Gathering Insights: Evaluation at trivago

Saad Mahamood, Lead Data Scientist & NLG Expert, 16.12.2022, Utrecht, Netherlands

Agenda

1. About Me
2. About trivago
3. Content & Evaluations
4. C-Testing
5. User Research & Evaluations
6. Summary

1. About Me

About Me

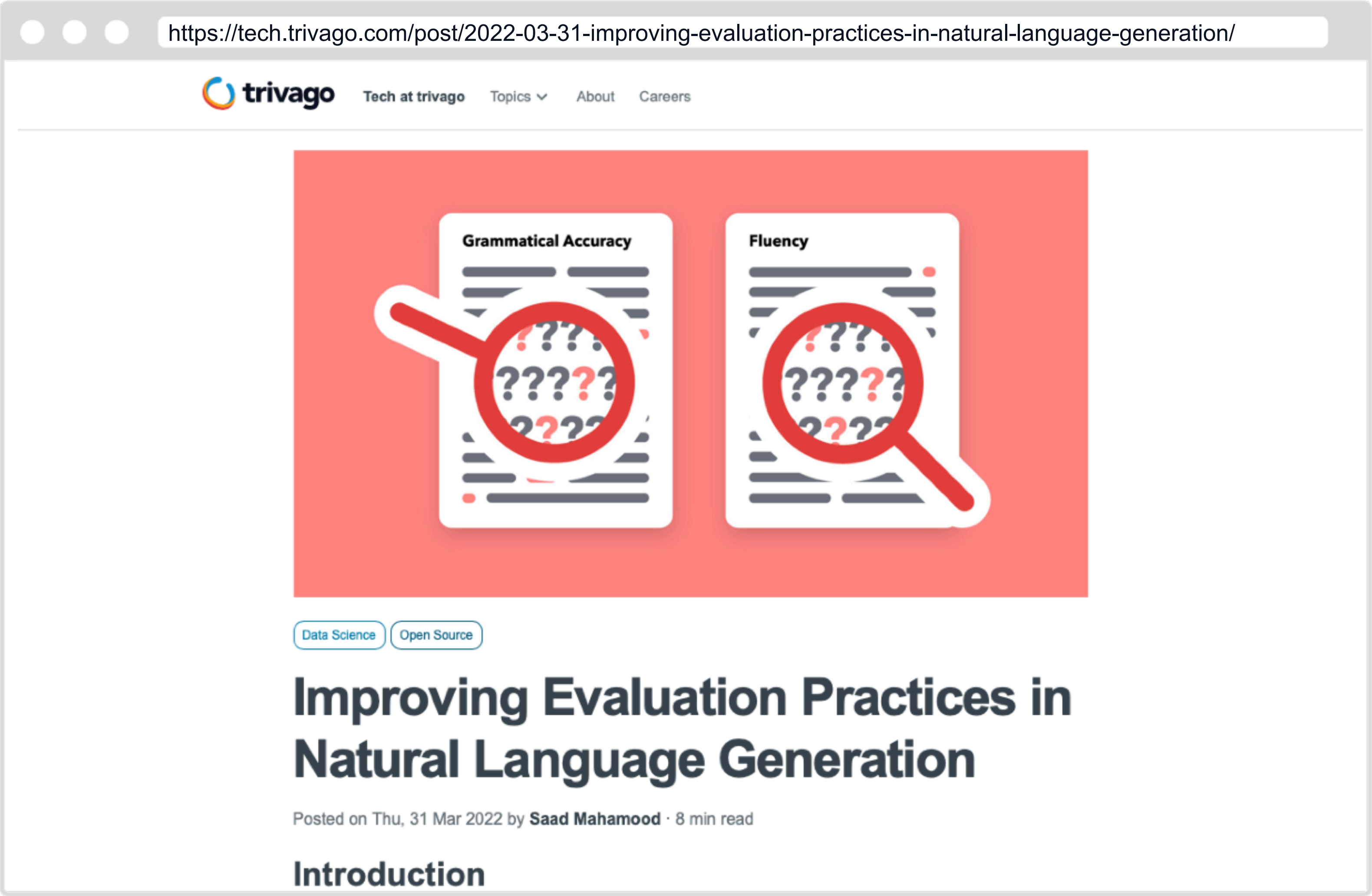
- Been active the field of NLG for ~14 years.
- 6 years at Aberdeen University.
  - 4 years PhD
  - 2 years PostDoc
- 5 years at Arria NLG.
- 4 years now at trivago.
  - Released ‘Hotel Scribe’ to generate automated descriptions of accommodations [Mahamood and Zembrzuski, 2019].
  - Currently lead a team of four data scientists.
  - Focus on a range of content based data science topics such as Image Tagging, Geospatial, Accommodation Metadata, and Data Quality problems.
  - Actively, participating in NLG research at trivago, including research projects. In particular, recently focused on the topic of evaluations.

NLG Evaluation Work

- ~2010: First aware of issues in NLG evaluation, due to Reiter's & Belz's paper on investigating the validity of evaluations [Reiter & Belz, 2009]
- 2015: Worked with Dimitra Gkatzia on generating a snapshot of NLG evaluation practices for the last 10 years [Gkatzia & Mahamood, 2015].
- 2020: Collaborated with multiple researchers on an extensive 20-year overview of how NLG human evaluations are conducted [Howcroft et al., 2020].

NLG Evaluation Work

- 2021: Worked on several evaluation related research initiatives:
    - Explored Commonsense human NLG evaluations [Clinciu et al., 2021]
    - Underreporting of errors in NLG output [Miltenberg et al, 2021]
    - Reproducing an earlier NLG experiment [Mahamood, 2021]
    - Automatic construction of evaluation test sets [Mille et al, 2021]
- 2022: Active participation in the ReproHum project.



https://tech.trivago.com/post/2022-03-31-improving-evaluation-practices-in-natural-language-generation/

