Segmentation, Mapping, Text

UCSD MGT 100 Week 02

Kenneth C. Wilbur and Dan Yavorsky

Segmentation

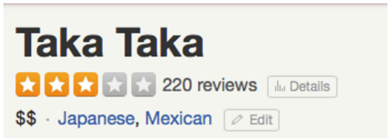
- Putting the S into STP
- Note: I don’t duplicate the reading

Heterogeneity

- A fancy way to say that customers differ, e.g.
- Product needs–usage intensity, frequency, context; loyalty

```
- Most important dimension of heterogeneity, by far
- This perspective differs from the reading
```

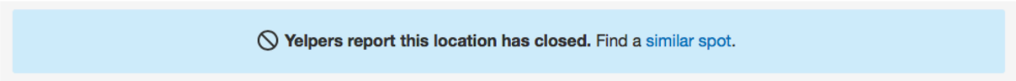
- Demographics–often overrated as predictors of behavior
- Psychographics–Orientation to Art, Status, Religion, Family, …
- Location
- Experience
- Information
- Attitudes
- Differences may predict purchases, wtp, usage, satisfaction, retention, …



Taka Taka

★★★☆☆ 220 reviews Details

$$ · Japanese, Mexican Edit



⃠ **Yelpers report this location has closed.** Find a similar spot.

Market Segmentation

- *Segments*: distinct customer groups with similar attributes *within* a segment, different attributes *between* segments

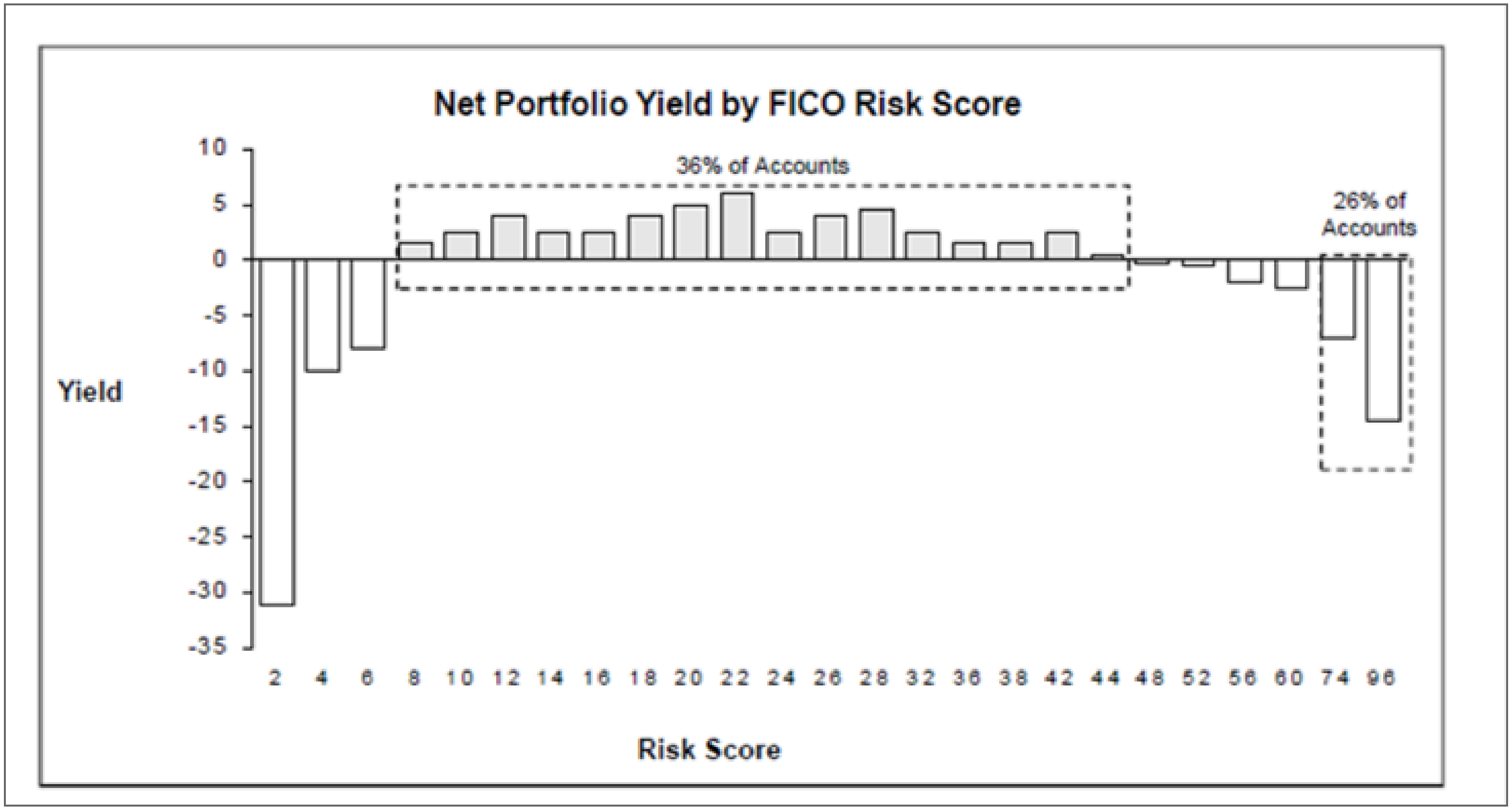
```
- Fundamental since the 1960s
- Numerous segmentation techniques exist
- Customer Response Profiles embody segments
- B2B segments: customer needs, size, profitability, internal structure
```

- Segmentation should drive most customer decisions

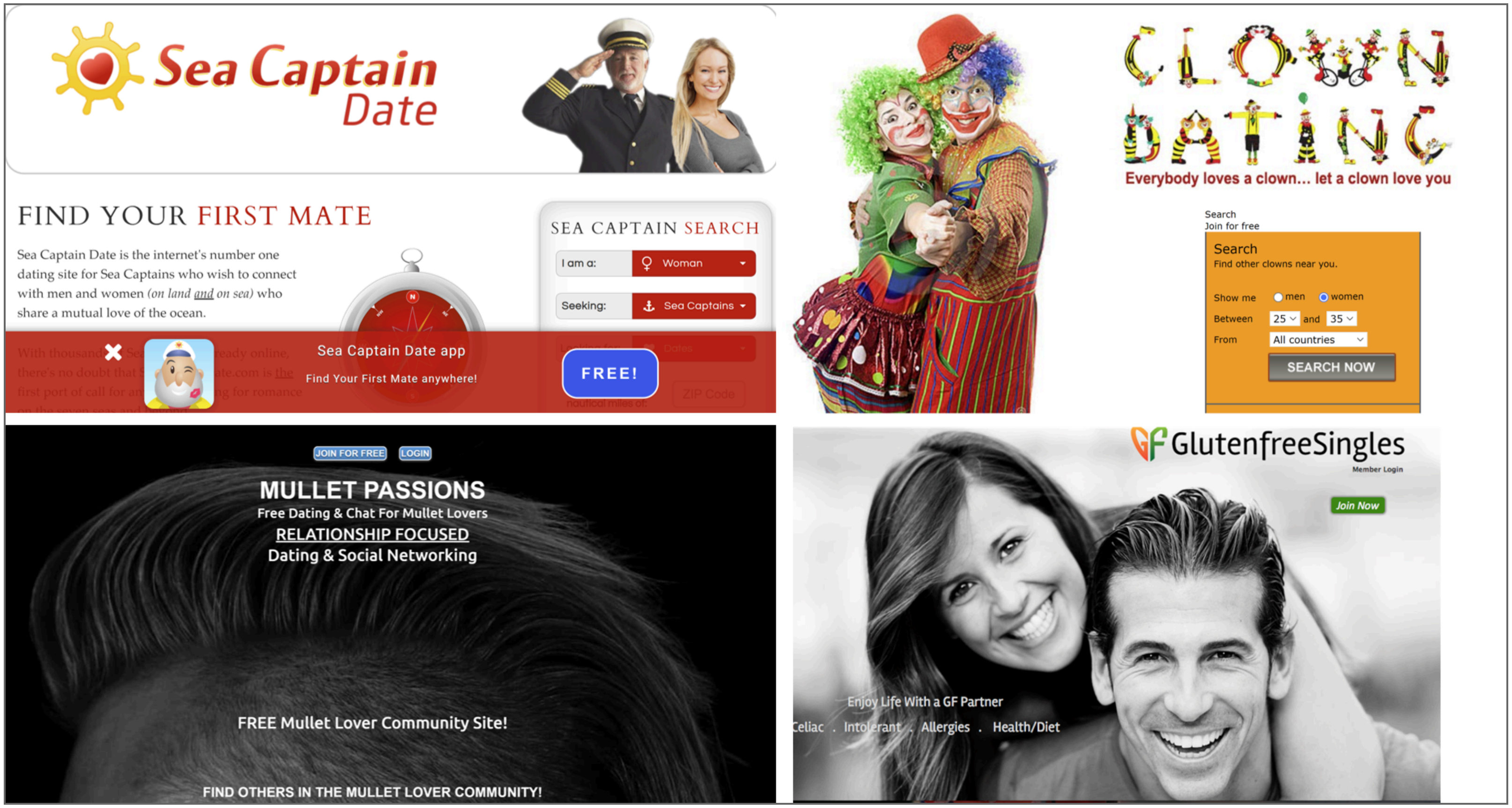
Segments in this class, ranked by size

1. BE majors uninterested in Customers/Marketing
2. BE majors interested in Customers/Marketing
3. Non-BE majors, often minoring in Business, Marketing or Business Analytics

Measurable


Net Portfolio Yield by FICO Risk Score
36% of Accounts
26% of Accounts
Yield
10
5
0
-5
-10
-15
-20
-25
-30
-35
2
4
6
8
10
12
14
16
18
20
22
24
26
28
32
36
38
42
44
48
52
56
60
74
96
Risk Score

Substantial



Sea Captain Date

FIND YOUR FIRST MATE

Sea Captain Date is the internet's number one dating site for Sea Captains who wish to connect with men and women *(on land <u>and</u> on sea)* who share a mutual love of the ocean.

With thousand Sea ready online, there's no doubt that S ate.com is the first port of call for an ng for romance on the seven seas and beyond

SEA CAPTAIN SEARCH

I am a: Woman

Seeking: Sea Captains

Looking for: Dates

nautical miles of: ZIP Code

Sea Captain Date app

Find Your First Mate anywhere!

FREE!

CLOWN DATING

Everybody loves a clown... let a clown love you

Search
Join for free

Search

Find other clowns near you.

Show me ○ men ● women

Between 25 and 35

From All countries

SEARCH NOW

JOIN FOR FREE LOGIN

MULLET PASSIONS

Free Dating & Chat For Mullet Lovers

<u>RELATIONSHIP FOCUSED</u>

Dating & Social Networking

FREE Mullet Lover Community Site!

FIND OTHERS IN THE MULLET LOVER COMMUNITY!

GlutenfreeSingles

Member Login

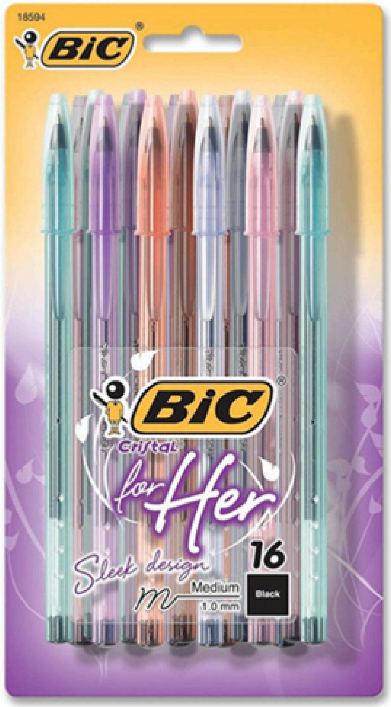
Join Now

Enjoy Life With a GF Partner

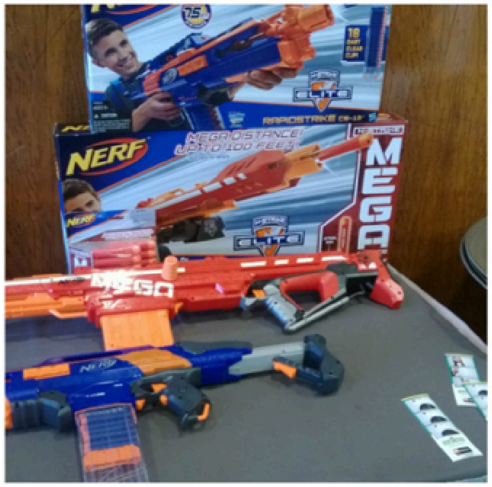
Celiac . Intolerant . Allergies . Health/Diet

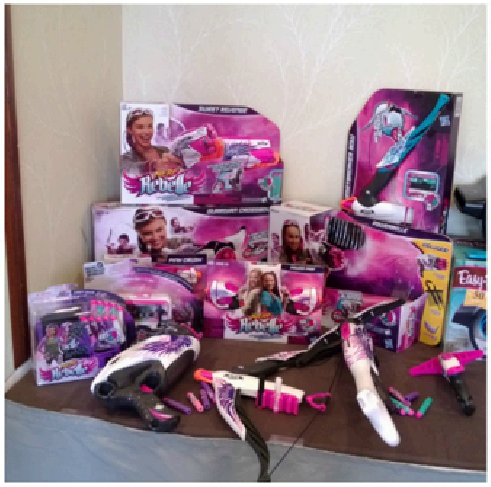
Is segmenting by gender sexist?

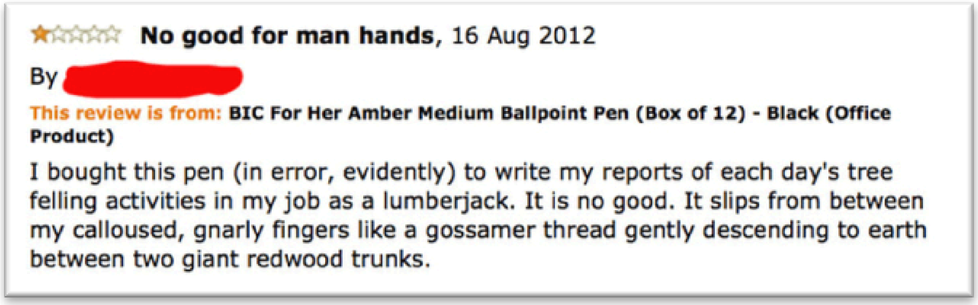
Is it sexist to market different products to different genders?



