Computing cocktail flavors: Text-mining user-generated websites for sensory data

Jacob Lahne, PhD (jlahne@vt.edu)

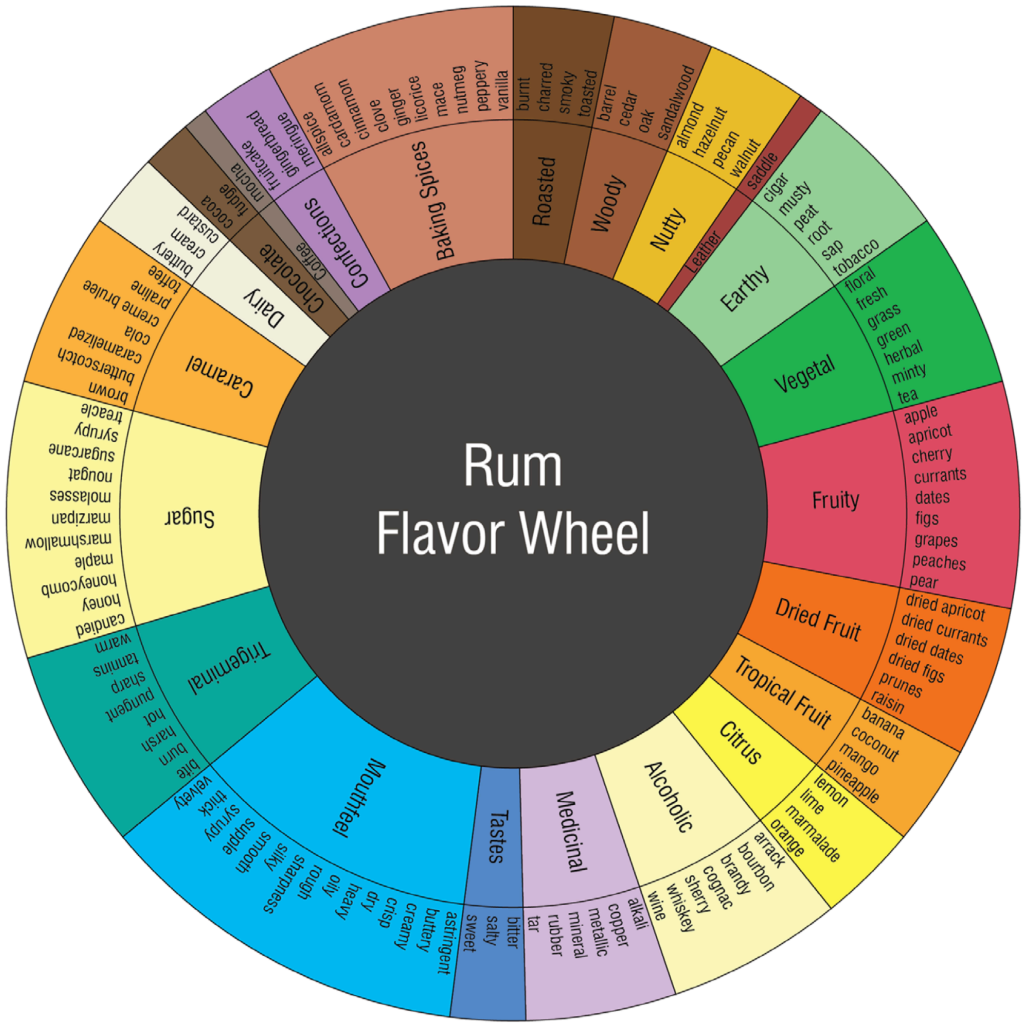
Department of Food Science & Technology, Virginia Tech, USA

