I 202: Information Organization & Retrieval Fall 2025

Class 17: Grounded Coding

From Previous Class: Ethical Concerns of Data Collection & Image Classification

- The main issues:
  - *Intellectual property concerns (week 15)*
  - *Bias in representativeness of images (week 15)*
  - *Labor practices on crowdwork platforms (INFO 203)*
  - *Privacy concerns for applications of image recognition (I 203, I 205)*

Today’s Outline

Upcoming Assignments & Lectures

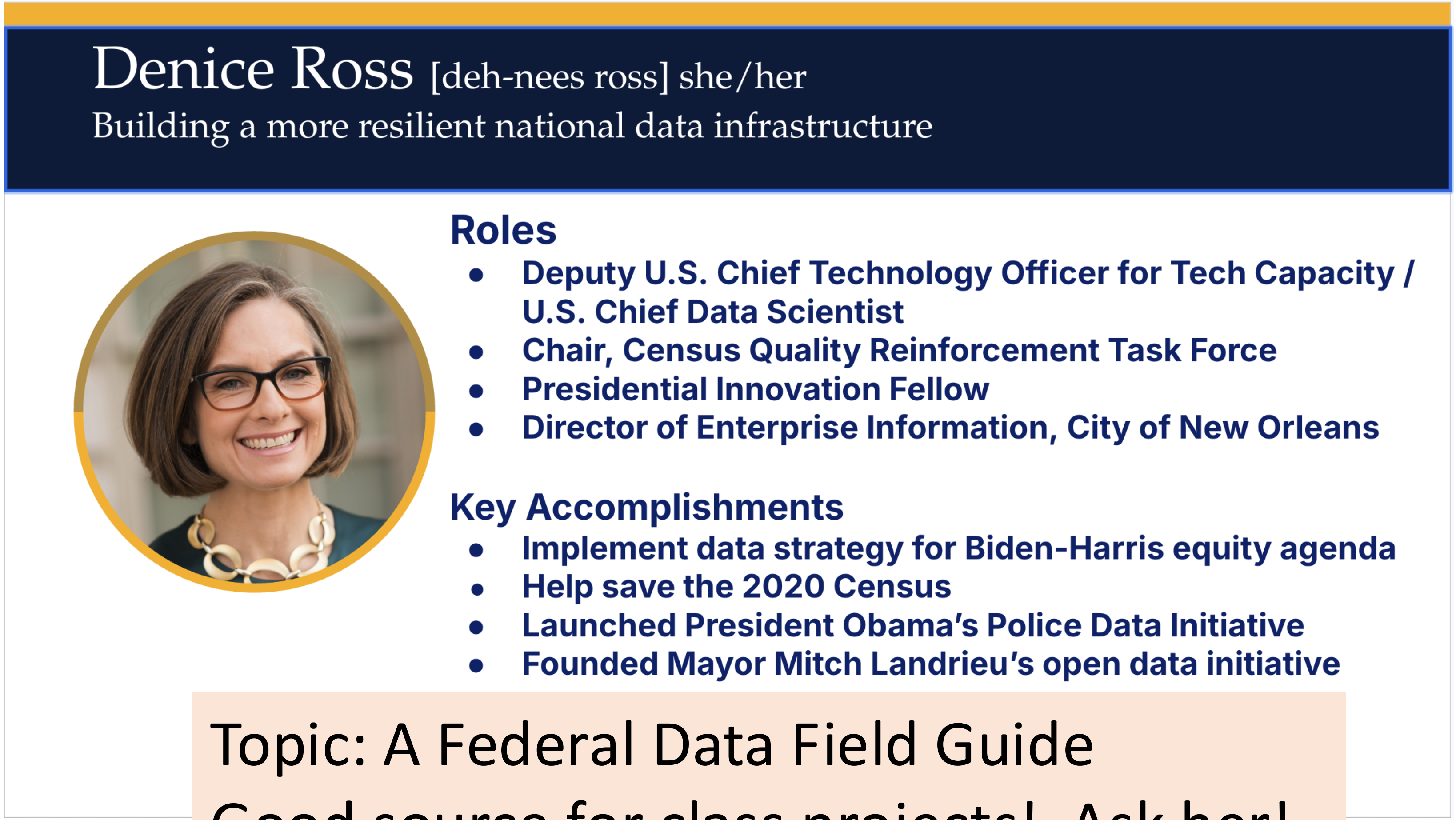
Revisit Card Sorting

Grounded Coding

Inter-annotator Agreement

Upcoming Lecture Topics

- Week 10: Information seeking, information foraging
- Week 11: Guest lectures:
  - *Federal Datasets*
  - *Rulemaking Standards*
- Week 12: How search engines work; search evaluation
- Week 13: Generative AI; LLMs and search
- Week 14: Misinformation (Thanksgiving Week)
- Week 15: Fairness & Bias in Search; Intellectual Property



Denice Ross [deh-nees ross] she/her

Building a more resilient national data infrastructure

Roles

- Deputy U.S. Chief Technology Officer for Tech Capacity / U.S. Chief Data Scientist
- Chair, Census Quality Reinforcement Task Force
- Presidential Innovation Fellow
- Director of Enterprise Information, City of New Orleans

Key Accomplishments

- Implement data strategy for Biden-Harris equity agenda
- Help save the 2020 Census
- Launched President Obama’s Police Data Initiative
- Founded Mayor Mitch Landrieu’s open data initiative

Topic: A Federal Data Field Guide

Good source for class projects! Ask her!

Judy Brewer (pronunciation: ju’dee bru’wr; pronouns: she/her)

Former OSTP Assistant Director for Accessibility

Previous roles include:

- WH OSTP Assistant Director for Accessibility
- PPS Digital Accessibility Expert
- MIT Principal Research Scientist, CSAIL
- W3C Director of Web Accessibility Initiative

Key Accomplishments include:

- Strengthened implementation of digital accessibility across federal, state, and local governments, including in science, education, and health care
- Promoted AI accessibility and equity
- Led development of globally recognized international standards for web and mobile accessibility

lied Linguistics

y

Topic: Data Standards Creation; Digital Accessibility

Upcoming Assignments

- This week: Grounded coding individually
- Week 10: Grounded coding with a partner
- Week 11-12: Final project proposal
- Week 13: Search assignment; may include questions about the gues lectures
- Week 14-15: Work on final project

Final Projects (Individual)

- **Goal**: synthesize concepts from throughout the semester and apply them to a real-world topic.
- **Sizable:** (we expect it to take approximately 4 weeks to complete, including proposal writing), but we are not expecting you to perform original research or propose new methods.
- **Choice** of:
  - **Paper** (3200-6400 words, cite at least 3 sources)
  - **Implementation project** (hosted online; also requires a writeup)
  - **Design project** (**i**ncludes creating information architectures and other category systems, and should substantially engage with concepts from the class, and requires an evaluation)

Today’s Outline

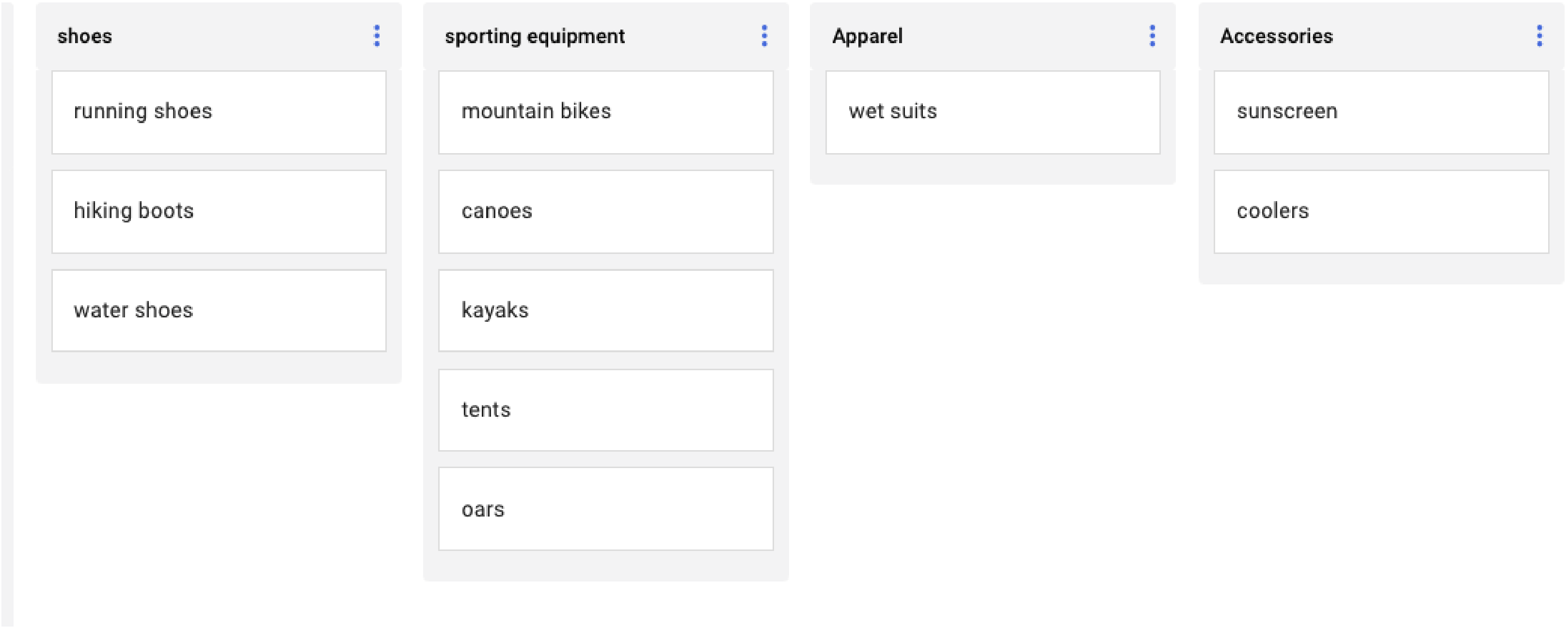
Upcoming Assignments & Lectures

Revisit Card Sorting

Grounded Coding

Inter-annotator Agreement

Previously: Card Sorting



shoes

- running shoes
- hiking boots
- water shoes

sporting equipment

- mountain bikes
- canoes
- kayaks
- tents
- oars

Apparel

- wet suits

Accessories

- sunscreen
- coolers

Why Do We Do Card Sorting?

