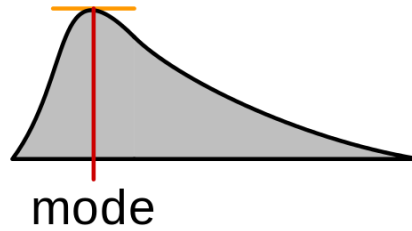


LOG-NORMAL

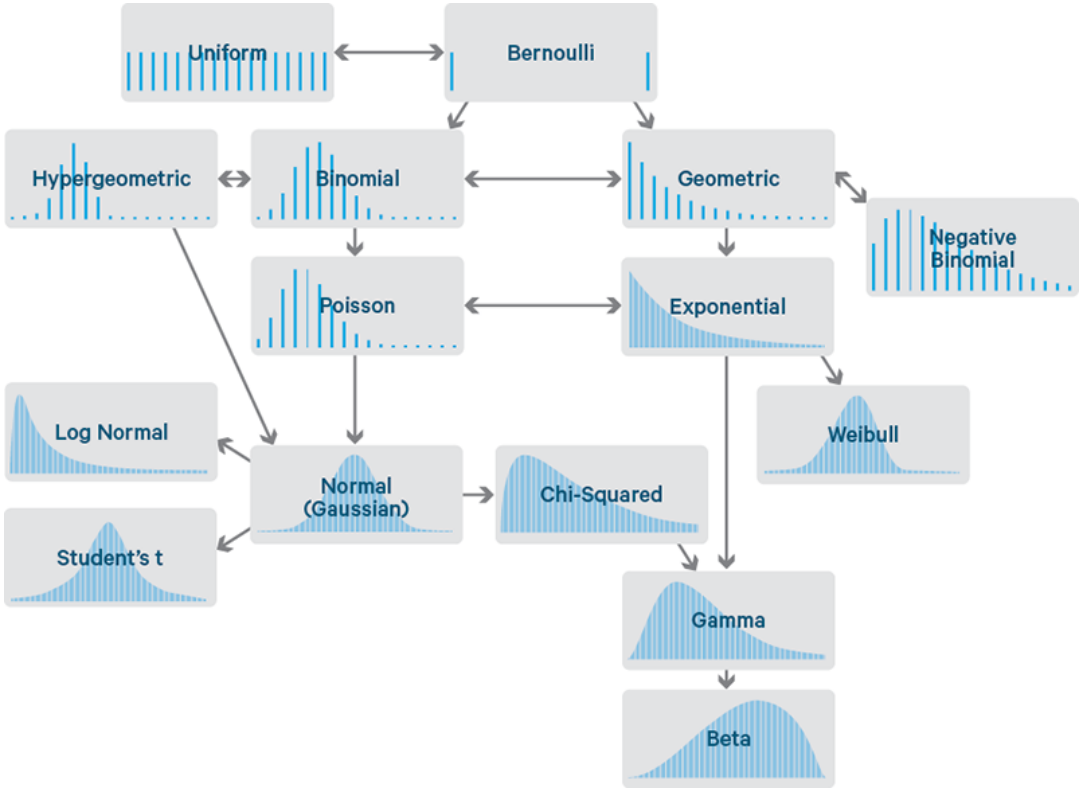
MANY BIOLOGICAL AND SOCIAL PROCESSES
(TISSUE GROWTH, CITY SIZE, # OF FOLLOWERS)



