Utrecht University

UMAP 2021 Tutorial

Multi-Method Evaluation for Adaptive Systems

Christine Bauer

Utrecht University, The Netherlands
Department of Information and Computing Sciences
Division Interaction | Human-Centered Computing Group

✉ c.bauer@uu.nl
https://christinebauer.eu
@christine_bauer

Utrecht University

Christine Bauer

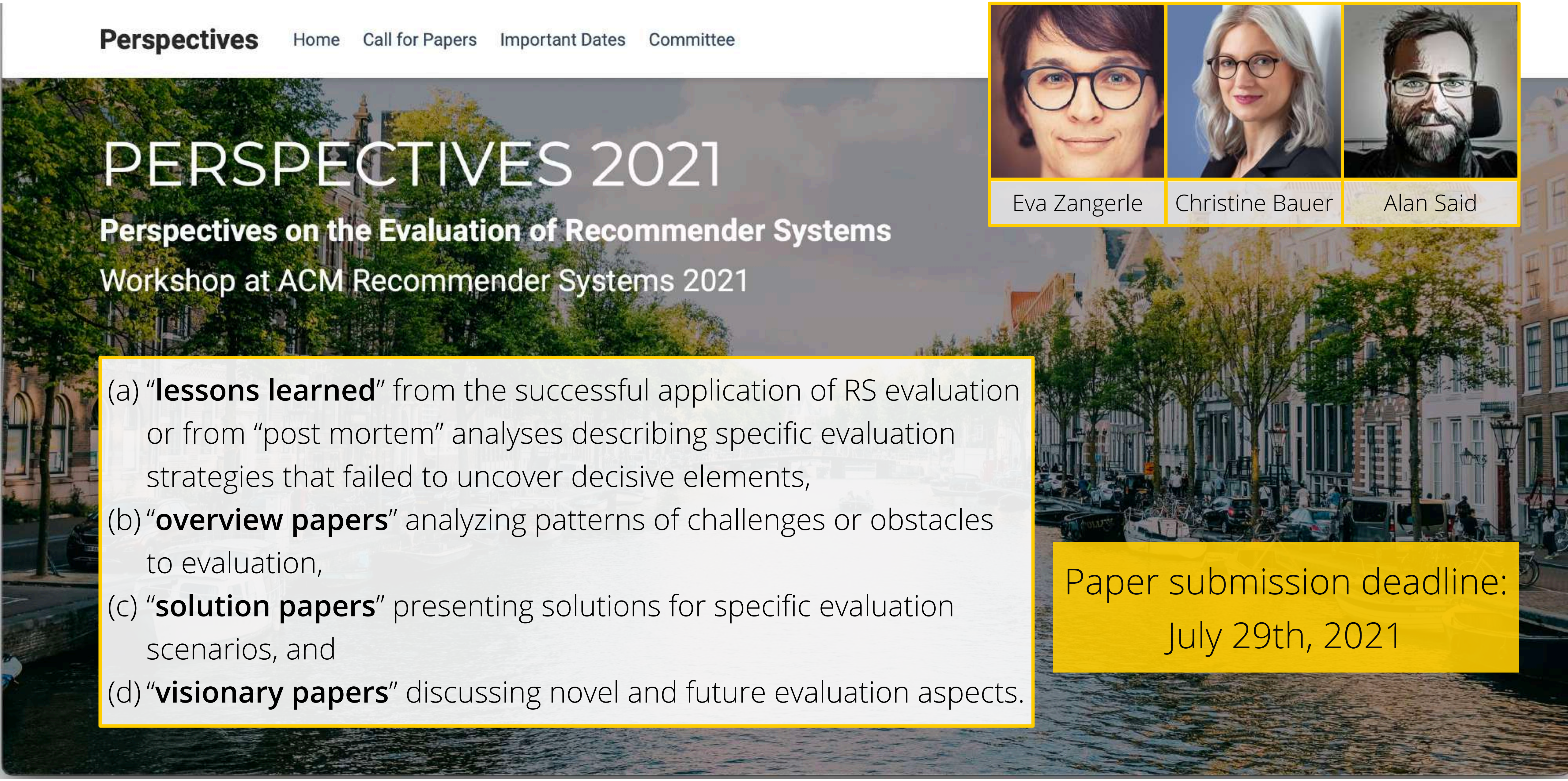
Assistant Professor

Utrecht University, The Netherlands
Department of Information and Computing Sciences
Division Interaction | Human-Centered Computing Group

Interactive intelligent systems; context-adaptivity

Context-aware recommender systems
Music sector

✉ c.bauer@uu.nl
https://christinebauer.eu
@christine_bauer



PERSPECTIVES 2021

Perspectives on the Evaluation of Recommender Systems

Workshop at ACM Recommender Systems 2021

(a) "**lessons learned**" from the successful application of RS evaluation or from "post mortem" analyses describing specific evaluation strategies that failed to uncover decisive elements,

(b) "**overview papers**" analyzing patterns of challenges or obstacles to evaluation,

(c) "**solution papers**" presenting solutions for specific evaluation scenarios, and

(d) "**visionary papers**" discussing novel and future evaluation aspects.

Paper submission deadline:
July 29th, 2021

https://perspectives-ws.github.io/2021/

https://multimethods.info

maintained in collaboration with
Eva Zangerle

Learning objectives

- participants are aware of and familiar with the wide spectrum of opportunities how an adaptive or personalized system may be evaluated
- participants are able to come up with evaluation designs that comply with the four basic options of multi-methods evaluation

- stimulate critical reflection of one's on evaluation practices and those of the community at large

Agenda

- Overview: Potentially relevant evaluation goals, perspectives, properties,…
- The tradition of evaluation approaches
- Blind spots
- Introduction to multi-method evaluation
- Overview: 4 basic options of integrating multiple methods
- Wrap up of first part of the tutorial
- Group work
  - Multi-method design in break-out rooms
  - Discussion of elaborations in plenum
  - Presentation of potential solutions
- Challenges of multi-method evaluation
- Where do we go from here?—Discussion
- Summing up and take away

What aspects define whether a personalized/adaptive system is "good"?

Potentially relevant evaluation goals, perspectives, properties,...

What exactly do I want/need to find out?

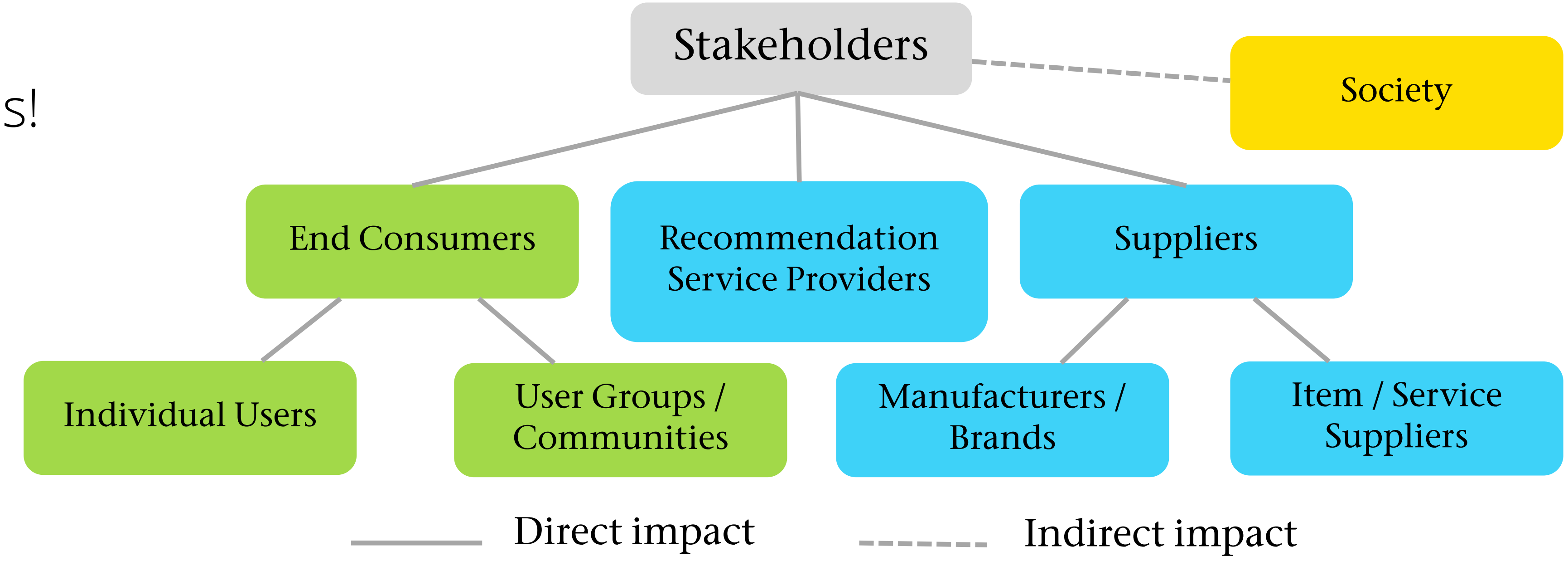
- *Is it relevant?*
- *Does it matter in practice?*

Stakeholders

- What is my target group?
- Who else is affected/involved?

- Also consider sub-groups!

Example for recommender systems



Stakeholders

Society

End Consumers

Recommendation Service Providers

Suppliers

Individual Users

User Groups / Communities

Manufacturers / Brands

Item / Service Suppliers

Direct impact

Indirect impact

Dietmar Jannach & Christine Bauer (2020). Escaping the McNamara Fallacy: Toward More Impactful Recommender Systems Research. AI Magazine, 41(4), pp 79-95. DOI: 10.1609/aimag.v41i4.5312

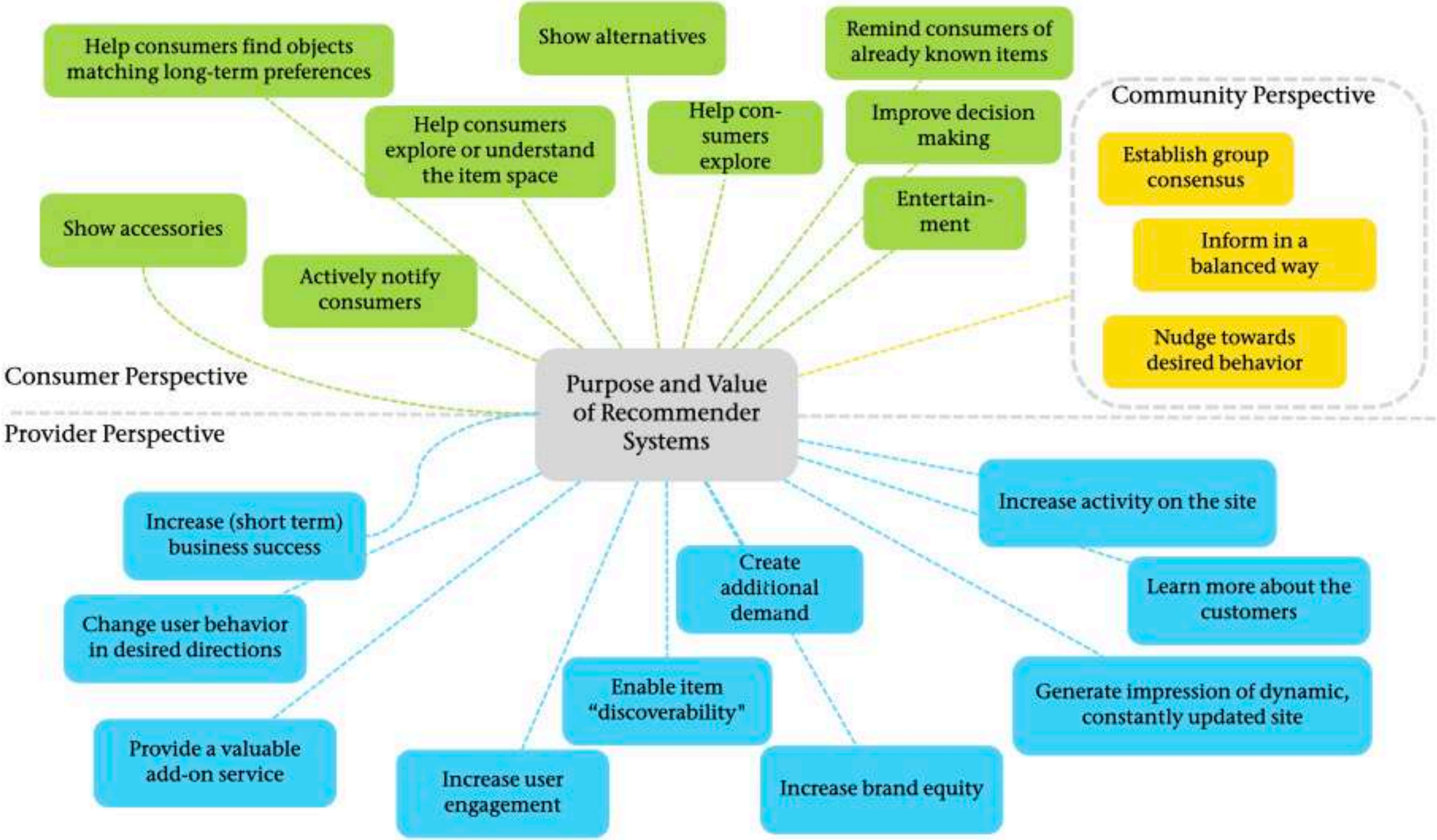
Task, intent, goal, need

- What is the user’s task?
- What is the user’s intent?
- What does the user want?
- What does the user need?
- Are multiple tasks, intents, demands, needs?

