Data Aggregation/Summarization and Business Implications

Dai Yao
Associate Professor of Marketing
Department of Management and Marketing
dai@yaod.ai
http://www.yaod.ai

SKKU
Business School

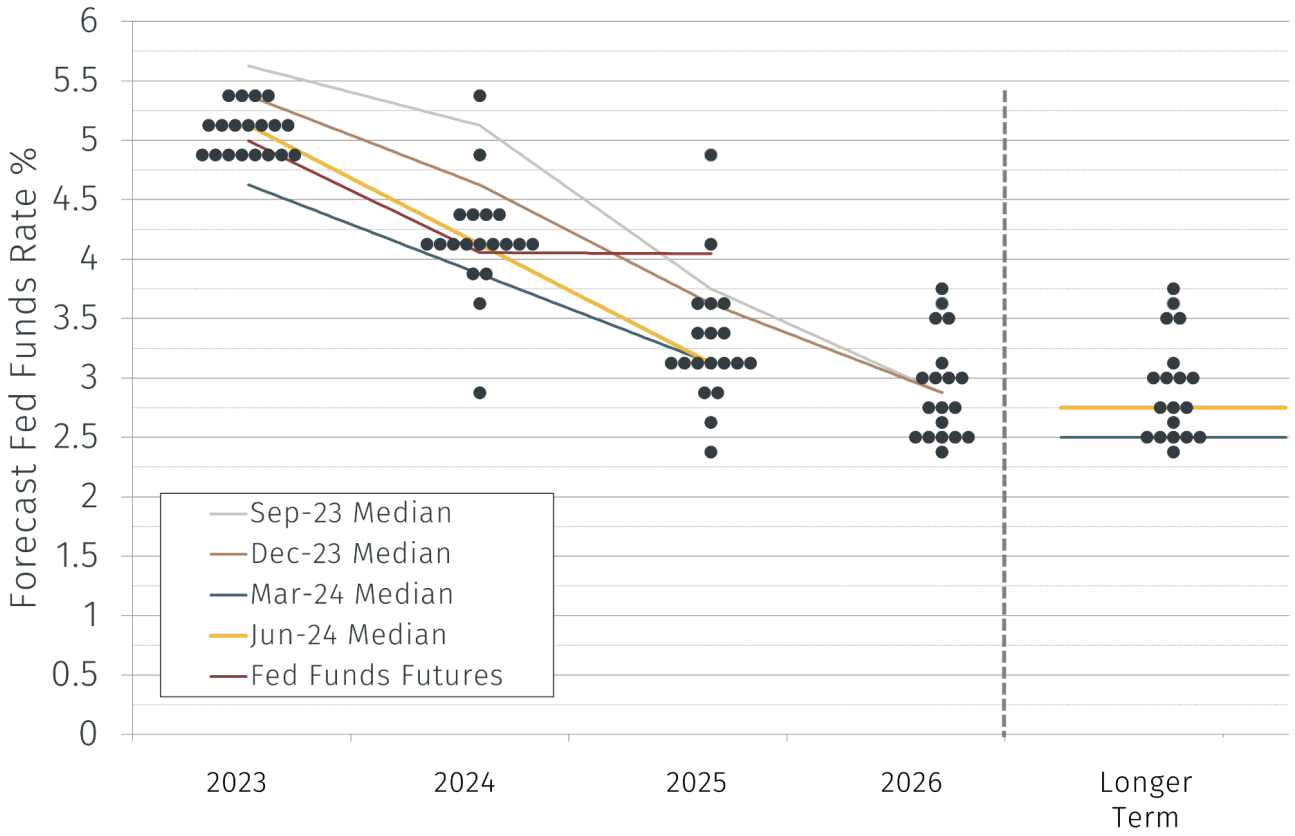
Outline

Background	One Entity	Multi Entities	Relationships	Implications
• Prevalence of data or even big data • Urgency to make fast and good decisions • Limited cognitive resource and attention	• Basic summary statistics • Bias in observability of estimates and forecasts • Bias in observed estimates and forecasts (e.g., fat tails)	• Dimension reduction • Market structure analysis with structured data • Market structure analysis with unstructured data	• Variable selection in one linear regression • Variable selection in multiple linear regressions	• Optimal aggregation rules • Impact of reviews on product sales • Impact of AIGS on review activities and content consumption

The first class

The second class

Prevalence of data or even big data in our lives



Prevalence of data or even big data in our lives

Prevalence of data or even big data in our lives

Prevalence of data or even big data in our lives

- Traditional data newly made available because of advancement in technology
- E.g., shopping trajectories

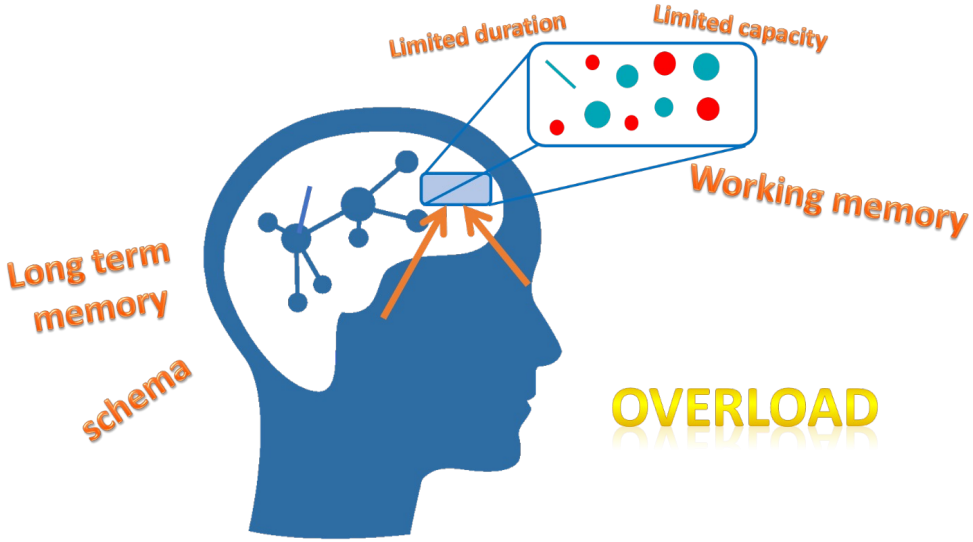
- New business contexts and interactions
- E.g., live streaming

- New data generated by AI trained using human data;
- New business contexts and interactions facilitated by AIGC

- New data by AI in the physical world
- E.g., self-driving cars, robots, drones

https://www.youtube.com/watch?v=58BcWGojWHU#t=34m5s

Limited cognitive resources and attention



Stereotyped bias applies a widely held but fixed, oversimplified, and generalized image or idea of a particular type of person or thing.

Similarity bias leads people to connect with others who share similar interests, experiences, and backgrounds.

Anchoring bias causes us to rely too heavily on the first piece of information about a person or topic.

Availability bias distorts our view based on the most immediate information that readily comes to mind.

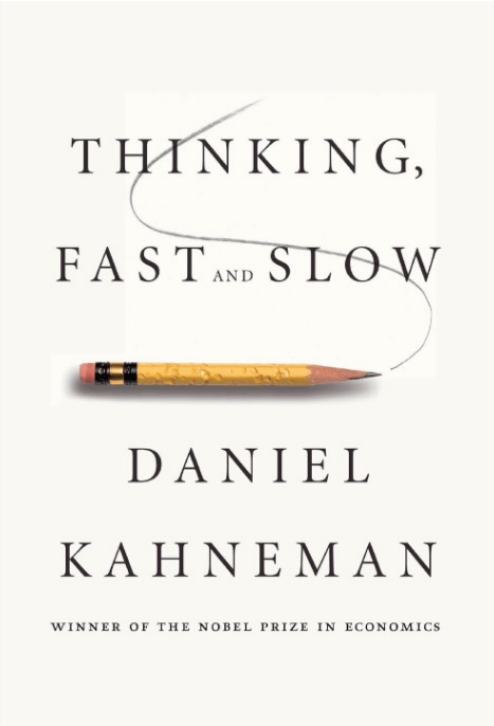
The **bandwagon effect** is groupthink or herd mentality, where people adopt behaviors and views of a dominant group.

Confirmation bias tends to interpret new evidence to validate one's existing beliefs.

Fundamental attribution error tends to assume someone's actions are because of their character rather than circumstances.