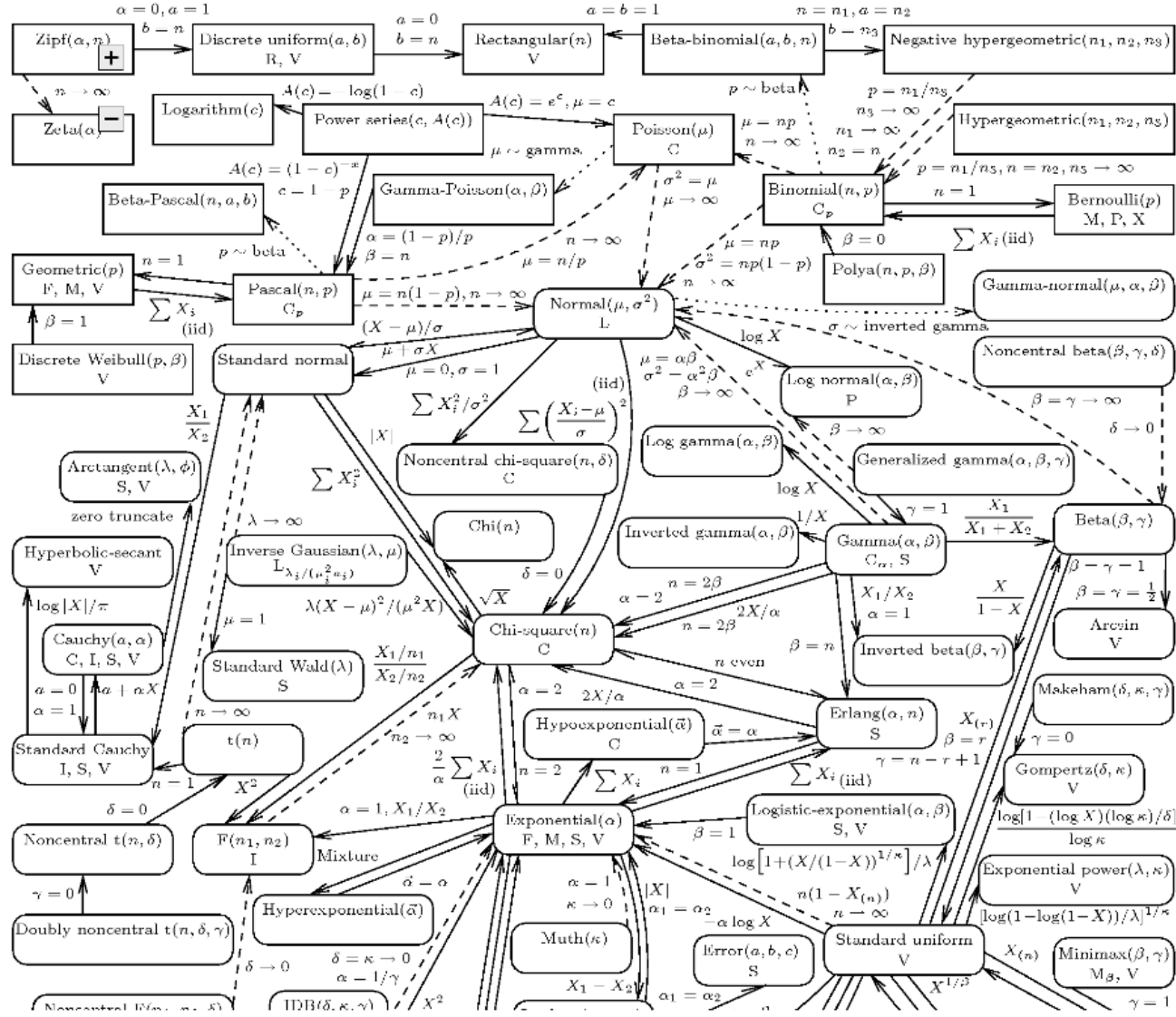
DISTRIBUTIONS AND SOME COMMON MEASURES OF CENTRAL TENDENCY



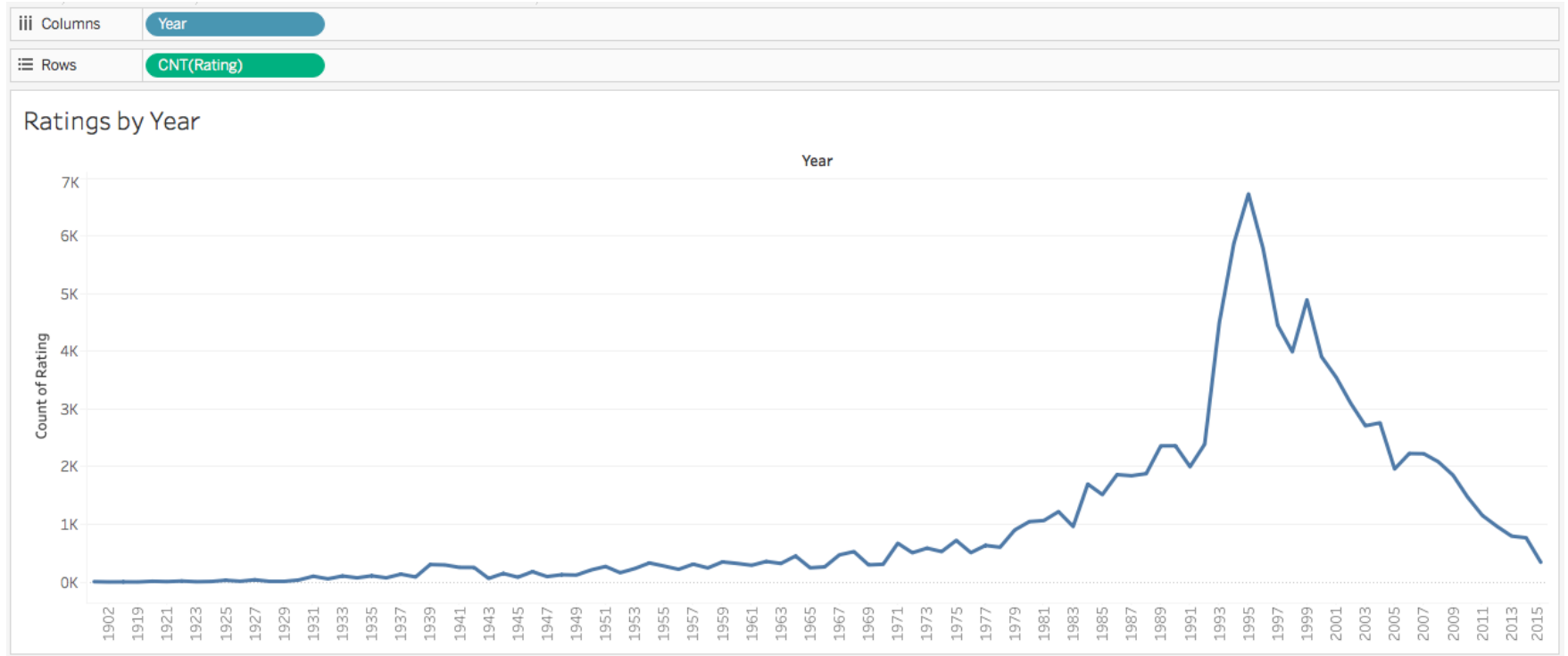
COMMON DISTRIBUTIONS



