Responsible Design of Information Access Systems

Asia Biega

MAX PLANCK INSTITUTE
FOR SECURITY AND PRIVACY

This lecture

- Interactive: If you have questions, feel free to interrupt
- People might have different backgrounds, it’s OK to ask if you don’t know!
- We’ll be fluidly switching between computational, social, and organizational perspectives
- Plan:
    - Why IA systems ar sociotechnical (and not purely “technical”)
    - Potential negative impacts of IA systems + sources of harm
    - Mitigation strategies (both computational and non-computational)
    - How to operationalize responsibility concepts
- Source of all photos in this presentation: Unsplash

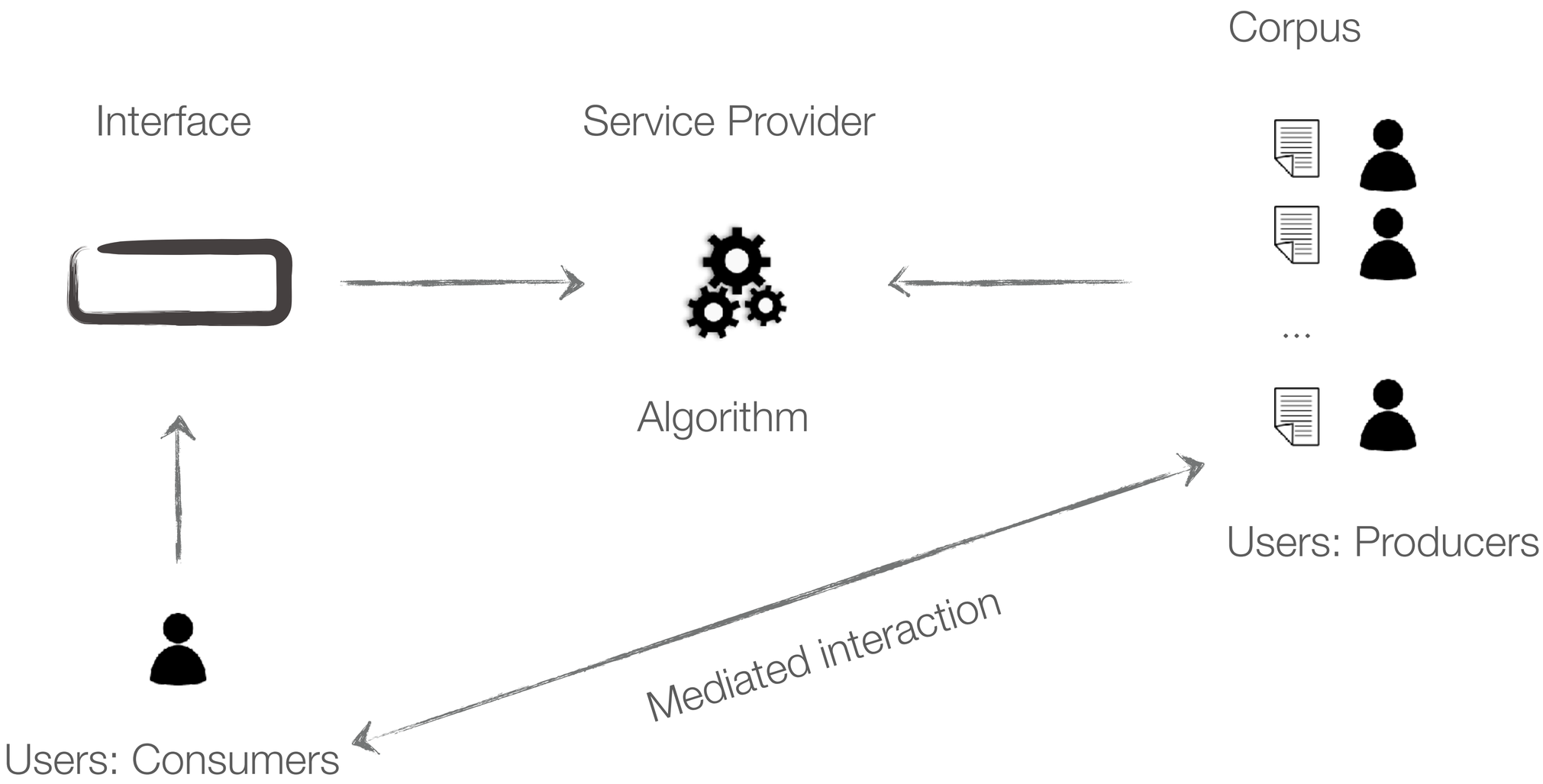
Information Access

Information Access Systems

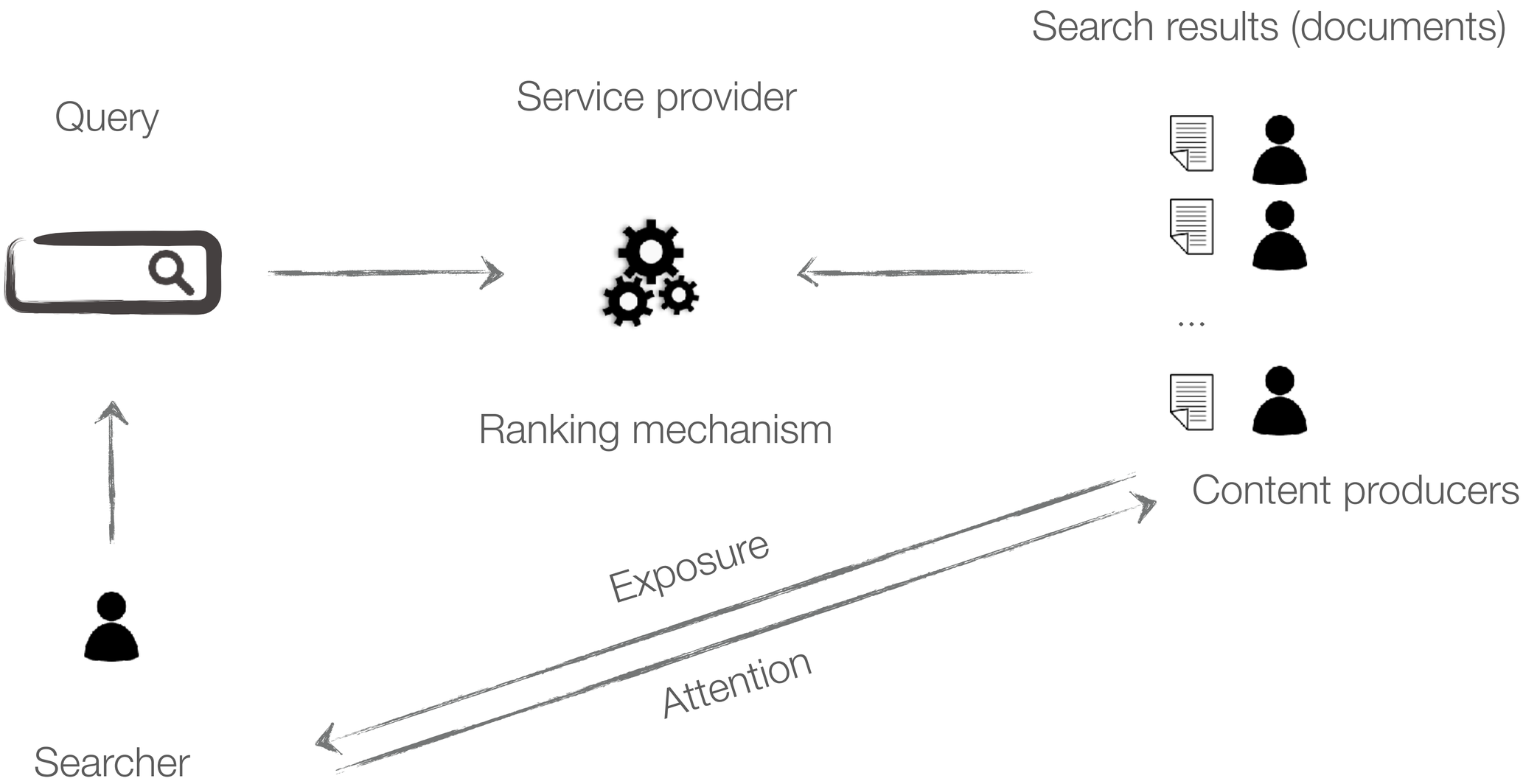
- Web search engines (Google, Bing, …)
- Social media search engines (Twitter, Facebook, …)
- Streaming (Music, Movies, …)
- Product Recommendation (Books, …)
- News recommendations
- Hiring (Recruiter -> JobSeekers, Job Seeker -> Job Openings)
- …