Bic/Amazon





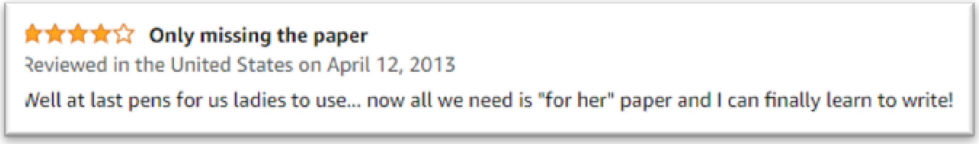


★☆☆☆☆ **No good for man hands**, 16 Aug 2012

By [redacted]

This review is from: **BIC For Her Amber Medium Ballpoint Pen (Box of 12) - Black (Office Product)**

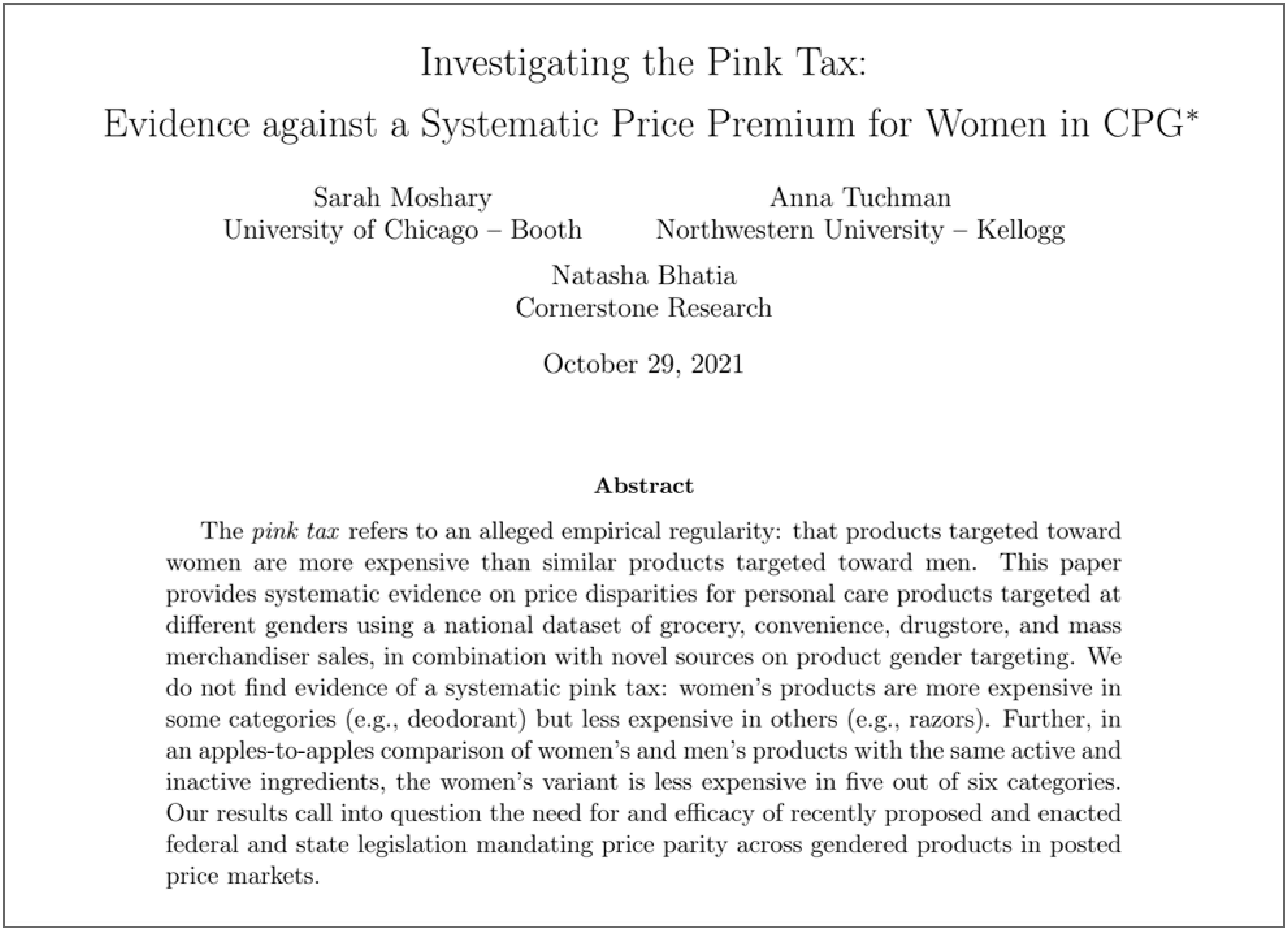
I bought this pen (in error, evidently) to write my reports of each day's tree felling activities in my job as a lumberjack. It is no good. It slips from between my calloused, gnarly fingers like a gossamer thread gently descending to earth between two giant redwood trunks.



★★★★☆ **Only missing the paper**

Reviewed in the United States on April 12, 2013

Well at last pens for us ladies to use... now all we need is "for her" paper and I can finally learn to write!



Investigating the Pink Tax: Evidence against a Systematic Price Premium for Women in CPG*

Sarah Moshary
University of Chicago – Booth

Anna Tuchman
Northwestern University – Kellogg

Natasha Bhatia
Cornerstone Research

October 29, 2021

Abstract

The *pink tax* refers to an alleged empirical regularity: that products targeted toward women are more expensive than similar products targeted toward men. This paper provides systematic evidence on price disparities for personal care products targeted at different genders using a national dataset of grocery, convenience, drugstore, and mass merchandiser sales, in combination with novel sources on product gender targeting. We do not find evidence of a systematic pink tax: women's products are more expensive in some categories (e.g., deodorant) but less expensive in others (e.g., razors). Further, in an apples-to-apples comparison of women's and men's products with the same active and inactive ingredients, the women's variant is less expensive in five out of six categories. Our results call into question the need for and efficacy of recently proposed and enacted federal and state legislation mandating price parity across gendered products in posted price markets.

Customer demographics

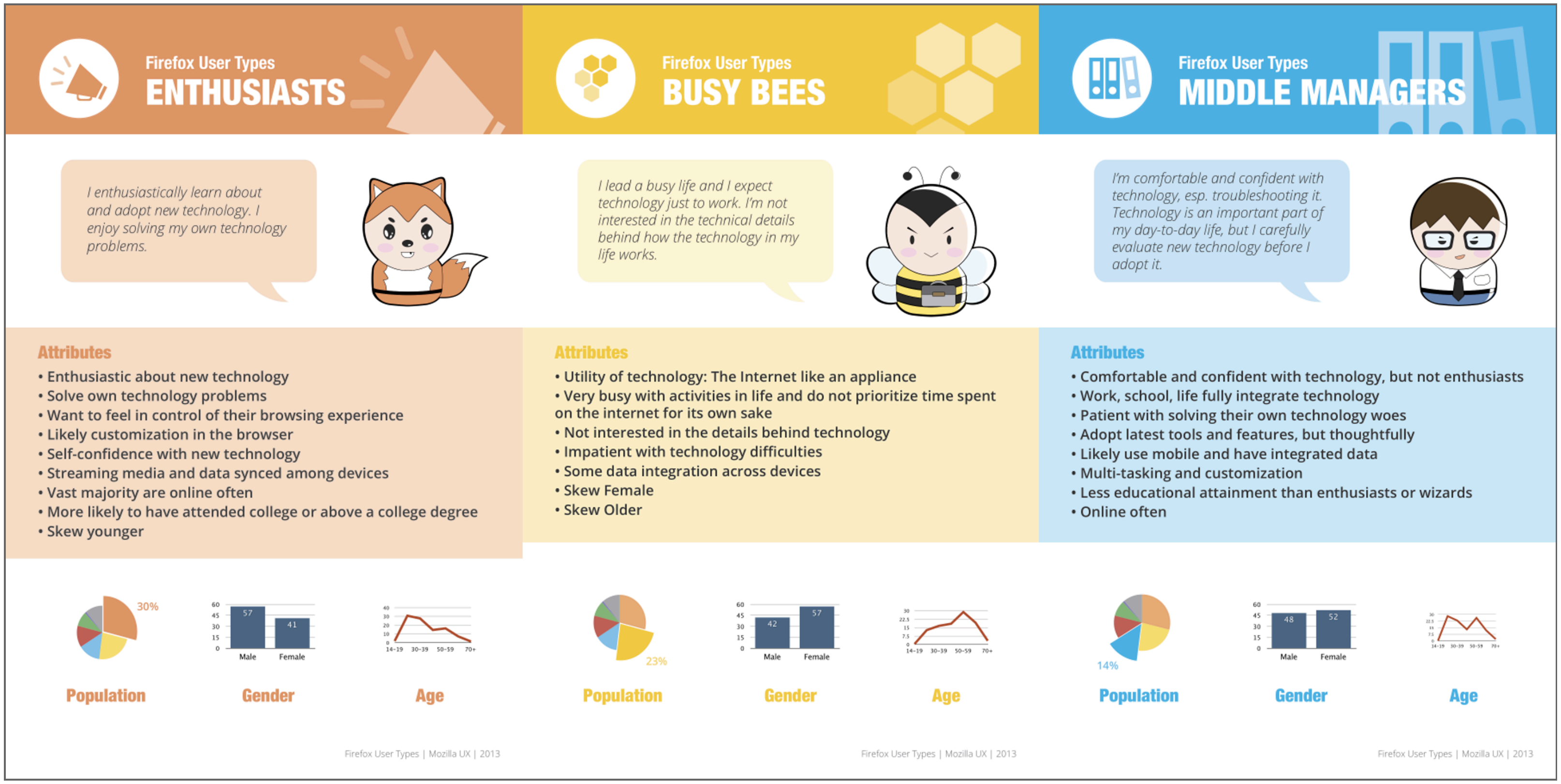
- In some markets– makeup, diapers, sports, shoes– demographics correlate strongly with behaviors
- In most markets– smartphones, universities, software, cars– demographics correlate weakly with behaviors
- Demographics don’t typically cause purchases, except when they predict real differences in customer needs
- Why do we so often overrate demographics as predictors of behavior?
- **Gender-Based Pricing in Consumer Packaged Goods: A Pink Tax?**

Segmentation in Action

- Who does it
- Browser example
- Why we keep it quiet

- *Nearly every* large business segments its markets

Firefox User Types



Firefox User Types
ENTHUSIASTS

I enthusiastically learn about and adopt new technology. I enjoy solving my own technology problems.

Attributes
- Enthusiastic about new technology
- Solve own technology problems
- Want to feel in control of their browsing experience
- Likely customization in the browser
- Self-confidence with new technology
- Streaming media and data synced among devices
- Vast majority are online often
- More likely to have attended college or above a college degree
- Skew younger

Population Gender Age

Firefox User Types | Mozilla UX | 2013

Firefox User Types
BUSY BEES

I lead a busy life and I expect technology just to work. I'm not interested in the technical details behind how the technology in my life works.

Attributes
- Utility of technology: The Internet like an appliance
- Very busy with activities in life and do not prioritize time spent on the internet for its own sake
- Not interested in the details behind technology
- Impatient with technology difficulties
- Some data integration across devices
- Skew Female
- Skew Older

Population Gender Age

Firefox User Types | Mozilla UX | 2013

Firefox User Types
MIDDLE MANAGERS

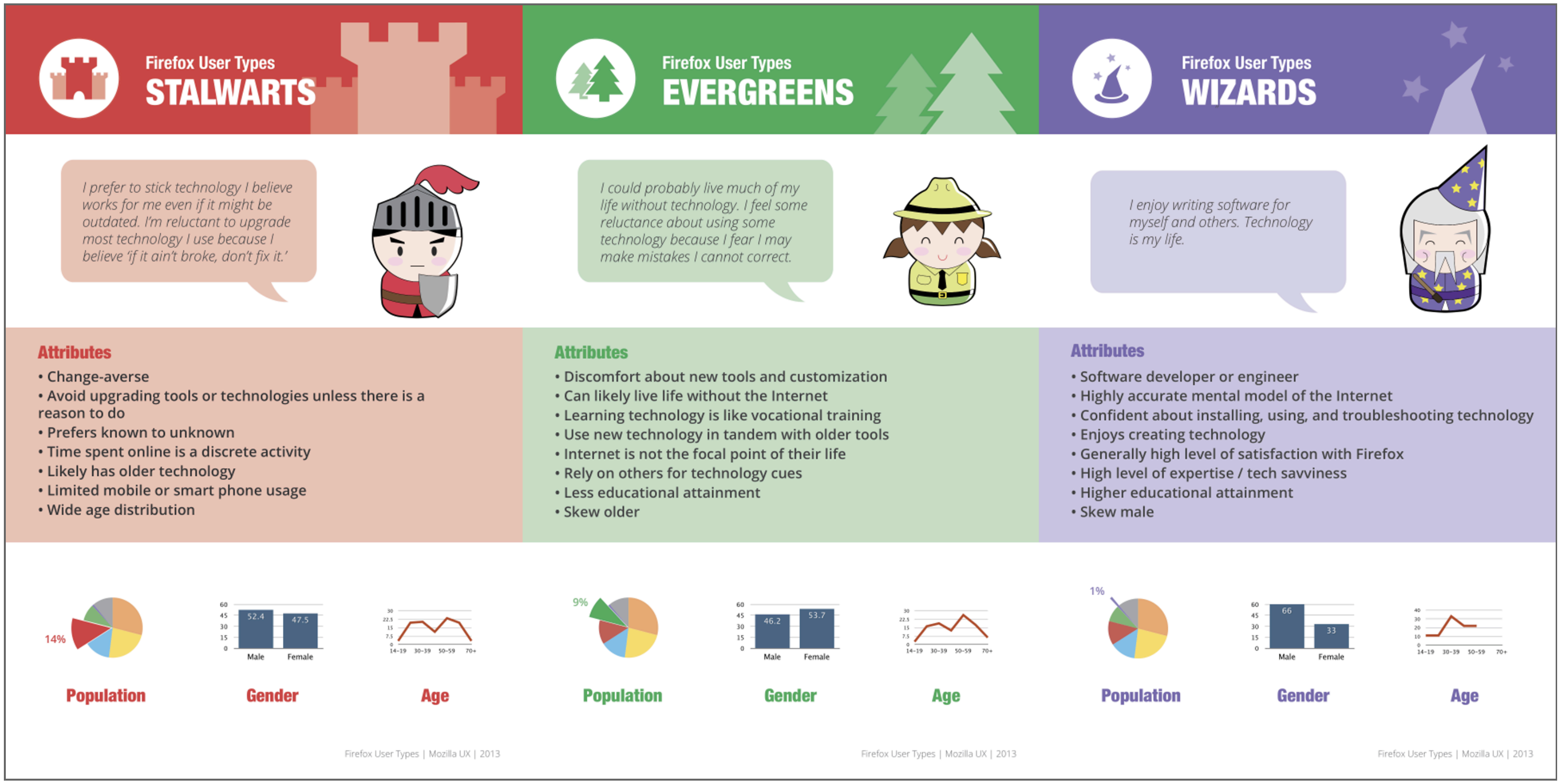
I'm comfortable and confident with technology, esp. troubleshooting it. Technology is an important part of my day-to-day life, but I carefully evaluate new technology before I adopt it.