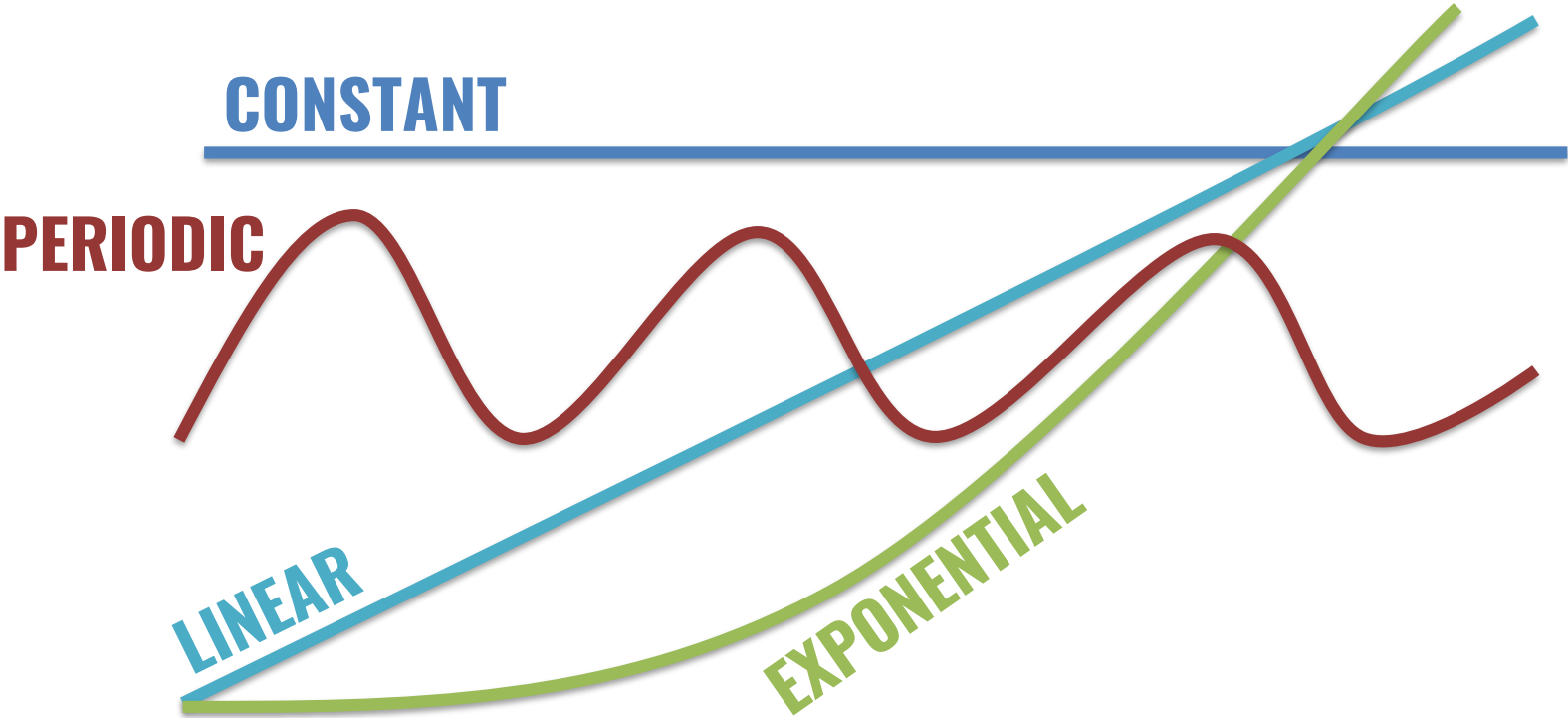
CLUDERA



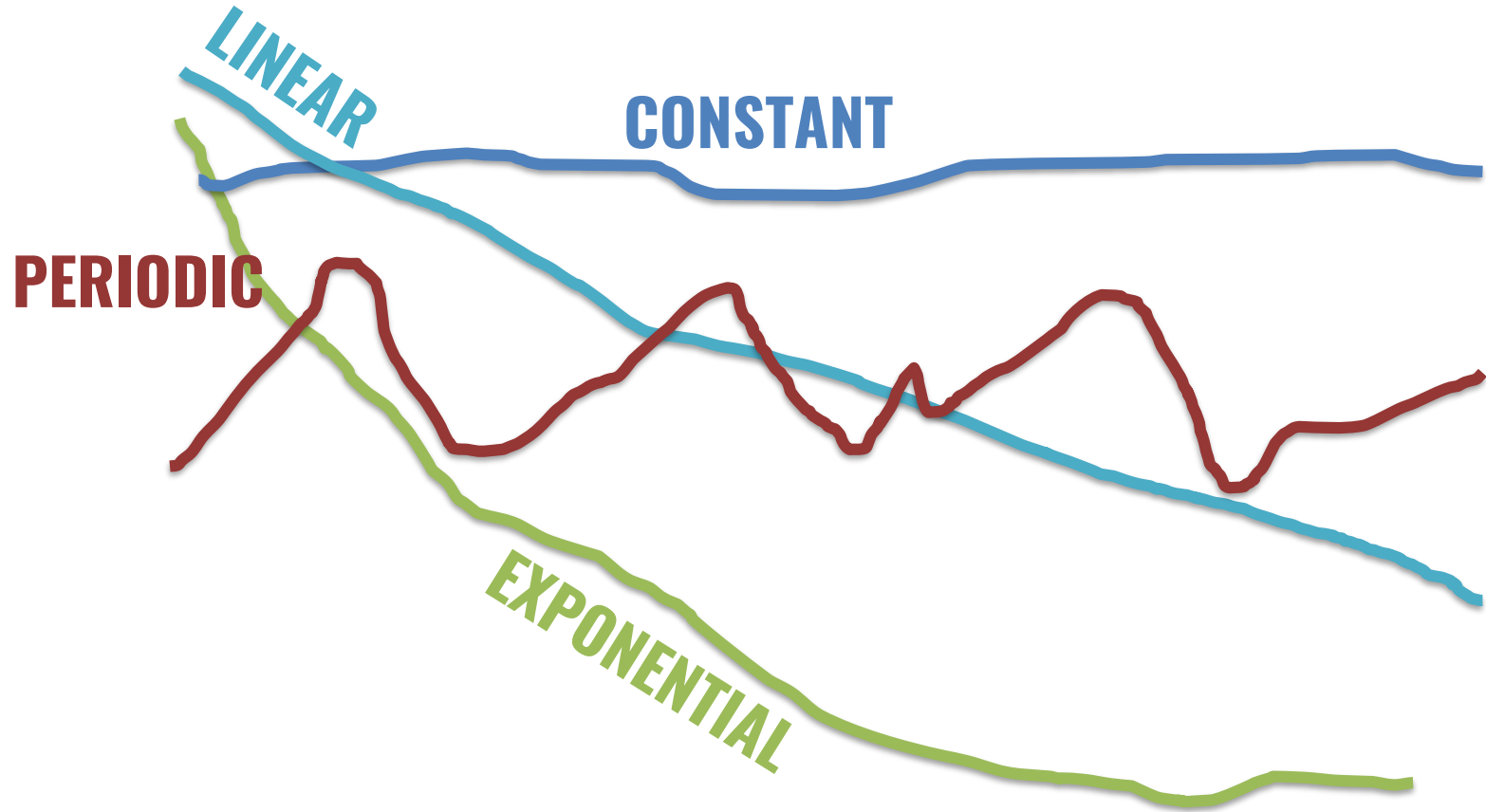
DISTRIBUTIONS OVER TIME



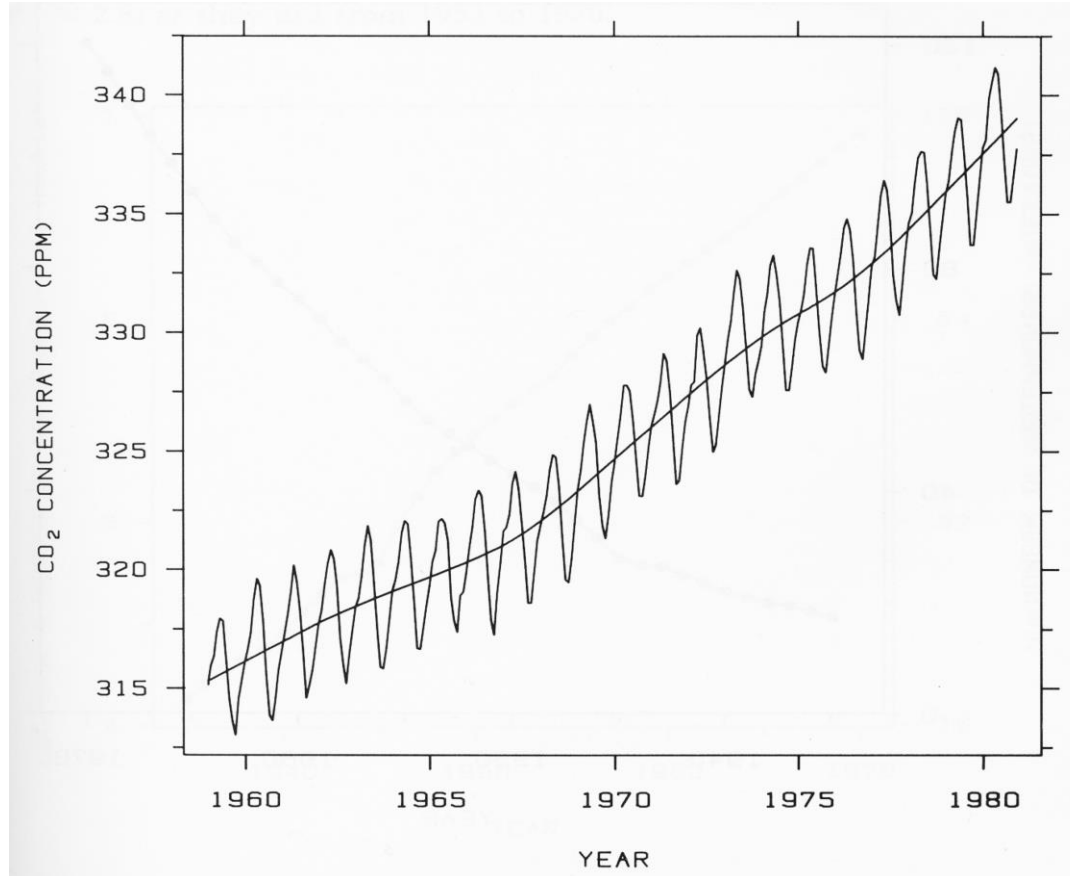
IDENTIFYING TRENDS



IDENTIFYING TRENDS

