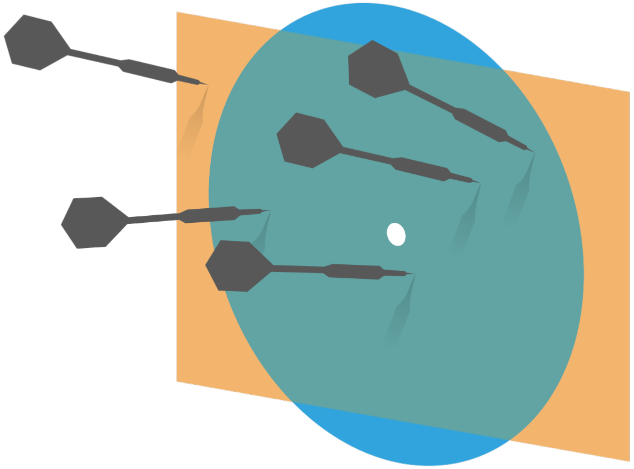
Probability and Statistics for Computer Science



Credit: wikipedia

"Unsupervised learning is arguably more typical of human and animal learning…"--- Kelvin Murphy, former professor at UBC

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 04.29.2021

Last time

- Curse of dimensions
- Unsupervised learning
- Clustering

Objectives

Q. Is k-means clustering deterministic?

A. Yes

B. No

K-means clustering example: Portugal consumers

- The dataset consists of the annual grocery spending of 440 customers

- Each customer's spending is recorded in 6 features:
  - fresh food, milk, grocery, frozen, detergents/paper, delicatessen

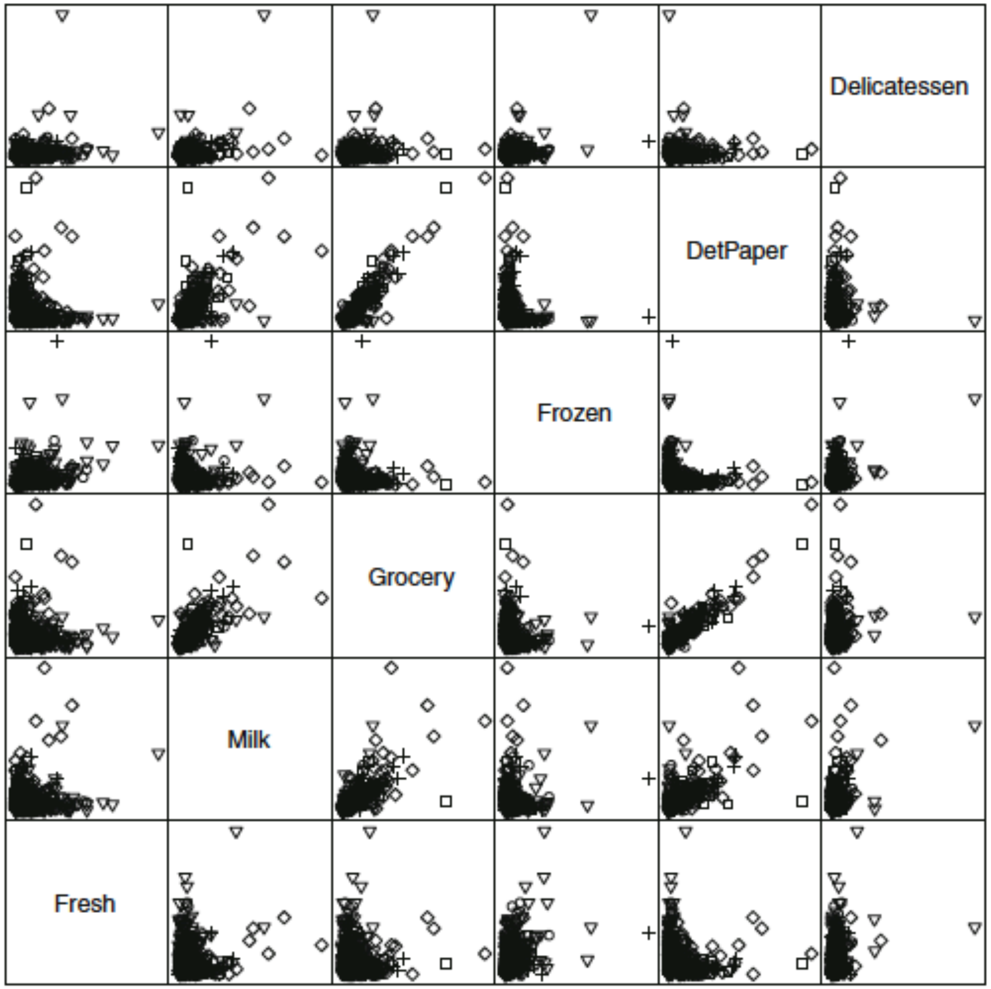
- Each customer is labeled by: 6 labels in total
  - Channel (Channel 1 & 2) (Horeca 298, Retail 142)
  - Region (Region 1, 2 &3) (Lisbon 77, Oporto 47, Other 316)

Lisbon, Portugal

Oporto, Portugal

Visualization of the data

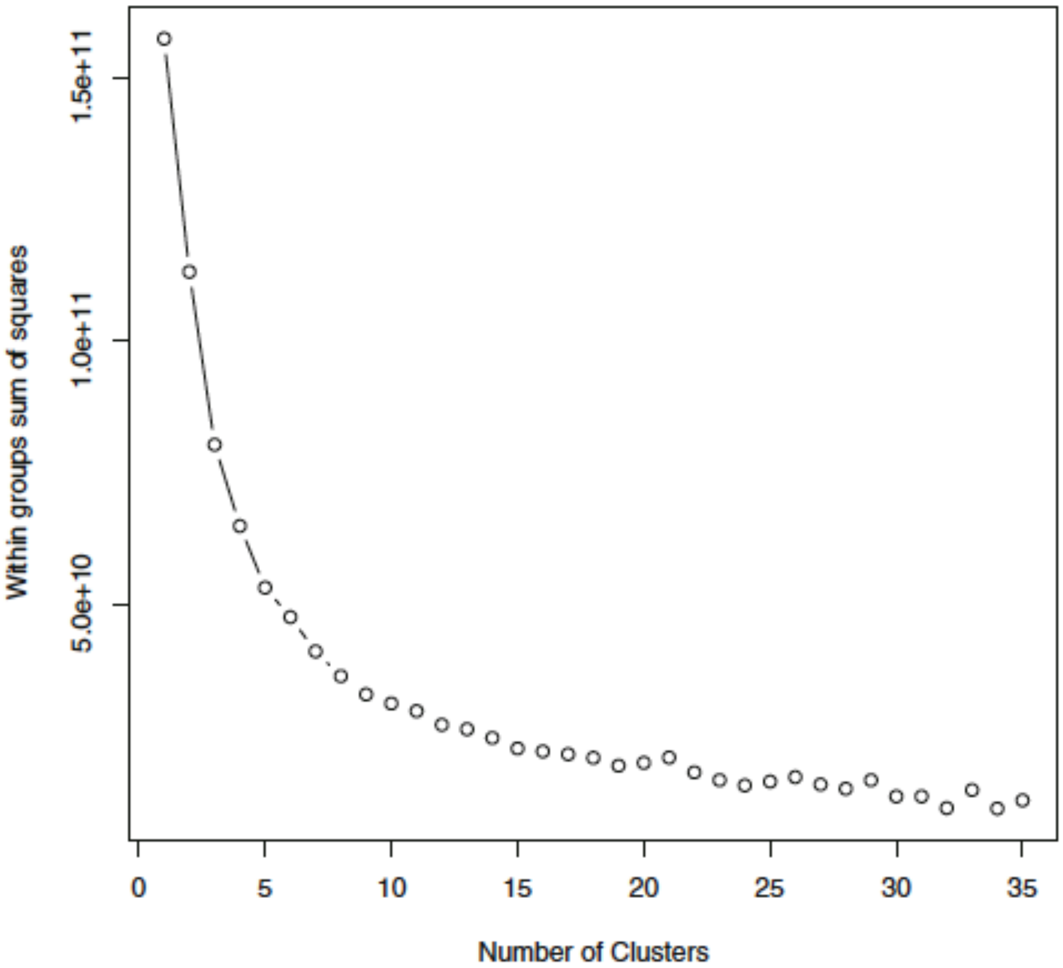
- Visualize the data with scatter plots
- We do see that some features are correlated.
- But overall we do not see significant structure or groups in the data.



