Representing Data

Dr Peaks Krafft (they/them), Sociology

THE UNIVERSITY of EDINBURGH
informatics

THE UNIVERSITY of EDINBURGH
School of Social
& Political Science

Challenge of The Course

- Every Week:
  Understanding complex social phenomena using big data
- This Week:
  The first step of working with data.
  Representing it in a computer!

Learning Goals

- Describe the range of data sources social scientists use for empirical work
- Choose suitable types of data for different kinds of research questions
- Evaluate the pros and cons of quantification and computational approaches versus common alternatives

About Me

- UG in Mathematics & Statistics
- Wasn't fond of social science
- Took a seminar on Computational Social Science
- Went to work with author of CSS paper in Science
- PhD in Computer Science at MIT
- Advised not to read Sociology..

Now

- Lecturer in Computational Sociology
- Co-direct the MSc Digital Sociology
- Researching in the Sociology of Knowledge

Sociology of Knowledge

- Ways of knowing
- How people come to know things or believe things about the world
- Methods – what you do
- Methodology – the study of methods

Computational Social Science

- Informatics studies abstraction, efficiency, formal models
- Sociology studies power, inequality
- What do computational social scientists study?
- What defines it as a distinct field?

Computational Social Science

- CSS is a methodological area
  (not a social science in its own right)
- Characterised by the inventive use of data and methods
- Today: Methods of data collection

Basic Research Framework

- Research Question & Hypothesis
  - RQ: What makes persuasion effective?
  - Hypothesis: Evidence makes persuasion effective
- Data Collection
  - Download Posts & Replies from Change My View Subreddit
- Sample Population
  - Redditors in Change My View as proxy for “people engaging in persuasion online”
- Methods & Analysis
  - Identify “Evidence” in replies (e.g., search for “http” links)
- Measure & Report Outcome
  - Persuasion rate of replies with evidence vs. without evidence

This Week: Representing Data

- Research requires thoughtful planning
- Setting a research question is a crucial part of the planning process. Tj discussed last week. Will recur throughout semester
- Planning data collection is usually the next step.

This Week: Representing Data

- Using Big Data for Computational Social Science:
  Need to have an idea of what kind of data might be helpful
- Also importantly:
  Big Data is not always the most useful method!
- Must be aware of alternative options

Poll

- How many people here have worked with some type of data?

Survey

- Can you tell me about some other types of data?

Survey

- Why do you think a lot of social scientists don't like Big Data so much?

Plan for Today

1. **What types of data are there?**
2. Where does data come from?
3. What is data useful for?

What types of data are there?

- Qualitative Data: Interviews, archival texts, historical records
- Quantitative Data: Surveys, opinion polls, census counts

What types of data are there?

- Qualitative Data: Interviews, archival texts, historical records
- Quantitative Data: Surveys, opinion polls, census counts
- "Big Data": web data, location data, purchase data, etc.

What types of data are there?

- Last week (and most of the course):
  Focus is "Big Data", web data especially
- Today:
  Bigger picture view

Quantification

- "quantitative", "quantity" = "measurement"
- "data", "datum" = "to give", "to present"
- To give measurement to phenomena (people, events, etc.)

Quantification

(Not) Quantification – Art School

What do you see in this image?

(Not) Quantification – Art School

What do you see in this image?

What does it mean to you?

(Not) Quantification – Art School

What do you see in this image?

- Trees, cars, crosswalk, buildings, road, pavement, people, suits, symmetry

What does it mean to you?

- Time period, band members, order, law, hippies, confidence, going somewhere.. Together?

Quantification

Quantification

Image ID	George	Paul	Ringo
AbbeyRoad.png	1	1	1

Quantification

What is not in the image?

Quantification

Image ID	People
AbbeyRoad.png	4

Quantification

Image ID	People	Dogs
AbbeyRoad.png	4	0

(Not) Quantification – Art School

What is not in the image?

- Band name
- Album name
- Final Beatles album recorded
- First song: "Come Together"
- A time of major conflict in the band
- Most spoofed album cover ever?

Qualitative Data

- "qualitative", "quality" = "nature", "character"
- "data", "datum" = "to give", "to present"
- To present the nature of phenomena (people, events, etc.)

