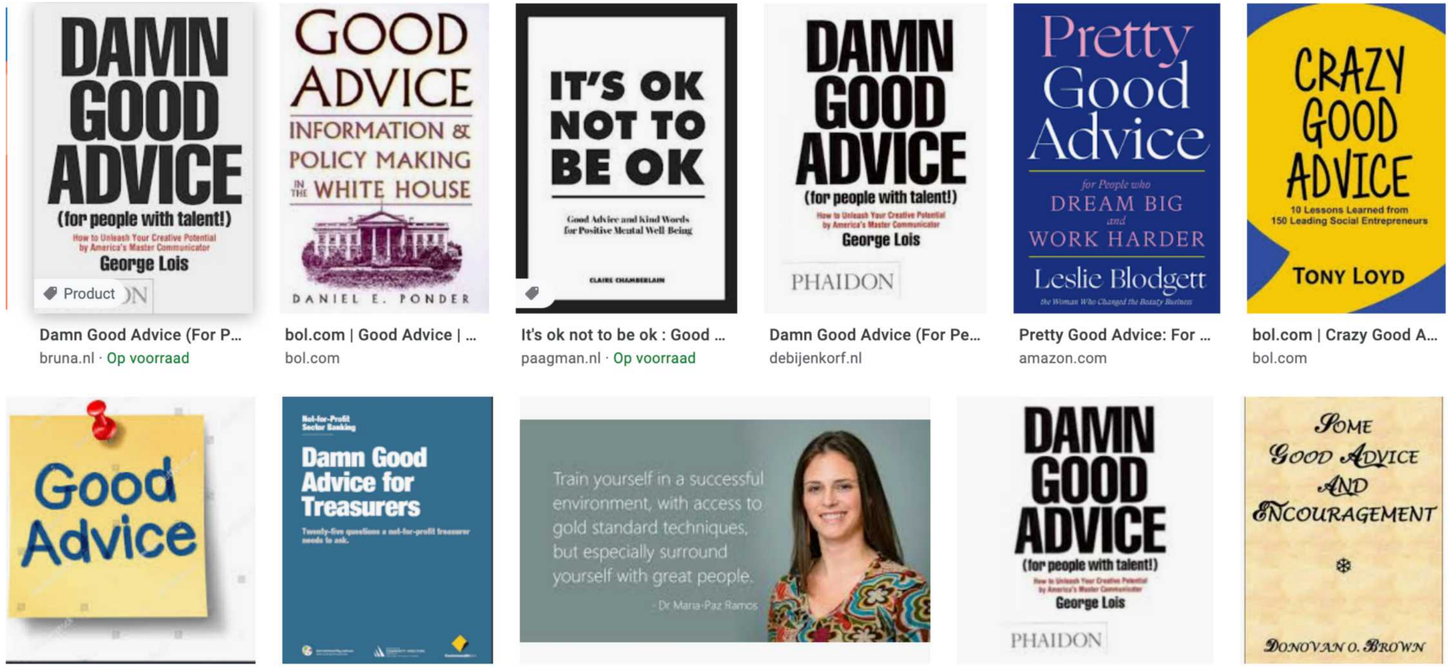
HUMAN EXPERIENCE & MEDIA DESIGN

HEMD

METRICS – WEEK 4

TALKING ABOUT METRICS



Main question: when are recommendations 'good'

CLASSIC APPROACH

You calculate in various ways:

- do the recommendations that show up match the **mathematically expected recommendation**?

CLASSIC APPROACH

You calculate in various ways:

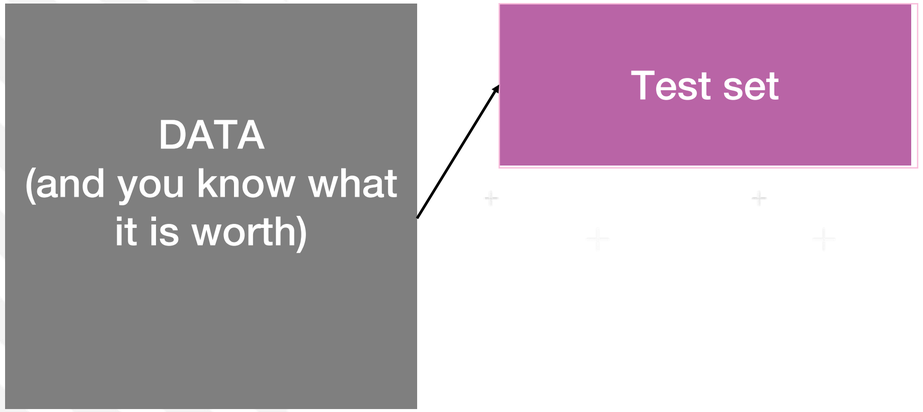
- do the recommendations that show up match the **mathematically expected recommendation**?

- the result is a **metric**

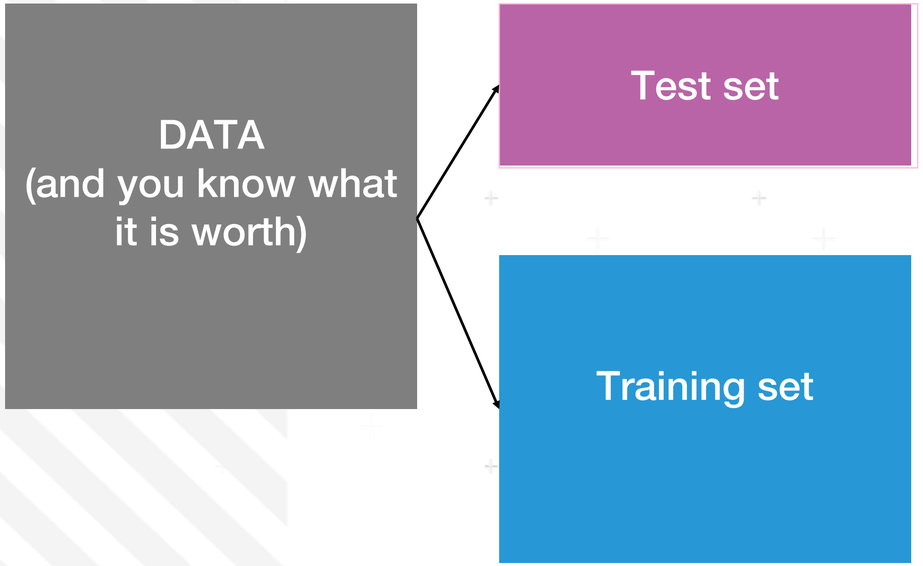
HOW PEOPLE BUILD ALGORITHMS (ONE OPTION)

DATA
(and you know what it is worth)

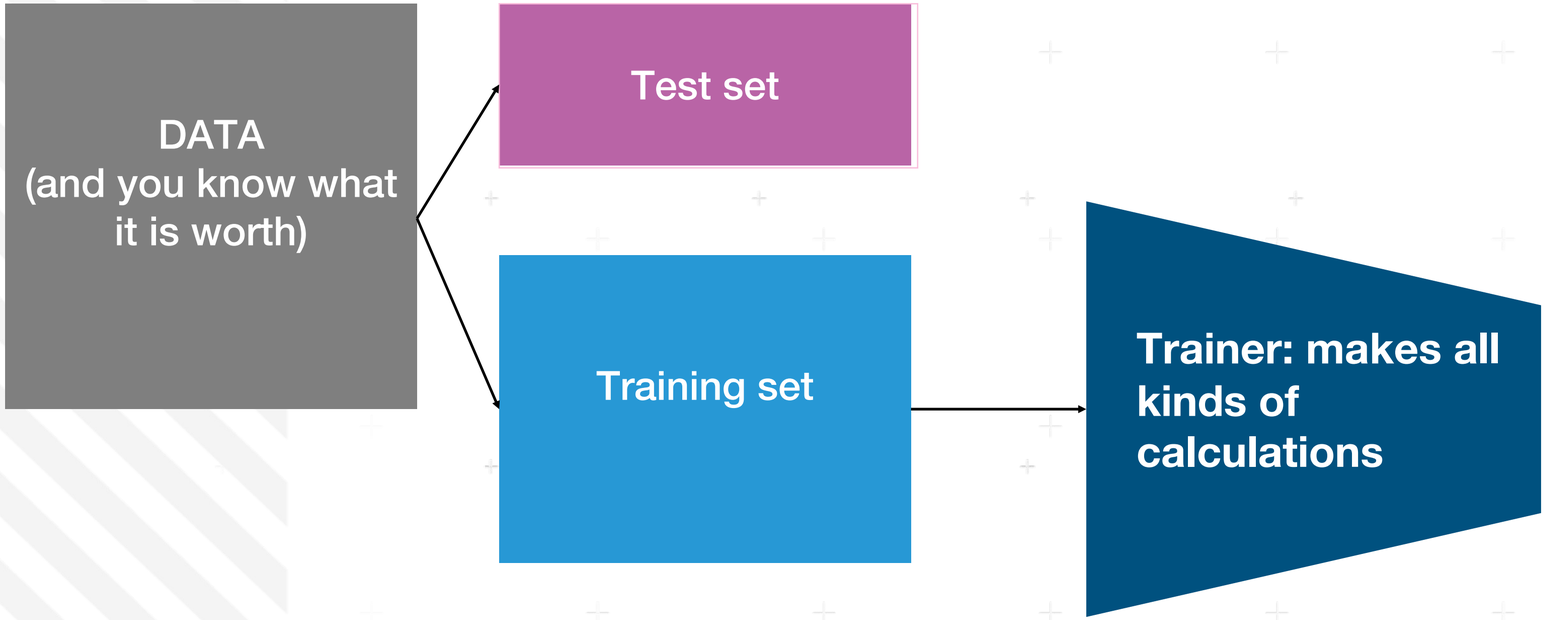
HOW PEOPLE BUILD ALGORITHMS (ONE OPTION)



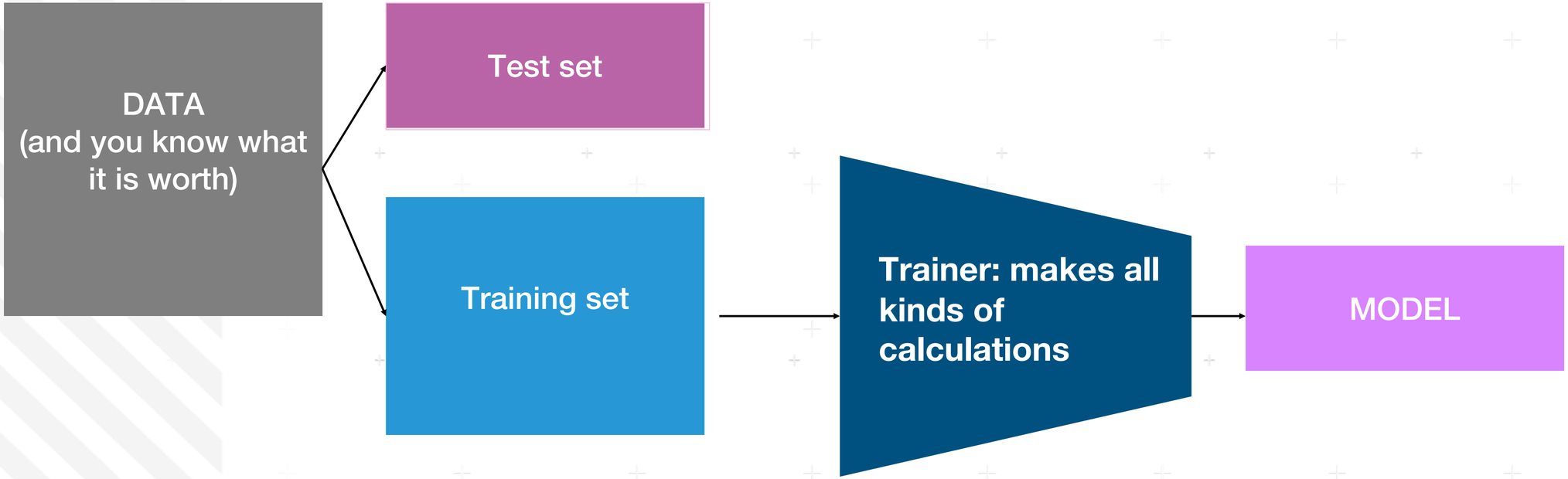
HOW PEOPLE BUILD ALGORITHMS (ONE OPTION)



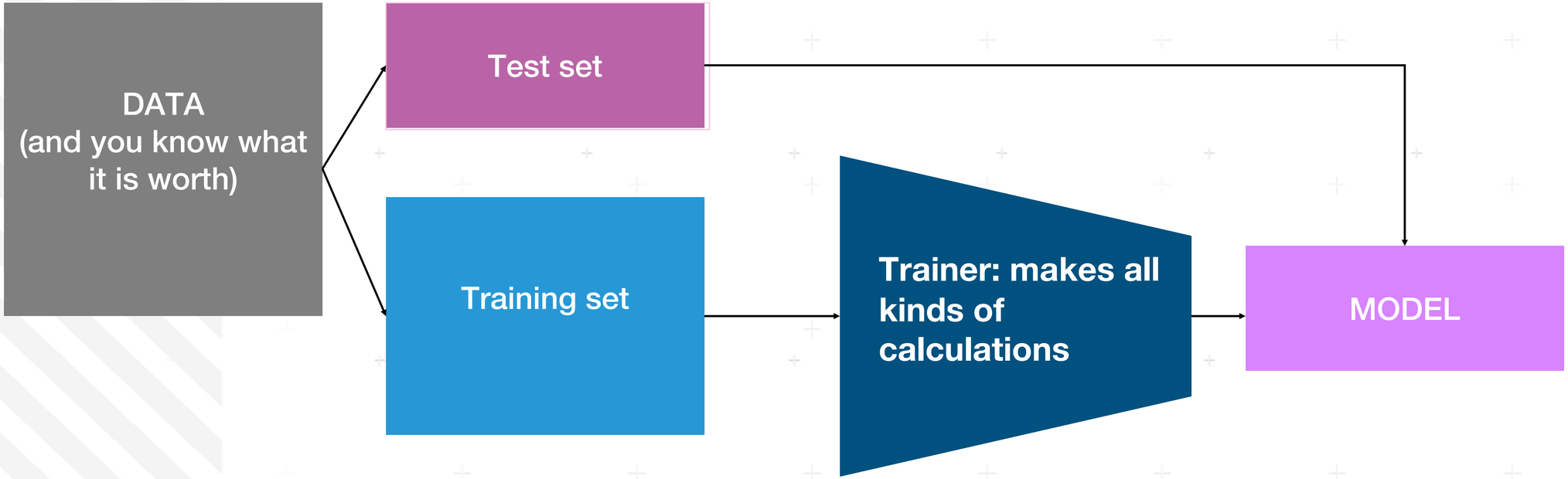
HOW PEOPLE BUILD ALGORITHMS (ONE OPTION)



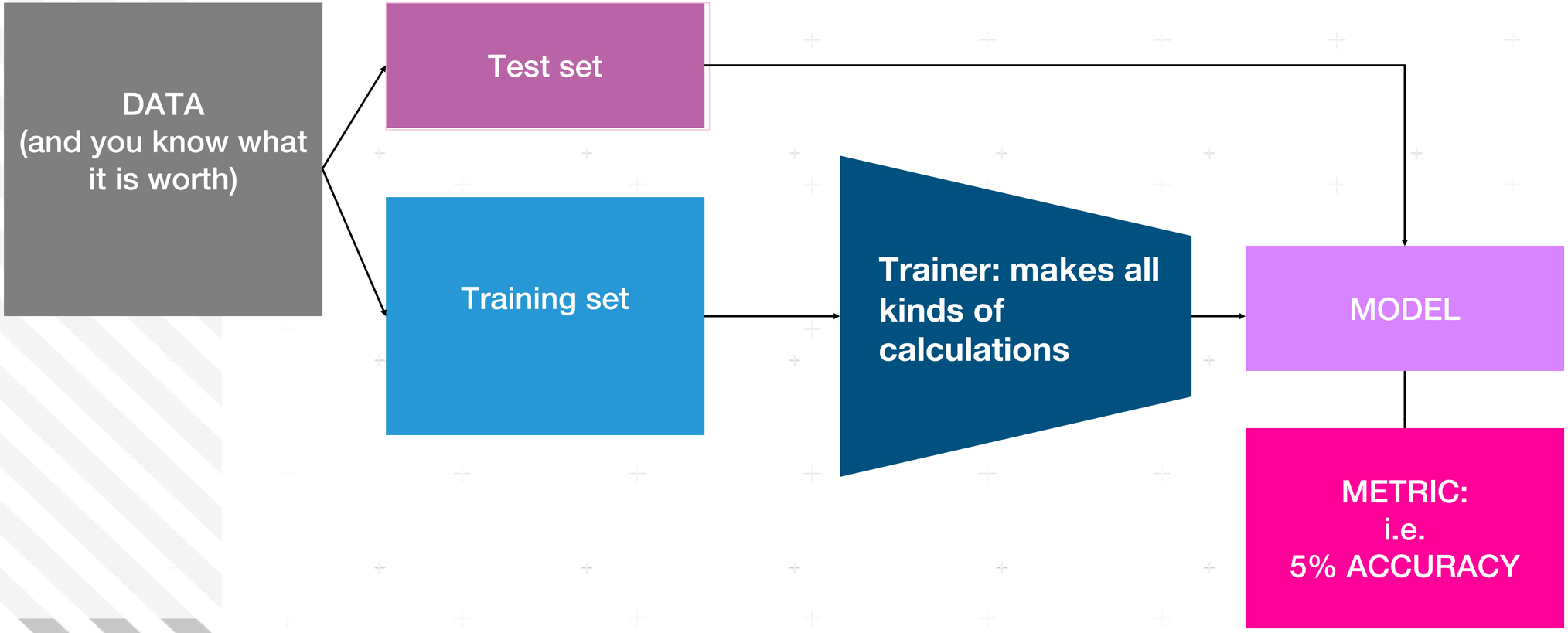
HOW PEOPLE BUILD ALGORITHMS (ONE OPTION)



HOW PEOPLE BUILD ALGORITHMS (ONE OPTION)



HOW PEOPLE BUILD ALGORITHMS (ONE OPTION)



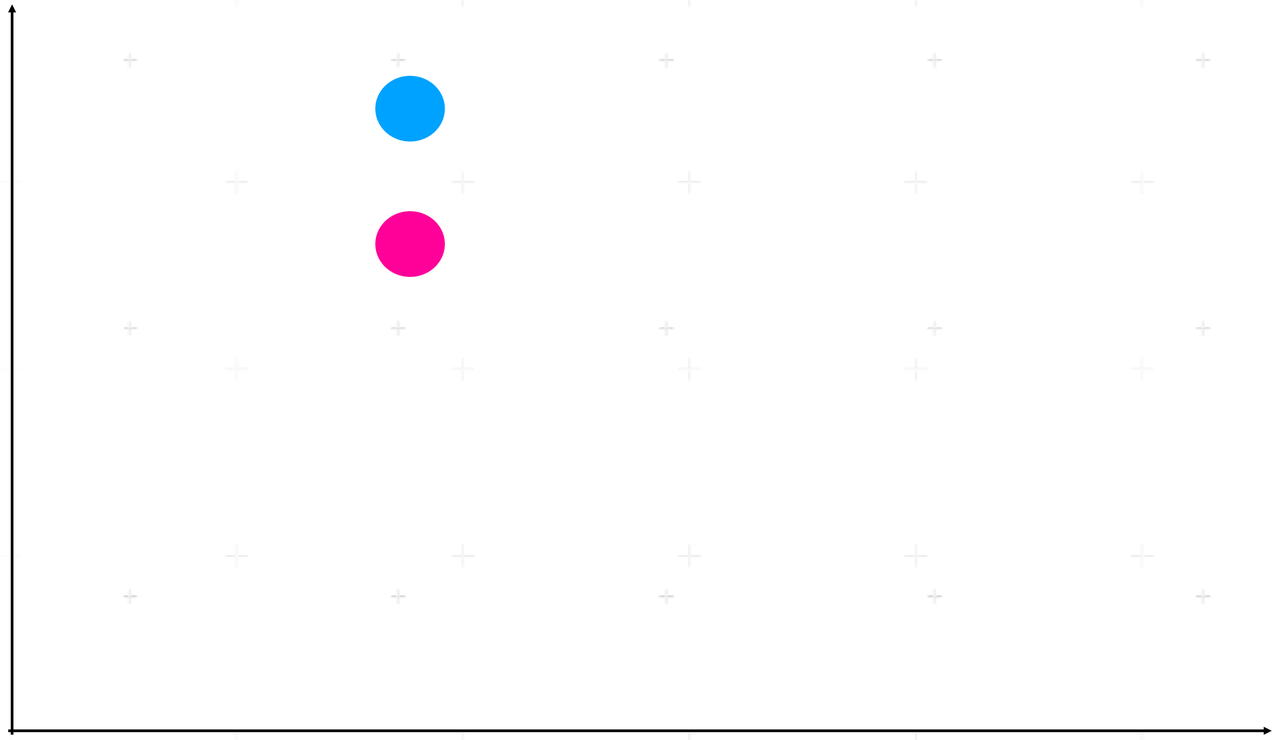
CLASSIC OFFLINE METRICS (WITHOUT USERS)