Horns and halo effect biases perceptions of someone because of opinions of that person's other traits.

Limited cognitive resources and attention



System 1 and 2 thinking

System 1 "Fast"

DEFINING CHARACTERISTICS
Unconscious
Effortless
Automatic

WITHOUT Self-Awareness or Control

"What You See Is All There Is"

ROLE
Assess the Situation
Deliver Updates

System 2 "Slow"

DEFINING CHARACTERISTICS
Deliberate and Conscious
Effortful
Controlled Mental Process

WITH Self-Awareness or Control

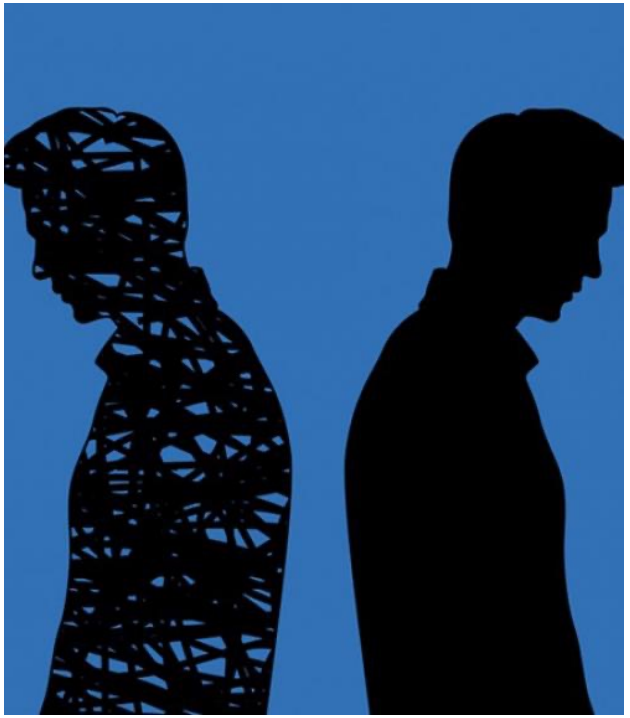
Logical and Skeptical

ROLE
Seeks New Information
Makes Decisions

Limited cognitive resources and attention

Attention/Eyeball Economy

Urgency to make fast and ideally good decisions



Background	One Entity	Multi Entities	Relationships	Implications
• Prevalence of data or even big data • Urgency to make fast and good decisions • Limited cognitive resource and attention	• Basic summary statistics • Bias in observability of estimates and forecasts • Bias in observed estimates and forecasts (e.g., fat tails)	• Dimension reduction • Market structure analysis with structured data • Market structure analysis with unstructured data	• Variable selection in one linear regression • Variable selection in multiple linear regressions	• Optimal aggregation rules • Impact of reviews on product sales • Impact of AIGS on review activities and content consumption

Examples

Customer reviews

4.7 out of 5

146 global ratings

5 star	87%
4 star	6%
3 star	2%
2 star	2%
1 star	3%

Angela

Owl Diaries

Reviewed in the United States on February 6, 2024

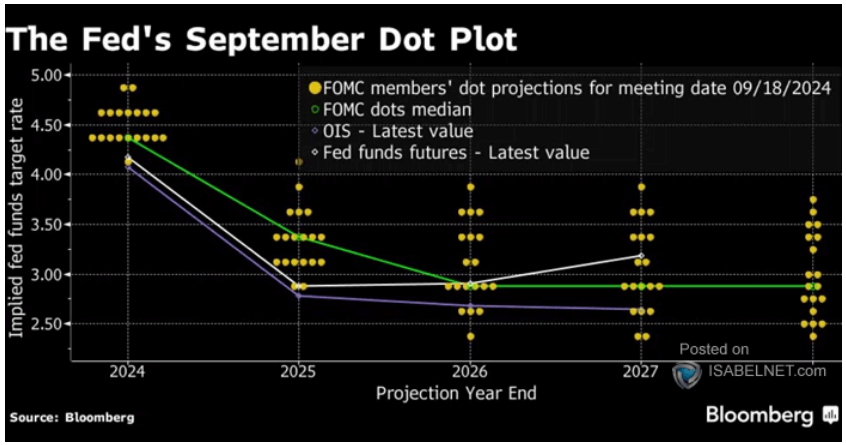
Verified Purchase

My daughter loves these books. It has a series that's she's been reading since kindergarten on her own. I think it's a great way to start your child off with chapter books. I just love the funny names of the characters and the words they use. My daughter is in the first grade and she wants to read these over and over.

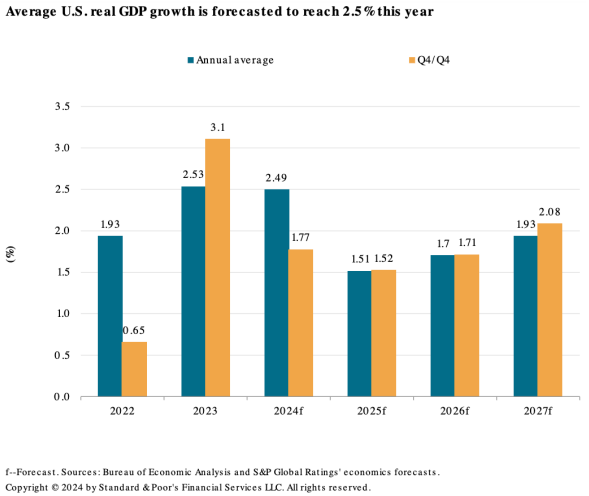
One person found this helpful

Helpful | Report

https://www.amazon.com/Eva-President-Branches-Book-Diaries/dp/1338880276/



https://www.isabelnet.com/implied-fed-funds-target-rate/



https://www.spglobal.com/ratings/en/research/articles/240326-economic-outlook-u-s-q2-2024-heading-for-an-encore-13048486

The Hong Kong Polytechnic University 香港理工大學

Basic summary statistics

- Number of observations
- Average of the observations
- Standard deviation of the observations
- *Correlation coefficient if observations of more than one variables*

Order statistics, oftentimes five-number summary

- Minimum of the observations
- The first quartile of the observations
- Median of the observations
- The third quartile of the observations
- Maximum of the observations

Example:

(14 73 28 42 9 85 31 19 67 51 22 98 11 46 75 39 88 25 59 49 13 81 36 68 55 29 92 17 63 79)

Summary Statistics

Statistic	Value
Number of Observations	30
Average	48.2
Standard Deviation	25.1
Minimum	9
Maximum	98
Median	46.5

Not all estimates/ratings are observable

MARKETING SCIENCE
Vol. 31, No. 3, May–June 2012, pp. 372–386
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.1110.0662

Online Product Opinions: Incidence, Evaluation, and Evolution

Wendy W. Moe
Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20472, wmoe@rhsmith.umd.edu
David A. Schweidel
Wisconsin School of Business, University of Wisconsin–Madison, Madison, Wisconsin 53706, dschweidel@bus.wisc.edu

$$\Pr(z_{ijk}=1 \mid \text{experience}_j=1) = \Phi\left(\beta_{i0}+\beta_{i,1:N}X_{j,k(i)}+\beta_{i,N+1}GT_{j,k(i)}+\delta_1 V_{ij}+\delta_2 V_{ij}^2\right).$$

Figure 5 Role of Postexperience Evaluation in Rating Incidence

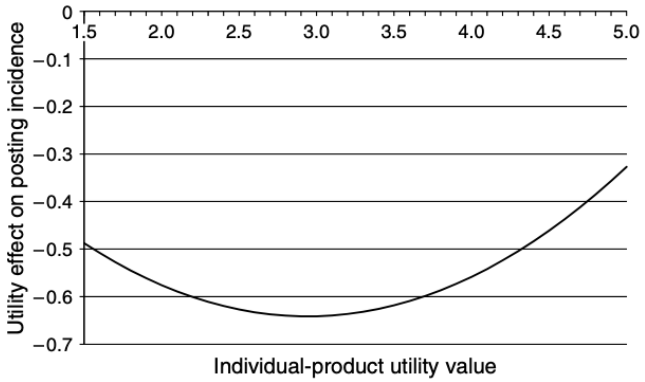
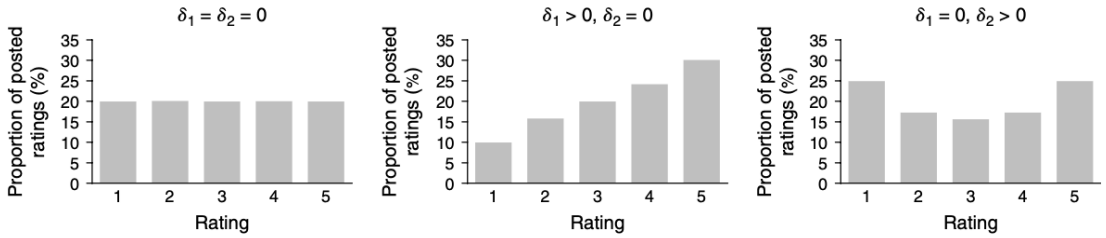