Qualitative Data

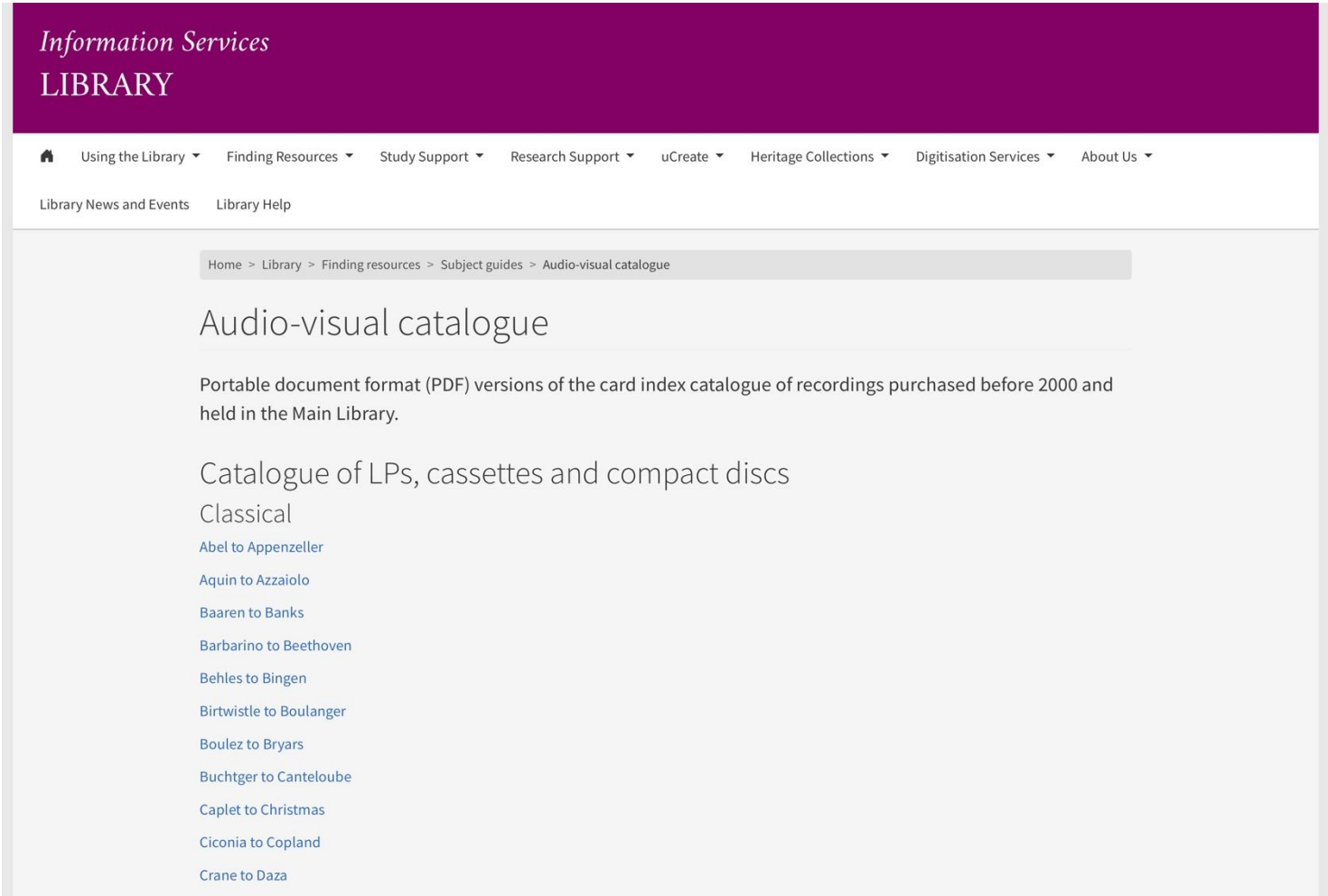
- Is there a common vocabulary of symbols that album covers deploy to signify cultural meanings and contexts?

How might you try to answer this question?