- **MAE**: Mean Absolute Error (the lower the better)
- **RMSE**: Root Mean Square Error (ditto, plus: higher penalty for high errors)
- **HR**: Hit Rate (the higher the better)
  - **cHR**: Cumulative hit Rate, - **rHR**: Rating Hit Rate

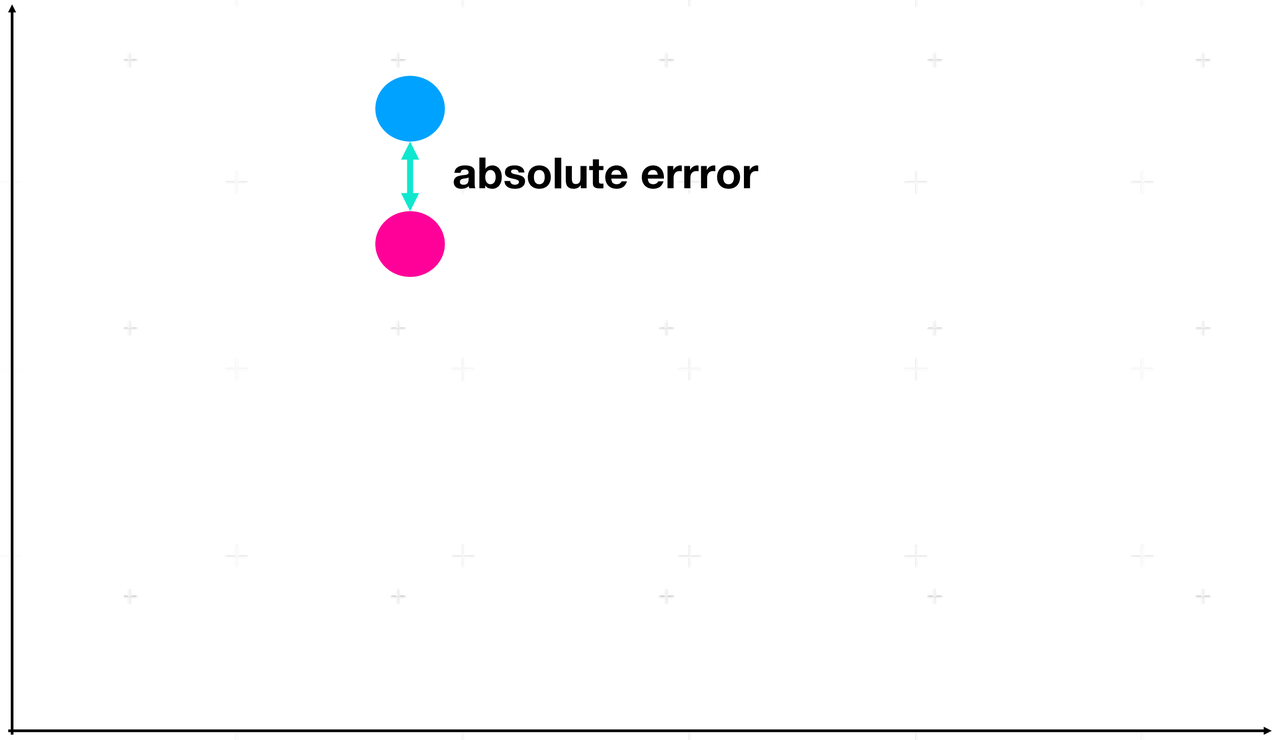
- **Coverage:** measures the ability to recommend long tail items
- **Diversity:** measures the variety of the items
- **Novelty**: how new are the recommendations? (trade-off with trust)
- **Churn:** how sensitive are recommendations to new data
- **Responsiveness**: how sensitive are recommendations to new behavior?

https://medium.com/fnplus/evaluating-recommender-systems-with-python-code-ae0c370c90be

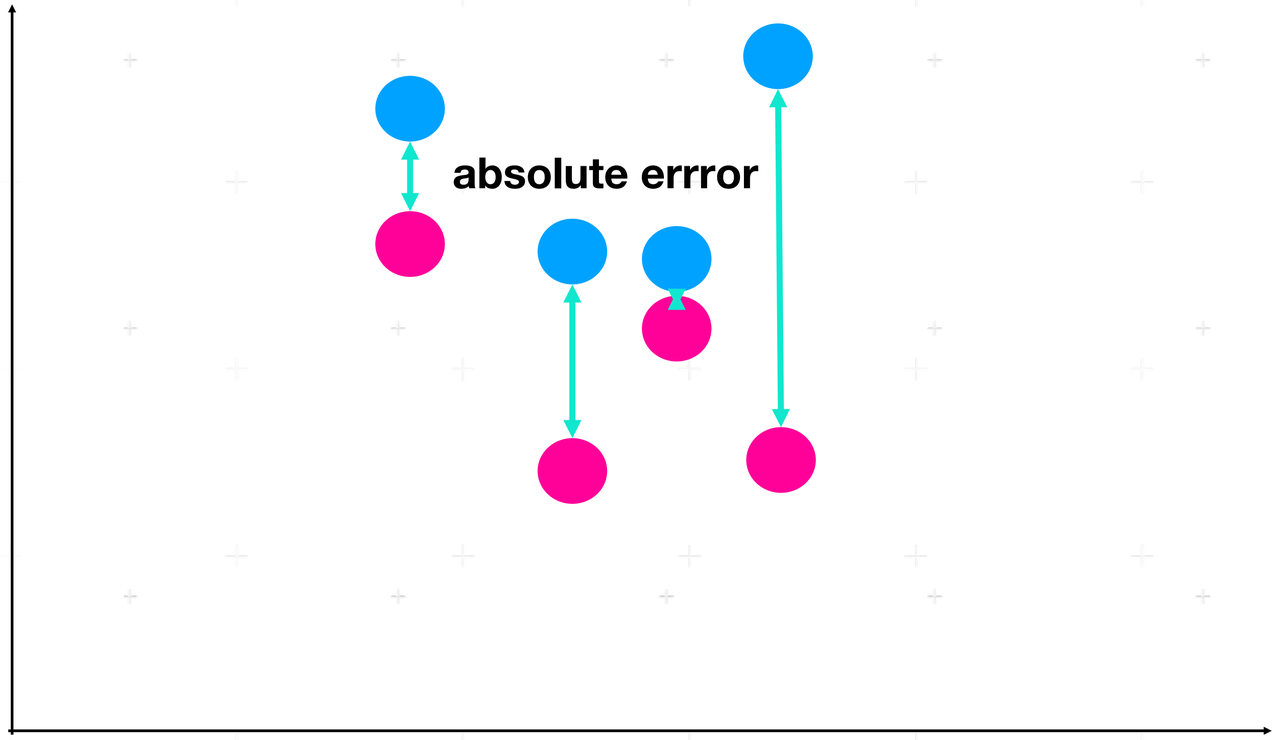
MAE



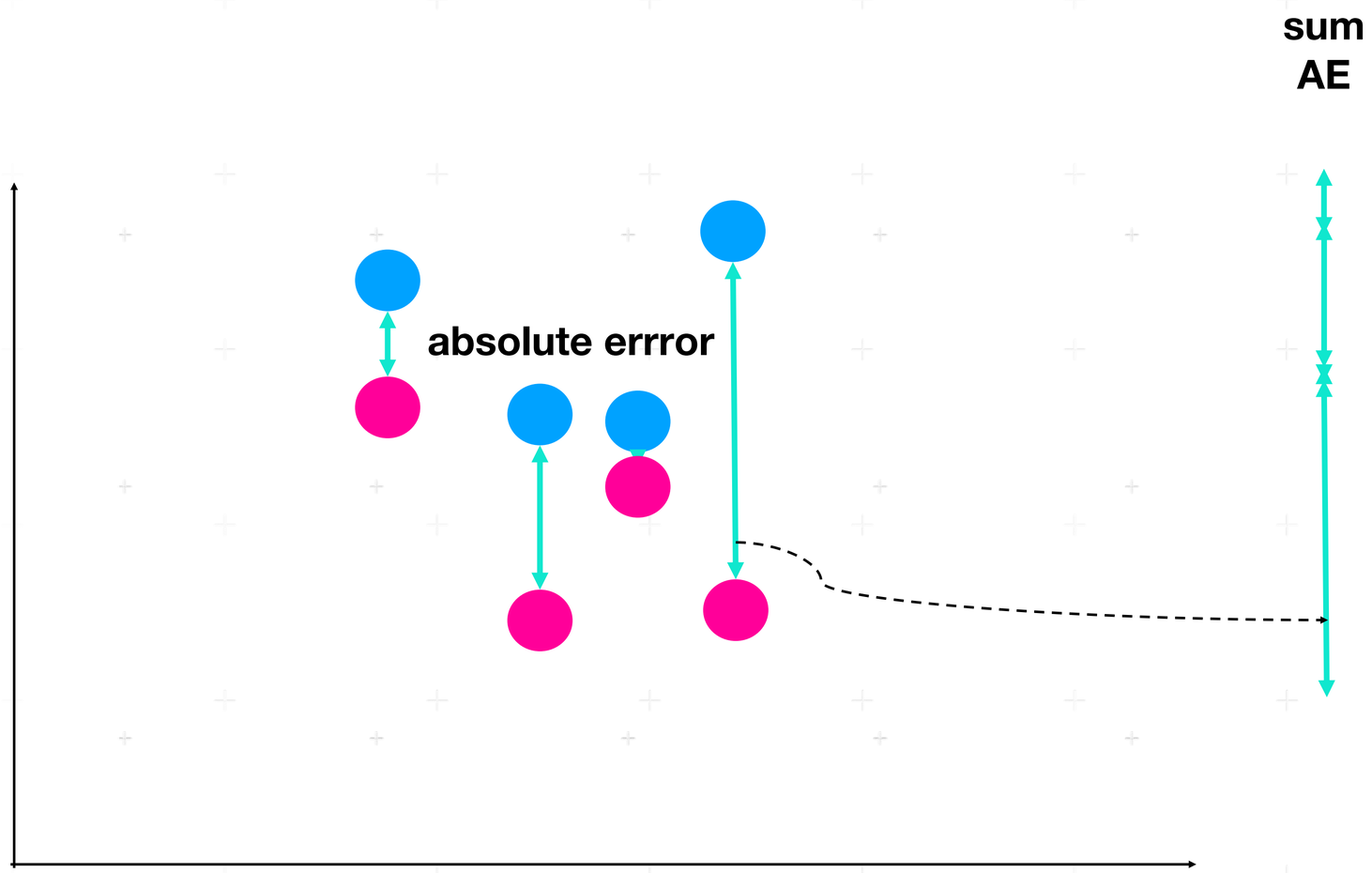
MAE



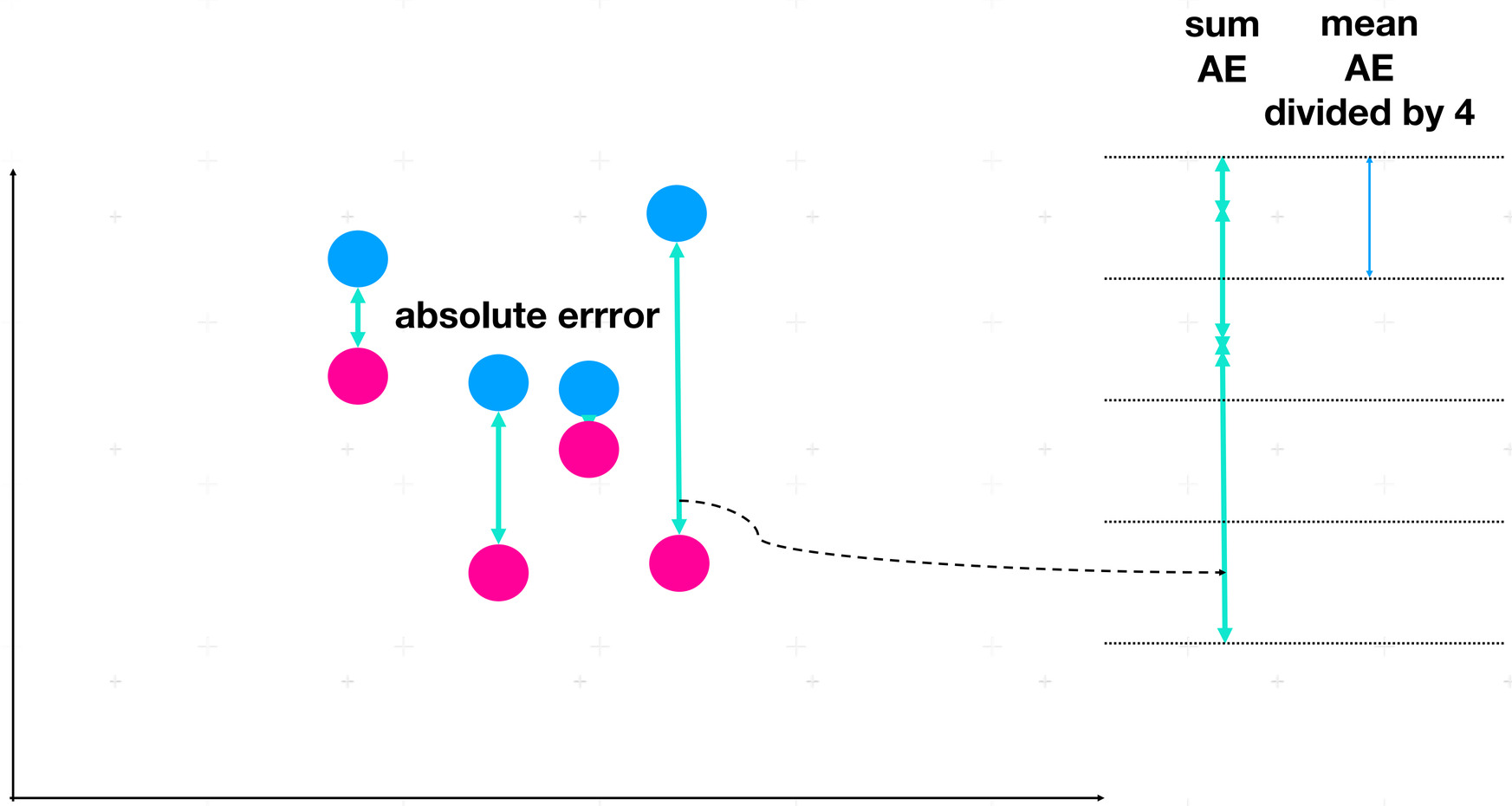
MAE



MAE



MAE



CLASSIC OFFLINE METRICS

- **MAE**: Mean Absolute Error (the lower the better)
- **RMSE**: Root Mean Square Error (ditto, plus: higher penalty for high errors)
- **HR**: Hit Rate (the higher the better)
  - **cHR**: Cumulative hit Rate, - **rHR**: Rating Hit Rate

- **Coverage:** measures the ability to recommend long tail items
- **Diversity:** measures the variety of the items
- **Novelty**: how new are the recommendations? (trade-off with trust)
- **Churn:** how sensitive are recommendations to new data
- **Responsiveness**: how sensitive are recommendations to new behavior?

https://medium.com/fnplus/evaluating-recommender-systems-with-python-code-ae0c370c90be

CLASSIC OFFLINE METRICS

Lower or higher score

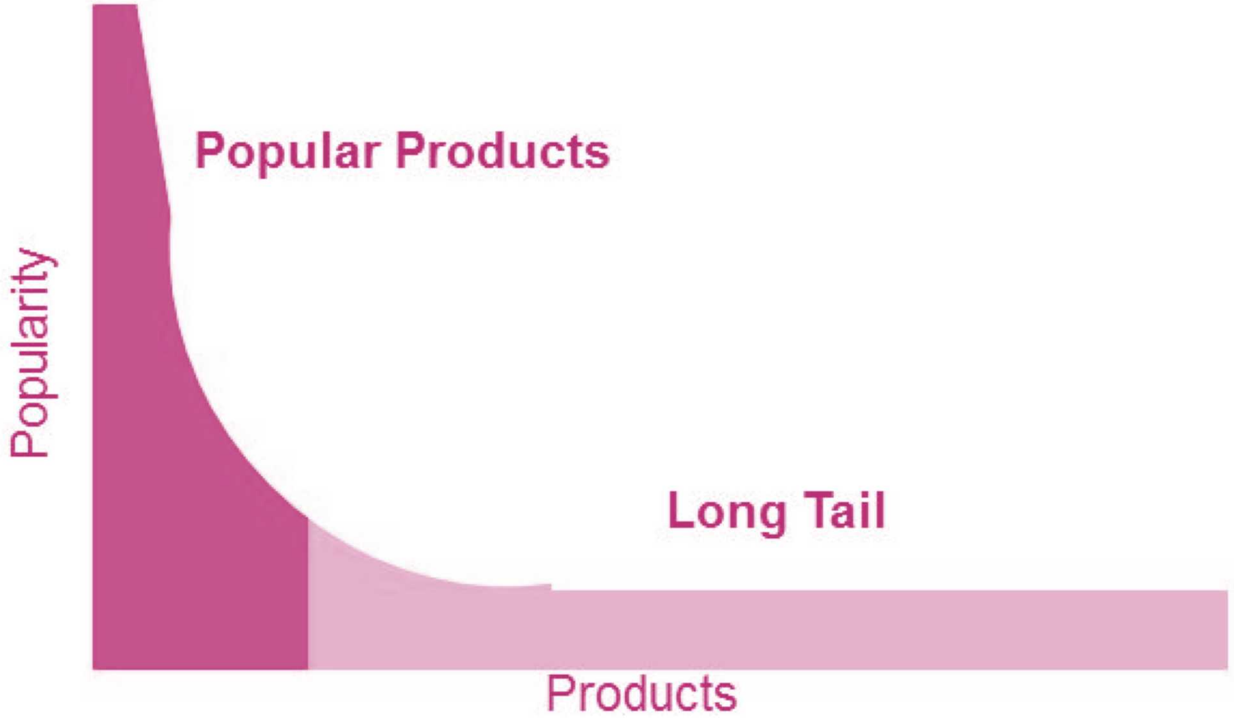
- **MAE**: Mean Absolute Error (the lower the better)
- **RMSE**: Root Mean Square Error (ditto, plus: higher penalty for high errors)
- **HR**: Hit Rate (the higher the better)
  - **cHR**: Cumulative hit Rate, - **rHR**: Rating Hit Rate

Strike a balance

- **Coverage:** measures the ability to recommend long tail items
- **Diversity:** measures the variety of the items
- **Novelty**: how new are the recommendations? (trade-off with trust)
- **Churn:** how sensitive are recommendations to new data
- **Responsiveness**: how sensitive are recommendations to new behavior?

https://medium.com/fnplus/evaluating-recommender-systems-with-python-code-ae0c370c90be

LONG TAIL



CLASSIC OFFLINE METRICS

Lower or higher score

- **MAE**: Mean Absolute Error (the lower the better)
- **RMSE**: Root Mean Square Error (ditto, plus: higher penalty for high errors)
- **HR**: Hit Rate (the higher the better)
  - **cHR**: Cumulative hit Rate, - **rHR**: Rating Hit Rate

Strike a balance

- **Coverage:** measures the ability to recommend long tail items
- **Diversity:** measures the variety of the items
- **Novelty**: how new are the recommendations? (trade-off with trust)
- **Churn:** how sensitive are recommendations to new data
- **Responsiveness**: how sensitive are recommendations to new behavior?

https://medium.com/fnplus/evaluating-recommender-systems-with-python-code-ae0c370c90be

ONLINE METRICS ARE MORE IMPORTANT (USER INPUT)

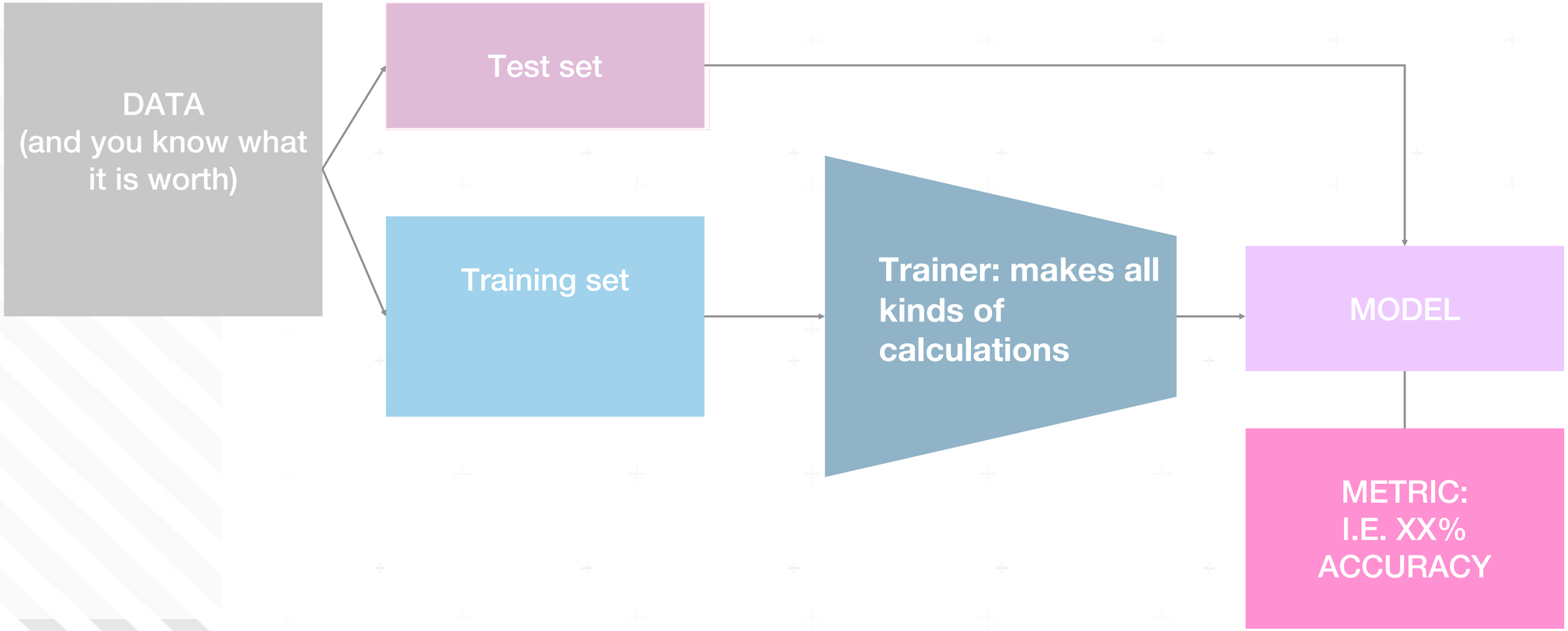
How often do users

- click on your recommendations?
- read/view/listen to them?
- rate them?
- rate them positively?

