An exploration of recipes

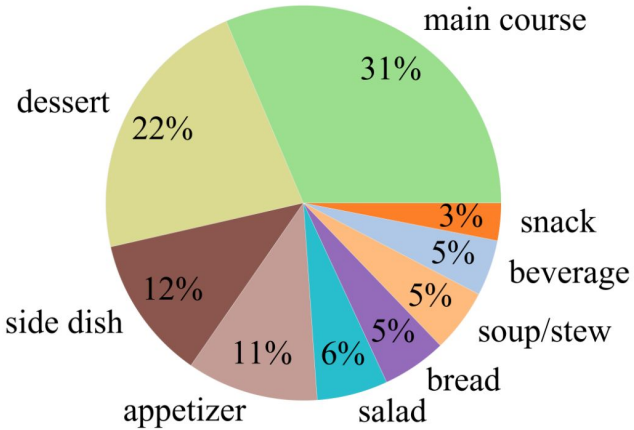
- Andrei-Calin Furtuna
- Paul-Adrian Gafton
- Sorin-Sebastian Mircea
- Alexandru Mocanu

EPFL

Dataset – Recipes 1M

- contains one million structured cooking recipes with 13M associated images
- used subset of 51k recipes
- the attributes of the recipes that we use are the recipe name, its ingredients, the quantities used from each ingredient, the nutritional values per 100g and cooking instructions

- aim to gain more insights into how various dishes are related and also into the recipes themselves



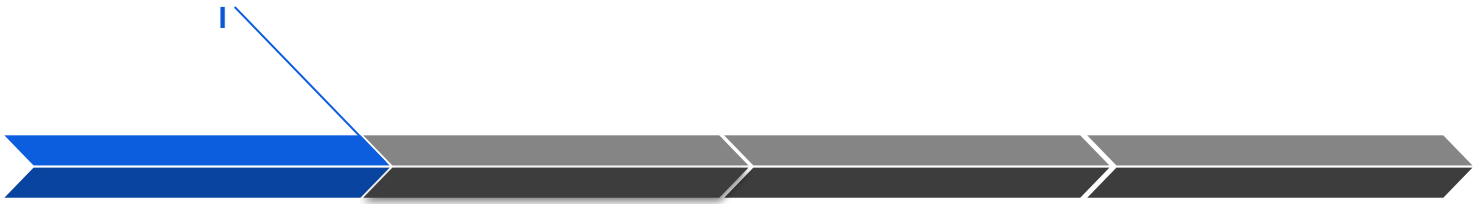
Dataset - Recipes 1M

- The dataset includes a few errors due to the used simplification algorithms. For example sometimes `olive oil` is incorrectly mapped to [*olive*, `oil`, `olive oil`]. Additionally some ingredients may be wrongly split up: `yellow`, `dry`, `prepared`, `other`

Example ingredients from recipes

```
In [2]: ingredients[recipes[0]]
Out [2]: array(['basil leaves', 'focaccia', 'leaves', 'mozzarella', 'pesto',
                'plum tomatoes', 'rosemary', 'sandwiches', 'sliced', 'tomatoes'])
```