- What is the providers goal, need or intent?
- Do these overlap with the users’ perspective? Do they contradict?

- Is this fair?
- Is this desirable?
- Who says that?

Purpose and value of recommenders


Purpose and Value of Recommender Systems
Consumer Perspective
Help consumers find objects matching long-term preferences
Show alternatives
Remind consumers of already known items
Help consumers explore or understand the item space
Help consumers explore
Improve decision making
Entertainment
Show accessories
Actively notify consumers
Community Perspective
Establish group consensus
Inform in a balanced way
Nudge towards desired behavior
Provider Perspective
Increase (short term) business success
Change user behavior in desired directions
Provide a valuable add-on service
Increase user engagement
Enable item "discoverability"
Create additional demand
Increase brand equity
Increase activity on the site
Learn more about the customers
Generate impression of dynamic, constantly updated site

Dietmar Jannach & Christine Bauer (2020). Escaping the McNamara Fallacy: Toward More Impactful Recommender Systems Research. AI Magazine, 41(4), pp 79-95. DOI: 10.1609/aimag.v41i4.5312

Dietmar Jannach & Gediminas Adomavicius (2016). Recommendations with a Purpose. In Proceedings of the 10th ACM Conference on Recommender Systems, RecSys 2016, 7–10. New York: ACM. DOI: 10.1145/2959100.2959186

Variables of interest, their conceptualization, their measurement

- e.g., increase in sales, feeling good, time spent on platform, balanced usage

- Why are these variables interesting?
- Who says/defines that?

Applicable metrics

- Predication accuracy, accuracy again?, again accuracy??
- How to measure "satisfaction"?
- Does it matter in practice?

Feasibility

- Access to skills for the method

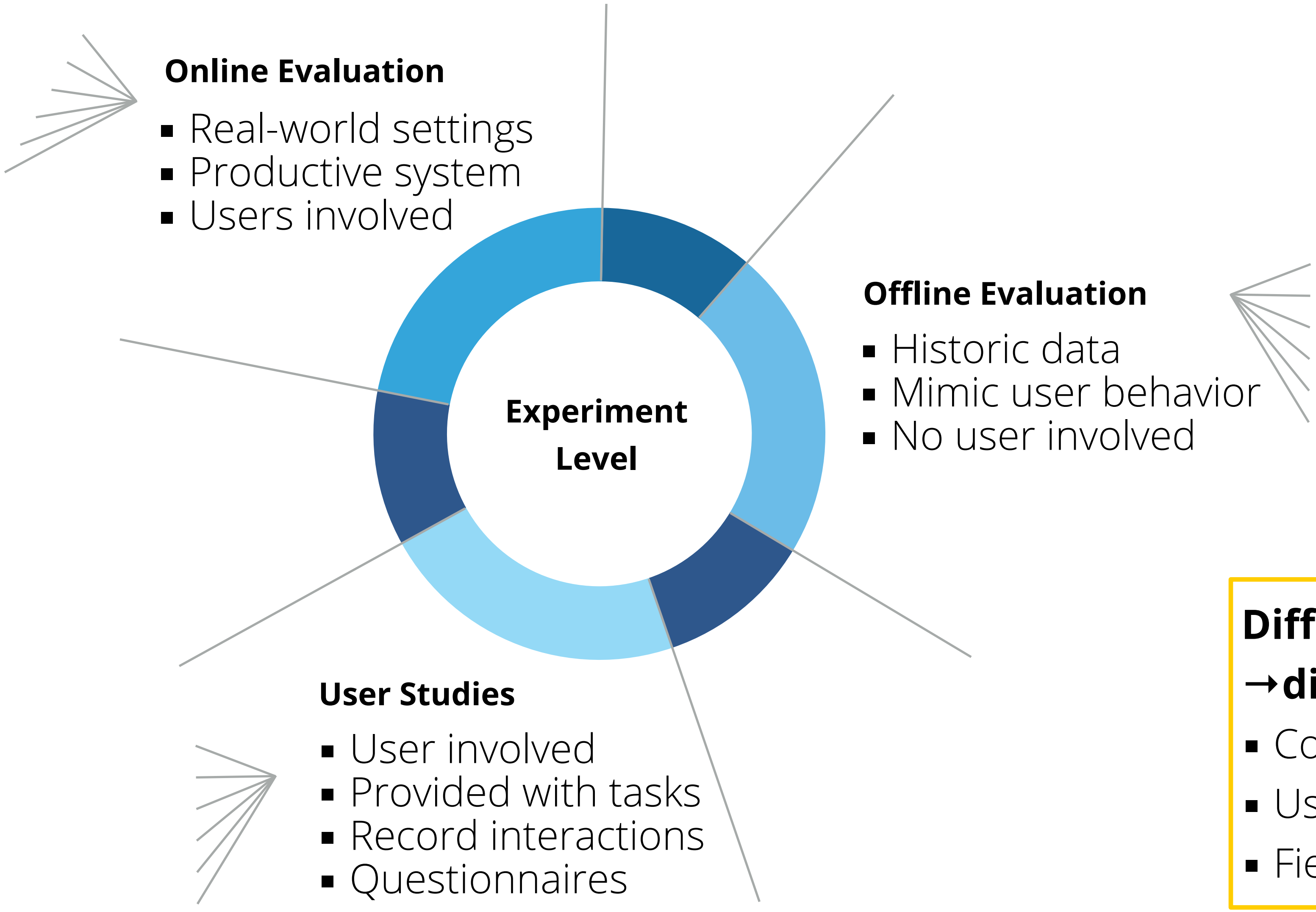
  ➡ no skills yet is no excuse for doing a bad evaluation

- Access to resources

  ➡ limited resources are no excuse for doing a bad evaluation

The tradition of evaluation approaches

Tradition of evaluation approaches



Online Evaluation

- Real-world settings
- Productive system
- Users involved

Offline Evaluation

- Historic data
- Mimic user behavior
- No user involved

User Studies

- User involved
- Provided with tasks
- Record interactions
- Questionnaires

Different (sub-)communities
→different terminology

- Computational or algorithmic approaches
- User studies (in the lab or online)
- Field studies (using a real-world system)

Blind spots

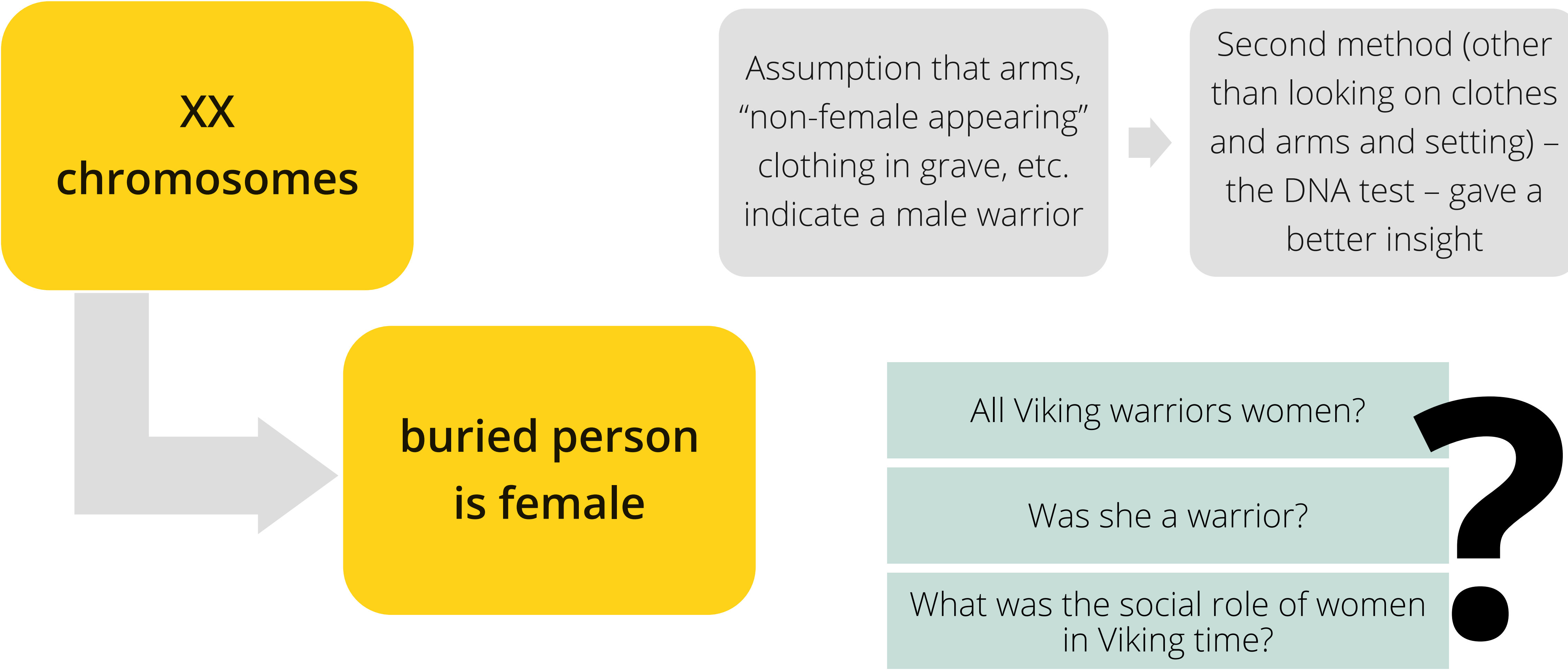
In 1878 in Birka (Southeastern Sweden), unburied Viking settlement from about 750 to 950

High-status,
Viking warrior,
male.

Weapons found in the grave suggest the occupant was a high-status warrior.
(Image credit: Neil Price, Charlotte Hedenstierna-Jonson, Torun Zachrisso, Anna Kjellström;
Copyright : Antiquity Publications Ltd.)

Illustration how the burial might have looked just before it was closed in Viking times.
(Image credit: Drawing by Þórhallur Þráinsson; Copyright Antiquity Publications Ltd.)

2017, DNA test



XX chromosomes

buried person is female

Assumption that arms, “non-female appearing” clothing in grave, etc. indicate a male warrior

Second method (other than looking on clothes and arms and setting) – the DNA test – gave a better insight

All Viking warriors women?

Was she a warrior?

What was the social role of women in Viking time?

Seminal example of choice overload

less attractive

30% sales

more attractive

3% sales

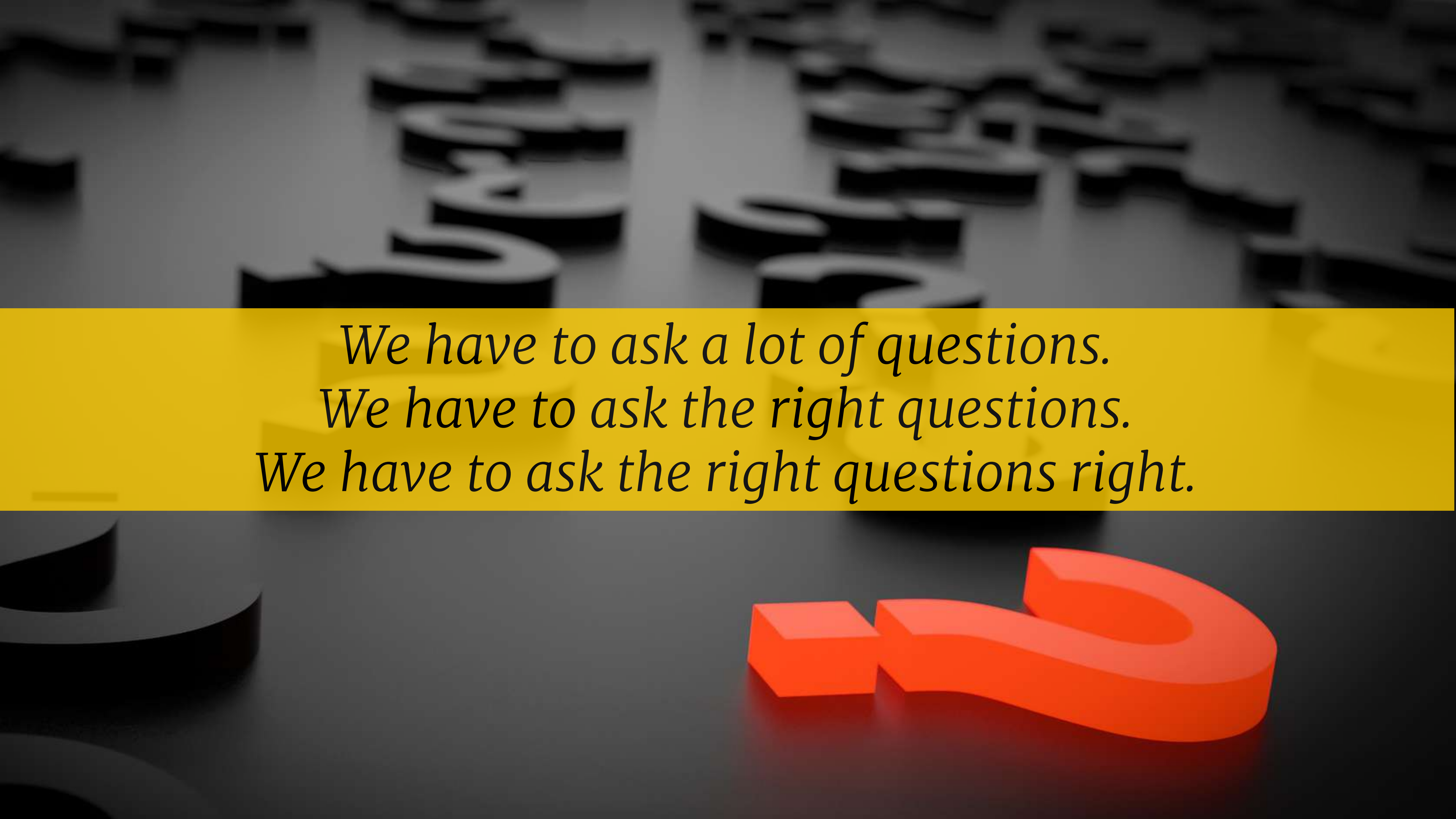
higher purchase satisfaction

Is the goal to increase sales?