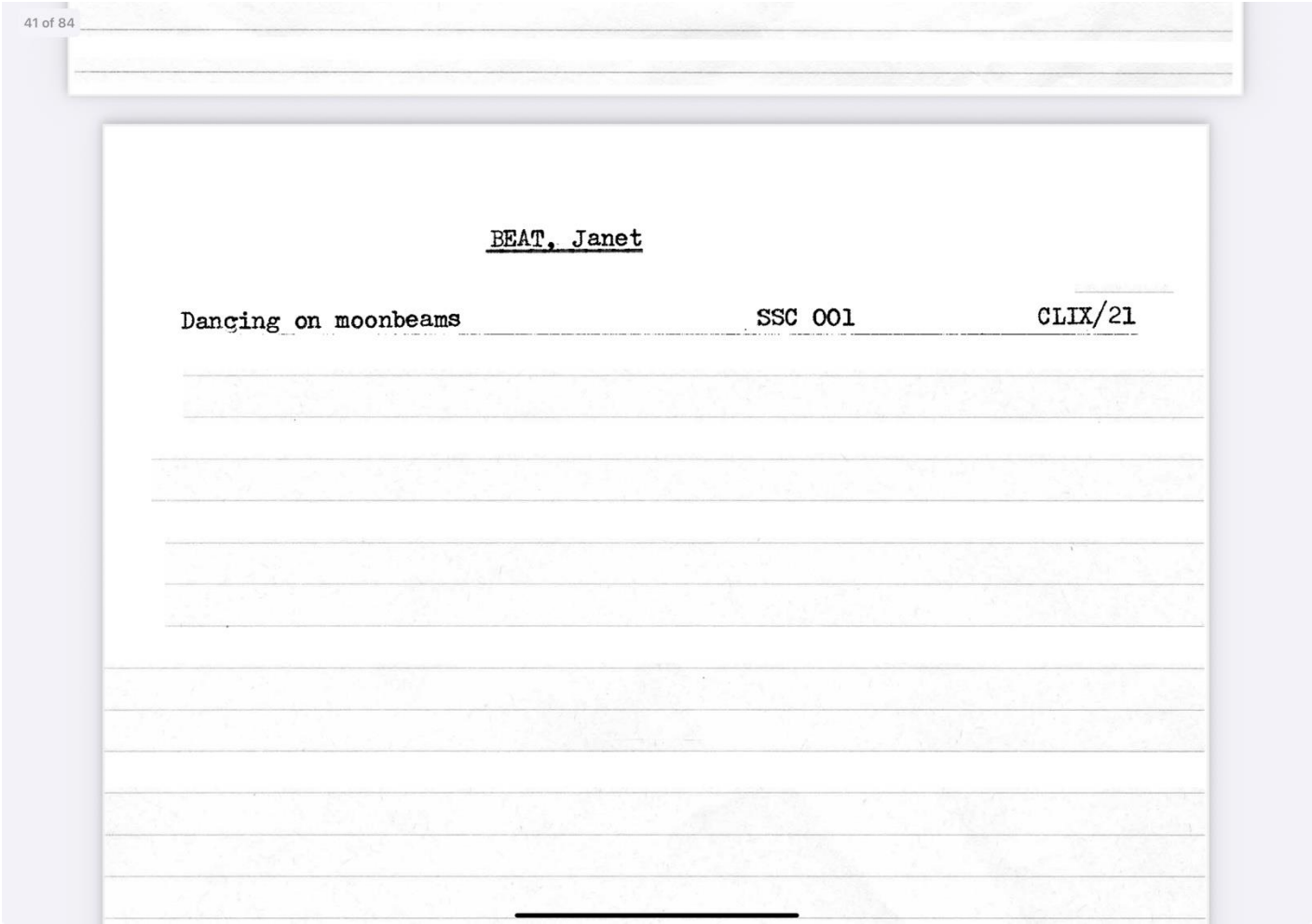
Qualitative Data

- Is there a common vocabulary of symbols that album covers deploy to signify cultural meanings and contexts?
- The library has a large collection of Compact Disks (CDs)

Qualitative Dataset



Qualitative Dataset



Qualitative Data

- Usually ends up focusing on examples, cases, use of data in context of arguments
- E.g., refine research question: What does it say about an album if its title is not on the album cover image?
- Analyse a set of cases, make an argument

Big Data Solution

- How might you use a Big Data/Computational approach to the same question?

Is there a common vocabulary of symbols that album covers deploy to signify cultural meanings and contexts?

Big Data Solution

- Is there a common vocabulary of symbols that album covers deploy to signify cultural meanings and contexts?
- Spotify data? Wikipedia data?
- Pre-defined set of symbols?
- Ask AI?
- Get auxiliary data: Album sales?
- Refine research question: How is the album cover including the album title or not correlated with sales?

"Mixed Methods" Solution

Visual Narratives and Collective Memory across Peer-Produced Accounts of Contested Sociopolitical Events

EMILY PORTER, University of Washington
P. M. KRAFFT*, University of Oxford, UK
BRIAN KEEGAN, University of Colorado

Studying cultural variation in recollections of sociopolitical events is crucial for achieving diverse understandings of such events. To date, most studies in this area have focused on analyzing variation in texts describing events. Here, we analyze variation in image usage across Wikipedia language editions to understand if, like text, visual narratives reflect distinct perspectives in articles about culturally-tethered events. We focus on

"Mixed Methods" Solution

Table 2. Framework for Coding Crisis Imagery by Sub-Code, Count of Usage Across Images (Multiple Codes Can be Applied Per Image), Inter-Rater Reliability (Cohen's Kappa), and Description of Sub-Codes for $N = 761$ Images from Coup and Crisis Articles

Code	Sub-codes	Count	IRR	Descriptions
Activity	Active	206	0.719	Protesting or fighting
	Passive	549	0.679	Portraits, meetings, or landscapes
Demographics	Adult male	447	0.897	Primary subject(s) of image is an adult man or men
	Adult female	81	0.846	Primary subject(s) of image is an adult woman or women
	Child male	7	0.492	Primary subject(s) of image is a young boy or boys
	Child female	15	0.697	Primary subject(s) of image is a young girl or girls
	Cannot discern*	68	0.777	There is no primary subject in the image
	Non-human	224	0.896	There are no humans in the image
Facial expression	Frown	33	0.488	Primary subject(s) of image are frowning or distressed
	Smile	114	0.842	Primary subject(s) of image are smiling or pleased
	Unclear*	299	0.667	Primary subject(s) of image have ambiguous expressions
Foreign influence	Foreign leaders	31	0.905	Images of leaders from different countries meeting
	Foreign flag	15	0.931	Images of symbols from different countries
	Alleged atrocities	14	0.819	Images of dead bodies
	Buildings	84	0.828	Images of government, religious, etc. structures