Improving Evaluation Practices in Natural Language Generation

Posted on Thu, 31 Mar 2022 by **Saad Mahamood** · 8 min read

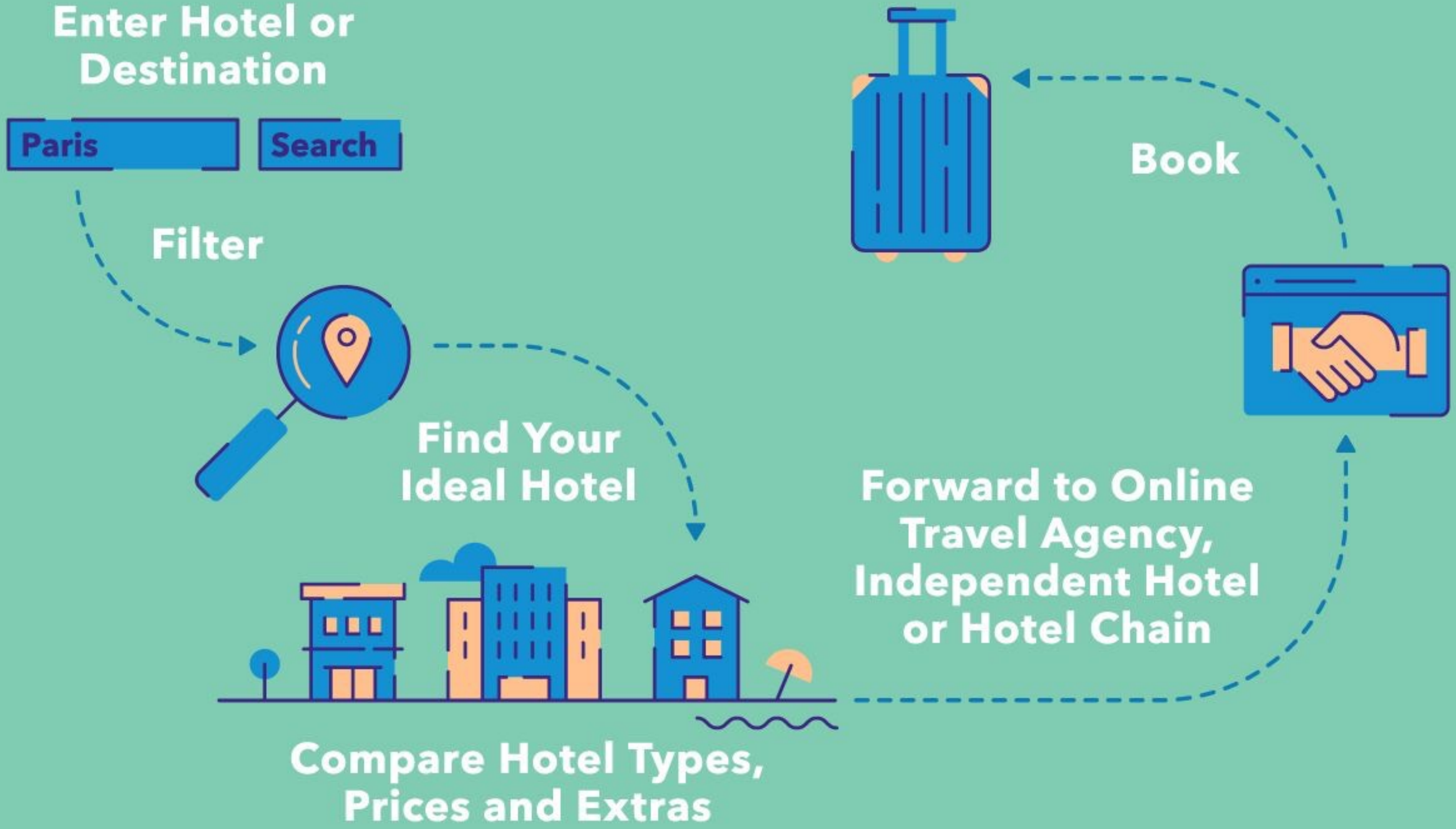
Introduction

2. About trivago

About trivago

Since being founded in 2005, trivago has become a leading global accommodation search website:

- Focus on helping millions of travellers to search for and compare hotels and other accommodations.
- Conducts business in 190 countries, with 5 million hotels and alternative accommodation, across 53 localised platforms in 31 languages.
- Based in Düsseldorf, Germany with around 800 employees.



Enter Hotel or Destination

Filter

Find Your Ideal Hotel

Compare Hotel Types, Prices and Extras

Forward to Online Travel Agency, Independent Hotel or Hotel Chain

Book

About Connectivity Insights

Connectivity Insights is one several Data Science teams within trivago…

- Team composed of four data scientists lead by myself.
- Focused on content related Data Science problems. In particular, we focus on:
    - **Images** — e.g. Image tagging, image quality, image selection, etc.
    - **Matching Accommodations from Partners and Destination Assignment**
    - **Accommodation Metadata, Descriptions, Reviews, and Ratings**
    - **Pricing Data** — e.g. Fetching prices, price components, etc.

3. Content & Evaluations

Content & Evaluations

Content based evaluations are mostly intrinsic evaluations…

- We seek to evaluate the performance of a given solution/approach by evaluating either against:
    - Internally collected Human annotated ground-truth datasets.
        - Requires extensive effort to align annotators to annotate consistently and correctly.
    - External ground-truth datasets provided by third-party providers.
        - Where possible try and use multiple providers and look for consensus.
    - Comparative evaluations against an existing solution.
        - “How does X compare to current Y.”

Content & Evaluations

In addition to evaluations, the business impact of content related changes is also equally important…

- Impact can be measured in terms of:
    - The number of accommodations impacted with a given change.
    - The types of accommodations impacted.
        - Some accommodations are more important than others.
    - The countries or platforms impacted by a given change.
- Some examples of evaluations done for some recent projects:
    - Image Tagging
    - Accommodation Type Assignment
    - Destination Assignment