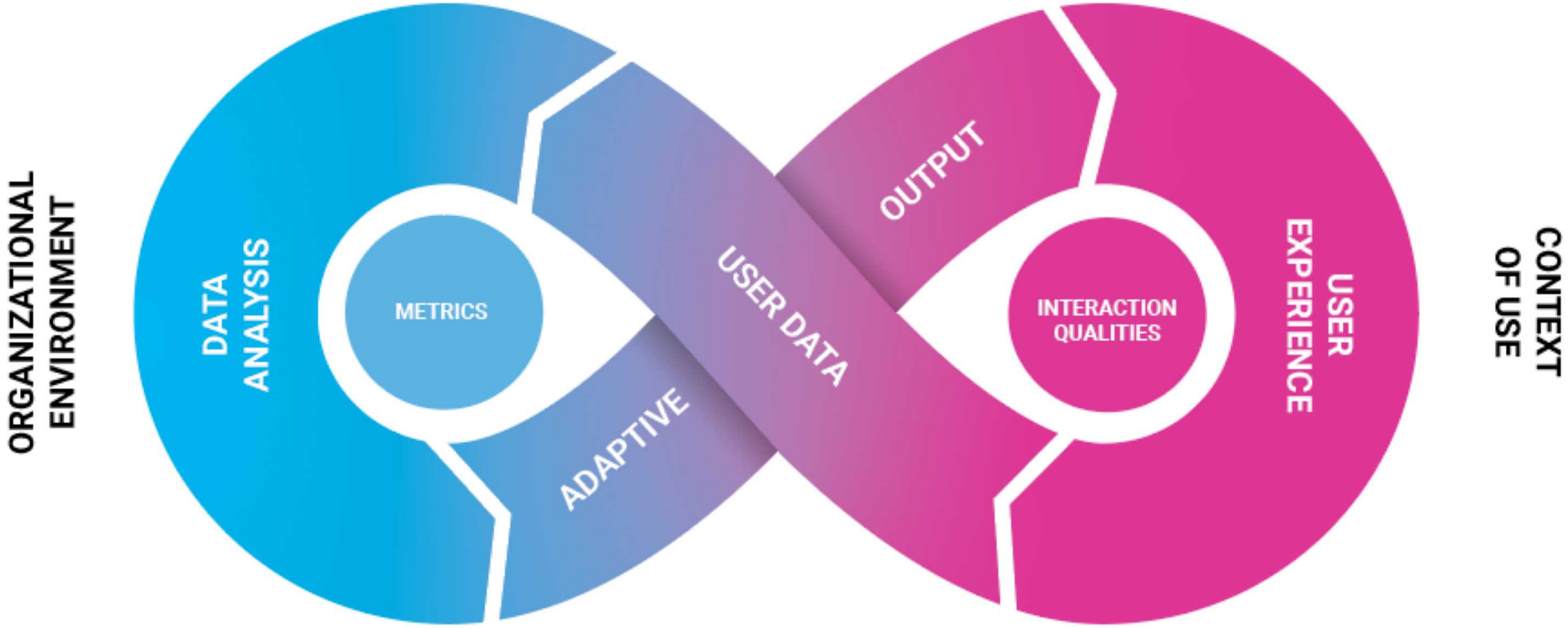
Too much similarity is repetitive and boring

e.g. McNee, S. M., Riedl, J., & Konstan, J. A. (2006, April). Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems* (pp. 1097-1101).