"Mixed Methods" Solution

Table 3. The Photo on the Left Containing Boris Yeltsin on a Tank is Representative of an "Outsider" Image

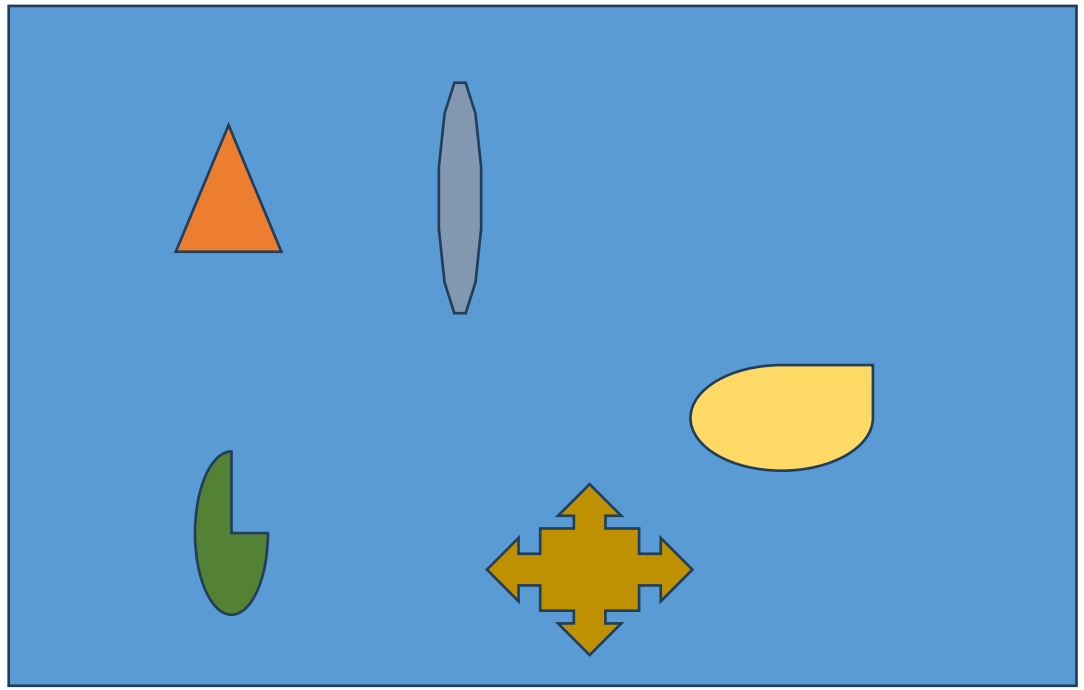
Topic	**1991 Soviet coup d'état attempt**	
		
File name	"Boris Yeltsin 19 August 1991-1.jpg"	"1991 CPA 6369.jpg"
Codes	VI-Non-violent Image; IT-Passive Image; IF-Group; FE-Unclear; D-Adult Male; TOP-Military or Political Leaders; TOP-Civilians; TOP-Journalists; GT-Military Technology; GT-Diplomacy; GT-Buildings	VI-Non-violent Image; IT-Passive Image; D-Adult Male; IF-Individual; FE-Smile; D-Adult Male; TOP-Military or Political leaders; GT-Portraits; GT-Signs and Symbols
Languages	Arabic; Bavarian; Bulgarian; Catalan; Czech; German; English; Esperanto; Spanish; Estonian; Basque; Persian; Finnish; French; Galician; Hebrew; Croatian	Bulgarian; Czech; Chuvash; Norwegian; Polish; Romanian; Russian; Ukrainian
Other articles	1991; Boris Yeltsin; History of the Soviet Union; Moscow; Stalin (1992 film); Russia; White House	None

Qual versus Quant Data

- What is lost in quantification?