Scatter Plot Matrix

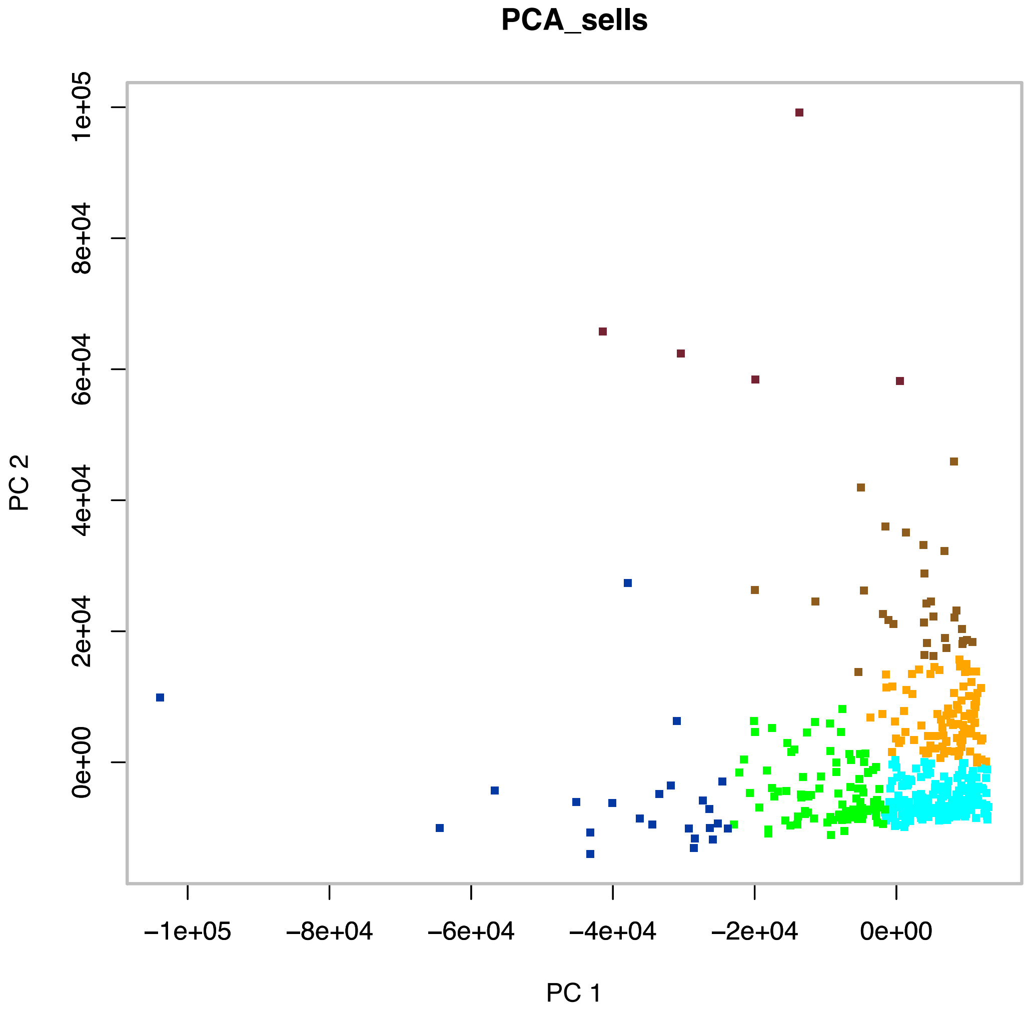
Do kmeans and choose k through the cost function

It’s good to pick a **k** around the knee:
I choose 6 for it matches the number of labels



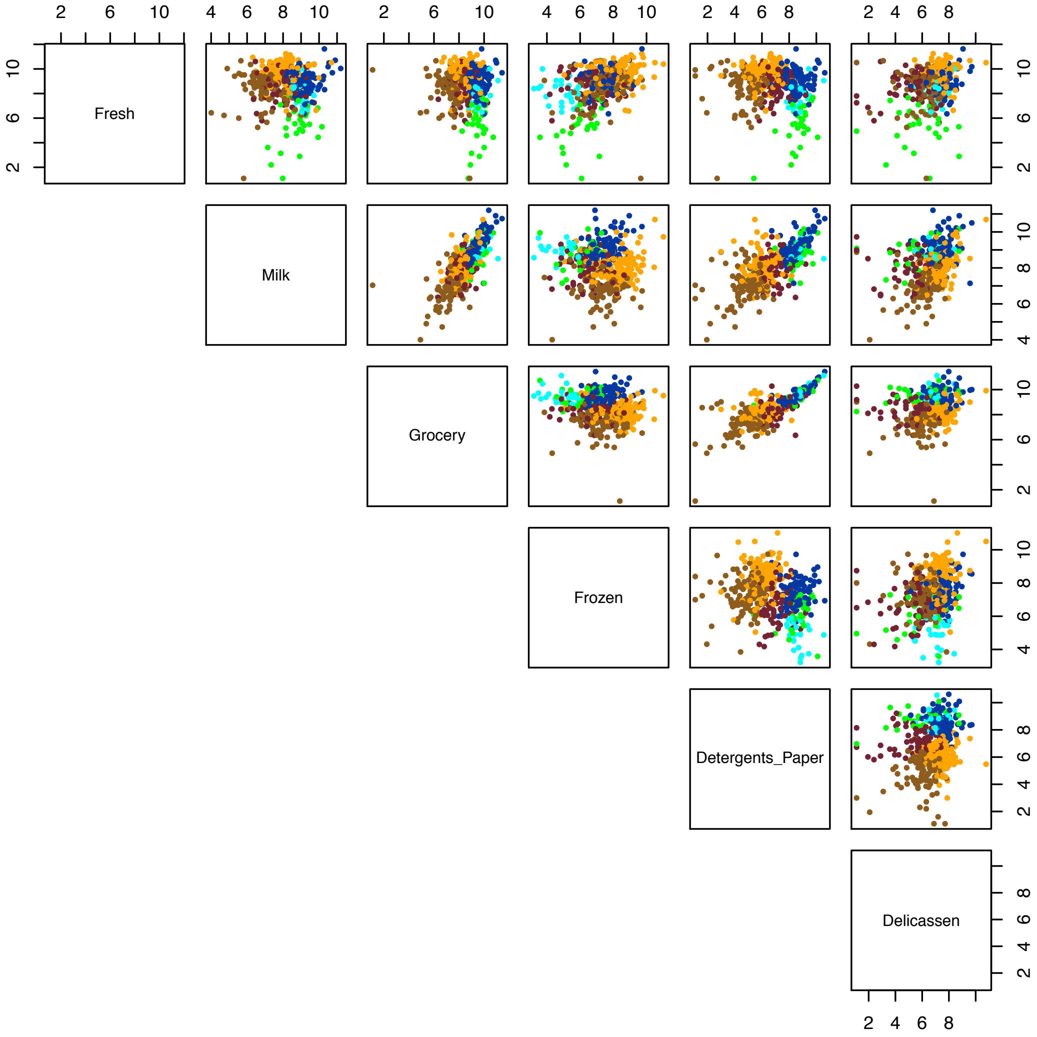
Visualization of the data (PCA)

- PCA does show some separation. **Colors are the clusters**
- Data points show large range of dynamics!