- Uncover people’s underlying assumptions/intuitions about how concepts organize
- **Open sort**: given a set of terms, group terms & name the groups
  - *See how much agreement there is for categories for a set of items*
- **Closed sort**: given a set of terms and some pre-defined categories, place the terms into the categories
  - *See how well people’s intuitions agree with a set of categories*

Grounded Coding

A technique for creating or analyzing surveys, interviews, and other qualitative info

Grounded Coding

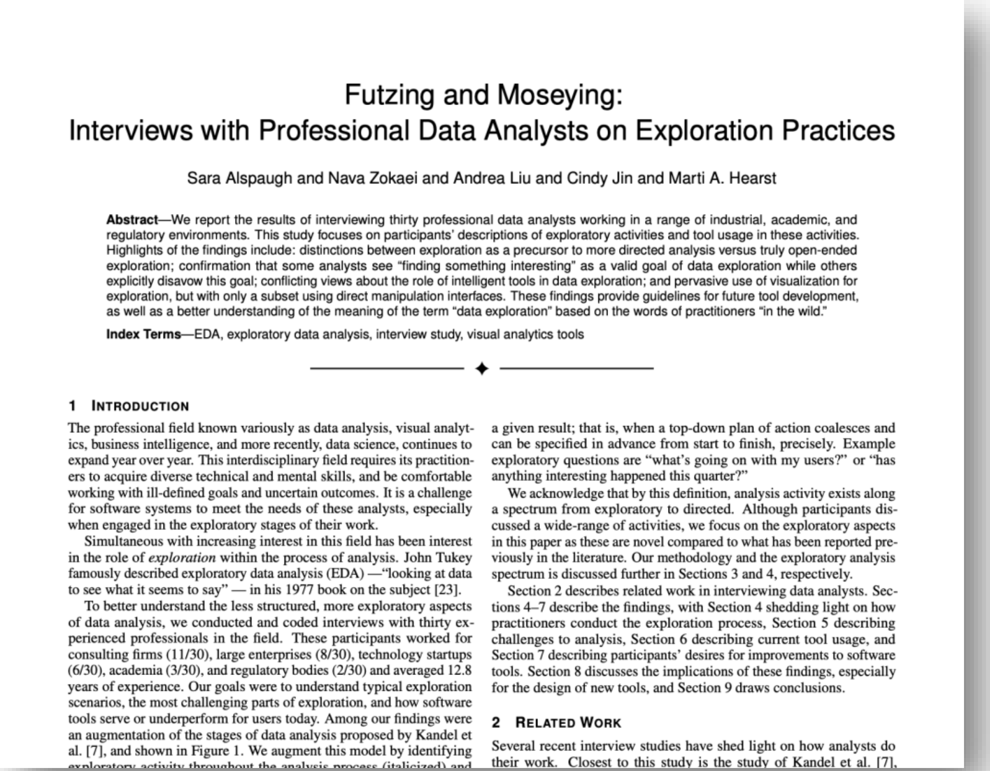
What: A method for assigning categories to qualitative data

Why: Make sense of qualitative data (surveys, tweets, interviews)

How:

- Iterative
- Build consensus from a team of categorizers ("coders")
- Categories based on data (grounded) instead of theory
- ***After*** rounds of data-driven category formation, use theories to motivate which categories to make prominent

Example: Survey of Data Analysts



Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices

Sara Alspaugh and Nava Zokaei and Andrea Liu and Cindy Jin and Marti A. Hearst

Abstract—We report the results of interviewing thirty professional data analysts working in a range of industrial, academic, and regulatory environments. This study focuses on participants' descriptions of exploratory activities and tool usage in these activities. Highlights of the findings include: distinctions between exploration as a precursor to more directed analysis versus truly open-ended exploration; confirmation that some analysts see "finding something interesting" as a valid goal of data exploration while others explicitly disavow this goal; conflicting views about the role of intelligent tools in data exploration; and pervasive use of visualization for exploration, but with only a subset using direct manipulation interfaces. These findings provide guidelines for future tool development, as well as a better understanding of the meaning of the term "data exploration" based on the words of practitioners "in the wild."

Index Terms—EDA, exploratory data analysis, interview study, visual analytics tools

1 Introduction

The professional field known variously as data analysis, visual analytics, business intelligence, and more recently, data science, continues to expand year over year. This interdisciplinary field requires its practitioners to acquire diverse technical and mental skills, and be comfortable working with ill-defined goals and uncertain outcomes. It is a challenge for software systems to meet the needs of these analysts, especially when engaged in the exploratory stages of their work.

Simultaneous with increasing interest in this field has been interest in the role of *exploration* within the process of analysis. John Tukey famously described exploratory data analysis (EDA) —"looking at data to see what it seems to say" — in his 1977 book on the subject [23].

To better understand the less structured, more exploratory aspects of data analysis, we conducted and coded interviews with thirty experienced professionals in the field. These participants worked for consulting firms (11/30), large enterprises (8/30), technology startups (6/30), academia (3/30), and regulatory bodies (2/30) and averaged 12.8 years of experience. Our goals were to understand typical exploration scenarios, the most challenging parts of exploration, and how software tools serve or underperform for users today. Among our findings were an augmentation of the stages of data analysis proposed by Kandel et al. [7], and shown in Figure 1. We augment this model by identifying exploratory activity throughout the analysis process (italicized) and

a given result; that is, when a top-down plan of action coalesces and can be specified in advance from start to finish, precisely. Example exploratory questions are "what's going on with my users?" or "has anything interesting happened this quarter?"

We acknowledge that by this definition, analysis activity exists along a spectrum from exploratory to directed. Although participants discussed a wide-range of activities, we focus on the exploratory aspects in this paper as these are novel compared to what has been reported previously in the literature. Our methodology and the exploratory analysis spectrum is discussed further in Sections 3 and 4, respectively.

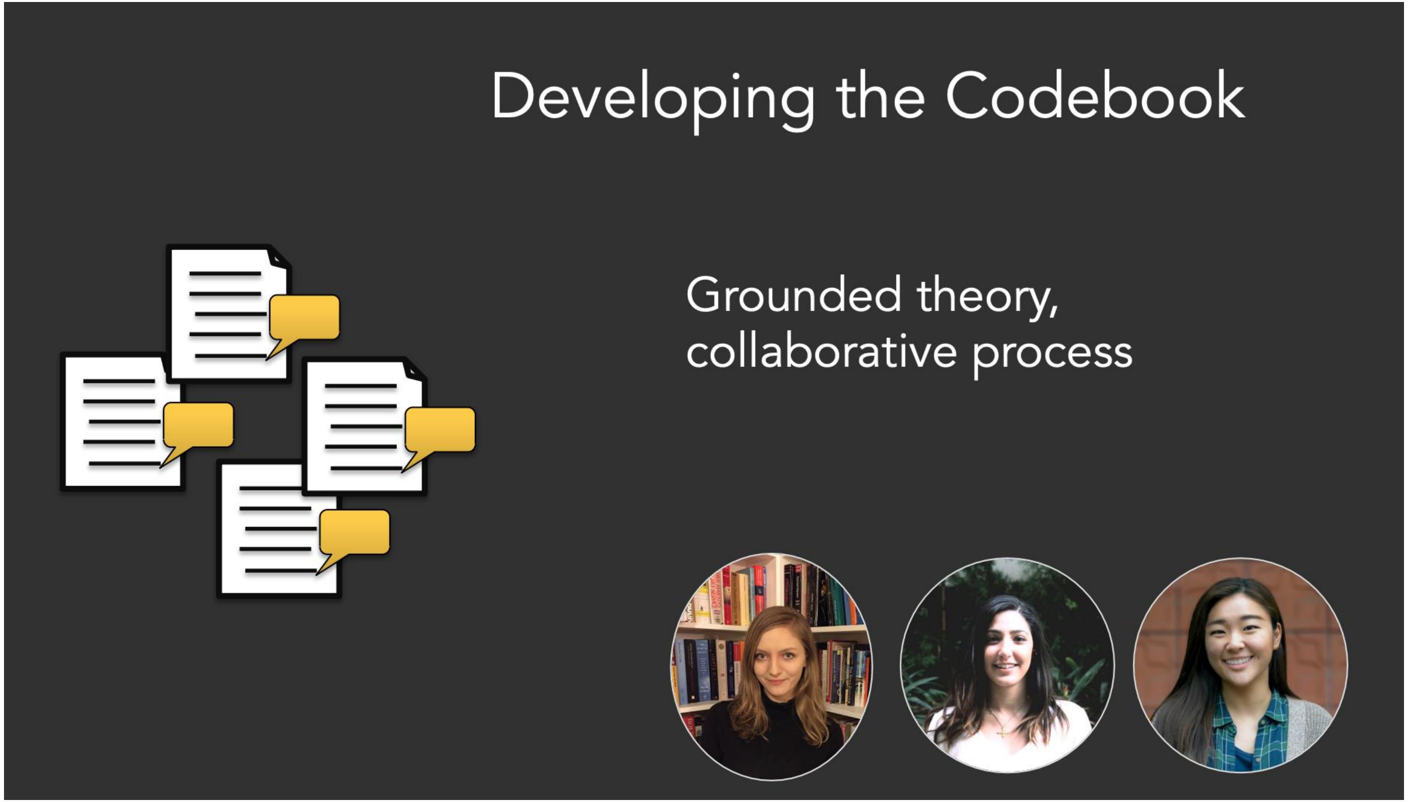
Section 2 describes related work in interviewing data analysts. Sections 4–7 describe the findings, with Section 4 shedding light on how practitioners conduct the exploration process, Section 5 describing challenges to analysis, Section 6 describing current tool usage, and Section 7 describing participants' desires for improvements to software tools. Section 8 discusses the implications of these findings, especially for the design of new tools, and Section 9 draws conclusions.

2 Related Work

Several recent interview studies have shed light on how analysts do their work. Closest to this study is the study of Kandel et al. [7],

Alspaugh et al., Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices, IEEE TVCG 2019

Example: Survey of Data Analysts



Developing the Codebook

Grounded theory,
collaborative process

Alspaugh et al., Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices, IEEE TVCG 2019

The Codebook

Codes hierarchically organized

75 codes

Top levels:

- Background
- Workflow stage (exploration goals)
- Tools used
- Desired tools and features
- Homegrown automation

Alspaugh et al., Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices, IEEE TVCG 2019

Labeling Utterances

Pass 1: using codebook, 2 coders independently labeled each utterance.