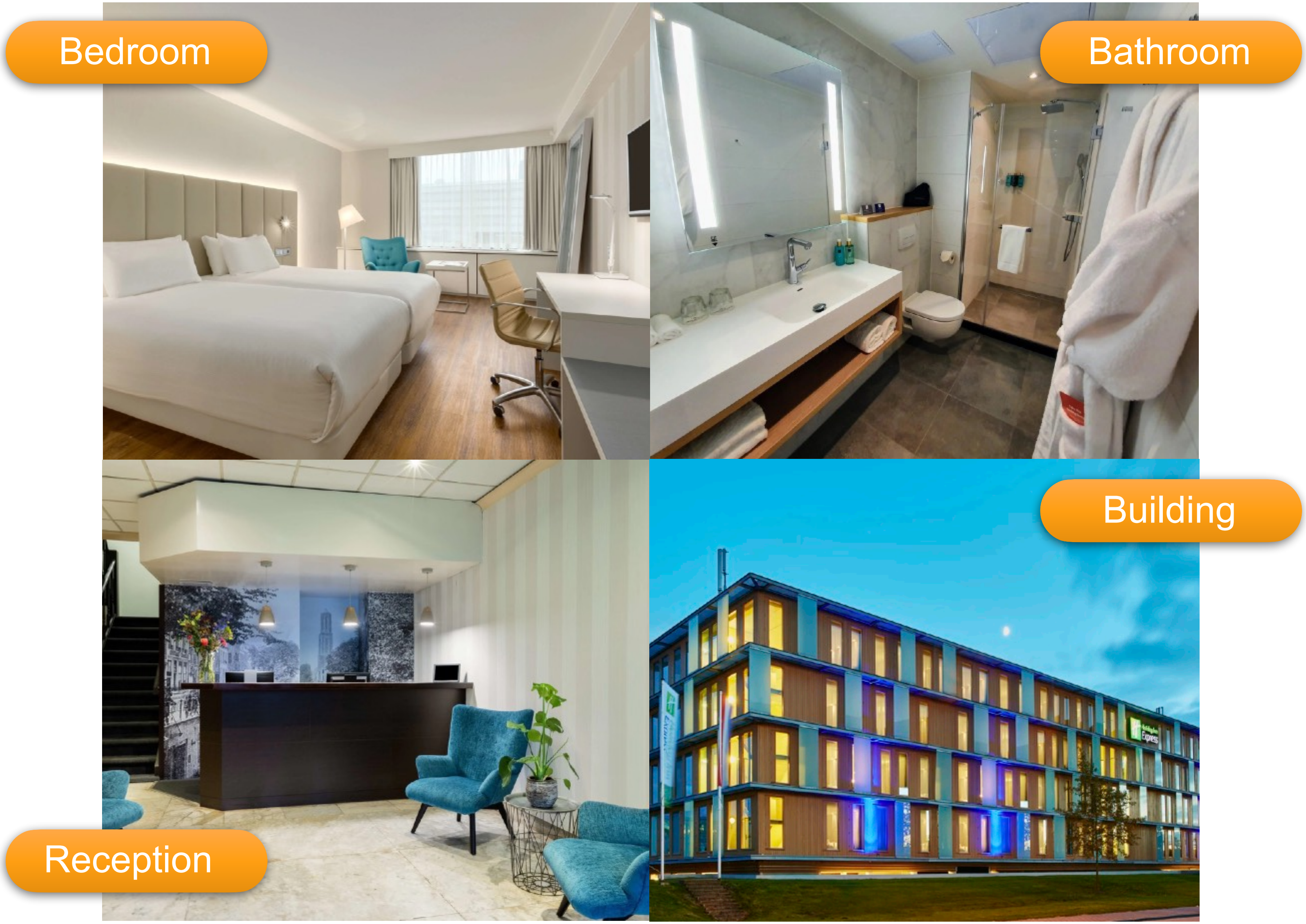


Image Tagging

Omayma Said

Danti Ramadanti

Image tagging allows us to give semantic “meaning” to images and enables downstream opportunities…

- Developed an in-house model after evaluating third-party image tagging models from an external provider.
- Evaluation is done by using a withheld human annotated dataset and computing precision and recall on a per tag basis.
  - In addition, we manually inspect the tagged images for correctness at random.
- Greatest challenge is not the evaluation, but the human dataset collection for training, testing, and validation:
  - Need robust and concrete definitions per tag.
  - Must extensively align annotators to tag images consistently and correctly.