THIS ALSO ASSUMES THAT YOU KNOW ‘CORRECT’ ANSWERS

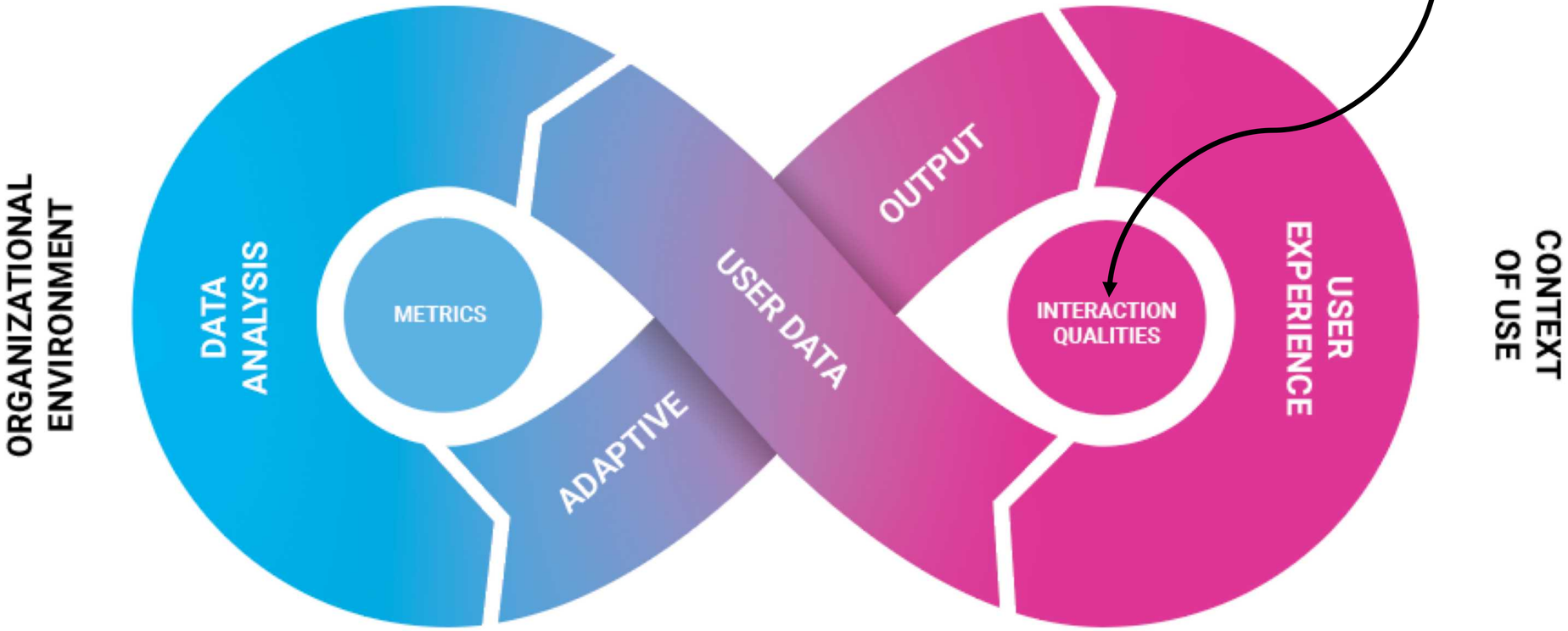


INTERACTION QUALITIES - METRICS



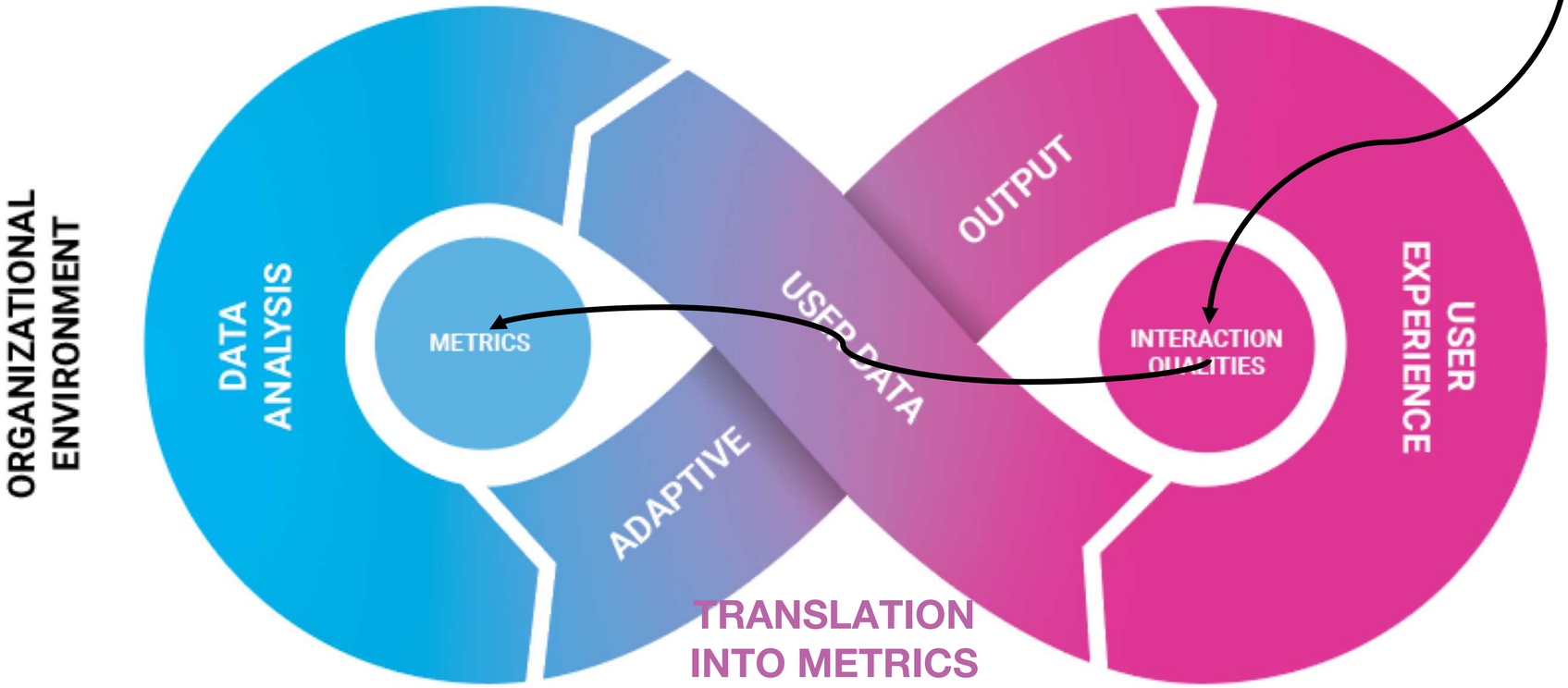
INTERACTION QUALITIES - METRICS

EFFICIENT
SAFE
TRUSTWORTHY
DIVERSE



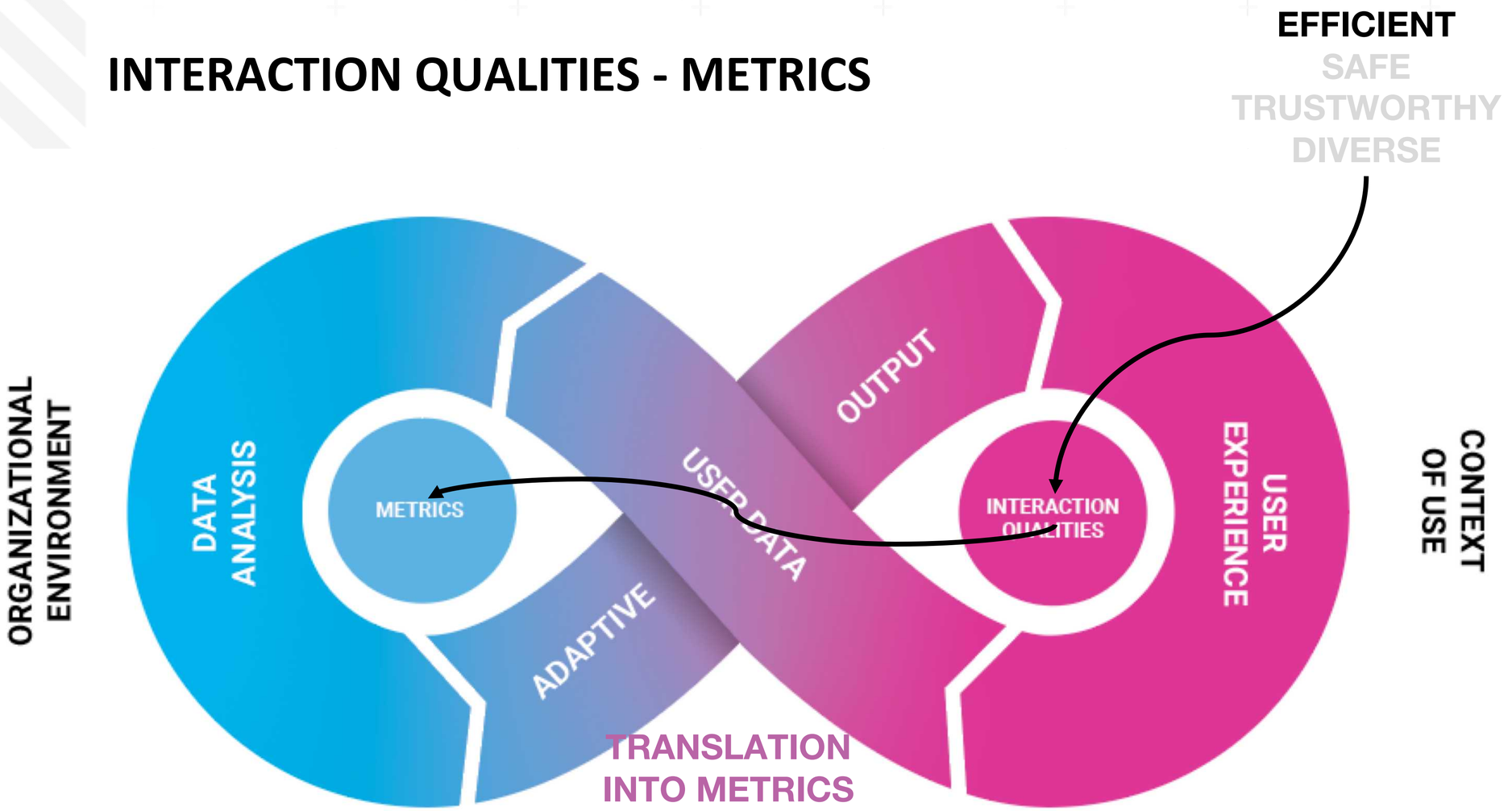
INTERACTION QUALITIES - METRICS

EFFICIENT
SAFE
TRUSTWORTHY
DIVERSE



TRANSLATION
INTO METRICS

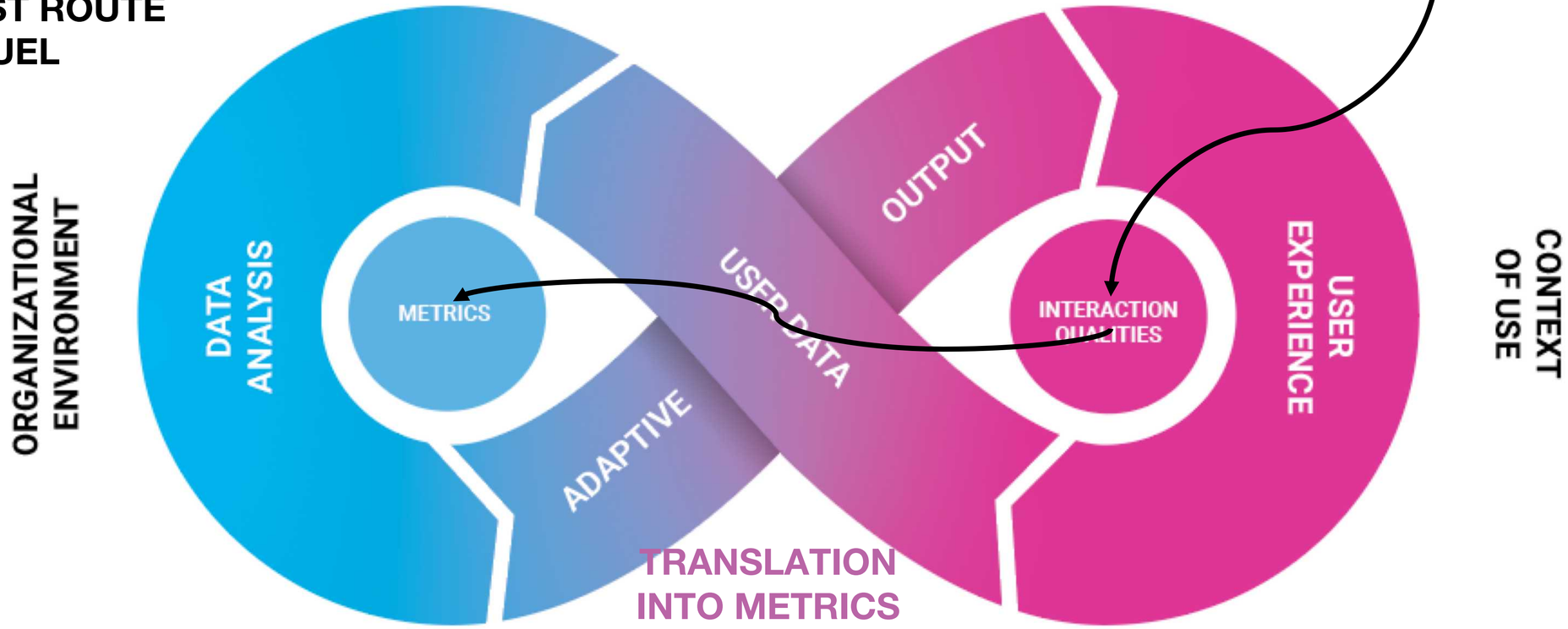
INTERACTION QUALITIES - METRICS



INTERACTION QUALITIES - METRICS

- SHORTEST ROUTE
- QUICKEST ROUTE
- LEAST FUEL

EFFICIENT
SAFE
TRUSTWORTHY
DIVERSE

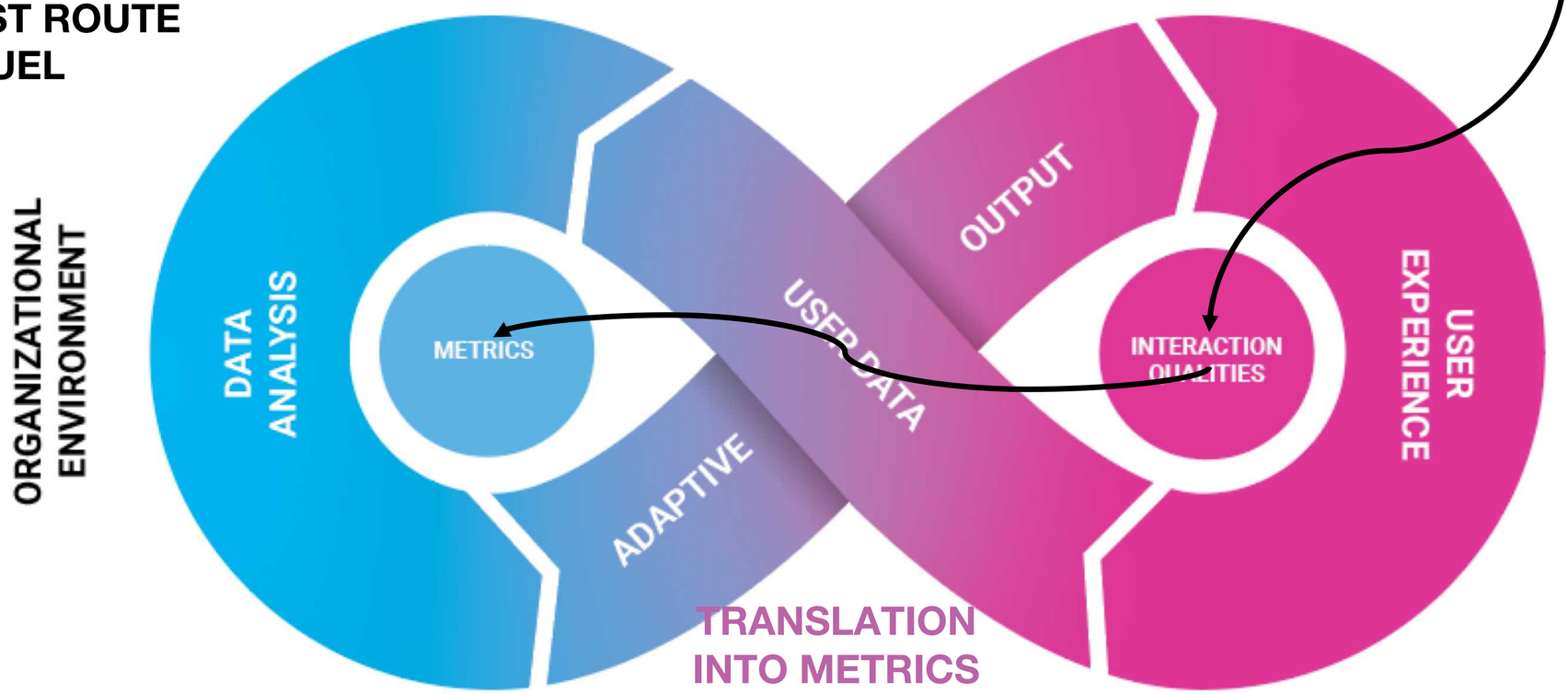


TRANSLATION
INTO METRICS

INTERACTION QUALITIES - METRICS

- SHORTEST ROUTE
- QUICKEST ROUTE
- LEAST FUEL

EFFICIENT
SAFE
TRUSTWORTHY
DIVERSE

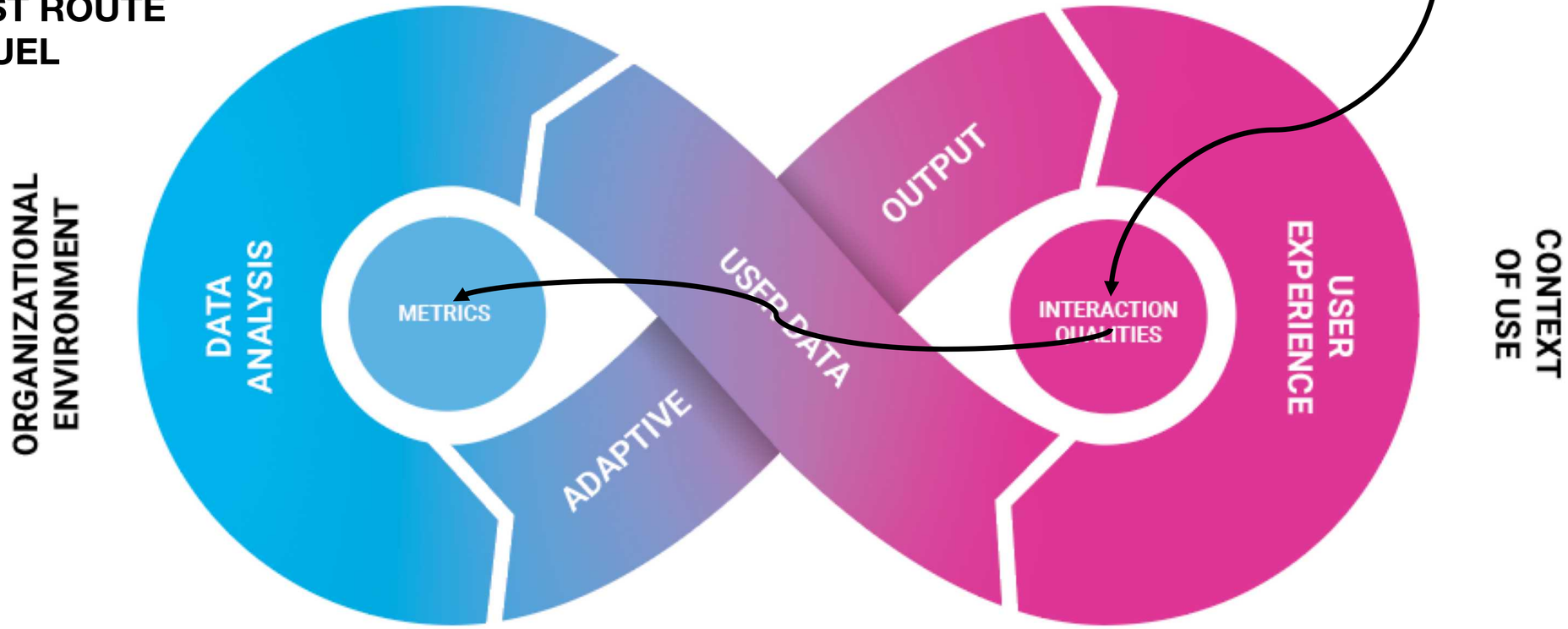


TRANSLATION INTO METRICS

INTERACTION QUALITIES - METRICS

- SHORTEST ROUTE
- QUICKEST ROUTE
- LEAST FUEL

EFFICIENT
SAFE
TRUSTWORTHY
DIVERSE

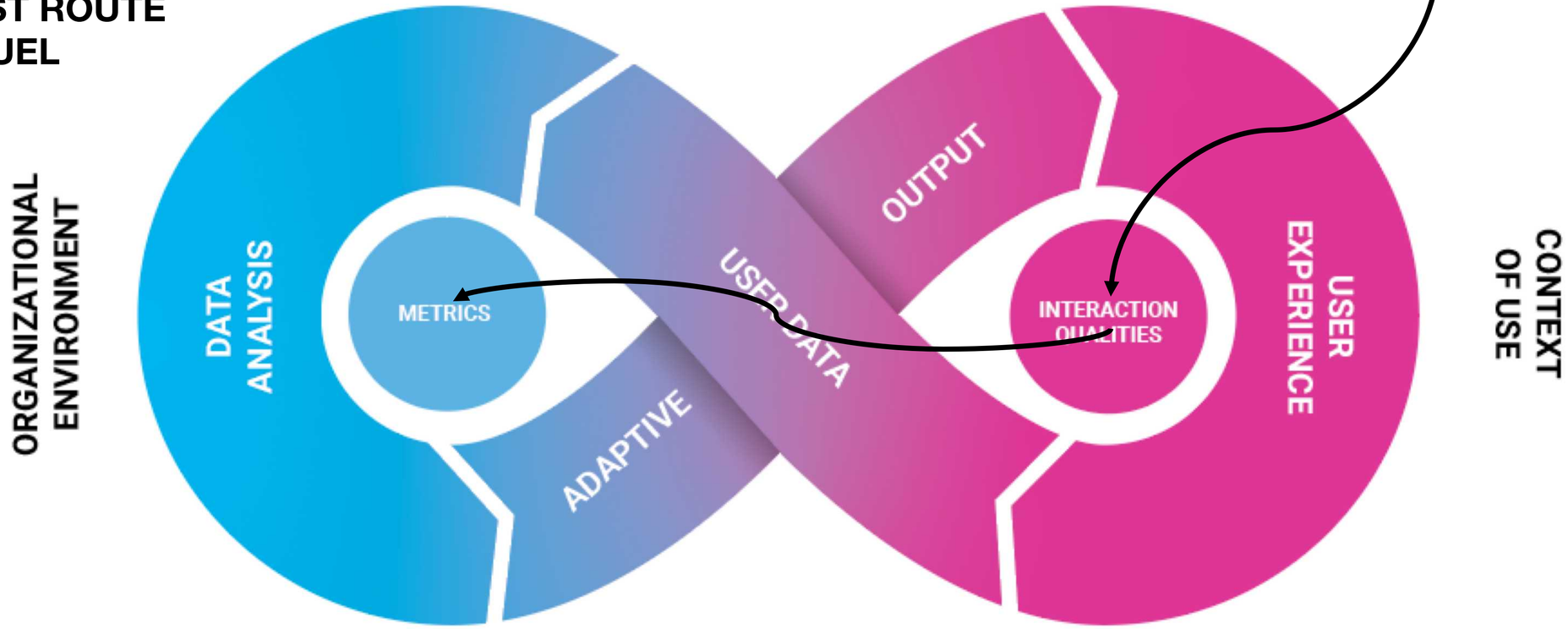


the selection determines relevant data
the selection determines how your analysis steps through the data

INTERACTION QUALITIES - METRICS

EFFICIENT
SAFE
TRUSTWORTHY
DIVERSE

- SHORTEST ROUTE
- QUICKEST ROUTE
- LEAST FUEL

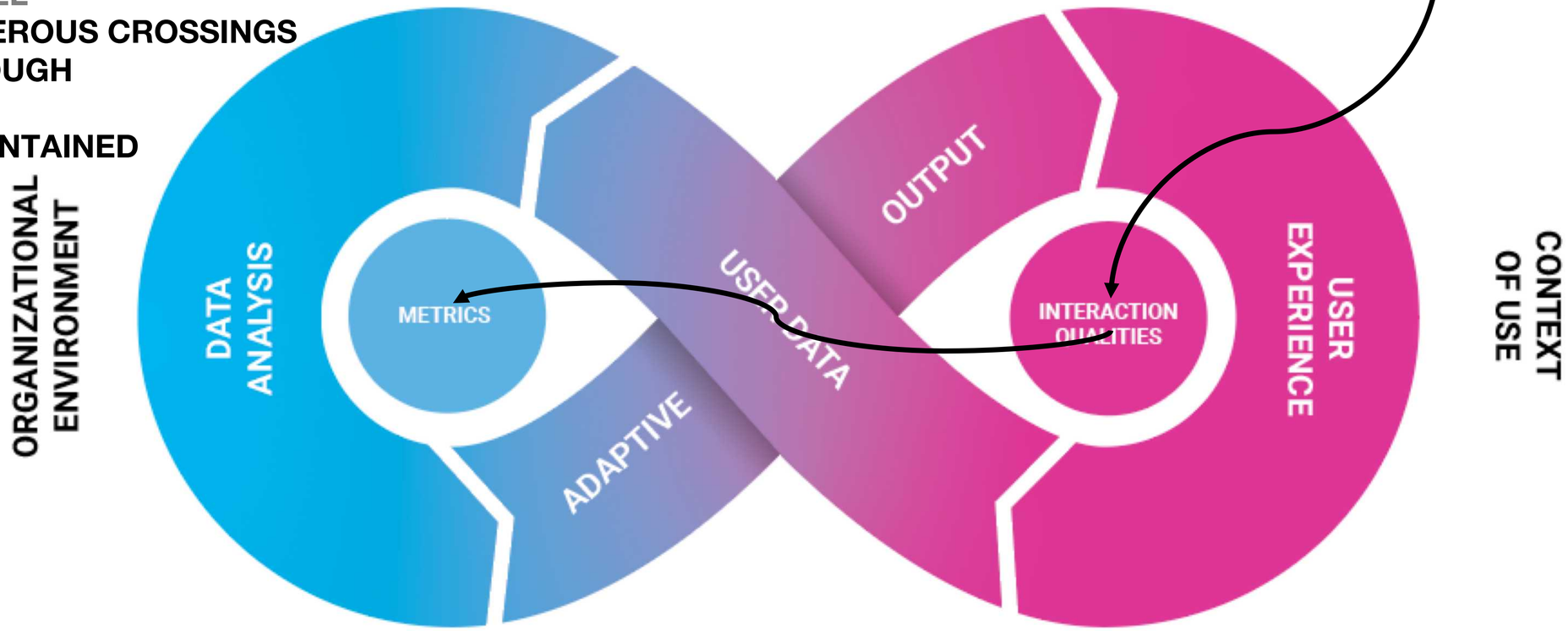


the selection determines relevant data
the selection determines how your analysis steps through the data
determination of weight of each value

INTERACTION QUALITIES - METRICS

- SHORTEST ROUTE
- QUICKEST ROUTE
- LEAST FUEL
- NO DANGEROUS CROSSINGS
- BUSY ENOUGH
- WELL-LIT
- WELL-MAINTAINED

EFFICIENT
SAFE
TRUSTWORTHY
DIVERSE



the selection determines relevant data
the selection determines how your analysis steps through the data
determination of weight of each value

SIDE NOTE: IT IS ALWAYS A TRADE-OFF

ALSO IN THE LITERAR INTERACTION

SIDE NOTE: IT IS ALWAYS A TRADE-OFF

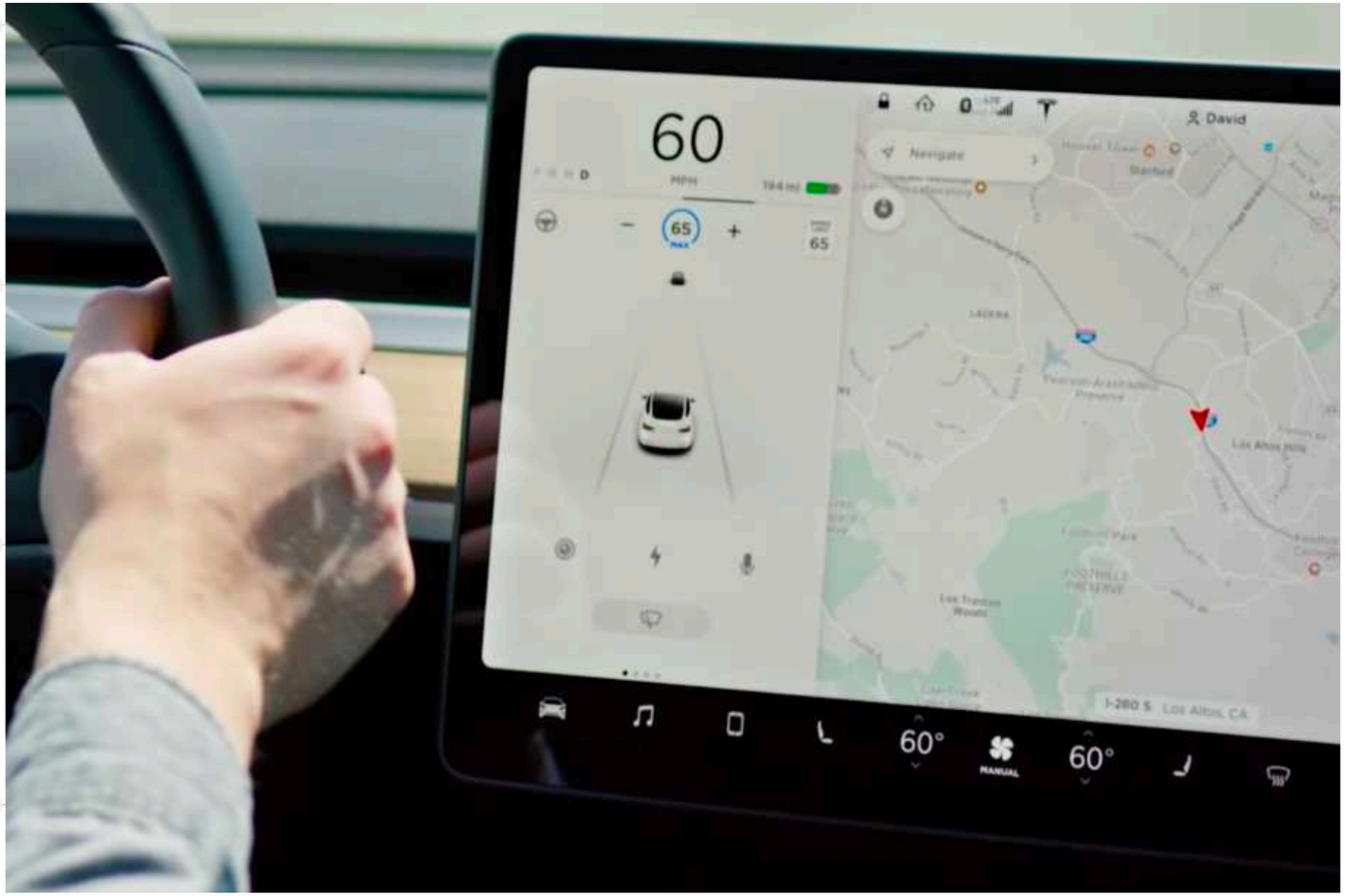
ALSO IN THE LITERAR INTERACTION

Supersafe
Annoyingly inefficient

SIDE NOTE: IT IS ALWAYS A TRADE-OFF

ALSO IN THE LITERAR INTERACTION

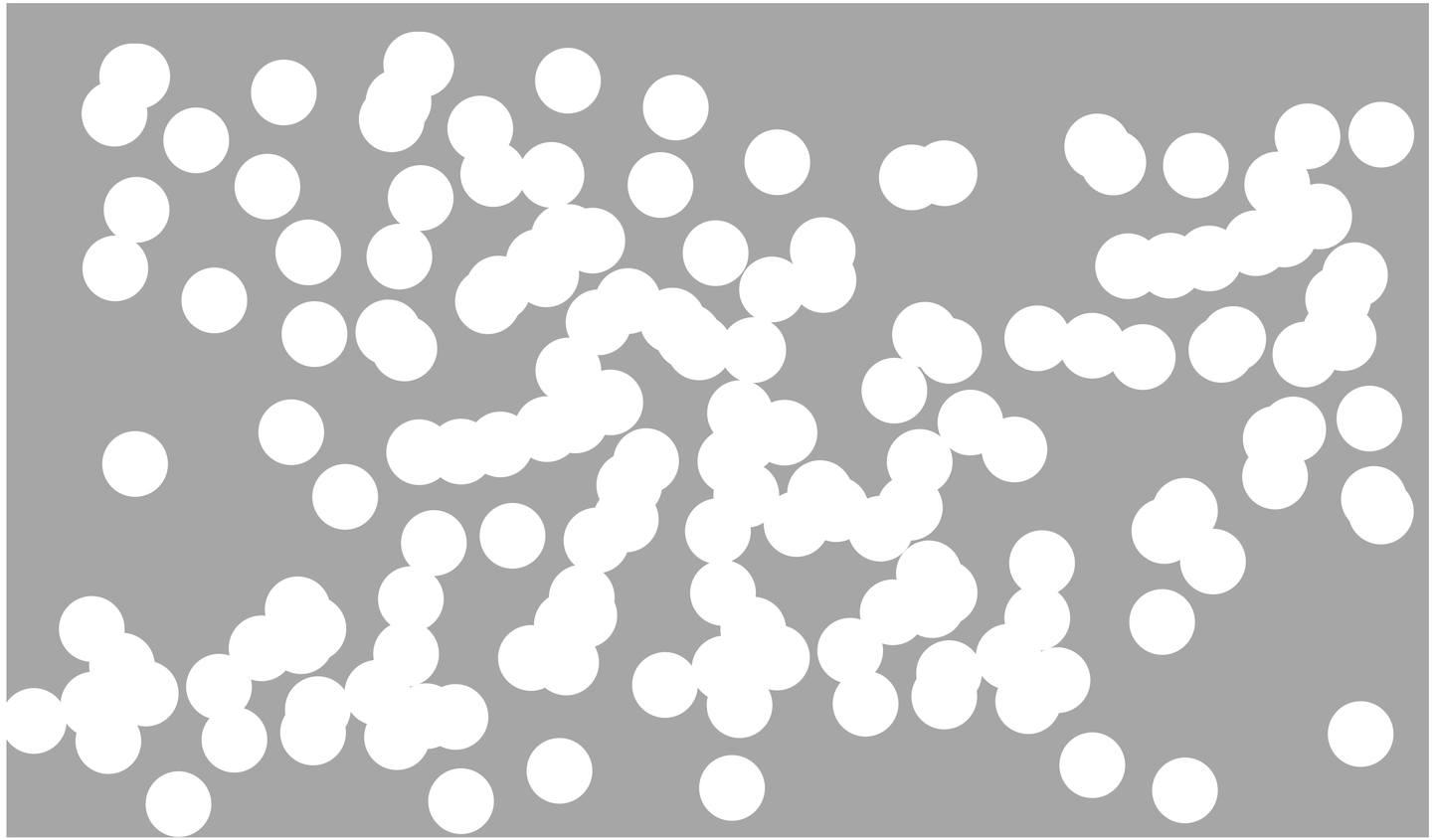
Superefficient
Kind of dangerous



LET’S UNPACK “METRICS”