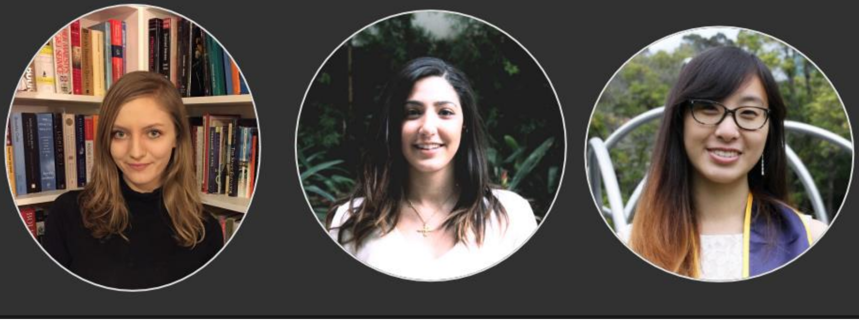
Every utterance gets ≥ 2 coders

Pass 2: each reviewed the other's codes, considered changing if conflict.

Pass 3: any remaining differences were tie-broken by the third coder.

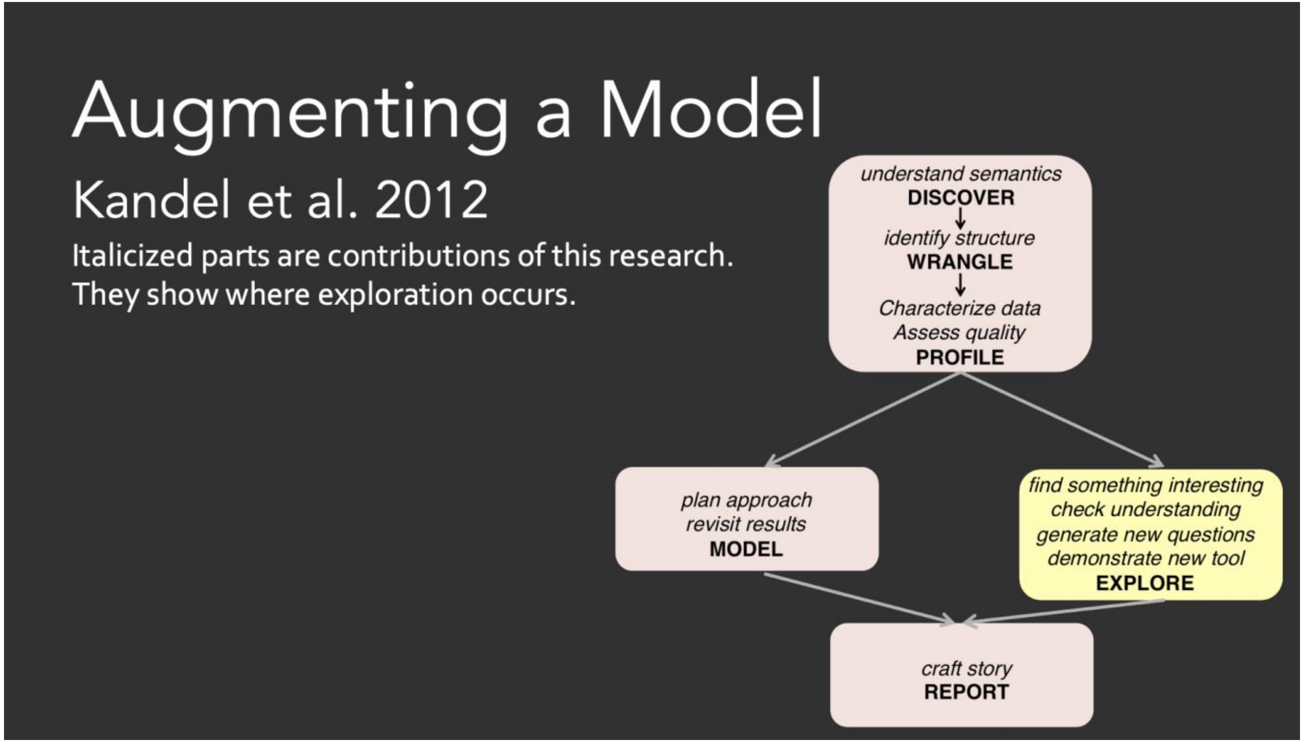
8683 total

Cohen's Kappa=0.91



Alspaugh et al., Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices, IEEE TVCG 2019

Example: Survey of Data Analysts



Augmenting a Model

Kandel et al. 2012

Italicized parts are contributions of this research.
They show where exploration occurs.

understand semantics
DISCOVER
↓
identify structure
WRANGLE
↓
Characterize data
Assess quality
PROFILE

plan approach
revisit results
MODEL

find something interesting
check understanding
generate new questions
demonstrate new tool
EXPLORE

craft story
REPORT

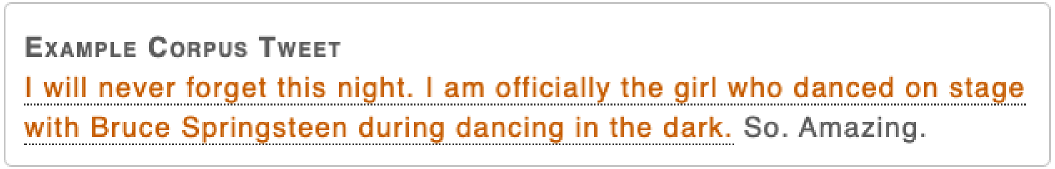
Alspaugh et al., Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices, IEEE TVCG 2019

Example: Grounded Coding of Tweets

Goal: better understand how students are writing outside the classroom

Approach: use tweets to analyze the writing practices of fans of Bruce Springsteen

Data: tweets before, during, and after a concert in 2012



EXAMPLE CORPUS TWEET

I will never forget this night. I am officially the girl who danced on stage with Bruce Springsteen during dancing in the dark. So. Amazing.

Baby, We Were Born to Tweet: Springsteen Fans, The Writing Practices of *In Situ* Tweeting, and the Research Possibilities for Twitter, **Wolff, Kairos journal, 19.3, 2015**
https://web.archive.org/web/20180212012518/http://kairos.technorhetoric.net/19.3/topoi/wolff/index.html

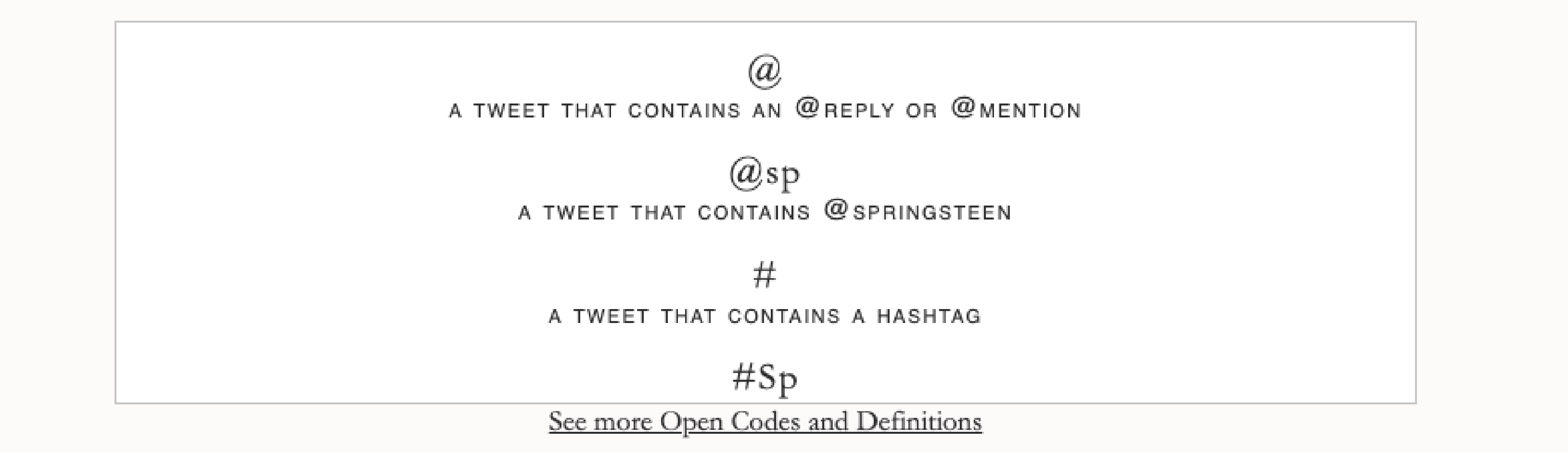
wikipedia

First Stage: Open Coding

Notice small details

You can use any code naming scheme you like

Make brief definitions for each code



@

A TWEET THAT CONTAINS AN @REPLY OR @MENTION

@sp

A TWEET THAT CONTAINS @SPRINGSTEEN

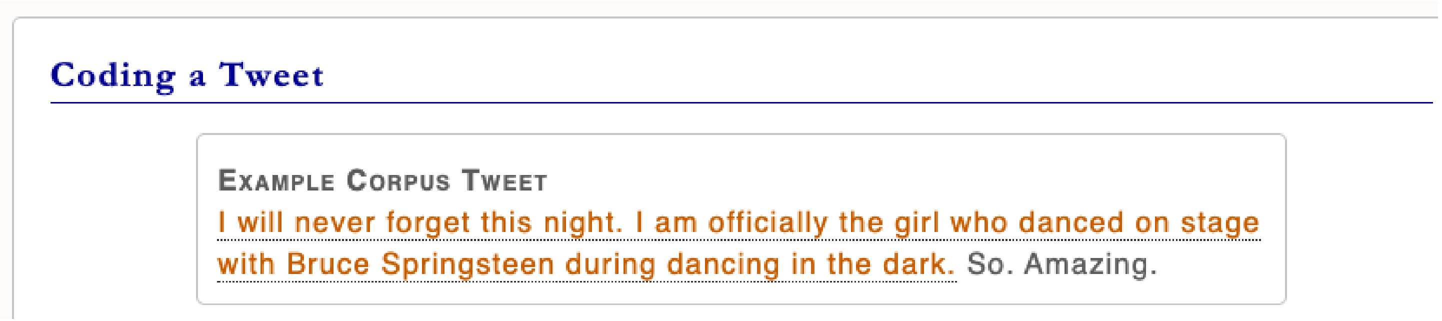
#

A TWEET THAT CONTAINS A HASHTAG

#Sp

See more Open Codes and Definitions

Baby, We Were Born to Tweet: Springsteen Fans, The Writing Practices of *In Situ* Tweeting, and the Research Possibilities for Twitter, **Wolff, Kairos journal, 19.3,**



Coding a Tweet

EXAMPLE CORPUS TWEET

I will never forget this night. I am officially the girl who danced on stage with Bruce Springsteen during dancing in the dark. So. Amazing.