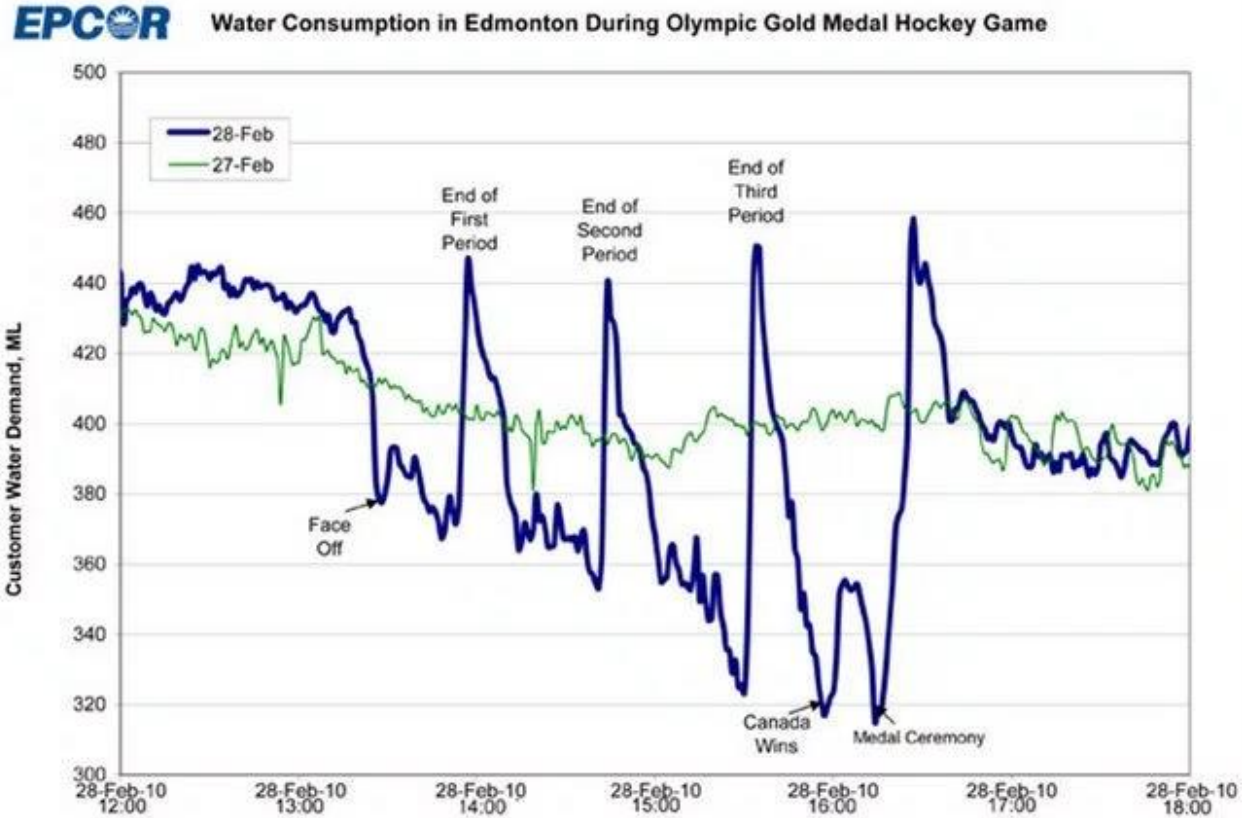


COMBINATIONS

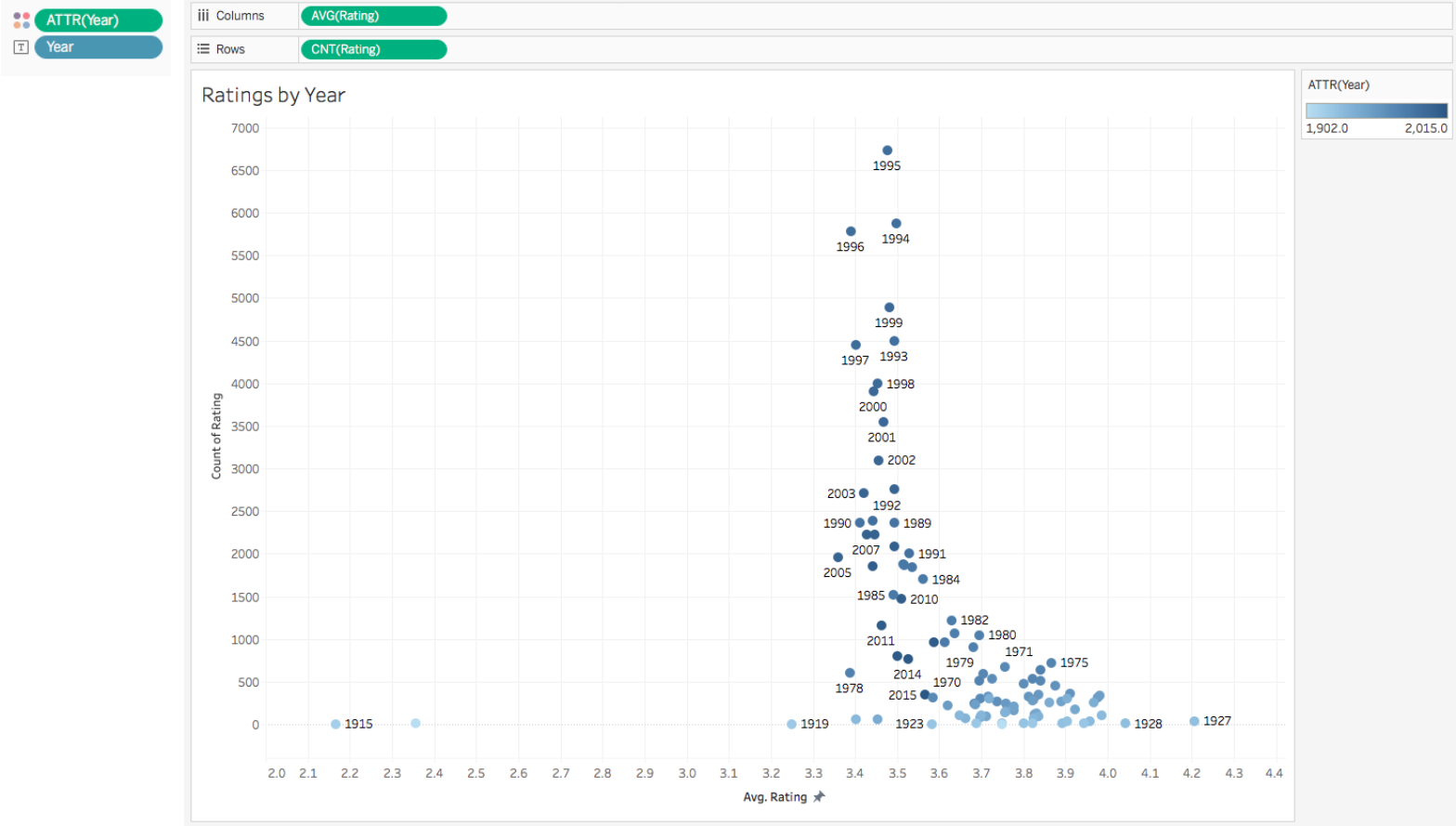


YEARLY CO₂ CONCENTRATIONS
CLEVELAND 85

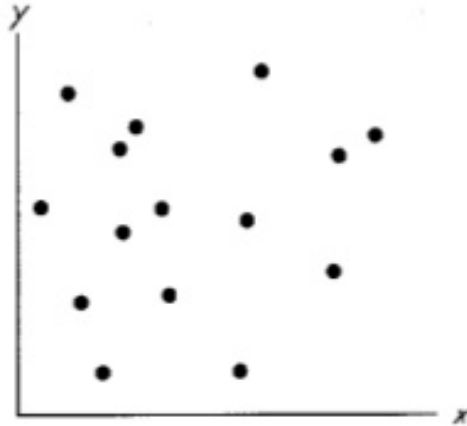