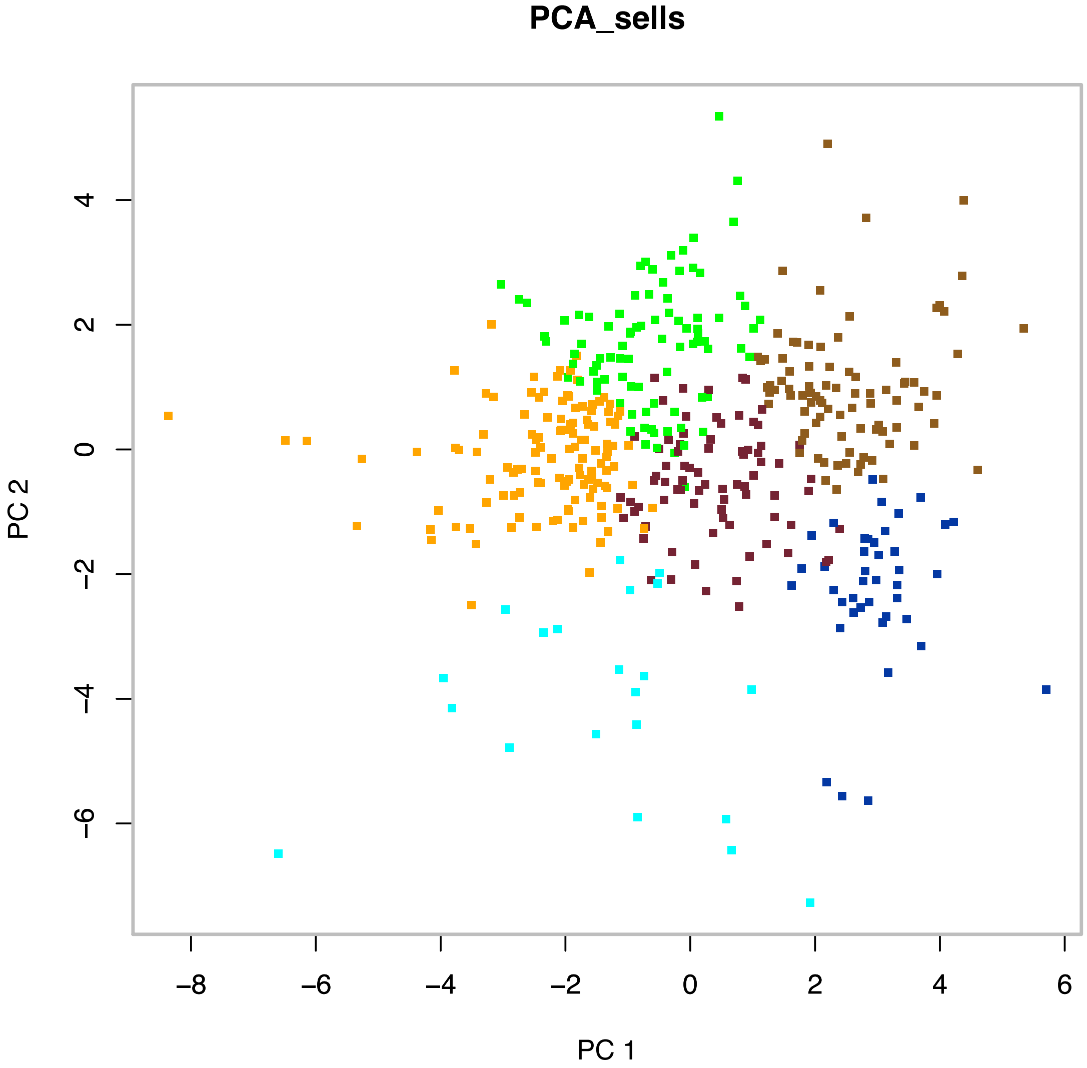
Do log transform of the data

- Log transform the data
- Do scatter plot matrix after the log transform
- Do the kmeans and color the clusters identified by k-means



PCA after log transformation: Clusters

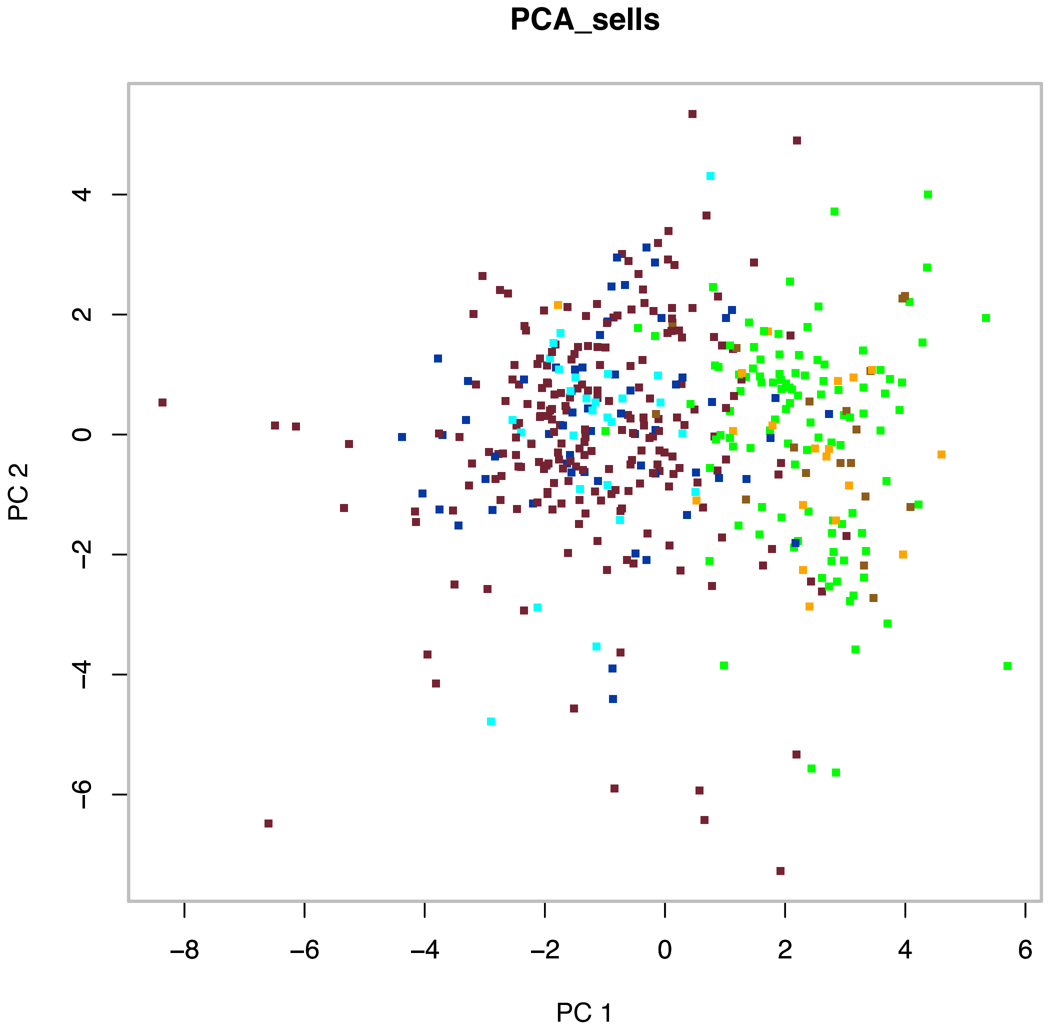
Colors show the **clusters** identified by k-means