Is the goal to have an attractive offer?

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing?. Journal of personality and social psychology, 79(6), 995.

http://www.ted.com/talks/sheena_iyengar_choosing_what_to_choose.html (at 1:22)



We have to ask a lot of questions.
We have to ask the right questions.
We have to ask the right questions right.

There are blind spots in single method evaluation with one metric.

Examples

Evaluating a music recommender system

Spotify
YouTube
pandora

. . .

example

Focus: Music consumer’s perspective

Offline evaluation
with focus on the music consumer

It can show that users' historic listening behavior can be simulated (e.g., high accuracy).

- Does the user want to listen to these familiar songs in the future?
- Would the user be satisfied with the same number/ proportion of unfamiliar songs?
- Is the user interested in discovering (more) new, unfamiliar songs?
- …

Online evaluation with focus on the music consumer

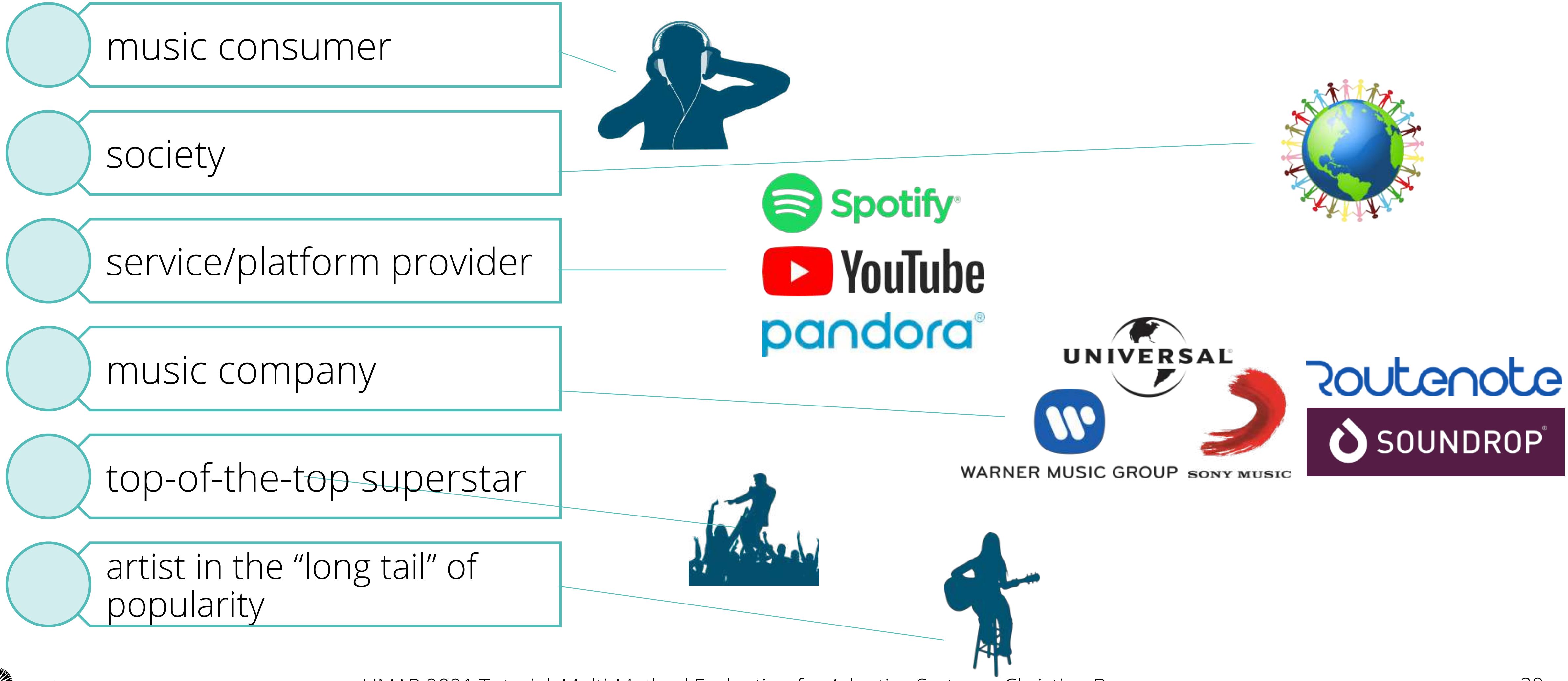
It can show that users click or skip recommended songs; or stay on platform for longer/shorter than usually.

- Does the user want to listen to the recommended songs in the future?
- Is the user is satisfied with the number/proportion of unfamiliar songs recommended?
  e.g., wants more discovery; skipped songs did not meet preferences; not in the mood for unfamiliar songs
- …

What does all that mean for evaluation?

Examples

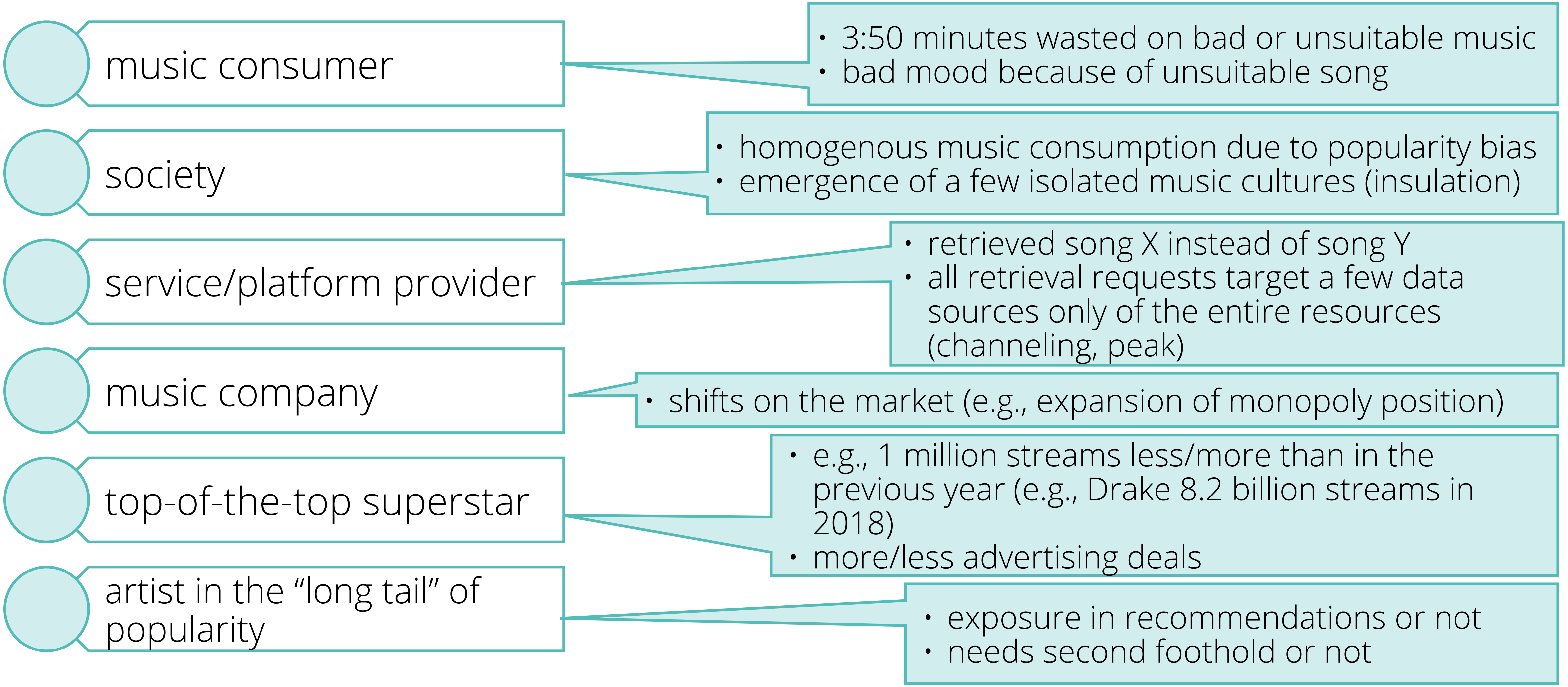
There are various stakeholder involved.
Example of the music recommender ecosystem.



- music consumer
- society
- service/platform provider
- music company
- top-of-the-top superstar
- artist in the "long tail" of popularity

Christine Bauer & Eva Zangerle (2019). Leveraging Multi-Method Evaluation for Multi-Stakeholder Settings. Proceedings of the 1st Workshop on the Impact of Recommender Systems (ImpactRS '19). Copenhagen, Denmark, 19 September.

What happens if recommendations go wrong?



music consumer
- 3:50 minutes wasted on bad or unsuitable music
- bad mood because of unsuitable song

society
- homogenous music consumption due to popularity bias
- emergence of a few isolated music cultures (insulation)

service/platform provider
- retrieved song X instead of song Y
- all retrieval requests target a few data sources only of the entire resources (channeling, peak)

music company
- shifts on the market (e.g., expansion of monopoly position)

top-of-the-top superstar
- e.g., 1 million streams less/more than in the previous year (e.g., Drake 8.2 billion streams in 2018)
- more/less advertising deals

artist in the "long tail" of popularity
- exposure in recommendations or not
- needs second foothold or not

We have to consider all stakeholders.
We have to involve all stakeholders.

Christine Bauer & Eva Zangerle (2019). Leveraging Multi-Method Evaluation for Multi-Stakeholder Settings. Proceedings of the 1st Workshop on the Impact of Recommender Systems (ImpactRS '19). Copenhagen, Denmark. 19 September.

Results of “traditional evaluation”

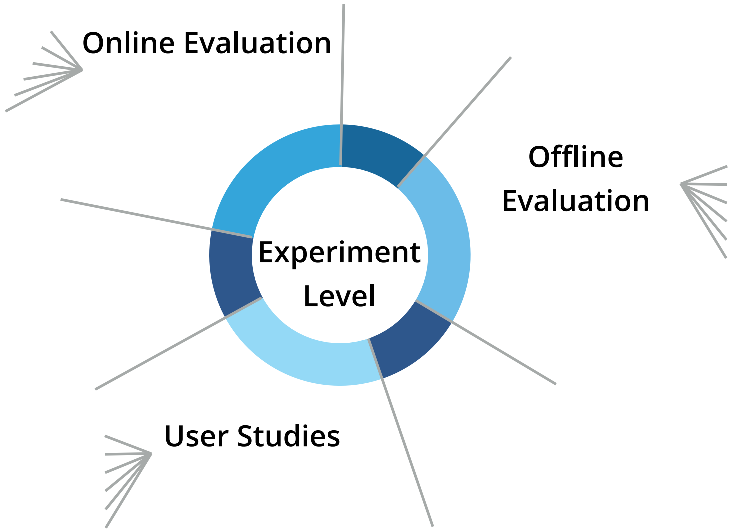
Focus on one single perspective

Incomplete picture: blind spots

Small set of metrics;
often picked from one perspective only

Evaluation results may differ

e.g., user satisfaction does not always correlate with high recommender accuracy
offline evaluations of accuracy are not always meaningful for predicting relative performance of different techniques



We need to thoughtfully configure the evaluation design space.
And we have to do this on several levels for a comprehensive evaluation.