Attributes
- Comfortable and confident with technology, but not enthusiasts
- Work, school, life fully integrate technology
- Patient with solving their own technology woes
- Adopt latest tools and features, but thoughtfully
- Likely use mobile and have integrated data
- Multi-tasking and customization
- Less educational attainment than enthusiasts or wizards
- Online often

Population Gender Age

Firefox User Types | Mozilla UX | 2013

Firefox User Types



Firefox User Types
STALWARTS

I prefer to stick technology I believe works for me even if it might be outdated. I'm reluctant to upgrade most technology I use because I believe 'if it ain't broke, don't fix it.'

Attributes

- Change-averse
- Avoid upgrading tools or technologies unless there is a reason to do
- Prefers known to unknown
- Time spent online is a discrete activity
- Likely has older technology
- Limited mobile or smart phone usage
- Wide age distribution

Firefox User Types | Mozilla UX | 2013

Firefox User Types
EVERGREENS

I could probably live much of my life without technology. I feel some reluctance about using some technology because I fear I may make mistakes I cannot correct.

Attributes

- Discomfort about new tools and customization
- Can likely live life without the Internet
- Learning technology is like vocational training
- Use new technology in tandem with older tools
- Internet is not the focal point of their life
- Rely on others for technology cues
- Less educational attainment
- Skew older

Firefox User Types | Mozilla UX | 2013

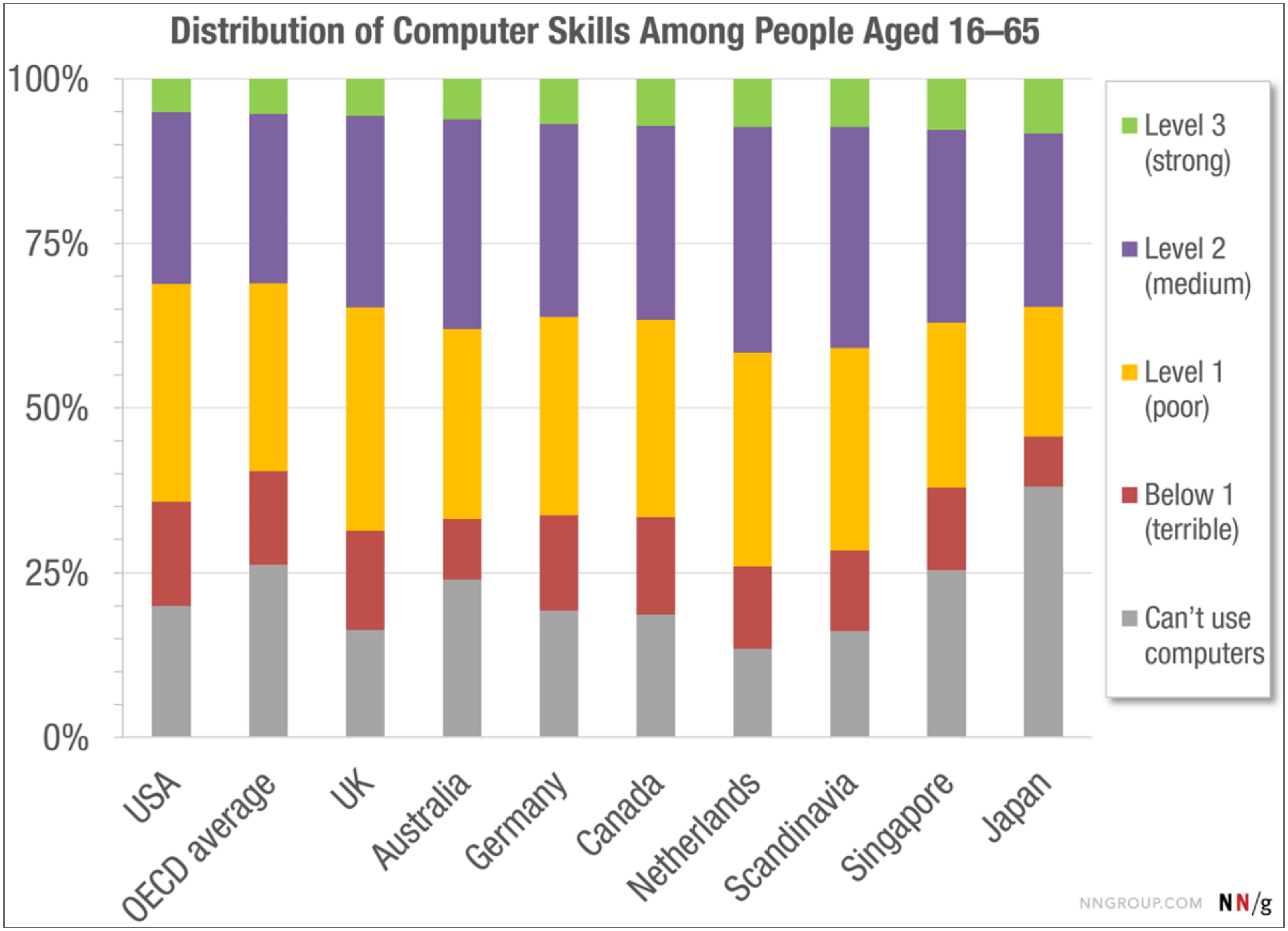
Firefox User Types
WIZARDS

I enjoy writing software for myself and others. Technology is my life.

Attributes

- Software developer or engineer
- Highly accurate mental model of the Internet
- Confident about installing, using, and troubleshooting technology
- Enjoys creating technology
- Generally high level of satisfaction with Firefox
- High level of expertise / tech savviness
- Higher educational attainment
- Skew male

Firefox User Types | Mozilla UX | 2013


Distribution of Computer Skills Among People Aged 16–65
100%
75%
50%
25%
0%
USA
OECD average
UK
Australia
Germany
Canada
Netherlands
Scandinavia
Singapore
Japan
🟩 Level 3 (strong)
🟪 Level 2 (medium)
🟨 Level 1 (poor)
🟥 Below 1 (terrible)
⬜ Can't use computers
NNGROUP.COM NN/g

Firms don’t publicize segments

- UO website: “We stock our stores with what we love, calling on our — and our customer’s — interest in contemporary art, music and fashion. …
- “We offer a lifestyle-specific shopping experience for the educated, urban-minded individual in the 18 to 30 year-old range…”

URBAN OUTFITTERS

Firms don’t publicize segments

- Earnings call: “Our customer is from traditional homes and advantage, but this offers them the benefit of rebellion…
- Our customer is exposed to new ideas and philosophies. This can be a real involvement and work, or it could be just talk.
- Irreverence and concern can live together. Often products sell well that represent the concerns they have but also can speak to their irreverence.
- Our customer leads a pretty cloistered existence although they deem themselves worldly…they believe that they’re right and they believe that everything that’s happening to them is what’s happening everywhere.
- Our customer is highly involved in mating and dating behavior…one of the primary drives for their spending behavior…they work hard to postpone adulthood… ’’

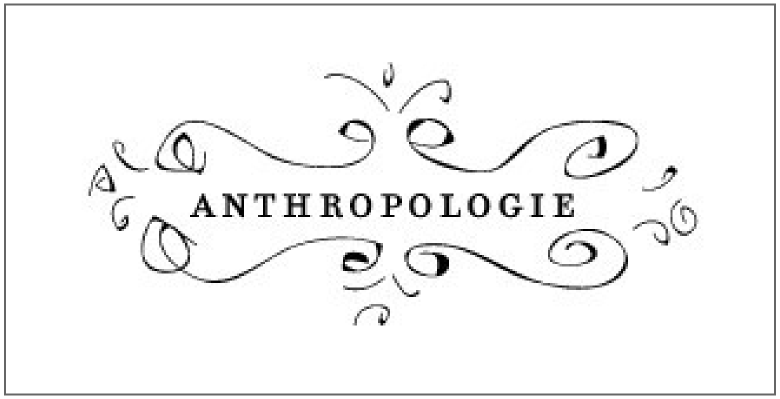
URBAN OUTFITTERS

Firms don’t publicize segments

- A. website: “a lifestyle brand that catered to creative, educated and affluent 30-45 year-old women…
- “Our customer is a creative-minded woman, who wants to look like herself, not the masses. She has a sense of adventure about what she wears, and although fashion is important to her, she is too busy enjoying life to be governed by the latest trends.’’

Firms don’t publicize segments

- Earnings call: “We don’t think of her in terms of age or affluence or even location. We try to think of her in her life stage and her sensibilities.
- “She’s recently wed. She’s settling down. She’s very interested less in the mating rituals and actually has been trying and building and creating an environment she wants to live in for herself and family.
- “She loves art and culture… And clothing and her living environment to her are canvases in which she’s able to express and control her life, whereas workplace and those things around her, she may not control.
- “We believe in many ways that’s what’s touched her and connect her to Anthropologie and why she is more loyal to us than to most retailers.’’



-source

How we segment

- Customer Data Platform (CDP)
- Data Marketplaces

Customer Data Platform (CDP) - 4 jobs

1. Data collection

```
Intake data from numerous disparate sources:
In-house, direct partners, data brokers, public data
```

2. Data unification or harmonization

```
Authenticate and de-duplicate rows and columns
```

3. Data comprehension

```
Generate inferences, test hypotheses, make predictions, estimate models
Covers descriptive, diagnostic & predictive analytics
```

4. Data activation

```
Prescriptive analytics: Use data to inform and automate marketing actions
```

Data Marketplaces

- Automated platforms for transacting & transmitting data e.g. **AWS Data Exchange**, **Databricks Marketplace**, **Snowflake Marketplace**

```
Relatively recent phenomenon
```

- Upsides

```
Many data types and sources
Easy subscriptions, automatic updates, automated wrangling
Competitive marketplace may lower prices
Enable complementarities with cloud storage, etc
```

- Some caveats

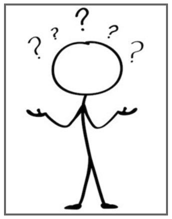
```
Low barriers to entry
Questionable validation, trustworthiness, purchaser control
May lead to a lemons/peaches market, but reviews may help
```

- Suppose we segment the smartphone market according to each customer’s desired brand.
- Is this a good approach?

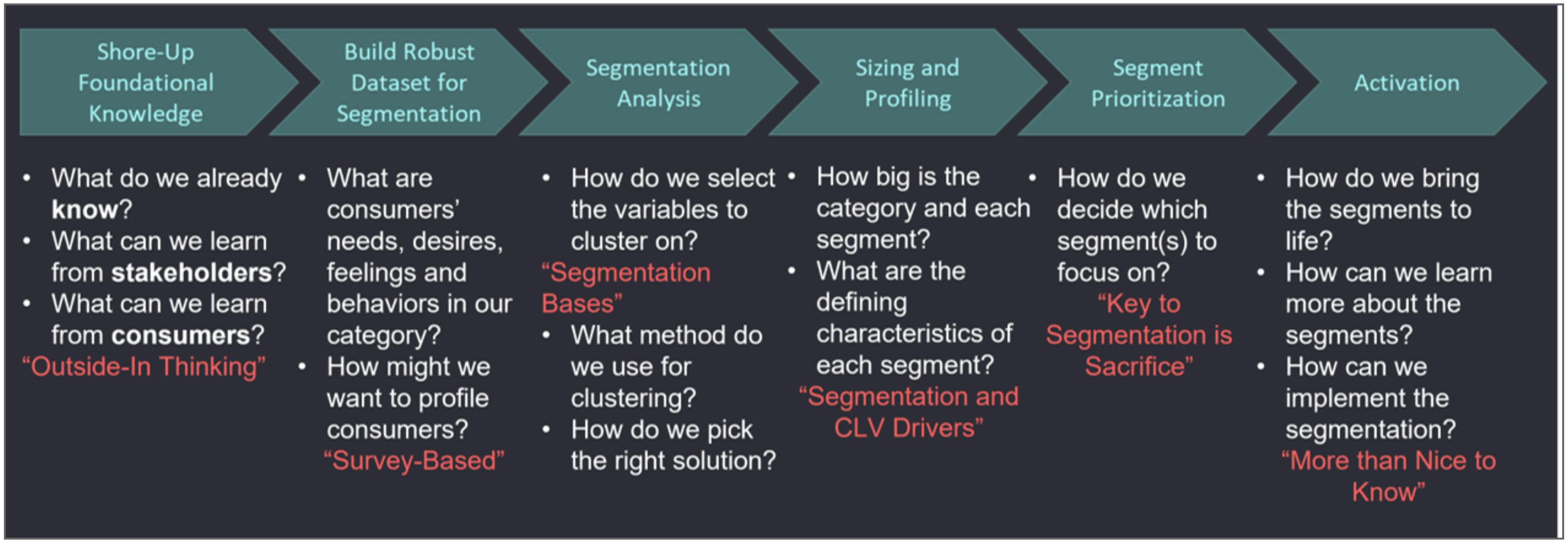
How to pick attributes?

- We want to segment based on attributes that drive sales, profit, retention. But how?

1. Theory, experience
2. Market research
3. Customer database
4. Consult customer experts (salespeople)
5. Find out what other firms are doing
6. Let sales data pick for us (het. logit)



How GBK segments



Shore-Up Foundational Knowledge

- What do we already **know**?
- What can we learn from **stakeholders**?
- What can we learn from **consumers**?

"Outside-In Thinking"

Build Robust Dataset for Segmentation

- What are consumers' needs, desires, feelings and behaviors in our category?
- How might we want to profile consumers?

"Survey-Based"

Segmentation Analysis

- How do we select the variables to cluster on?

"Segmentation Bases"

- What method do we use for clustering?
- How do we pick the right solution?

Sizing and Profiling

- How big is the category and each segment?
- What are the defining characteristics of each segment?

"Segmentation and CLV Drivers"

Segment Prioritization

- How do we decide which segment(s) to focus on?