Information access systems: Where are the people?

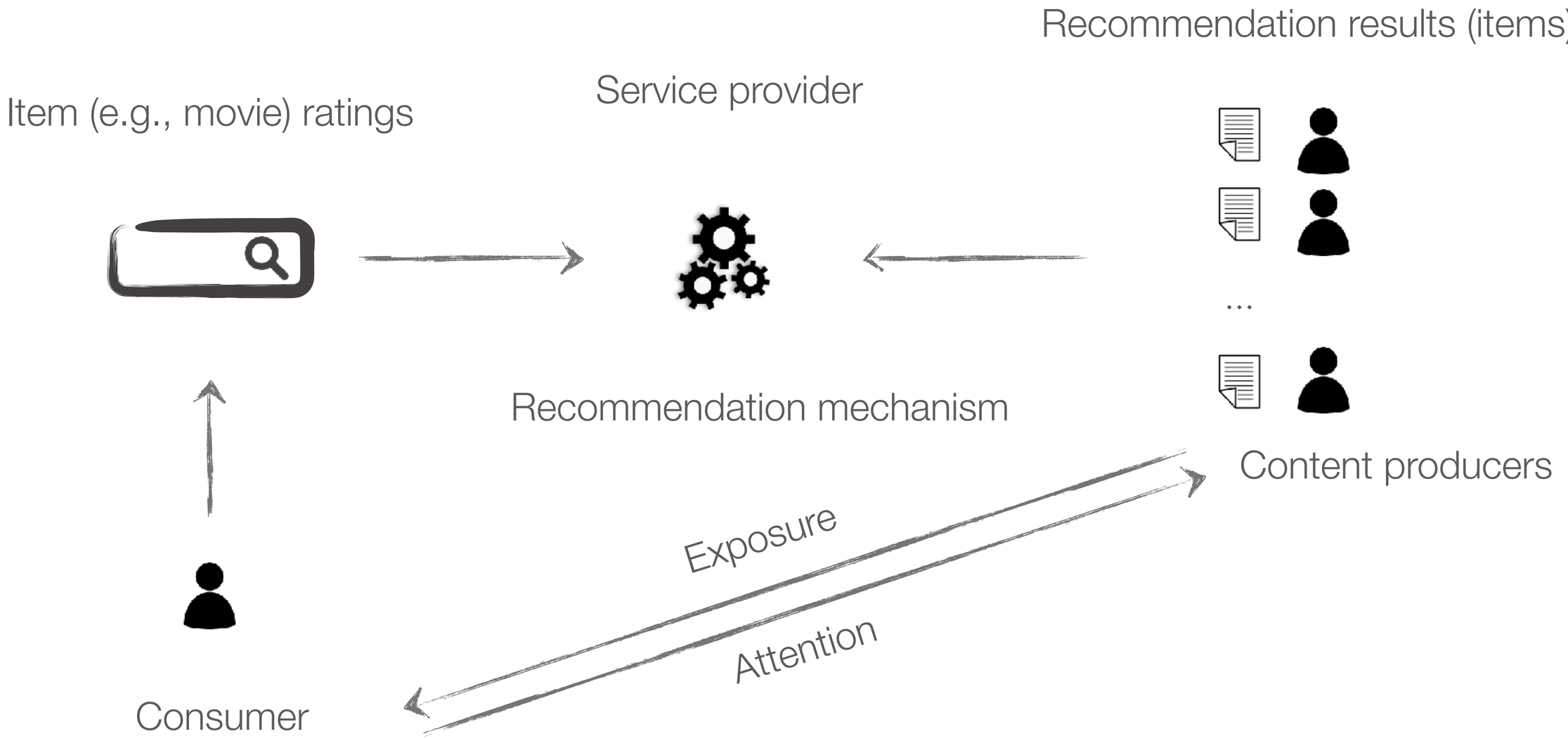
Information Access Systems



Search Systems



Recommender Systems



Information access systems: Where are the people?