Figure 2 Illustrative Distributions of Posted Ratings



THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Estimates/ratings are oftentimes biased (1)

Reviewing Experts' Restraint from Extremes and Its Impact on Service Providers

PETER NGUYEN
XIN (SHANE) WANG
XI LI
JUNE COTTE

This research investigates reviewing experts on online review platforms. The main hypothesis is that greater expertise in generating reviews leads to greater restraint from extreme summary evaluations. The authors argue that greater experience generating reviews facilitates processing and elaboration and enhances the number of attributes implicitly considered in evaluations, which reduces the likelihood of assigning extreme summary ratings. This restraint-of-expertise hypothesis is tested across different review platforms (TripAdvisor, Qunar, and Yelp), shown for both assigned ratings and review text sentiment, and demonstrated both between (experts vs. novices) and within reviewers (expert vs. pre-expert). Two experiments replicate the main effect and provide support for the attribute-based explanation. Field studies demonstrate two major consequences of the restraint-of-expertise effect. (i) Reviewing experts (vs. novices), as a whole, have less impact on the aggregate valence metric, which is known to affect page-rank and consumer consideration. (ii) Experts systematically benefit and harm service providers with their ratings. For service providers that generally provide mediocre (excellent) experiences, reviewing experts assign significantly higher (lower) ratings than novices. This research provides important caveats to the existing marketing practice of service providers incentivizing reviewing experts and provides strategic implications for how platforms should adopt rating scales and aggregate ratings.

Keywords: online word-of-mouth, expertise, user rating average, platform strategy, text analysis, sentiment analysis

Peter Nguyen (pnguyen@miamioh.edu) is an assistant professor of marketing at Miami University, Oxford, OH 45056, USA. Xin (Shane) Wang (xwang@ivey.ca) is an associate professor of marketing and statistics, MBA '80 Faculty Fellow, at Ivey Business School, Western University, London, ON N6G 0N1, Canada. Xi Li (xili44@cityu.edu.hk) is an assistant professor of marketing at the City University of Hong Kong, Kowloon Tong, Hong Kong, China. June Cotte (jcotte@ivey.ca) is the professor of marketing, Scott & Melissa Beattie Professorship in Marketing, at Ivey Business School, Western University, London, ON N6G 0N1, Canada. Please address correspondence to Peter Nguyen. This article is based on the first author's dissertation. The authors thank seminar participants at Duke University, Queen's University, and the University of Connecticut, as well as participants in the AMA-CBSIG conference in Bern, for their valuable feedback on earlier versions of this article. This work was supported by the Social Sciences and Humanities Research Council (435-2019-0597). Supplementary materials are included in the web appendix accompanying the online version of this article.

Editor: J. Jeffrey Inman

Associate Editor: Andrew T. Stephen

Advance Access publication 15 July 2020

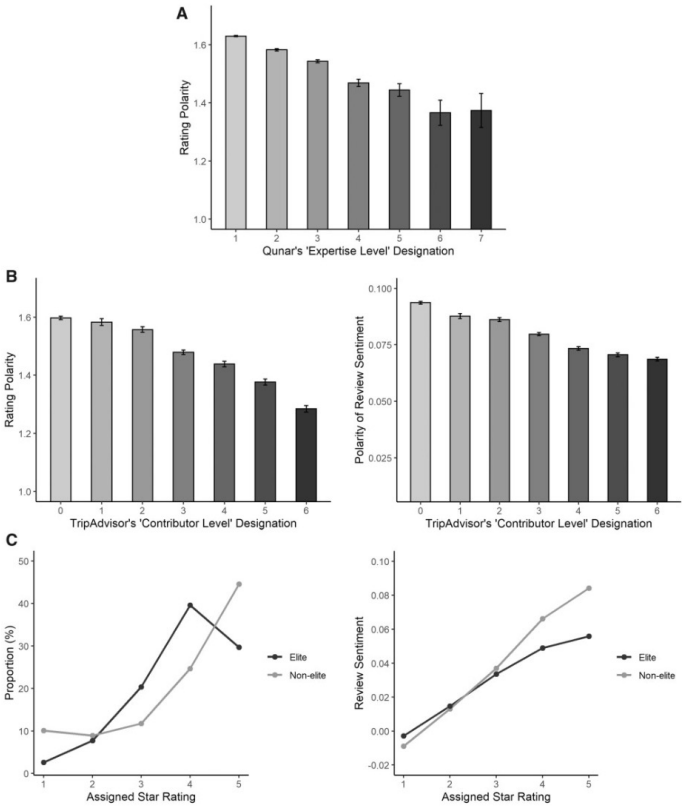
Consumers rely on the opinions and recommendations of others. Many of these recommendations have come from expert professionals (e.g., sommeliers, movie critics). Over the past couple of decades, the world has seen the rise of online reviews, where consumers not only rely on other consumers' experiences but also share their own. Online review platforms now recognize their top users as reviewing "experts." For example, Yelp has its "Elite" status, TripAdvisor has its "Contributor Level," Google has its "Local Guide" badges, and Amazon has its "Amazon Vine Program." Given that consumers are increasingly both sharing and consuming reviews, understanding the nature of reviewing experts has become an important topic in consumer research.

The study of online reviewing experts is particularly important for *service providers*, such as hotels and restaurants. Many businesses incentivize, by quite literally

DOI: 10.1093/jcr/ucaa037

FIGURE 1

POLARITY OF EVALUATIONS AS A FUNCTION OF PLATFORM-DEFINED REVIEWING EXPERTISE. (A) QUNAR (STUDY 1). (B) TRIPADVISOR (STUDY 3). (C) YELP (STUDY 4) ON



The Hong Kong Polytechnic University
香港理工大學

Estimates/ratings are oftentimes biased (2)

Fat Tails in Human Judgment:

Empirical Evidence and Implications for

the Aggregation of Estimates and Forecasts

Miguel Sousa Lobo* Dai Yao†

September 18, 2024

Abstract

How frequent are large disagreements in human judgment? The substantial literature relating to expert assessments of real-valued quantities and their aggregation almost universally assumes that judgment errors follow a jointly normal distribution. We investigate this question empirically using 73 data sets from 4 different sources that include over 169,000 estimates and forecasts. We find incontrovertible evidence for excess kurtosis, that is, of fat tails. Despite the diversity of the analyzed data, varying in how much uncertainty there is in the quantity being assessed and varying in the level of expertise and of sophistication of those making the assessments, we find surprising consistency in the frequency with which an individual judgment is in large disagreement with the consensus. Fitting a generalized normal distribution to the data, we find most estimates for the shape parameter to be between 0.9 and 1.6 (where 1 is the double-exponential distribution, and 2 is the normal distribution). This has important implications, in particular for the aggregation of expert estimates and forecasts and for the construction of confidence intervals. We describe optimal Bayesian aggregation with fat tails, and propose a simple average-median average heuristic (AMA) that performs well for the range of empirically observed distributions.

*miguel.lobo@insead.edu, Associate Professor of Decision Sciences at INSEAD.

†dai@yaod.ai, Associate Professor at The Hong Kong Polytechnic University.