Objectives

Hypothesis

Stakeholders

Offline Study

User Study

Online Study

Data

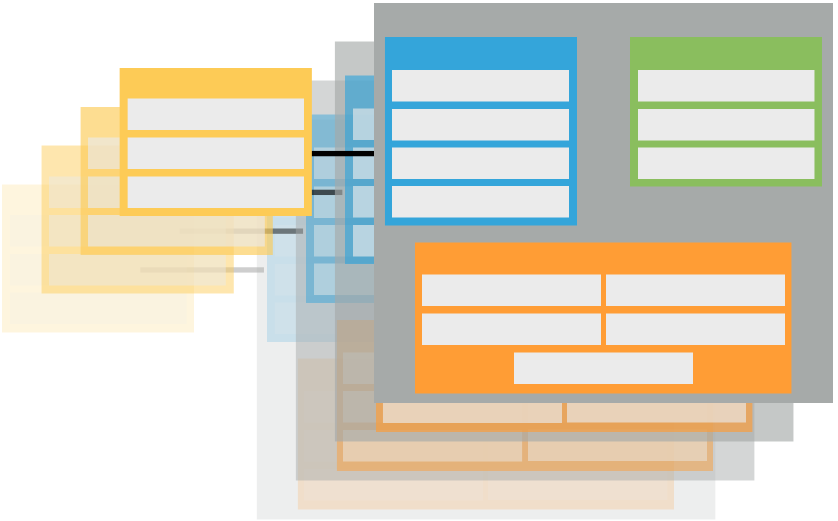
Data Quality and Biases

Evaluation Metrics

Control Variables

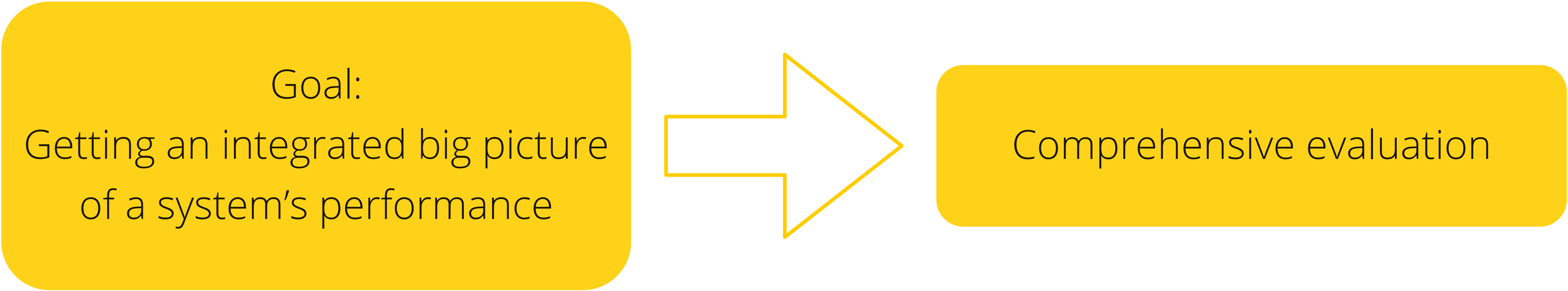
Reliability

Generalization Power



Eva Zangerle & Christine Bauer (under review). Evaluating Recommender Systems: Survey and Framework.

Introduction to multi-method evaluation



Goal:
Getting an integrated big picture
of a system's performance

Comprehensive evaluation

Multi-method evaluation

Similarities with mixed methods research

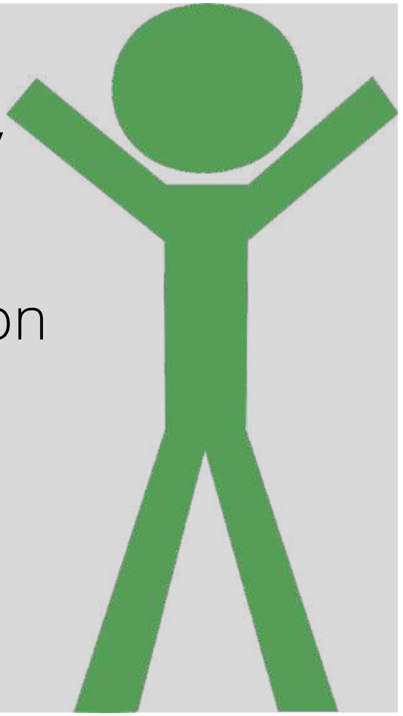
- mixed methods research
  - ➡ 3rd paradigm
  - ➡ combination of a quantitative and qualitative method
- multi-method evaluation
  - ➡ not restricted to qual+quant combination
  - ➡ focus on evaluation

Why combining multiple evaluation methods?

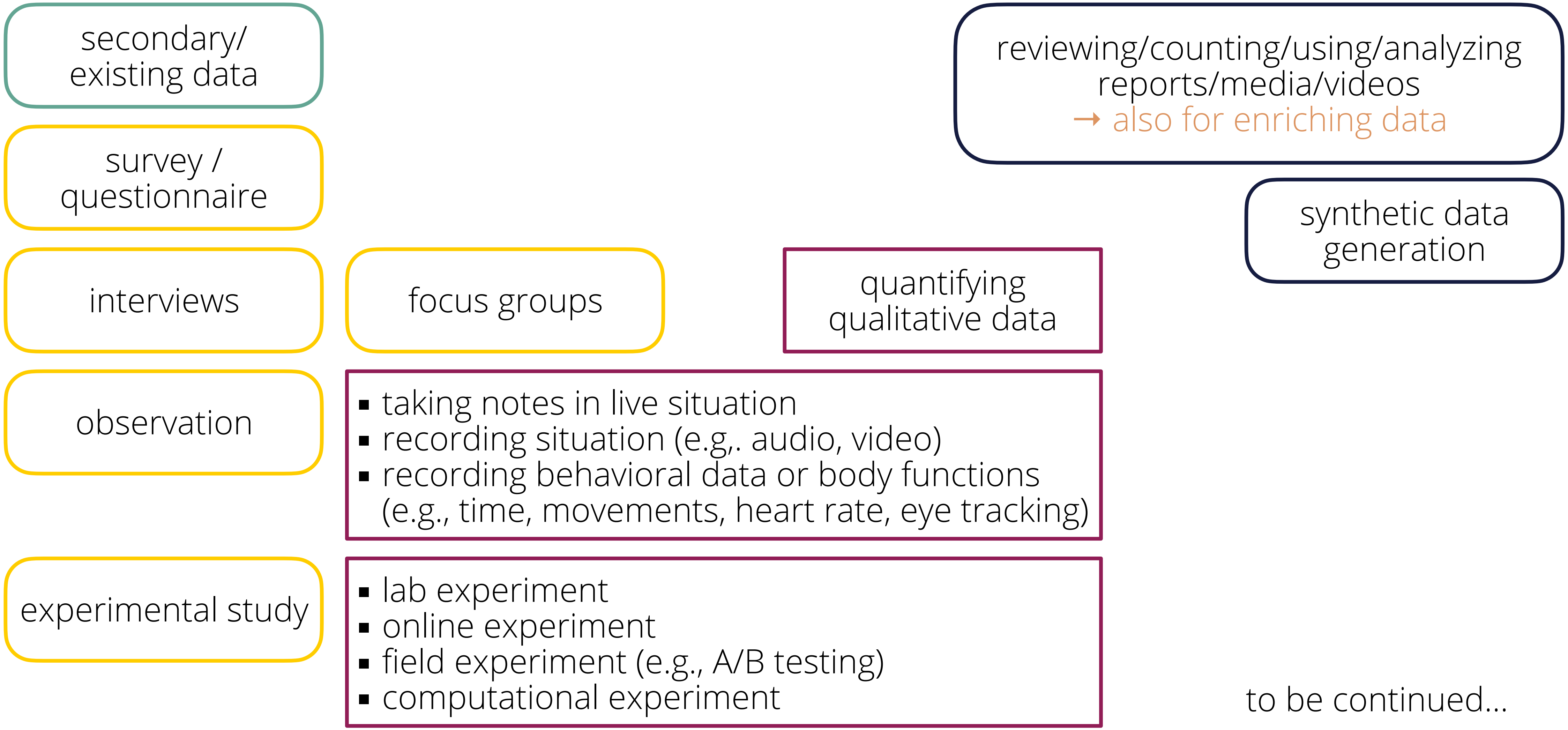
- To capture the same phenomenon from different angles
- To capture diverse, but complementary phenomena
- To resolve conflicting findings
- To get an integrated picture of performance in the context of use
- To triangulate quality

Benefits

- Explore sophisticated issues more holistically and widely
- Capture diverse, but complementary phenomena
- Apply diverse methods to capture the same phenomenon from possibly different angles
- Resolve conflicting findings
- Neutralize biases inherent to evaluation approaches



Data collection / elicitation



secondary/ existing data

survey / questionnaire

interviews

focus groups

observation

- taking notes in live situation
- recording situation (e.g,. audio, video)
- recording behavioral data or body functions (e.g., time, movements, heart rate, eye tracking)

experimental study

- lab experiment
- online experiment
- field experiment (e.g., A/B testing)
- computational experiment

quantifying qualitative data

reviewing/counting/using/analyzing reports/media/videos
→ also for enriching data

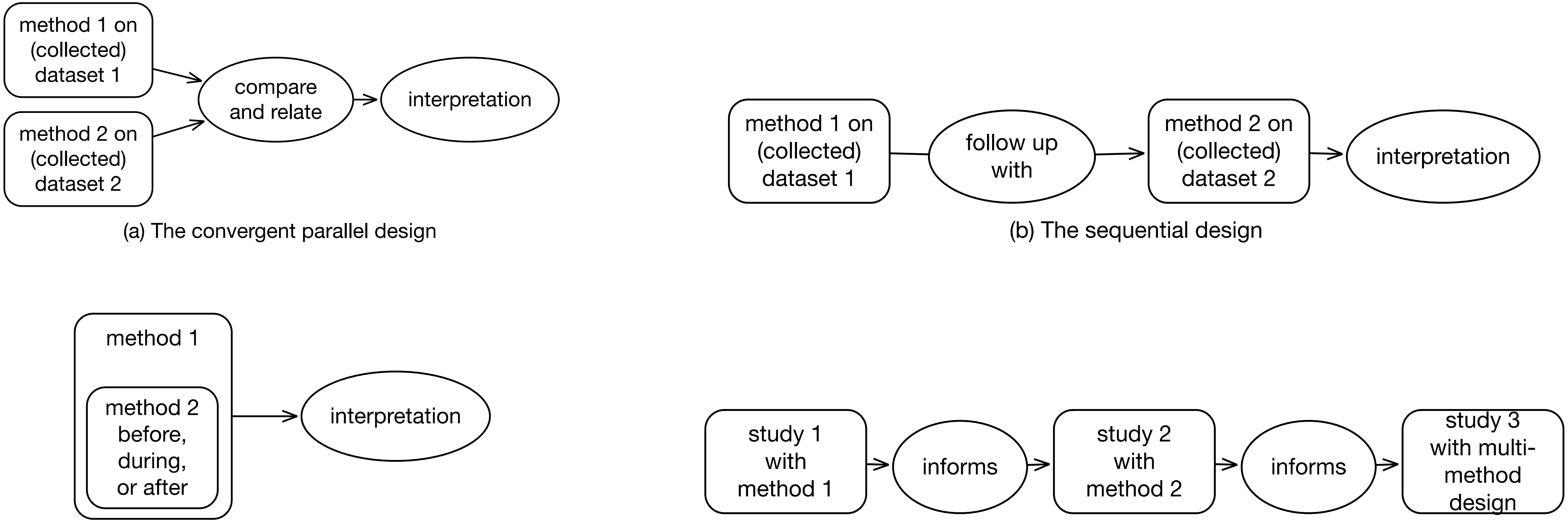
synthetic data generation

to be continued…

The four basic options of integrating multiple methods

There are several strategies for multi-method evaluation

John W Creswell and Vicki L. Plano Clark. 2011. Designing and conducting mixed methods research. Sage Publications, Los Angeles, CA, USA.