- Data generation (ratings, …)
- Interactions with the system
- Query construction (users: consumers)
- Corpus creation (users: producers)
- Evaluation: relevance labels (annotators)
- Algorithm selection, parameter tuning, data cleaning, … (developers)
- …

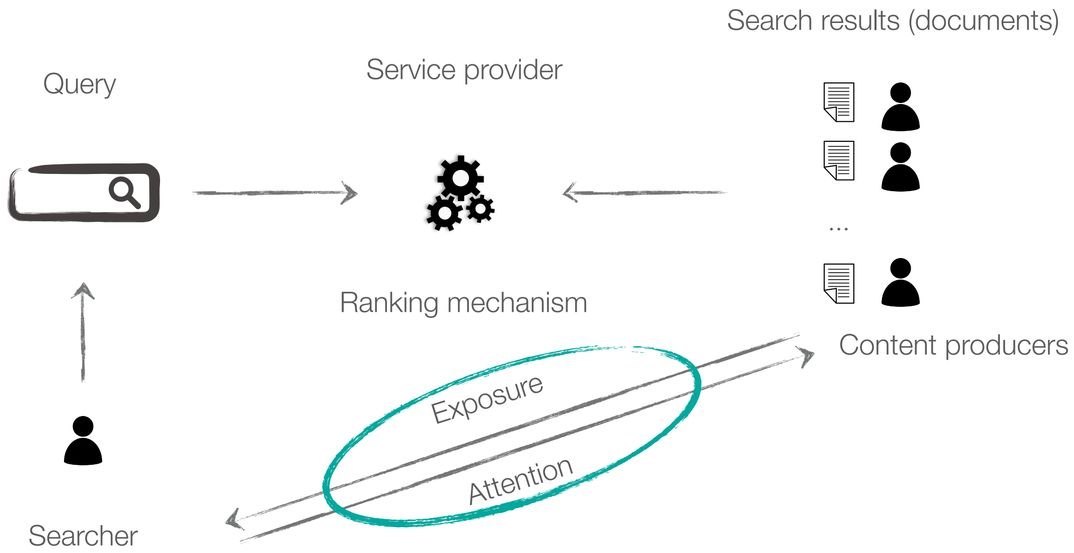
Why IA systems are sociotechnical?

- Used by people: producers + consumers
- Learn from user behaviour and explicit feedback
- Use people's creations as underlying corpora
- Created by people: researchers, engineers, designers, PMs, …
- Impact societal & economic outcomes
- …

Negative Impacts

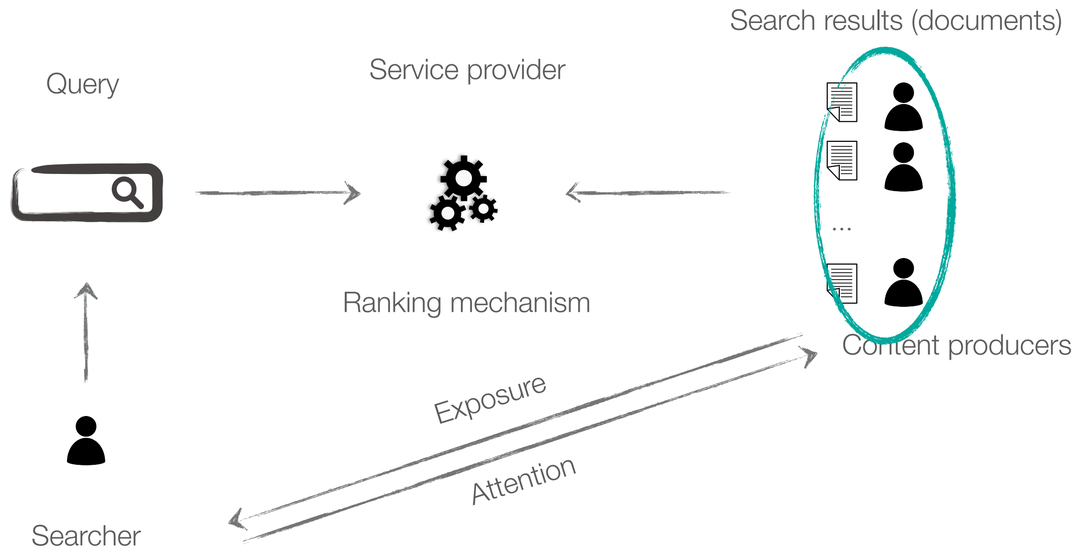
Distributional harms

- **Unfair distribution of a “resource”** to individuals or demographic groups:
  - **Unfair exposure** of producers to consumers in the system —> exposure often translates to resources offline, e.g. jobs, income, etc.
  - **Unfair performance** of a system for consumers



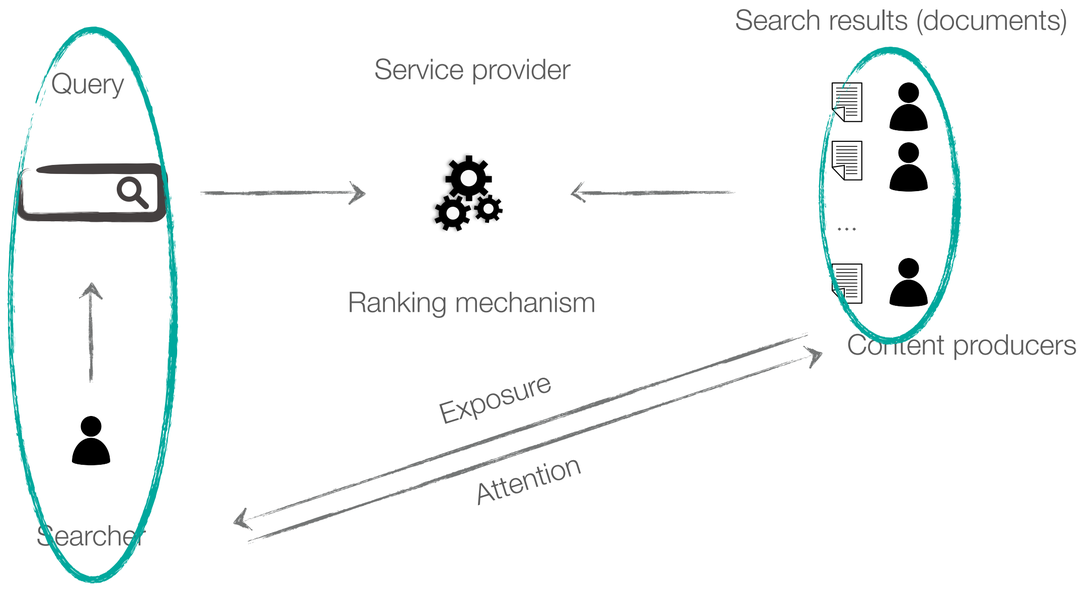
Representational harms

- **Who is represented** in the results?
  - E.g., Underrepresentation of women in results of queries about professions
- **How are groups** represented in the results?
  - The first woman represented in the results of query “CEO” is/was a barbie doll
- Can be an issue in various modalities: images text (how are groups described)



Privacy loss

- Consumers: Increased data collection
  - Revealing many personal details
  - Risk of inference of additional information
- Producers (search): Exposure in sensitive context



Influencing individuals

- Search and recommendation results can impact or manipulate:
  - User worldview and beliefs
  - User behaviour
    - Using search/reco as a source of advice
    - Zero-query search/reco

Influencing societies

- Search and recommendation results can impact:
  - Economy
  - Politics
  - Culture
  - Education (digital divide)
  - Social cohesion (polarization)
  - Values (globalization/westernization)