COMPARISON AGAINST A KNOWN BASELINE



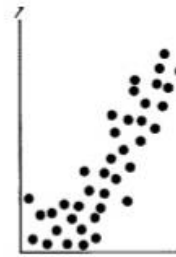
COMPARING TWO MEASURES



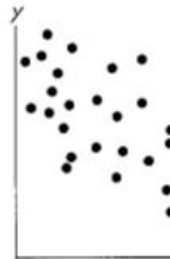
IDENTIFYING CORRELATIONS



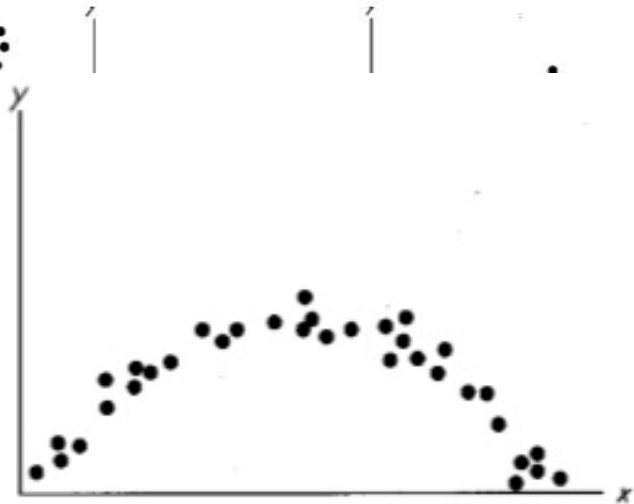
(g) No correlation between x and y



(a) Positive correlation between x and y



(d) Negative correlation between x and y