Image Tagging

Accom. Type Assignment

B&B

House / Apartment

Hostel

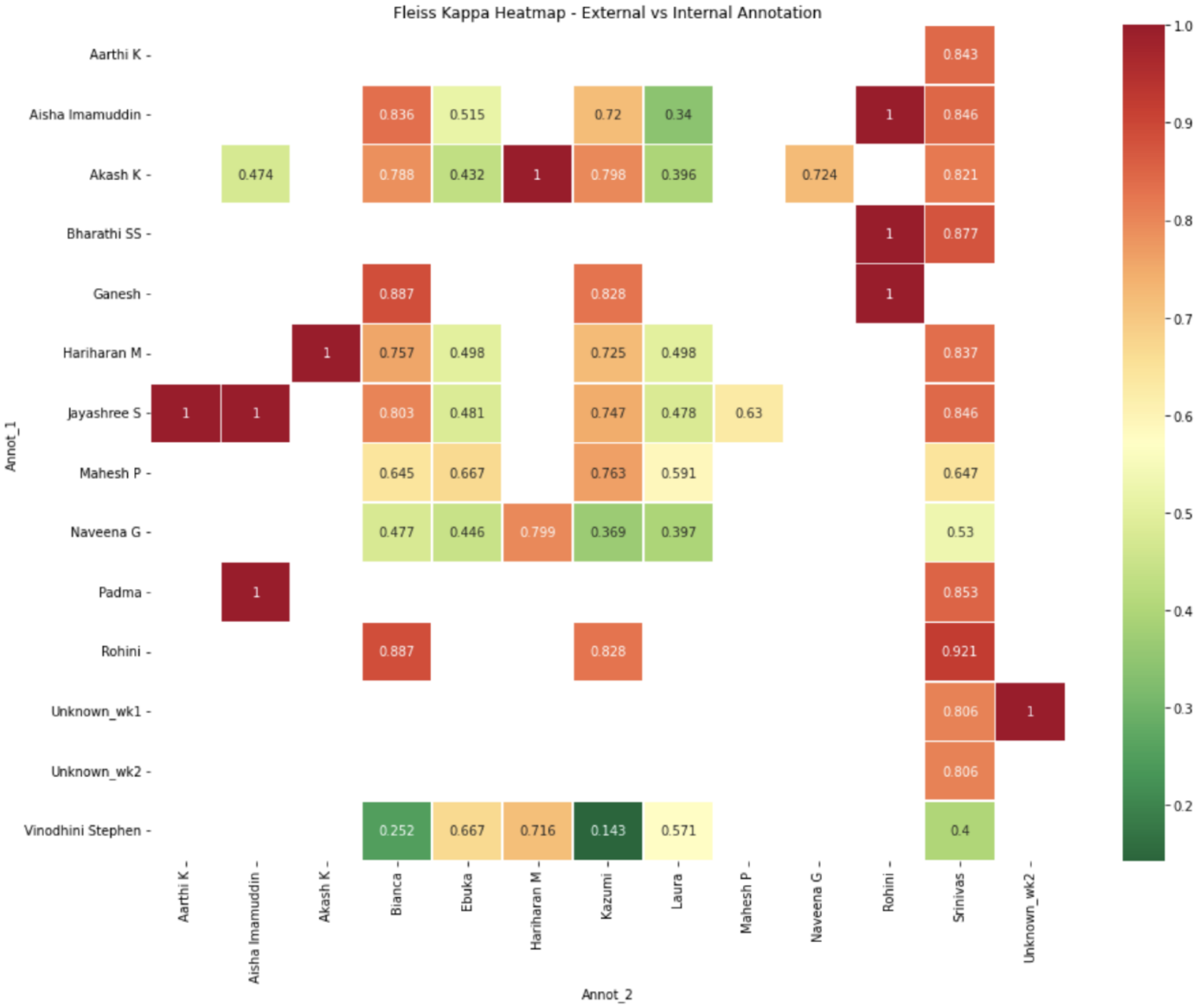
Hotel

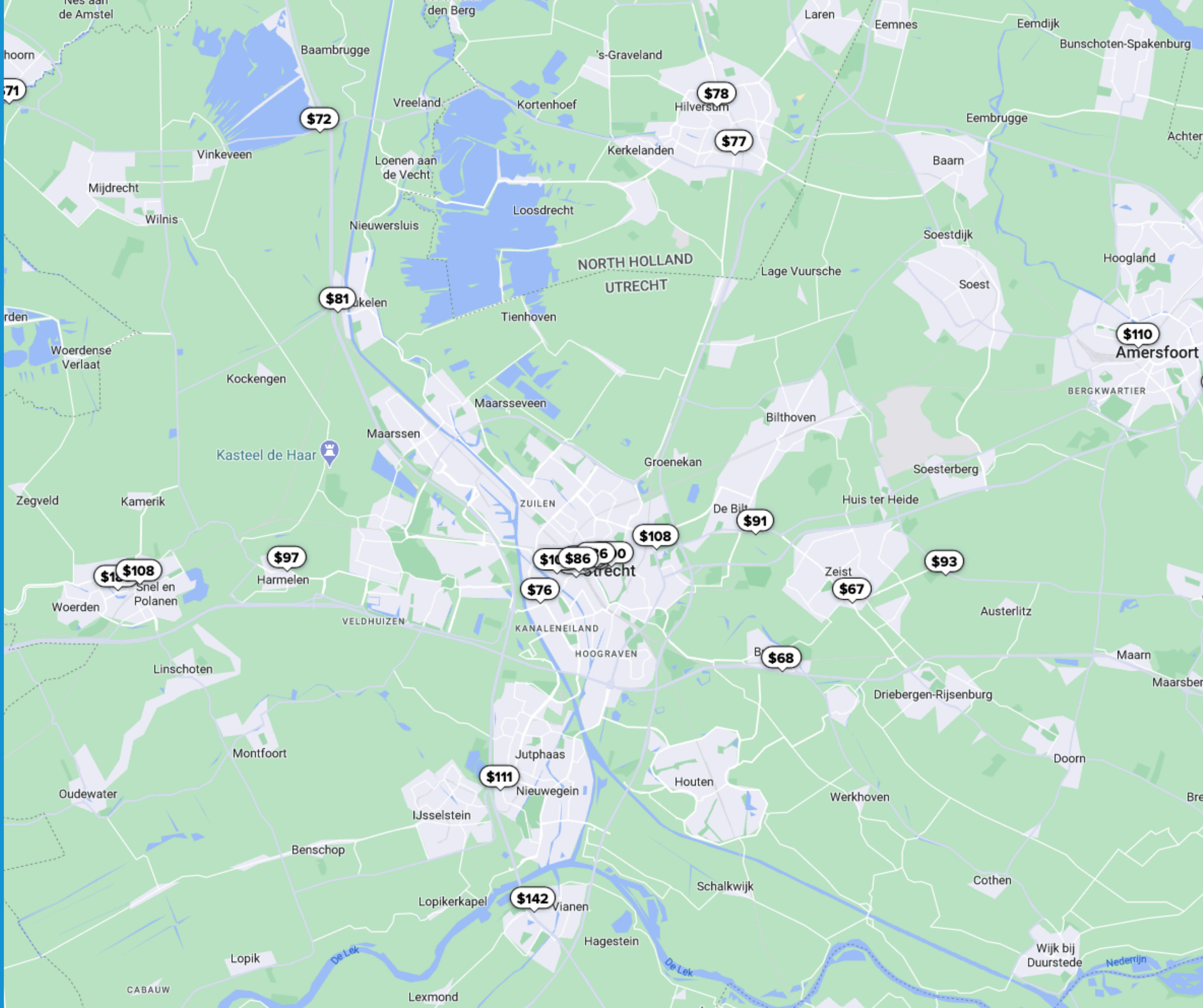
Srinivas Ramesh
Kamath

Accommodation type assignment is the process of assigning a type for a given item…

- Normally, this data is provided by the partner who is sending the accommodation to trivago.
- However, in some cases this information is missing or differs from our definitions and we must try and infer the accommodation type.
- Developed a robust rule-based model to replace a non-functioning neural model for type assignment.
- Like image tagging, evaluation was done against high quality human annotations over 12,000 items.

Accom. Type Assignment


Fleiss Kappa Heatmap - External vs Internal Annotation
Annot_1
Annot_2
Aarthi K
Aisha Imamuddin
Akash K
Bharathi SS
Ganesh
Hariharan M
Jayashree S
Mahesh P
Naveena G
Padma
Rohini
Unknown_wk1
Unknown_wk2
Vinodhini Stephen
Aarthi K
Aisha Imamuddin
Akash K
Bianca
Ebuka
Hariharan M
Kazumi
Laura
Mahesh P
Naveena G
Rohini
Srinivas
Unknown_wk2
0.843
0.836
0.515
0.72
0.34
1
0.846
0.474
0.788
0.432
1
0.798
0.396
0.724
0.821
1
0.877
0.887
0.828
1
1
0.757
0.498
0.725
0.498
0.837
1
1
0.803
0.481
0.747
0.478
0.63
0.846
0.645
0.667
0.763
0.591
0.647
0.477
0.446
0.799
0.369
0.397
0.53
1
0.853
0.887
0.828
0.921
0.806
1
0.806
0.252
0.667
0.716
0.143
0.571
0.4
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2



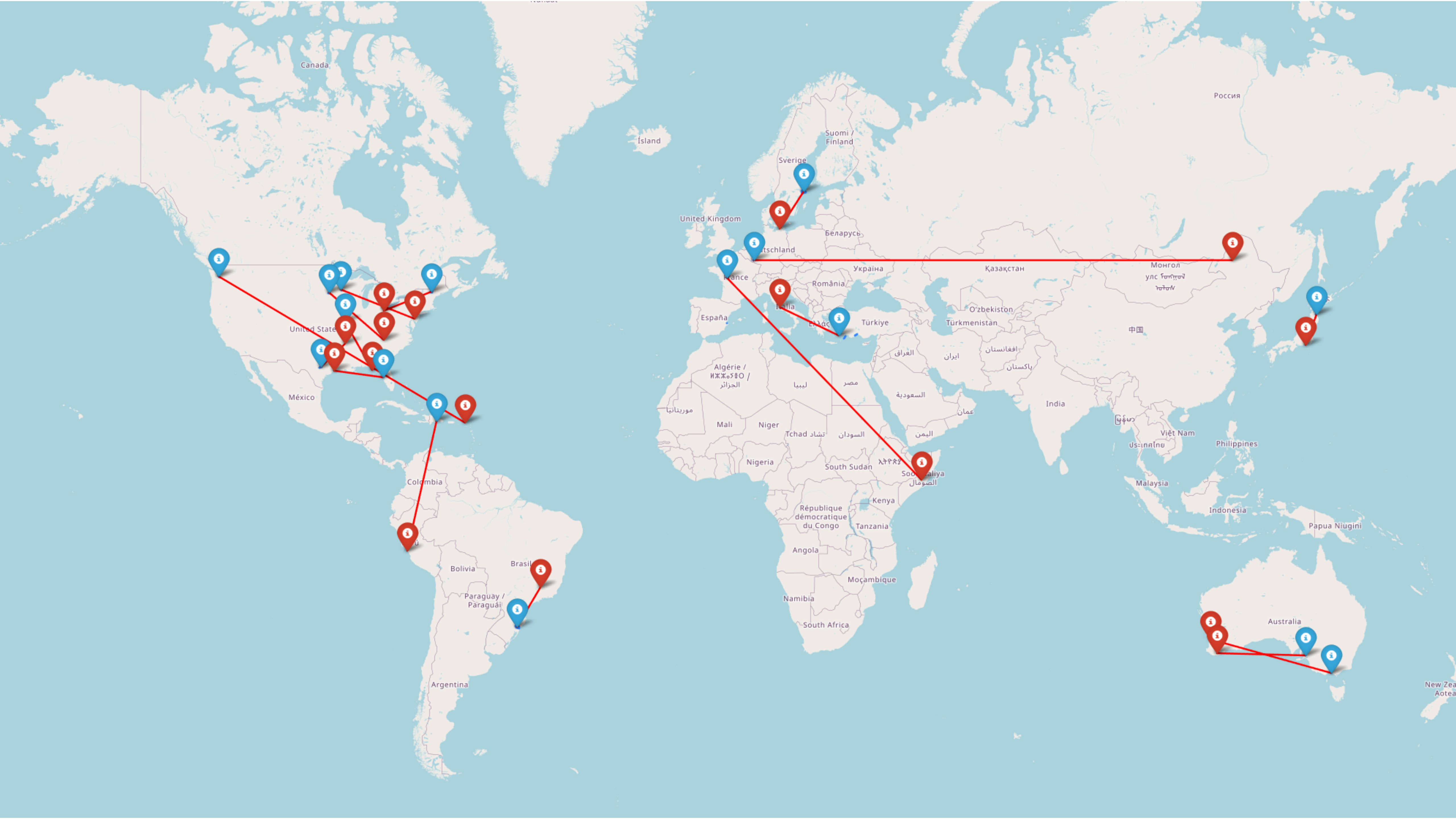
Destination Assignment

Veronica Gonzalez Solano

Destination Assignment

Destination assignment is the process of assigning accommodation to a location such as a city, town, or attraction.