Sources of harm

Data Biases

- Issues with data:
  - Not representative of the society / user population
  - Encoding historical/societal bias
  - Emergent bias when users change their ways of using a system
  - Sudden data distribution shifts (e.g., Covid suddenly changed data distributions on job platforms)
  - …
- http://www.aolteanu.com/SocialDataLimitsTutorial/
- https://nissenbaum.tech.cornell.edu/papers/biasincomputers.pdf
- R. Baeza-Yates's lecture, "Bias on the Web", CACM 2018

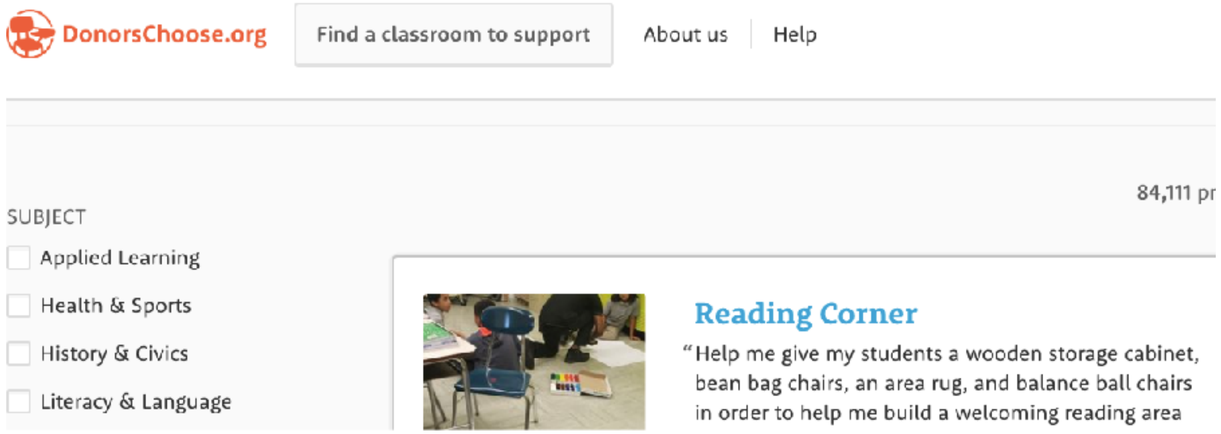
Behavioral Biases

- Cognitive (interaction) biases:
  - Position bias: users won't provide behavioural feedback on items that are ranked lower
  - Automation bias: users assume items are relevant because an automated system must know well
  - …
- Societal biases:
  - Annotators: Gender-biased annotations of job seeker profiles
  - Users: Biased behaviour of users themselves (e.g., women not inspecting high-profile job openings)
- "Cognitive Biases in Search: A review and reflection of cognitive biases in Information Retrieval", Azzopardi, CHIIR 2021
- "Cognitive Biases in Crowdsourcing", Eickhoff, WSDM 2018

Algorithmic/Evaluation Designs

- Objective functions strengthen the impact of certain data points, and erase the impact of others
  - By summing over some errors and not others
  - By squaring errors
  - By taking an absolute value
- Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs, Barocas et al., AIES 2021

UI/UX Designs



- Interfaces impact outcomes on platforms
- Chakraborty et al. "On the Impact of Choice Architectures on Inequality in Online Donation Platforms", WWW 2019

Experimentation

- A/B testing: measuring the effects of a system change on a sample of real platform traffic / user interactions
- System modifications can have tangible consequences for users
- "Exploring or exploiting? Social and ethical implications of autonomous experimentation in AI", Bird et al., FATML 2016
- Who's the Guinea Pig? Investigating Online A/B/n Tests in-the-Wild, FAT* 2019

Feedback Loops

- People influence systems <-> systems influence people
- Over time, negative impacts may be reinforced and amplified
- Over time, the effects of our interventions might be unexpected

Interventions

Responsibility

- Responsibility norms:
  - Social and ethical
  - Legal
    - (GDPR, CCPA, anti-discrimination laws, AI Act, Digital Services Act, …)

EU’s General Data Protection Regulation (GDPR)

1. Lawfulness, **fairness**, **transparency**
2. **Purpose limitation**
3. **Data minimisation**
4. **Accuracy**
5. Storage limitation
6. Integrity and confidentiality

Operationalizing Fairness

(Where do fairness metrics come from?)

What does it mean to operationalize?

Concept
(fairness)

Operationalizations

Concept
(fairness)

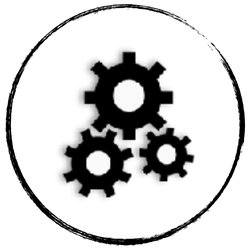
↓

Definition
= goal, what is fair?

↓

Measure
= how far are we from the goal?

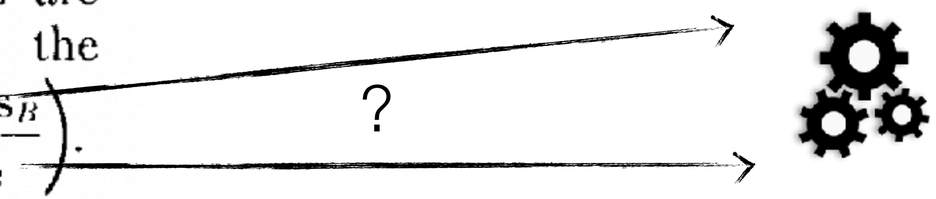
→



Intervention
= how do we achieve the goal?

Getting to a fairness definition

We define an "equitable relationship" to exist when the person scrutinizing the relationship (i.e., the scrutineer—who could be Participant A, Participant B, or an outside observer) perceives that all participants are receiving equal relative outcomes from the relationship $\left(\text{i.e., } \frac{\text{Outcomes}_A}{\text{Inputs}_A} = \frac{\text{Outcomes}_B}{\text{Inputs}_B}\right)$.



- Walster et al., "New directions in equity research", 1973

Getting to a fairness definition: stakeholders

- Identifying the relevant stakeholders:
  - Consumers / searchers
  - Producers / content creators
  - People who are depicted in the results

Getting to a fairness definition: resource

- Distributional fairness: fairly distribute a “resource”
- What are some resources we could consider in IA systems?
  - Consumers: fair quality of service
  - Producers: fair exposure
  - Entities: fair coverage

Getting to a fairness definition: resource

- Consumers -> quality of service:
  - Use established IR/RecSys metrics
- Producers -> exposure:
  - Click probability, inspection time (eye tracking)
  - Use established user browsing models

Proxies

- A “resource” we choose is typically only a proxy:
  - Exposure in an HR platform as a proxy for e.g. getting a job
  - Because the real construct is hard to measure or we don’t have access to the right data
- Important to investigate the reliability and validity of our proxies
  - To what extend are we capturing the intended construct?
  - What are the caveats of our selected proxy?
- Jacobs and Wallach, “Measurement and Fairness”, FAccT 2021