method 1 on (collected) dataset 1

method 2 on (collected) dataset 2

compare and relate

interpretation

(a) The convergent parallel design

method 1 on (collected) dataset 1

follow up with

method 2 on (collected) dataset 2

interpretation

(b) The sequential design

method 1

method 2 before, during, or after

interpretation

(c) The embedded design

study 1 with method 1

informs

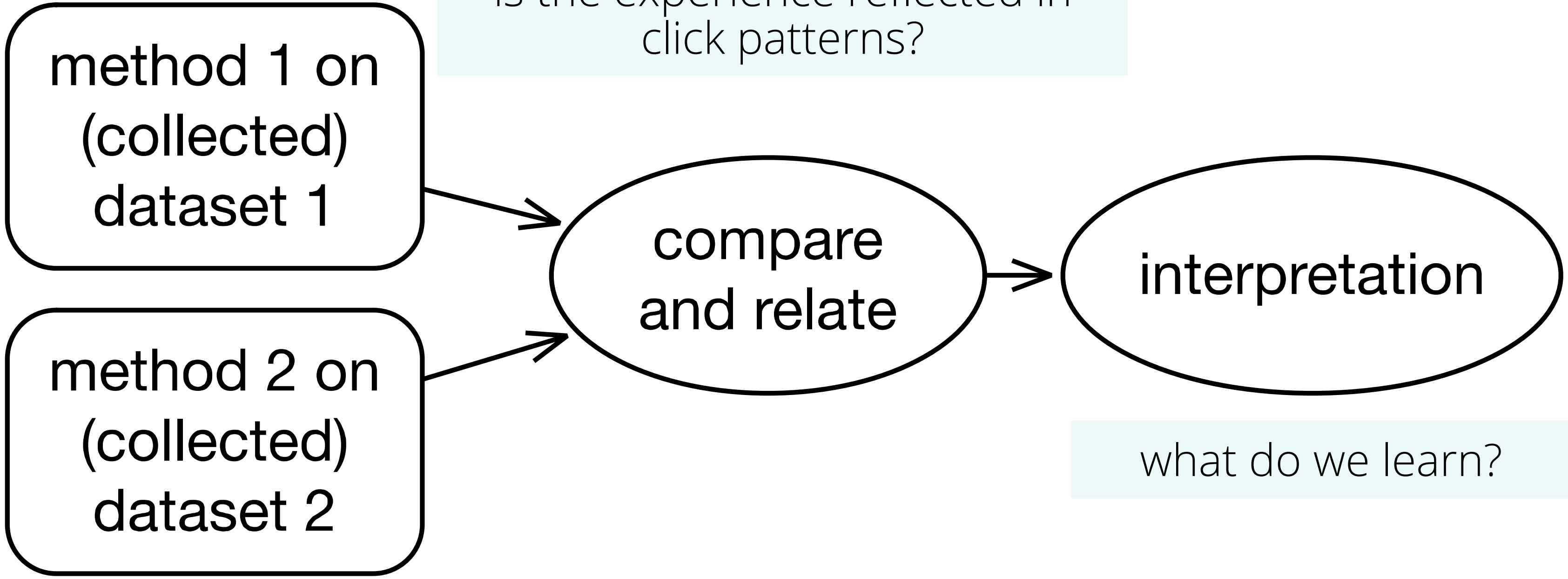
study 2 with method 2

informs

study 3 with multi-method design

(d) The multi-phase design

(a) The convergent parallel design


method 1 on (collected) dataset 1
method 2 on (collected) dataset 2
compare and relate
interpretation

prediction experiment based on clicks (implicit feedback); focus on rankings

survey or interviews with selected users on experience with rankings

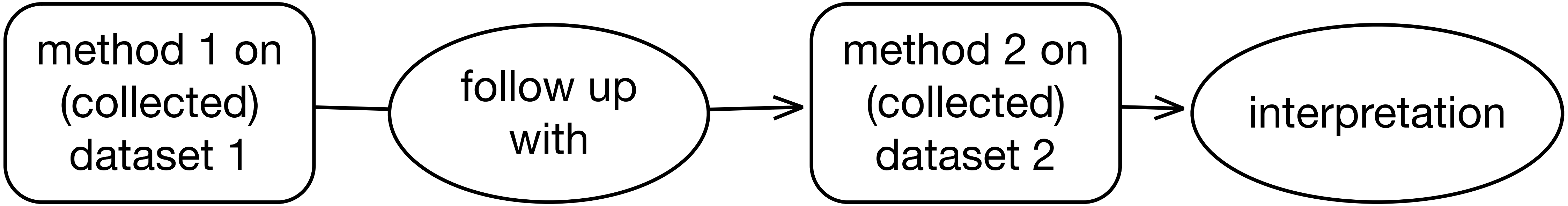
results are merged only here.
is the experience reflected in click patterns?

what do we learn?

(b) The sequential design

prediction experiment based on clicks (implicit feedback);
focus on rankings

laboratory experiment to test different interface designs;
click patterns



method 1 on (collected) dataset 1

follow up with

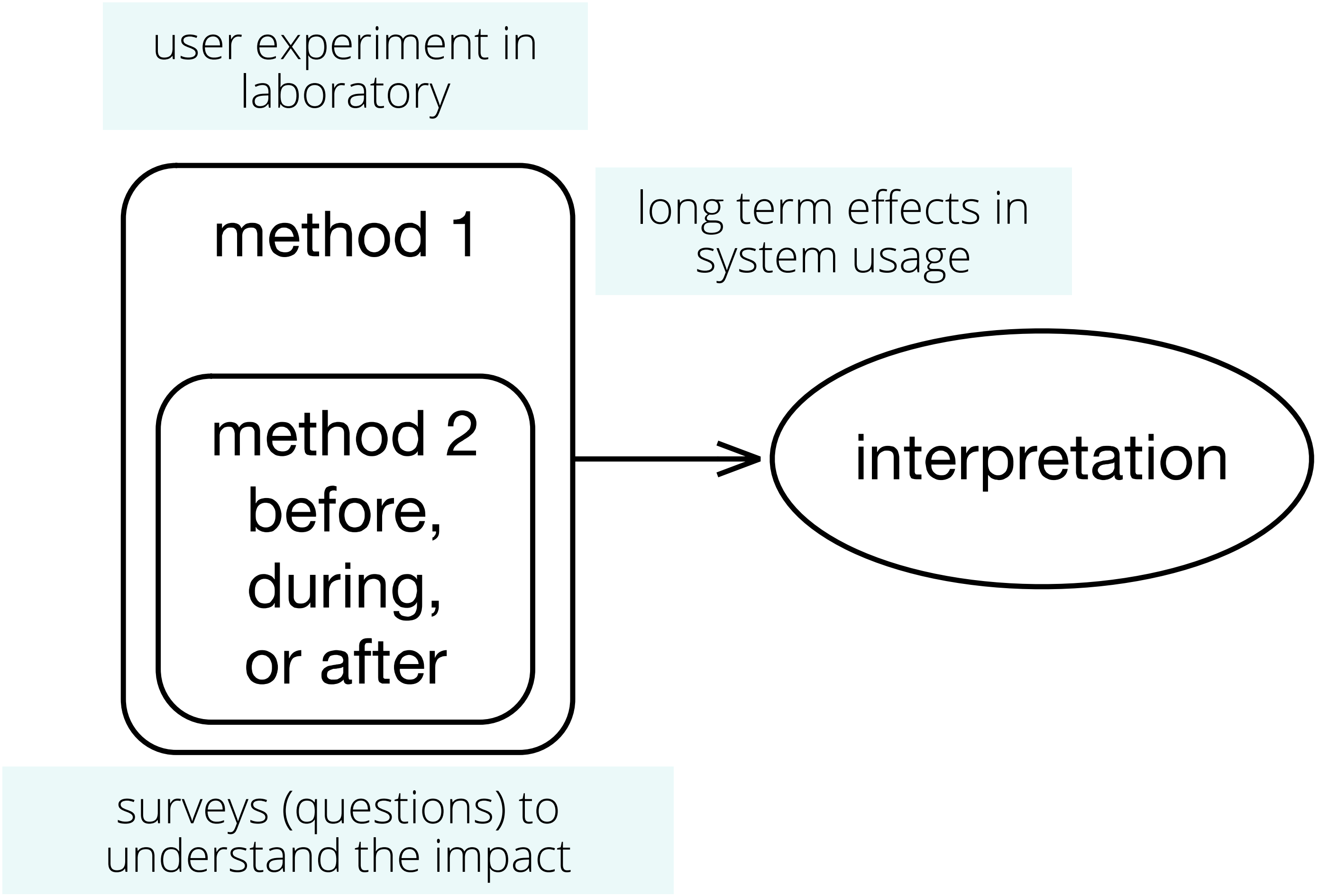
method 2 on (collected) dataset 2

interpretation

what do we learn from the two studies altogether?

(c) The embedded design

Purpose is to answer different questions that require different types of data.



user experiment in laboratory

long term effects in system usage

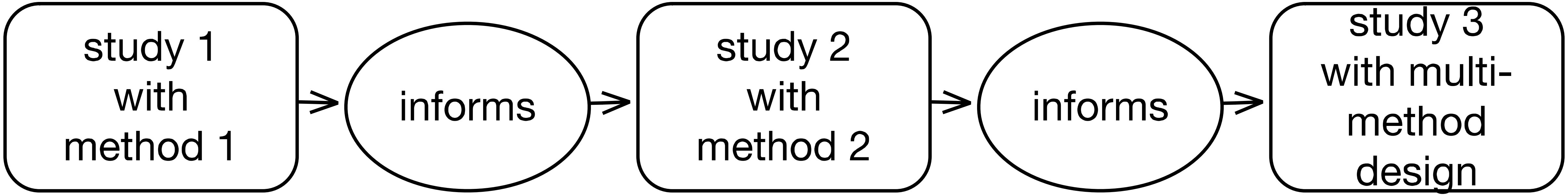
surveys (questions) to understand the impact

(d) The multi-phase design

online study focusing on click patterns of users using a recsys

think-aloud study with selected users with goal to find out why they click on which item or quit

experiment to test influencing factors on different clicking behavior with additional survey



study 1 with method 1

informs

study 2 with method 2

informs

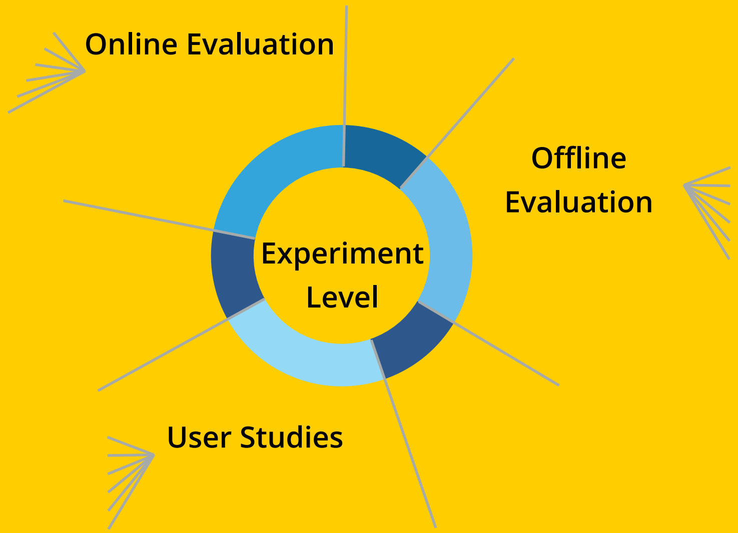
study 3 with multi-method design

we find out that users click mostly on the first three items, then they quit the platform

what do we learn from this altogether?

Break

Wrap up



Objectives

Hypothesis

Stakeholders

Offline Study

User Study

Online Study

Data

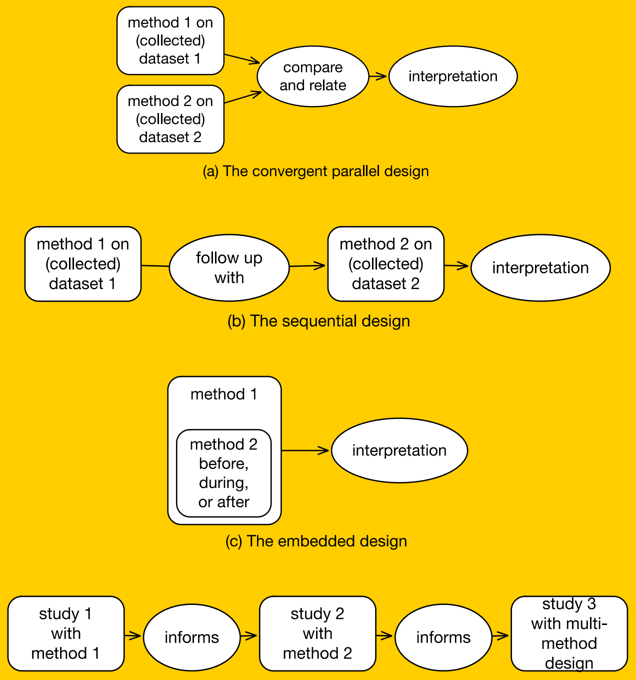
Data Quality and Biases

Evaluation Metrics

Control Variables

Reliability

Generalization Power


method 1 on (collected) dataset 1
method 2 on (collected) dataset 2
compare and relate
interpretation
method 1 on (collected) dataset 1
follow up with
method 2 on (collected) dataset 2
interpretation
method 1
method 2 before, during, or after
interpretation
study 1 with method 1
informs
study 2 with method 2
informs
study 3 with multi-method design