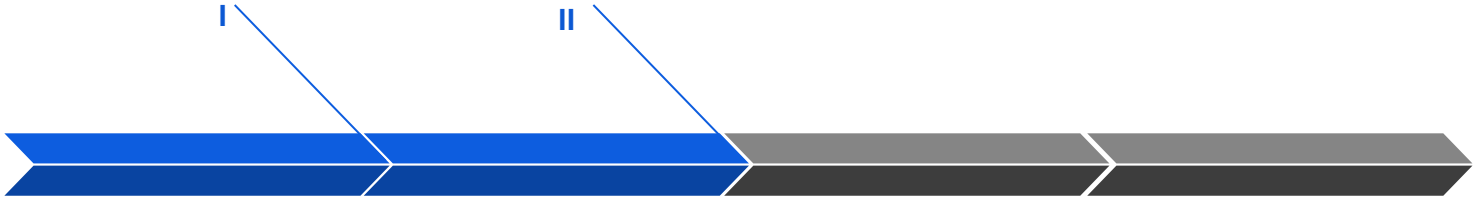
Data preprocessing and graph generation



Ingredient names

Extract relevant
names for ingredients

Data preprocessing and graph generation



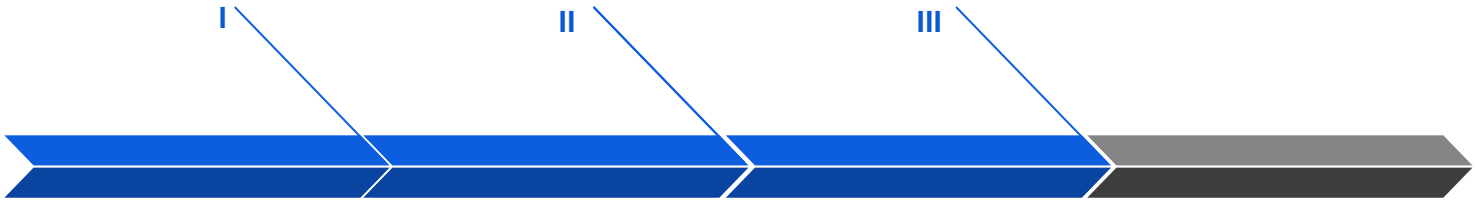
Ingredient names

Extract relevant names for ingredients

Dataframes

Generate ingredient and recipe dataframes with information for exploration and exploitation

Data preprocessing and graph generation



Ingredient names

Extract relevant names for ingredients

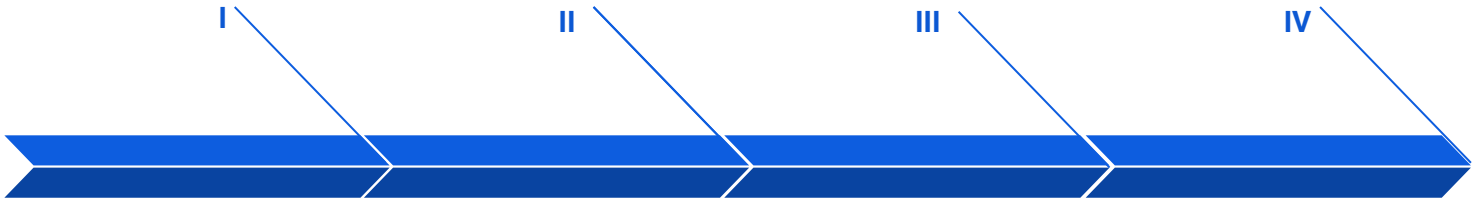
Dataframes

Generate ingredient and recipe dataframes with information for exploration and exploitation

Ingredients graph

Create ingredients graph by connecting ingredients occurring together

Data preprocessing and graph generation



Ingredient names

Extract relevant names for ingredients

Dataframes

Generate ingredient and recipe dataframes with information for exploration and exploitation

Ingredients graph

Create ingredients graph by connecting ingredients occurring together

Recipes graph

Create recipes graph by ignoring common ingredients and downsampling edges

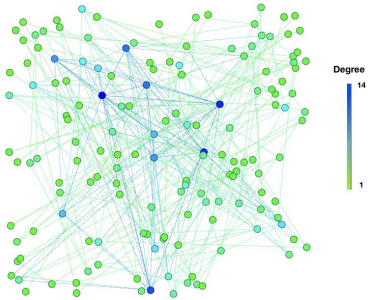
Graphs

Recipes graph

- a recipe graph, where vertices are associated to recipes and the edges are weighted by a sum of tf-idf-like terms

$$t_{ij,k} = \frac{2}{n_i + n_j} \log\left(\frac{N}{N_k}\right)$$

Nodes	Edges
51 K	12.8 M

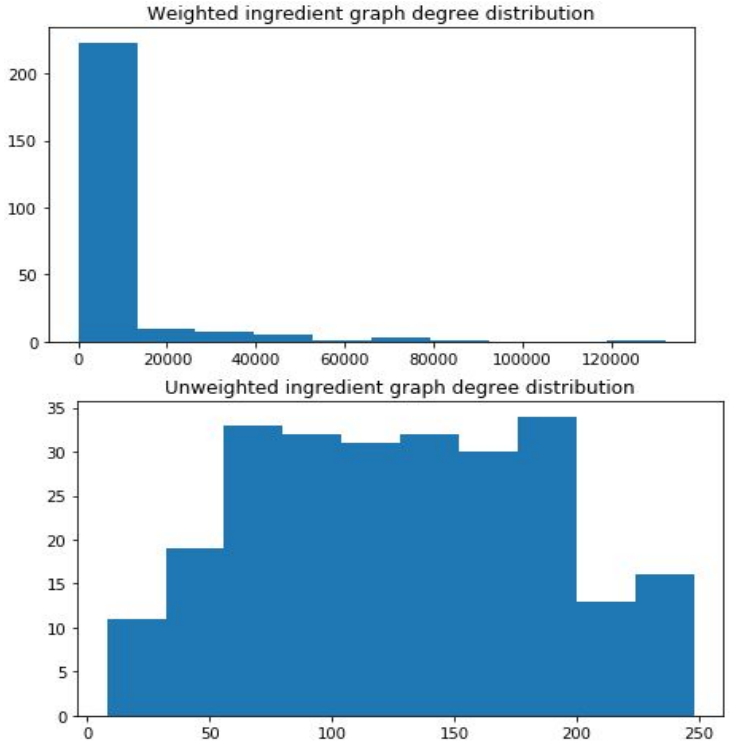


Ingredients graph

- an ingredient graph, where vertices are associated to ingredients, and the edges are weighted by the number of recipes in common