"Key to Segmentation is Sacrifice"

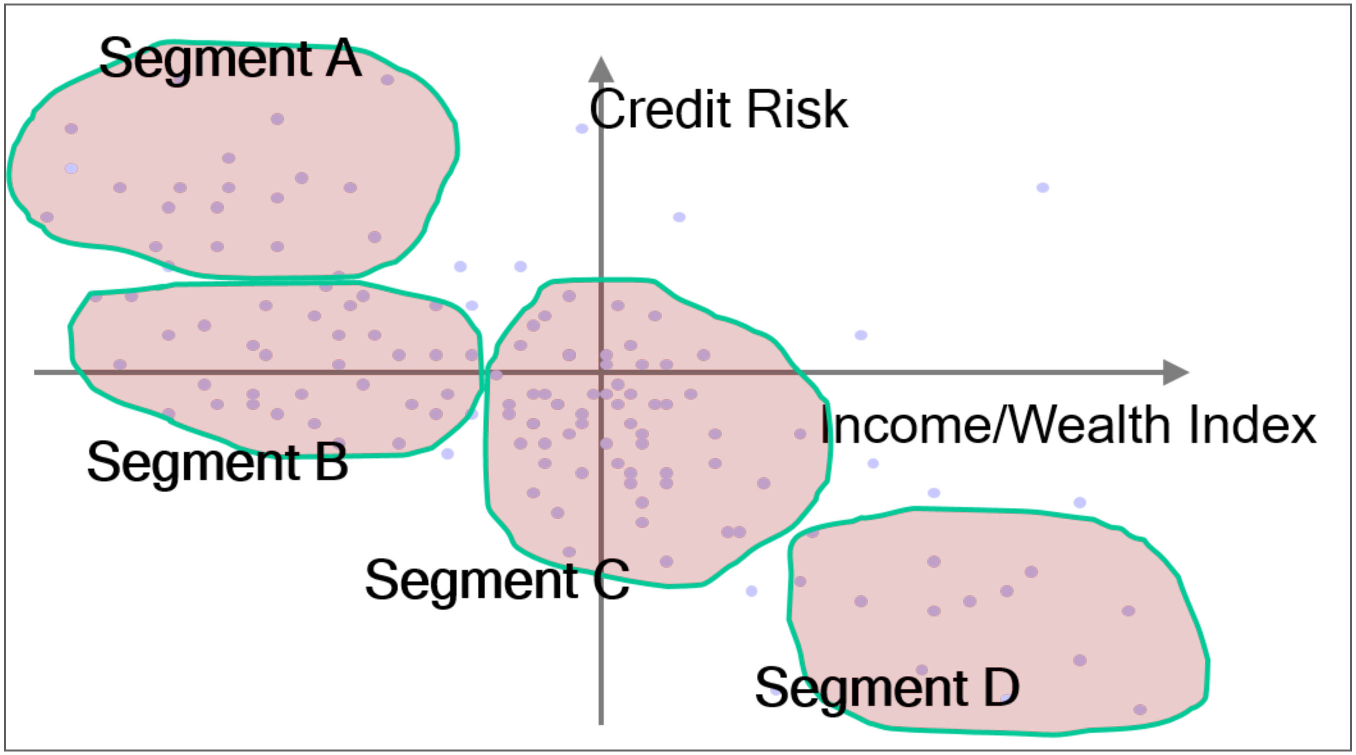
Activation

- How do we bring the segments to life?
- How can we learn more about the segments?
- How can we implement the segmentation?

"More than Nice to Know"

Cluster analysis

- “Unsupervised learning”:
  techniques to classify, describe, summarize unlabeled data



K-Means

- Simple, elegant approach to define $k = 1, \ldots, K$ segments
- Main idea: Choose K centroids $\{C_1, \ldots, C_K\}$ to minimize total within-segment variation:

$$\min \sum_{k=1}^{K} W(C_k)$$

- where $W(C_k)$ measures variation among customers assigned to segment k

K-Means

- Most common $W(C_k)$ function is Euclidean distance
- Given a set of $i \in I_k$ customers in segment k, each with $p = 1, \ldots, P$ measured attributes x_{ip},

$$W(C_k) = \sqrt{\sum_{i \in I_k} \sum_{p} (x_{ip} - \bar{x}_{kp})^2}$$

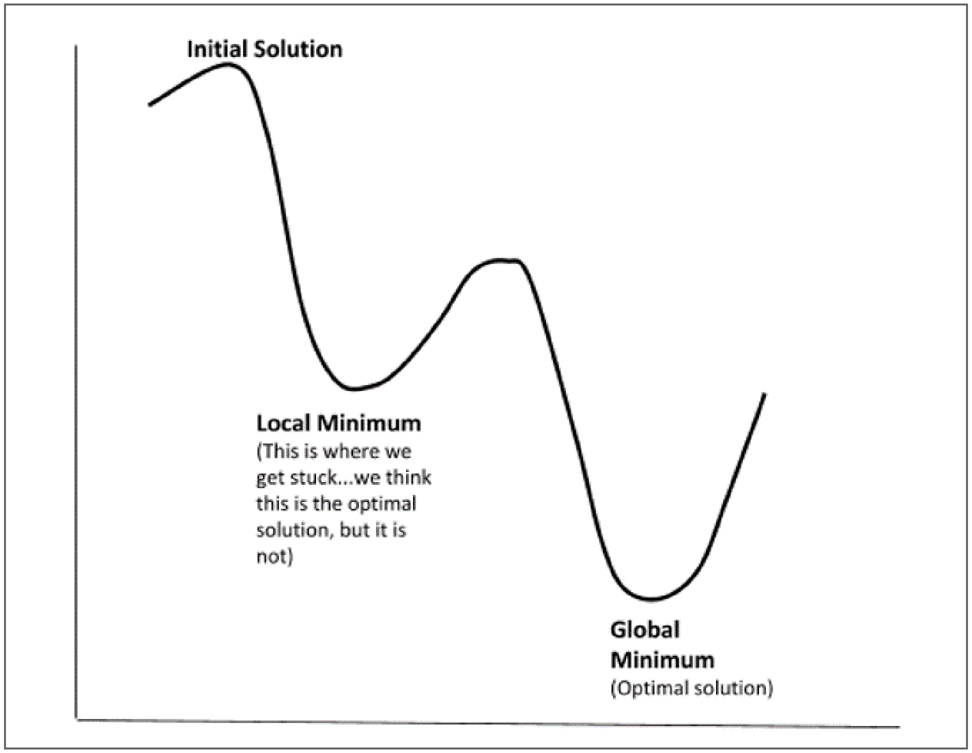
where $\bar{x}_{kp}$ is the average of $\bar{x}_{ip}$ for all $i \in I_k$, and the centroid is $C_k = (\bar{x}_{k1}, \ldots, \bar{x}_{kP})$

K-Means Algorithm

- How do we assign customers to segments?
- There are nearly K^n ways to partition n obs into K clusters
- Happily, a simple algorithm finds a local optimum:

1. Randomly choose K centroids
2. Assign every customer to nearest centroid
3. Compute new centroids based on customer assignments
4. Iterate 2-3 until convergence
5. (Optional) Repeat 1-4 for many random centroids

```
You will run this in Week 2 script
Let's illustrate it in class
```


Initial Solution
Local Minimum
(This is where we get stuck...we think this is the optimal solution, but it is not)
Global Minimum
(Optimal solution)

- We usually use 'hillclimbers' to optimize numeric functions
- Imagine I asked you to find the highest point on campus; how? On Earth?
- W(C_k) is not globally concave, so we can't guarantee a global minimum
- Thus, we pick many starting points, and see which offers the lowest W(C_k)
- Note: Some algos promise to find global minimum, but this is only provable for a globally convex function. This claim can be a

```
'tell'
```

- Meet your study group. Create a group chat. Schedule to meet weekly in person to discuss homework.
- Previously called “CRM systems,” “data warehouses,” “data lakes”

Market Mapping

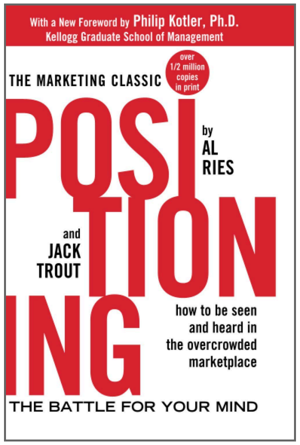
- Positioning in attribute space
- Economic theories of differentiation: Vertical, horizontal
- PCA & Perceptual maps

Marketing strategy

- *S*egmentation: How do customers differ
- *T*argeting: Which segments do we seek to attract and serve
- *P*ositioning

```
    - What value proposition do we present
    - How do our product's objective attributes compare to competitors
    - Where do customers perceive us to be
    - How do we want to influence consumer perceptions
```

- Market mapping helps with Positioning


With a New Foreword by Philip Kotler, Ph.D.
Kellogg Graduate School of Management
THE MARKETING CLASSIC
over 1/2 million copies in print
POSITIONING
by AL RIES
and JACK TROUT
how to be seen and heard in the overcrowded marketplace
THE BATTLE FOR YOUR MIND

Market Maps

- *Market maps* use customer data to depict competitive situations. Why?

```
- Understand brand/product positions in the market
- Track changes
- Identify new products or features to develop
- Understand competitor imitation/differentiation decisions
- Evaluate results of recent tactics
- Cross-selling, advertising, identifying complements or substitutes, bundles...
```

Market maps

- We often lack ground truth data

```
- Using a single map to set strategy is risky
```

- Repeated mapping builds confidence (“Movies, not pictures”)
- Many large brands do this regularly

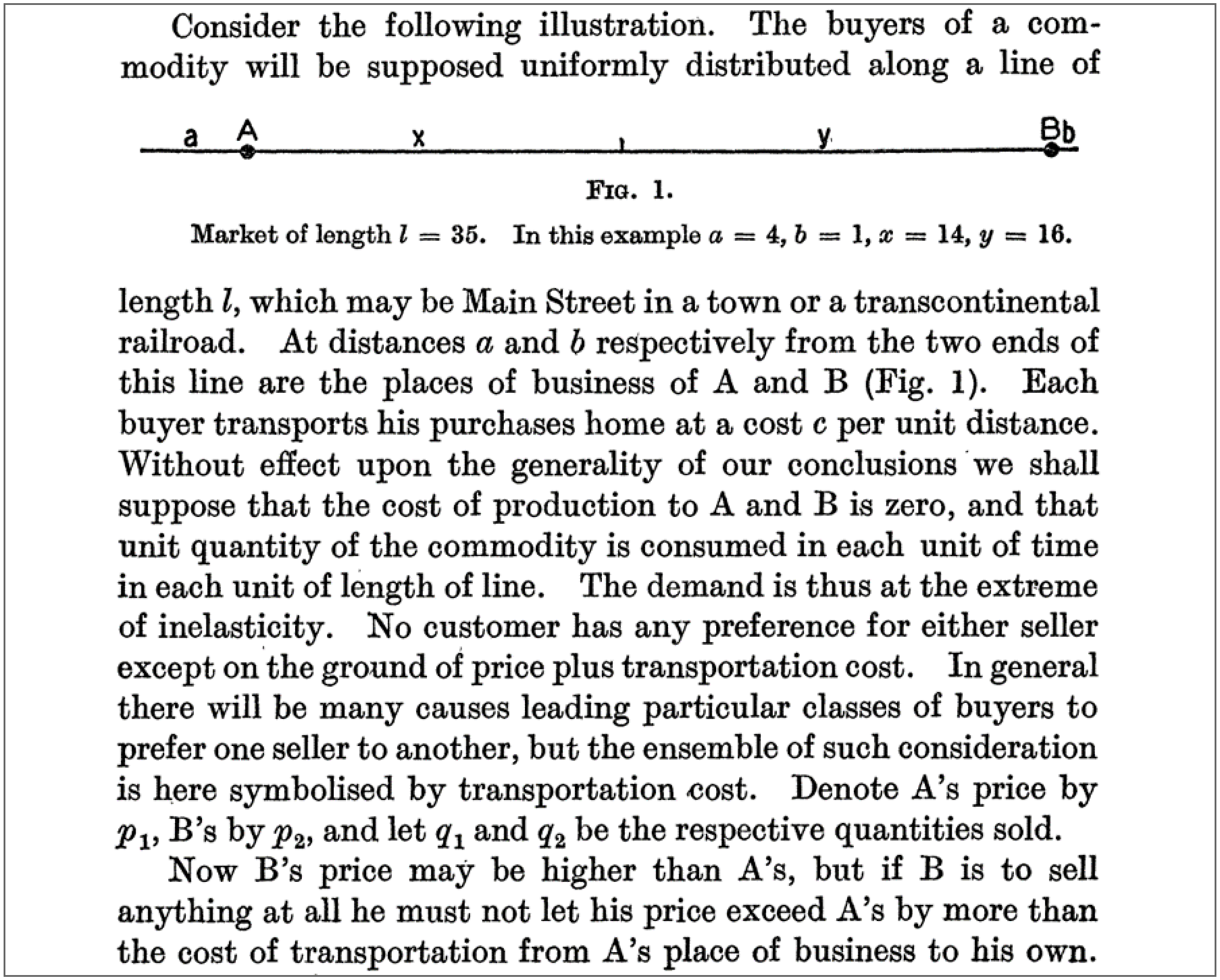
Vertical Diff., AKA quality

- Product attributes where more is better, all else constant
  - Efficacy, e.g. CPU speed or horsepower
  - Efficiency, e.g. power consumption
  - Input good quality (e.g. clothes, food)
- Important: not everyone buys the better option (why not?)

Horizontal Diff., AKA fit or match

- Product attributes w heterogeneous valuations
  - Physical location
  - Familiarity, e.g. what you grew up with
  - Taste, e.g. sweetness or umami
  - Brand image, e.g. Tide, Jif, Coca-Cola
  - Complements, e.g. headphones or charging cables

Hotelling (1929)



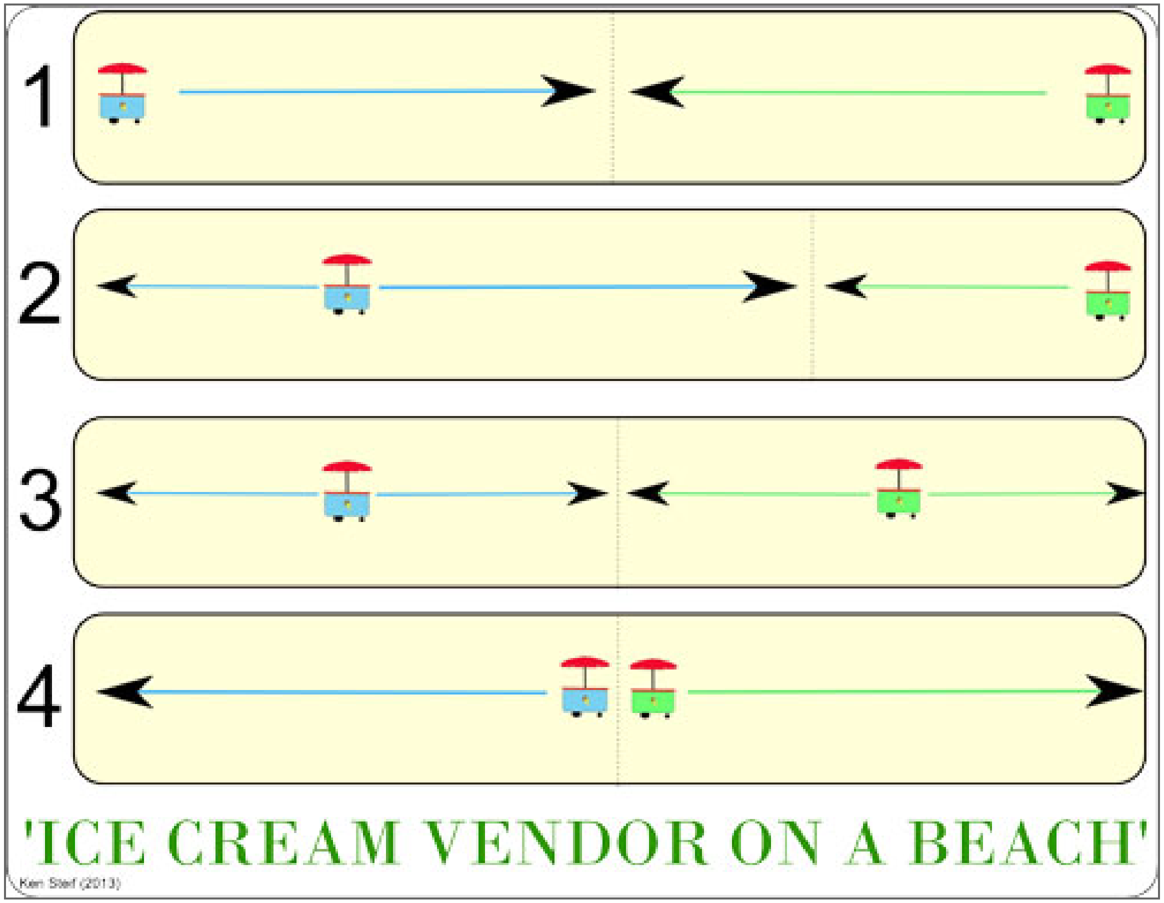
Consider the following illustration. The buyers of a commodity will be supposed uniformly distributed along a line of length l, which may be Main Street in a town or a transcontinental railroad. At distances a and b respectively from the two ends of this line are the places of business of A and B (Fig. 1). Each buyer transports his purchases home at a cost c per unit distance. Without effect upon the generality of our conclusions we shall suppose that the cost of production to A and B is zero, and that unit quantity of the commodity is consumed in each unit of time in each unit of length of line. The demand is thus at the extreme of inelasticity. No customer has any preference for either seller except on the ground of price plus transportation cost. In general there will be many causes leading particular classes of buyers to prefer one seller to another, but the ensemble of such consideration is here symbolised by transportation cost. Denote A's price by p_1, B's by p_2, and let q_1 and q_2 be the respective quantities sold.