(a) The convergent parallel design

(b) The sequential design

(c) The embedded design

(d) The multi-phase design

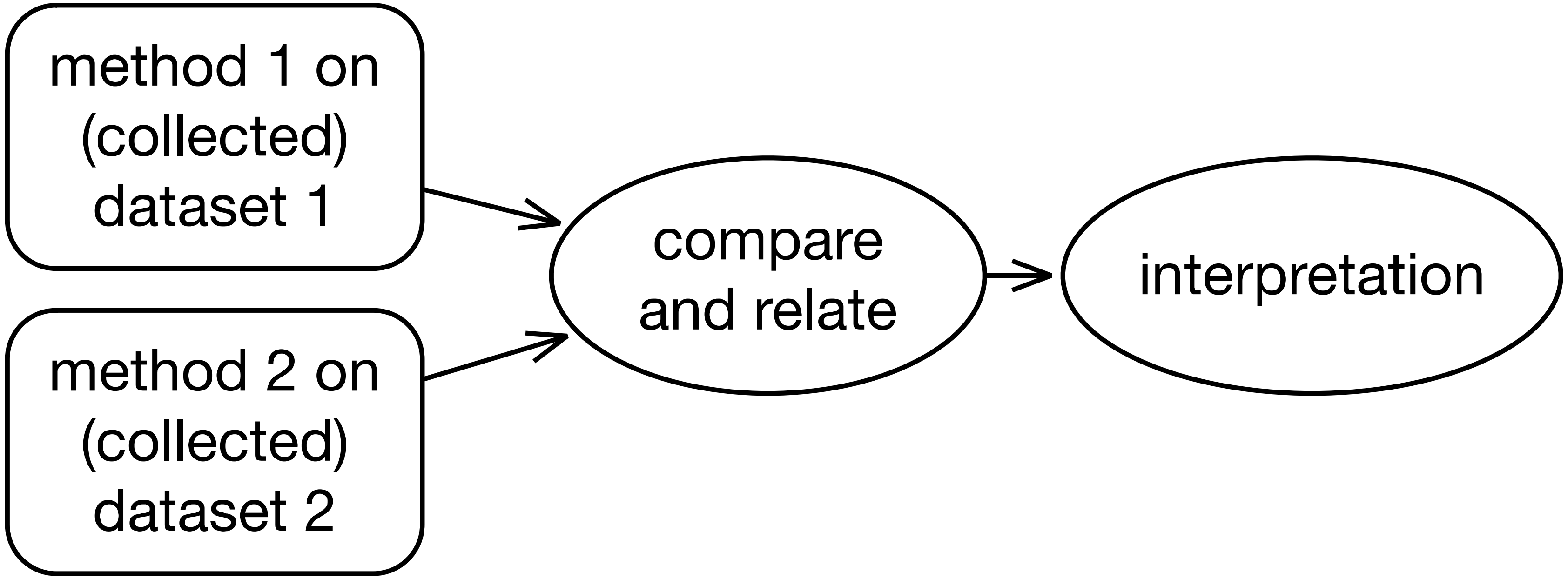
Let's get started

Task

Let’s imagine:
An UMAP 2022 configuration system

- adapts to preferences and needs of all people interested in UMAP 2022
- we need to evaluate this system

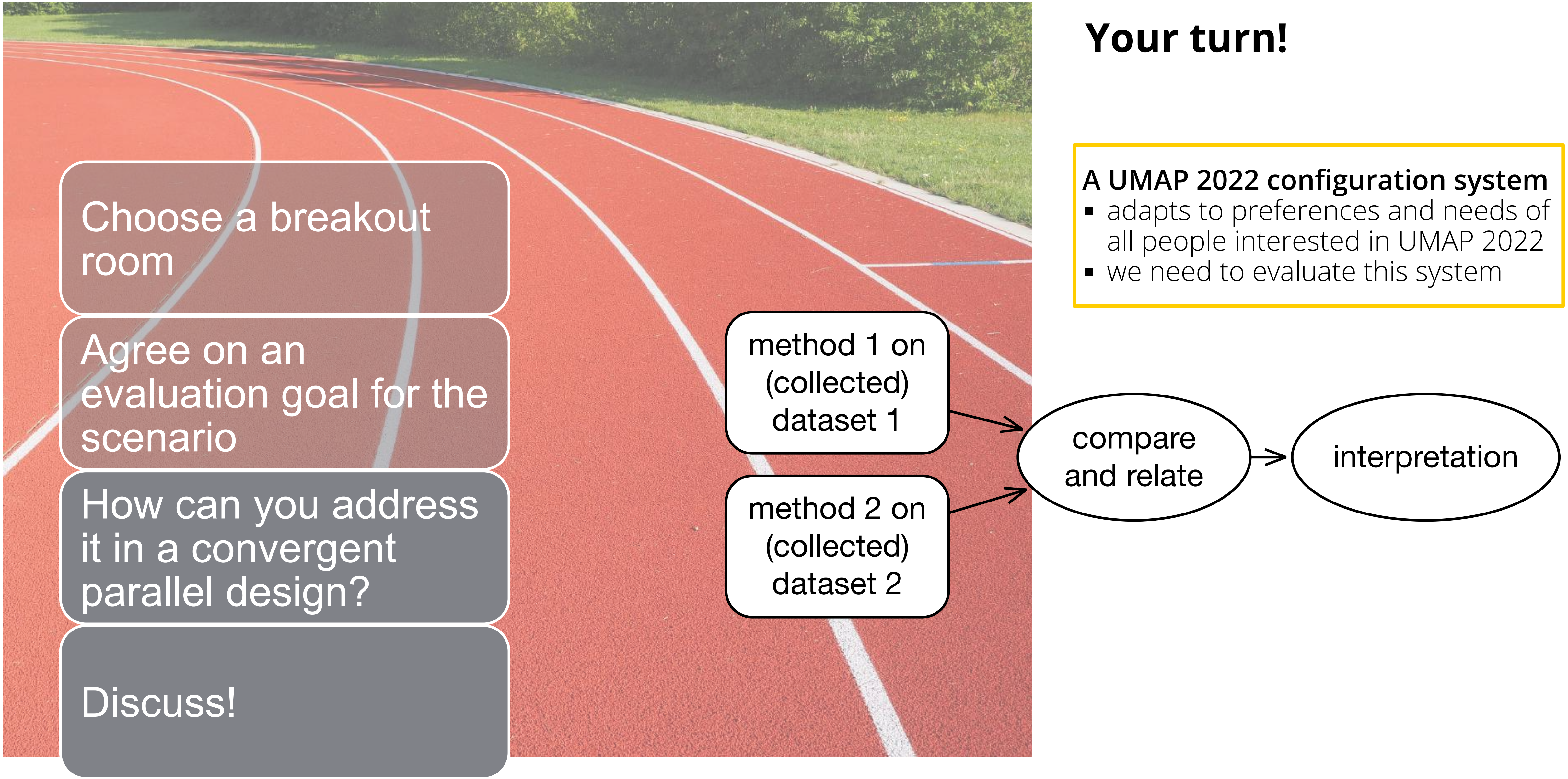
(a) The convergent parallel design


method 1 on (collected) dataset 1
method 2 on (collected) dataset 2
compare and relate
interpretation

Your turn!

A UMAP 2022 configuration system

- adapts to preferences and needs of all people interested in UMAP 2022
- we need to evaluate this system



Choose a breakout room

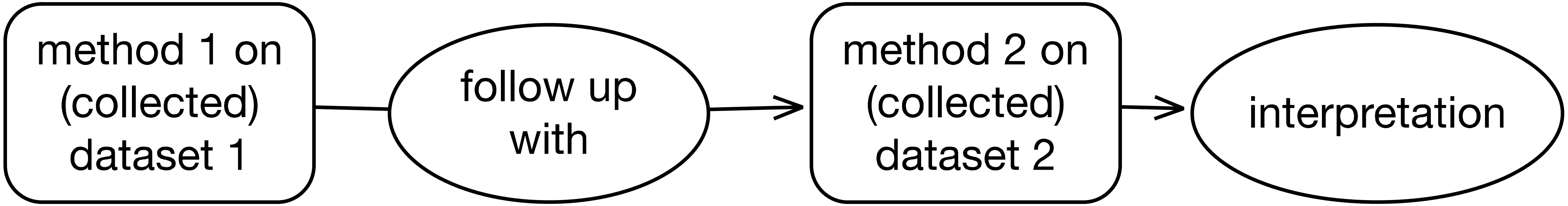
Agree on an evaluation goal for the scenario

How can you address it in a convergent parallel design?

Discuss!

method 1 on (collected) dataset 1
method 2 on (collected) dataset 2
compare and relate
interpretation

(b) The sequential design


method 1 on (collected) dataset 1
follow up with
method 2 on (collected) dataset 2
interpretation

Your turn!

A UMAP 2022 configuration system

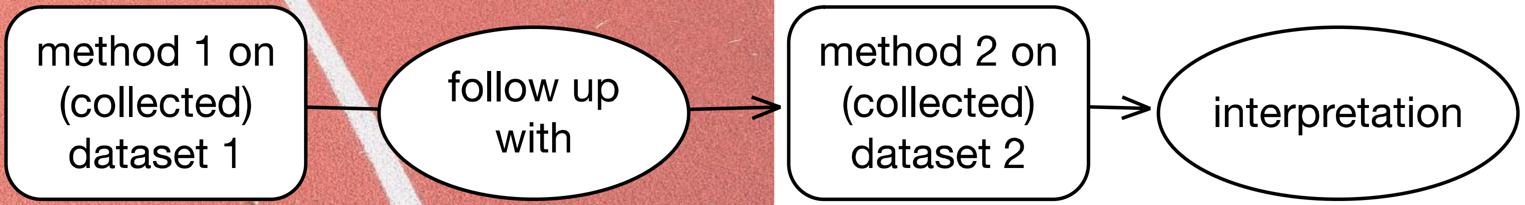
- adapts to preferences and needs of all people interested in UMAP 2022
- we need to evaluate this system

Choose a breakout room

Agree on an evaluation goal for the scenario

How can you address it in a sequential design?

Discuss!