Eurosense 2020

Sensory data from free text

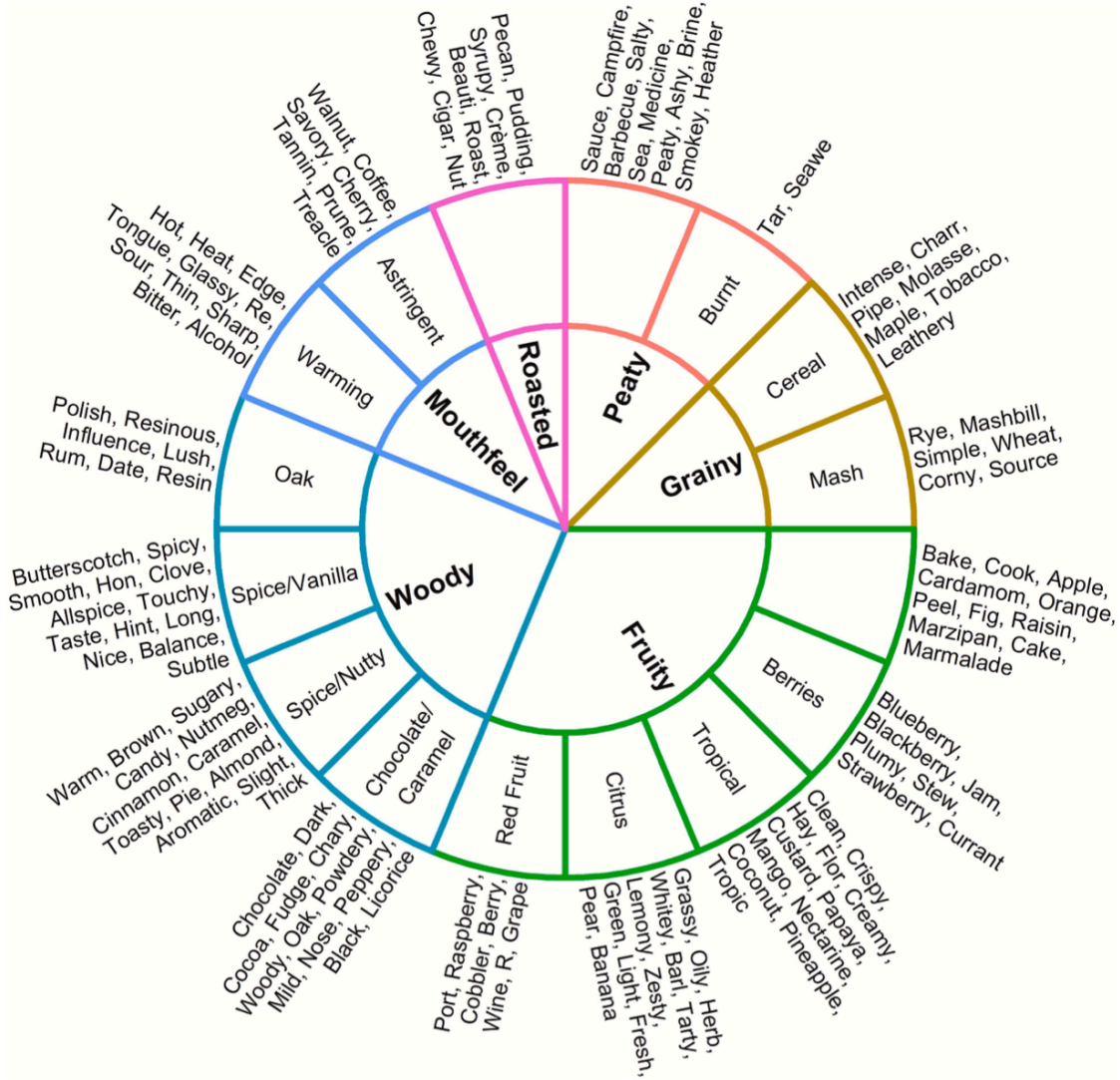
- Recently, sensory scientists have begun to investigate **unstructured, free text as a source of sensory information**
  - **Affect and sentiment**: Visalli et al. (2020), Luc et al. (2020)
  - **Description**: Bécue-Bertaut et al. (2008), Ickes et al. (2017), Hamilton and Lahne (2020), Mahieu et al. (2020)
- A **range of sources** have been used to generate free-text data
  - Solicited, open-text comments in survey forms
  - Published, topic-relevant texts, usually by experts (e.g., wine guides)
  - Unsolicited comments and reviews (usually from websites)
- Dealing with free text requires new approaches:
  - Prioritizes **emic** sensory knowledge that requires **different theory and analysis**
  - **Actually obtaining these data can be a real challenge for sensory scientists**

Over 400 websites manually processed for sensory descriptions of rum.



