IC2S2 2021 Tutorial

Understanding Group Membership from Language Use

Elahe Naserian

Miriam Koschate

Tutorial timeline

Introductions	Understanding social groups	**Data selection & preparation**	Training & cross-validation	Domain adaptation	Model validation	Application	**Q&A**
5 min	25 min	15 min	15 min	*30 min*	30 mins	10 min	15 min
	Short break	*Short break*	*Longer break*	*Short break*	*Short break*		

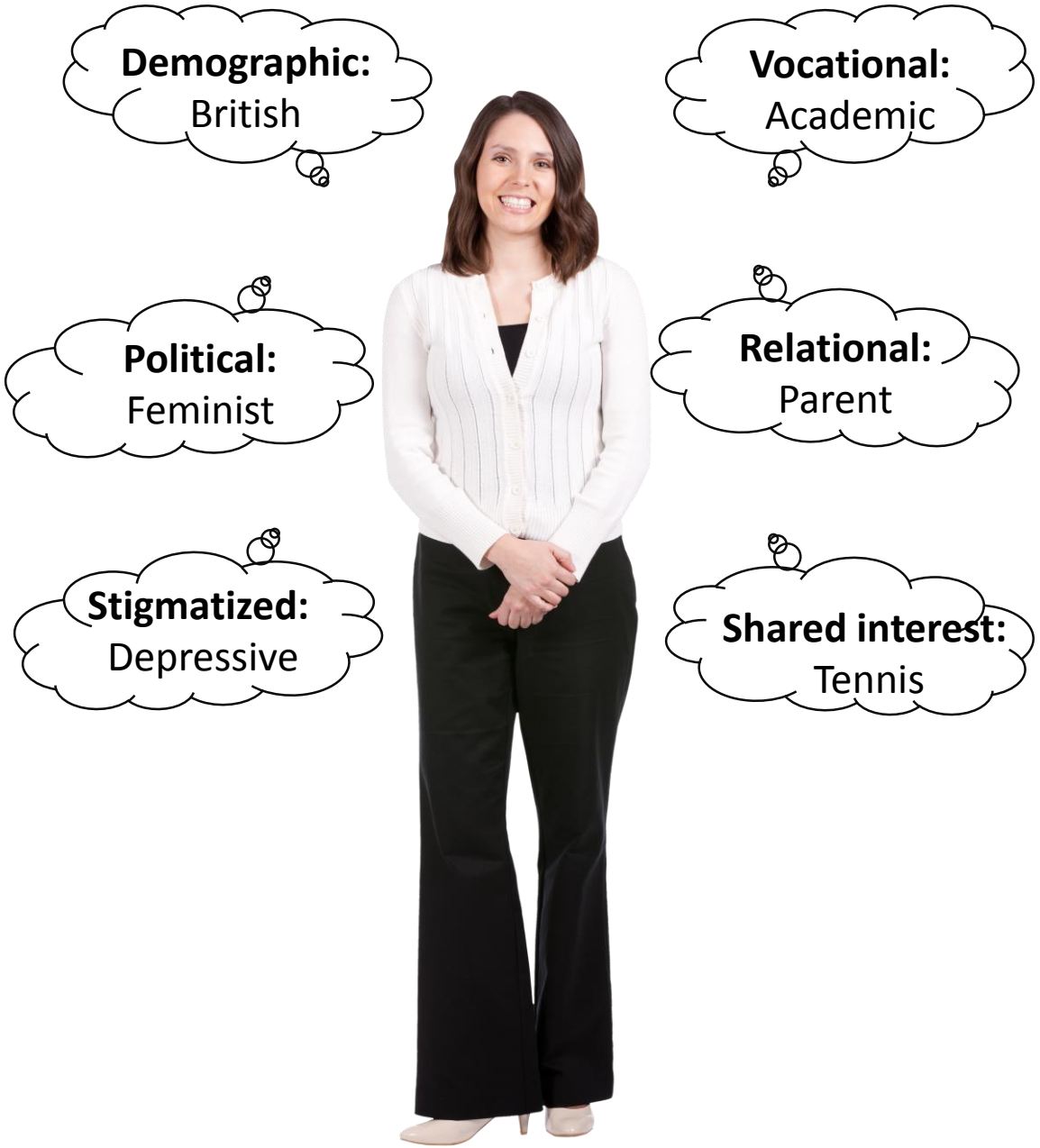
Welcome to our tutorial!

Please briefly introduce yourself:

- Who are you?
- Why are you interested in the linguistic analysis of groups?