Getting to a fairness definition: goal

- Parity: distribute resource equally
- Proportionality: distribute resource
  - in proportion to their presence in a corpus
  - in proportion to their relevance to a query

Getting to a fairness definition: granularity

- Individual fairness
- Group fairness
  - Groups defined by socially salient groups (gender, ethnicity, …)
  - Groups defined by business needs (e.g., producers of different popularity)

Measuring (un)fairness

- From a fairness goal to a measure:

$$\tilde{\mathbf{P}}_g = \frac{|\pi_{\leq k} \cap \mathcal{D}_g|}{k}$$

$$\Delta_{\text{diff}} = \sum_{g \in \mathcal{G}} (\mathbf{P}_g - \tilde{\mathbf{P}}_g) \qquad \text{Difference}$$

$$\Delta_{\text{abs}} = \sum_{g \in \mathcal{G}} |\mathbf{P}_g - \tilde{\mathbf{P}}_g| \qquad \text{Absolute Difference}$$

$$\Delta_{\text{sq}} = \sum_{g \in \mathcal{G}} \left(\mathbf{P}_g - \tilde{\mathbf{P}}_g\right)^2 \qquad \text{Squared Difference}$$

$$\Delta_{\text{KL}} = \sum_{g \in \mathcal{G}} \mathbf{P}_g \log\left(\frac{\mathbf{P}_g}{\tilde{\mathbf{P}}_g}\right) \qquad \text{KL Divergence}$$

- Kirnap et al., Estimation of Fair Ranking Metrics with Incomplete Judgments, The WebConf 2021

Current limitations of fairness measurement

- Missing/limited group membership annotations
- Calibrated “merit” scores
- Accurate estimation of exposure

Operationalizations

Concept
(fairness)

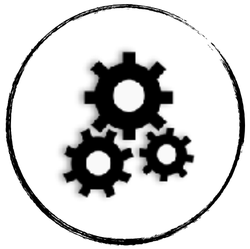
↓

Definition
= goal, what is fair?

↓

Measure
= how far are we from the goal?

→



Intervention
= how do we achieve the goal?

(Algorithmic) Interventions

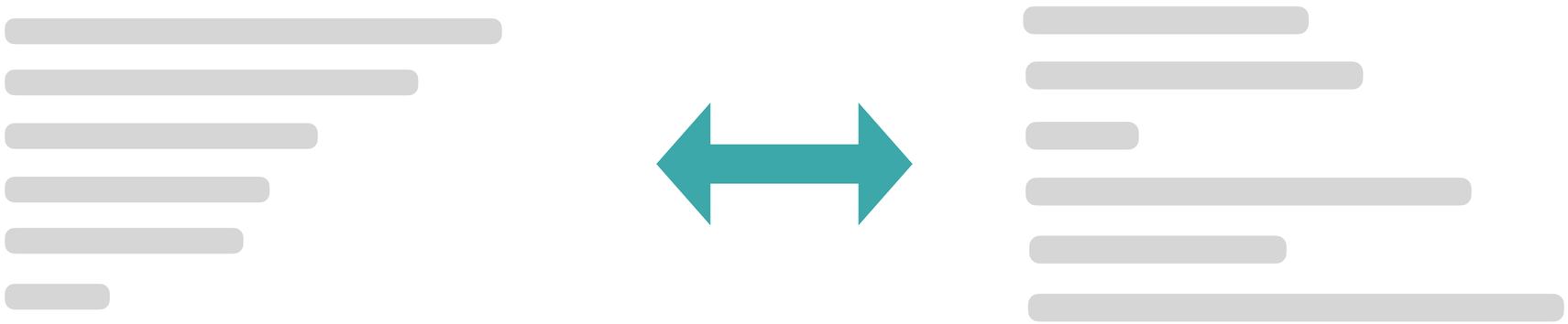
Algorithmic Interventions

- Pre-processing
  - Fix issues (e.g., underrepresentation) in the data
- In-processing
  - Modify the learning objective [Zehlike&Castillo, 2020]:

$$L_{DELTR}\left(y^{(q)}, \hat{y}^{(q)}\right) = L\left(y^{(q)}, \hat{y}^{(q)}\right) + \gamma U\left(\hat{y}^{(q)}\right)$$

- Post-processing
  - Modify a model's output

Post-processing interventions



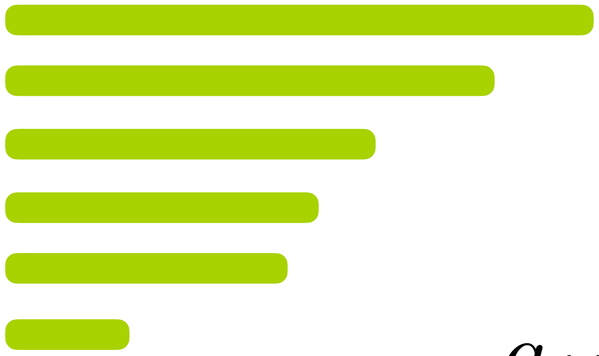
Sequences/distributions of rankings

ρ_1 $\dots$ ρ_m

minimize	$\text{unfairness}(\rho_1^*, ..., \rho_m^*)$
subject to	$\text{utility\_loss}(\rho_j^*, \rho_j) \leq \theta$
over a series of rankings	ρ_j