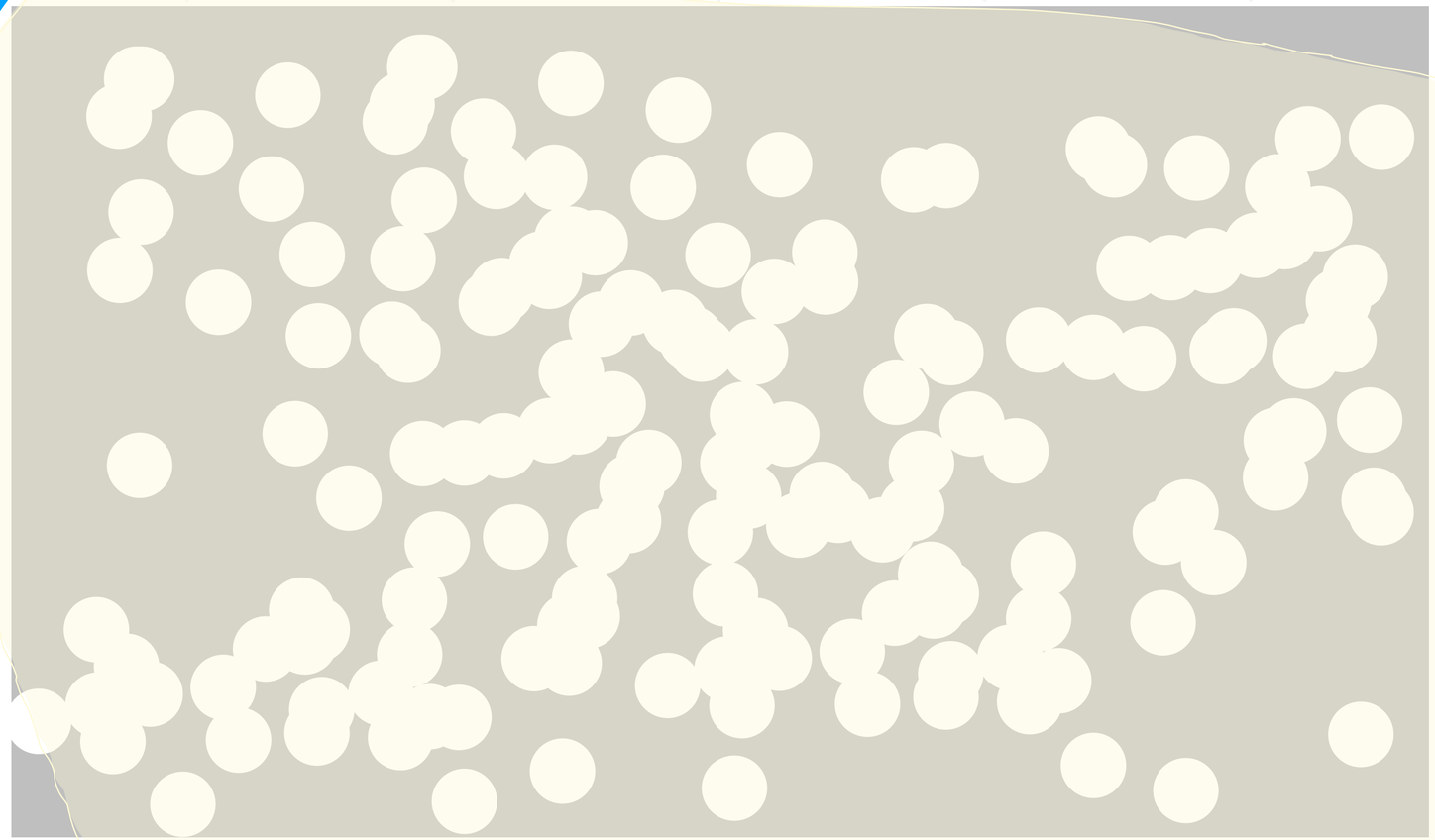
DATA
(and you know what it is worth)