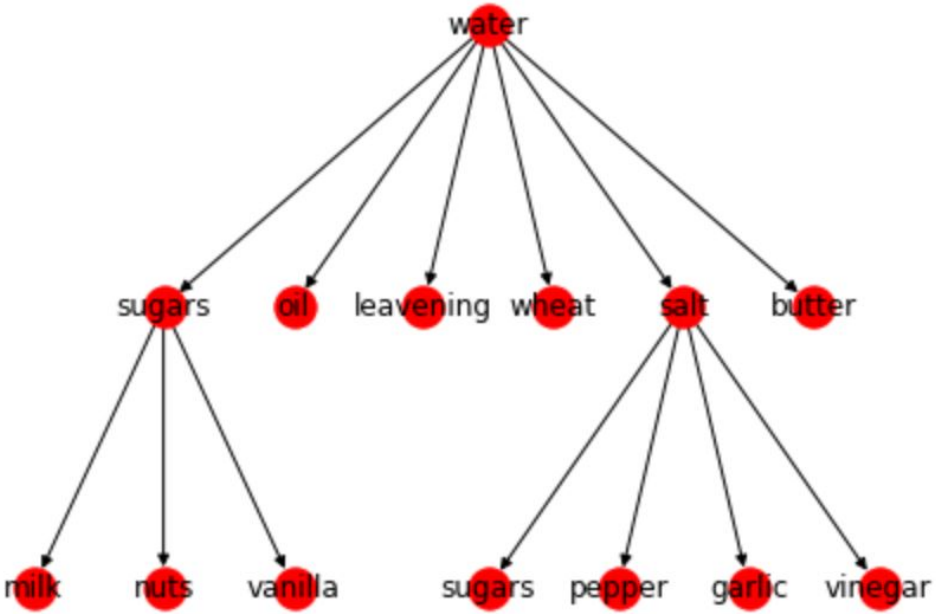
Nodes	Edges
251	16 K

Ingredients Graph Exploration

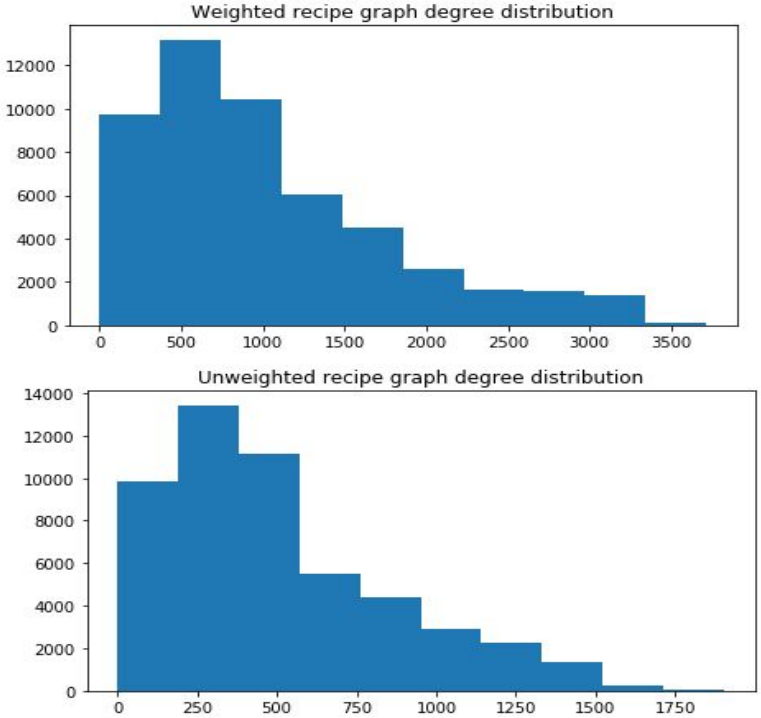


- the unweighted graph has a fairly uniform degree distribution, suggesting a dense graph