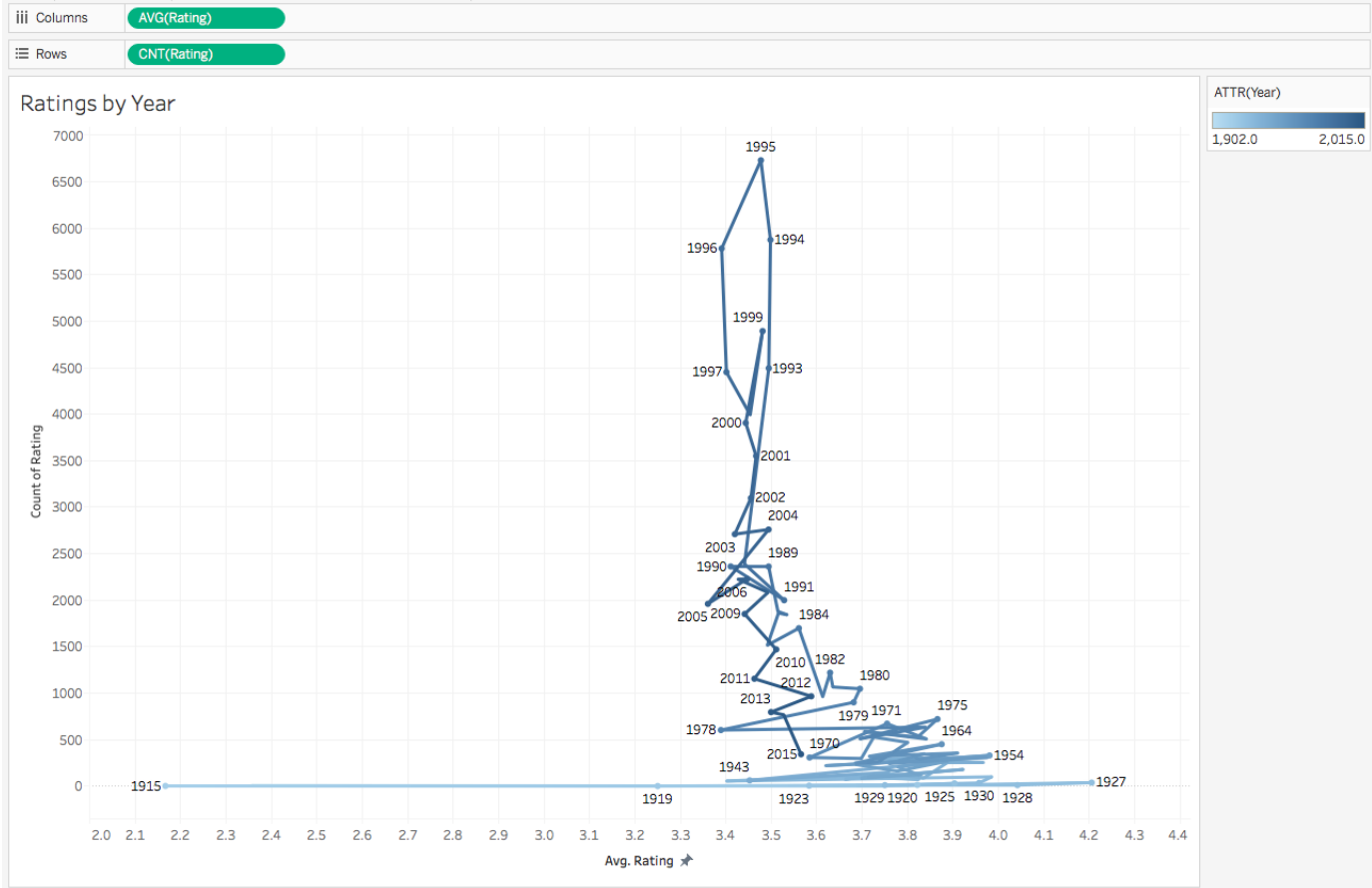
(h) Nonlinear correlation between x and y

correlation between x and y

correlation between x and y

COMPARING TWO MEASURES

ATTR(Year)
Year



ANY HYPOTHESES?