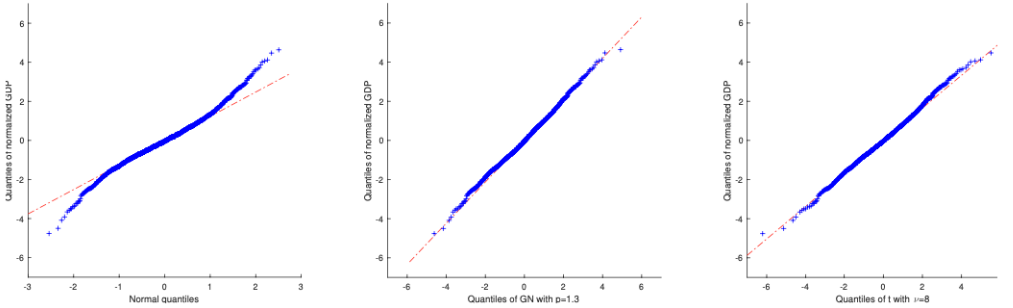


Figure 1: Quantile-quantile plots of USA gross domestic product growth forecasts by economists (after each quarter's forecasts were normalized to have zero mean and unit variance), against a normal distribution, against a generalized normal distribution with shape parameter $p = 1.3$, and against a Student's t-distribution with the number of degrees of freedom $\nu = 8$.

$$f(x) = \frac{1}{2s\Gamma\left(1+1/p\right)} \exp\left\{-\left|\frac{x-u}{s}\right|^p\right\}.$$

Data set	N	n	Skew.	Kurt.	p_{kurt}	k	p_{Bayes}	p_{low}	p_{high}
UNHI	2017	18	-0.91	5.3	1.09	3	0.95	0.86	1.04
UNLO	2022	18	-0.72	4.1	1.36	3	1.52	1.41	1.64
FXHI	2003	18	0.38	3.5	1.64	39	1.69	1.54	1.85
FXLO	2025	18	-0.27	4.1	1.36	39	1.35	1.22	1.49
SCALE	3618	34	0.42	3.9	1.43	71	1.66	1.55	1.77

Table 6: Estimates of the GN shape parameter p for trivia questions.

				Homogeneous judges				Heterogeneous judges			
Data set	N	m	n	k	p_{Bayes}	p_{low}	p_{high}	k	p_{Bayes}	p_{low}	p_{high}
GDP	3682	125	70	141	1.23	1.15	1.31	395	1.81	1.65	1.96
GDPa	2158	99	42	85	1.12	1.03	1.21	287	1.67	1.48	1.89
NFARM	8783	127	170	341	1.17	1.12	1.22	599	1.64	1.55	1.74

Table 7: Estimates of the GN shape parameter p for economic forecasts, controlling for heterogeneity across forecasters.

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

How about summaries of data in unstructured form

Gheen

★★★★★ **Beauty -- In Every Way**

Reviewed in the United States on November 18, 2024

Size: 42 Inch | Style: TV Only | Verified Purchase

Buying a TV for any room has few considerations - but all important. They are the appropriate size of the TV to its service space, the quality of color production, the quality of sound, ease and efficiency of controls and the ability to set install and setup the TV without having to employ and engineer. This LG Class OLED exudes excellence in all areas at a reasonable cost. Wall mount installation and operational setup took 15 minutes each. The color quality is stunningly beautiful, and the program walks you through making fine adjustments to meet your color tastes. The sound quality was excellent, and so beautiful that a plan to add a sound bar were scrapped. The TV controller is easy to manage and instantly responsive. At age 74 and having owned a plethora of TVs across the years, the LG is the best and most beautiful buy.

11 people found this helpful

Helpful | Report

Tony S

★★★★★ **Ultimate Gaming Experience – Stunning Visuals and Smart Features!**

Reviewed in the United States on October 22, 2024

Size: 48 Inch | Style: TV Only | Verified Purchase

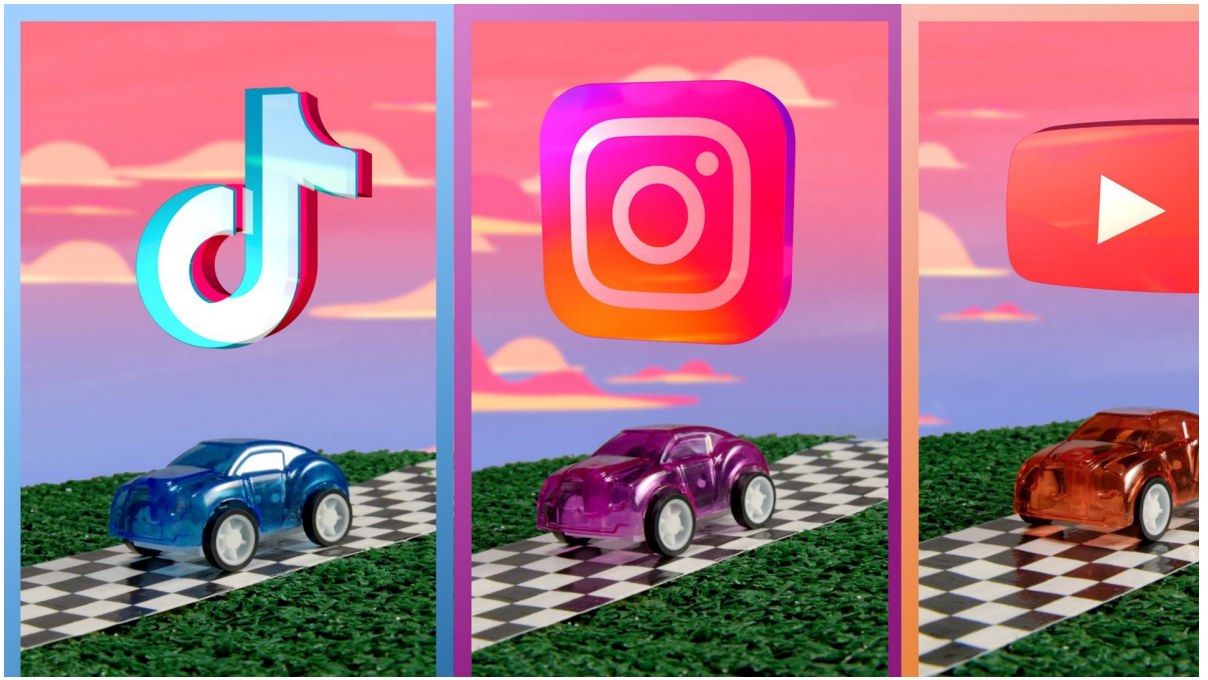
The LG 48-Inch OLED evo C4 is hands down the best TV I've used for gaming. The picture quality is incredible, with deep blacks, vibrant colors, and smooth motion that make every game look stunning. The 4K processor ensures everything runs seamlessly, and the OLED technology eliminates motion blur, making it perfect for fast-paced games.

What really impressed me is the low input lag and the built-in gaming features like NVIDIA G-Sync and FreeSync support. It provides a responsive experience, giving me an edge in competitive play. The size is also ideal for immersive gaming without being overwhelming.

The AI-powered Magic Remote and Alexa integration make it incredibly convenient to control both the TV and my smart home devices. Switching between gaming and streaming apps is quick, and the user interface is intuitive.

For the price, this TV offers exceptional value. Whether you're a casual or serious gamer, the LG OLED evo C4 delivers a next-level experience. Highly recommend it!

Texts



Short and long videos

Amazon ’s solution: AI-Generated Summaries (AIGS)

Customer reviews

4.5 out of 5

261 global ratings

5 star 78%

4 star 7%

3 star 6%

2 star 3%

1 star 7%

How customer reviews and ratings work

Review this product

Share your thoughts with other customers

Write a customer review

Customers say

Customers like the picture quality, ease of setup, and sound quality of the television. They mention it's divine, vibrant, and 4K content looks stunning. Some appreciate the value for money, saying it gives great bang for the buck. Customers also appreciate the build quality and speed.