- Biggest challenges when performing destination assignment are the following:
  - Missing, incorrect, or misleading Geocodes.
  - Missing, incorrect, or contradictory addressing data.
- Therefore we need to not only locate where an accommodation should be, but then assign it a destination.
- However, this all depends on the correctness of our data and processes to perform the above.


Canada
Ísland
Sverige
Suomi / Finland
United Kingdom
Беларусь
Deutschland
France
Україна
România
Italia
España
Türkiye
Қазақстан
Монгол улс
Россия
O'zbekiston
Türkmenistan
中国
United States
México
Colombia
Bolivia
Brasil
Paraguay / Paraguái
Argentina
Algérie / ⵍⵣⵣⴰⵢⴻⵔ / الجزائر
ليبيا
مصر
موريتانيا
Mali
Niger
Tchad تشاد
السودان
Nigeria
South Sudan
Soomaaliya الصومال
Kenya
République démocratique du Congo
Tanzania
Angola
Moçambique
Namibia
South Africa
العراق
ایران
افغانستان
پاکستان
السعودية
عمان
اليمن
India
Việt Nam
ประเทศไทย
Philippines
Malaysia
Indonesia
Papua Niugini
Australia
New Zealand / Aotearoa

Destination Assignment

We focused on evaluating the correctness of our destination ground truth data…

- In particular, check that we have correct data for where each destination is located w.r.t their geocodes.
- Ensure that geographical polygons have the right destination.
- Evaluation consists of:
    - Checking multiple data providers e.g. Google, OSM, TomTom, ArcGIS, etc.
        - Looking of either complete or majority consensus.
    - Using human annotators when there is **no consensus** between data providers.
        - Highly challenging as there is high degree of subjectivity involved.

4. C-Testing

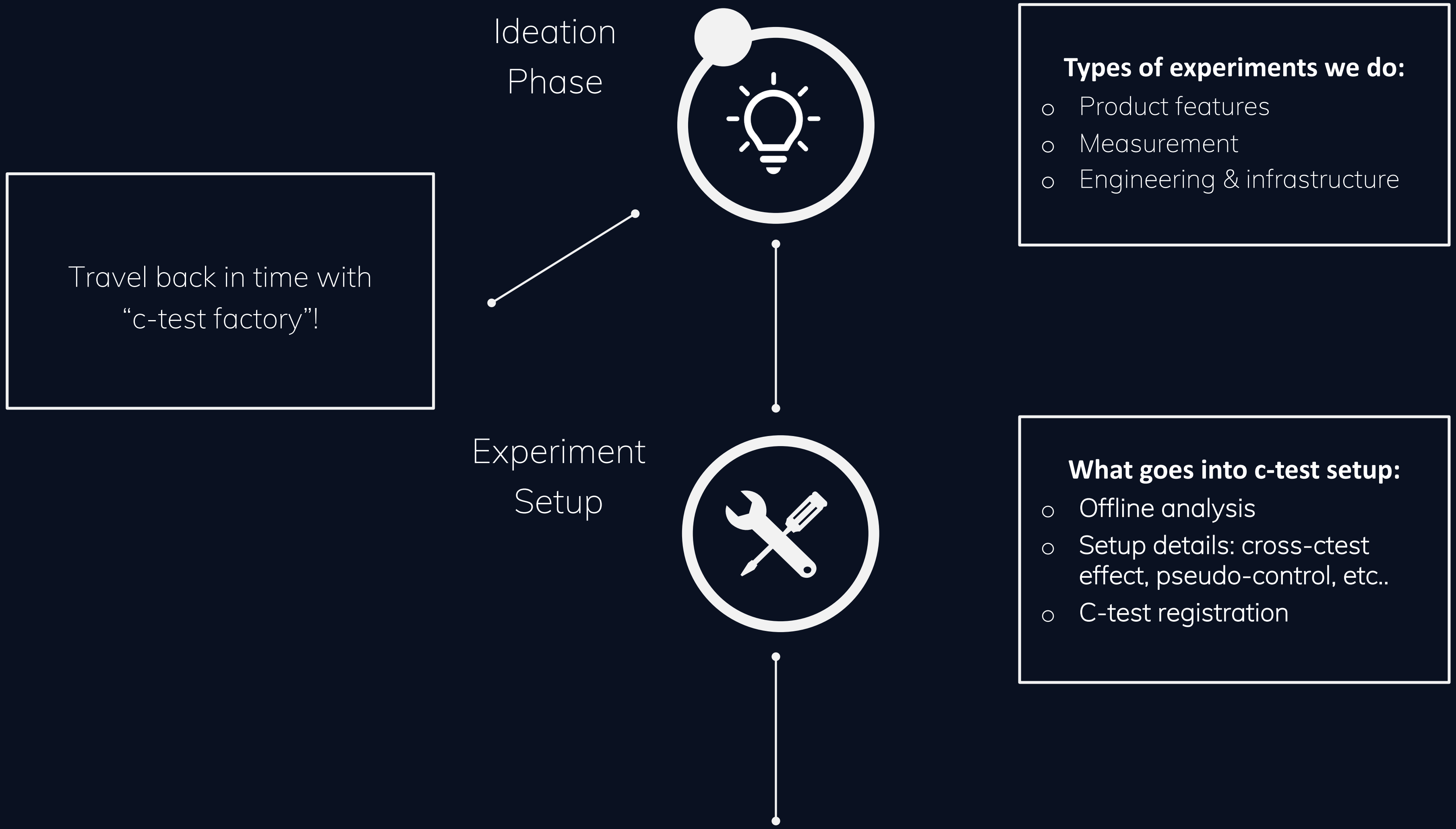
C-Testing

At trivago we make extensive use of C-testing to perform multi-group testing…

- We use C-testing to for evaluating product changes with our users.
    - Done both for front end UI/UX changes as well as backend changes as well (e.g. changes to the ranking model).
- C-testing is transparent to users and we can test specific groups such as by platform (web browser vs. Mobile Apps), locale, and by specific accommodations.
- Changes are significant if top line KPIs change more than 1% either way.
- C-tests are repeated to check if the result is reproducible.

The Art of Decision Making

C-testing lies at the heart of how we make decisions at trivago.