During open coding for the above tweet, the following open codes were applied:

POT

A TWEET POSTED AFTER THE CONCERT OR ABOUT POST-CONCERT EVENTS

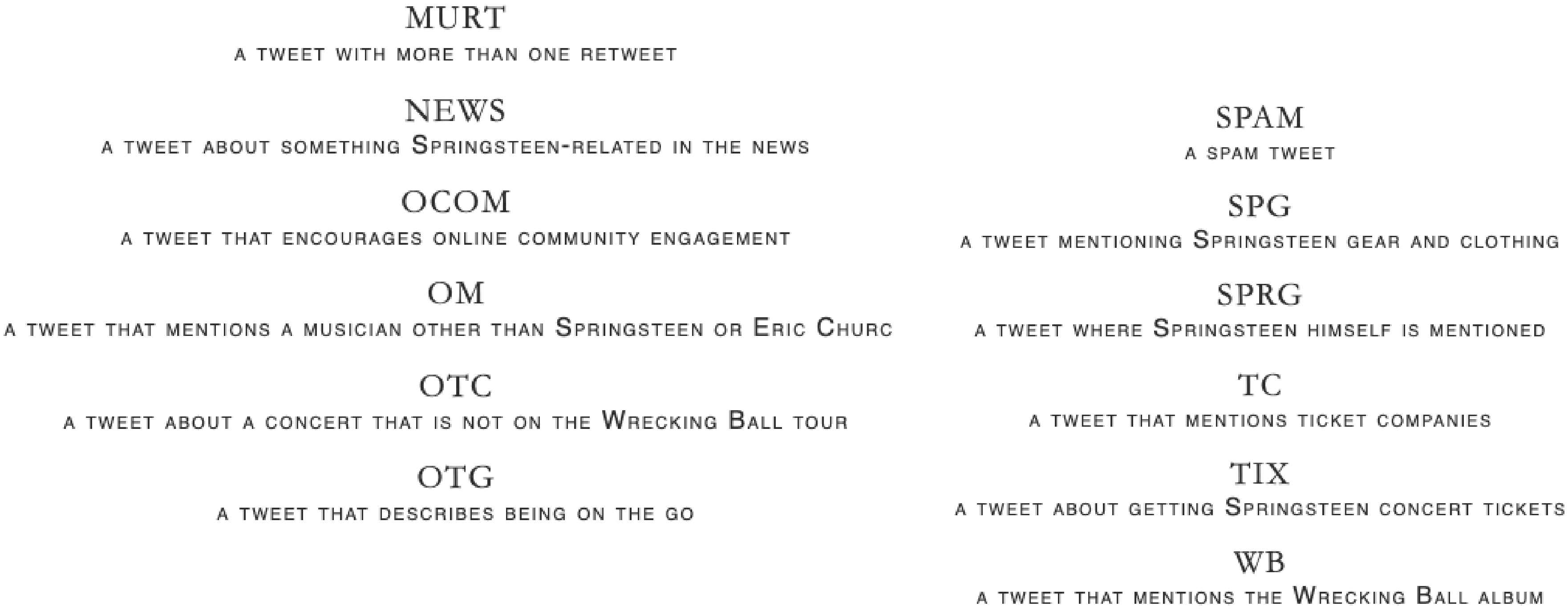
SOT

A TWEET THAT CONTAINS A SPRINGSTEEN SONG TITLE

SPRG

A TWEET WHERE SPRINGSTEEN HIMSELF IS MENTIONED

Open Codes



MURT

a tweet with more than one retweet

NEWS

a tweet about something Springsteen-related in the news

OCOM

a tweet that encourages online community engagement

OM

a tweet that mentions a musician other than Springsteen or Eric Churc

OTC

a tweet about a concert that is not on the Wrecking Ball tour

OTG

a tweet that describes being on the go

SPAM

a spam tweet

SPG

a tweet mentioning Springsteen gear and clothing

SPRG

a tweet where Springsteen himself is mentioned

TC

a tweet that mentions ticket companies

TIX

a tweet about getting Springsteen concert tickets

WB

a tweet that mentions the Wrecking Ball album

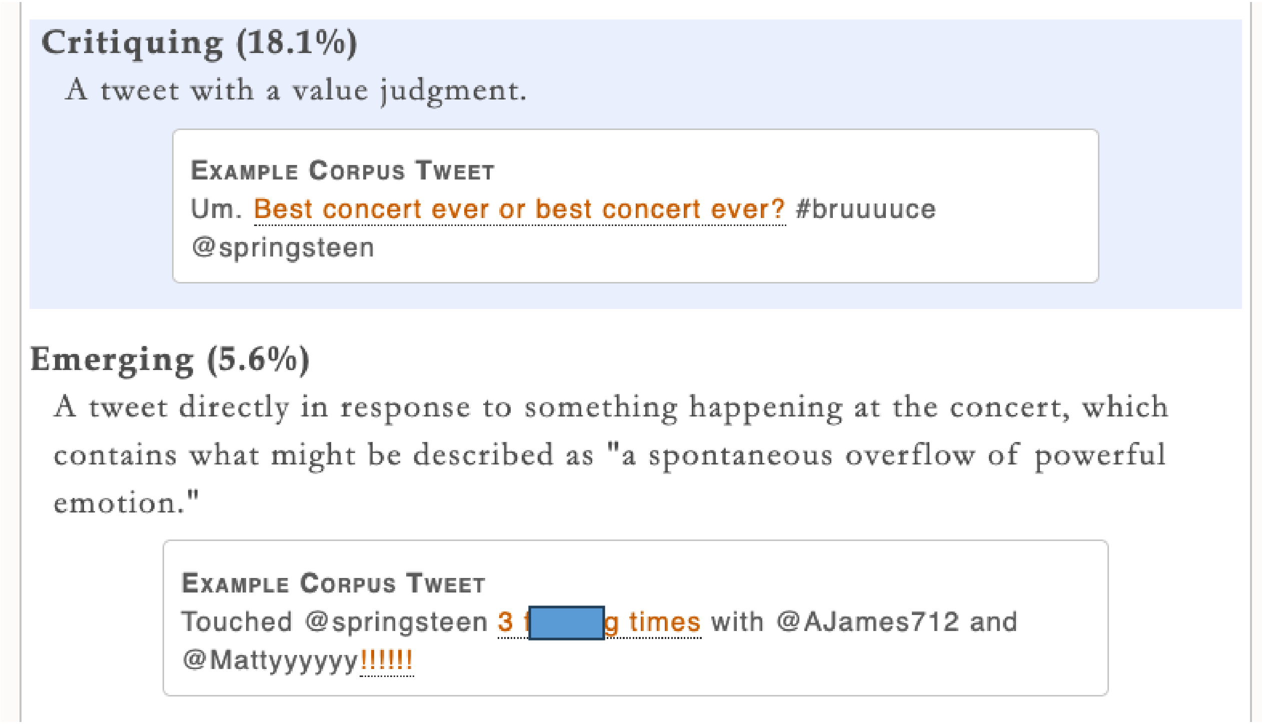
Baby, We Were Born to Tweet: Springsteen Fans, The Writing Practices of *In Situ* Tweeting, and the Research Possibilities for Twitter, **Wolff, Kairos journal, 19.3, 2015**

AXIAL CODING

- **Goal**: generate categories related to the focus; capture higher-level phenomena
- In this case study, the author decided to focus on pre, during, and post concert tweets because these were the most complex and embodied many of the other categories
- **Example**:
  - LYRIC: mention lyrics: ALB: mentioned album: SOT: song title
  - Intertextual: Tweets containing LYRIC, ALB, and/or SOT

Baby, We Were Born to Tweet: Springsteen Fans, The Writing Practices of *In Situ* Tweeting, and the Research Possibilities for Twitter, **Wolff, Kairos journal, 19.3, 2015**

Axial Codes



Critiquing (18.1%)

A tweet with a value judgment.

Example Corpus Tweet
Um. Best concert ever or best concert ever? #bruuuuce @springsteen

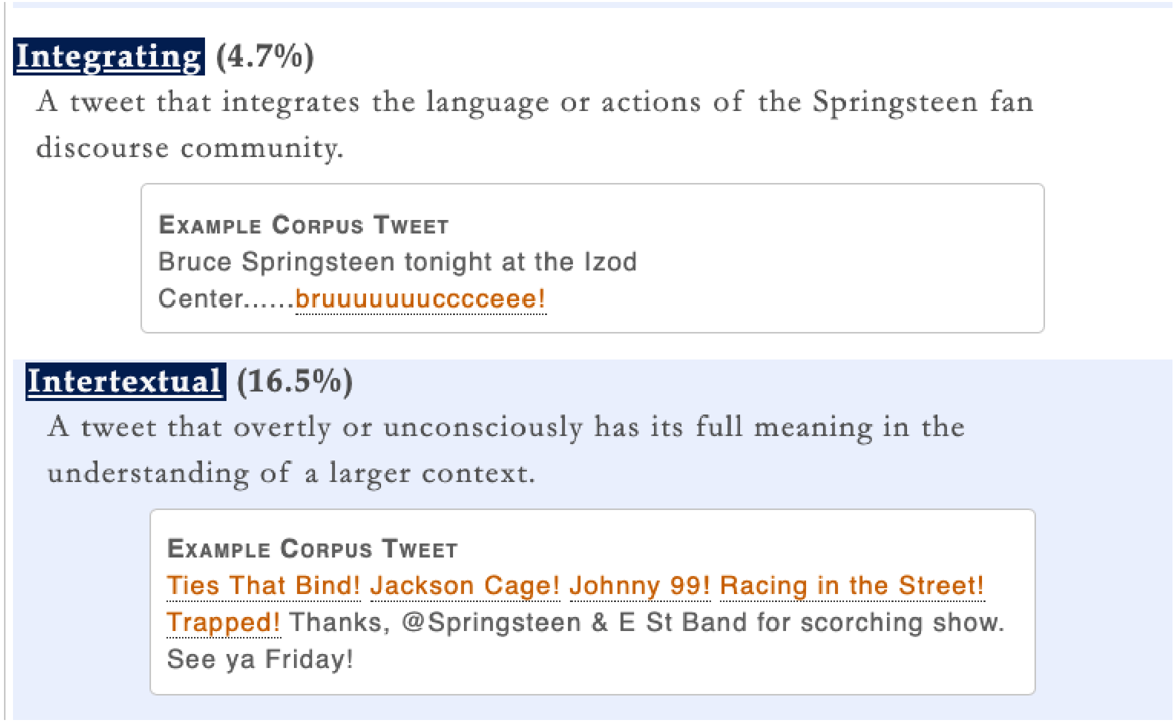
Emerging (5.6%)

A tweet directly in response to something happening at the concert, which contains what might be described as "a spontaneous overflow of powerful emotion."

Example Corpus Tweet
Touched @springsteen 3 f[redacted]g times with @AJames712 and @Mattyyyyyy!!!!!!