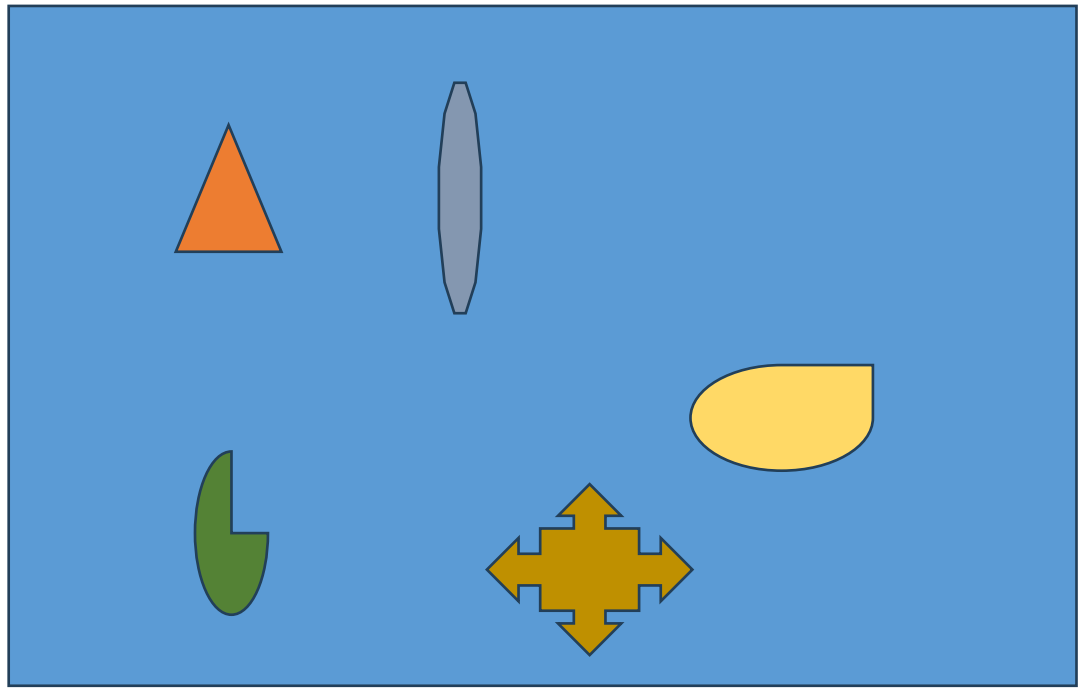
Qual vs Quant: Big Picture

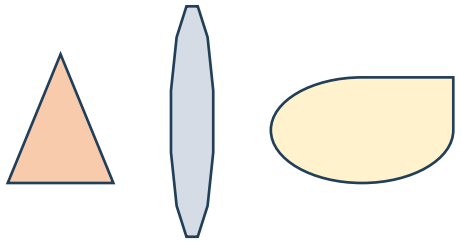
The Phenomenal Universe



Qualitative Data: Big Picture

The Phenomenal Universe



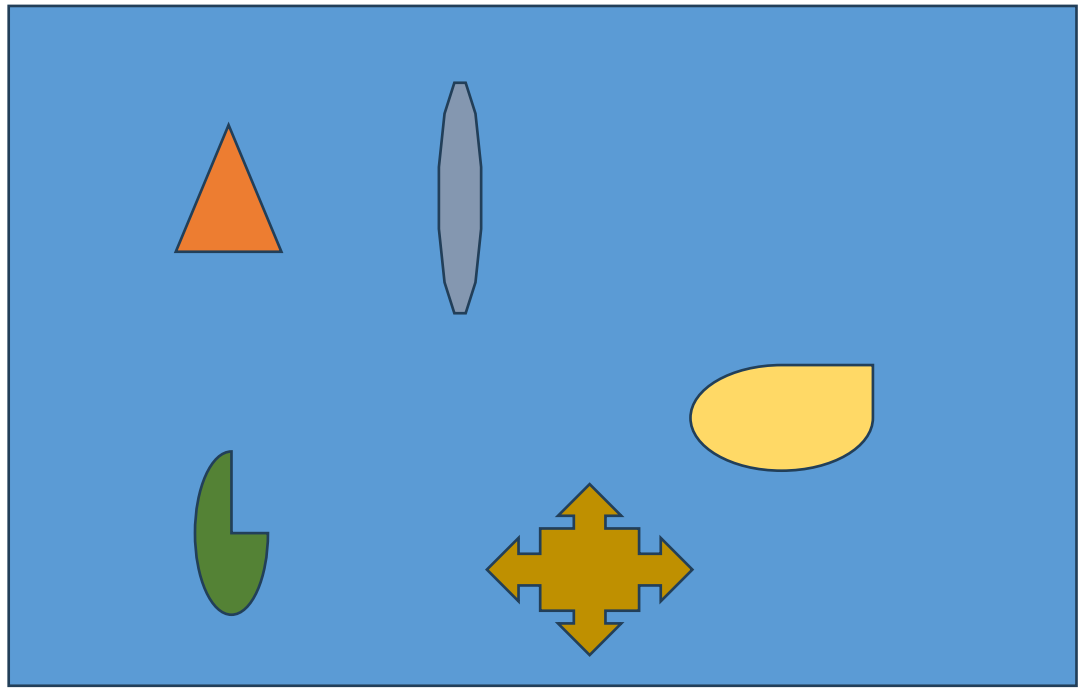


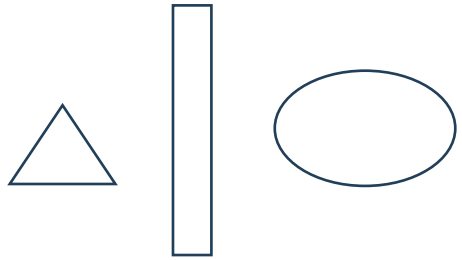
A pale reflection

Curating qualitative data (taking photos, collecting interviews, documentary evidence, etc.)