PCA after log transformation

Colors show the **Channel-region labels**

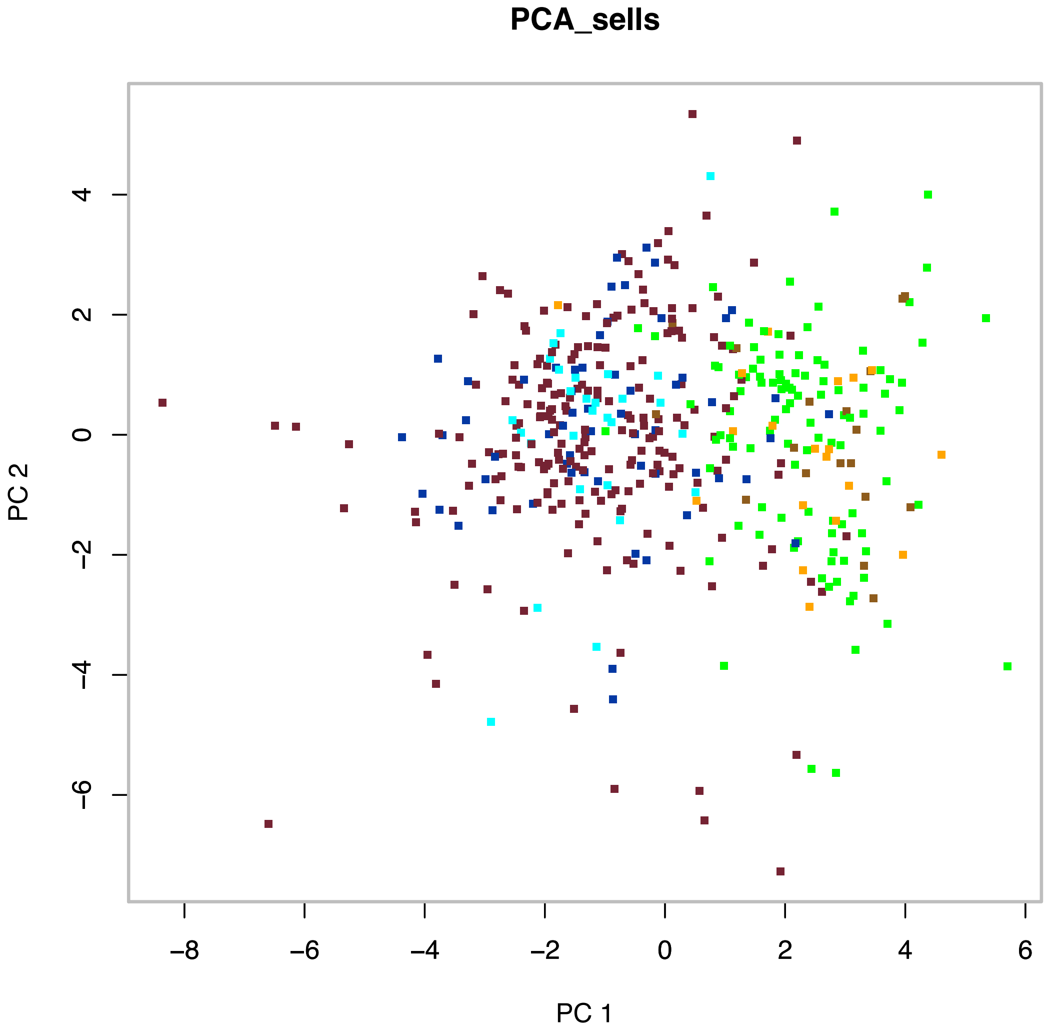
What does this tell us?



PCA after log transformation

Colors show the **Channel-region labels**

Channels differ a lot



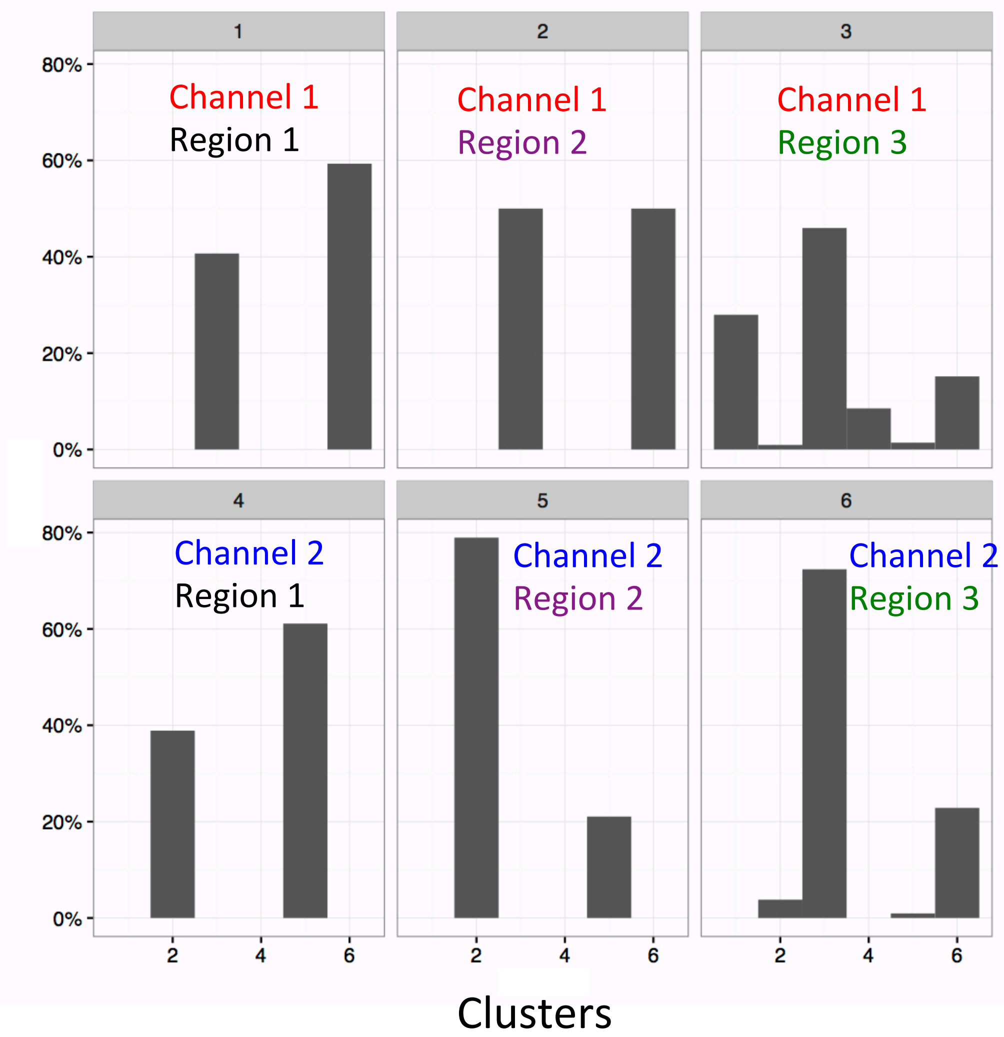
Cluster center histogram of the Portugal grocery spending data

- For each channel/ region, we make a histogram of customers that map to each of the **6 cluster centers**.