Baby, We Were Born to Tweet: Springsteen Fans, The Writing Practices of *In Situ* Tweeting, and the Research Possibilities for Twitter, **Wolff, Kairos journal, 19.3, 2015**

Axial Codes



Integrating (4.7%)

A tweet that integrates the language or actions of the Springsteen fan discourse community.

EXAMPLE CORPUS TWEET

Bruce Springsteen tonight at the Izod Center......bruuuuuuucccceee!

Intertextual (16.5%)

A tweet that overtly or unconsciously has its full meaning in the understanding of a larger context.

EXAMPLE CORPUS TWEET

Ties That Bind! Jackson Cage! Johnny 99! Racing in the Street! Trapped! Thanks, @Springsteen & E St Band for scorching show. See ya Friday!

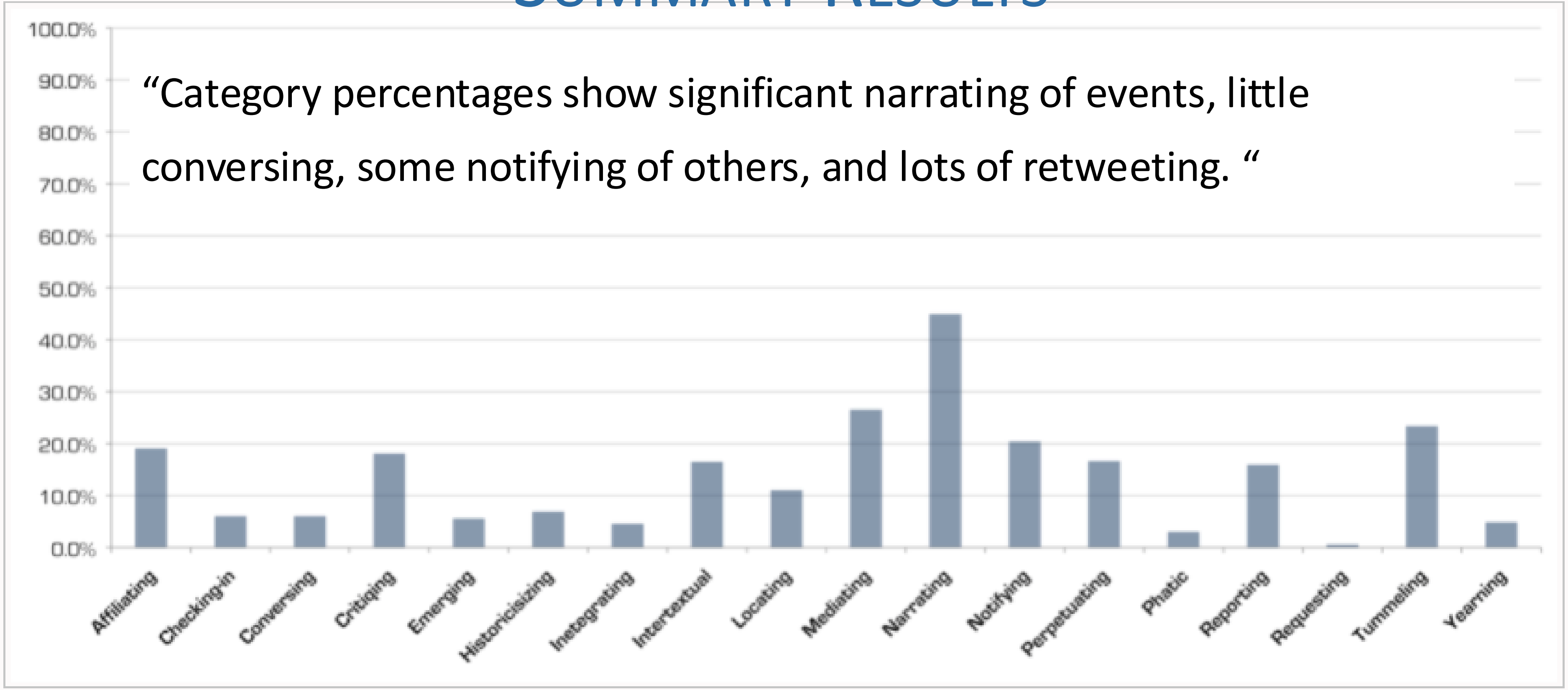
Baby, We Were Born to Tweet: Springsteen Fans, The Writing Practices of *In Situ* Tweeting, and the Research Possibilities for Twitter, **Wolff, Kairos journal, 19.3, 2015**

Selective Coding

- **Goal**: generate theories about the phenomena
- **Method**: make connections between the categories defined during axial coding
- **Example**:
  - Practicing: "fandom involves a particular set of critical and interpretive practices"
  - Composed of 5 out of the 18 axial codes

Baby, We Were Born to Tweet: Springsteen Fans, The Writing Practices of *In Situ* Tweeting, and the Research Possibilities for Twitter, **Wolff, Kairos journal, 19.3, 2015**

SUMMARY RESULTS

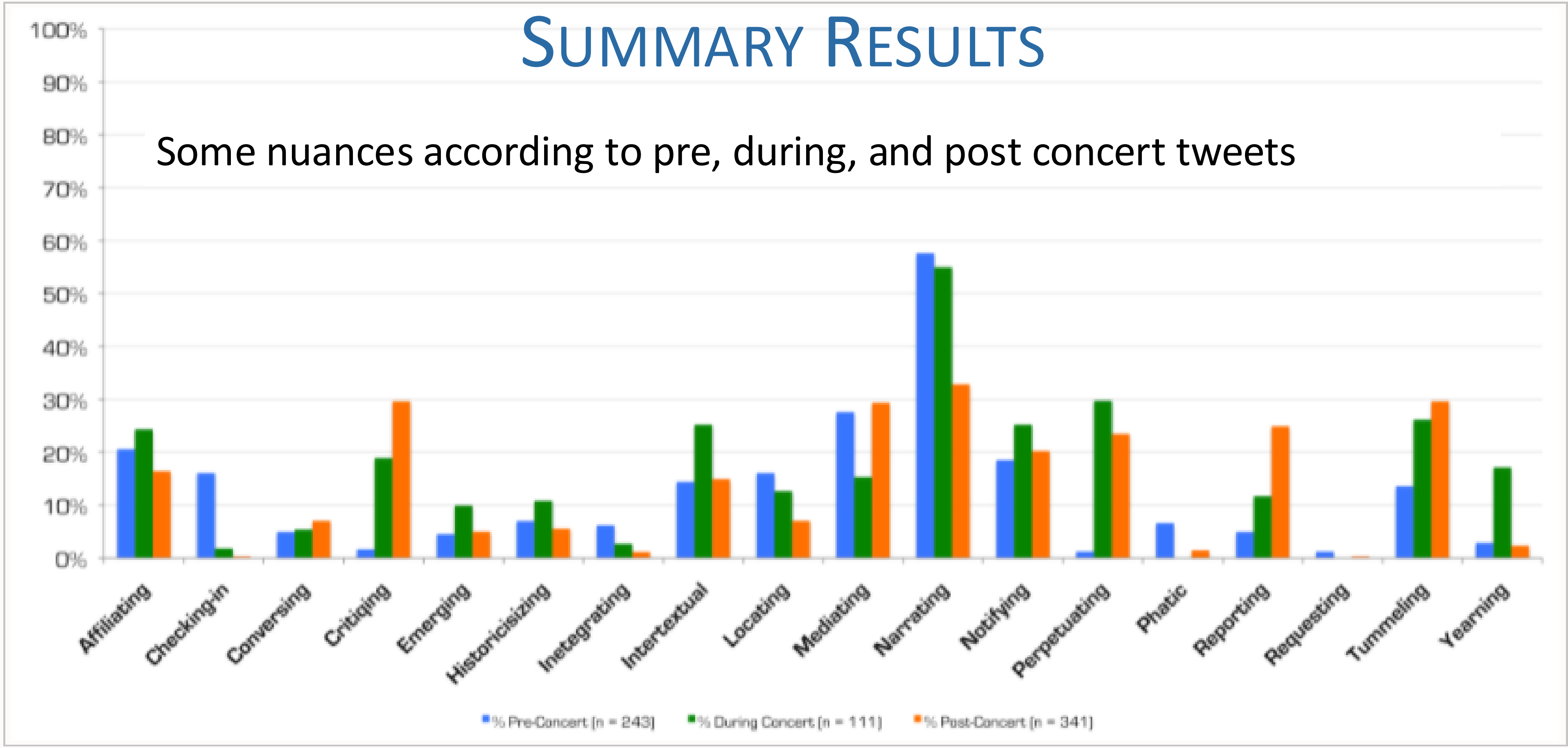


“Category percentages show significant narrating of events, little conversing, some notifying of others, and lots of retweeting. “

Baby, We Were Born to Tweet: Springsteen Fans, The Writing Practices of *In Situ* Tweeting, and the Research Possibilities for Twitter, **Wolff, Kairos journal, 19.3, 2015**

SUMMARY RESULTS

Some nuances according to pre, during, and post concert tweets



Baby, We Were Born to Tweet: Springsteen Fans, The Writing Practices of *In Situ* Tweeting, and the Research Possibilities for Twitter, **Wolff, Kairos journal, 19.3, 2015**

Assignment: Practice Making Open Codes
Sentences about Climate Change

1	supported	Global sea level rose about 8 inches in the last century.
2	supported	Extreme melting and changes to the climate like this has released pressure on to the continent, allowing the ground to rise up.
3	supported	The Great Barrier Reef is experiencing the most widespread bleaching ever recorded
4	refuted	Human additions of CO2 are in the margin of error of current measurements and the gradual increase in CO2 is mainly from oceans degassing as the planet slowly emerges from the last ice age.
5	refuted	The rate of warming according to the data is much slower than the models used by the IPCC

INTER-ANNOTATOR AGREEMENT

Calculate this to see if you coding strategy is reliable

What to do when your coders don’t agree?

Try, try again!