FIG. 1.

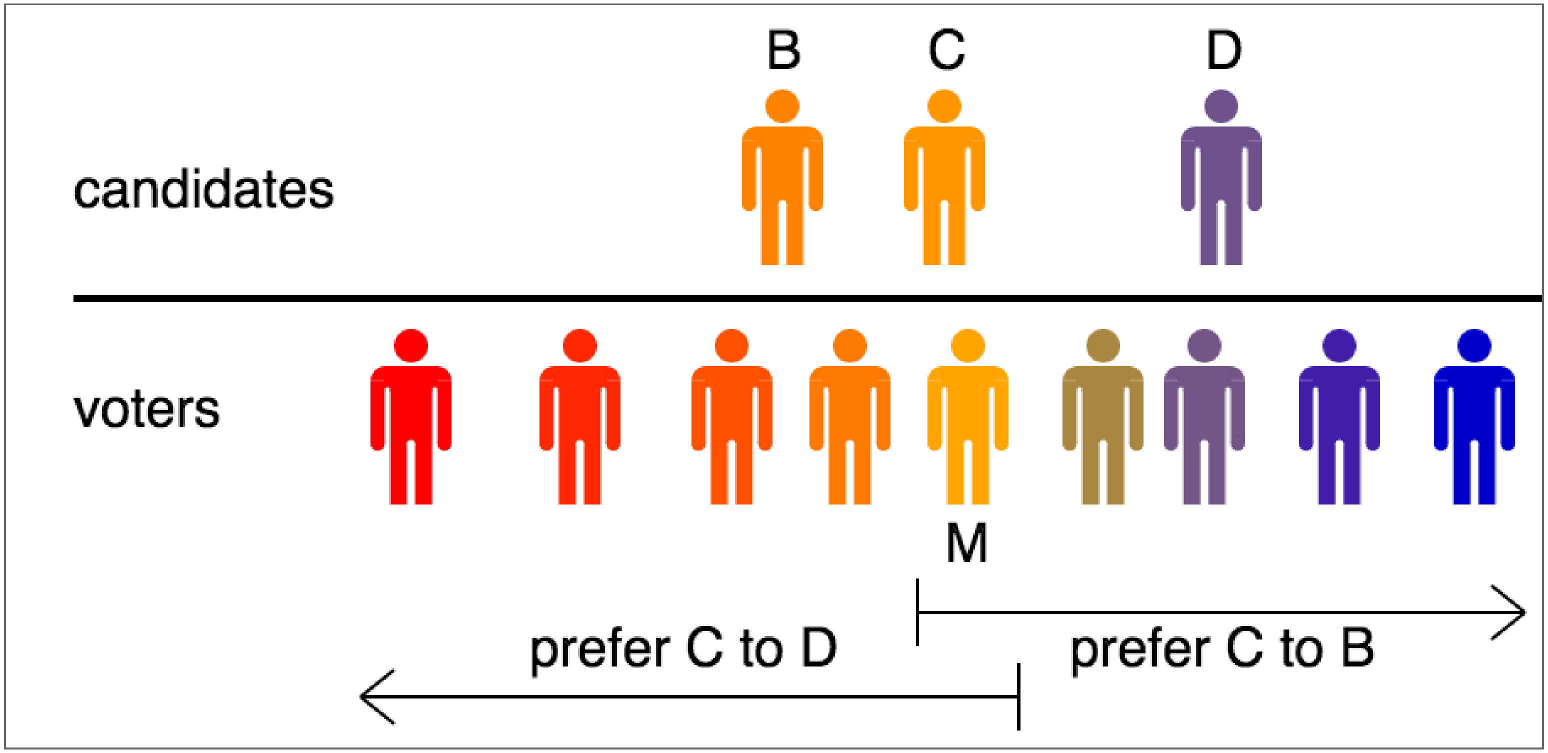
Market of length $l = 35$. In this example $a = 4$, $b = 1$, $x = 14$, $y = 16$.

Now B's price may be higher than A's, but if B is to sell anything at all he must not let his price exceed A's by more than the cost of transportation from A's place of business to his own.

Ice cream vendors

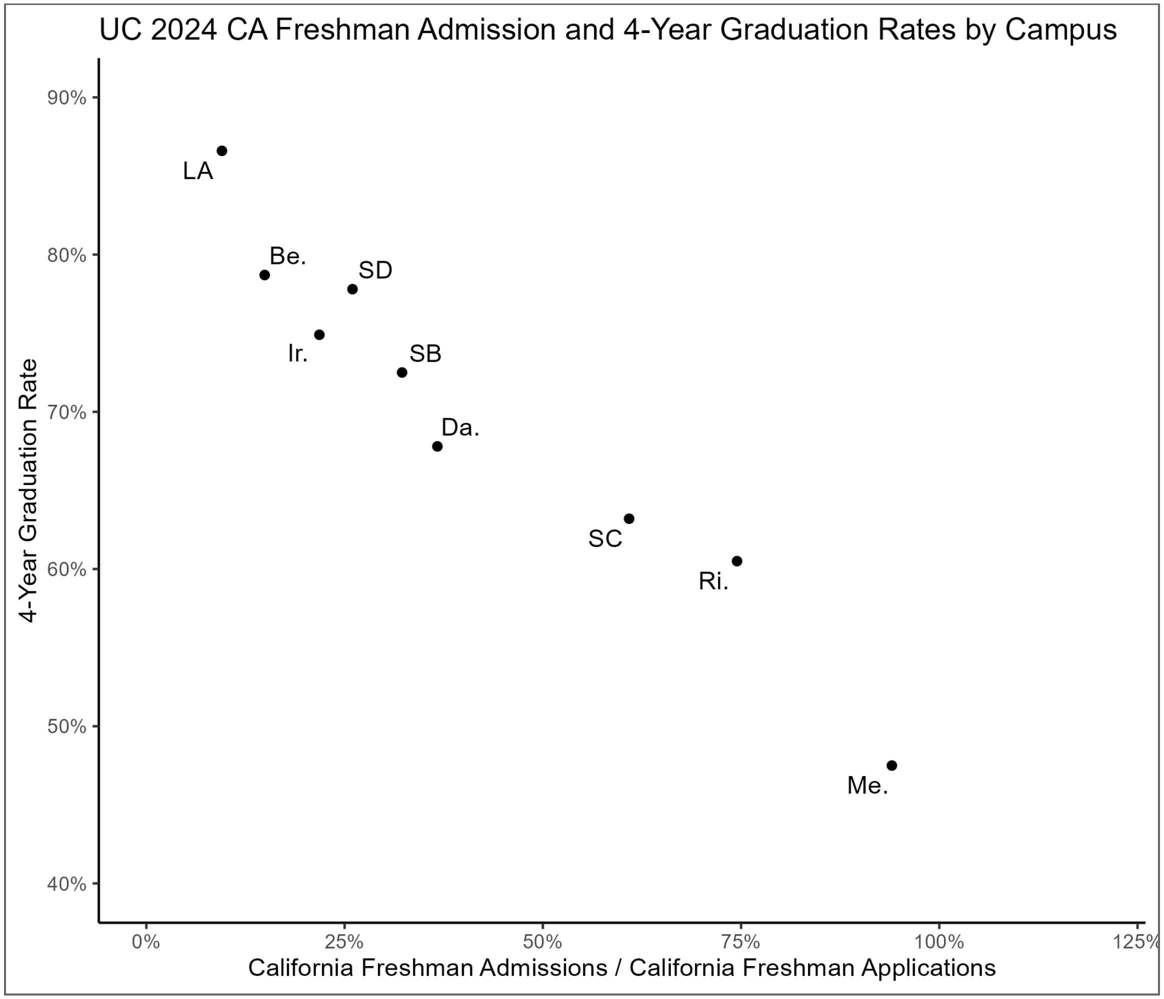


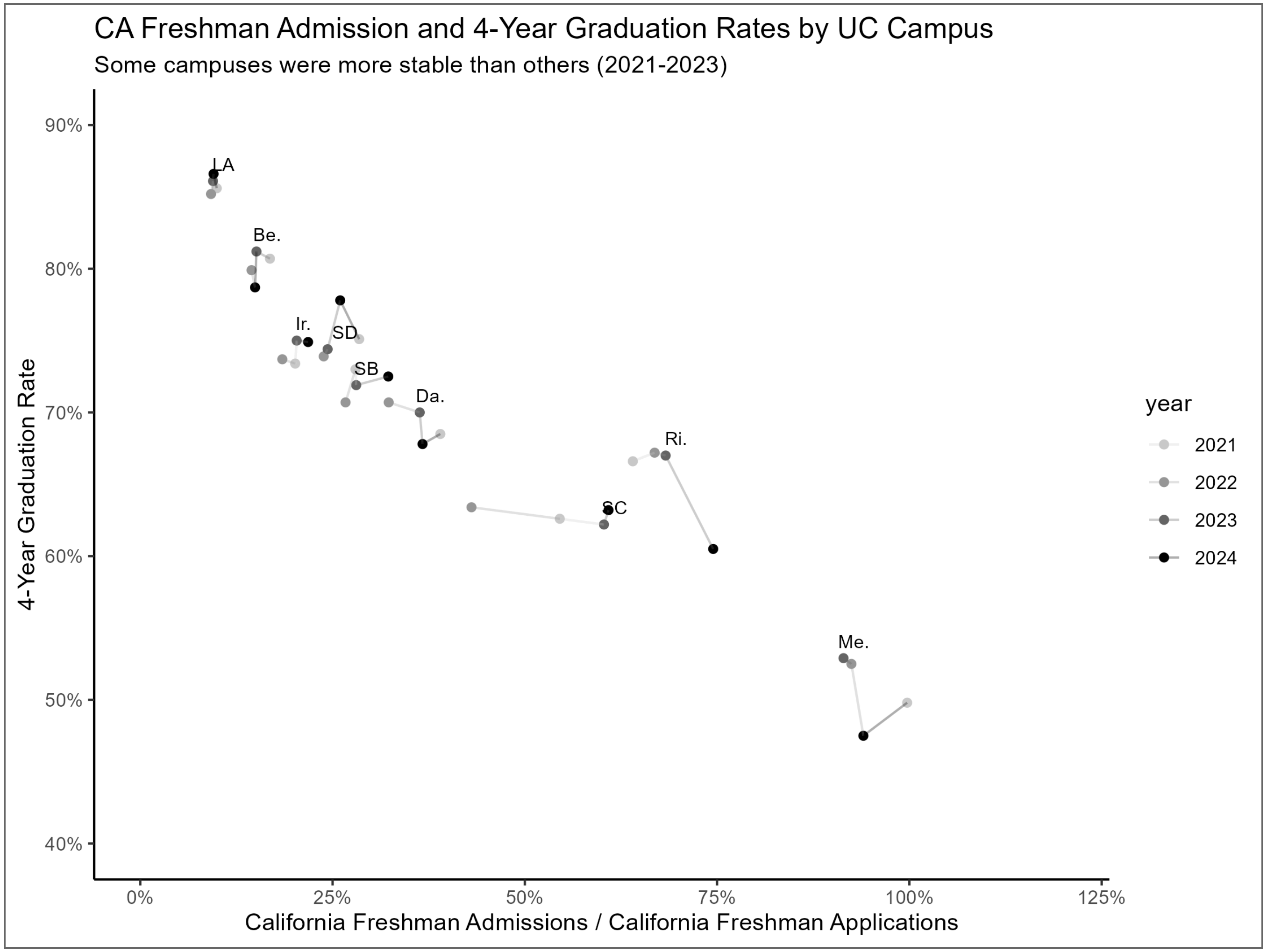
Median voter theorem



- Suppose you are the UCSD Chancellor, tasked with increasing in-state freshman enrollments
- You want to map UC campuses in the market for California freshman applicants
- You posit that selectivity and time-to-degree matter most

```
- Students want to connect with smart students
- Students want to graduate on time
```


UC 2024 CA Freshman Admission and 4-Year Graduation Rates by Campus
4-Year Graduation Rate
90%
80%
70%
60%
50%
40%
0%
25%
50%
75%
100%
125%
California Freshman Admissions / California Freshman Applications
LA
Be.
SD
Ir.
SB
Da.
SC
Ri.
Me.


CA Freshman Admission and 4-Year Graduation Rates by UC Campus
Some campuses were more stable than others (2021-2023)
4-Year Graduation Rate
90%
80%
70%
60%
50%
40%
0%
25%
50%
75%
100%
125%
California Freshman Admissions / California Freshman Applications
LA
Be.
Ir.
SD
SB
Da.
Ri.
SC
Me.
year
2021
2022
2023
2024

What if there are too many product attributes to graph?