Quantification: Big Picture

The Phenomenal Universe





1x triangle, 1x rectangle, 1x oval

Choosing what to measure, abstracting through metrics

Questions?

Break

- 10-minute break

Plan for Today

1. What types of data are there?
2. **Where does data come from?**
3. What is data useful for?

Data Representations

- Data, datum = "to give", "to present"
- Representation = "re"-presentation
- Presenting again in a different way

Data Representations

- Data, datum = "to give", "to present"
- Representation = "re"-presentation
- Presenting again in a different way
- Where does data come from?

Data Representations

- Presenting again in a different way
- Where does data come from?
  What is being presented?
  What is data re-presenting?

Data Representations

- Presenting again in a different way
- What is being represented?

Data Representations

- Presenting again in a different way
- What is being represented?
- Phenomena, experience, identity

Data Representations

- What is being represented?
- Phenomena, experience, identity
- Quantitative = re-presenting via numbers (for a computer)
- Qualitative = re-presenting via curation

Where does data come from?

Quantitative:

- Surveys, Digital Trace Data, Experiments

Qualitative:

- Interviews, Participant-observation, Archival research

Examples from Last Week

- Reddit data (Digital Trace Data)
- Reddit data (Experiment)
- Slurp data (Survey)

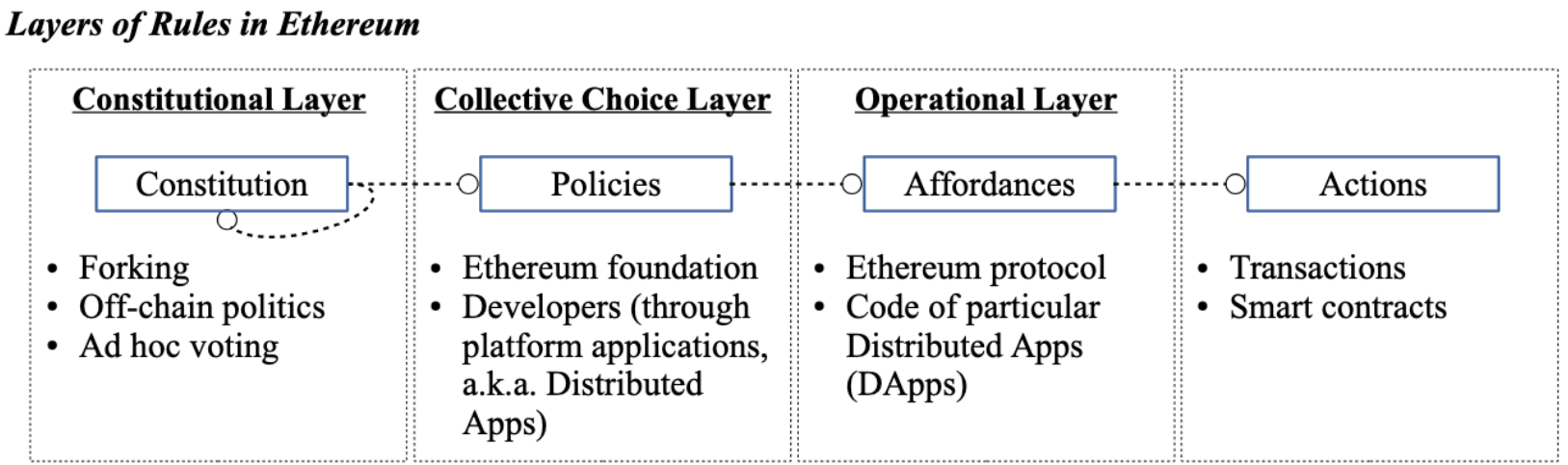
Example

Research Question: Is crypto a scam?