- **What do you see?**

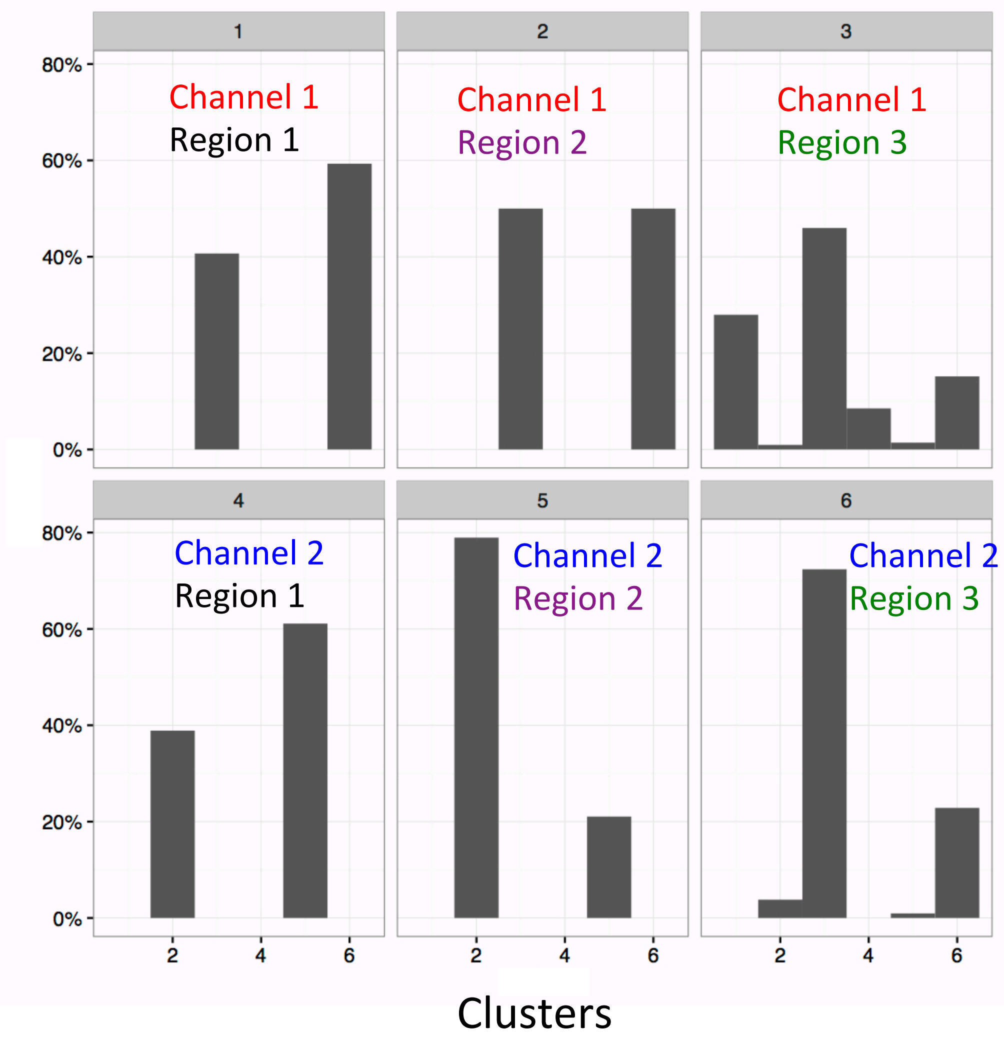
Channel1: Horeca
Channel2: Retail

Region1: Lisbon
Region2: Oporto
Region3: Other



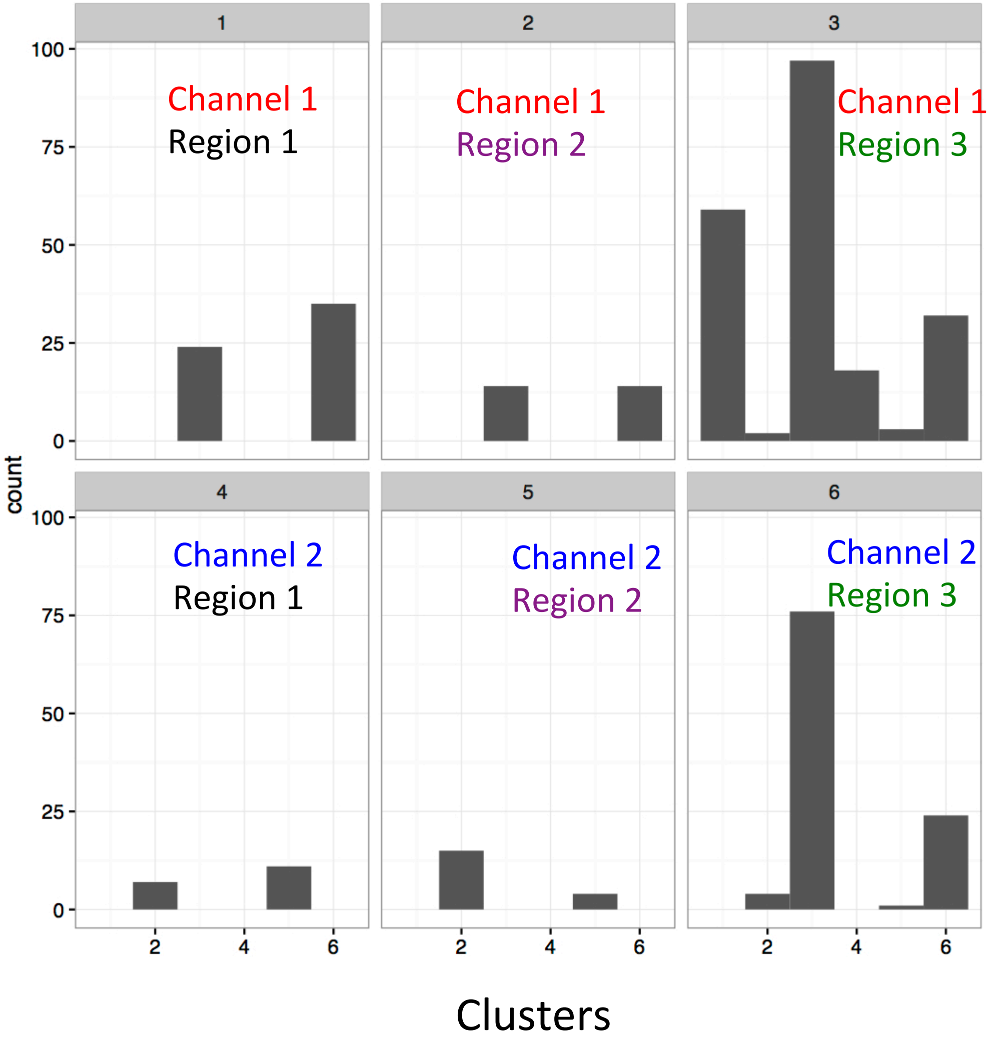
Cluster center histogram of the Portugal grocery spending data

- For each channel/ region, we make a histogram of customers that map to each of the 6 cluster centers.
- **Channels are significantly different!**
- **Region 3 is special**
- **Is it enough to plot the percentage?**



Cluster center histogram of the Portugal grocery spending data

- For each channel/ region, we make a histogram of customers that map to each of the 6 cluster centers.
- **Channels are significantly different!**
- **Region 3 is special**
- **Count matters depending on the purpose**



Q. What can we do with cluster center histograms?

A. investigate the feature patterns of data groups

B. Classify new data with the cluster center histograms.

C. Both A and B.

Markov Chain

- Motivation
- Definition of Markov model
- Graph representation – Markov chain
- Transition probability matrix
- The stationary Markov chain
- The pageRank algorithm