Ideation Phase

Experiment Setup

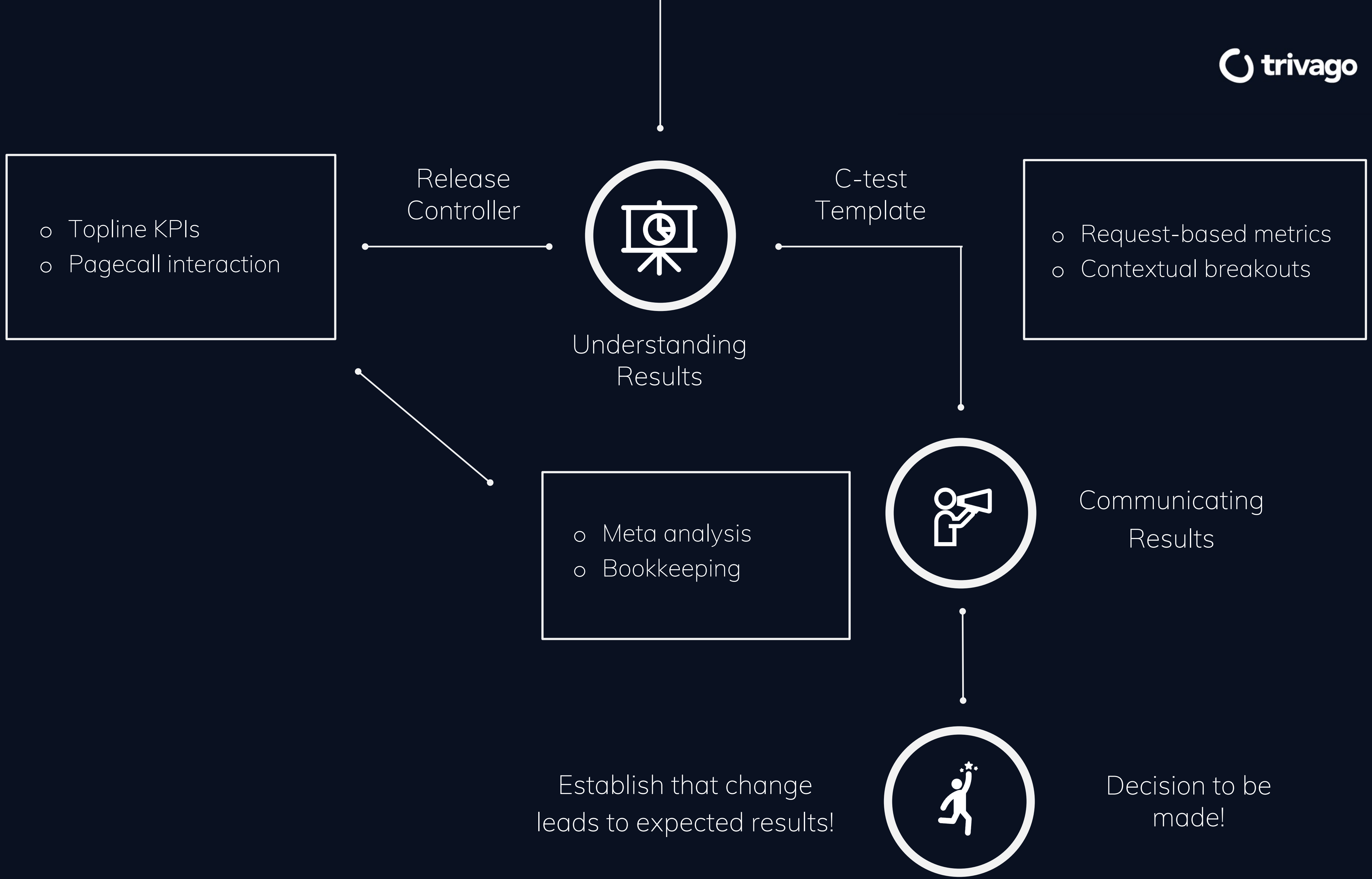
Travel back in time with “c-test factory”!

Types of experiments we do:

- Product features
- Measurement
- Engineering & infrastructure

What goes into c-test setup:

- Offline analysis
- Setup details: cross-ctest effect, pseudo-control, etc..
- C-test registration



- Topline KPIs
- Pagecall interaction

Release Controller

Understanding Results

C-test Template

- Request-based metrics
- Contextual breakouts

- Meta analysis
- Bookkeeping

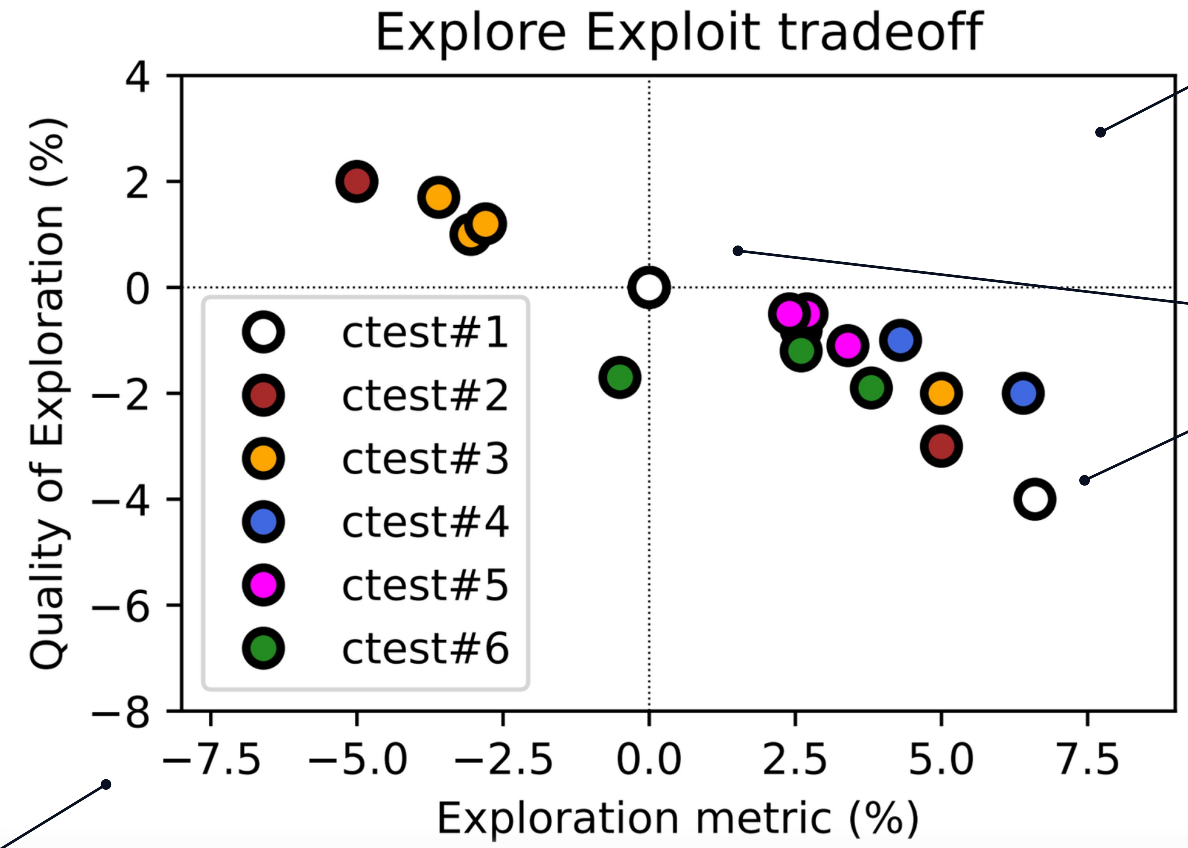
Communicating Results

Establish that change leads to expected results!