Three different approaches

Example: Qualitative

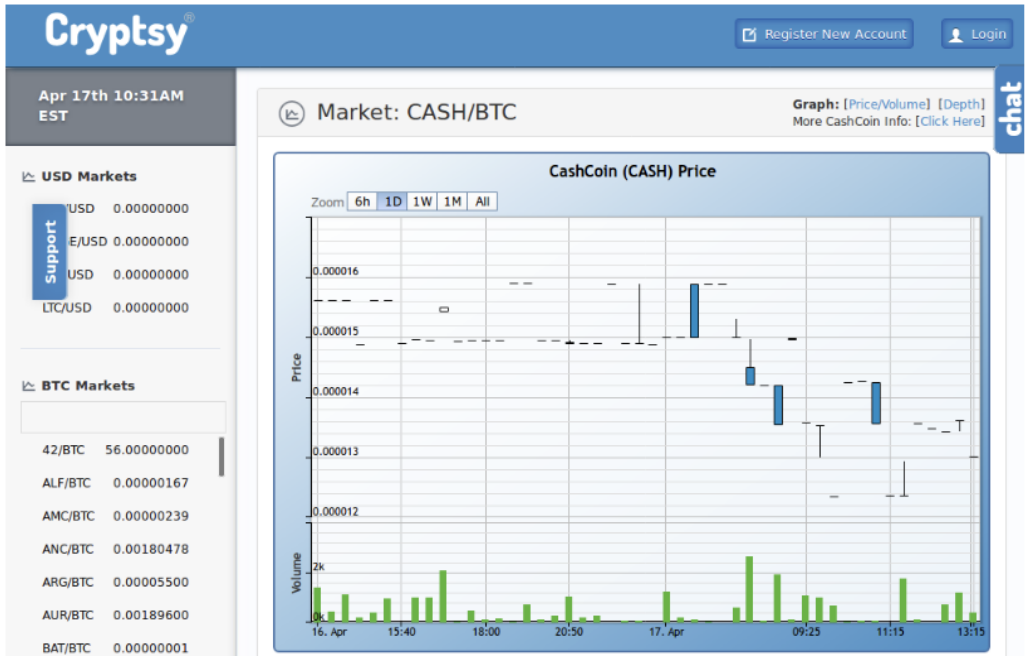
Focus is on understanding context and detail



Frey, Krafft, and Keegan (2019)

Example: Quantitative (Experiment)

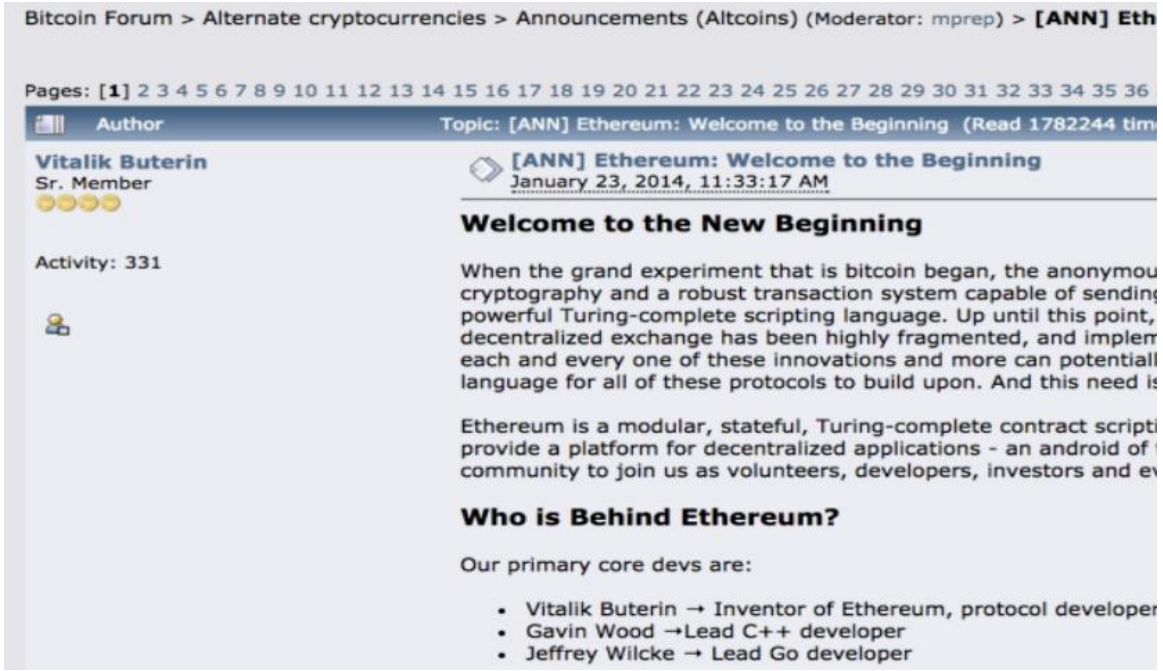
Focus is on narrow testable hypothesis



Krafft, Della Penna, and Pentland (2018)

Example: Big Data (Digital Trace)

Often more exploratory



Jahani, Krafft, Suhara, Moro, and Pentland (2018)