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



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?**

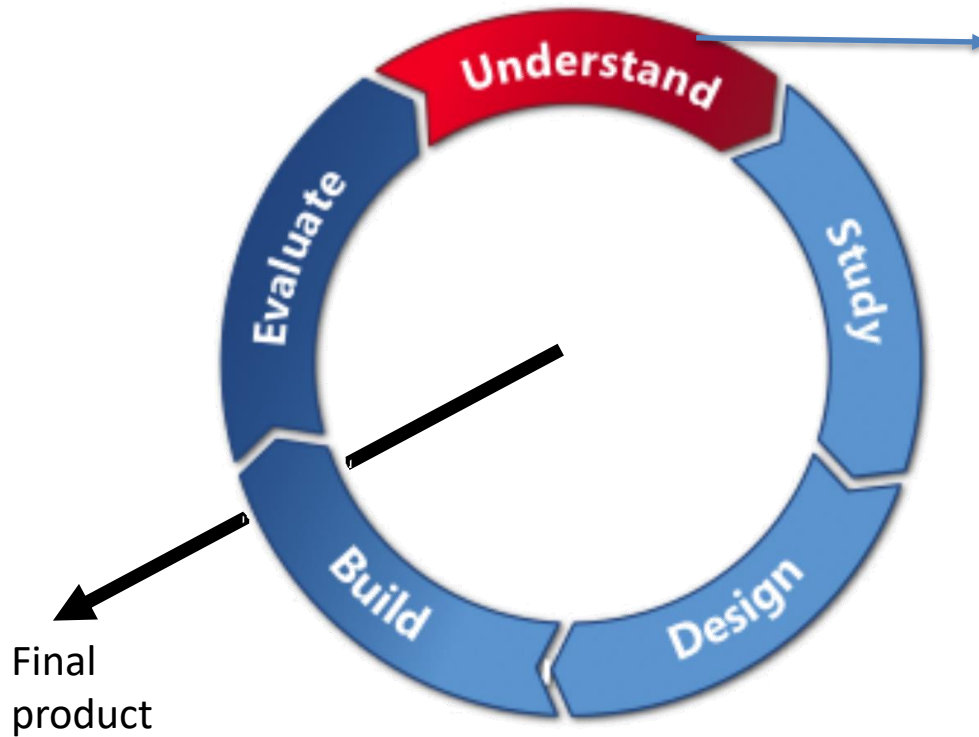
THE USER-CENTERED APPROACH

- 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?

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 your 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?

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

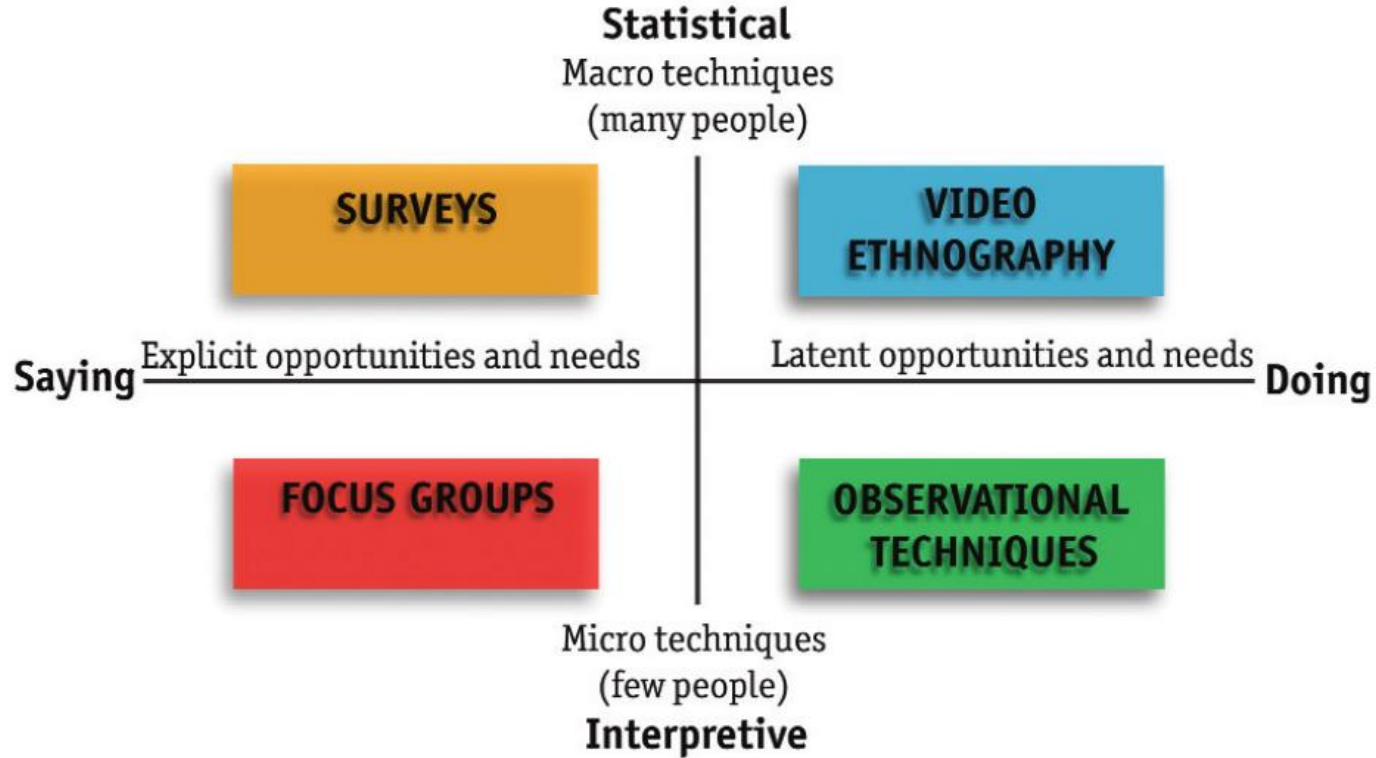
IF ALL ELSE FAILS

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 why people are acting the way they do, or what their goals, needs, or feelings are



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?



**IDENTIFY DATA & VARIABLES FOR
YOUR ANALYSIS**

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