The Social Psychology of Groups

- ✓ We are part of many different social categories and groups
- ✓ The social context activates the most relevant identity in our minds



Group Norms and Values

Salient group membership affects:

- **Individual Behaviour**

e.g., Voting (Bryan et al., 2011)

- **Collective Behaviour**

e.g., Crowd behavior (Alnabulsi & Drury, 2014)

- **Perception**

e.g., Olfactory judgements (Coppin et al., 2016)

- **Cognition**

e.g., Selective forgetting (Coman & Hirst, 2015)

- **Attitudes**

e.g., Sexism (Wang & Dovidio, 2017)

Personal Identity Salience

Social Identity Salience

Golden Agers

More proto-typical

Youngsters

More proto-typical

Social Identity Switching

Business(wo)man

Student

Homogeneity within

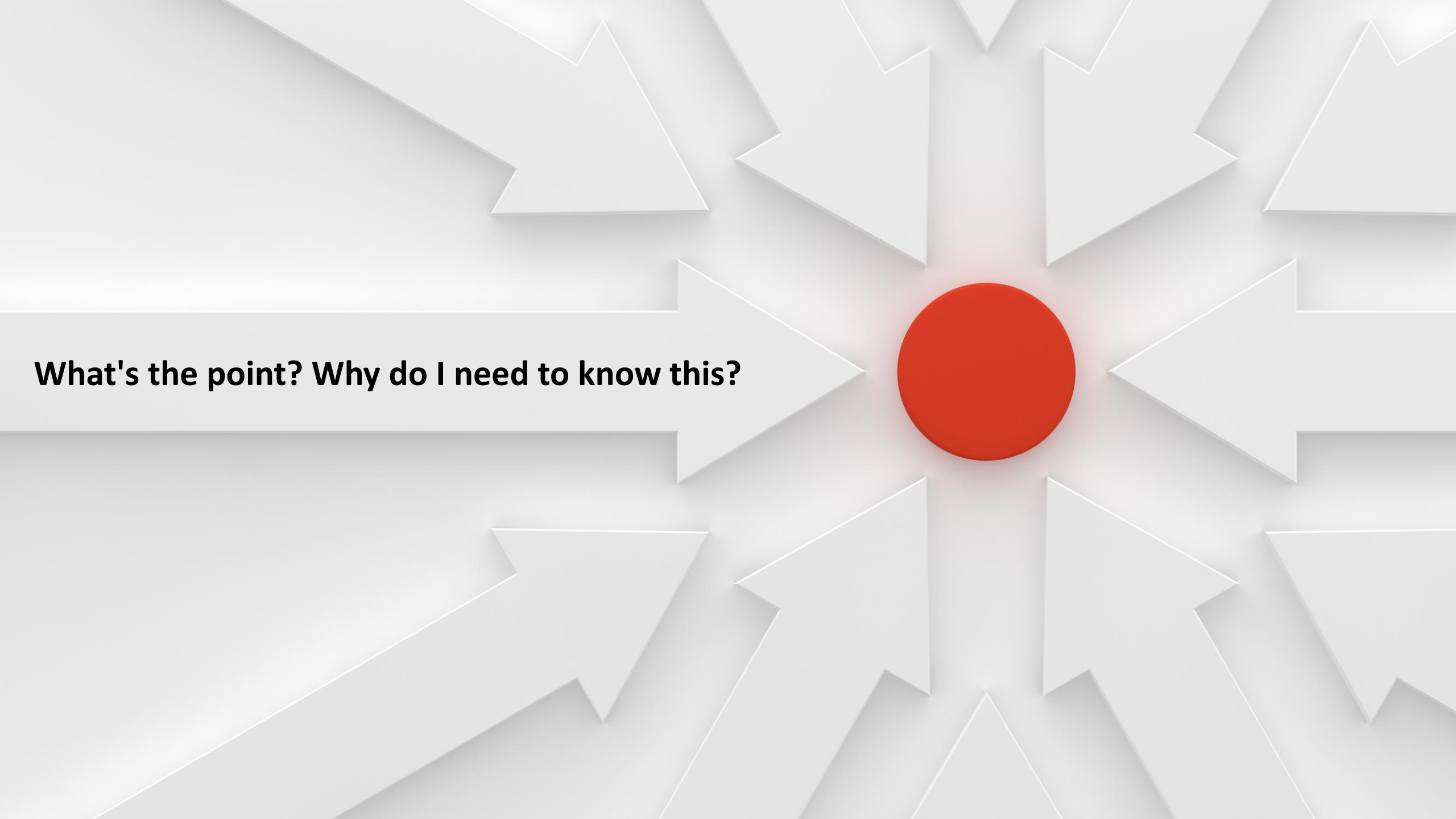
Heterogeneity between

Homogeneity within

Quick Recap

- **Multiple** group memberships
- **Social context activates** relevant group membership
- Active group **informs behaviour** through set of **norms & values**
- We **switch** between groups – and their norms & values
- Some members are more **typical**
- **Homogeneity** within but **differentiation** to other groups

What's the point? Why do I need to know this?