Post-processing interventions

Relevance



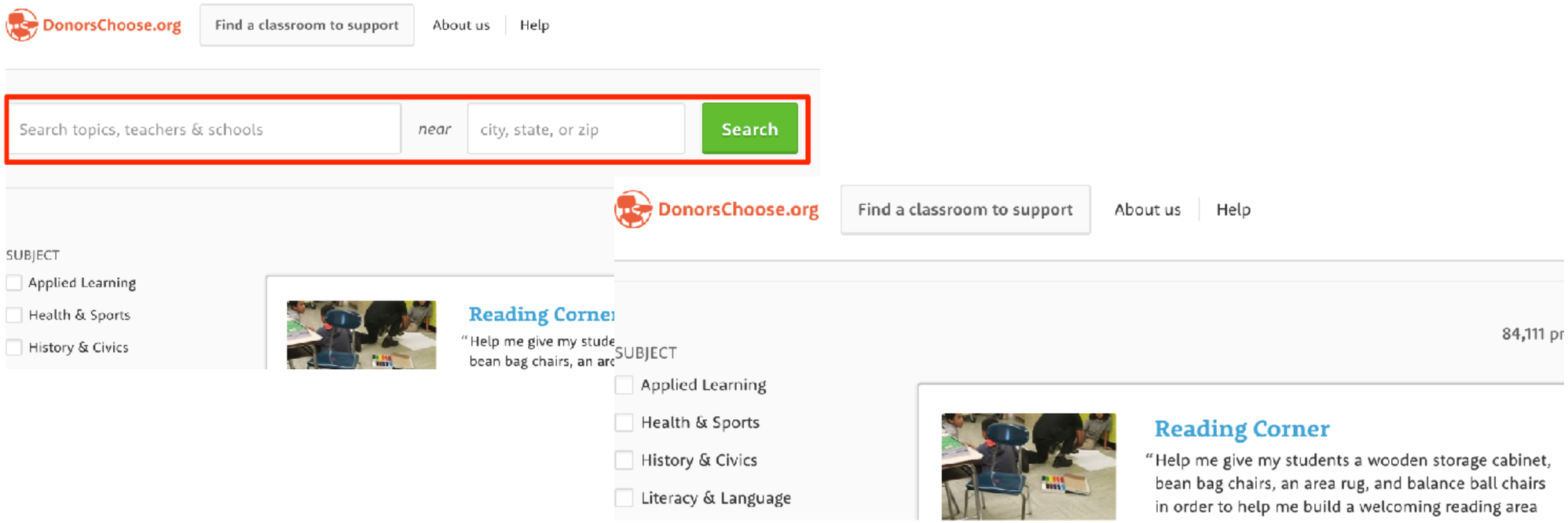
Attention

Position bias

$$\frac{a_{i1}}{r_{i1}} = \frac{a_{i2}}{r_{i2}}, \ \forall u_{i1}, u_{i2}.$$

What's the problem?

Design Interventions



- Nudging users toward equitable outcomes
- Mota et al. "On the Desiderata for Online Altruism: Nudging for Equitable Donations", WWW 2019

Balancing the interests of different stakeholders

- There are many stakeholders in an information access system:
  - Consumers
  - Producers
  - Service providers
  - Advertisers
  - Regulators
  - …
- Stakeholders have different (sometimes conflicting) incentives and goals
- How to balance the interests of different stakeholders in practice?
  - Multi-objective optimization

Operationalizing Data Minimization

DATA MINIMIZATION

Article 5(1)(c) GDPR:
personal data shall be 'adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed'

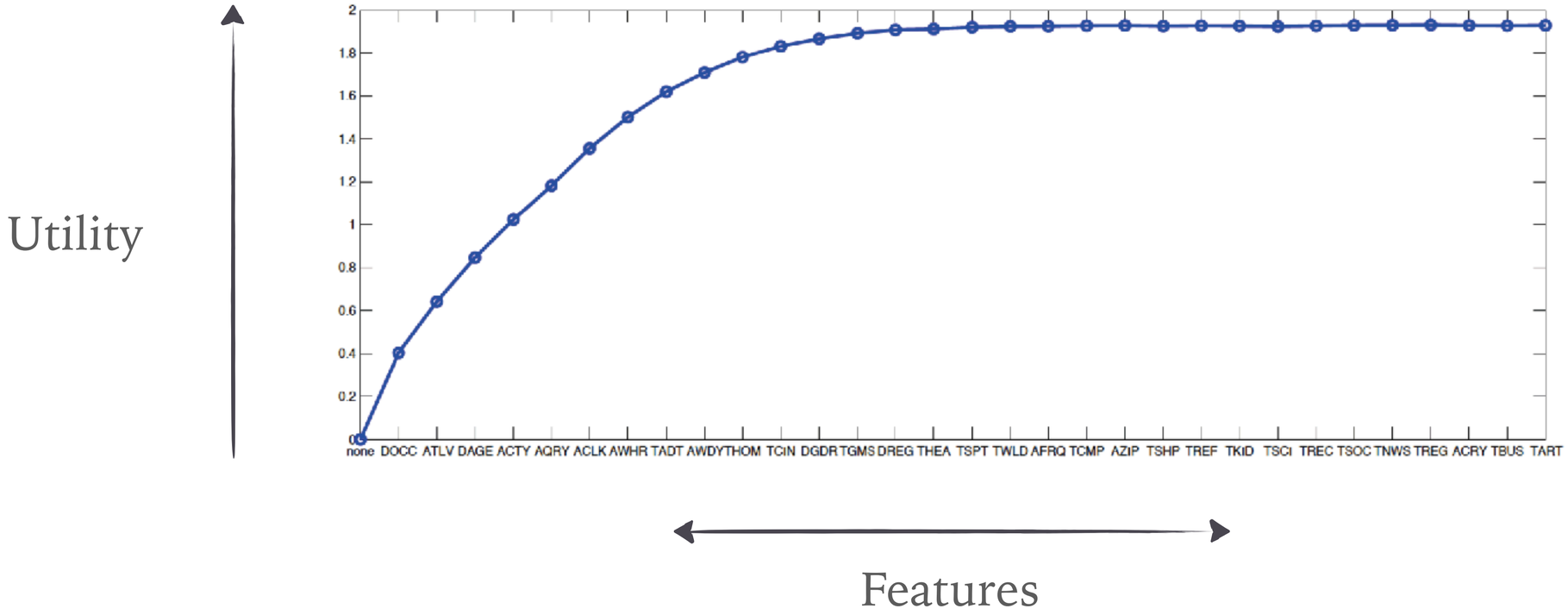
DATA MINIMIZATION: WHY?

- Avoiding the unnecessary risks of personal data processing.
- Preventing inference of personal information.
- Increased user control over data, agency, understanding.
- Limited surveillance, targeting.
- …

DATA MINIMIZATION: WHY NOT?

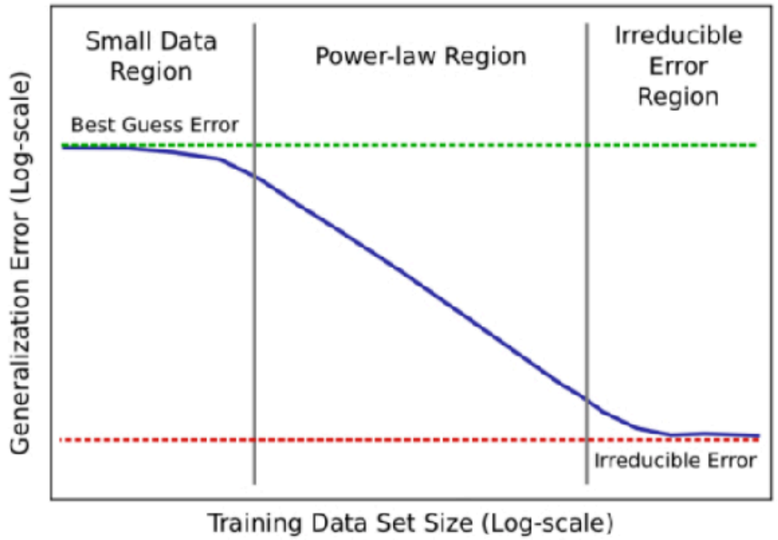
- Curbing innovation.
- Limiting societally desirable tech.
- Limiting competitiveness of the European economy.
- Just impossible in the world of Big Data.
- …