Ickes et al. 2017

Over 6500 online reviews processed automatically to obtain sensory descriptors for whisk(e)y.



Hamilton & Lahne 2020

Presentation agenda

- This presentation will provide a basic walkthrough on **how to obtain and process free-text data from a website**
- We are going to be looking at a favorite website: **Cocktail Virgin**
  - "Cocktail blog" documenting the Boston bartending scene since 2007.
  - Publishes ~1 drink/day with recipe + commentary
  - Standardized format: Title / recipe / picture / **commentary** / ingredients
  - I have permission from the website owner (Frederic Yarm) to scrape the site
- The following topics with small demos **in `R`**:
  - Parsing and scraping HTML/CSS (will not cover JavaScript) using **`rvest`**
  - Breaking up text using Regular Expressions using **`tidytext`**

Let’s look at a website!

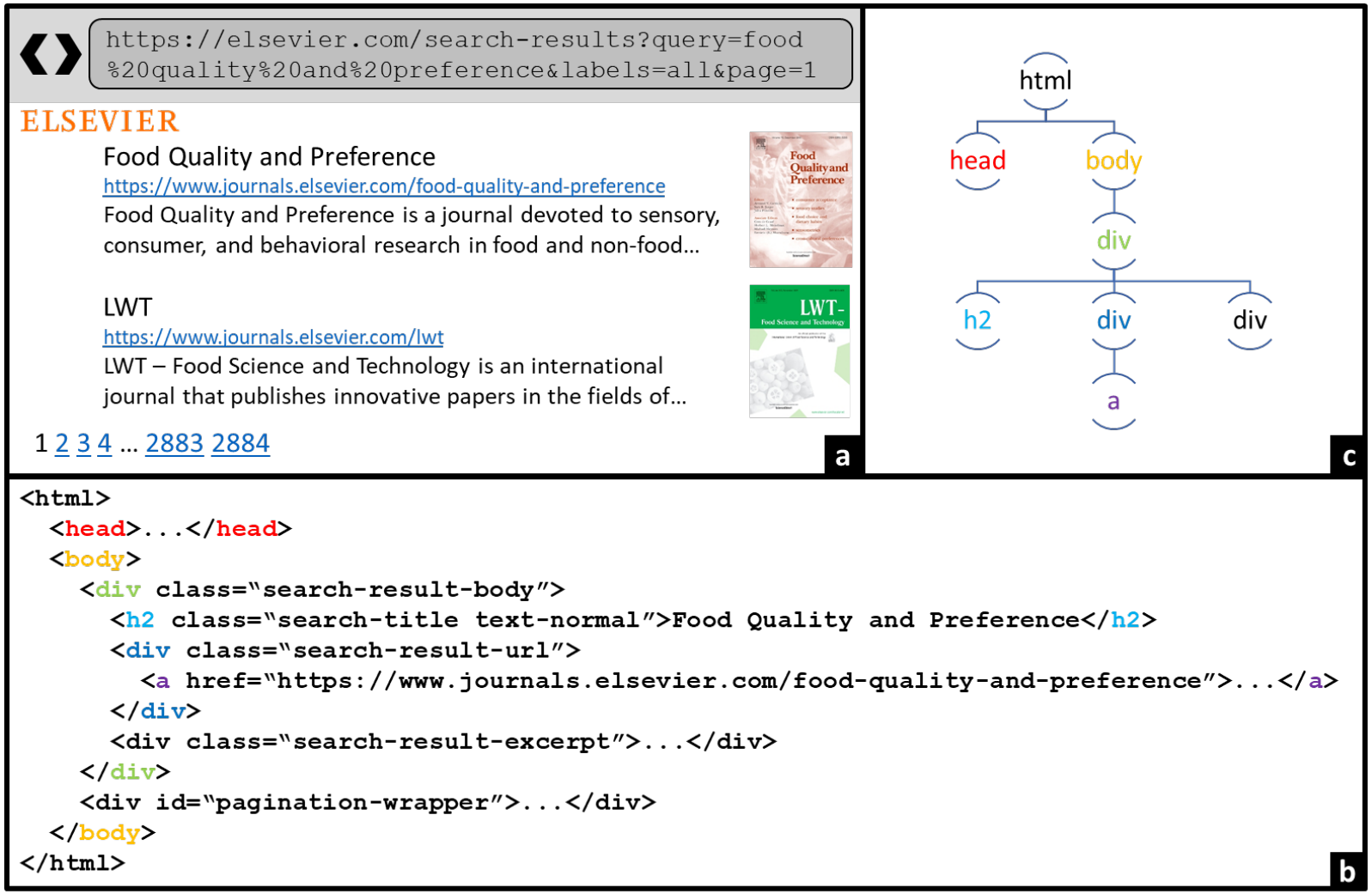
Cocktail Virgin (most recent posts)