Decision to be made!

Meta Analysis

In the context of ranking accommodation at trivago-
show the ones you know perform well or ones that have never been shown before?



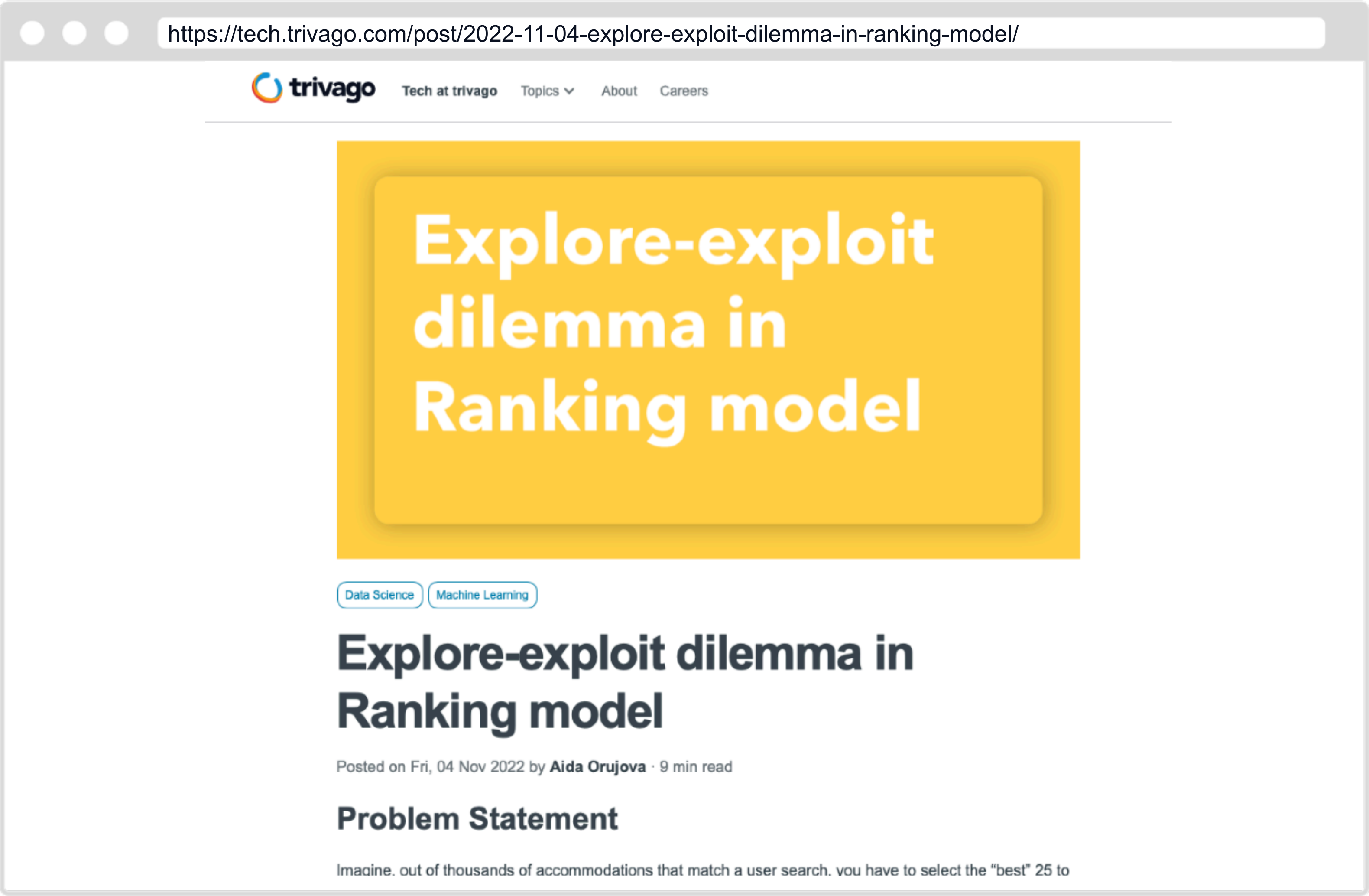
High exploration rate with high quality of the newly exposed accommodations

More exploration led to a drop in quality

Same color illustrate different levels of exploration for a particular test

Trade-off between extent and quality of exploration

Aida Orujov



https://tech.trivago.com/post/2022-11-04-explore-exploit-dilemma-in-ranking-model/

Explore-exploit dilemma in Ranking model

Explore-exploit dilemma in Ranking model

Posted on Fri, 04 Nov 2022 by **Aida Orujova** · 9 min read

Problem Statement

Imagine, out of thousands of accommodations that match a user search, you have to select the "best" 25 to

5. User Research & Evaluations

User Research & Evaluation

User Research is our main approach of evaluating changes qualitatively with users

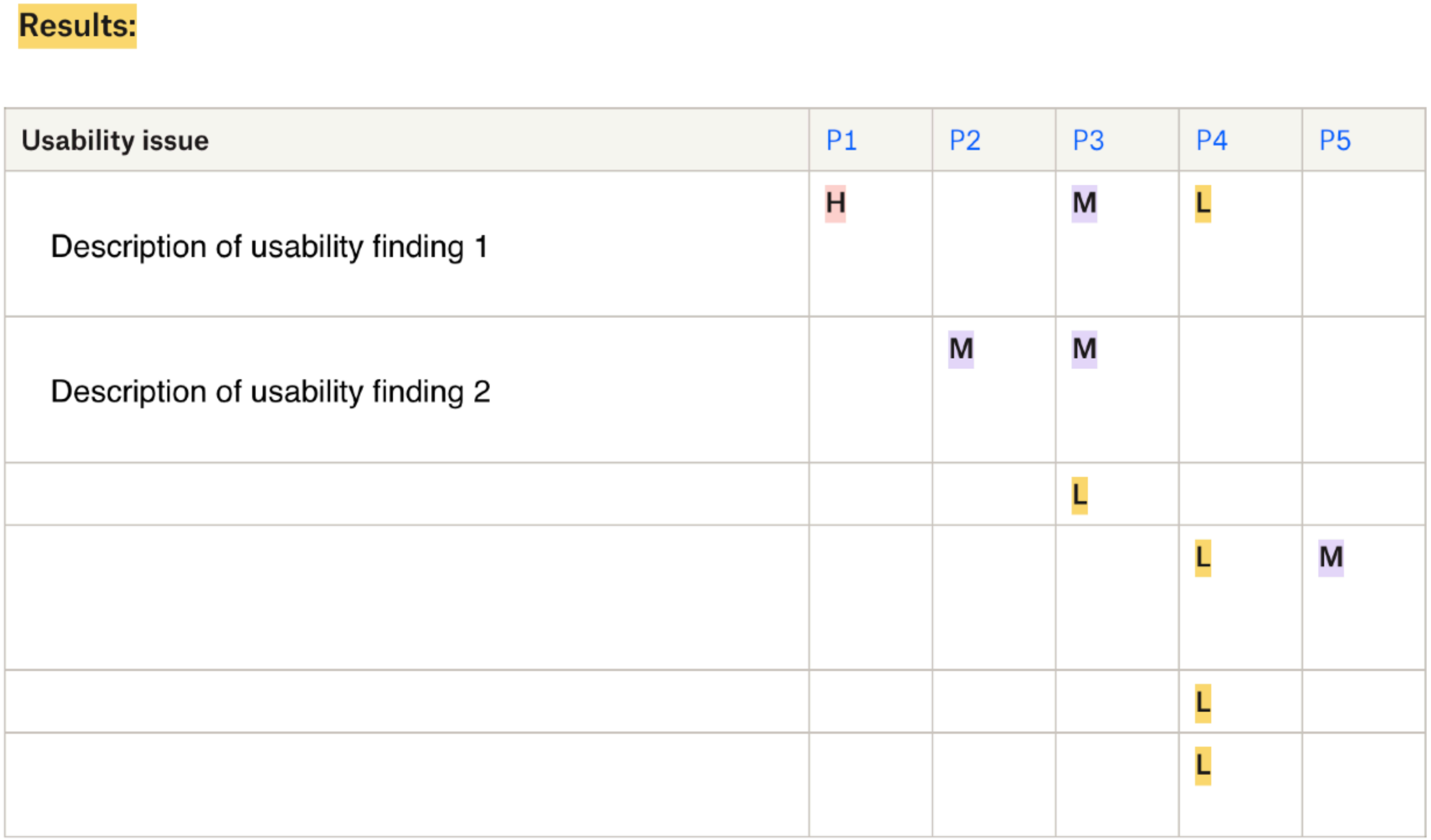
- We have several techniques for performing user research:
    1. Usability Testing
    2. Continuous Interviewing
    3. Diary Studies
        - Ask users to document their experiences over a period of time.
    4. User Interaction Recordings
        - Recorded for a small number of evaluation participants (1%).
        - Main interest is quantitatively seeing what users interact with.
    5. Surveys