AI-generated from the text of customer reviews

Select to learn more

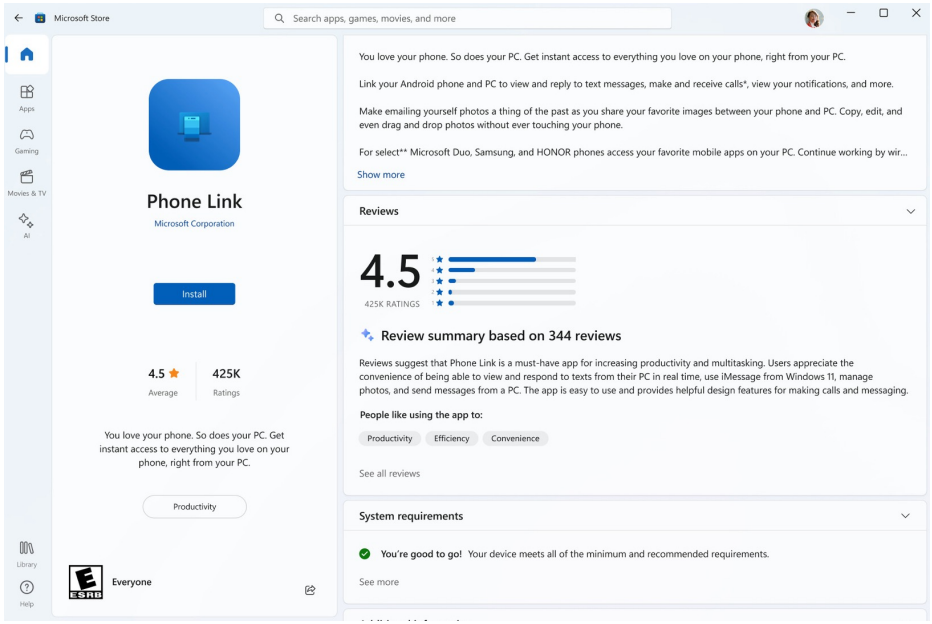
✓ Picture quality | ✓ Ease of setup | ✓ Sound quality | ✓ Value for money | ✓ Brightness | ✓ Build quality | ✓ Speed

Reviews with images

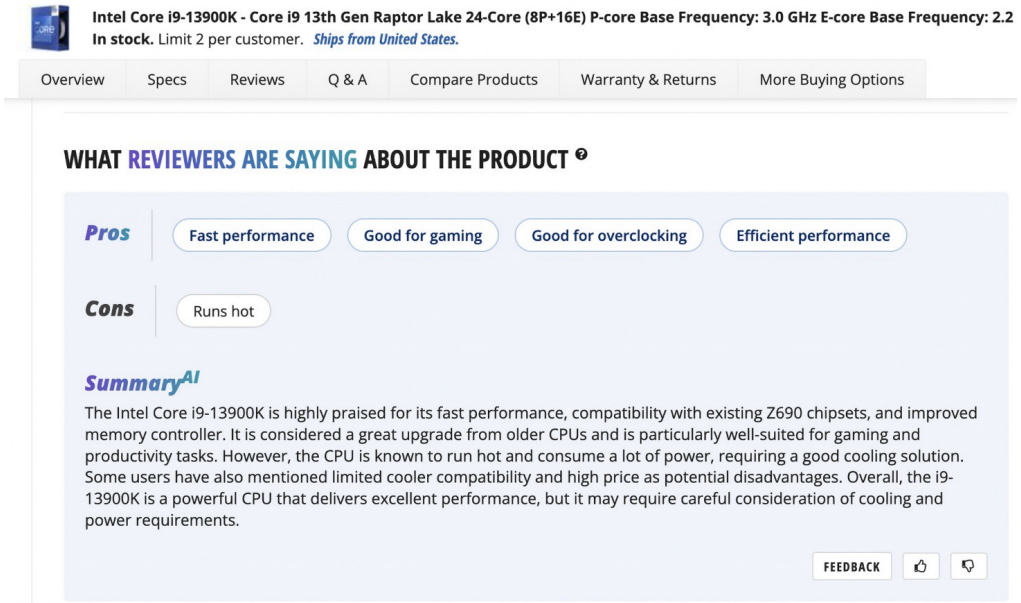
See all photos ›

https://www.amazon.com/LG-55-Inch-Processor-AI-Powered-OLED55C4PUA/dp/B0CVRDK4P6/

Microsoft Store, Newegg are also embracing the trend



https://www.theverge.com/2023/5/23/23732821/microsoft-store-bing-ai-hub-apps-build



https://www.theverge.com/2023/8/9/23825805/newegg-chatgpt-review-summaries-ai

Video platforms are also embracing it

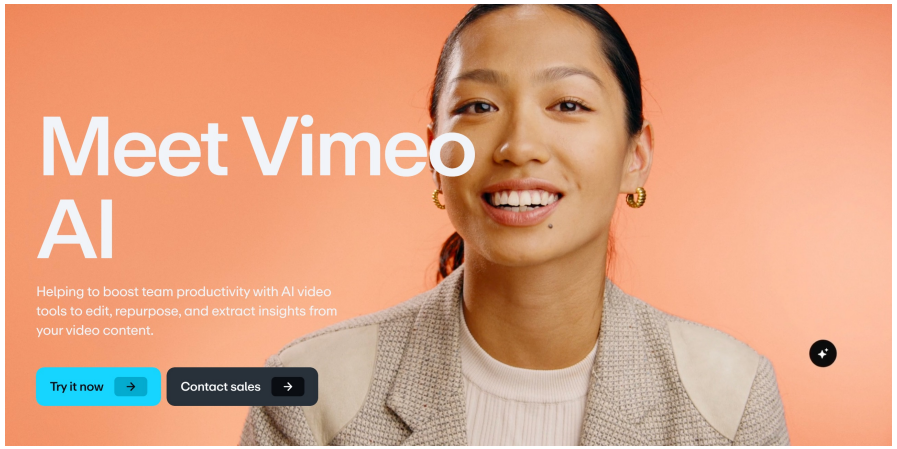
https://www.ithome.com/0/728/123.htm

A lot of tools available within/beyond main platforms

YouTube Video Summarizer

Get YouTube transcript and use AI to summarize YouTube videos in one click for free online with NoteGPT's YouTube summary tool.

Want to summarize local Video/Audio files? Go to Workspace to summarize



YouTube Summary with ChatGPT Online and Free

Summarize YouTube videos in seconds

Paste a YouTube video url here

Summarize

Vimeo Summarizer

Summarize Vimeo videos in seconds with ScreenApp's AI

Paste a URL (YouTube, Vimeo, Facebook, TikTol

Import from URL

Loved by over 1 million users

Summary of “ live” content: Zoom

https://www.zoom.com/en/blog/zoom-ai-companion-getting-started-guide/

zoom

NEW 使用AI Companion的会议摘要现已支持其他语言的预览。

以下项目的会议摘要： PolyU EM Brownbag Series #24 (11/25/2024)

快速回顾

Angela讨论了影响者在发布赞助内容时对他们声誉的影响，质疑了对影响者声誉的衡量方法，以及研究中未测

后续步骤

-
- 。

摘要

影响者声誉的衡量

Angela讨论了两篇论文，重点关注了影响者在发布赞助内容时对他们声誉的影响。她指出，虽然赞助内容可能

Video Content Embedding and Measurement

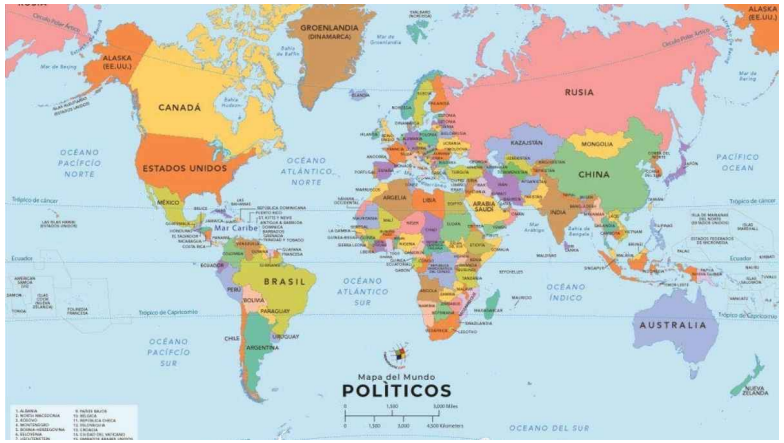
Angela and Yao discussed the embedding and measurement of video content. They agreed that video conte

网红与产品匹配度

Angela讨论了她关于短视频和直播带货的研究，重点是网红和产品之间的匹配度对销售的影响。她概述了研究

Background	One Entity	Multi Entities	Relationships	Implications
• Prevalence of data or even big data • Urgency to make fast and good decisions • Limited cognitive resource and attention	• Basic summary statistics • Bias in observability of estimates and forecasts • Bias in observed estimates and forecasts (e.g., fat tails)	• Dimension reduction • Market structure analysis with structured data • Market structure analysis with unstructured data	• Variable selection in one linear regression • Variable selection in multiple linear regressions	• Optimal aggregation rules • Impact of reviews on product sales • Impact of AIGS on review activities and content consumption