First Post

Resources: CSS Diner (basic selection), Interneting Is Hard (advanced/comprehensive)

Looking at web data—HTML/CSS

- Observation + trial and error helps find HTML/CSS tags that identify relevant information
- Use existing software—here `R` + `rvest`—to access and extract what we want



Hamilton & Lahne 2020

Code for getting data from our website

Let’s look at a post with the data we really want

An in-depth and friendly tutorial to writing scrapers in R

scan for code from this presentation! →

Extracting information (parsing the text)

- We used control loops (like `while()`) to get all posts using our basic tools (not shown)
- We have obtained data in the form we wanted: **unstructured, free text**
- How can we make something useful from these data?
  - We will break up our ingredients field into individual ingredients using a fixed string
  - We will use a regular expression ("**regex**") to split descriptions from recipes
  - We will use **`spacyr`** to identify adjectives as potential descriptors
- A larger analysis would act on the full data set (~5000 posts at this date) and involve far more pre- and post-processing

A toy dataset for examination

We will play with a bit of a larger dataset from Cocktail Virgin in R

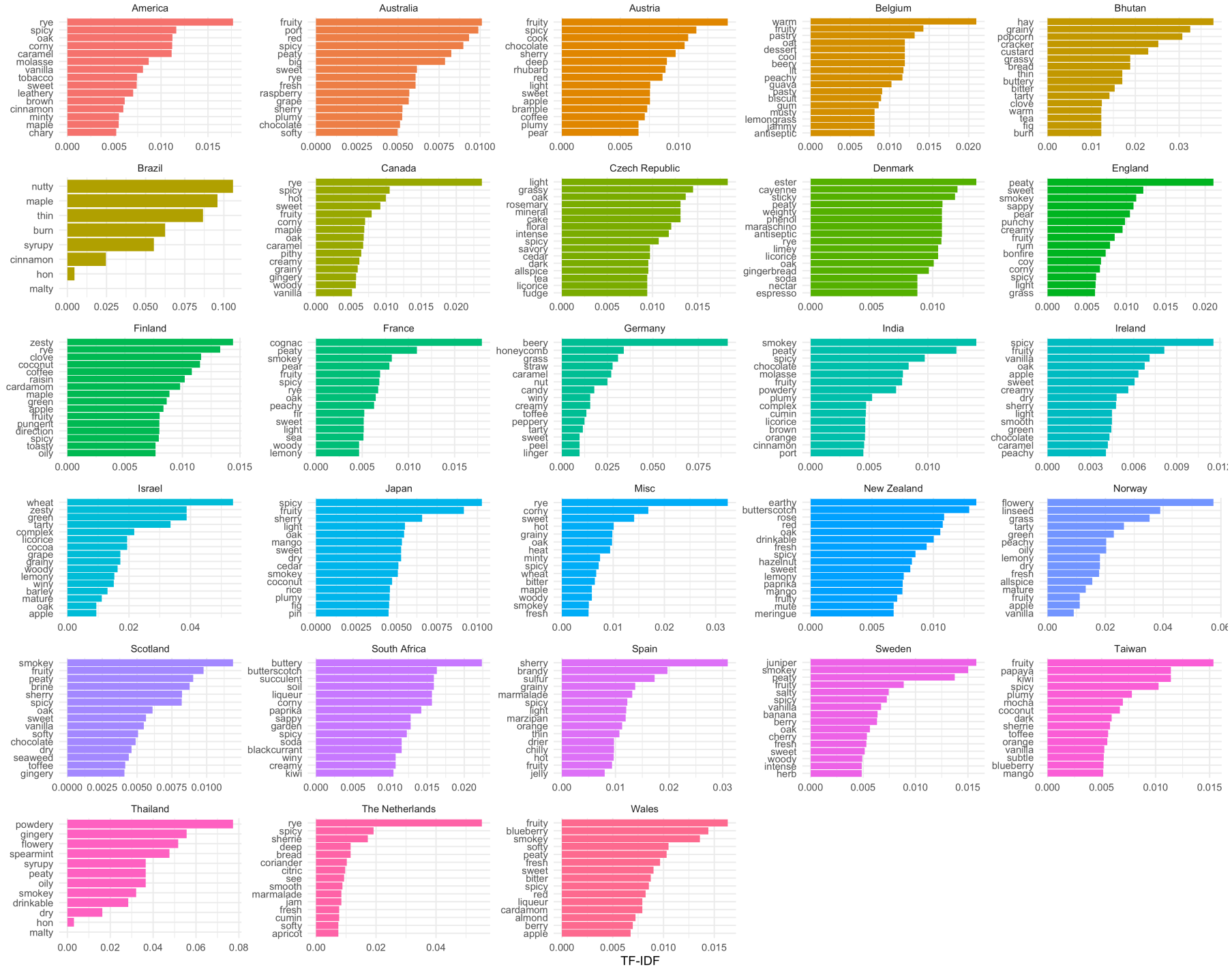
Resources: Text Mining with R (great intro text), RegexOne (interactive regex tutorial)