method 1 on (collected) dataset 1

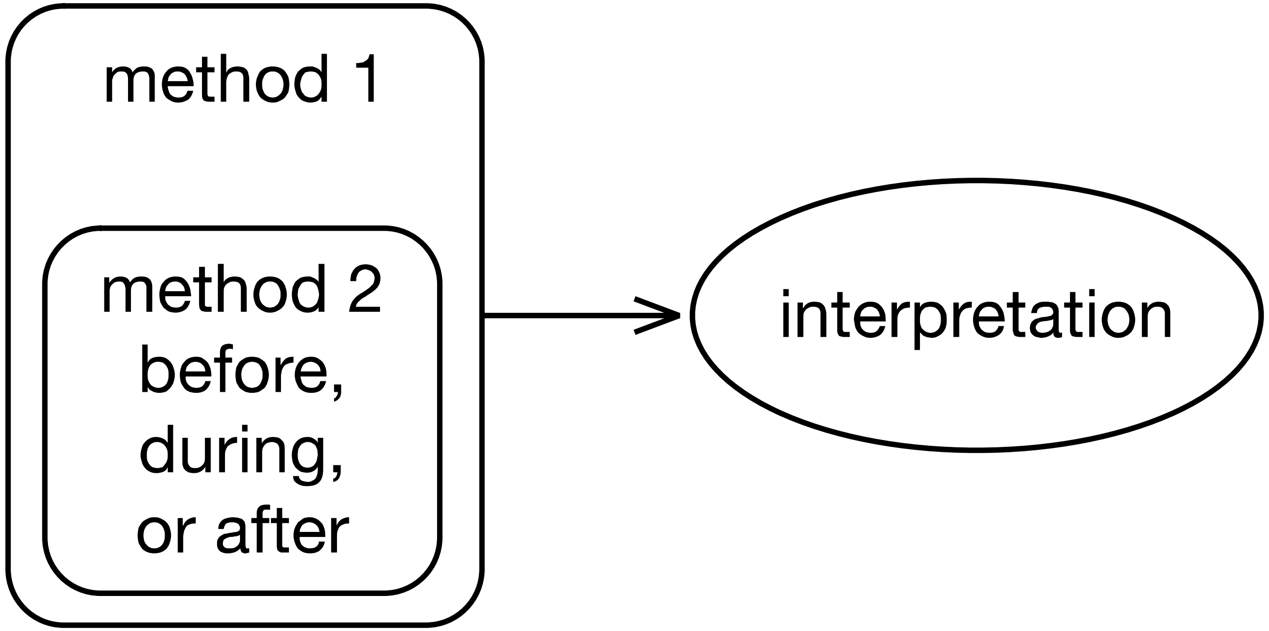
follow up with

method 2 on (collected) dataset 2

interpretation

(c) The embedded design

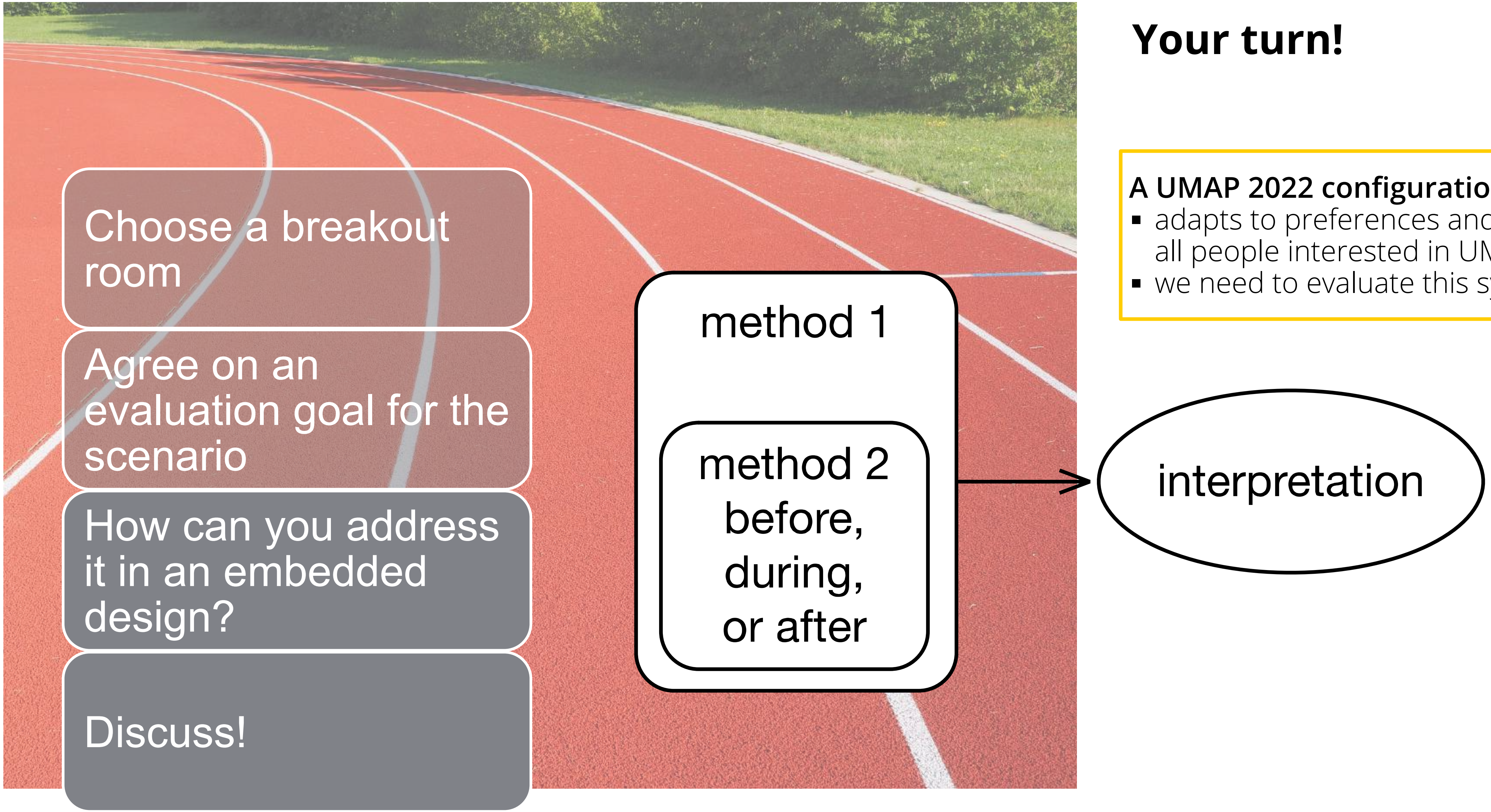
Purpose is to answer different questions that require different types of data.



Your turn!

A UMAP 2022 configuration system

- adapts to preferences and needs of all people interested in UMAP 2022
- we need to evaluate this system



Choose a breakout room

Agree on an evaluation goal for the scenario

How can you address it in an embedded design?

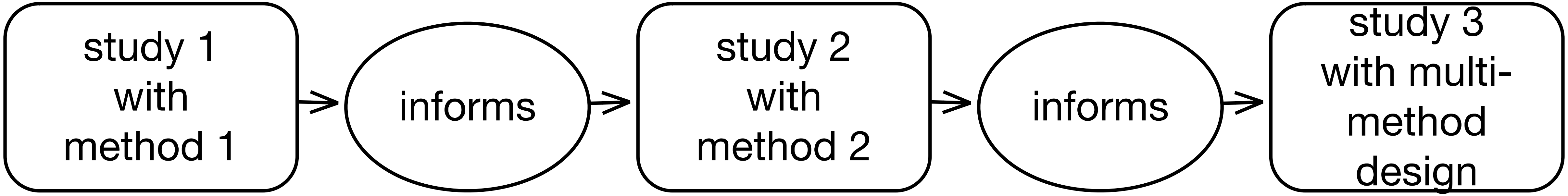
Discuss!

method 1

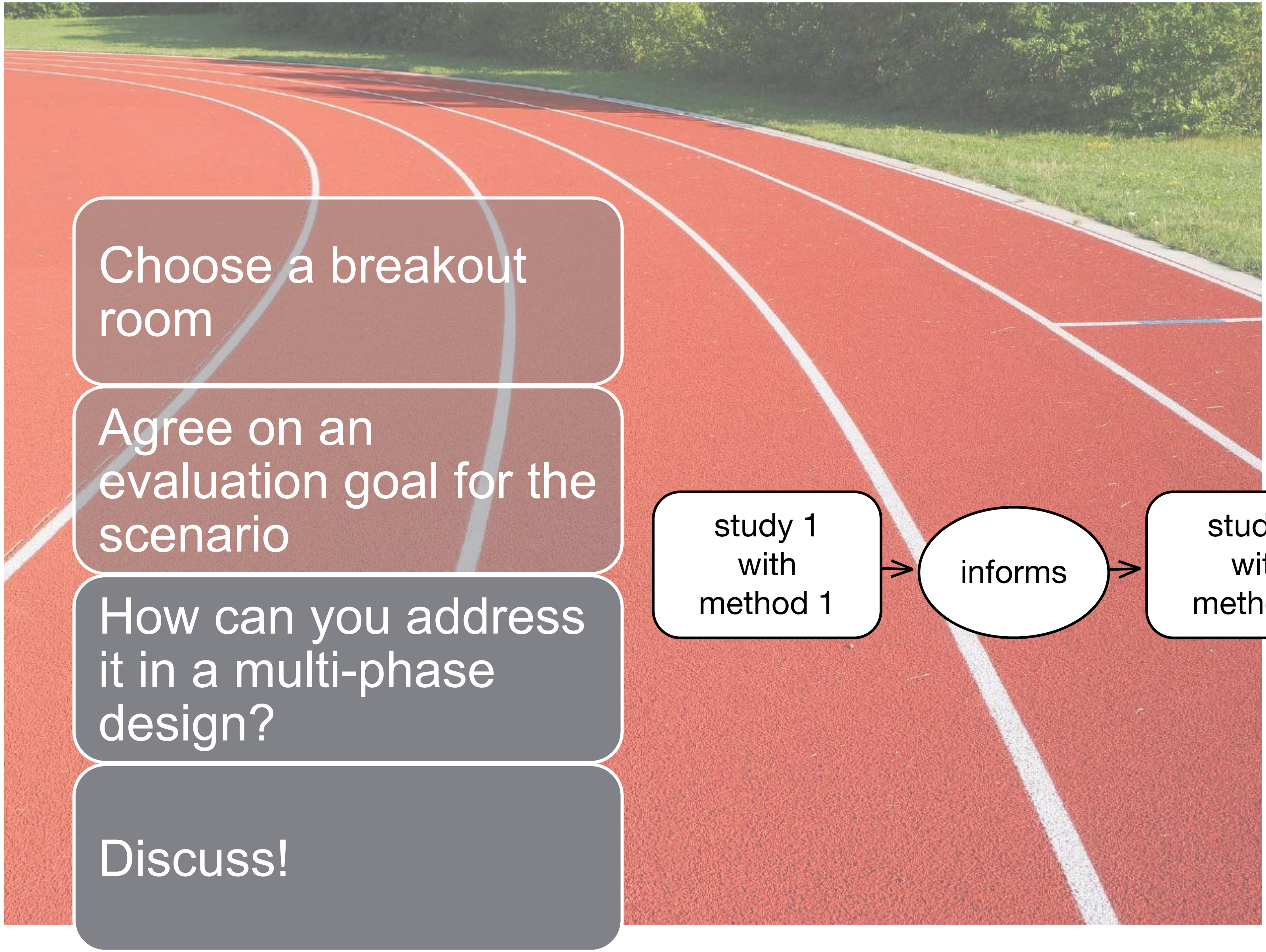
method 2 before, during, or after

interpretation

(d) The multi-phase design



Your turn!



Choose a breakout room

Agree on an evaluation goal for the scenario

How can you address it in a multi-phase design?

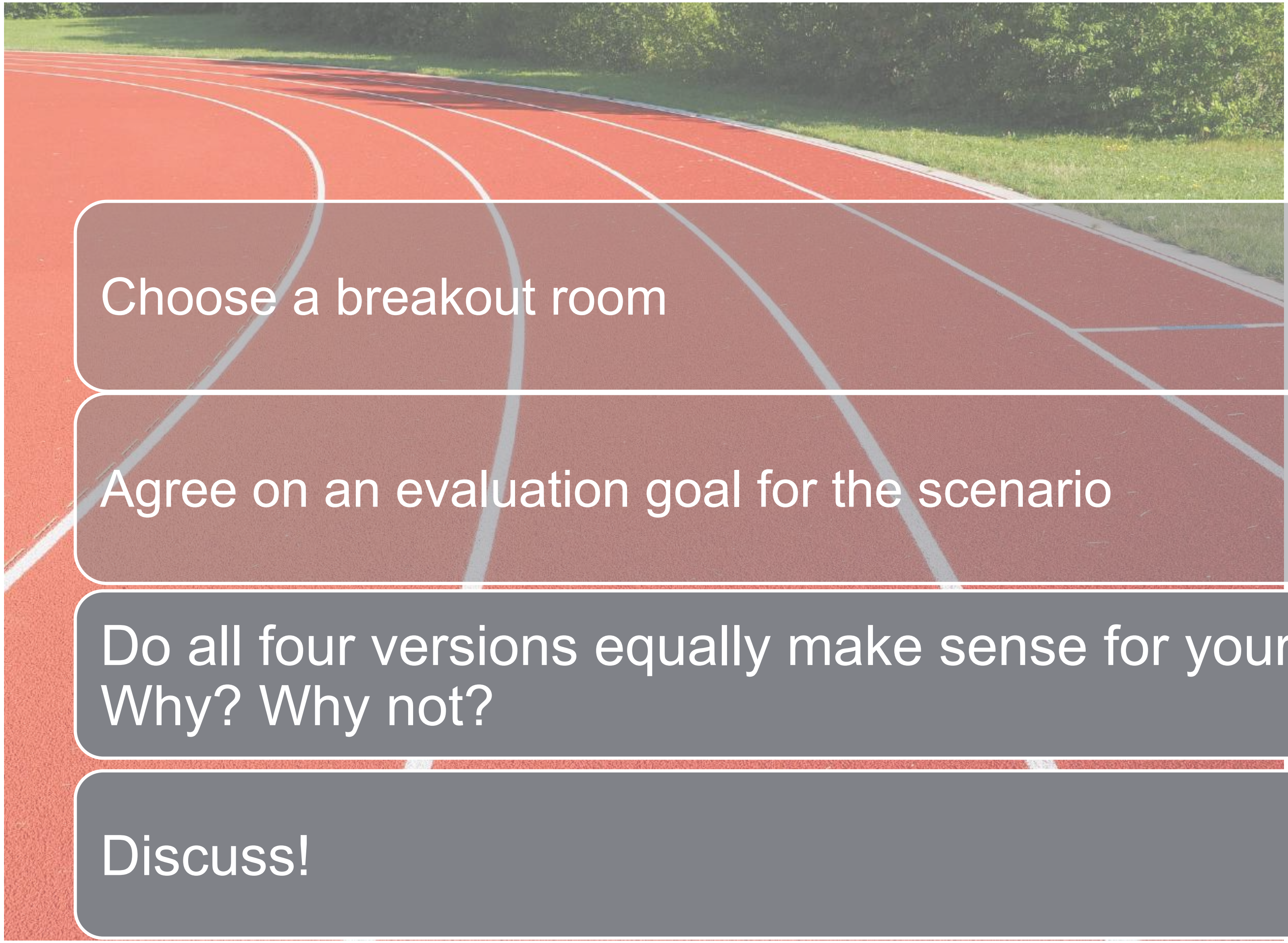
Discuss!

A UMAP 2022 configuration system

- adapts to preferences and needs of all people interested in UMAP 2022
- we need to evaluate this system

study 1 with method 1 → informs → study 2 with method 2 → informs → study 3 with multi-method design

Your turn!



Choose a breakout room

Agree on an evaluation goal for the scenario

Do all four versions equally make sense for your research goal? Why? Why not?

Discuss!

The challenges of multi-method evaluation

Factors to consider when choosing one method over another?