Examples



LANEIGE AMUSE Cosmetics TONY MOLY erborian KOREAN SKIN THERAPY PARIS · SEOUL AMOREPACIFIC MISSHA

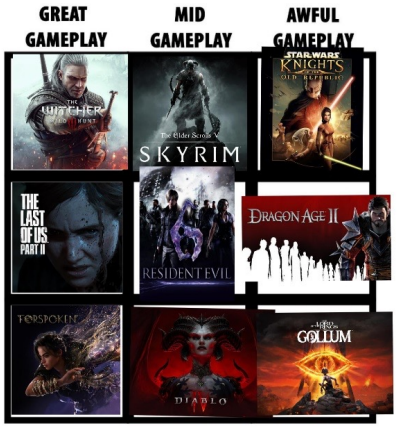
Dr.G Kaja ETUDE HOUSE peripera HERA JOAH

HOLIKA HOLIKA 3CE STYLENANDA rom&nd OFFICIAL CLIO innisfree

WELL WRITTEN

MID WRITING

TERRIBLY WRITTEN



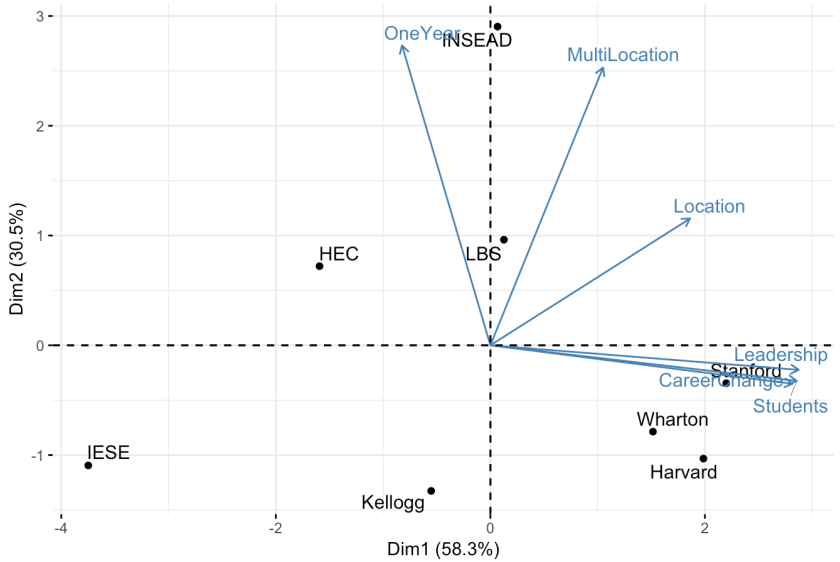
THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

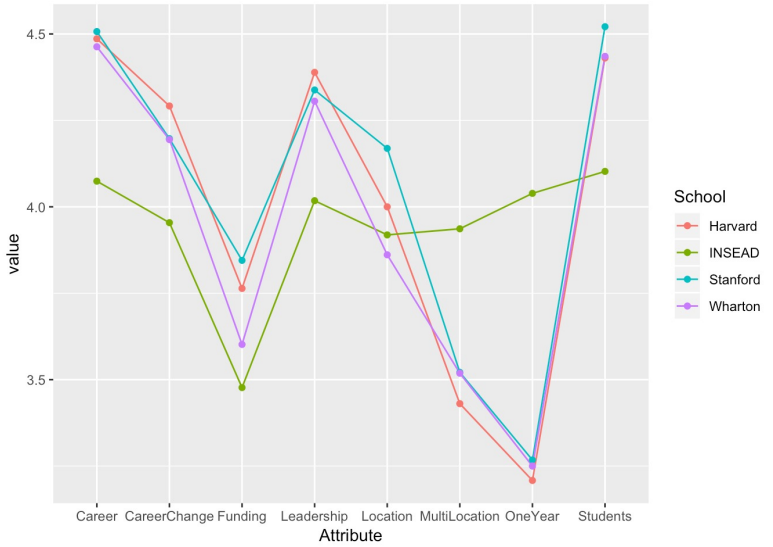
Aggregation of information about multiple entities

percep.raw.data

SchoolRated	Q17_1	Q17_2	Q17_3	Q17_4	Q17_5	Q17_6	Q17_7	Q17_8	Q17_9	Q17_10	Q17_11	Q17_12	Q17_13
8	3	5	5	4	4	4	2	2	3	4	5	3	5
7	4	5	5	4	4	2	2	2	3	4	4	3	5
1	2	5	5	5	5	5	4	2	5	5	5	3	5
4	5	5	5	5	3	5	3	5	5	5	4	5	5
3	3	3	3	3	3	3	3	3	3	3	3	3	3
4	3	3	3	3	3	3	3	3	3	3	3	3	3
6	3	5	5	5	5	5	5	3	3	5	4	3	4
8	3	5	5	5	5	3	5	3	3	5	4	3	4
4	3	4	4	5	3	5	3	3	4	4	3	4	4

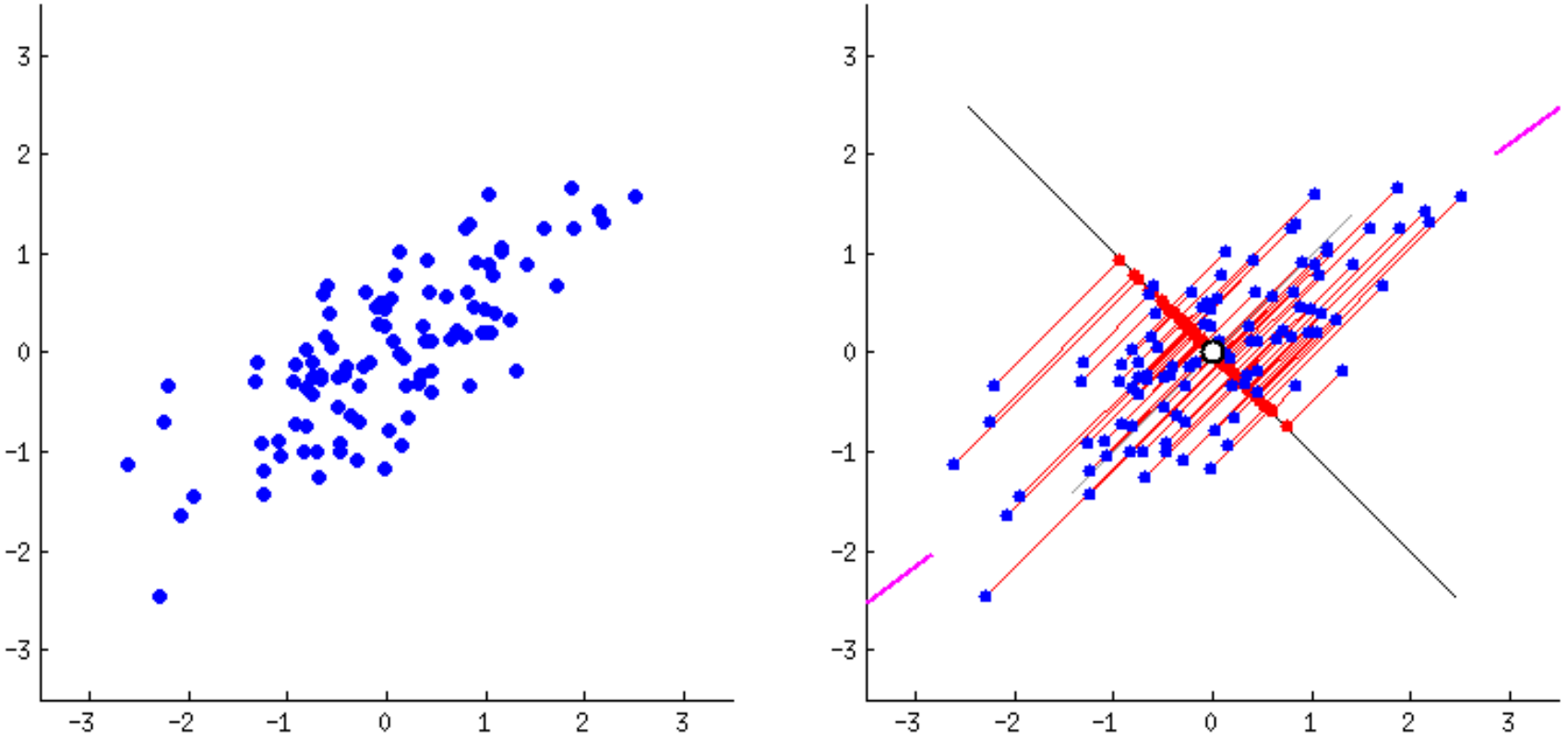
Attribute	Harvard	HEC	IESE	INSEAD	Kellogg	LBS	Stanford	Wharton
OneYear	3.208333	3.605263	3.40	4.038869	3.354839	3.679487	3.267606	3.250000
Students	4.430556	3.745614	3.58	4.102473	4.064516	4.102564	4.521127	4.435185
Leadership	4.388889	3.807018	3.42	4.017668	4.010753	3.974359	4.338028	4.305556
CareerChange	4.291667	3.605263	3.39	3.954064	3.978495	3.858974	4.197183	4.194444
Location	4.000000	3.973684	3.55	3.918728	3.709677	4.269231	4.169014	3.861111
MultiLocation	3.430556	3.508772	3.23	3.936396	3.258065	3.461538	3.521127	3.518519





THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Principal component analysis



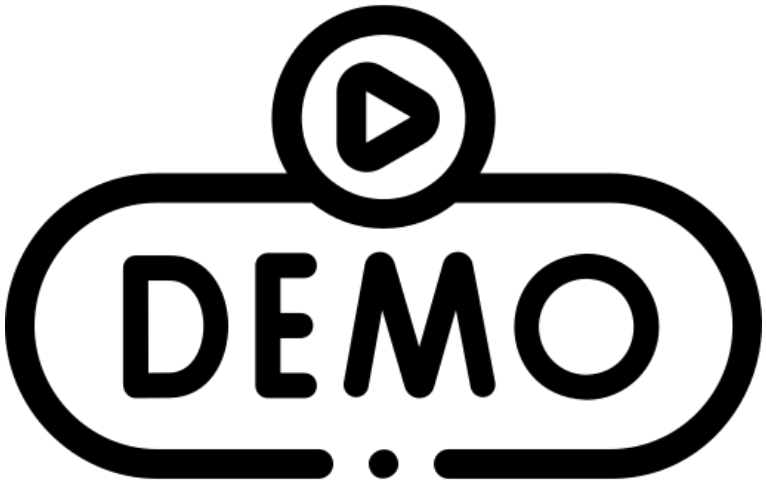
which is very similar to factor analysis, though the two are fundamentally different.
https://www.theanalysisfactor.com/the-fundamental-difference-between-principal-component-analysis-and-factor-analysis/

Principal component analysis: An illustration

Who ’s #1? INSEAD, Harvard, Wharton, LBS?

Who's who?			***Who's who?***		
"The Business School for the World"	"Impact the way the world does business"	"Global focus and learning experience throughout programs"	INSEAD The Business School for the World®	London Business School	The University of Melbourne | Melbourne Business School
"We educate leaders who make a difference in the world"	"Our ultimate goal: to create a global world that has meaning for us all"	"The Global Business School"	Harvard Business School	ESSEC Business School	HULT International Business School

Principal component analysis: An illustration



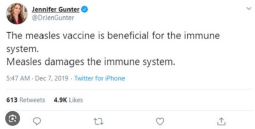
Two main approaches of market structure analysis

Brands as collections of *disaggregate* attributes, using **principal component analysis**

Basic data structure

Attribute	Harvard	HEC	IESE	INSEAD	Kellogg	LBS	Stanford	Wharton
OneYear	3.208333	3.605263	3.40	4.038869	3.354839	3.679487	3.267606	3.250000
Students	4.430556	3.745614	3.58	4.102473	4.064516	4.102564	4.521127	4.435185
Leadership	4.388889	3.807018	3.42	4.017668	4.010753	3.974359	4.338028	4.305556
CareerChange	4.291667	3.605263	3.39	3.954064	3.978495	3.858974	4.197183	4.194444
Location	4.000000	3.973684	3.55	3.918728	3.709677	4.269231	4.169014	3.861111
MultiLocation	3.430556	3.508772	3.23	3.936396	3.258065	3.461538	3.521127	3.518519

John, Loken, Kim, Monga (JMR 2006)



Culotta and Cutler (MKSC 2016)

Liu, Dzyabura, Mizik (MKSC 2020)
Dzyabura, Peres (JM 2021)

Li, Castelo, Katona, Sarvary (MKSC 2024)

Brands as *aggregate* objects, using **clustering or multidimensional scaling** based on **similarity** matrix

Kim, Albuquerque, Bronnenberg (JMR 2011)
Ringel, Skiera (MKSC 2016)

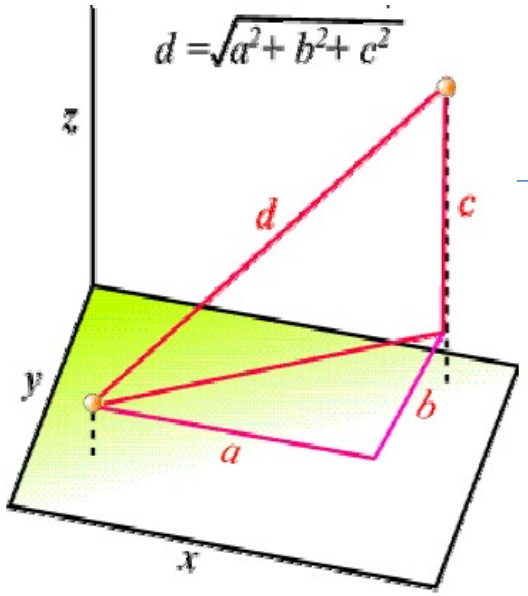
Gabel, Guhl, Klapper (JMR 2019)
Ruis, Athey, Blei (AAS 2020)

Chintagunta (MKSC 1998)
France and Ghose (MKSC 2016)

Lee and Bradlow (JMR 2011)
Netzer, Feldman, Goldenberg, Fresko (MKSC 2012)
Tirunillai and Tellis (JMR 2014)
Matthe, Ringel, Skiera (MKSC 2022)
Zhang, Kim, Xing (KDD 2015)

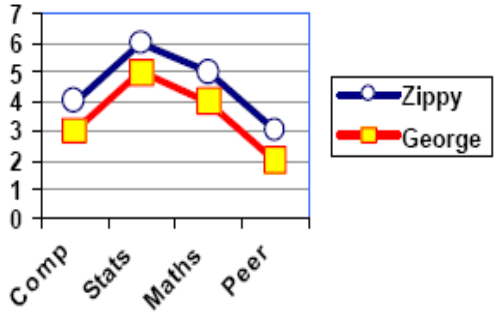
Yang, Zhang, Kannan (JM 2021)

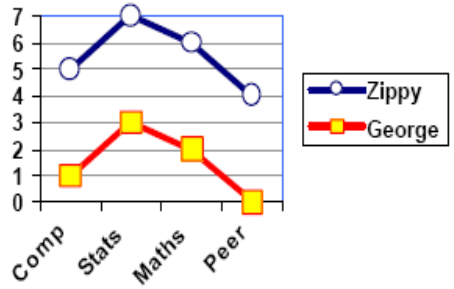
Clustering – Similarity measures



$$\mathbf{D} = \begin{bmatrix} 0 & d_{12} & d_{13} \\ d_{21} & 0 & d_{23} \\ d_{31} & d_{32} & 0 \end{bmatrix}$$