scan for code from this presentation! →

Automatic flavor identification for **whiskeys**

Generated using the methods detailed in this presentation:

1. Text scraping
2. Text pre-processing
3. Text tokenizing
4. (+ NN/LSTM magic)
5. Term Frequency – Inverse Document Frequency



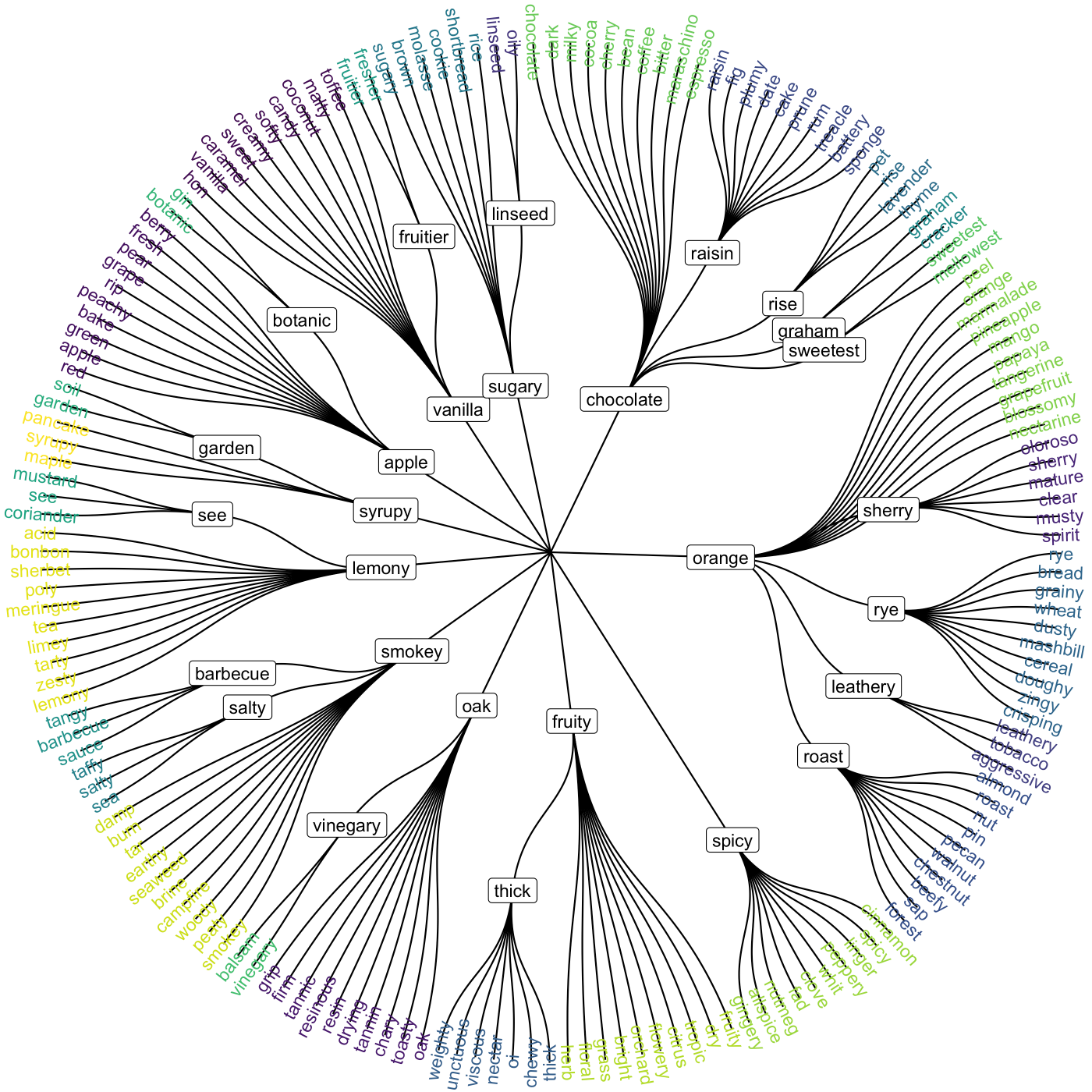
Automatic flavor wheel for **whiskeys**

Following up the previous workflow with tools from **network analysis** allows for examination of the relationship between pairs of descriptors.

The “flavor wheel” here is generated without any human input on similarity among descriptors.

Group labels are generated by raw frequency account (this can be improved in future iterations).