Balance between strengths and weaknesses associated with each method

Time for data collection and analysis

- observation or interview method helps to collect richer information, but it takes time
- survey helps you collect more data quickly, yet it may lack details

Feasibility of data acquisition / access to data

- dataset available that really fits the research goal
  (e.g., MovieLens again? Yes/no? Why/why not?)
- access to target group
  (access to specific user groups may be challenging; e.g., children, experts in a field)
- privacy and ethical concerns (institutional review board (IRB))

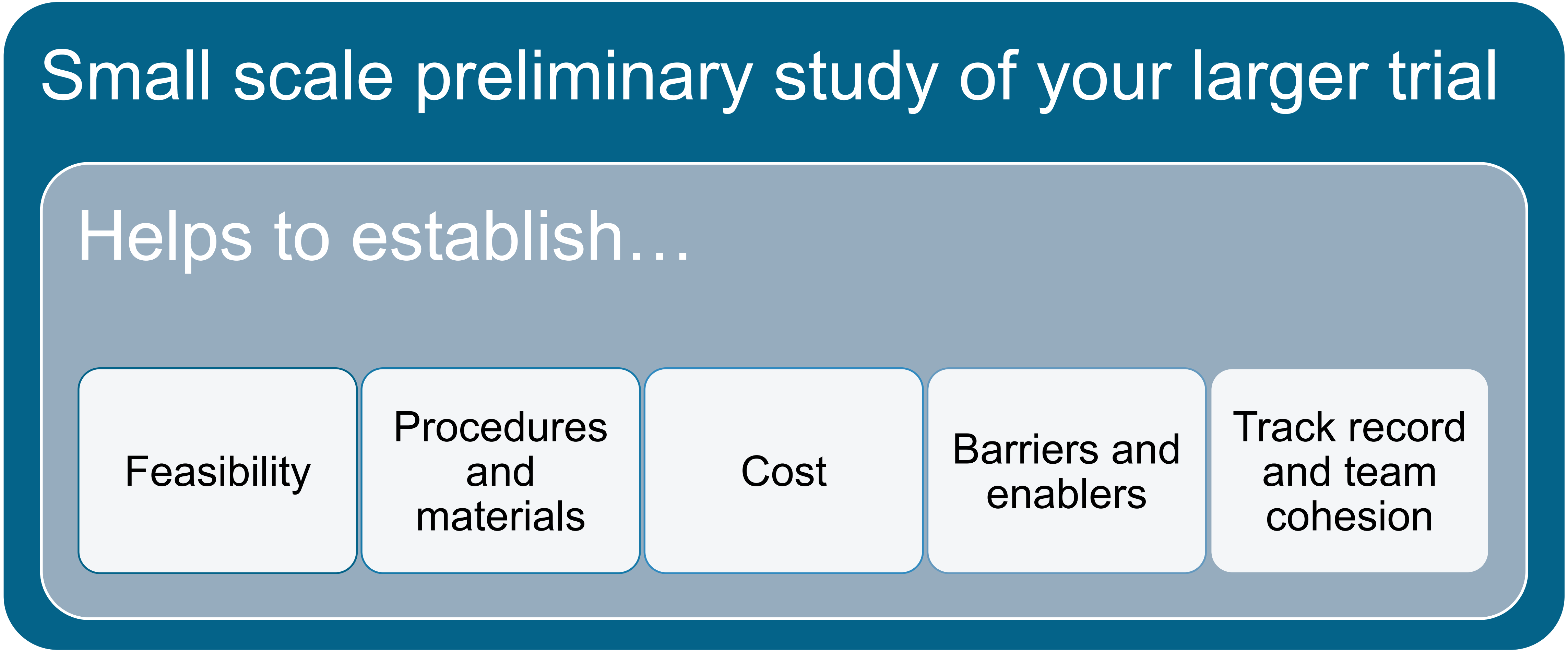
Access to skills for the method

- being non-skilled is not an excuse!!
- learning takes time
- identifying and getting involved skilled co-contributors takes time

...

What is feasible?

Pilot studies



Small scale preliminary study of your larger trial

Helps to establish…

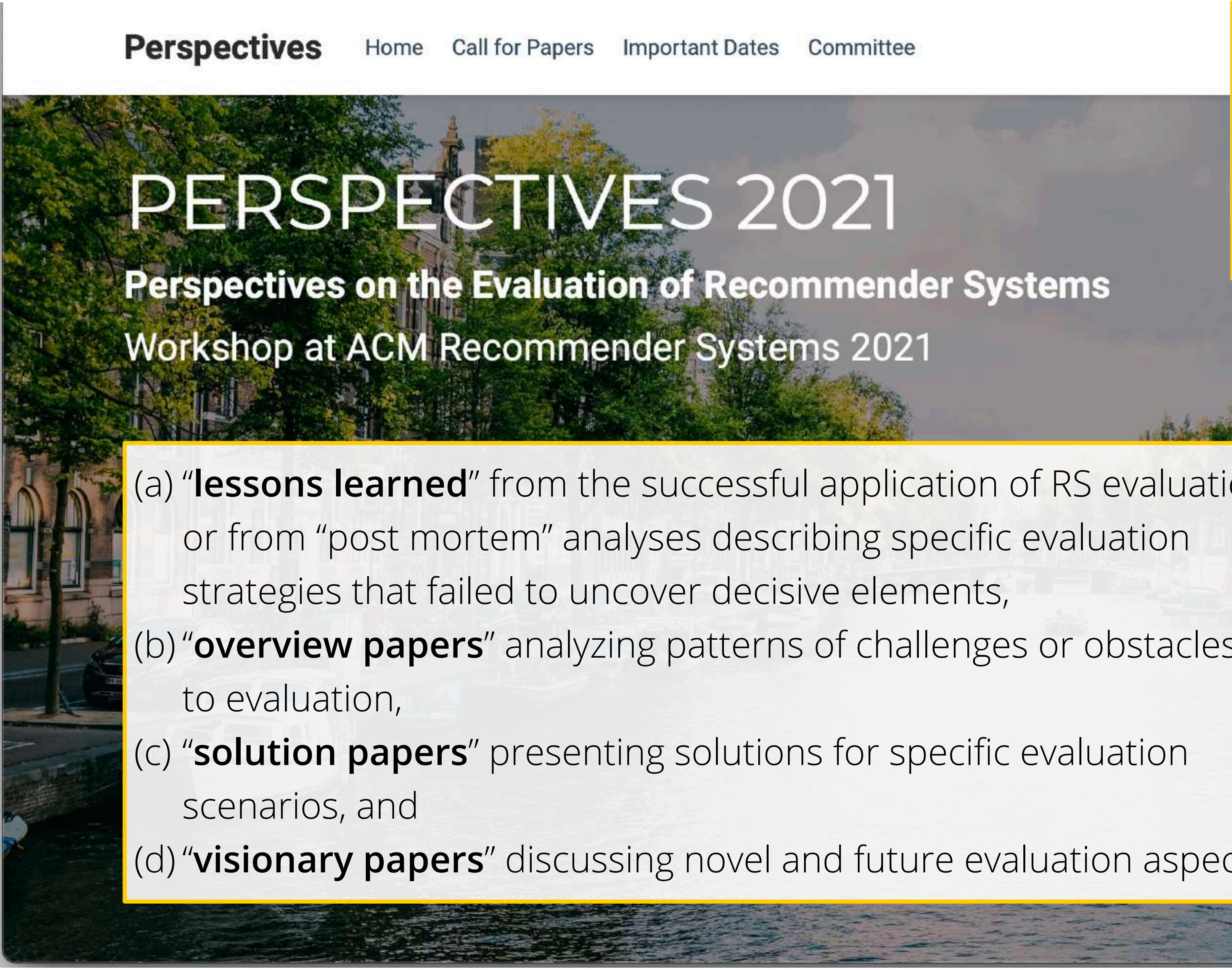
- Feasibility
- Procedures and materials
- Cost
- Barriers and enablers
- Track record and team cohesion

Things to remember

- Select a study design that allows you to answer your research question
- Select a design that provides the highest level of evidence possible – but is also feasible
- Conduct a pilot
- Pay attention to the finer details

Where do we go from here?

→ **menti.com**



PERSPECTIVES 2021

Perspectives on the Evaluation of Recommender Systems

Workshop at ACM Recommender Systems 2021

Eva Zangerle Christine Bauer Alan Said

(a) “**lessons learned**” from the successful application of RS evaluation or from “post mortem” analyses describing specific evaluation strategies that failed to uncover decisive elements,

(b) “**overview papers**” analyzing patterns of challenges or obstacles to evaluation,

(c) “**solution papers**” presenting solutions for specific evaluation scenarios, and

(d) “**visionary papers**” discussing novel and future evaluation aspects.

Paper submission deadline: July 29th, 2021

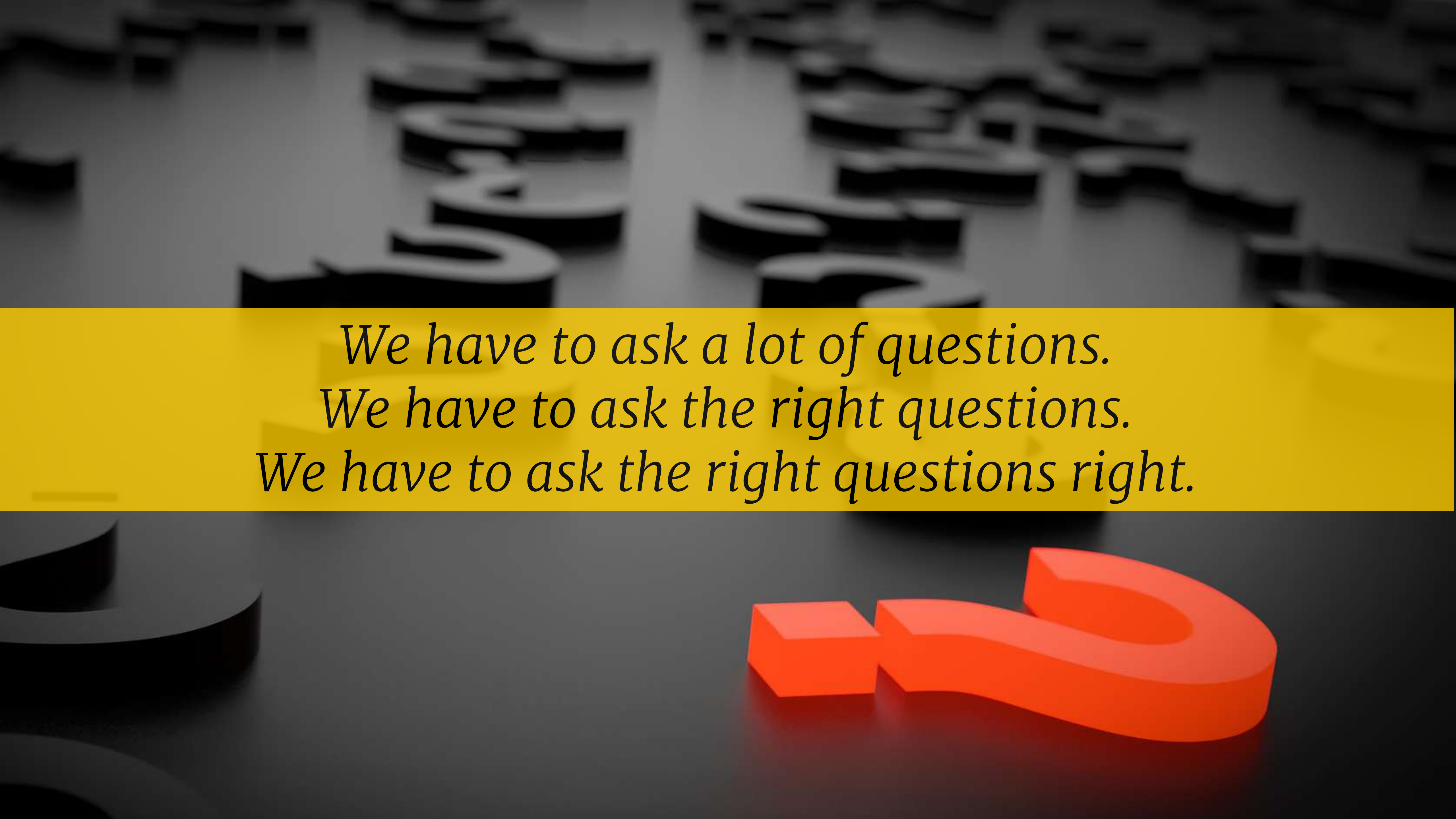
https://perspectives-ws.github.io/2021/

https://multimethods.info

maintained in collaboration with
Eva Zangerle

Take away

→ **reusable!!**



We have to ask a lot of questions.
We have to ask the right questions.
We have to ask the right questions right.

Things to remember

- Look at phenomena from different angles
- If your research is related to users, involve them!
- When focusing,
  have the overall picture in mind
- When having the overall picture in mind,
  keep your focus