LET’S UNPACK “METRICS”: HERE IS YOUR DATA



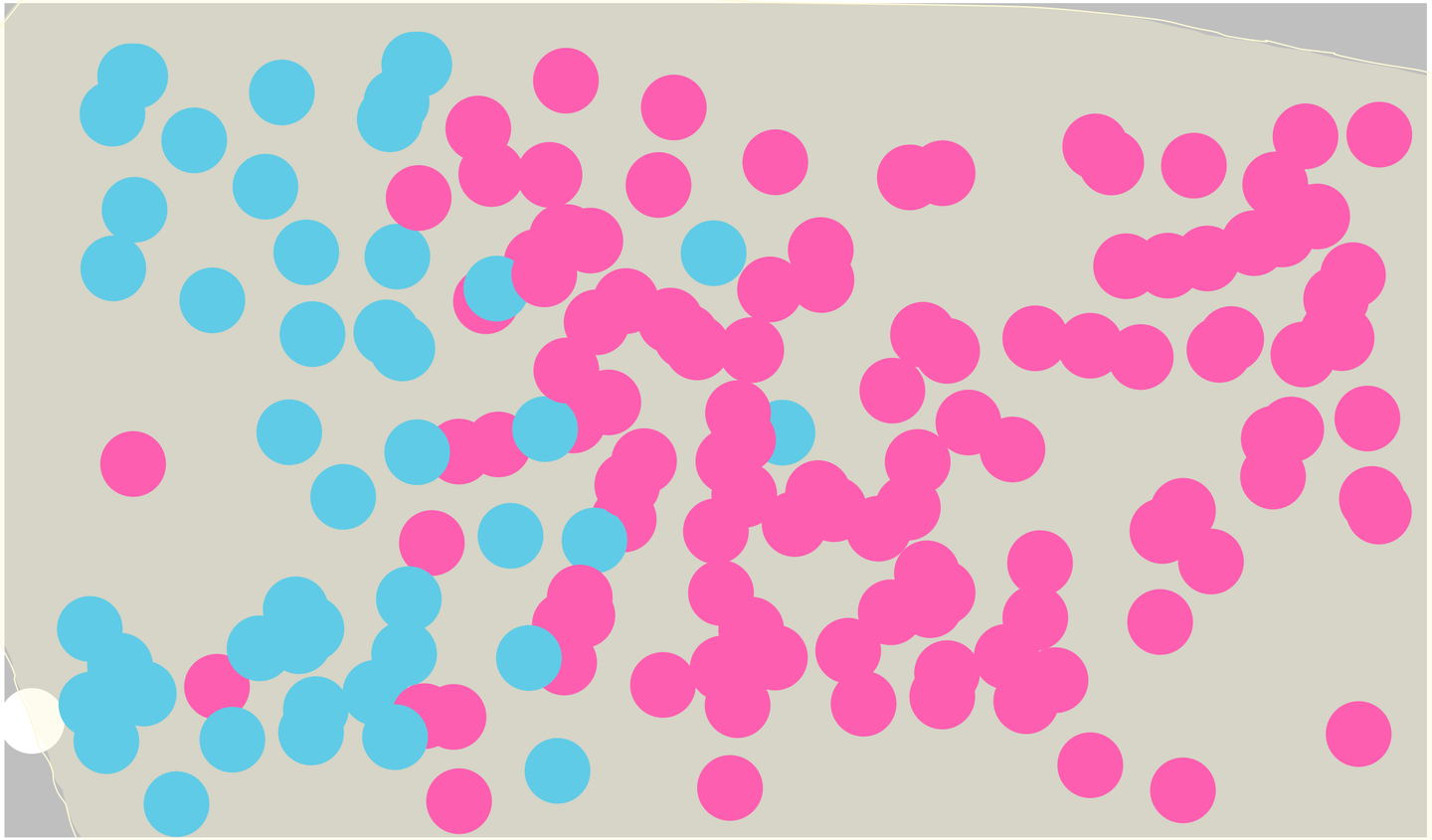
LET’S UNPACK “METRICS”: THIS IS A LENS THAT DETERMINES + OR -

LENS: 1



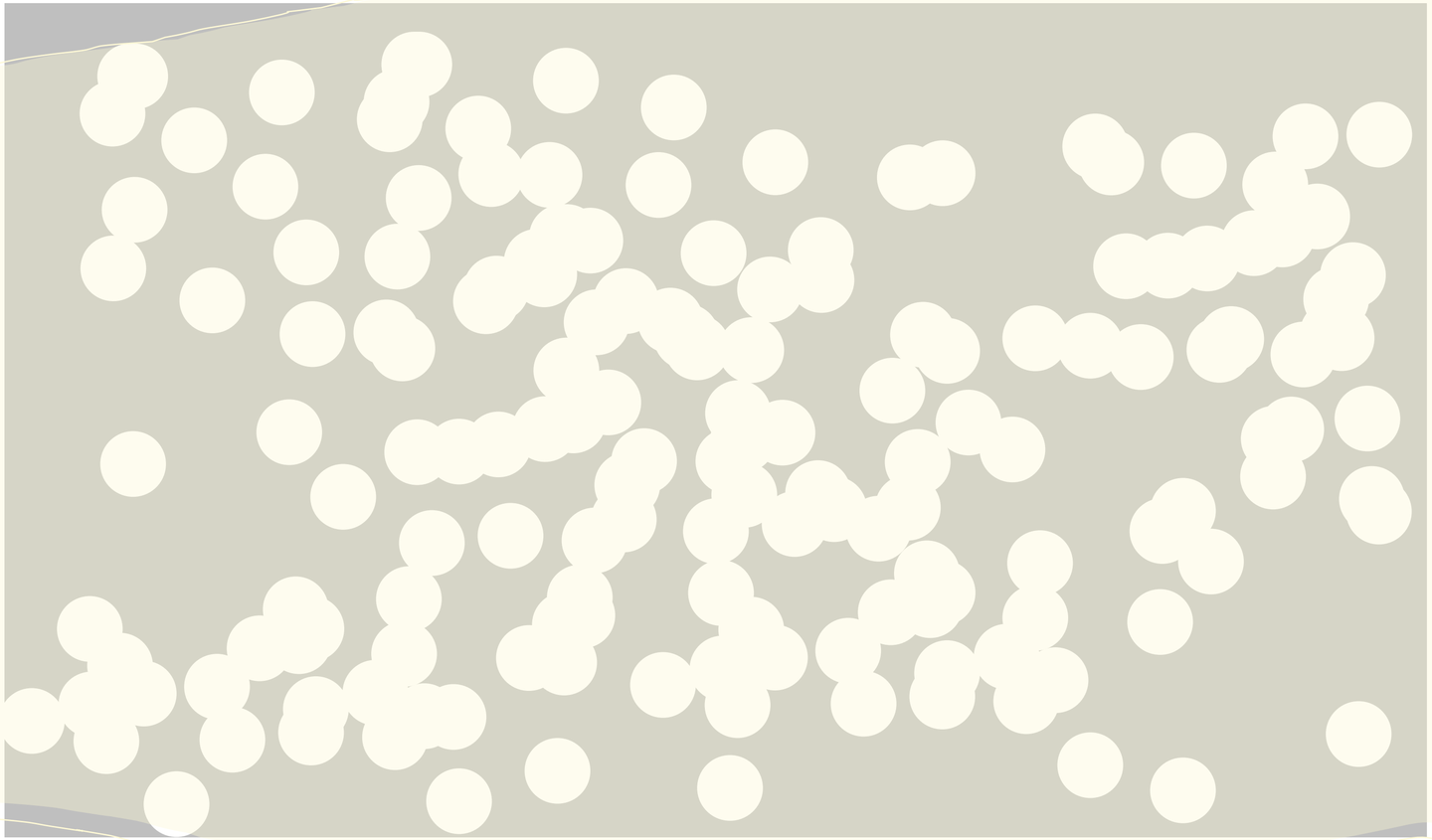
LET’S UNPACK “METRICS”: THIS IS A LENS THAT DETERMINES + OR -

LENS: 1



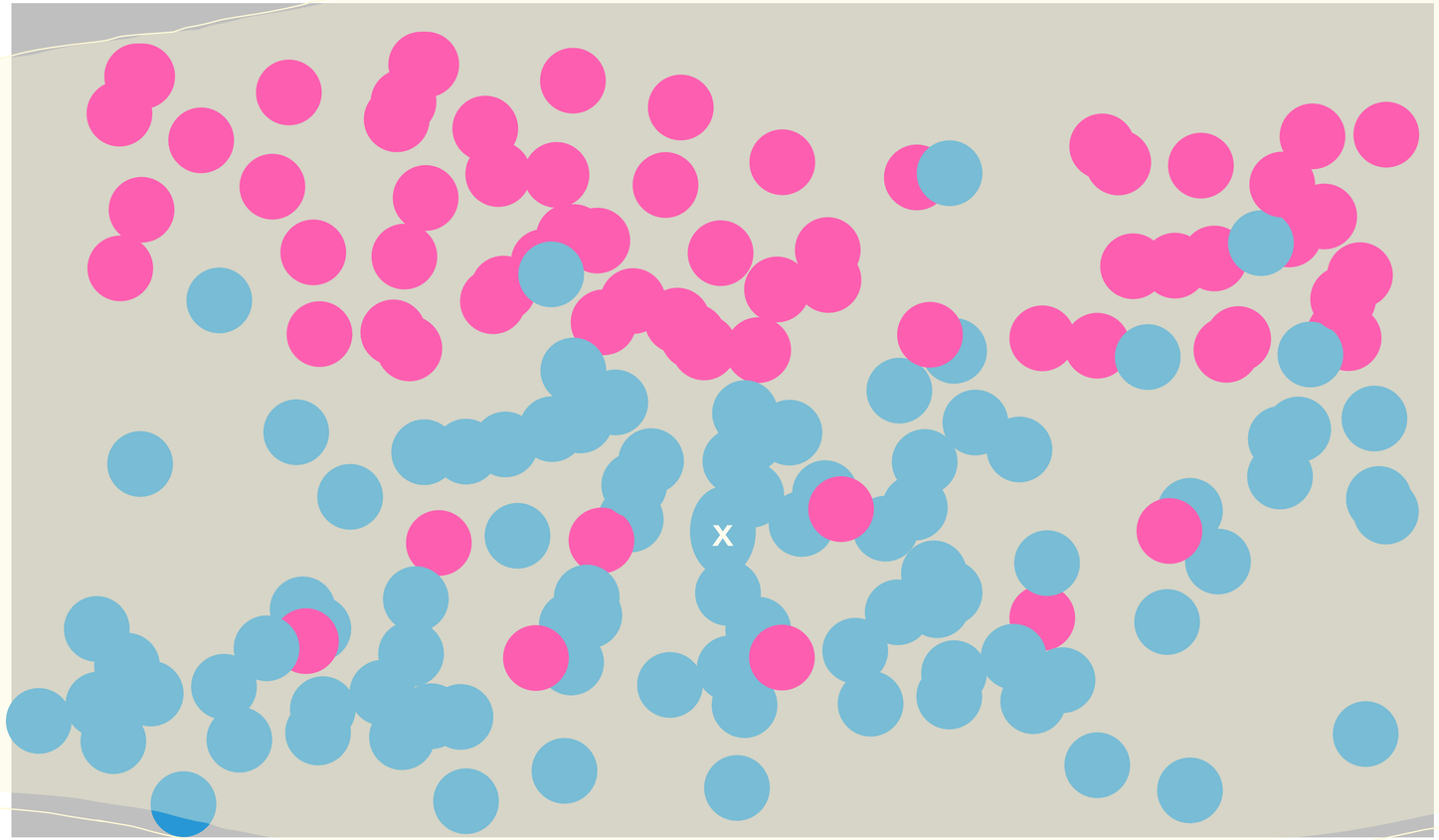
LET’S UNPACK “METRICS”: THIS IS ANOTHER LENS

LENS: 2

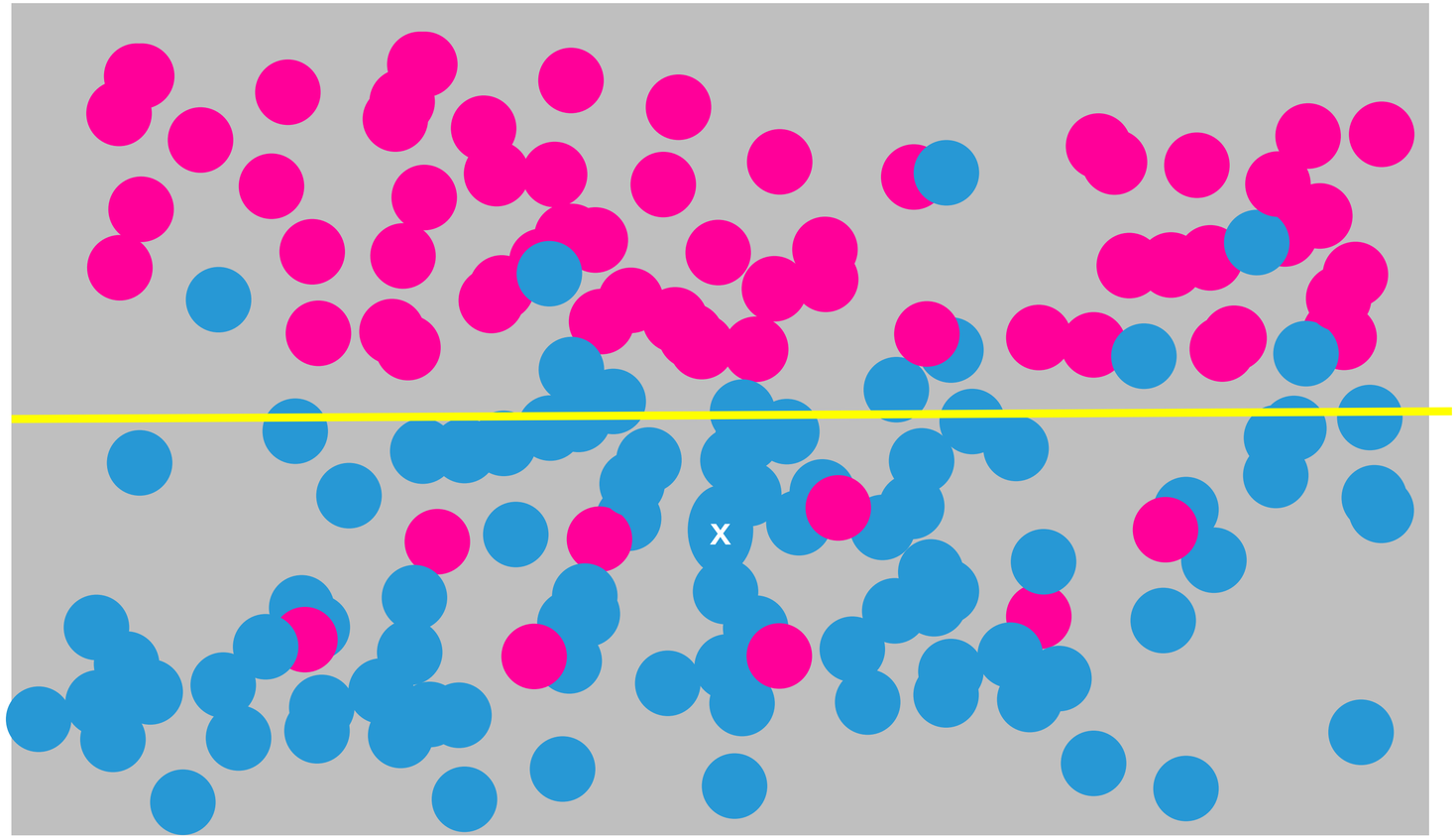


LET’S UNPACK “METRICS”: AND THUS DIFFERENT + AND -

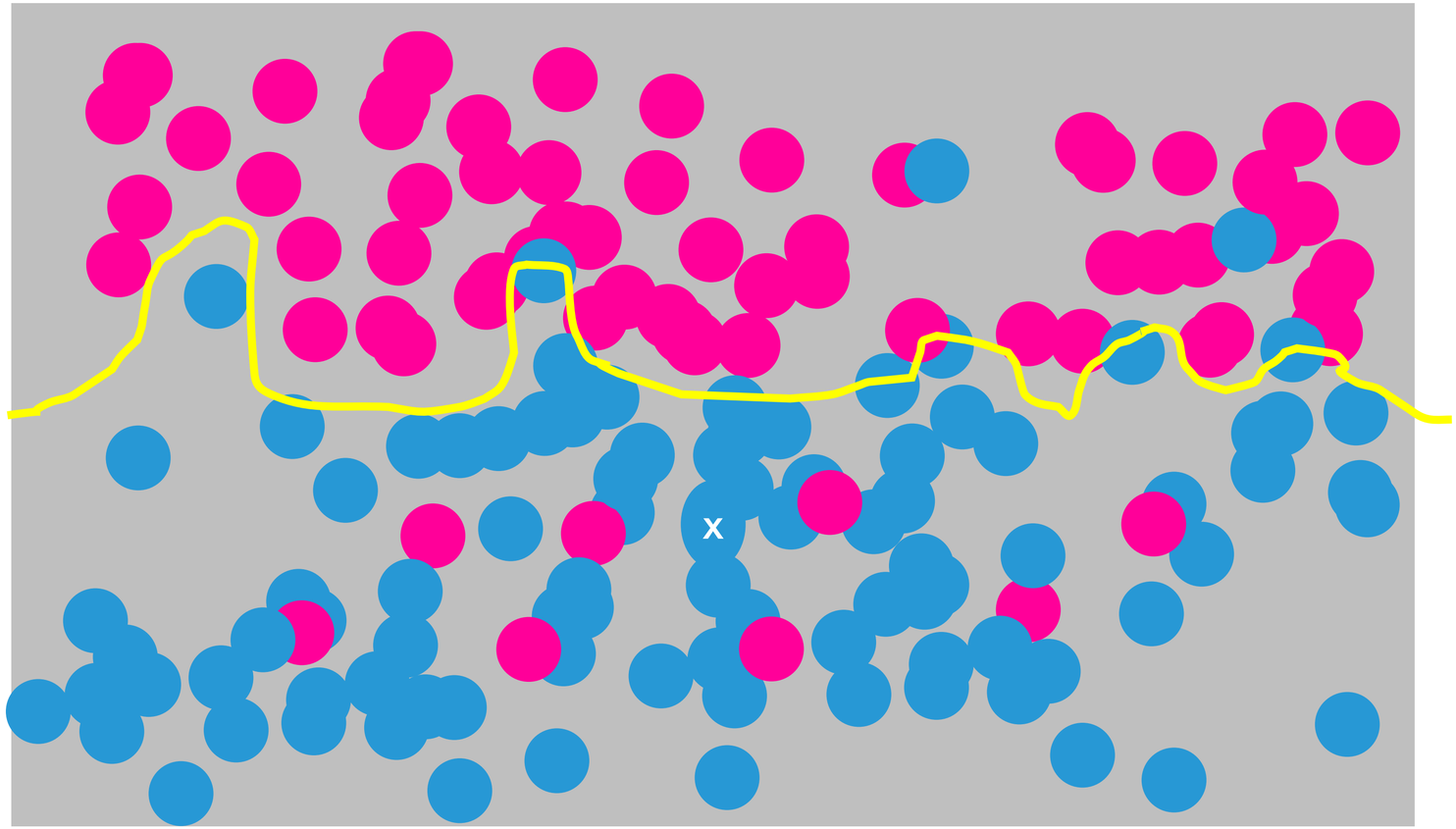
LENS: 2



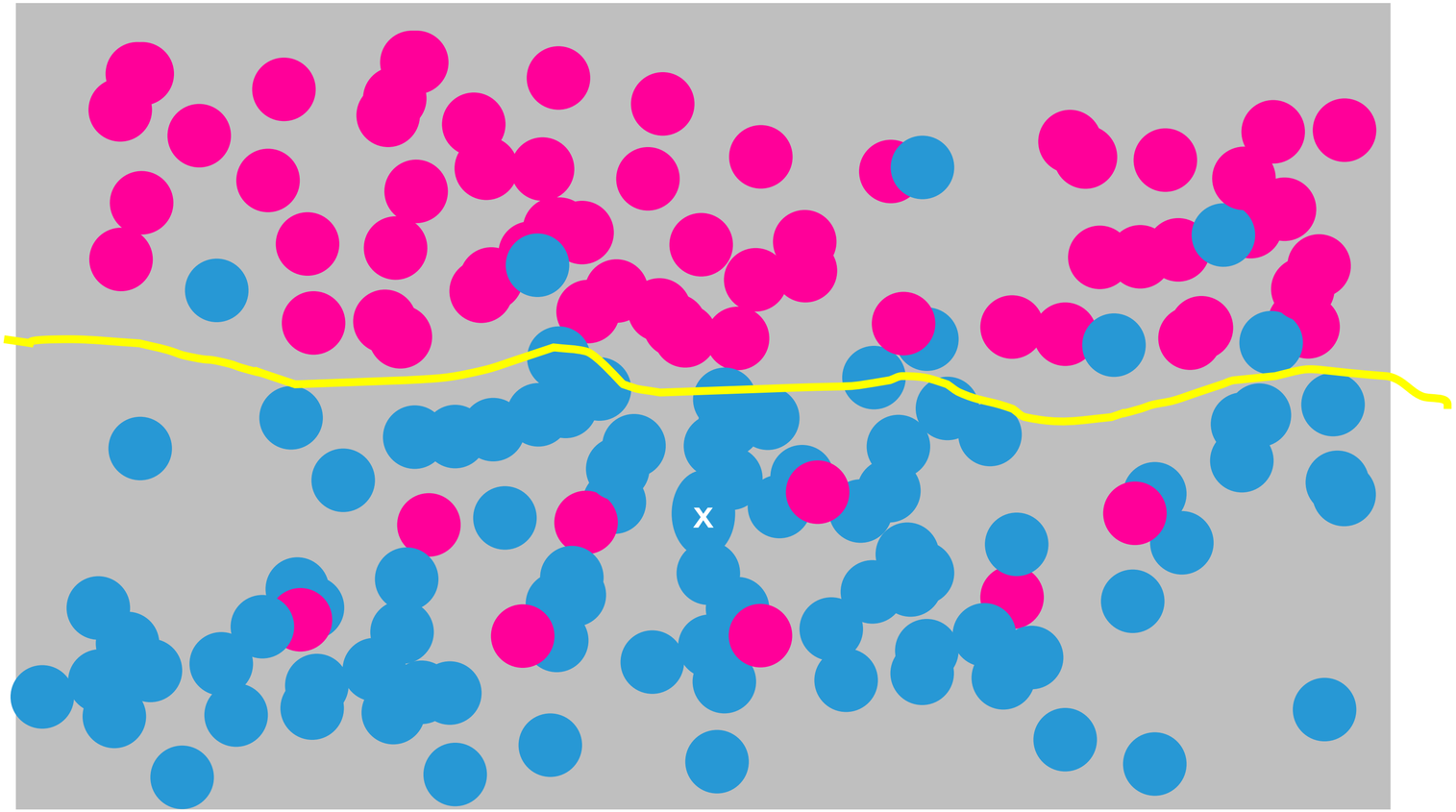
NOW YOU NEED A ‘MODEL’



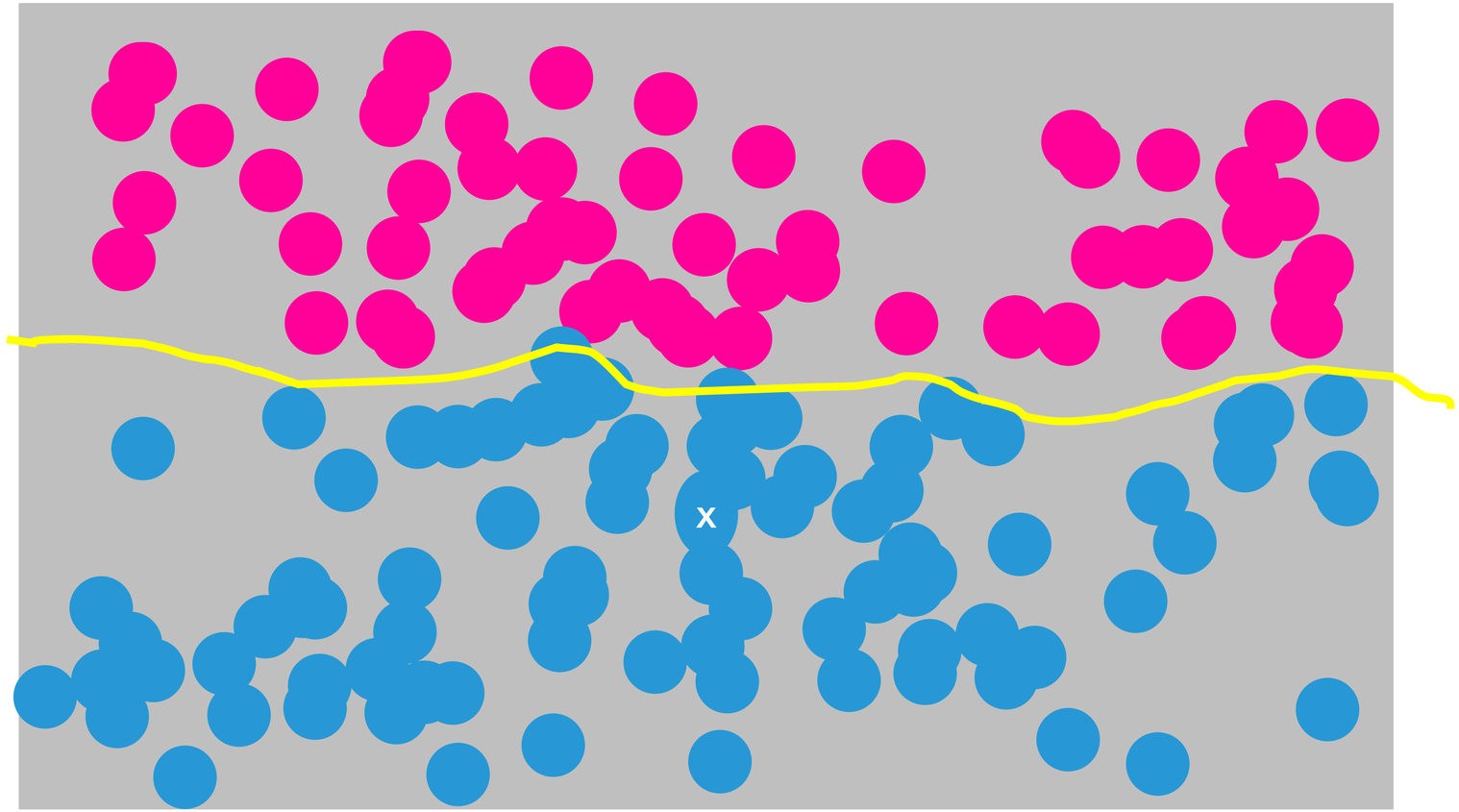
NOW YOU NEED A ‘MODEL’



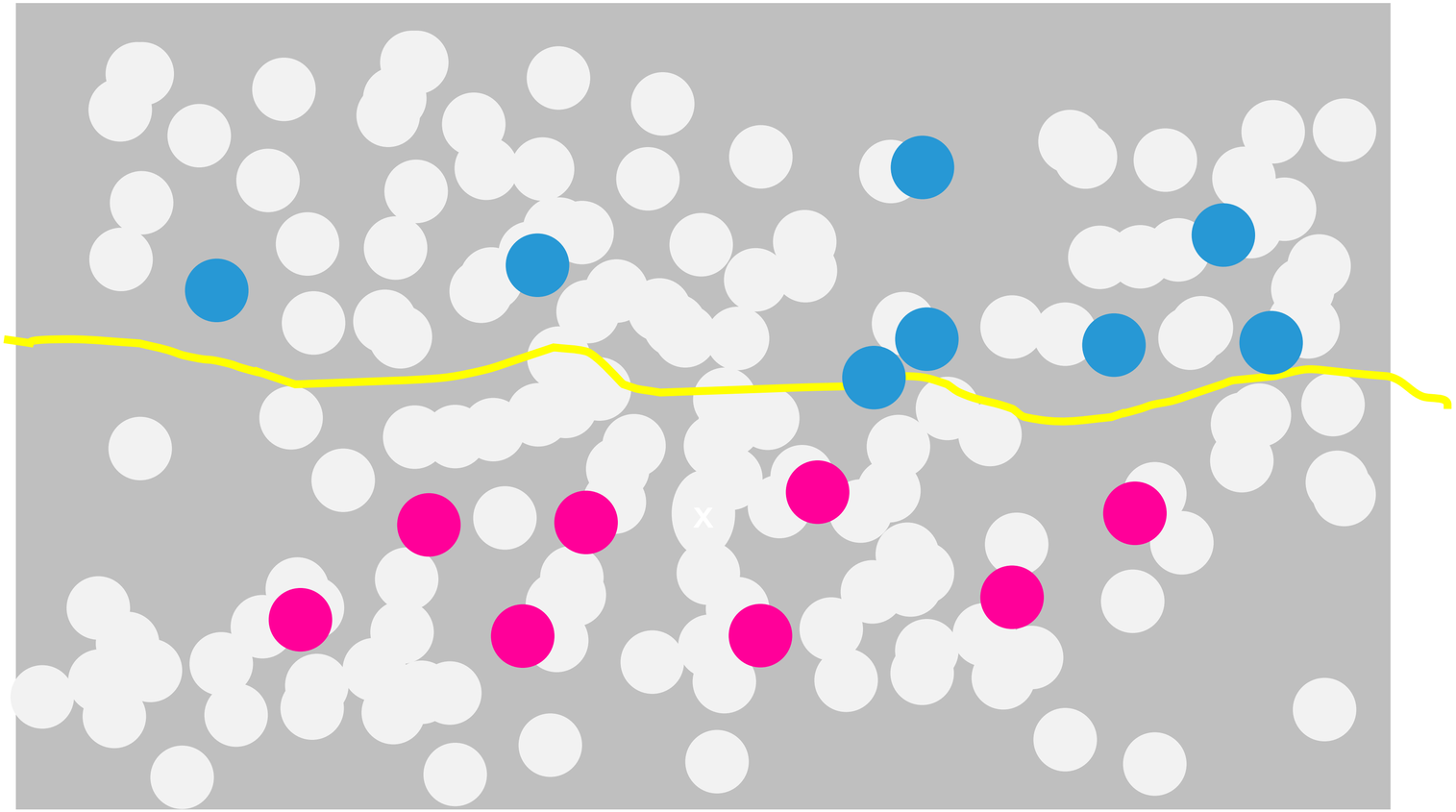
NOW YOU NEED A ‘MODEL’



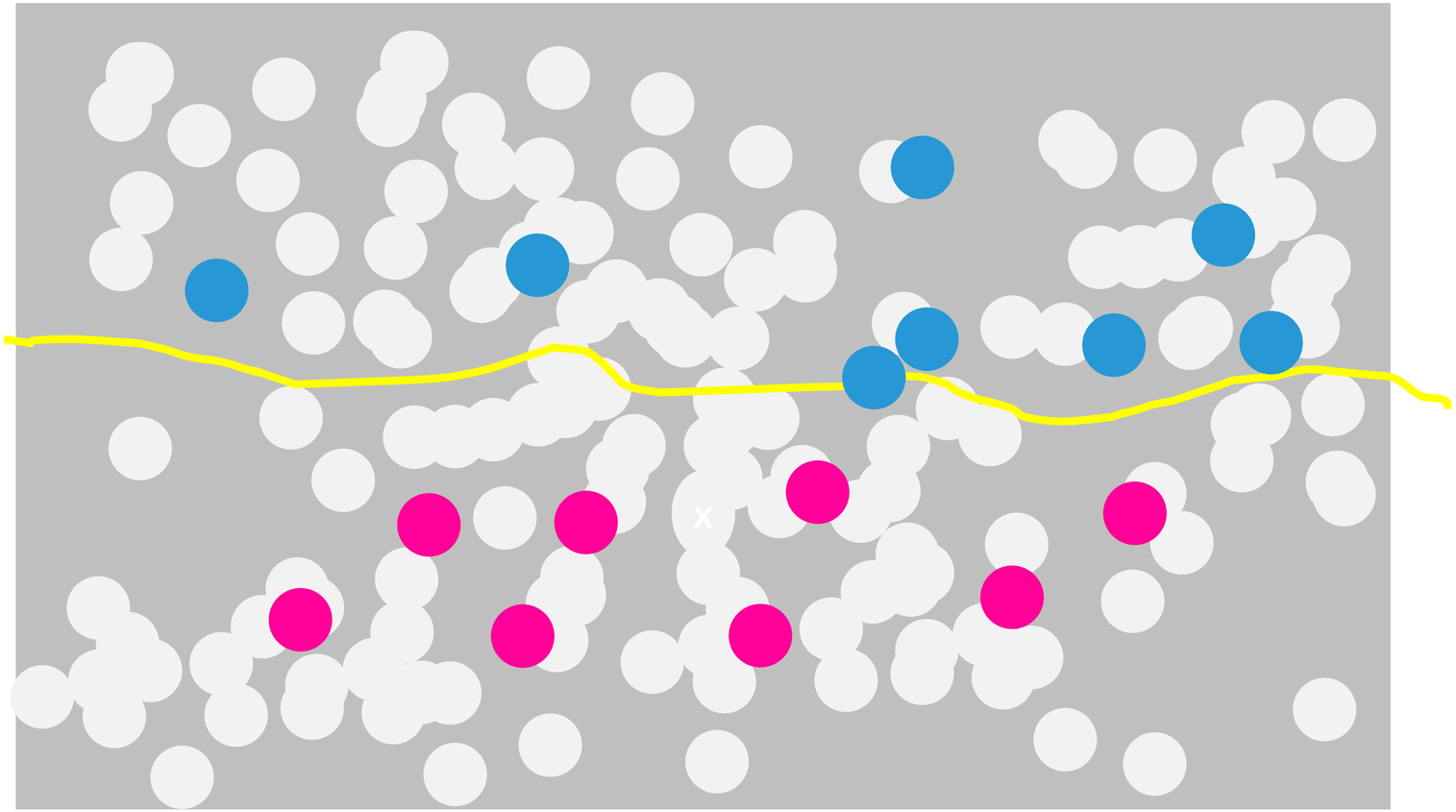
THESE ARE ALL CORRECT



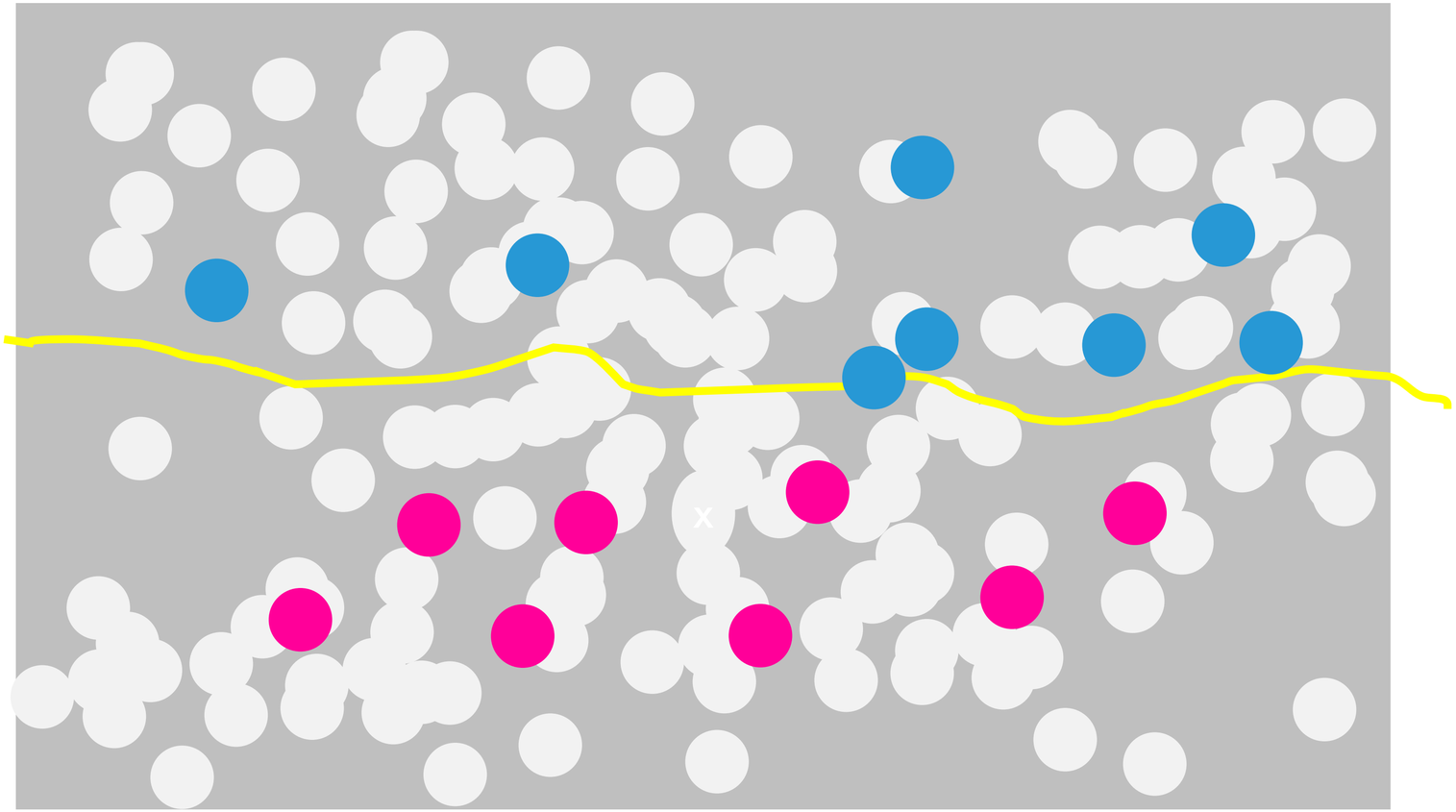
THESE ARE ALL “INCORRECT”