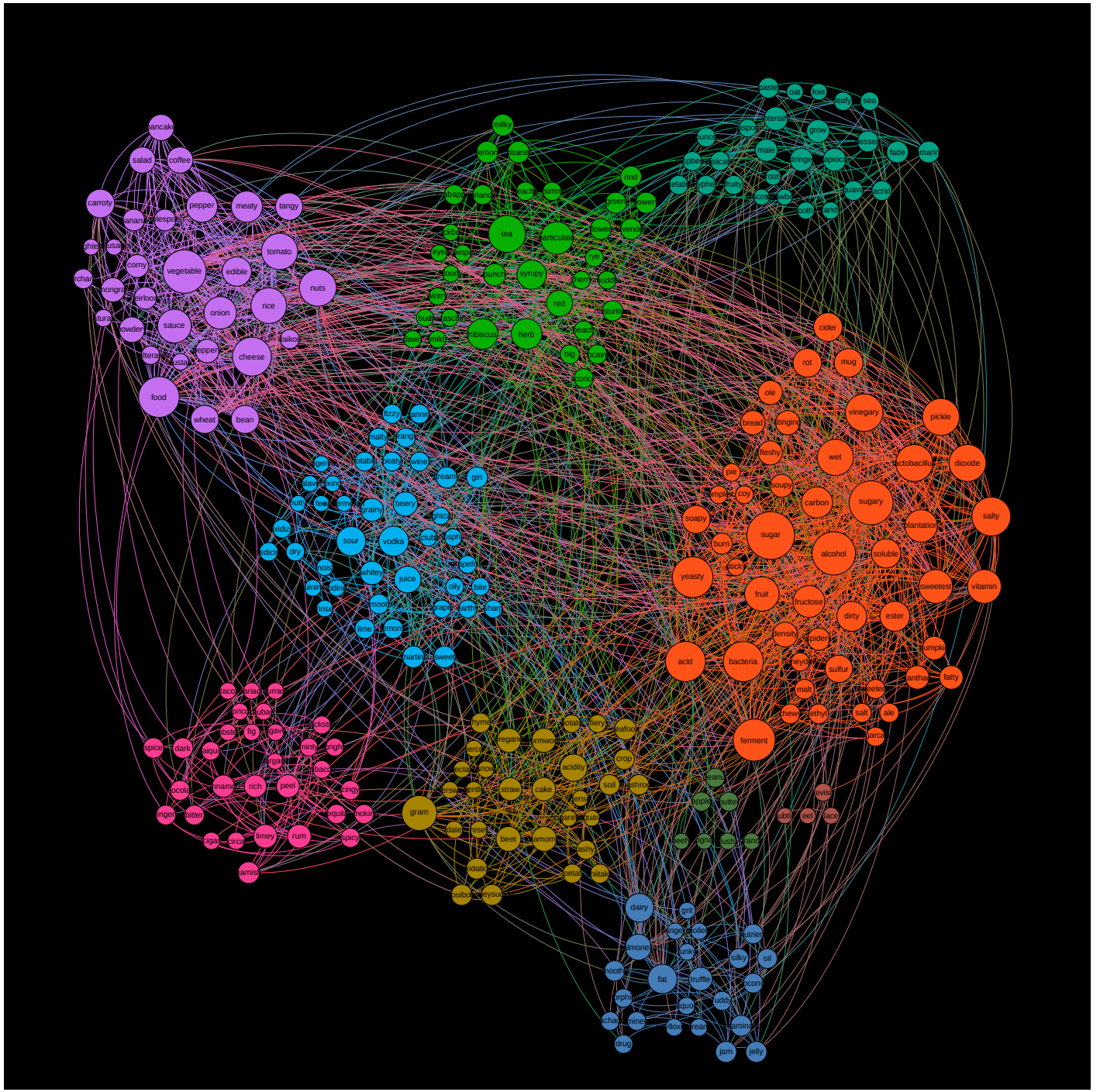
Clear patterns are evident.



Conclusions and future work

- Acquiring and processing free-text data can be a barrier for sensory scientists, but there are abundant resources to support this activity
- Free-text data (especially of unsolicited comments/text) can offer types of scale and insight that are unavailable with traditional methods
  - Potentially thousands of products
  - Emic and observer-uninfluenced reporting
- An algorithmic (coding) approach to these data can make analysis of these data feasible
- Future sensory science will incorporate both closed- and open-ended analyses to produce better insights

Questions? Email me: jlahne@vt.edu



Cited work

1. Visalli, M., Mahieu, B., Thomas, A., & Schlich, P. (2020). Automated sentiment analysis of Free-Comment: an indirect liking measurement? *Food Quality and Preference*, 103888. https://doi.org/10.1016/j.foodqual.2020.103888