Assessing Group Membership

1. Group norms and values affect behaviour:

Writing = behaviour :-)

Indirect measure that can be used for the analysis of naturally occurring data (e.g., forum posts)

Content and/or style?

- Content can help us to differentiate between groups – BUT:
  - May produce trivial results
  - Highly dependent on topic
- Style can help us to differentiate between groups
  - ✓ Sociolinguistics: Style shifts/code switching
  - ✓ More automatic/less control
  - ✓ Less dependent on topic

Assessing Group Membership

2. Group should be relevant to social context

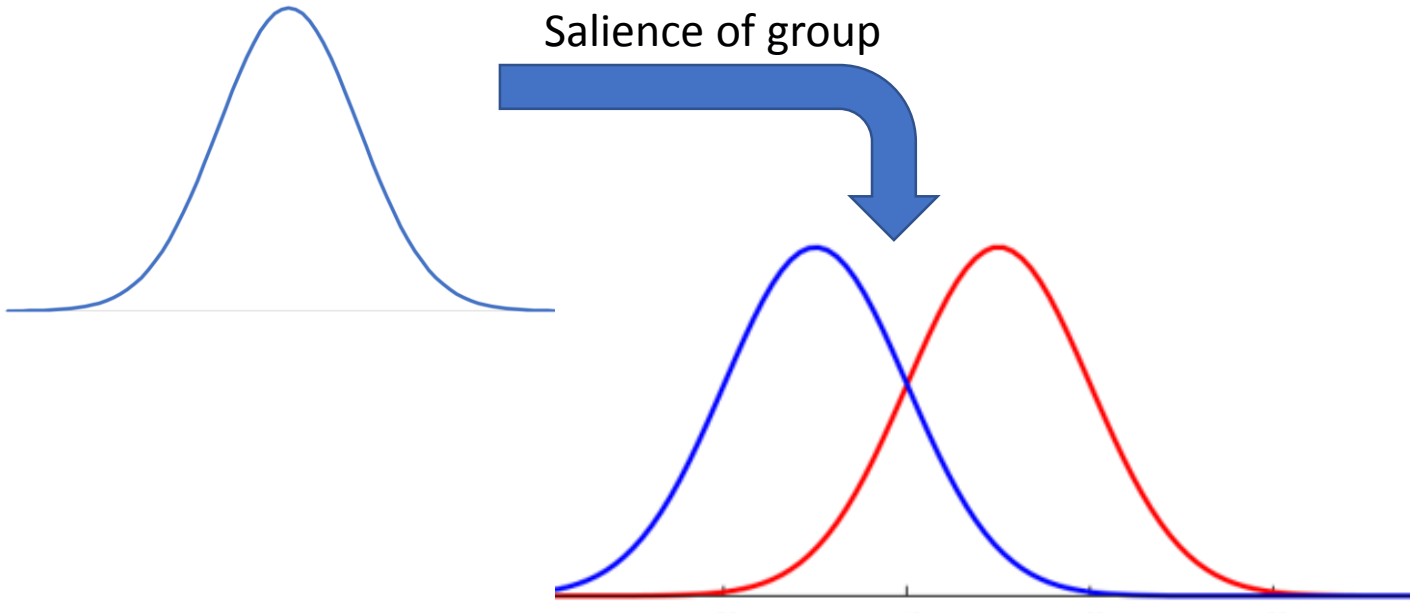
- We cannot assume that groups affect us all of the time/are a stable characteristic:
- Assessing gender, political affiliation, social class, etc. should be much harder "out of context"!

Do businesswomen write like "women"? Or do they switch to a "business person" style?

Assessing Group Membership

3. Homogeneity within, heterogeneity between

- **Ideal for a (binary) classification model!**



Assessing Group Membership

4. Variation due to typicality

We expect that some members write in a more group-typical way than other members

<u>Continuous</u> outcome variable rather than just member of Group A or B (correct classification)

Highly typical member of Group A/ untypical member of B

Both/neither/ unclear

Highly typical member of Group B/ untypical member of A

Mutually Exclusive Classifiers

- Binary gender and age groups (e.g., Rashid et al., 2013)
- Democrats/Republicans (e.g., Pennacchiotti & Popescu, 2011)
- Christians/Atheists (e.g.., Ritter et al., 2014)
- Many more and increasingly sophisticated...

