FACULTY OF ENGINEERING
AND ARCHITECTURE

DEPARTMENT OF INFORMATION TECHNOLOGY
IDLAB - Internet technology and data science lab

Machinaal Leren: project

Pieter Simoens - Sam Leroux - Wei-Cheng Wang - Sander De Coninck

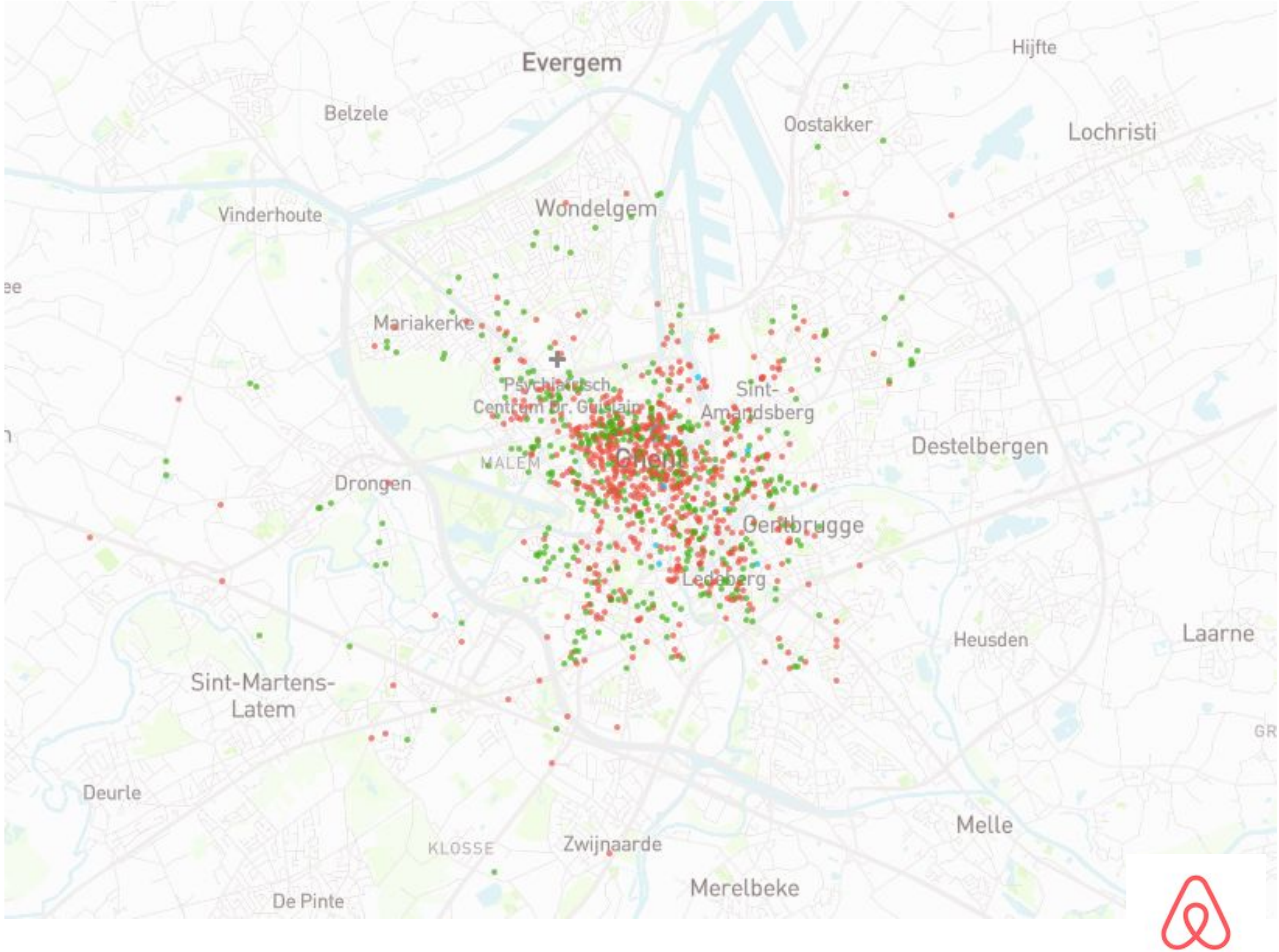
GHENT
UNIVERSITY

Project overview

- ~1000 AirBnb listings in Ghent
- Tabular, text and image data
- Explore this data and train predictive models

https://cloud.ilabt.imec.be/index.php/s/fXz9Q3EfAqe4zJc (Tabular data)

https://cloud.ilabt.imec.be/index.php/s/9t7xmxnxT7tmbnr (Image data: ~450MB)



Available data

Tabular data

ID	name	room_type	accommodates	bathrooms	bedrooms	beds	price
203806	Flat close to Sint-Pieters Station!	Entire home/apt	2	1	1	2	79