2. Luc, A., Lê, S., & Philippe, M. (2020). Nudging consumers for relevant data using Free JAR profiling: An application to product development. *Food Quality and Preference*, *79*(August 2019), 103751. https://doi.org/10.1016/j.foodqual.2019.103751
3. Bécue-Bertaut, M., Álvarez-Esteban, R., & Pagès, J. (2008). Rating of products through scores and free-text assertions: Comparing and combining both. *Food Quality and Preference*, *19*(1), 122–134. https://doi.org/10.1016/j.foodqual.2007.07.006
4. Ickes, C. M., Lee, S.-Y., & Cadwallader, K. R. (2017). Novel creation of a rum flavor lexicon through the use of web-based material. *Journal of Food Science*, *82*(5), 1216–1223.
5. Hamilton, L. M., & Lahne, J. (2020). Fast and automated sensory analysis: Using natural language processing for descriptive lexicon development. *Food Quality and Preference*, *83*, 103926. https://doi.org/10.1016/j.foodqual.2020.103926
6. Mahieu, B., Visalli, M., Thomas, A., & Schlich, P. (2020). Free-comment outperformed check-all-that-apply in the sensory characterisation of wines with consumers at home. *Food Quality and Preference*, *84*, 103937. https://doi.org/10.1016/j.foodqual.2020.103937