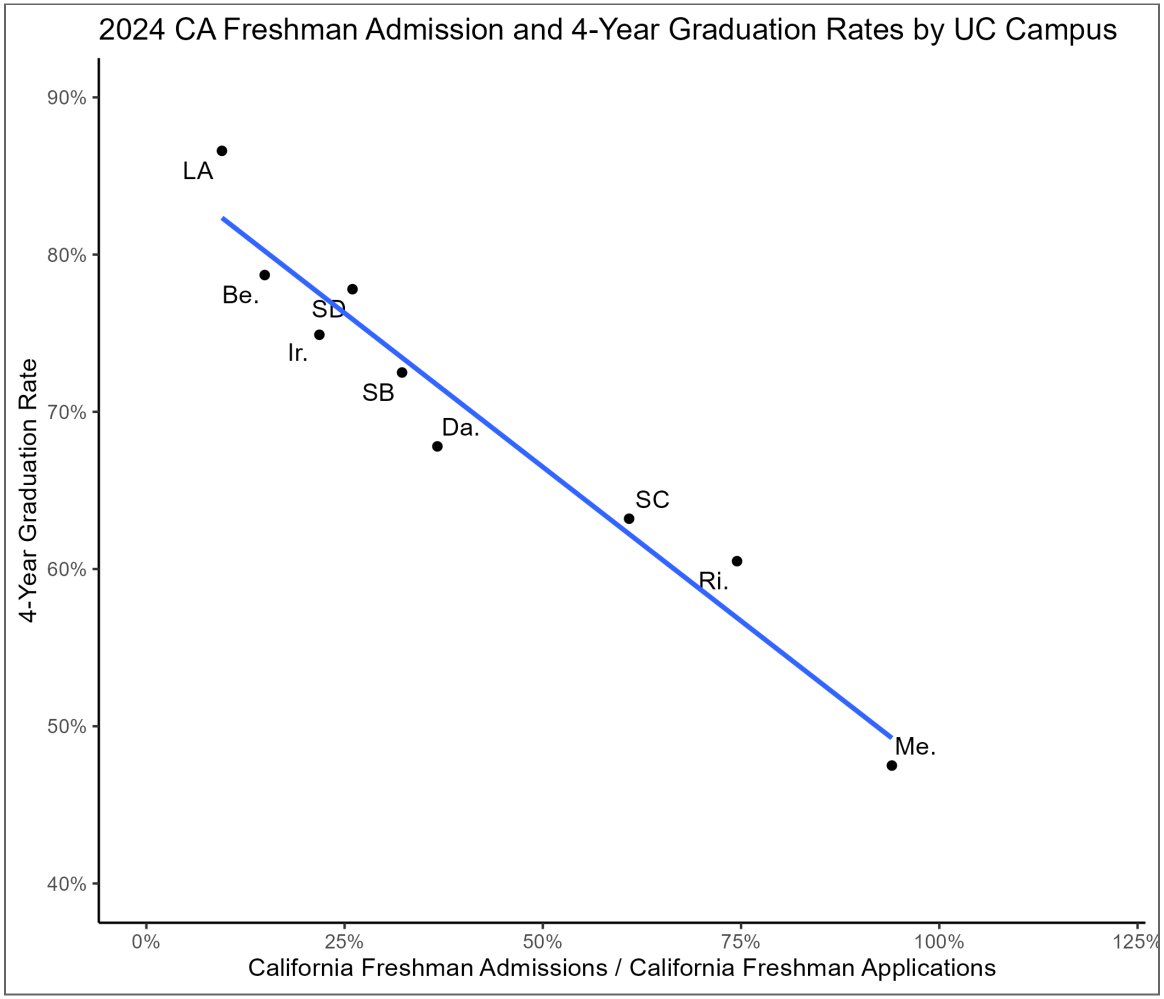
- Enter Principal Components Analysis

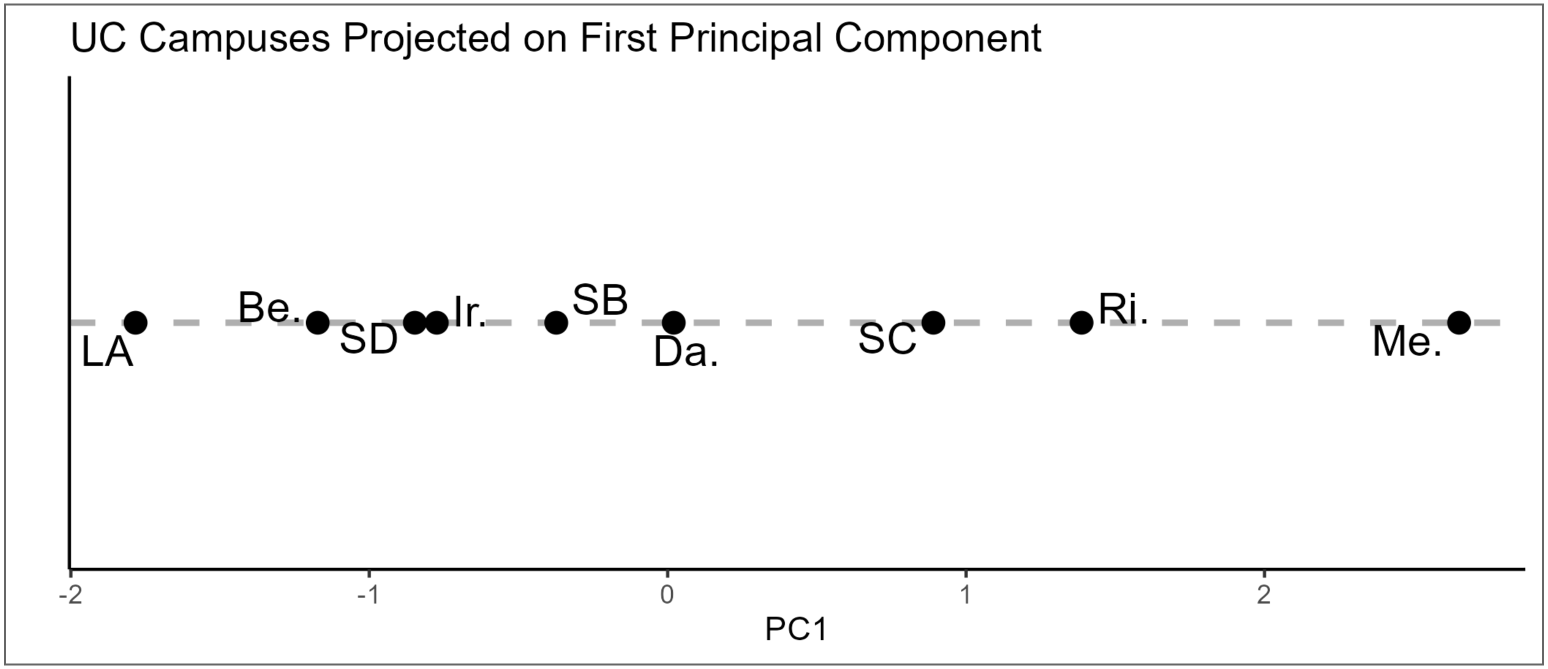
```
- Powerful way to summarize data
- Projects high-dimensional data into a lower dimensional space
- Designed to minimize information loss during compression
- Pearson (1901) invented; Hotelling rediscovered (1933 & 36)
```

Principal Components Analysis (PCA)

1. Store K continuous attributes for $J > K$ products in X, a $J \times K$ matrix
2. Consider X a K-dimensional space containing J points
3. Calculate $X'X$, a $K \times K$ covariance matrix of the attributes
4. 1st n eigenvectors of the attribute covariance matrix give unit vectors to map products in n-dimensional space

```
- We'll use first 1 or 2 eigenvectors for visualization
```


2024 CA Freshman Admission and 4-Year Graduation Rates by UC Campus
4-Year Graduation Rate
90%
80%
70%
60%
50%
40%
LA
Be.
SD
Ir.
SB
Da.
SC
Ri.
Me.
0%
25%
50%
75%
100%
125%
California Freshman Admissions / California Freshman Applications


UC Campuses Projected on First Principal Component
LA
Be.
SD
Ir.
SB
Da.
SC
Ri.
Me.
-2
-1
0
1
2
PC1

PCA FAQ

1. How do I interpret the principal components?

```
- Each principal component is a linear combination of the larger space's axes
- Principal components are the "new axes" for the newly-compressed space
- Principal components are always orthogonal to each other, by construction
```

2. What are the main assumptions of PCA?

```
- Variables are continuous and linearly related
- Principal components that explain the most variation matter most
- Drawbacks: information loss, reduced spatial interpretability, outlier sensitivity
```

3. How do I choose the # of principal components?

```
- Business criteria: 1 or 2 if you want to visualize the data
- Business criteria: Or, value of compressed data in subsequent operations
- Statistical criteria: Cume variance explained, scree plot, eigenvalue > 1
```

4. What are some similar tools to PCA?

```
- Factor analysis, linear discriminant analysis, independent component analysis...
```

How does PCA relate to K-means?

- K-Means identifies clusters within a dataset

```
- K-Means augments a dataset by identifying similarities within it
- K-Means never discards data
```

- PCA combines data dimensions to condense data with minimal information loss

```
- PCA is designed to optimally reduce data dimensionality
- PCA facilitates visual interpretation but does not identify similarities
```

- Both are unsupervised ML algos

```
- Both have "tuning parameters" (e.g. # segments, # principal components)
- They serve different purposes & can be used together
- E.g. run PCA to first compress large data, then K-Means to group points
- Or, K-Means to identify clusters, then PCA to visualize them in 2D space
```

Conceptual organization

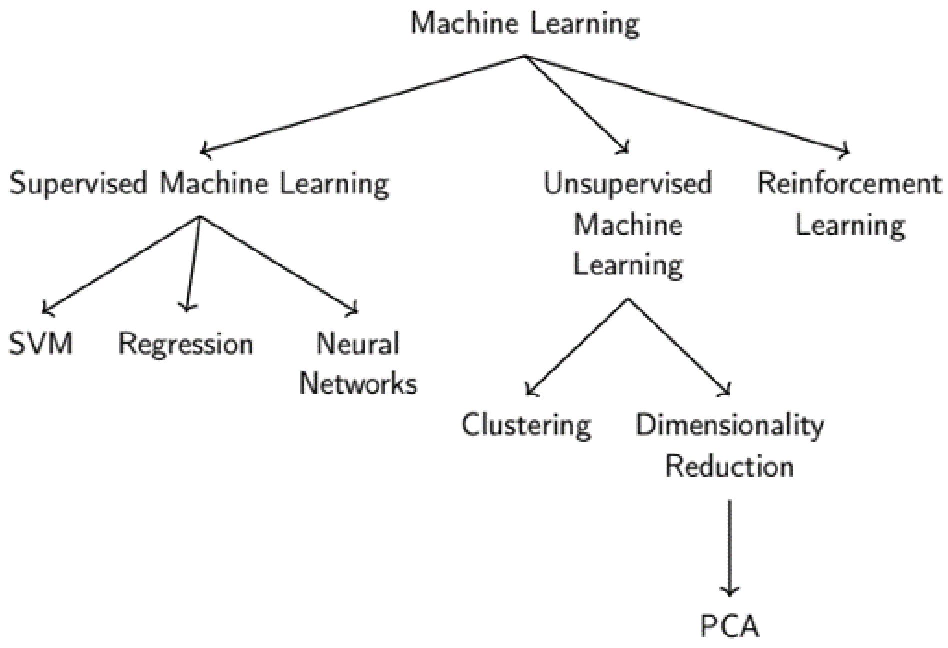


Figure 15: *Machine learning task solution space and model families*

Mapping Practicalities

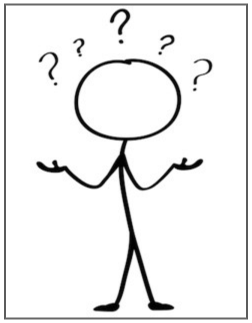
1. How to measure intangible attributes like trust?

```
- Ask consumers, e.g. "How much do you trust this brand?"
- Marketing Research techniques measure subjective attributes and perceptions
```

2. What if we don’t know, or can’t measure, the most important attributes?

```
- Multidimensional scaling
```

3. How should we weigh attributes?



Do we know the most important attributes?

- *Multidimensional scaling* draws *perceptual maps*

1. Suppose you can measure product similarity
2. For J products, populate the $J \times J$ matrix of similarity scores

```
   - With J brands, we have J points in J dimensions. Each dimension j indicates similarity to
brand j. PCA can projects J dimensions into 2D for plotting
```

3. Use PCA to reduce to a lower-dimensional space

```
   - Pro: We don't need to predefine attributes
   - Con: Axes can be hard to interpret
```

Multidimensional scaling

MDS Intuition, in 2D space

```
      - With a ruler and map, measure distances between 20 US cities ("similarity")
      - Record distances in a 20x20 matrix: PCA into 2D should recreate the map
      - But, we don't usually know the map we are recreating, so we look for ground-truth comparisons to indicate credibility and reliability
```

Examples:

```
      - Poli Sci: Political candidate positioning, eg left to right
      - Psychologists: Understand perceptions and evaluation of personality traits
      - Marketers: how consumers perceive brands or products
```

Example: Netzer et al. (2012)

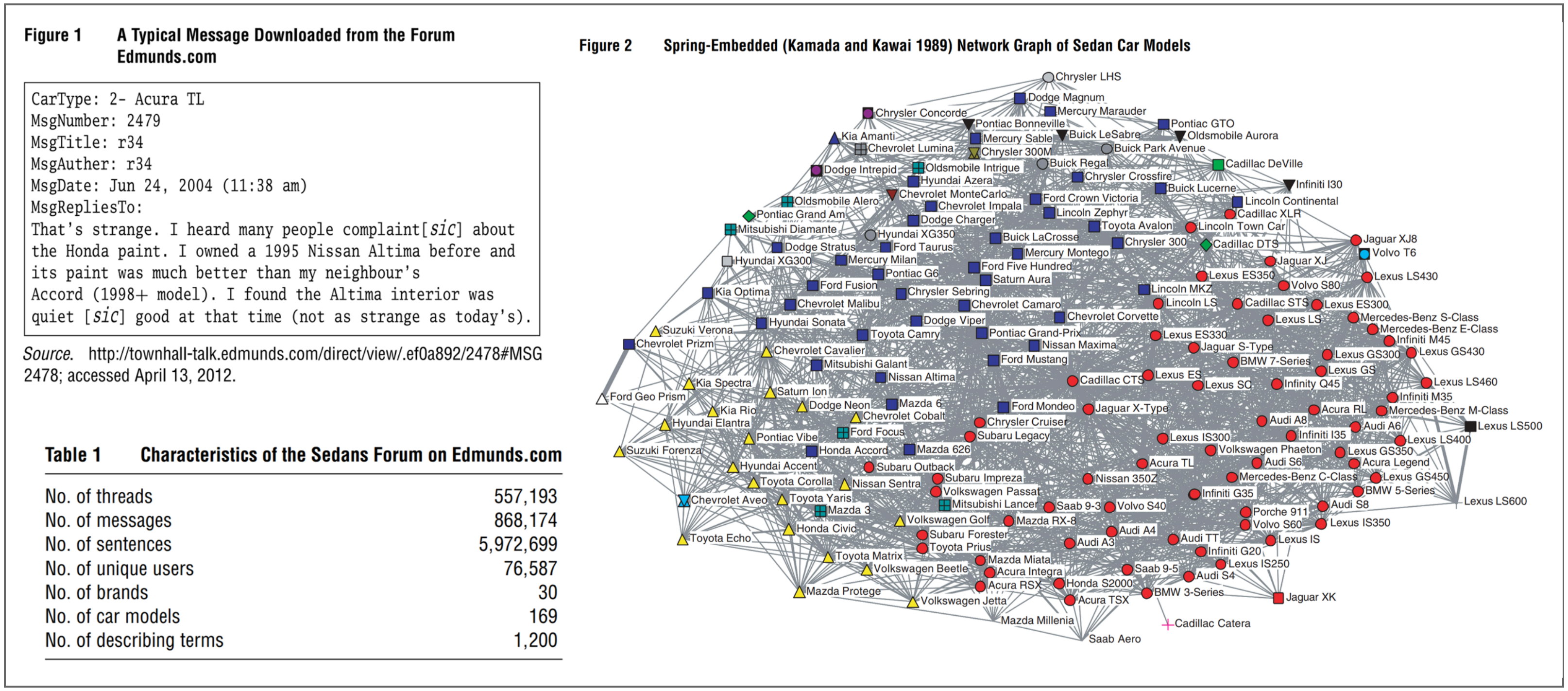


Figure 1 A Typical Message Downloaded from the Forum Edmunds.com

```
CarType: 2- Acura TL
MsgNumber: 2479
MsgTitle: r34
MsgAuther: r34
MsgDate: Jun 24, 2004 (11:38 am)
MsgRepliesTo:
That's strange. I heard many people complaint[sic] about
the Honda paint. I owned a 1995 Nissan Altima before and
its paint was much better than my neighbour's
Accord (1998+ model). I found the Altima interior was
quiet [sic] good at that time (not as strange as today's).
```

Source. http://townhall-talk.edmunds.com/direct/view/.ef0a892/2478#MSG2478; accessed April 13, 2012.