Text description

This well-located and comfortable one-bedroom flat on the ground floor of a city house allows you to discover Ghent at ease, whether you come by train or car. The heart of the city is at walking distance while you sleep in a quiet street.<br /><br /><b>The space</b><br />Within the city-centre of Ghent - 5 min walk from the main railway station Sint-Pieters, nearby musea (SMAK, STAM, De Bijloke) or a 15 min walk to the Graslei, Korenmarkt, you can rent a beautifully renovated ground-floor flat in a city-house.<br /><br />I am the owner of the house and we live on the second and third floor of the house. This makes it easy in case you have questions and also for the check-in and check-out.<br /><br /><b>Guest access</b><br />The flat is easily accessible for everyone as it is on the ground floor. You will have your own private kitchen with laundry facilities, a comfortable bedroom with direct access to the garden and a spacious bathroom with a dressing, cloth-hangers, and all bathroom n

Review text

This is a perfect location for a weekend or short stay in Gent! We felt at home immediately. Lot's of space and a gorgious bathroom. The neighbourhood is quiet, at a 10-15 min. walk from the city centre.

Images

Sprint 1: Tabular data

Explore the dataset to extract useful insights and predict/suggest the price for a new listing

Possible tasks:

- *Thorough* exploratory data analysis, e.g.:
  - Are there substantial price differences between neighbourhoods ?
  - Are there hosts with more than one listing ? How does this impact the price ?
  - What is the correlation between the review score and the price ?
  - ...

    Not enough to just show a plot! Clearly describe WHAT question you investigated, WHY you think this is a relevant question and WHAT you deduce/conclude from the results of your data analysis
- Are there outliers ?
- A new Airbnb owner needs to pick an appropriate price:
  - Train a model to predict the price based on a selection of features
  - Find the most similar listings
- ...

Deliverable: Jupyter notebook containing

- A structured description (**textual explanations** + code + plots) of how you solved the tasks.
- A table that shows what each team member worked on.

GHENT UNIVERSITY

Sprint 2: Text data

What insights can we gain from the text data (title, description and reviews) ?

Possible tasks:

- Detect duplicate listings
- Extract keywords from reviews / descriptions
- Automatically make a list of positive and negative points for a listing based on the reviews
- Recognize listings from the same owner
- Detect anomalies (listings/ reviews that are very different from other listings/ reviews)
- Detect reviews that are very similar
- Perform sentiment analysis on a review
- ...

Deliverable: Jupyter notebook containing

- All the code needed for your analysis, interwoven with a textual description of your reasoning and findings
- A table that shows what each team member worked on.

Sprint 3: Multiple data sources

Use the images or external data sources (e.g. Ghent open data)

Possible tasks:

- Use the images to find similar listings (e.g. similar interior style).
- Cluster the images
- Automatically detect attributes of the listing (e.g. garden, bath, shower, washing machine, …) based on the images
- Find duplicate listings based on the images
- Detect anomalies (rooms that look very different)
- Predict which room a picture is taken in (bedroom, bathroom, outside, …)
- Combine with external data to better predict the price (e.g. location of public transport, proximity to attractions, ...)
- ...

Deliverable: Jupyter notebook containing

- All your code + a description of your approach, reasoning and findings
- A table that shows what each team member worked on.

Practicalities

- Groups of 3. **Subscribe to a group at Ufora - Ufora tools - Groups - select “project” as category**

- You decide what models you develop, what preprocessing you use, ...

- You can use any software package, pretrained models, public code, external data, blog posts, … you want
  - BUT: clearly mention this in your report
  - Do not use them as a black box, you should be able to give a technical description of how they work.
  - Go further than the typical “AirBnB EDA” blogpost

- It is fine if you restrict the problem (e.g. only train models for listings in one language, clean the data, ...). But you should clearly motivate your choices.

- The end result of each sprint should be a report that is understandable for someone with a limited ML background + an explanation of your thought process + why you decided to do certain experiments + all the code and technicalities needed to reproduce your findings.

- Focus on best practices and creativity
  - High “accuracy” does not necessarily imply high marks (e.g. if you just applied existing code without much tuning)
  - Low “accuracy” does not necessarily imply low marks (e.g. if you developed an experimental, novel, ambitious, creative approach.)

Compute resources

- Free cloud infrastructure: Binder.org, AWS, Azure, ...
- Google Colab (includes free GPU access)
- Contact us if you need more compute resources (CPU or GPU)

Feedback and evaluation

- 40% sprint 1 and sprint 2
- 20% sprint 3 + presentation
- 40% theory exam

- Intermediate progress reports: Jupyter Notebooks but focus on presenting your results clearly.

- Sprint deadlines:
  - Sprint 1: week of 19/10
  - Sprint 2: week of 16/11
  - Sprint 3: week of 20/12

- Final presentation in exam period. Details to be announced later.
  - Only about sprint 3
  - Tell us what you *learned/concluded* from the data, not only what you *did*

GHENT
UNIVERSITY