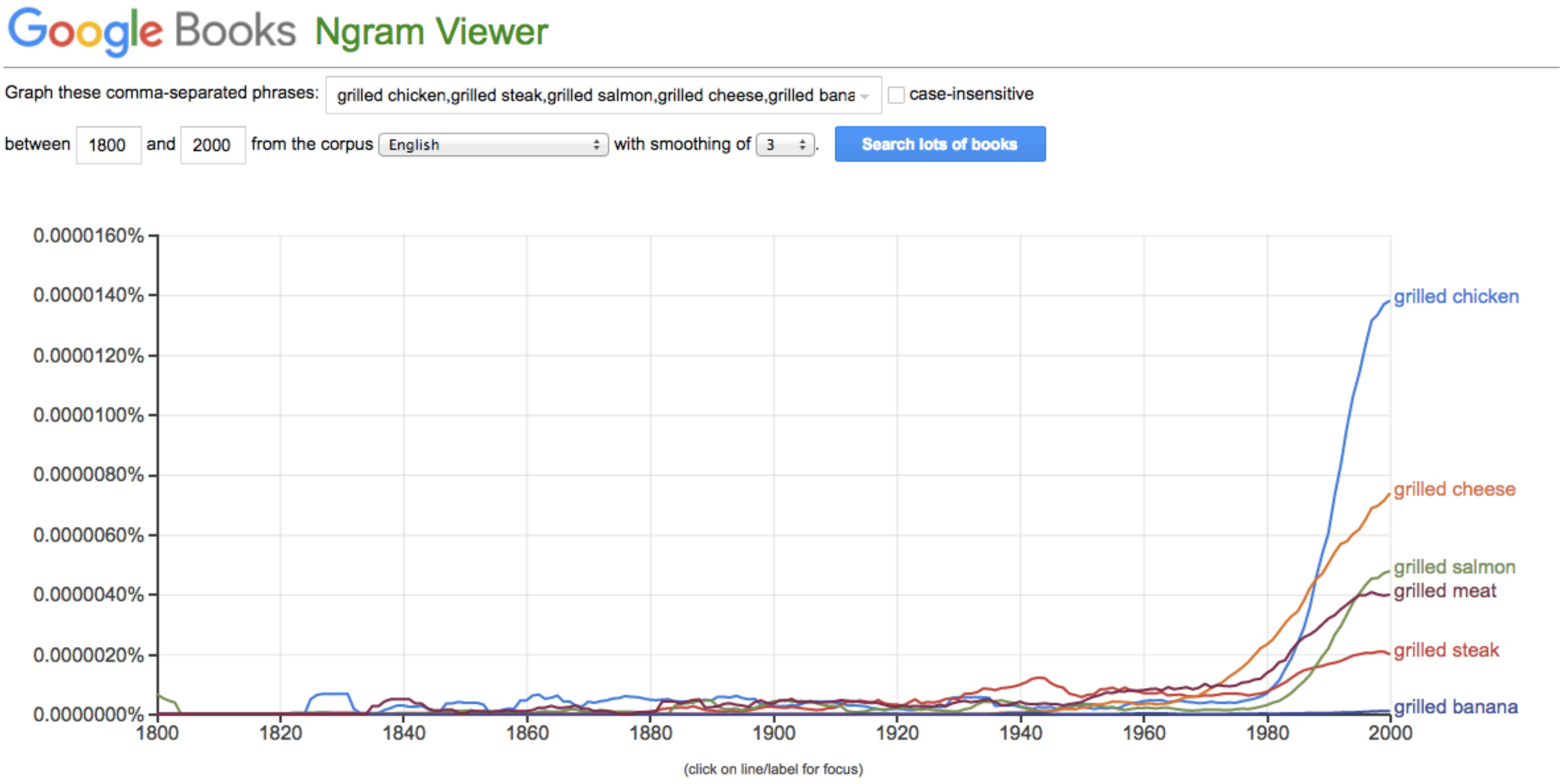
Motivation

- So far, the processes we learned such as **Bernoulli and Poisson** process are sequences of **independent** trials.
- There are a lot of real world situations where sequences of events are **Not independent** In comparison.
- Markov chain is one type of characterization of a series of **dependent** trials.

An example of dependent events in a sequence

I had a glass of wine with my grilled __________

An example of dependent events in a sequence



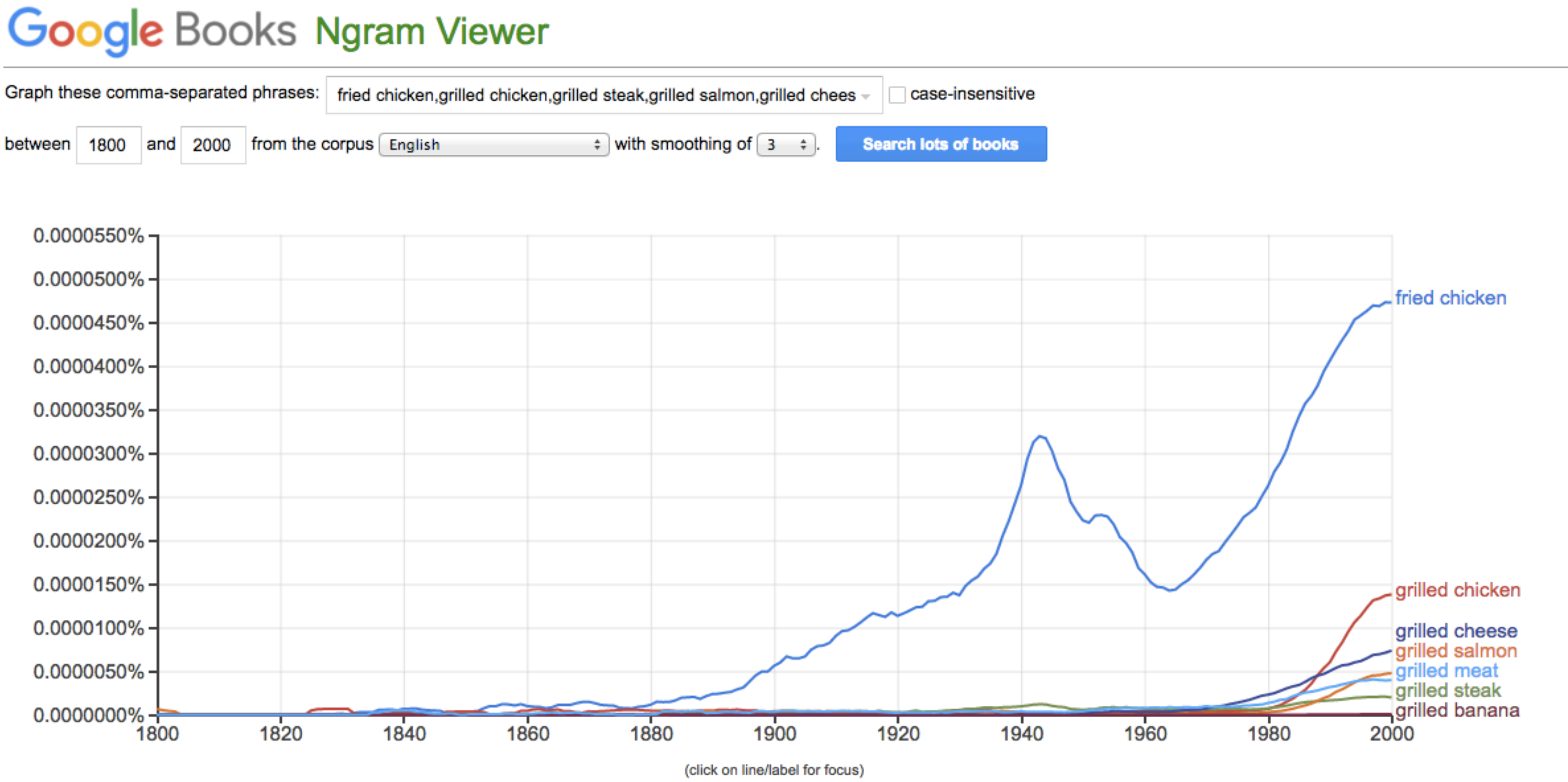
Google Books Ngram Viewer

Graph these comma-separated phrases: grilled chicken,grilled steak,grilled salmon,grilled cheese,grilled bana

□ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3 .