Table 1 Characteristics of the Sedans Forum on Edmunds.com

No. of threads	557,193
No. of messages	868,174
No. of sentences	5,972,699
No. of unique users	76,587
No. of brands	30
No. of car models	169
No. of describing terms	1,200

Figure 2 Spring-Embedded (Kamada and Kawai 1989) Network Graph of Sedan Car Models

Figure 3 **MDS Map of Discussion of Car Brands**

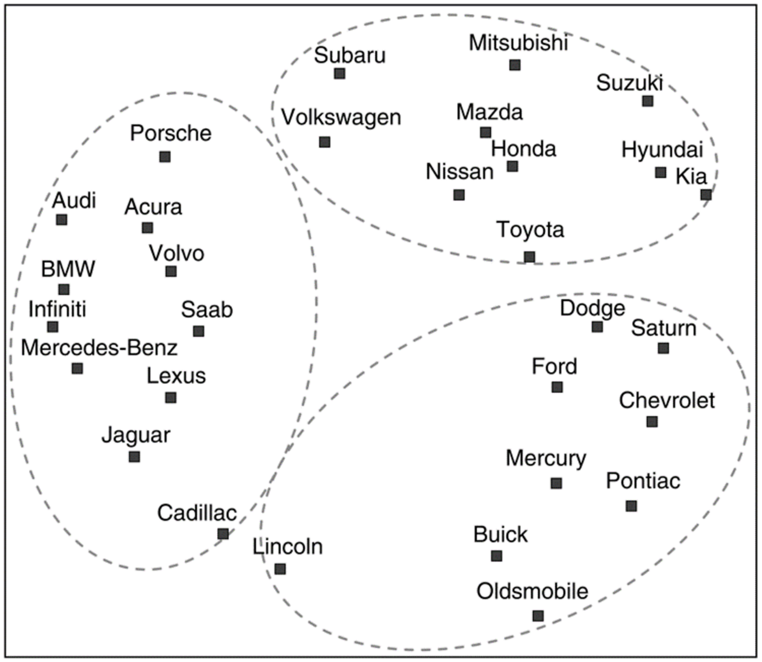


Figure 4 **MDS of Car Brands Using Car-Switching Data**

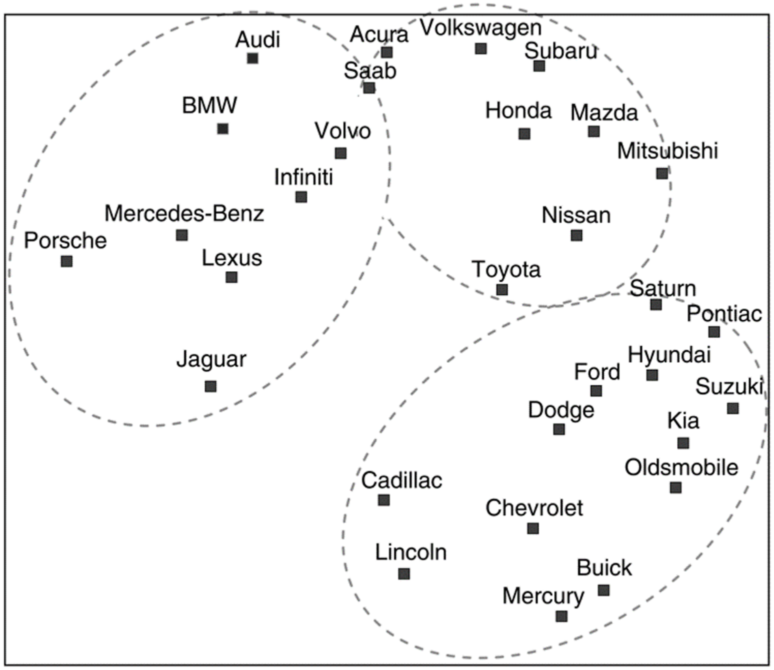
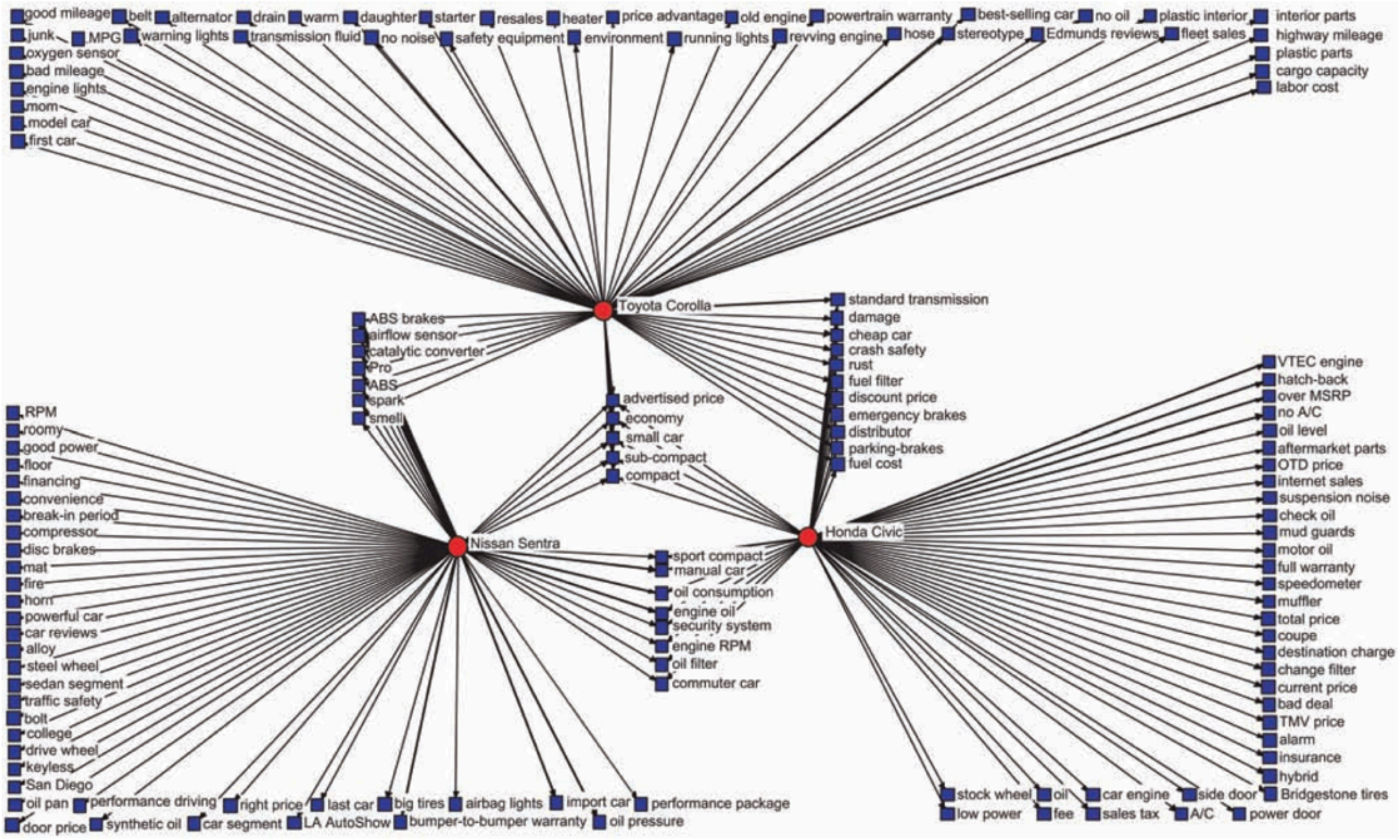


Figure 8 Terms Commonly Appearing with the Honda Civic, Nissan Sentra, and Toyota Corolla



How to weigh product attributes?

- *Demand modeling* uses product attributes and prices to explain customer purchases
- Now, just graph the 1-dim line with the original points projected onto it
- Do the relative conclusions hold up? What does PC1 mean?
- *Heterogeneous demand modeling* uses product attributes, prices and customer attributes to explain purchases
- How much information did we lose? Compare to previous graphs
-

```
      - "Revealed preferences": Demand models explain observed choices in uncontrolled market
  environments
```

-
- How do we reach equilibrium?

source code & data

Source: “Mine Your Business” by Netzer et al. (2012)

Text data

- The Challenge
- Embeddings
- LLMs: What are they doing
- What does it all mean?

The Challenge

- Suppose an English speaker knows n words, say $n = 10,000$
- How many unique strings of N words can they generate?

```
- N=1: 10,000
- N=2: 10,000^2=100,000,000
- N=3: 10,000^3=1,000,000,000,000=1 Trillion
- N=4: 10,000^4=10^16
- N=5: 10,000^5=10^20
- N=6: 10,000^5=10^24=1 Trillion Trillions
- ....
```

- Why do we make kids learn proper grammar?

```
- Average formal written English sentence is ~15 words
```

Embeddings

- represent words as vectors in high-dim space

```
    - Really, "tokens," but assume words==tokens for simplicity
```

- Assume W words, $A < W$ abstract concepts

```
    - Assume we have all text data from all history. Each sentence is a point in $W$-dimensional space
```

- We could run PCA to reduce from W to A dimensions

```
    - Assume we have infinite computing resources
    - We now have every sentence represented as a point in continuous A-space
```

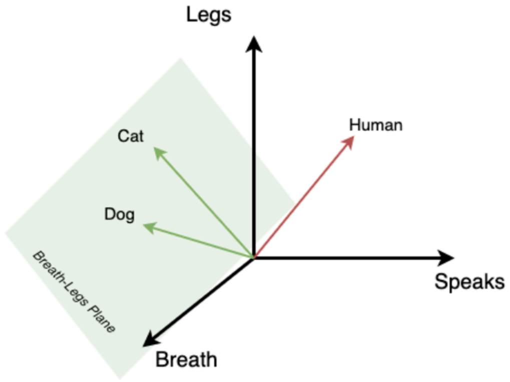
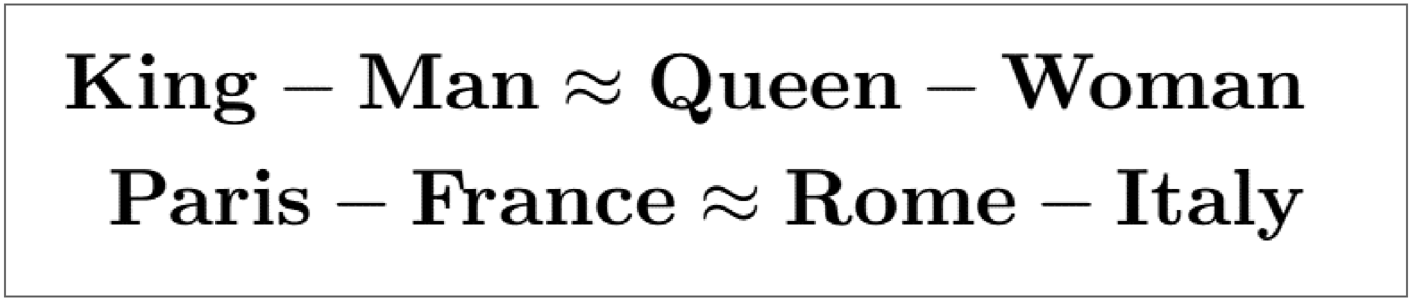


Figure created by Oleg Borisov. Axis denote the features that could have been learned by the embedding. As we can see Cat and Dog vectors can only lie in the Legs and Breath hyperplane, while a Human word-vector representation lies in the hyperplane of those 3 features.