Refine the categories until you get decent inter-annotator agreement

Cohen's Kappa

- **Purpose:** Measures the agreement between **two** raters for categorical data, correcting for chance.
- **Application:** Ideal for situations where two human annotators classify a set of items into predefined categories, such as rating movie reviews as "positive" or "negative".
- **Weighted Kappa:** A variation used for ordinal data (ranked categories). It accounts for the magnitude of disagreement, giving partial credit for "close" ratings.
- **Limitations:**
  - Limited to two annotators.
  - Can produce counterintuitive results (the "Kappa paradox") when data is highly skewed toward one category

Advantage: easiest to understand & compute

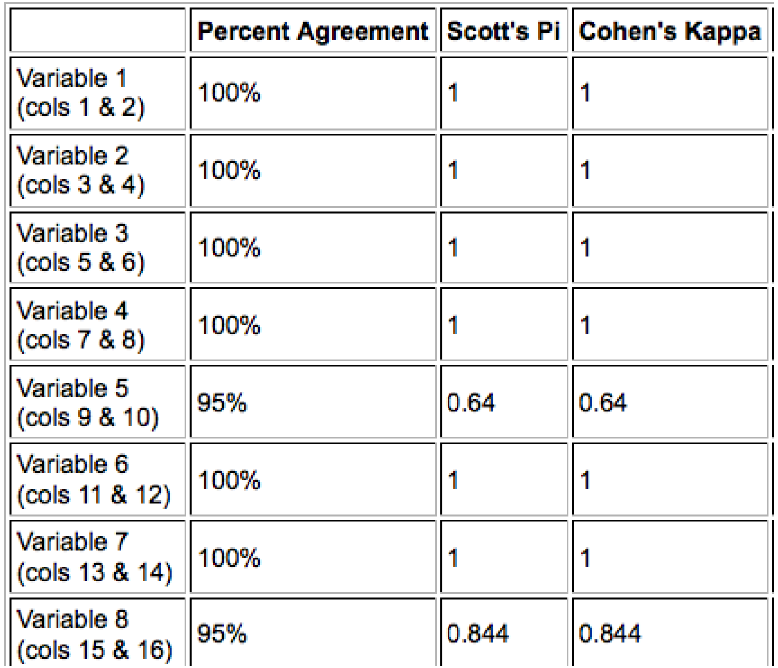
FLEISH'S KAPPA

- **Purpose:** An extension of Cohen's Kappa that measures the reliability of agreement for **three or more** raters with categorical data.
- **Application:** Suitable for larger annotation projects where multiple annotators are involved, such as assessing the severity of a medical condition or categorizing social media posts.
- **Assumption:** Assumes that the raters are randomly sampled from a population of raters. This means the same set of raters do not need to evaluate all items.
- **Key difference from Cohen's:** While Cohen's Kappa assumes the same two raters, Fleiss' Kappa can be used even if different raters evaluate different items.

KRIPPENDORFF'S ALPHA

- **Purpose:** A versatile reliability coefficient that can handle virtually any situation, including:
  - Any number of raters
  - Missing data
  - Any level of measurement (nominal, ordinal, interval, or ratio)
- **Application:** Often used in complex annotation tasks, such as coding open-ended text in content analysis or evaluating machine learning outputs.
- **Flexibility:** It is arguably one of the most robust and flexible inter-rater reliability measures available
- **Limitation**: More complex to understand (but supported by software packages)

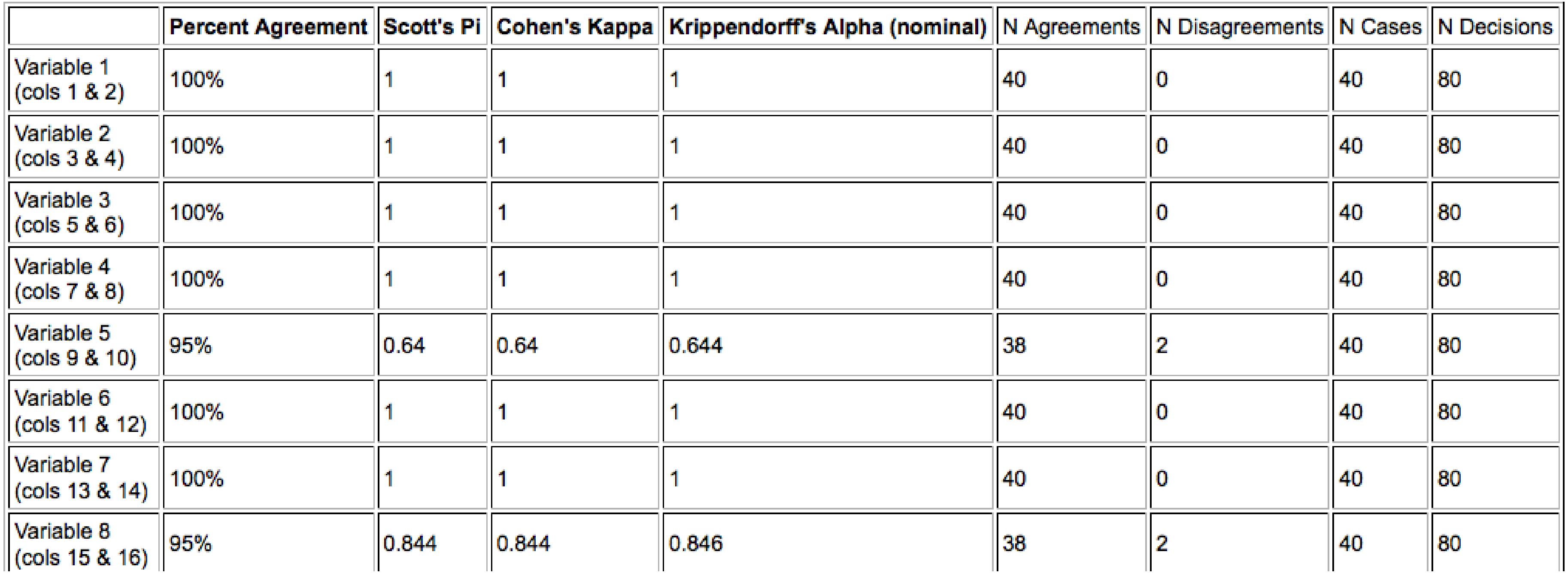
Inter-Annotator Agreement
2 coders; first 8 of 18 categories



	Percent Agreement	Scott's Pi	Cohen's Kappa
Variable 1 (cols 1 & 2)	100%	1	1
Variable 2 (cols 3 & 4)	100%	1	1
Variable 3 (cols 5 & 6)	100%	1	1
Variable 4 (cols 7 & 8)	100%	1	1
Variable 5 (cols 9 & 10)	95%	0.64	0.64
Variable 6 (cols 11 & 12)	100%	1	1
Variable 7 (cols 13 & 14)	100%	1	1
Variable 8 (cols 15 & 16)	95%	0.844	0.844

Baby, We Were Born to Tweet: Springsteen Fans, The Writing Practices of *In Situ* Tweeting, and the Research Possibilities for Twitter, **Wolff, Kairos journal, 19.3, 2015 kairos.technorhetoric.net/19.3/topoi/wolff/**

INTER-ANNOTATOR AGREEMENT
2 CODERS; FIRST 8 OF 18 CATEGORIES



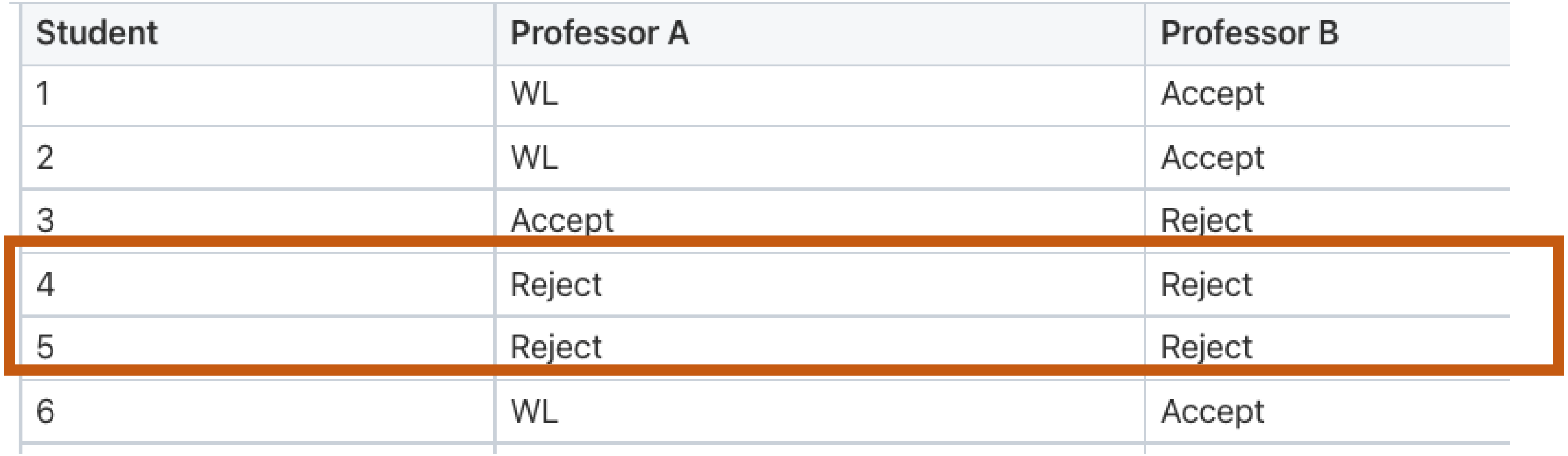
	Percent Agreement	Scott's Pi	Cohen's Kappa	Krippendorff's Alpha (nominal)	N Agreements	N Disagreements	N Cases	N Decisions
Variable 1 (cols 1 & 2)	100%	1	1	1	40	0	40	80
Variable 2 (cols 3 & 4)	100%	1	1	1	40	0	40	80
Variable 3 (cols 5 & 6)	100%	1	1	1	40	0	40	80
Variable 4 (cols 7 & 8)	100%	1	1	1	40	0	40	80
Variable 5 (cols 9 & 10)	95%	0.64	0.64	0.644	38	2	40	80
Variable 6 (cols 11 & 12)	100%	1	1	1	40	0	40	80
Variable 7 (cols 13 & 14)	100%	1	1	1	40	0	40	80
Variable 8 (cols 15 & 16)	95%	0.844	0.844	0.846	38	2	40	80

Baby, We Were Born to Tweet: Springsteen Fans, The Writing Practices of *In Situ* Tweeting, and the Research Possibilities for Twitter, **Wolff, Kairos journal, 19.3, 2015 kairos.technorhetoric.net/19.3/topoi/wolff/**

Measuring Inter-Annotator Agreement: Simple Proportions

Simple method: Simply compute the proportion of times the two annotators agree.