Usability Testing

Usability testing is our way of evaluating designs and getting feedback from users…

- We recruit participants based on set of key characteristics: age, gender, employment status, country, web expertise, etc.
- Users are screened before testing using a questionnaire for suitability.
- Participants are given a scenario and a set list of tasks that they must accomplish. Their screens are recorded during the evaluation.
- The recorded is reviewed and we seek to assess:
    - User feedback
    - Task accomplishment success
    - Time taken by the user
    - User confidence in completing the tasks
- Results are aggregated over many users for a given experiment to produce a set of usability findings.

Usability Testing



Results:

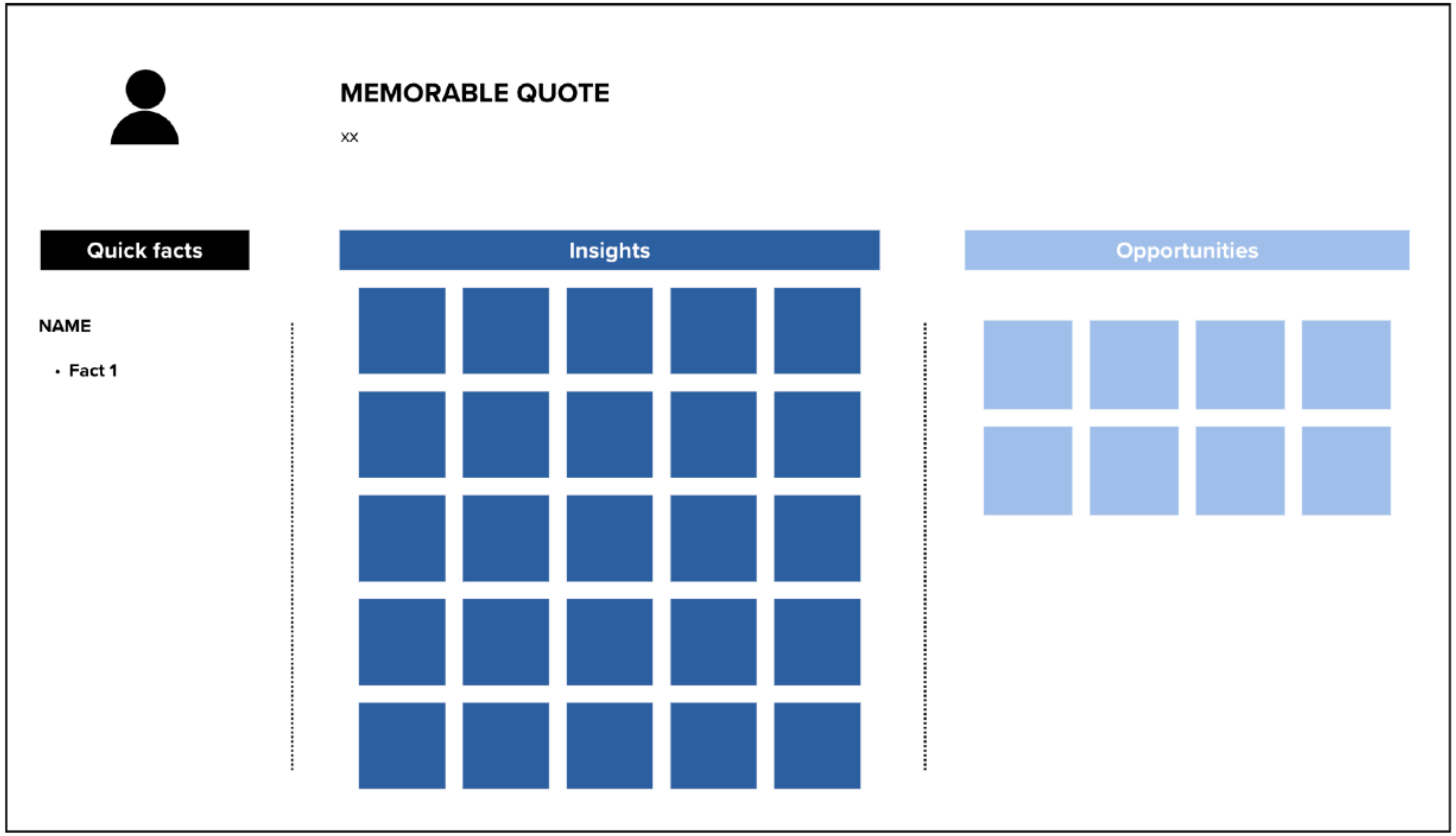
Usability issue	P1	P2	P3	P4	P5
Description of usability finding 1	H		M	L	
Description of usability finding 2		M	M		
			L		
				L	M
				L	
				L	

Continuous Interviewing

Another source feedback is through user interviews…

- We perform interviews with different cohort of users on a continual ongoing basis.
- Interviews are half-hour with a fixed interview script.
- Attempt to understand user needs and their pain points.
- Collate from the interviews recurrent themes and translate these themes into product action points.

Continuous Interviewing



MEMORABLE QUOTE

xx

Quick facts

NAME

- Fact 1

Insights

Opportunities

6. Summary

Summary

Evaluation within trivago is multi-faceted…

- Intrinsic quality evaluations using both human or third-party sources of ground-truth for our content based work.
- Extrinsic user-based evaluations using our C-testing framework to perform multi-group testing.
- Qualitative user-centric evaluations using methods such as interviews, diary studies, usability testing, etc.
- Key challenge with intrinsic evaluations is to create or find sources of truth to evaluate against.
- Evaluation result significance depends on business metrics or objectives.

trivago