Ingredient Co-occurrence

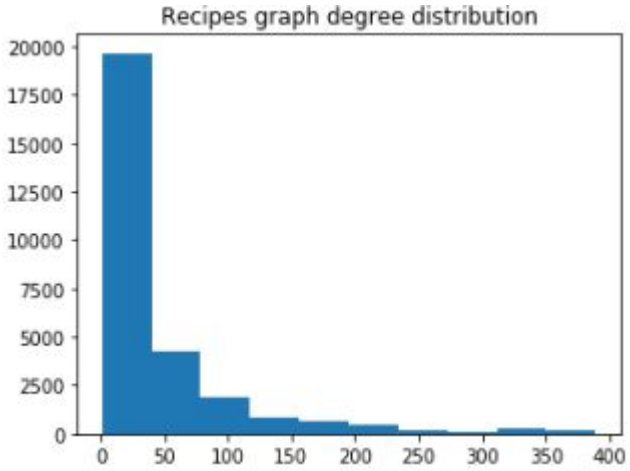


Recipe Graph Exploration

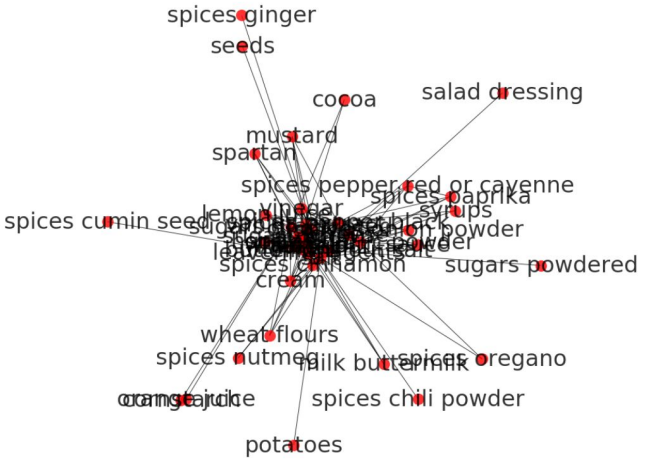


- the degree distribution does not show any hint of scale-free behaviour
- the distributions for both the weighted and unweighted graphs are similar, since most of the weights are around 2