In the table below, the two Profs agree 2 times out of 6, or .33



Student	Professor A	Professor B
1	WL	Accept
2	WL	Accept
3	Accept	Reject
4	Reject	Reject
5	Reject	Reject
6	WL	Accept

(WL is waitlist)

Measuring Inter-Annotator Agreement: Simple Proportions

This method has drawbacks

Example: Labeling photos;

Some distinctions are easier than others

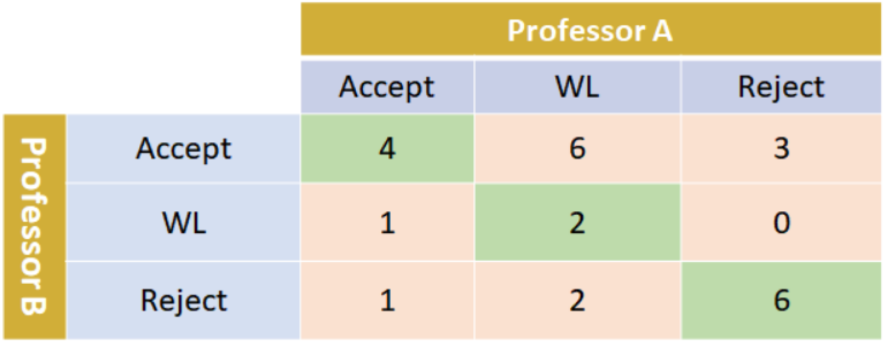
- cat vs porpoise is easy
- dolphin vs porpoise is difficult
- Therefore, two annotators are more likely to agree on cat/porpoise than dolphin/porpoise

Coder A	**Coder B**	**Agree?**
cat	cat	agree
cat	cat	agree
cat	cat	agree
cat	cat	agree
porpoise	dolphin	disagree
porpoise	dolphin	disagree
dolphin	dolphin	agree
dolphin	porpoise	disagree
dog	dog	agree
dog	cat	disagree

Cohen's Kappa is One Way to Measure Inter-Annotator Agreement

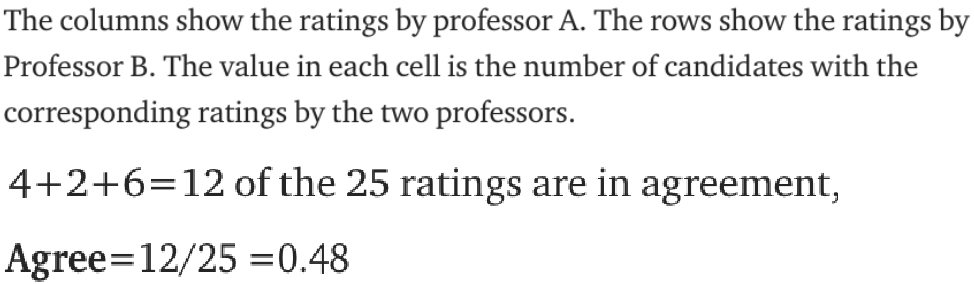
- It is a widely used measure.
  - *Krippendorff's alpha is preferred now, but tricker to compute*
- It takes into account that agreement can happen by chance.
- Ranges between 1 and -1
  - Values closer to 1 indicate high agreement.
  - Negative values indicate strong disagreement
- Ideally you achieve a score of around .8 or higher
  - However, scores are often lower if you have a lot of categories
  - Some categories are crisper than others

Example: Using Simple Proportions



		Professor A		
		Accept	WL	Reject
Professor B	Accept	4	6	3
	WL	1	2	0
	Reject	1	2	6

But: agreement can happen by chance!
Cohen’s Kappa gives you a measure of how good the agreement is taking chance agreement into account.



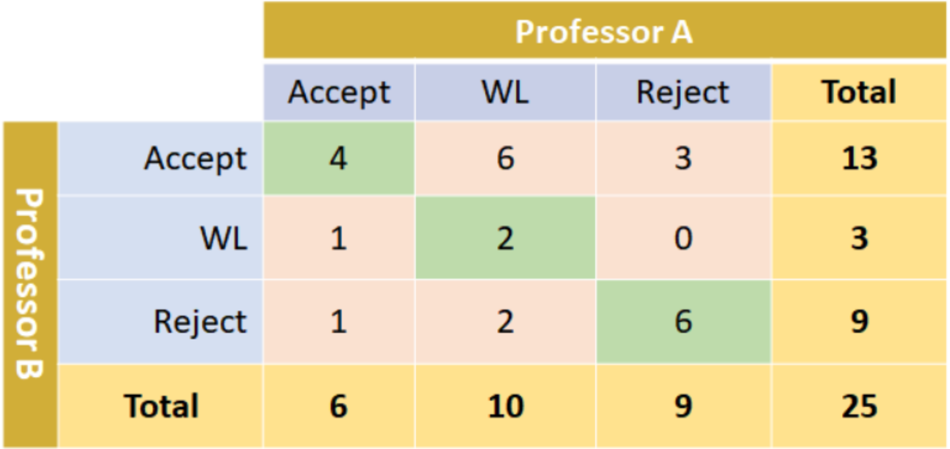
The columns show the ratings by professor A. The rows show the ratings by Professor B. The value in each cell is the number of candidates with the corresponding ratings by the two professors.

4+2+6=12 of the 25 ratings are in agreement,

Agree=12/25 =0.48

https://towardsdatascience.com/multi-class-metrics-made-simple-the-kappa-score-aka-cohens-kappa-coefficient-bdea137af09c

Example: Computing Cohen’s Kappa

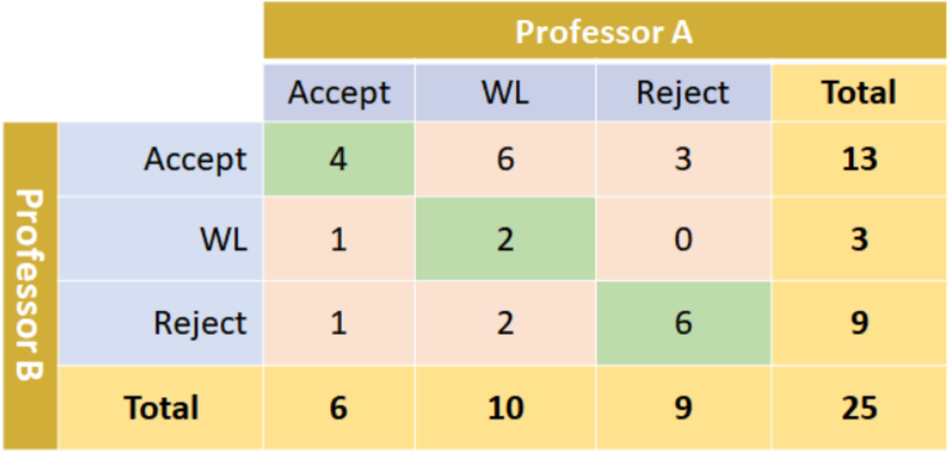


		Professor A			
		Accept	WL	Reject	Total
Professor B	Accept	4	6	3	13
	WL	1	2	0	3
	Reject	1	2	6	9
	Total	6	10	9	25

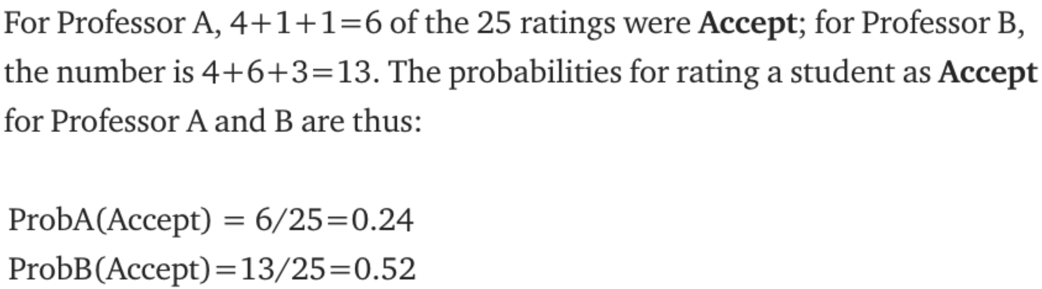
KappaScore = (Agree-ChanceAgree)/(1-ChanceAgree)

https://towardsdatascience.com/multi-class-metrics-made-simple-the-kappa-score-aka-cohens-kappa-coefficient-bdea137af09c

Cohen’s Kappa:
Compute Probability of Observed Agreement



		Professor A			
		Accept	WL	Reject	Total
Professor B	Accept	4	6	3	13
	WL	1	2	0	3
	Reject	1	2	6	9
	Total	6	10	9	25



For Professor A, 4+1+1=6 of the 25 ratings were **Accept**; for Professor B, the number is 4+6+3=13. The probabilities for rating a student as **Accept** for Professor A and B are thus:

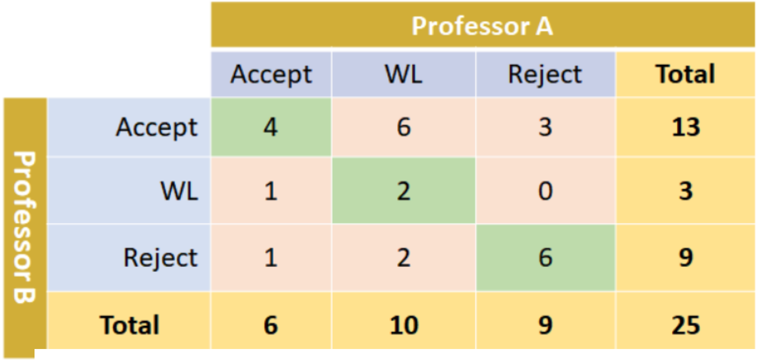
ProbA(Accept) = 6/25=0.24
ProbB(Accept)=13/25=0.52

Compute the observed probability of agreement for each label and coder

ProbA(Accept) = 6/25
ProbA(WL) = 10/25
ProbA(Reject) = 9/25
ProbB(Accept) = 13/25
ProbB(WL) = 3/25
ProbB(Reject) = 9/25

https://towardsdatascience.com/multi-class-metrics-made-simple-the-kappa-score-aka-cohens-kappa-coefficient-bdea137af09c

Cohen's Kappa:
Compute Probability of Chance Agreement



		Professor A			
		Accept	WL	Reject	Total
Professor B	Accept	4	6	3	**13**
	WL	1	2	0	**3**
	Reject	1	2	6	**9**
	Total	**6**	**10**	**9**	**25**

The probability of the chance of agreement for each label

Multiply the previous values for each label

The probability that both professors agree on an **Accept** *by chance* is equal to the product of ProbA and ProbB:

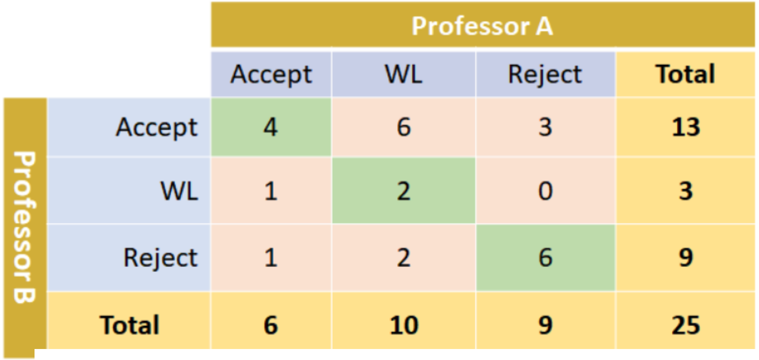
$$\text{ChanceAgree(Accept)} = \text{ProbA(Accept)} \times \text{ProbB(Accept)} = 0.24 \times 0.52 = \mathbf{0.1248}$$

$$\text{ChanceAgree} =$$
$$= \text{ChanceAgree(Accept)} + \text{ChanceAgree(WL)} + \text{ChanceAgree(Reject)}$$
$$= 0.1248 + 0.0480 + 0.1188 = \mathbf{0.3024}$$

$$\text{KappaScore} = \text{(Agree-ChanceAgree)/(1-ChanceAgree)}$$
$$= (0.48–0.3024)/(1–0.3024)$$
$$= 0.2546$$

https://towardsdatascience.com/multi-class-metrics-made-simple-the-kappa-score-aka-cohens-kappa-coefficient-bdea137af09c

Put the Cohen’s Kappa Formula Together



		Professor A			
		Accept	WL	Reject	Total
Professor B	Accept	4	6	3	**13**
	WL	1	2	0	**3**
	Reject	1	2	6	**9**
	Total	**6**	**10**	**9**	**25**

The Kappa Score is:

(ProbObservedAgreement – ChanceAgreement)
Divided by (1 – ChanceAgreement)

The probability that both professors agree on an **Accept** *by chance* is equal to the product of ProbA and ProbB:

ChanceAgree(Accept)=ProbA(Accept) × ProbB(Accept)=0.24×0.52=**0.1248**

ChanceAgree=
= ChanceAgree(Accept)+ChanceAgree(WL)+ChanceAgree(Reject)
= 0.1248+0.0480+0.1188=**0.3024**

KappaScore = (Agree-ChanceAgree)/(1-ChanceAgree)
= (0.48–0.3024)/(1–0.3024)
= 0.2546

Kappa should be above about .67
This shows very low agreement between the professors

https://towardsdatascience.com/multi-class-metrics-made-simple-the-kappa-score-aka-cohens-kappa-coefficient-bdea137af09c

WHAT ABOUT MULTIPLE LABELS PER ITEM?

- Situation: 2 people assigning codes to 3 sentences from interviews
- The coders want to assign more than one code per sentence

sentence	Coder A	Coder B
1	{Happy, Sad}	{Happy}
2	{Sad}	{Sad, Confused}
3	{Confused}	{Confused}

What about multiple labels per item?

- We expand this table out to assign a binary (0 or 1) to each combination

sentence	Coder A	Coder B
1	{Happy, Sad}	{Happy}
2	{Sad}	{Sad, Confused}
3	{Confused}	{Confused}

sentence	Code	A	B
1	Happy	1	1
1	Sad	1	0
1	Confused	0	0
2	Happy	0	0
2	Sad	1	1
2	Confused	0	1
3	Happy	0	0
3	Sad	0	0
3	Confused	1	1

Cohen's Kappa: Multiple Labels Per Item

Compute Observed Probability of Agreement

sentence	Code	A	B	Agree = 1
1	Happy	1	1	1
1	Sad	1	0	0
1	Confused	0	0	1
2	Happy	0	0	1
2	Sad	1	1	1
2	Confused	0	1	0
3	Happy	0	0	1
3	Sad	0	0	1
3	Confused	1	1	1

(Agree-ChanceAgree)/(1-ChanceAgree)

Agree 7/9 times = .778

Cohen’s Kappa: Multiple Labels Per Item

Compute Chance Probability of Agreement

sentence	Code	A	B	Agree = 1
1	Happy	1	1	1
1	Sad	1	0	0
1	Confused	0	0	1
2	Happy	0	0	1
2	Sad	1	1	1
2	Confused	0	1	0
3	Happy	0	0	1
3	Sad	0	0	1
3	Confused	1	1	1

$$(Agree-\boxed{ChanceAgree})/(1-\boxed{ChanceAgree})$$

Compute Chance Agreement
(also called Expected Agreement)

Compute probability that coder A ever said 1 (4/9)
Compute probability that coder B ever said 1 (4/9)
Compute probability that coder A ever said 0 (5/9)
Compute probability that coder B ever said 0 (5/9)

Multiply the chance for label 1 (4/9 * 4/9)
Multiply the chance for label 0 (5/9 * 5/9)
Add these together: .1975 + .308 = .506

Cohen’s Kappa: Multiple Labels Per Item
Compute Chance Probability of Agreement

sentence	Code	A	B	Agree = 1
1	Happy	1	1	1
1	Sad	1	0	0
1	Confused	0	0	1
2	Happy	0	0	1
2	Sad	1	1	1
2	Confused	0	1	0
3	Happy	0	0	1
3	Sad	0	0	1
3	Confused	1	1	1

(Agree-ChanceAgree)/(1-ChanceAgree)

Compute Chance Agreement
(also called Expected Agreement)

Compute probability that coder A ever said 1 (4/9)
Compute probability that coder B ever said 1 (4/9)
Compute probability that coder A ever said 0 (5/9)
Compute probability that coder B ever said 0 (5/9)

Multiply the chance for label 1 (4/9 * 4/9)
Multiply the chance for label 0 (5/9 * 5/9)
Add these together:

ChanceAgree = .1975 + .308 = .506

Multiple Labels Per Item
Cohen’s Kappa

sentence	Code	A	B	Agree = 1
1	Happy	1	1	1
1	Sad	1	0	0
1	Confused	0	0	1
2	Happy	0	0	1
2	Sad	1	1	1
2	Confused	0	1	0
3	Happy	0	0	1
3	Sad	0	0	1
3	Confused	1	1	1

(Agree-ChanceAgree)/(1-ChanceAgree)

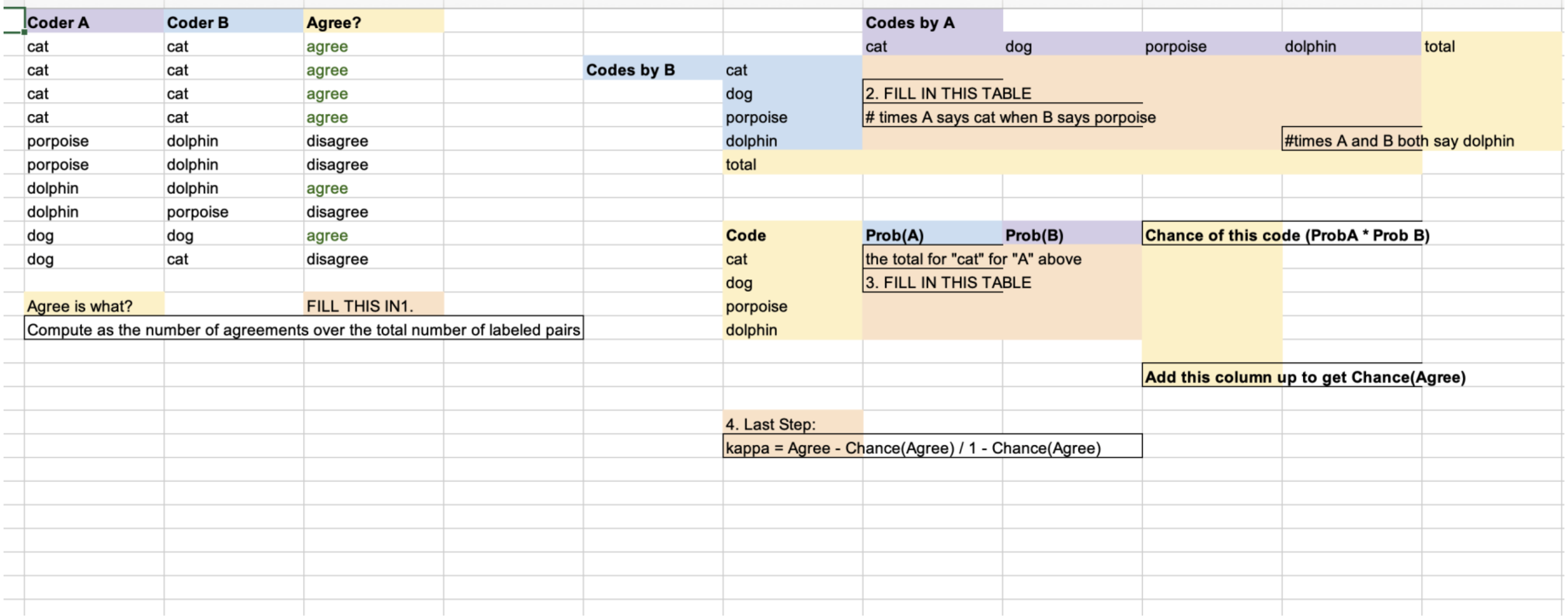
Agree = .778
ChanceAgree = .506

$$= \frac{0.778 - 0.506}{1 - 0.506} = \frac{0.272}{0.494} = 0.55$$

This is a middling score
The coders probably need to improve their codebook

Let’s Practice

Download spreadsheet on course website (next to lecture notes)



Coder A	**Coder B**	**Agree?**
cat	cat	agree
cat	cat	agree
cat	cat	agree
cat	cat	agree
porpoise	dolphin	disagree
porpoise	dolphin	disagree
dolphin	dolphin	agree
dolphin	porpoise	disagree
dog	dog	agree
dog	cat	disagree
Agree is what?		FILL THIS IN1.

Compute as the number of agreements over the total number of labeled pairs

		Codes by A				
		cat	dog	porpoise	dolphin	total
Codes by B	cat					
	dog	2. FILL IN THIS TABLE				
	porpoise	# times A says cat when B says porpoise				
	dolphin				#times A and B both say dolphin	
	total					

Code	**Prob(A)**	**Prob(B)**	**Chance of this code (ProbA * Prob B)**
cat	the total for "cat" for "A" above		
dog	3. FILL IN THIS TABLE		
porpoise			
dolphin			
			Add this column up to get Chance(Agree)

4. Last Step:

kappa = Agree - Chance(Agree) / 1 - Chance(Agree)

OTHER MEASURES

- Cohen’s Kappa is controversial today due to some limitations.
- Other measures can be better for more than 2 coders
- Other measures take into account that some categories are more important than others

See Antoine et al., Weighted Krippendorff’s alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation, EACL 2014
https://www.aclweb.org/anthology/E14-1058/

SUMMARY: GROUNDED CODING

- Is a method for creating systematic and consistent categories for qualitative data
- Is usually needed even if you are going to use machine learning, since you have to understand your target.
- Requires multiple annotators to be confident in results