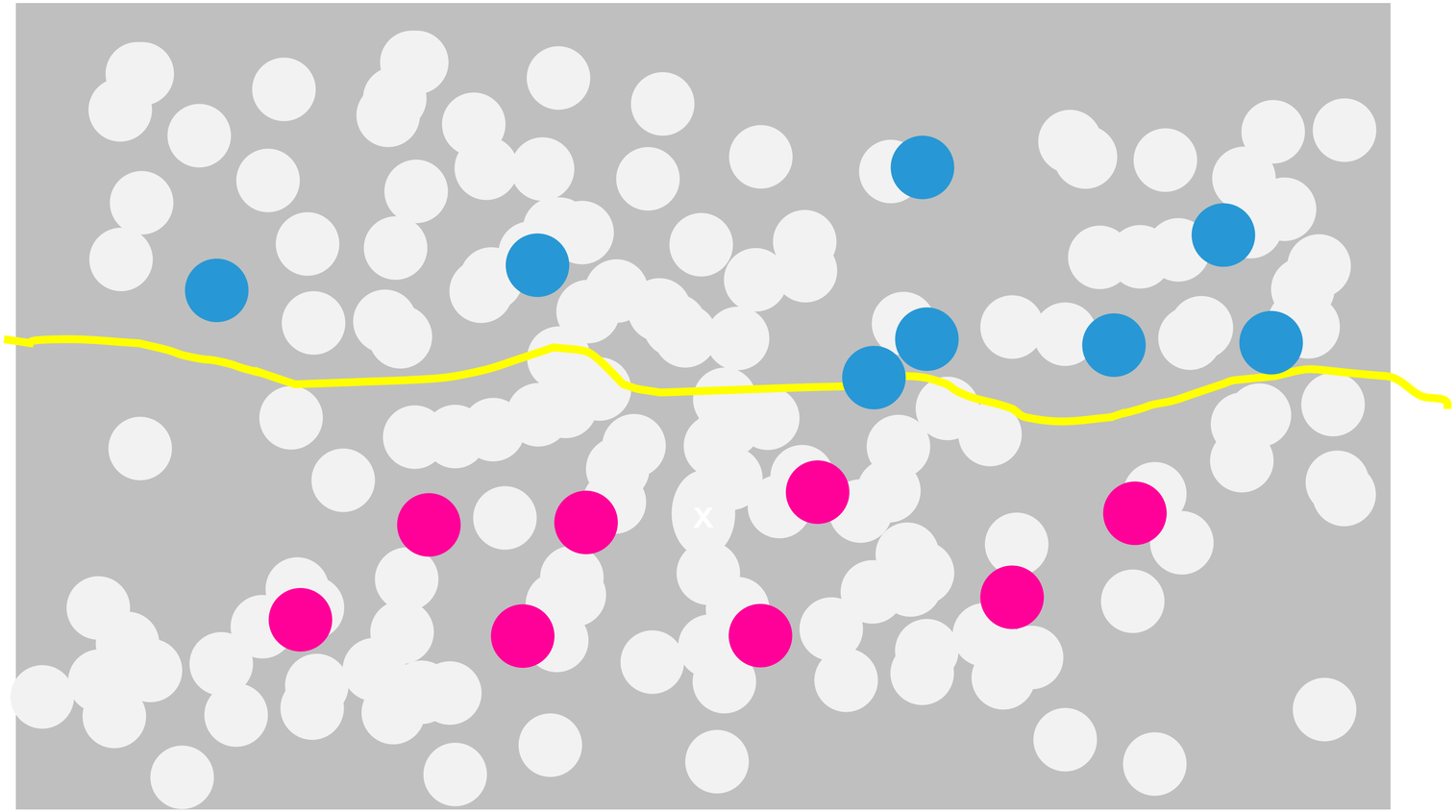
METRIC: INCORRECT ONES / CORRECT ONES



THRESHOLD: WHAT IS AN ACCEPTABLE NUMBER? 5%? 10%? 25%?

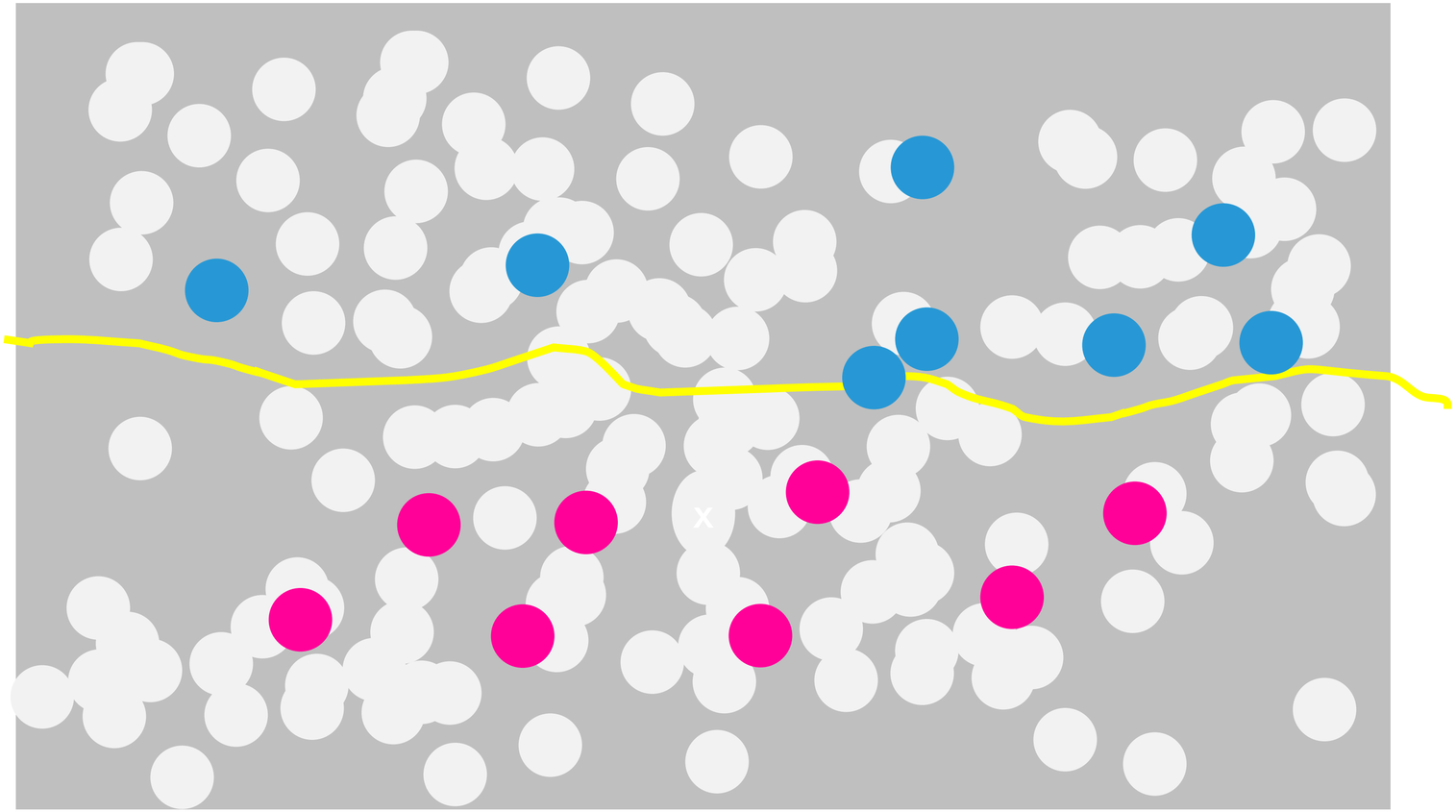


A METRIC IS: ONE NUMBER THAT SUMMARIZES YOUR DATA

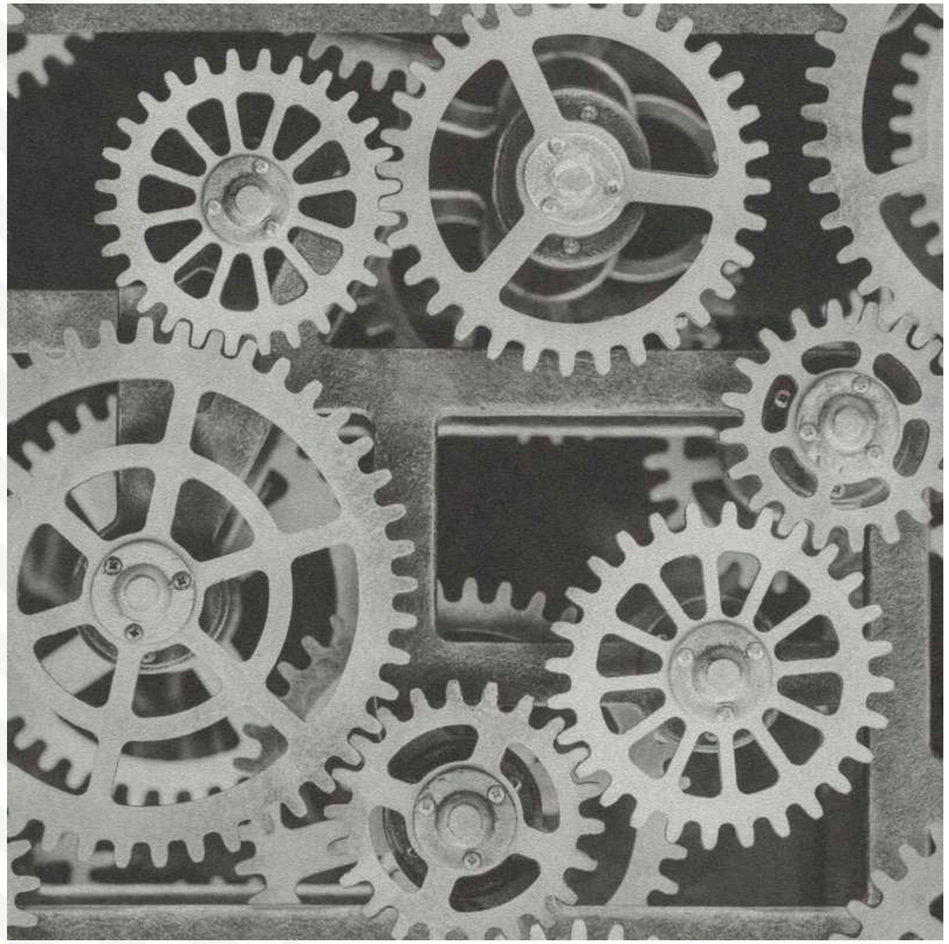


A METRIC IS: ONE NUMBER THAT SUMMARIZES YOUR DATA

A THRESHOLD IS: ONE NUMBER THAT YOU COMPARE YOU METRIC WITH



FIVE COMPLICATING FACTORS (NOT AN EXHAUSTIVE LIST)



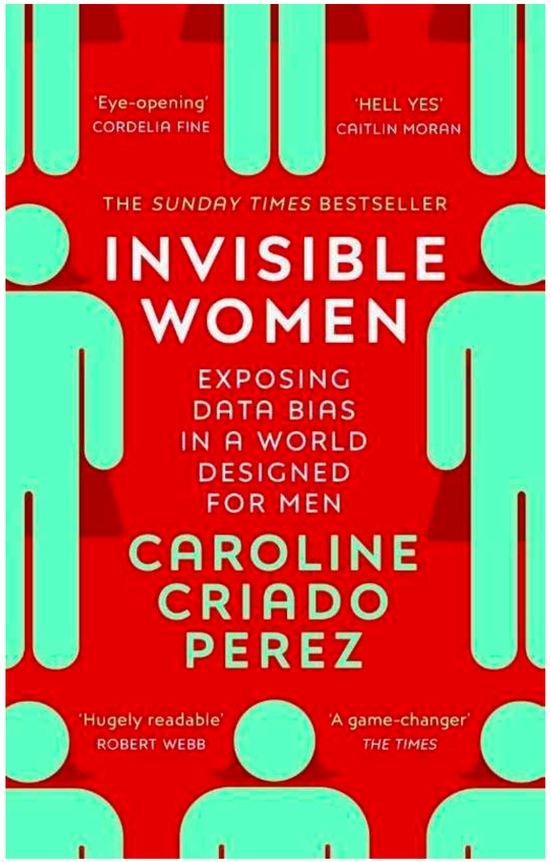
1. Metrics change truth
2. Data strategies change everything
3. Definitions are not neutral
4. Definitions are so so hard
5. Nothing is used in isolation

1. METRICS CHANGE ‘TRUTH’

“How did you sleep?”

“I don’t know, let me check”

2. Data strategies change everything



Which data?

a. Availabilty
b. Selection
c. Segmentation
d. Weighing

3. Definitions are not neutral

Exposure diversity

- "The content that the audience selects versus the available content"
- Immediate question: based on which properties?
- Supply side/Consumer side

Helberger, N., Karppinen, K., & D'acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191-207.

3. Definitions are not neutral

European Commission:

“defined media diversity not only in terms of the diversity of supply and distribution but also of use – although without specifying what the diverse use of information actually entails or how diverse such use would need to be in order to be considered sufficiently diverse”

Some kind of **lens**, some kind **interaction quality**, but **no metric** and **no threshold**

Helberger, N., Karppinen, K., & D’acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191-207.

3. Definitions are not neutral

British Ofcom:

"consumers actively multisource – such that the large majority of individuals consume a range of different news sources"

So some kind of **lens**, some kind **interaction quality**, a **metric** but **no threshold**

Helberger, N., Karppinen, K., & D'acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191-207.

3. Definitions are not neutral

British Ofcom:

"consumers actively multisource – such that the large majority of individuals consume a range of different news sources"

So some kind of **lens**, some kind **interaction quality**, a **metric** but **no threshold (no benchmark) (as Helberger et. al call it)**

Helberger, N., Karppinen, K., & D'acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191-207.

3. Definitions are not neutral

First things first: Why would you want diversity? (and there goes your neutrality)

1. Liberal perspective – people should have a choice
2. Deliberative perspective – facilitate a more inclusive society
3. Adversarial perspective – exposure to diverse viewpoints introduces valuable contestation

Helberger, N., Karppinen, K., & D'acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191-207.

3. Definitions are not neutral

First things first: Why would you want diversity? (and there goes your neutrality)

1. Liberal perspective – people should have a choice
2. Deliberative perspective – facilitate a more inclusive society
3. Adversarial perspective – exposure to diverse viewpoints introduces valuable contestation

Criteria: USER SATISFACTION and LEVEL OF AWARENESS OF OTHER OPTIONS
Metrics: session length, navigation behavior, likes/shares/clicks, attempts at adaptation, etc.

Helberger, N., Karppinen, K., & D'acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191-207.

3. Definitions are not neutral

First things first: Why would you want diversity? (and there goes your neutrality)

1. Liberal perspective – people should have a choice
2. Deliberative perspective – facilitate a more inclusive society
3. Adversarial perspective – exposure to diverse viewpoints introduces valuable contestation

Criteria: INCLUSIVENESS OF PUBLICE DISCOURSE and RECIPROCITY
Metrics: connections between people representing different ideologies, engagement with opposing views, etc.

Helberger, N., Karppinen, K., & D'acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191-207.

3. Definitions are not neutral

First things first: Why would you want diversity? (and there goes your neutrality)

1. Liberal perspective – people should have a choice
2. Deliberative perspective – facilitate a more inclusive society
3. Adversarial perspective – exposure to diverse viewpoints introduces valuable contestation

Criteria: PURPOSFULLY BIASED (TO CHALLENGE HEGEMONY)
Metrics: visibility of minority voices, visibility of controversial viewpoints, participation of earlier uninvolved groups

Helberger, N., Karppinen, K., & D'acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191-207.

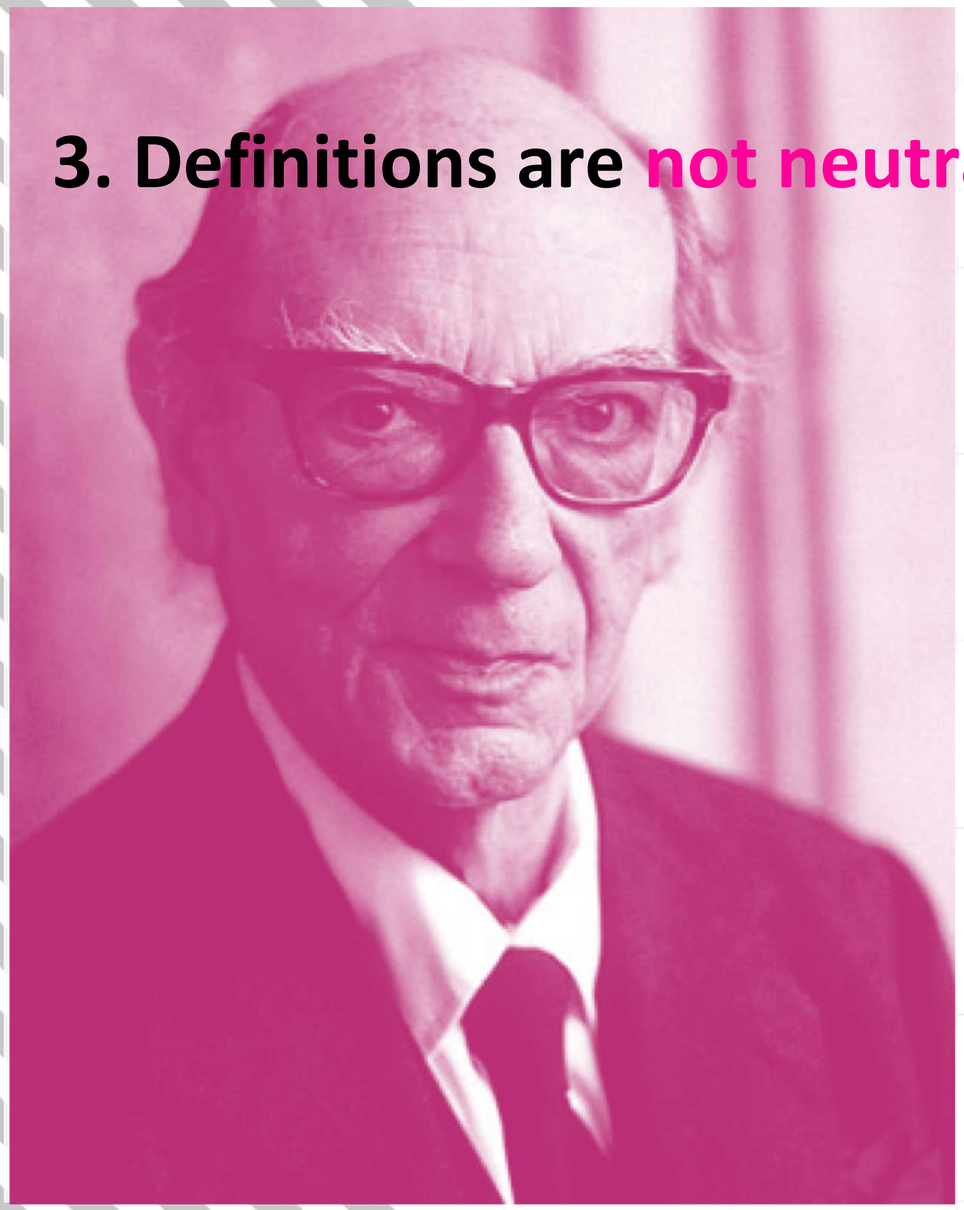
3. Definitions are not neutral

Examples of trade-offs that all affect the ‘neutrality’ of definitions:

- safety vs. efficiency
- flexibility vs. foolproof
- accuracy vs. diversity
- novelty vs. trust
- responsiveness vs. rigidity
- free market principle vs. regulation

Which side of the line you fall on **depends on your service**:
search engines? news outlets? streaming services for music and entertainment?