“JUST IMPOSSIBLE IN THE WORLD OF BIG DATA”



Krause and Horvitz, “A Utility-Theoretic Approach to Privacy and Personalization”, AAAI, 2008

“JUST IMPOSSIBLE IN THE WORLD OF BIG DATA”



Hestness et al. “Deep Learning Scaling is Predictable, Empirically”, 2017

“JUST IMPOSSIBLE IN THE WORLD OF BIG DATA”

Existing computational techniques:

- Feature selection
- Outlier detection
- Active learning
- …

OBSTACLES TO ADOPTION

- Scarcity of guidelines
  - “Understanding Software Developers' Approach towards Implementing Data Minimization”, Senarath and Arachchilage, SOUPS 2018

- Lack of operational definitions for *personalization, profiling, and decision-making systems*
  - UK, Norwegian data protection authorities’ guidelines:
    - Don’t talk about data-driven systems
    - Only mention techniques without detailed guidelines

DATA MINIMIZATION

Article 5(1)(c) GDPR:
personal data shall be 'adequate, relevant and **limited to what is necessary in relation to the purposes for which they are processed**'
(data minimisation)

OPERATIONALIZING PURPOSE LIMITATION AND DATA MINIMIZATION

In personalization, profiling, decision-making systems systems:

1. What is the purpose of personal data collection?
2. What does it mean to limit data in relation to the purpose?
3. What are the trade-offs and consequences of data minimization according to those interpretations?

Biega, Potash, Daumé III, Diaz, Finck, "Operationalizing the legal principle of data minimization for personalization", SIGIR 2020

INTERPRETING PURPOSE

1. What is the purpose of personal data collection?

Observation:

In personalization, personal data is collected not necessarily to deliver, but to improve the service.

Proposal:

Tie the processing purpose to improvements performance metrics.

INTERPRETING LIMITATION

2. What does it mean to limit data in relation to the purpose?

[Biega et al., SIGIR 2020]

Global data minimization:
limit subject to a threshold on the **loss in the global model performance**:

$$\min \quad k \quad \text{s.t.} \qquad \forall u, |\tilde{\mathcal{I}}_u| = k \quad \text{and} \qquad \mathbb{E}_{\mathcal{U}}[\sigma(\tilde{r}'_u)] - \mathbb{E}_{\mathcal{U}}[\sigma(\tilde{r}_u)] \leq \lambda$$

Per-user data minimization:
limit subject to a threshold on the **loss in the per-user model performance**:

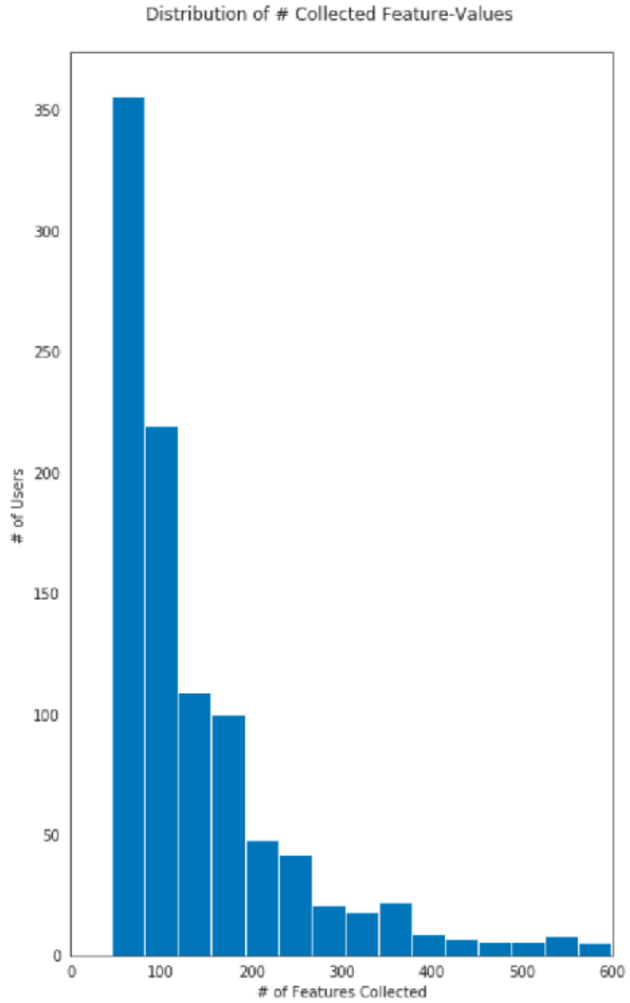
$$\min \quad k \quad \text{s.t.} \qquad \forall u, |\tilde{\mathcal{I}}_u| = k \quad \text{and} \qquad \forall u, \sigma(\tilde{r}'_u) - \sigma(\tilde{r}_u) \leq \lambda$$

[Shanmugam et al., FAccT 2022]

Performance-curve-slope data minimization:
stop collecting once the derivative of the perfomance curve drops below a threshold:

$$\frac{d\sigma_M}{dn}(|\mathcal{A}|) \leq t$$

IMPACT ON USERS WHEN MINIMIZING USING ACTIVE LEARNING



Shanmugam et al. "Learning to Limit Data Collection via Scaling Laws: A Computational Interpretation for the Legal Principle of Data Minimization", FAccT 2022

DATA MINIMIZATION AND FAIRNESS

- No hierarchy of principles in Art. 5 GDPR:
  - Data minimisation is not superior to fairness or accuracy, and vice-versa.