An example of dependent events in a sequence



Graph these comma-separated phrases: fried chicken,grilled chicken,grilled steak,grilled salmon,grilled chees

□ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3.

Markov chain

- Markov chain is a process in which outcome of any trial in a sequence is **conditioned by the outcome of the trial immediately preceding, but not by earlier ones.**
- Such dependence is called **chain dependence**

Andrey Markov (1856-1922)

Markov chain in terms of probability

- Let $X_0, X_1,$... be a sequence of discrete finite-valued random variables

- The sequence is a Markov chain if the probability distribution X_t only depends on the distribution of the immediately preceding random variable X_{t-1}

$$\boxed{P(X_t|X_0..., X_{t-1}) = P(X_t|X_{t-1})}$$

- If the conditional probabilities (transition probabilities) do **NOT change with time**, it's called **constant Markov chain.**

$$P(X_t|X_{t-1}) = P(X_{t-1}|X_{t-2}) = ... = P(X_1|X_0)$$

Coin example

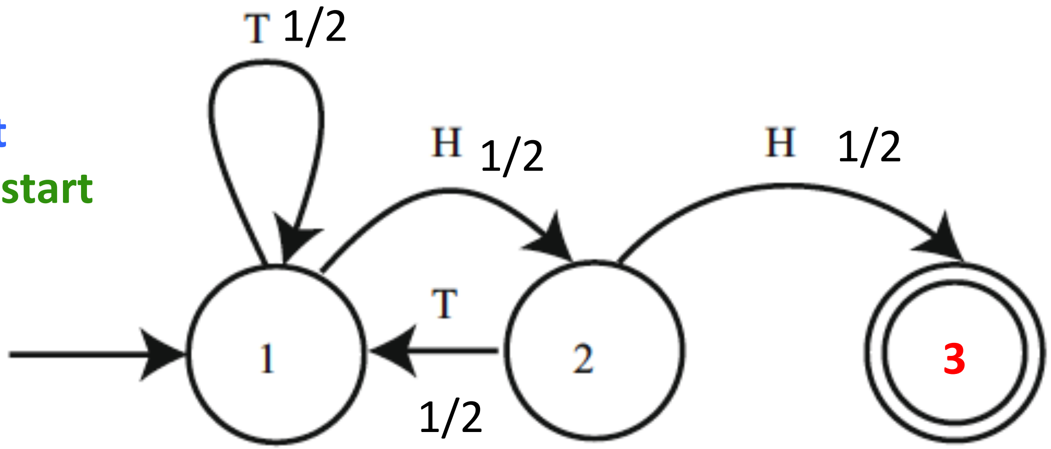
- Toss a fair coin until you see two heads in a row and then stop, what is the probability of stopping after exactly **n** flips?

- Use a state diagram, which is a **directed graph**. Circles are the states of likely outcomes. Arrow directions show the direction of transitions. Numbers over the arrows show transition probabilities.

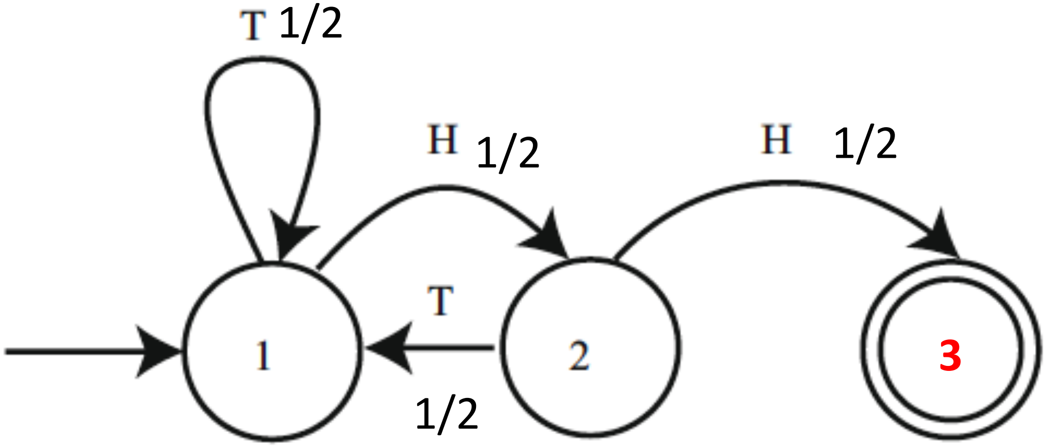
1 -> Start or just had tail/restart

2 -> had one head after start/restart

3 -> 2heads in a row/Stop



Is this a Markov chain? And why?



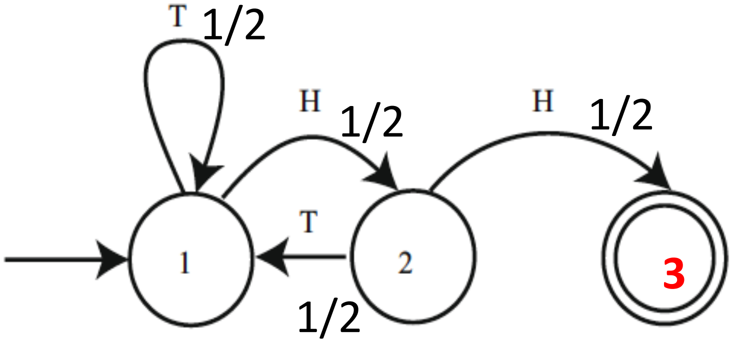
Is this a Markov chain? And why?

Yes. Because for each trial, the probability distribution of the outcomes is only conditioned on the previous trial.

The model helps form recurrence formula

- Let p_n be the probability of stopping after **n** flips

$$p_1 = 0 \quad p_2 = 1/4 \quad p_3 = 1/8 \quad p_4 = 1/8 \quad \ldots$$



The model helps form recurrence formula

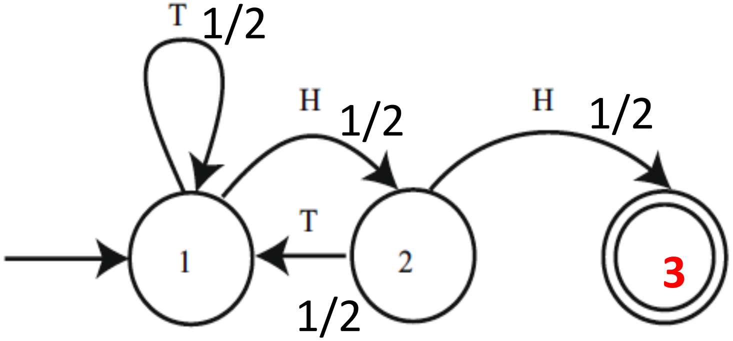
- Let p_n be the probability of stopping after **n** flips