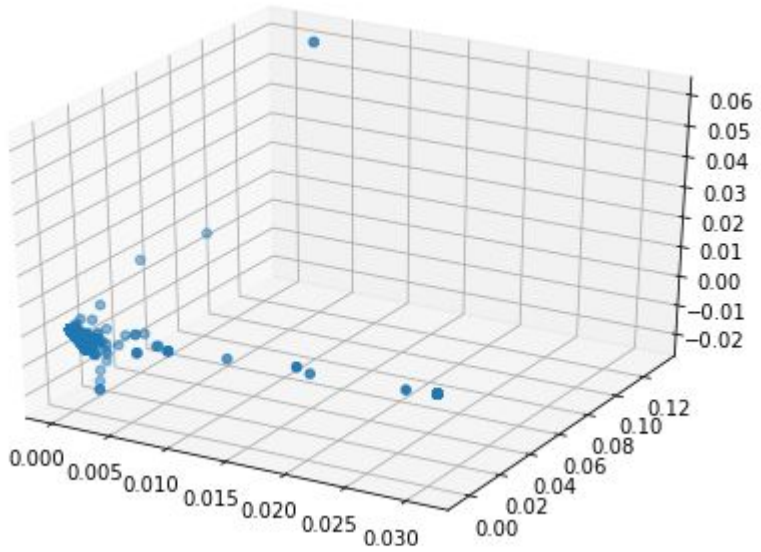
Recipe Graph Exploration



- downsampled the graphs



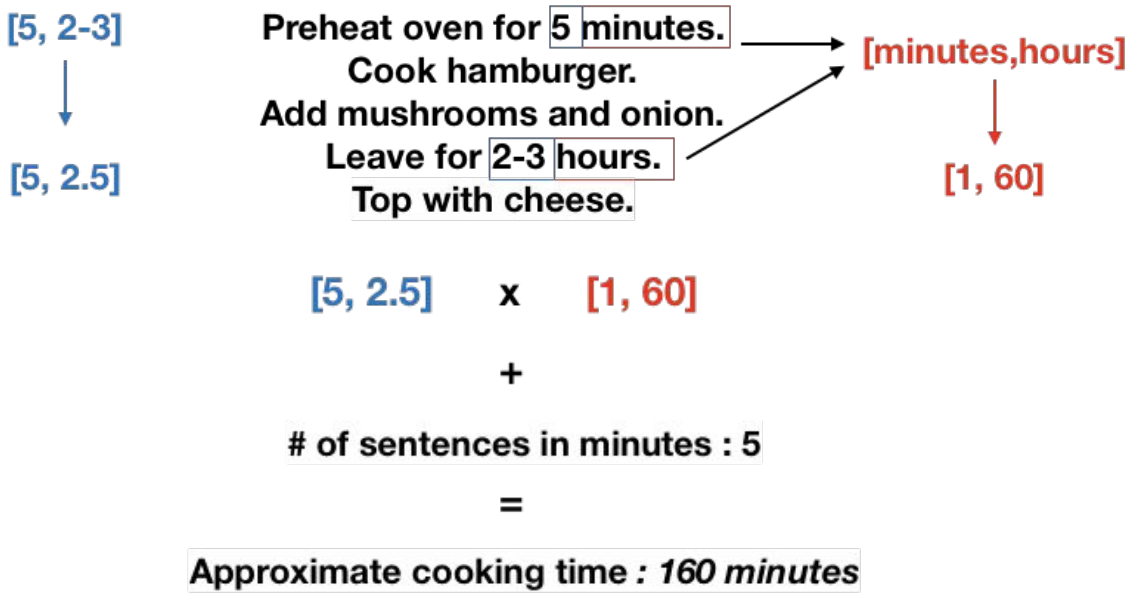
Looking at vanilla



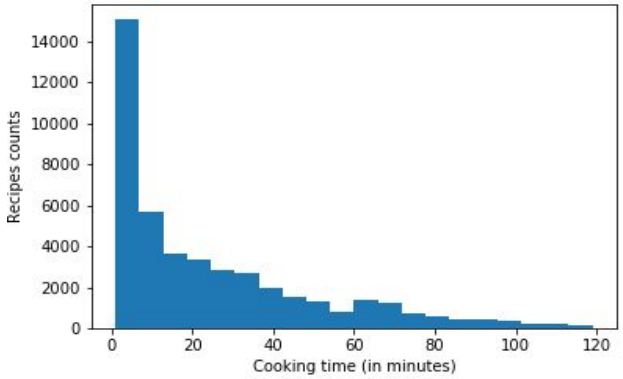
- 3D embedding of the subgraph of recipes containing vanilla

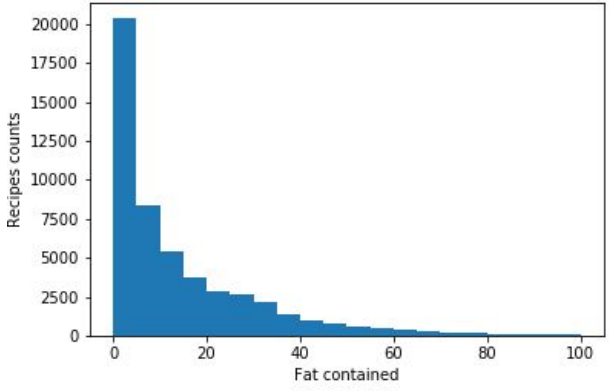
Cooking time and healthy food

Extract cooking time from recipes instructions



Cooking time and healthy food

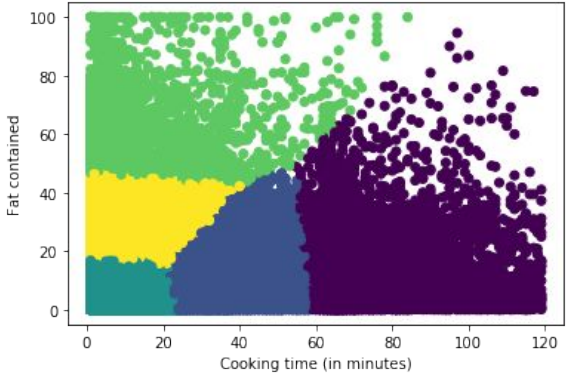




How to eat fast and healthy ?

- What is “healthy” ?
  - Calories a day for an adult : ~ 2.000
  - 20% to 35% calories should come from fat
    ⇒ < 80 grams of fat per day
    ⇒ ***~ 25 grams per meal***

- Identify healthy and fast recipes (~ 20 minutes)



- Smoothies
- Dips
- Salads
- Soups
- Deserts

Conclusions

Observations

- graphs construction
- graphs do not belong to some clear category
- clustering of recipes

Challenges

- relevant ingredient names generation
- size of the dataset ⇒ aggressive downsampling was needed
- graph structure usefulness