Problems from a social psychological view:

- Assume stable influence on writing (≠ salience)
- We cannot conclude that group membership is the driver of differences:
  - Confounded by demographics, personality, topic, audience
- Cannot assess switches between groups

Overlapping groups

A person can, in principle, be part of both groups of interest (e.g., Doctor and Black)

Advantages:

- **Control** for stable characteristics including demographics and personality
- Assess intra-individual **switches**

Automated Social Identity Assessment (ASIA)

Classification model:

Distinguishes between two overlapping identities based on stylistic features of language

Model: **Logisitic regression** (but other classification models may work just as well – see Cork et al., 2020)

Variables: **44 LIWC style features**

LIWC = Linguistic Inquiry and Word Count 2015 (Pennebaker et al., 2015)

WC **WPS** **Dict**	Word Count Words per Sentence % of words in LIWC	**Compare** **Interrog**	Comparisons (greater, best, after) Interrogatives (how, when, where)
Sixltr	Words longer than 6 characters	**Number** **Quantifiers**	Numbers (second, thousand) Quant (few, many, more)
Ppron	Personal pronouns (I, we, you, she/he, they, itself, them…)	**Posemo** **Negemo**	Positive emotions (love, sweet) Negative emotions (hurt, nasty)
Ipron	Impersonal pronouns (it, it's, those)	**Insight** **Cause**	Insight (think, know) Causation (because, effect)
Article	Articles (the, a, an)	**Tentat** **Certain**	Tentativeness (maybe, perhaps) Certainty (always, never)
Prep	Prepositions (before, of, to, toward, with…)	**Time** **Space**	Time (end, until, Monday) Space (down, in, small)
Auxverb	Auxiliary verbs (be, am, have, do,…)	**Swear** **Filler**	Swear words (damn,…) Filler words (hmm, uh)
Adj **Adverb**	Adjectives (e.g., free, happy, long) Adverbs (e.g., very, really)	**Comma** **SemiC**	Comma (,) Semi-colon (;)

Pennebaker, Boyd, Jordan, & Blackburn (2015). The development and psychometric properties of LIWC2015.

ASIA Validation Pathway

1. Ethical considerations
2. Selection of the training dataset
3. Quantifying stylistic features from text
4. Training the model
5. Cross-validating the model on within-data
6. Cross-platform validation
7. Experimental validation
8. Concurrent validation

Steps 6-8 depend on your research question!

ASIA Validation Pathway

1. **Ethical considerations**
2. Selection of the training dataset
3. Quantifying stylistic features from text
4. Training the model
5. Cross-validating the model on within-data
6. Cross-platform validation
7. Experimental validation
8. Concurrent validation

Steps 6-8 depend on your research question!

Ethical considerations

1) **Consequences of group membership**

- Stigma, Discrimination, Persecution
- Are you **exposing** a person's group membership in a way that may lead to negative consequences?
- Online data often do not have **informed consent** or **awareness** of the research being conducted

-> More (not less) ethical responsibility for the researcher

2) **Realistic expectation of privacy**

- Can you reasonably expect that the participant is aware that their **data is public**? (e.g., well-known issues with Facebook privacy settings)
- Can you ensure that the data is **not traced back** to an individual (irrespective of whether their name is known or just a username), e.g., by Googling?

ASIA Validation Pathway

1. Ethical considerations
2. **Selection of the training dataset**
3. Quantifying stylistic features from text
4. Training the model
5. Cross-validating the model on within-data
6. Cross-platform validation
7. Experimental validation
8. Concurrent validation

Data selection

- **Proof-of-concept case:**

  Parents v Feminists

- **Data for training/validation:**

Mumsnet UK: large parenting platform for mostly middle-class parents; also hosts one of the largest UK feminist forums

-> user ID to tell us who uses both

Reddit: forums (sub-reddits) for almost anything you can think of; international English-speaking platform with majority of US users; hosts parenting and feminist subreddits

ASIA Validation Pathway

1. Ethical considerations
2. Selection of the training dataset
3. **Quantifying stylistic features from text**
4. **Training the model**
5. **Cross-validating the model on within-data**
6. **Cross-platform validation**
7. **Experimental validation**
8. **Concurrent validation**