Cool things about embeddings

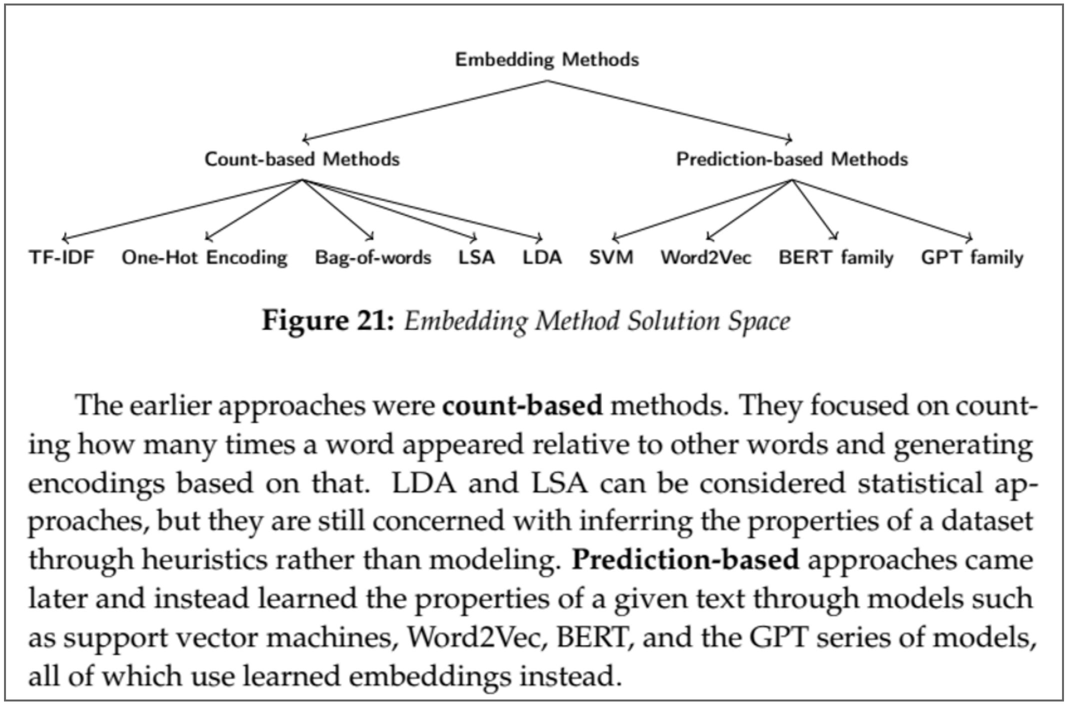
- Compression stores enormous textual data in a small space, other than human memory
- We can do math using words!



$$\textbf{King} - \textbf{Man} \approx \textbf{Queen} - \textbf{Woman}$$

$$\textbf{Paris} - \textbf{France} \approx \textbf{Rome} - \textbf{Italy}$$

Many ways to encode embeddings



Embedding Methods
Count-based Methods
Prediction-based Methods
TF-IDF
One-Hot Encoding
Bag-of-words
LSA
LDA
SVM
Word2Vec
BERT family
GPT family

Figure 21: *Embedding Method Solution Space*

The earlier approaches were **count-based** methods. They focused on counting how many times a word appeared relative to other words and generating encodings based on that. LDA and LSA can be considered statistical approaches, but they are still concerned with inferring the properties of a dataset through heuristics rather than modeling. **Prediction-based** approaches came later and instead learned the properties of a given text through models such as support vector machines, Word2Vec, BERT, and the GPT series of models, all of which use learned embeddings instead.

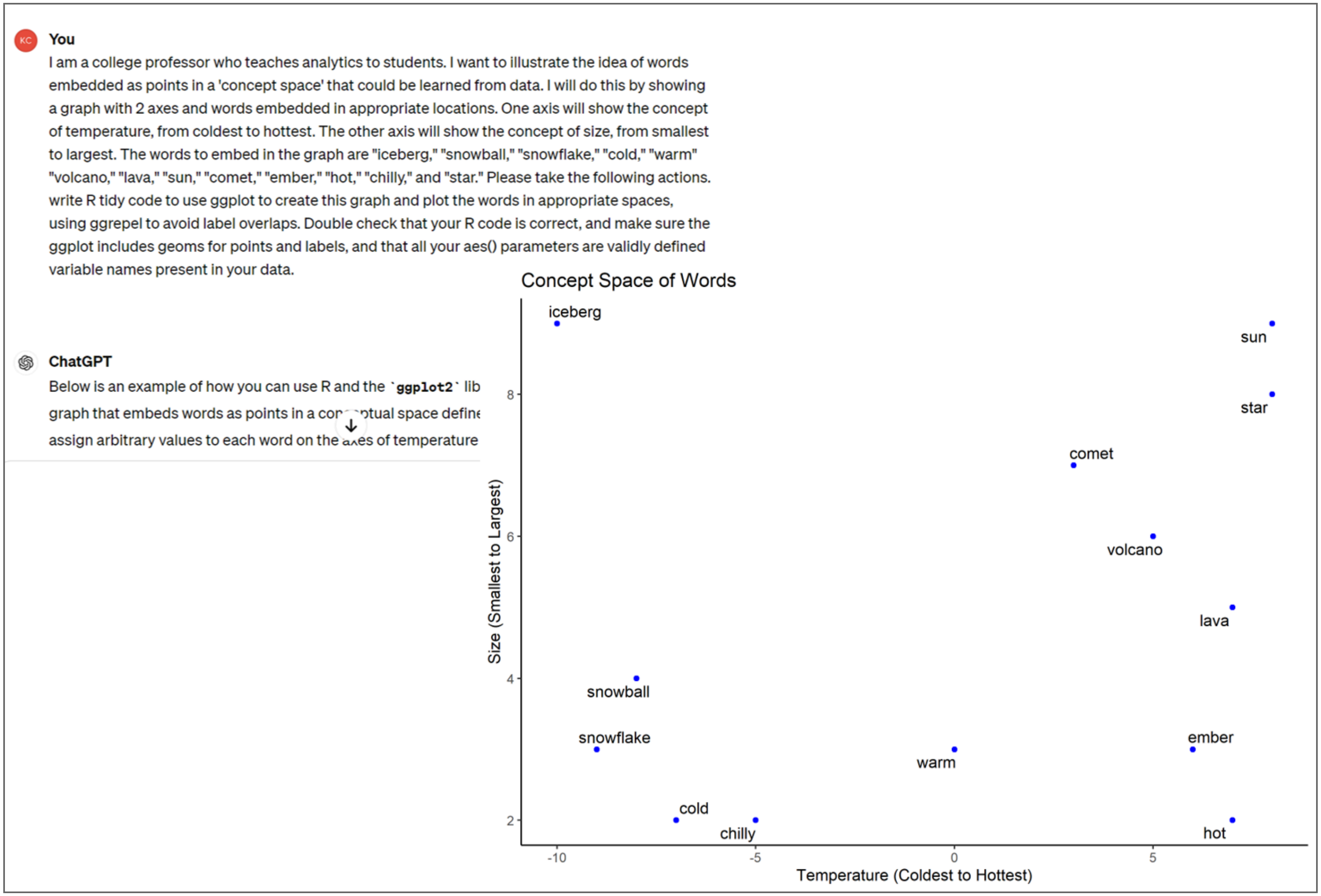
LLMs: Given a prompt,

1. Recode prompt to maximize contextual understanding

```
    - 'the bank of the river is steep' vs 'the bank near the river is solvent'
    - This is the 'attention' step you hear a lot about
    - Basically, modify every word's location based on every other word's position in the prompt
sequence
```

2. Feed recoded prompt into transformer as a *sequence* of points in concept-space
3. Predict the next point and add it to the sequence
4. Repeat step 3 until no more good predictions
5. Repeat steps 1-4 many many times, then hire humans to evaluate results, use evaluations for RLHF to refine the process
6. Add ‘reasoning’ via reinforcement learners, and ‘deep research’ via agentic tool use
7. Sell access to customers, then train a bigger LLM

Example: Concept Space



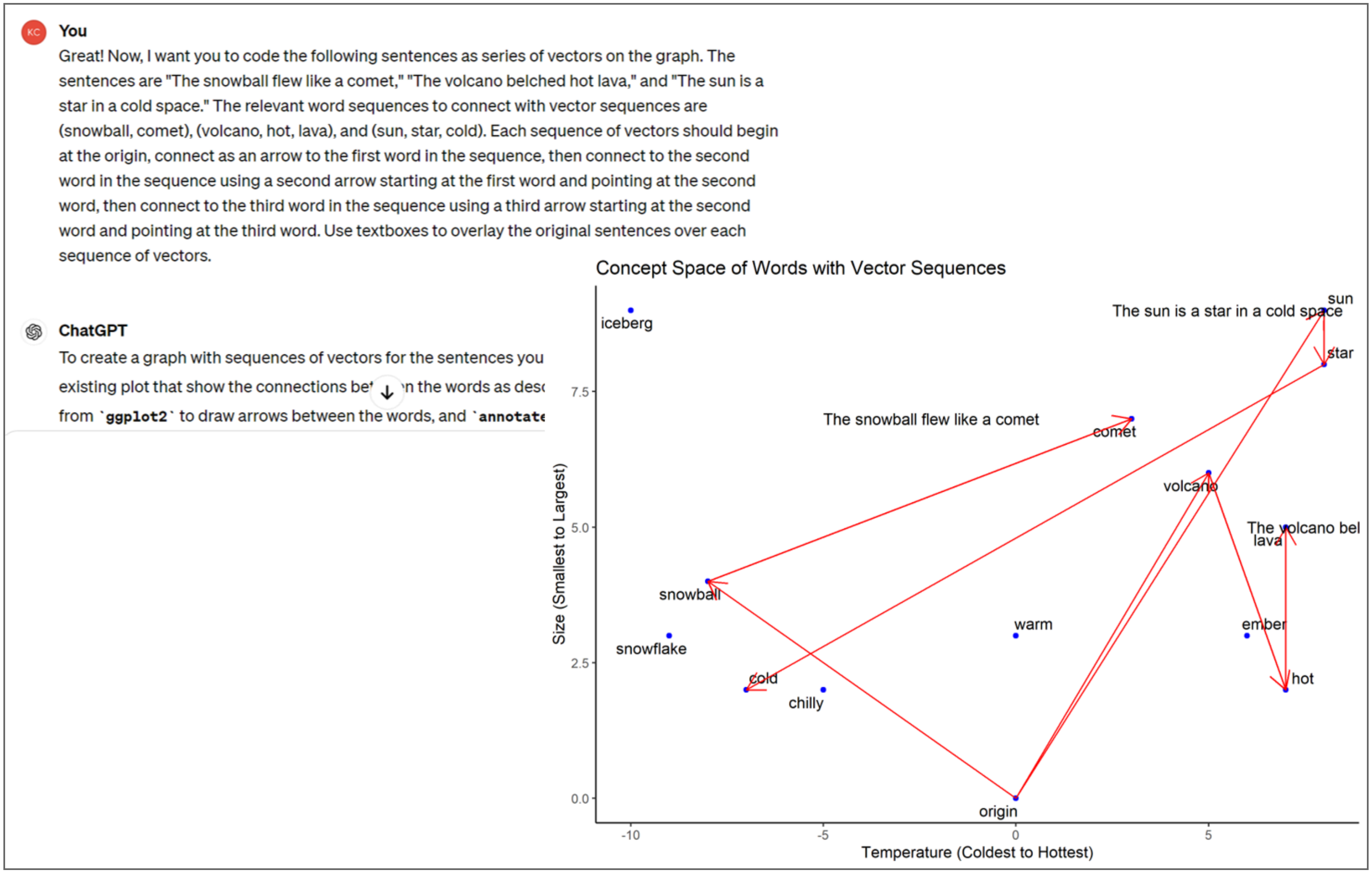
You

I am a college professor who teaches analytics to students. I want to illustrate the idea of words embedded as points in a 'concept space' that could be learned from data. I will do this by showing a graph with 2 axes and words embedded in appropriate locations. One axis will show the concept of temperature, from coldest to hottest. The other axis will show the concept of size, from smallest to largest. The words to embed in the graph are "iceberg," "snowball," "snowflake," "cold," "warm" "volcano," "lava," "sun," "comet," "ember," "hot," "chilly," and "star." Please take the following actions. write R tidy code to use ggplot to create this graph and plot the words in appropriate spaces, using ggrepel to avoid label overlaps. Double check that your R code is correct, and make sure the ggplot includes geoms for points and labels, and that all your aes() parameters are validly defined variable names present in your data.

ChatGPT

Below is an example of how you can use R and the `ggplot2` lib graph that embeds words as points in a conceptual space define assign arbitrary values to each word on the axes of temperature

Example: Sentences as Vector Sequences



You

Great! Now, I want you to code the following sentences as series of vectors on the graph. The sentences are "The snowball flew like a comet," "The volcano belched hot lava," and "The sun is a star in a cold space." The relevant word sequences to connect with vector sequences are (snowball, comet), (volcano, hot, lava), and (sun, star, cold). Each sequence of vectors should begin at the origin, connect as an arrow to the first word in the sequence, then connect to the second word in the sequence using a second arrow starting at the first word and pointing at the second word, then connect to the third word in the sequence using a third arrow starting at the second word and pointing at the third word. Use textboxes to overlay the original sentences over each sequence of vectors.

ChatGPT

To create a graph with sequences of vectors for the sentences you existing plot that show the connections bet n the words as desc from `ggplot2` to draw arrows between the words, and `annotate

KC
Concept Space of Words with Vector Sequences
iceberg
The sun is a star in a cold space
sun
star
The snowball flew like a comet
comet
volcano
The volcano bel
lava
snowball
warm
ember
snowflake
cold
chilly
hot
origin
Size (Smallest to Largest)
7.5
5.0
2.5
0.0
-10
-5
0
5
Temperature (Coldest to Hottest)

What LLMs Can and Can’t Do

- *Can* generate intelligible semantic sequences
- *Can* summarize large text training data sets
- *Can* help humans save time and effort in semantic tasks
- *Can* uncover previously unknown relations in training data
- *Can’t* distinguish truth from frequency in training data

```
- LLMs propagate popular biases in training data, unless taught otherwise
```

- *Can’t* reliably evaluate sequences absent from training data
- *Can’t* discover new relationships absent in training data
- *Can’t* think, reason, imagine, feel, want, question

```
- But might complement other components that do such things
```

What happens next?

- No way to know. The tech is far ahead of science

```
- LLMs are productive combinations of pre-existing components
- This is normal: Eng/ML/stats theory chases applications
- Spellchecker and calculator are not productive analogies
```

- My guesses

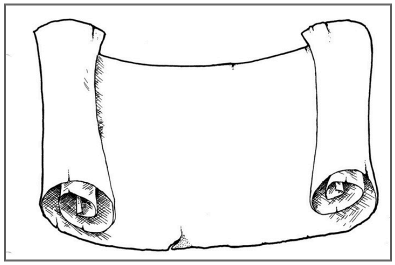
```
- "It's easy to predict everything, except for the future."
- Simple tasks: LLMs outcompete humans
- Medium-complexity tasks: LLMs help low-skill humans compete
- Complex tasks: Skillful LLM use requires highly skilled humans
- Law matters a LOT: Liability, copyright, privacy, disclosure
- In eqm, typical quality should rise; *not* using LLMs will handicap
- Long term: More automation, more products, more concentration of capital
- More word math techniques will be invented, some will be useful
```

- What future tech might complement LLMs?

```
- Reframes current argument about Sentient AI
- Robots? World models? Causal reasoning engines? Volition?
```

Class script

- Standardizing variables
- Iris example
- Running & graphing kmeans
- Use PCA to map the smartphone market
- Transformers: developed to translate languages, e.g. mapping ‘sombrero’ to



- l, they can transform data of generic type x to generic type y
- past 20 years: digital data, faster computers, better algos
- **goal code**
- graphic source

Wrapping up

Recap

- Segmentation should be based on customer needs
- Customer behavior best predicts behavior (not demos)
- Market maps depict competition, aide positioning
- PCA projects high-dim data into low-dim space w minimal information loss
- Embeddings represent words as points in concept-space, enabling word-math

Next week's reading helps avoid or reduce struggle

Going further

- Deep Dive into LLMs like ChatGPT by Karpathy (2025)
- Tracing the thoughts of a large language model
- K-Means Clustering: An Explorable Explainer
- Learning the k in k-means (Hammerly and Elkan 2003)