Plan for Today

1. What types of data are there?
2. Where does data come from?
3. **What is data useful for?**

Ground Truth

- "Ground truth" - a phrase you will hear a lot

Ground Truth

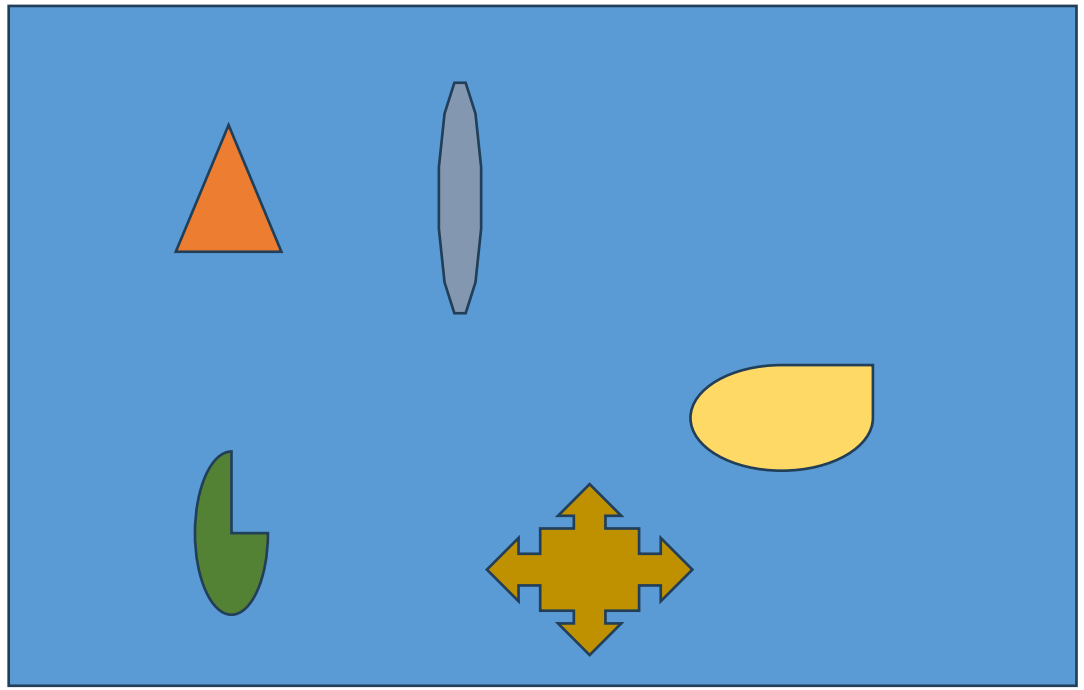
- "Ground truth" - a phrase you will hear a lot
- Data as a reflection of reality
- Quant envy
- Role in policy-making especially
- "Two cultures" divide humanities and natural sciences

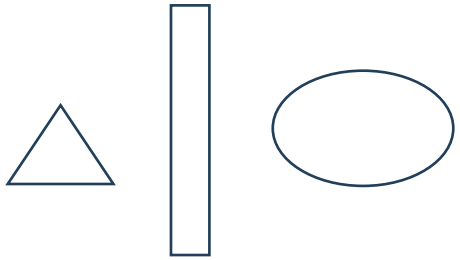
Can data speak?

- "The data speaks for itself."

Quantification: Big Picture

The Phenomenal Universe

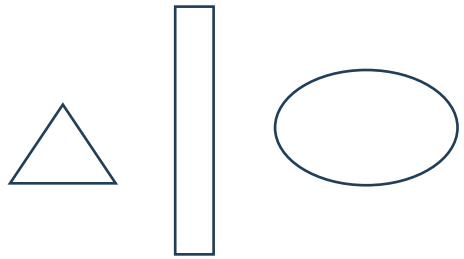




1x triangle, 1x rectangle, 1x oval

Choosing what to measure, abstracting through metrics

What the Computer Sees



1x triangle, 1x rectangle, 1x oval

Ground Truth

- "Raw data" - another phrase you will hear a lot

Ground Truth

- No such thing as "raw data"
- Data is always "cooked"

Conclusions

- As Haïlée from Slurp mentioned last week:
  Analysis and conclusions are always limited.

Conclusions

- As Haïlée from Slurp mentioned last week:
  Analysis and conclusions are always limited.
- Data can play a powerful role in story-telling...
- ...in persuasion
- ...getting you a job

Conclusions

- As Haïlée from Slurp mentioned last week:
  Analysis and conclusions are always limited.
- Data can play a powerful role in story-telling...
- ...in persuasion
- ...getting you a job
- **But don't confuse data with truth!**
- **You are the one who will have a voice, whose voice will be amplified by the power that data gives.**

Questions?

THE UNIVERSITY of EDINBURGH
informatics

THE UNIVERSITY of EDINBURGH
School of Social
& Political Science

THE UNIVERSITY of EDINBURGH
informatics

THE UNIVERSITY of EDINBURGH
School of Social
& Political Science