Helberger, N., Karppinen, K., & D’acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191-207.



3. Definitions are not neutral: ethical implications

For instance: **exposure diversity**

infringes on autonomy? or enhances autonomy?

Positive liberty: freedom to make positive choices
Negative liberty: freedom from interference
(Isaiah Berlin)

Helberger, N., Karppinen, K., & D'acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191-207.

3. Definitions are not neutral: how to ethically nudge

Table 1. Categorization of different types of nudges (Hansen & Jespersen, 2013, p. 23).

	Transparent	Non-transparent
System 2 Thinking	Transparent facilitation of consistent choice/prompting of reflected choices	Manipulation of choice
System 1 Thinking	Transparent influence (technical manipulation of behaviour)	Non-transparent manipulation of behaviour

Helberger, N., Karppinen, K., & D'acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191-207.

4. Definitions are so so hard

For instance:

What would be metrics for a potential romantic match?

- Dating apps used to go for similarity
- OKCupid allows user to decide (1) different or similar? (2) how different can it be? (3) how important is that question?
- user in the driver's seat (next week)

4. Definitions are so so hard

For instance:

What would be metrics for safe driving?

https://docs.google.com/document/d/1jVE426PmOOrmiyJ6wCmRZ2Bb822pm2gN-ZHjZfed6Xg/edit?usp=sharing

ANWB – AA (UK) - TRIPLE A (US) - ADAC (D) - ACP (P) – STARTER24 (PL)

Assistance
Traffic information
Holidays

ANWB – AA (UK) - TRIPLE A (US) - ADAC (D) - ACP (P) – STARTER24 (PL)

Assistance
Traffic information
Holidays

ANWB – AA (UK) - TRIPLE A (US) - ADAC (D) - ACP (P) – STARTER24 (PL)

Assistance
Traffic information
Holidays
Lobbying
Insurances

ANWB – AA (UK) - TRIPLE A (US) - ADAC (D) - ACP (P) – STARTER24 (PL)

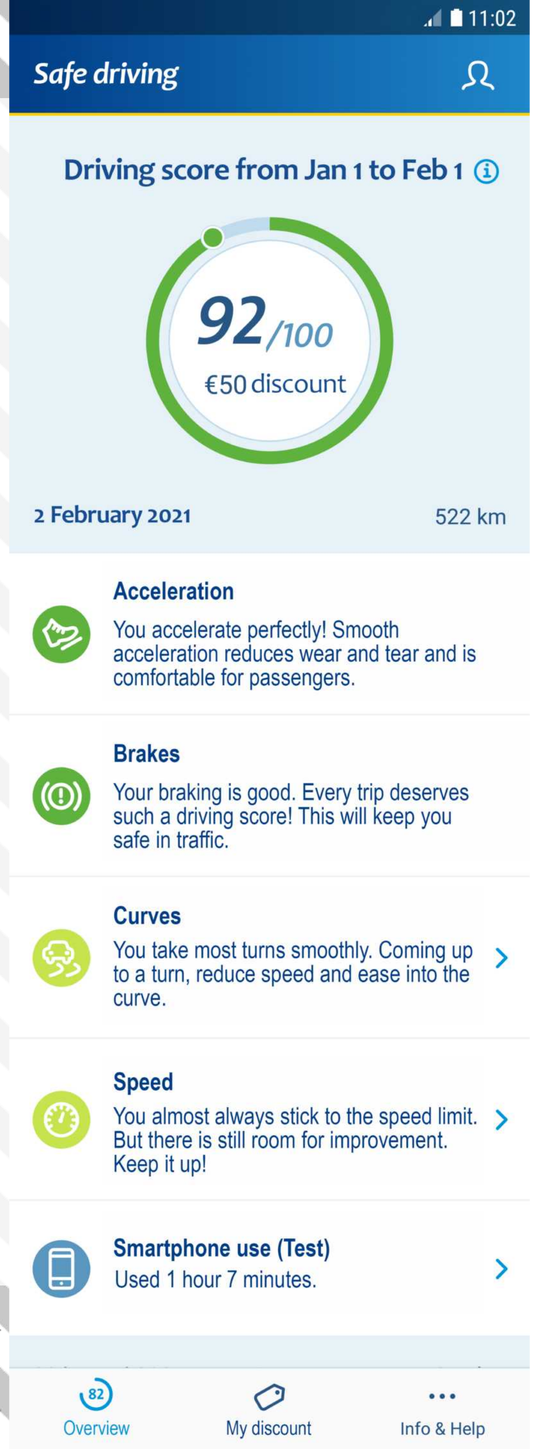
Assistance
Traffic information
Holidays
Lobbying
Insurances

ANWB'S METRICS FOR SAFE DRIVING

- Discount for Safe driving behavior
- Every 10 days: feedback
- Including recommendations

Other elements that play a role:

- interest insurance company
- solidarity
- personal safety/public safety

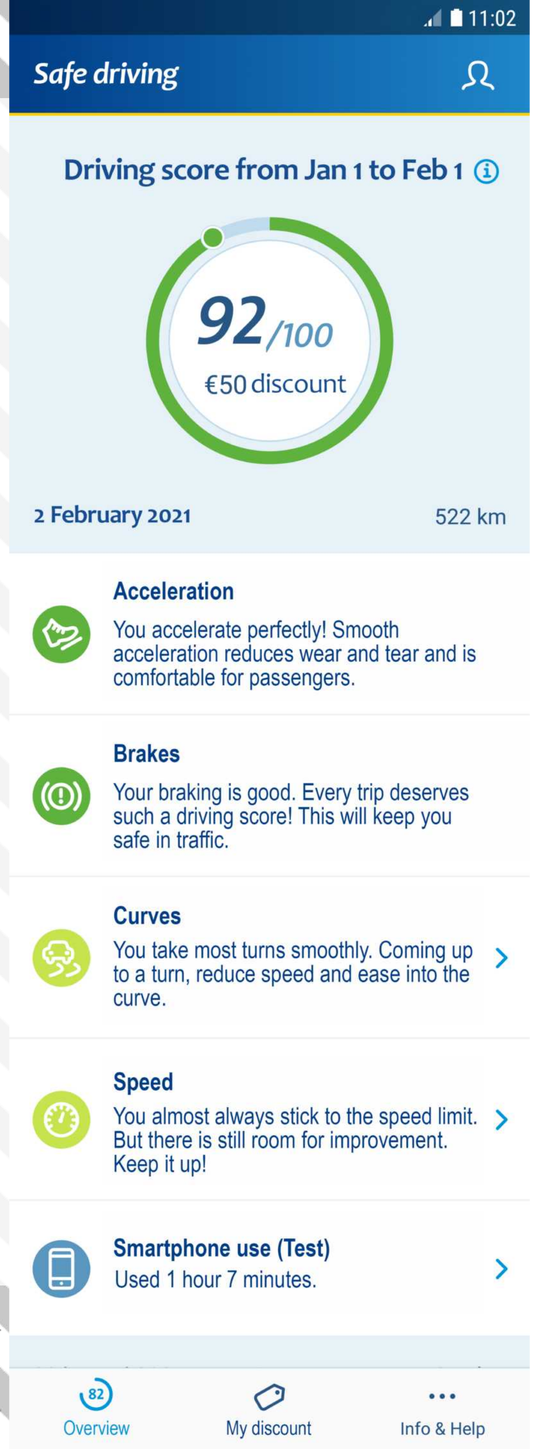


ANWB'S METRICS FOR SAFE DRIVING

- Acceleration
- Brakes
- Curves
- Speed
- (Test) use of the mobile phone

What do all these have in common?

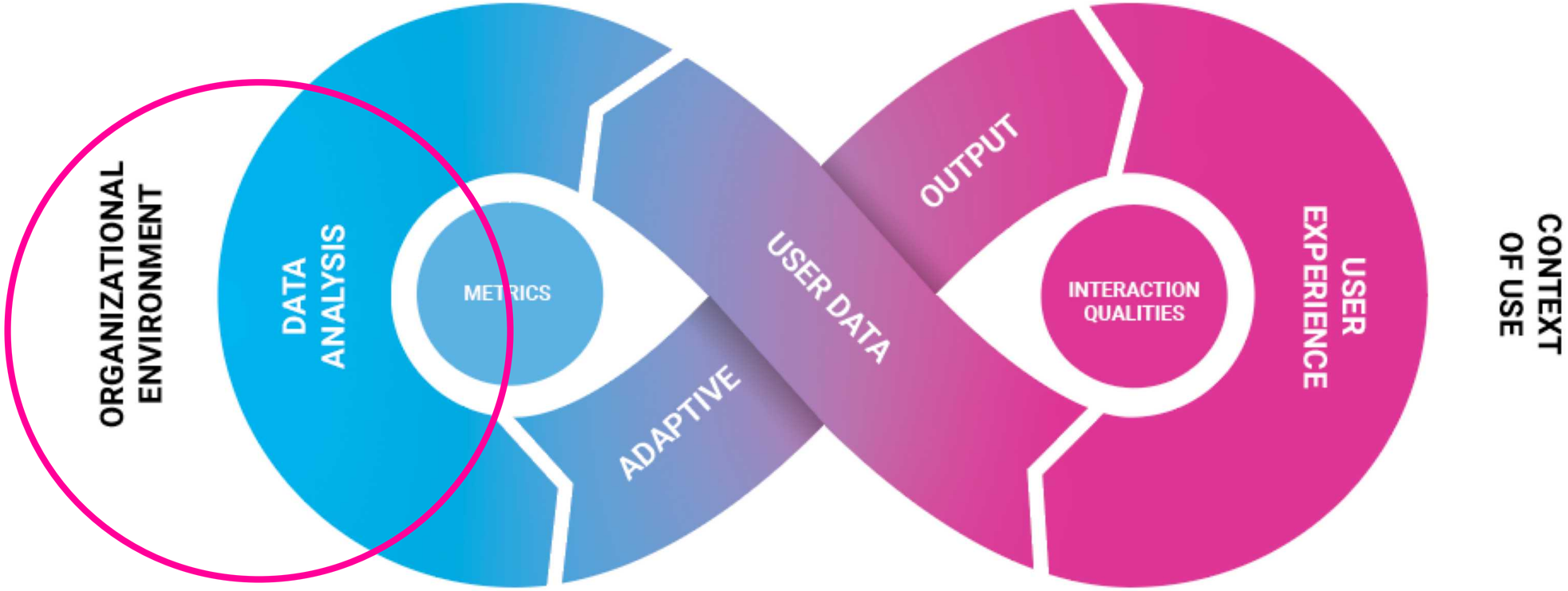
Am I a safe driver?



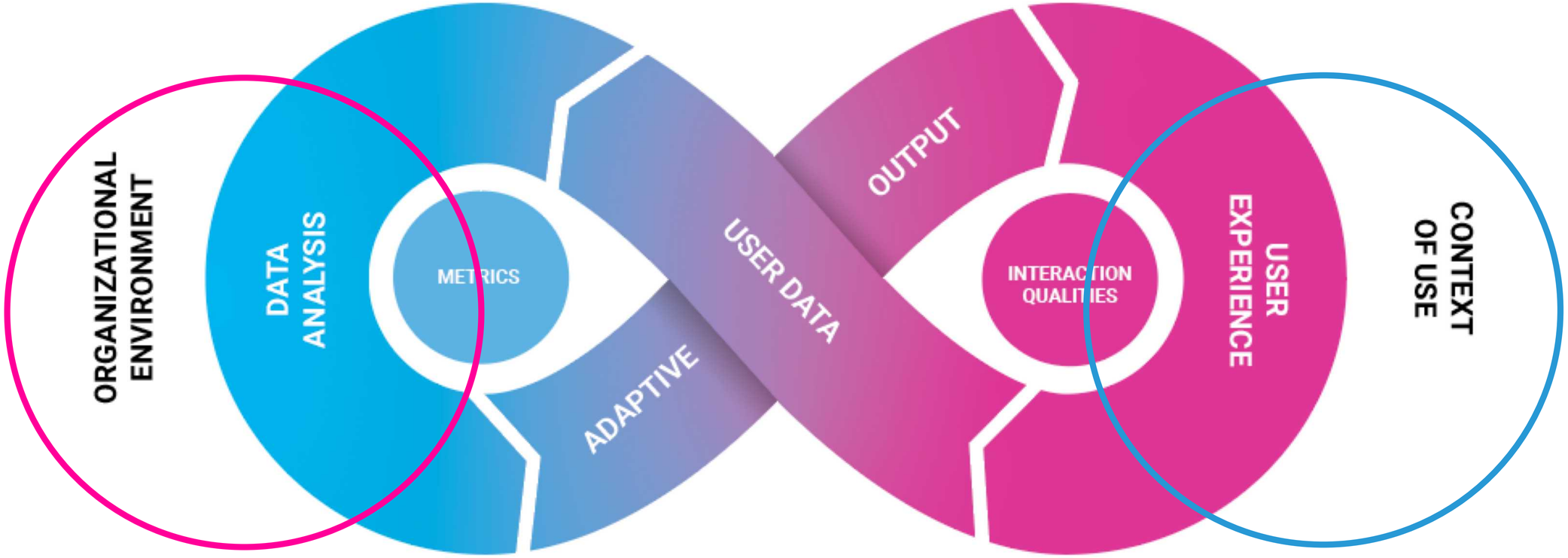
5. NOTHING IS USED IN ISOLATION

- These drivers are on the road
- There are other road users in all shapes and forms
- There are – in my case – hysterical cars that interfere

METRICS: WHAT ARE THE SOURCES?



METRICS: WHAT ARE THE SOURCES?



A USER-CENTRIC EVALUATION FRAMEWORK FOR REC SYS

ResQue

Pu, P., Chen, L., & Hu, R. (2011, October). A user-centric evaluation framework for recommender systems. In Proceedings of the fifth ACM conference on Recommender systems (pp. 157-164).

A USER-CENTRIC EVALUATION FRAMEWORK FOR REC SYS

ResQue

Pu, P., Chen, L., & Hu, R. (2011, October). A user-centric evaluation framework for recommender systems. In Proceedings of the fifth ACM conference on Recommender systems (pp. 157-164).

A USER-CENTRIC EVALUATION FRAMEWORK FOR REC SYS

ResQue aims at measuring:

- Quality of the recommended items
- System's usability
- System's usefulness
- System's interface
- System's interaction qualities
- Users' satisfaction with the systems
- Influence on users' behavioral intentions (including purchasing intentions of recommendations)
- Returning users.

Pu, P., Chen, L., & Hu, R. (2011, October). A user-centric evaluation framework for recommender systems. In Proceedings of the fifth ACM conference on Recommender systems (pp. 157-164).

A USER-CENTRIC EVALUATION FRAMEWORK FOR REC SYS

Some of these things are measured with the help by:

- understanding and identifying effective preference elicitation methods
- techniques that motivate users to rate items
- methods that generate diverse and more satisfying recommendation lists
- explanation interfaces
- trust formation with recommenders
- design guidelines for enhancing a recommender's interface layout
- other determinants that influence users' positive perception of recommender systems

Pu, P., Chen, L., & Hu, R. (2011, October). A user-centric evaluation framework for recommender systems. In Proceedings of the fifth ACM conference on Recommender systems (pp. 157-164).

PROCEDURE

1. Selection of constructs and their domains
   - perceived system qualities
   - users' beliefs as a result of these qualities
   - subjective attitudes
   - behavioral intentions

2. Validation of constructs by comparison to accepted instruments (TAM, UTAUT, SUMI)

3. Questions to operationalize the constructs (iterative procedure)

Pu, P., Chen, L., & Hu, R. (2011, October). A user-centric evaluation framework for recommender systems. In Proceedings of the fifth ACM conference on Recommender systems (pp. 157-164).

Layers and their operationalizations

1. **Perceived system qualities:**
   - recommendation quality
   - the interface adequacy
   - the interaction adequacy of the recommender

2. **Beliefs**

3. **Attitudes**

4. **Behavioral intentions**

Pu, P., Chen, L., & Hu, R. (2011, October). A user-centric evaluation framework for recommender systems. In Proceedings of the fifth ACM conference on Recommender systems (pp. 157-164).

Operationalizations layer 1 and their metrics (from a user's perspective!)

1. **Recommendation Quality:**
   Perceived accuracy, novelty, attractiveness, diversity, context compatibility
2. **Interface Adequacy:**
   Information sufficiency, interface label layout adequacy, clarity
3. **Interaction Adequacy**
   Initial preference elicitation, preference revision, system's ability to explain results

Pu, P., Chen, L., & Hu, R. (2011, October). A user-centric evaluation framework for recommender systems. In Proceedings of the fifth ACM conference on Recommender systems (pp. 157-164).

Layers and their operationalizations

1. **Perceived system qualities:**

2. **Beliefs**
   - Perceived usefulness
   - Perceived ease of use
   - Control and transparency

3. **Attitudes**

4. **Behavioral intentions**

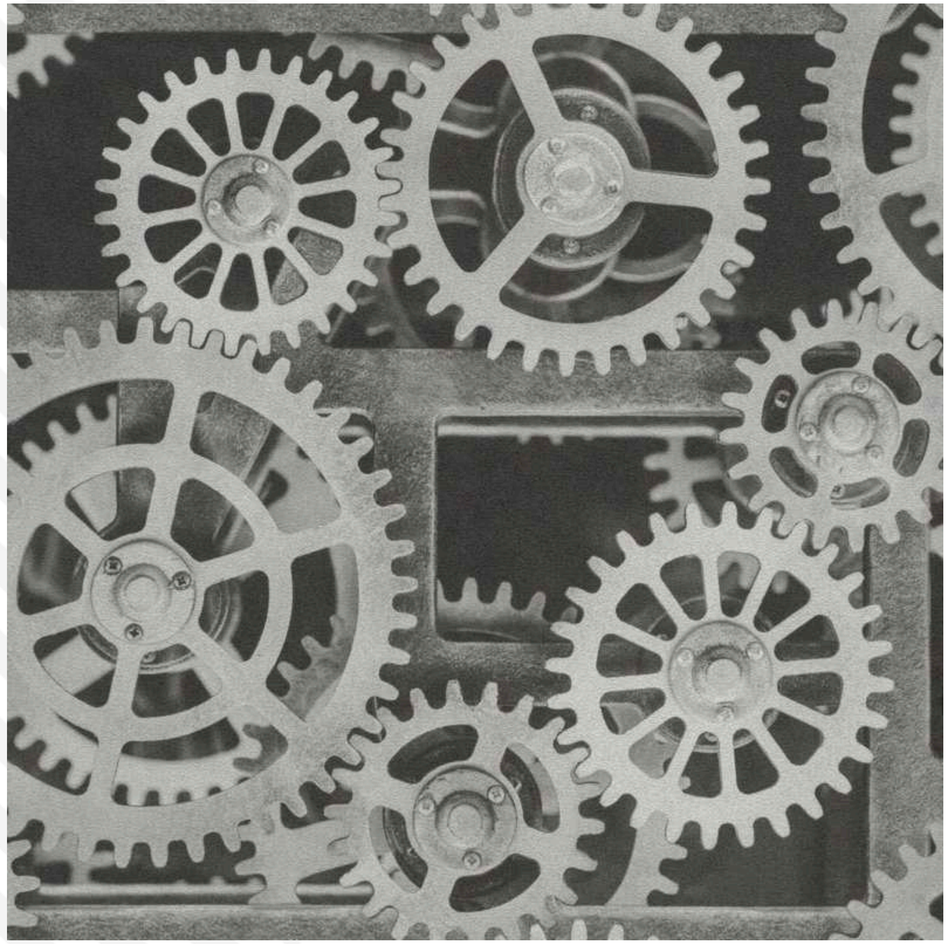
Pu, P., Chen, L., & Hu, R. (2011, October). A user-centric evaluation framework for recommender systems. In Proceedings of the fifth ACM conference on Recommender systems (pp. 157-164).

Layers and their operationalizations + metrics)

1. **Perceived system qualities:**
2. **Beliefs**
   - Perceived usefulness (decision support/decision quality
   - Perceived ease of use (efficiency)
   - Control and transparency (user control/transparency)
3. **Attitudes**
   - overall satisfaction, confidence inspiring, trust
4. **Behavioral intentions**
   - User loyalty, user ‘purchases’

Pu, P., Chen, L., & Hu, R. (2011, October). A user-centric evaluation framework for recommender systems. In Proceedings of the fifth ACM conference on Recommender systems (pp. 157-164).

FIVE COMPLICATING FACTORS (NOT AN EXHAUSTIVE LIST)



1. Metrics change truth
2. Data strategies change everything
3. Definitions are not neutral
4. Definitions are so so hard
5. Nothing is used in isolation