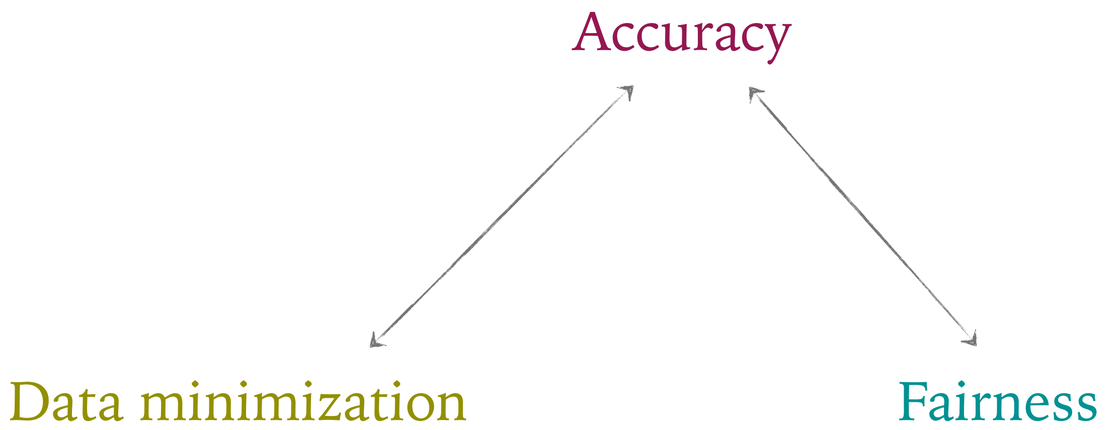
Accuracy

Data minimization

Fairness

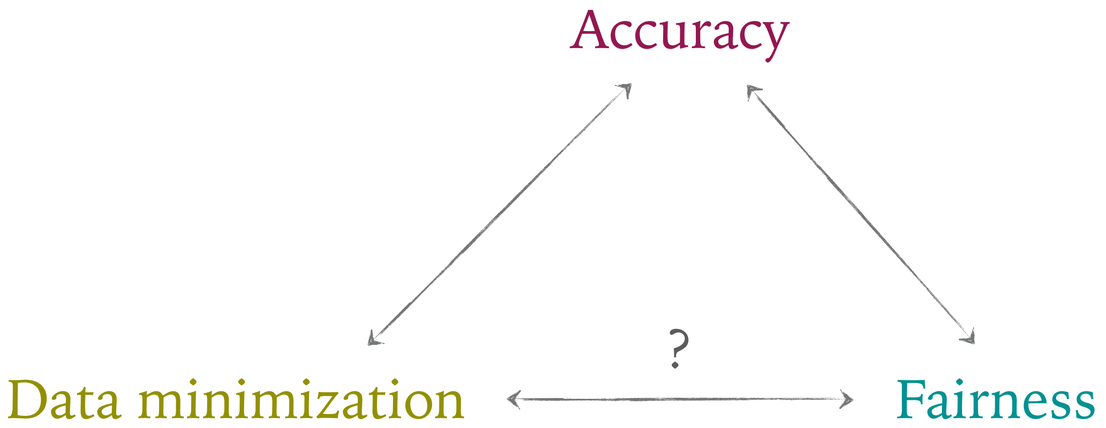
DATA MINIMIZATION AND FAIRNESS

- No hierarchy of principles in Art. 5 GDPR:
  - Data minimisation is not superior to fairness or accuracy, and vice-versa.



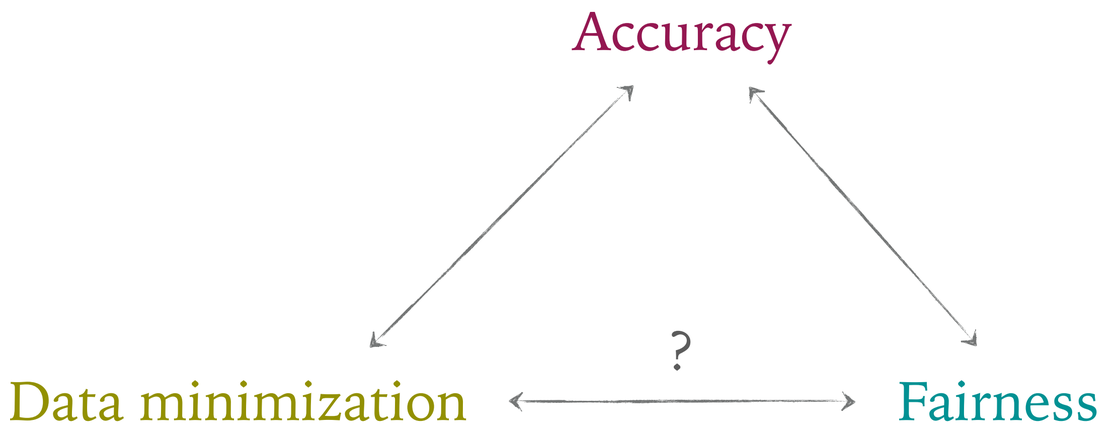
DATA MINIMIZATION AND FAIRNESS

- No hierarchy of principles in Art. 5 GDPR:
  - Data minimisation is not superior to fairness or accuracy, and vice-versa.



DATA MINIMIZATION AND FAIRNESS

- No hierarchy of principles in Art. 5 GDPR:
  - Data minimisation is not superior to fairness or accuracy, and vice-versa.



Quality-of-service impact on individuals and groups [Biega et al. SIGIR 2020]

Data collection impact on individuals and groups [Shanmugam et al. FAccT 2022]

Minimization as an obstacle to fairness audits [Clavell et al. AIES 2002]

…

Refusal

- When the Implication Is Not to Design (Technology), Baumer & Silberman, CHI 2007:
  - “Could the technology be replaced by an equally viable low-tech or non-technological approach to the situation?”
  - “Does a technological intervention result in more trouble or harm than the situation it’s meant to address?”
  - “Does a technology solve a computationally tractable transformation of a problem rather than the problem itself?”

Documentation

- Questions:
  - How is data: collected, sampled, generated
  - How was a model trained, where does it fail
- Datasheets for Datasets, Gebru et al.
- Model Cards for Model Reporting, Mitchell et al.
- “Nutrition labels” for models and data

Organizational Interventions

- Before any technical work happens: a team models a problem
  - How to think through that problem?
  - How to anticipate negative impacts?
- Structured frameworks/checklists for responsible AI in the industry

Research in Responsible IA Systems

- Related work distributed across multiple communities:
  - Information Retrieval, Computing&Society, Data Mining, Databases, HCI/ Social Computing, Machine Learning
- Data: Fabris et al. “Algorithmic Fairness Datasets: the Story so Far”, 2022
- TREC Fair Ranking (Semantic Scholar: scholarly search, Wikipedia: prioritizing articles for editing)
- Interdisciplinarity: Talk to people from other fields

Other useful resources

- Fairness and Discrimination in Information Access Systems, Ekstrand et al.
- Fairness in Ranking: A Survey, Zehlike et al.
- Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries, Olteanu et al.
- Fairness and Machine Learning: Limitations and Opportunities, Barocas et al.
- Law for Computer Scientists and Other Folk, Hildebrandt

Other useful resources

- Fairness and Discrimination in Information Access Systems, Ekstrand et al.
- Fairness in Ranking: A Survey, Zehlike et al.
- Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries, Olteanu et al.
- Fairness and Machine Learning: Limitations and Opportunities, Barocas et al.
- Law for Computer Scientists and Other Folk, Hildebrandt

Responsible Design of Information Access Systems

Asia J. Biega, asia.biega@mpi-sp.org

MAX PLANCK INSTITUTE
FOR SECURITY AND PRIVACY

- Papers referenced during the Q&A:
  - https://dl.acm.org/doi/abs/10.1145/3461702.3462602 / https://arxiv.org/abs/2012.00423
  - https://ojs.aaai.org/index.php/HCOMP/article/view/5281