$$p_1 = 0 \quad p_2 = 1/4 \quad p_3 = 1/8 \quad p_4 = 1/8 \quad \ldots$$

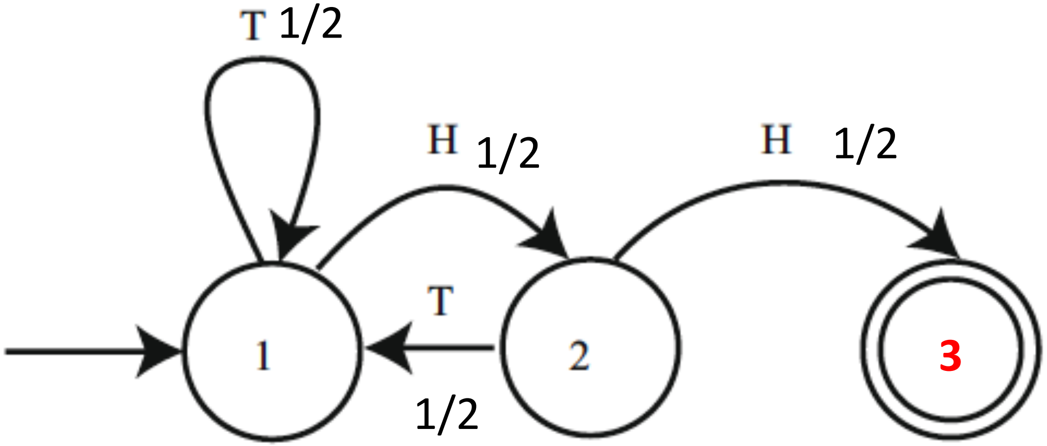
- If $n > 2$, there are two ways the sequence starts
  - Toss T and finish in n-1 tosses
  - Or toss HT and finish in n-2 tosses

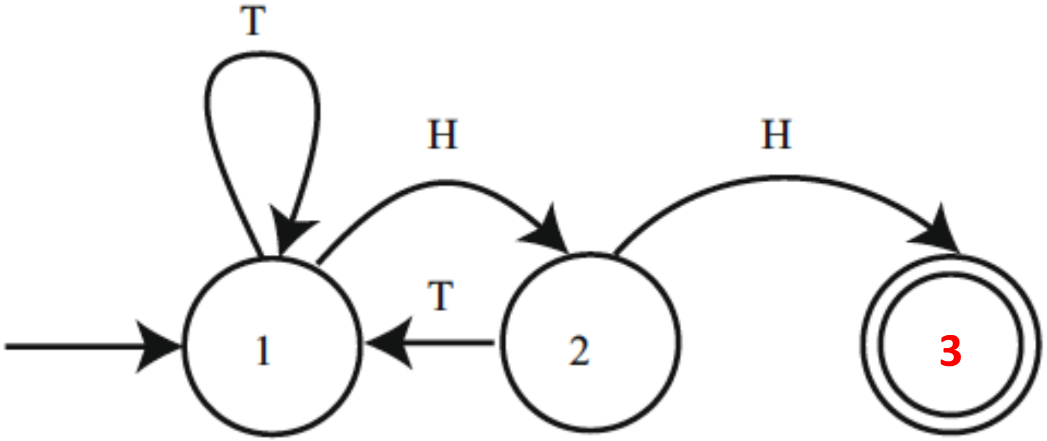
- So we can derive a recurrence relation

$$p_n = \underset{\substack{\uparrow \\ P(T)}}{\frac{1}{2}} p_{n-1} + \underset{\substack{\uparrow \\ P(HT)}}{\frac{1}{4}} p_{n-2}$$



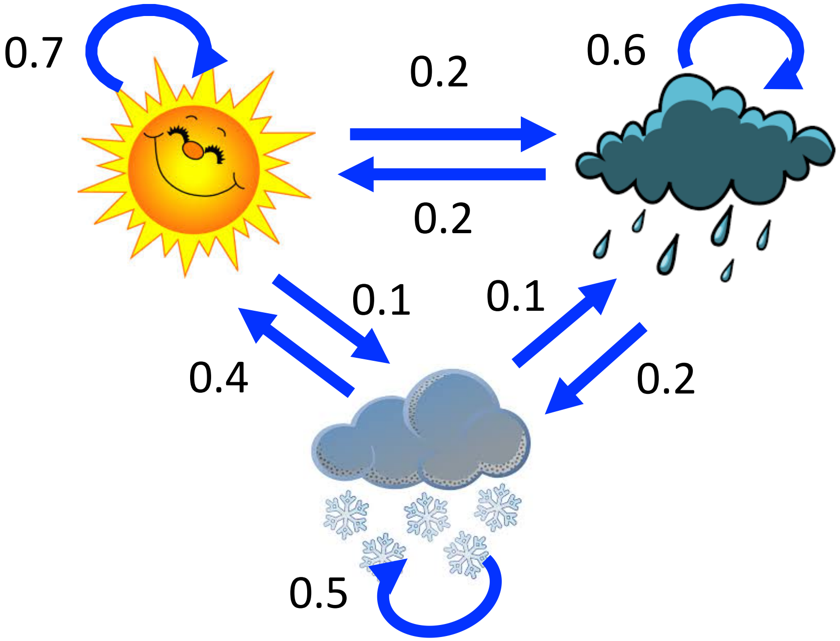
Transition probability btw states





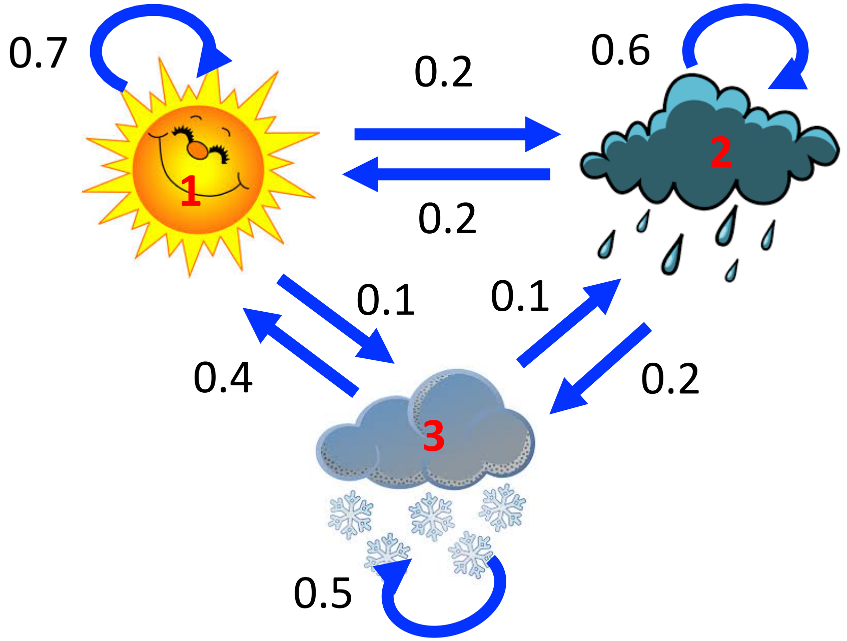
Transition probability matrix: weather model

- Let’s model daily weather as one of the three states (Sunny, Rainy, and Snowy) with Markov chain that has the transition probabilities as shown here.



Transition probability matrix: weather model

- Let’s model daily weather as one of the three states (Sunny, Rainy, and Snowy) with Markov chain that has the transition probabilities as shown here.



i, the current state at time point t

j, the next state at time point t+1

$$\boldsymbol{P} = \begin{array}{cc} \begin{array}{ccc}\text{Sunny} & \text{Rainy} & \text{Snowy}\end{array} & \\ \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} & \begin{array}{c}\text{Sunny}\\ \text{Rainy}\\ \text{Snowy}\end{array} \end{array}$$

The transition probability matrix

Q: The transition probabilities for a node sum to 1

A. Yes.

B. No.

Only the row sum is 1, that is: the probabilities associated with outgoing arrows sum to 1.

Additional References

- Robert V. Hogg, Elliot A. Tanis and Dale L. Zimmerman. “Probability and Statistical Inference”
- Kelvin Murphy, “Machine learning, A Probabilistic perspective”

See you next time

See You!