$$\mathbf{r} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix}$$

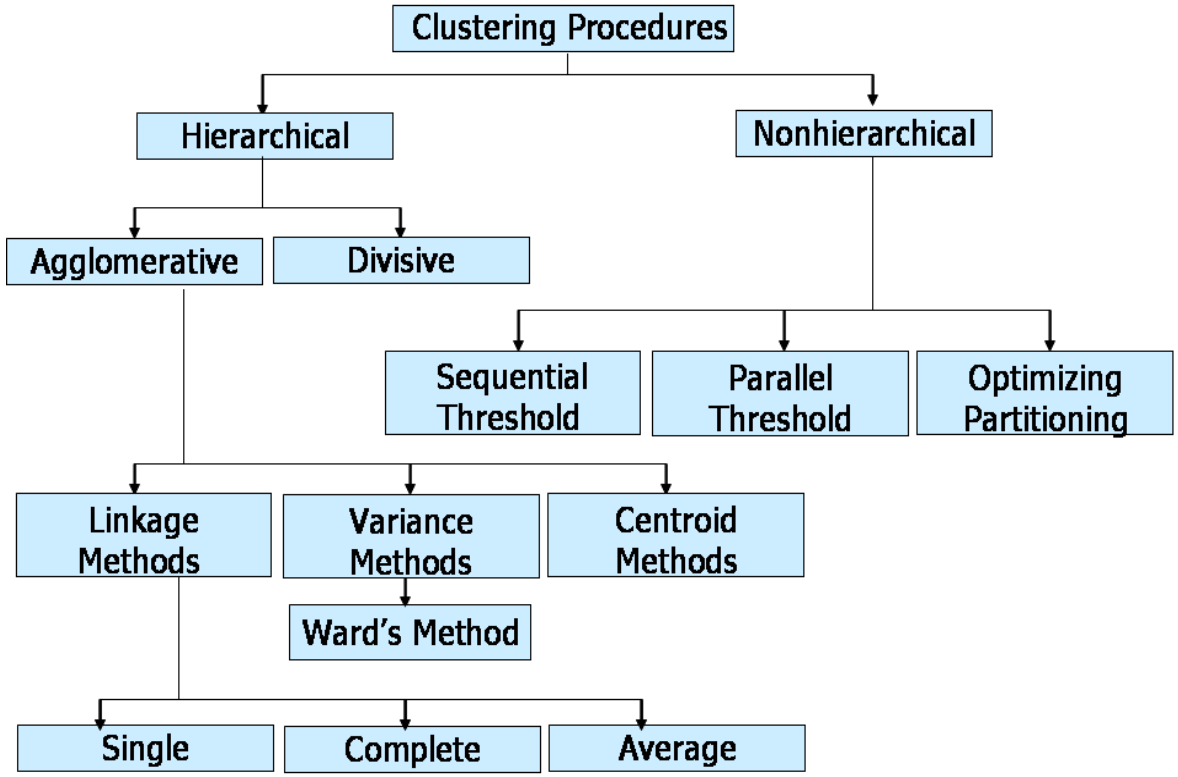




- Euclidian distance, city-block or Manhattan distance, Chebyshev distance
- Variable standardization before calibration

- Correlation is simple to calculate
- The elevation of score might be an issue

Clustering – Methods



Hierarchical clustering: is characterized by the development of a hierarchy or tree-like structure. Hierarchical methods can be agglomerative or divisive.

Nonhierarchical clustering: These methods are frequently referred to as *k*-**means clustering**. These methods include sequential threshold, parallel threshold, and optimizing partitioning.

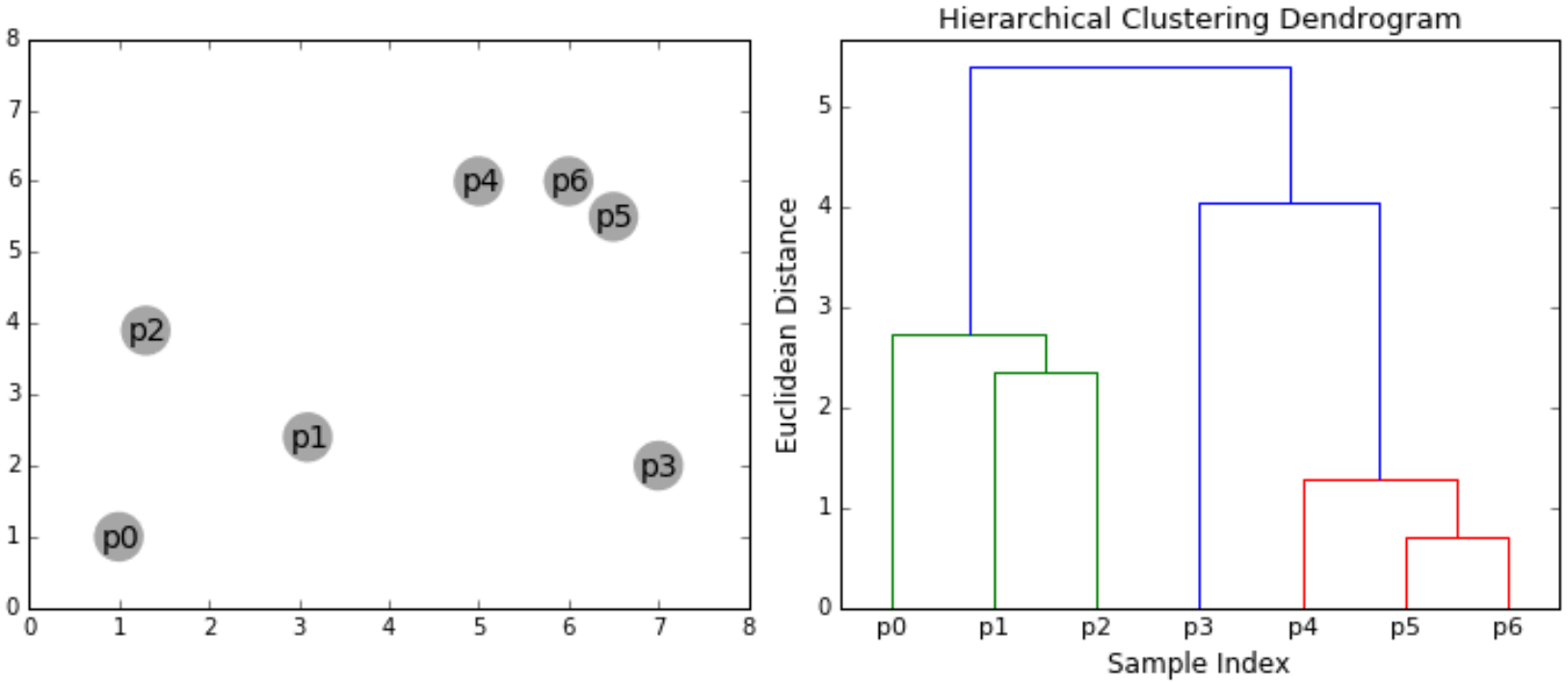
Agglomerative (hierarchical) clustering: starts with each object in a separate cluster. Clusters are formed by grouping objects into bigger and bigger clusters. This process is continued until all objects are members of a single cluster.

Divisive (hierarchical) clustering: starts with all the objects grouped in a single cluster. Clusters are divided or split until each object is in a separate cluster.

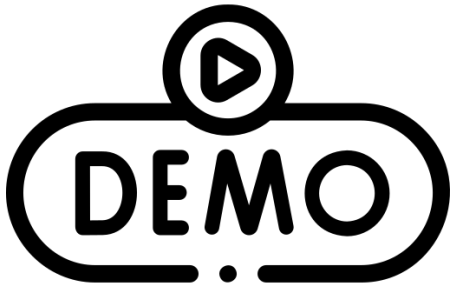
Agglomerative methods are commonly used in marketing research. They consist of

- linkage methods,
- variance methods or error sums of squares, and
- centroid methods.

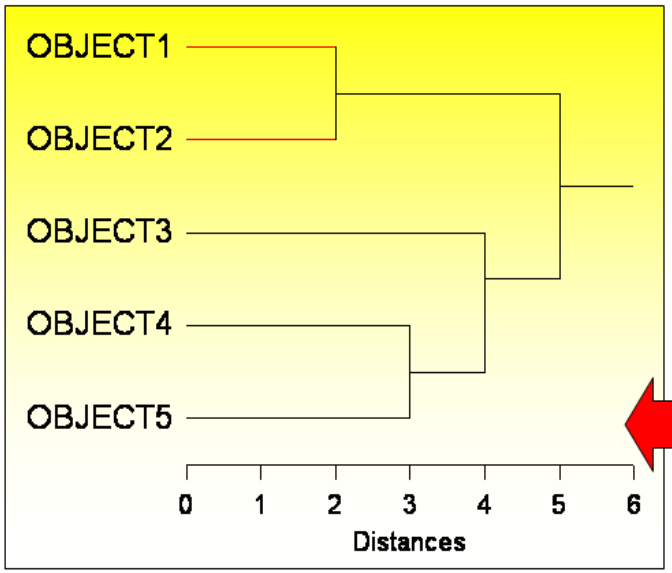
Clustering – hierarchical



Clustering – hierarchical: An illustration



Simple joining (nearest neighbour)

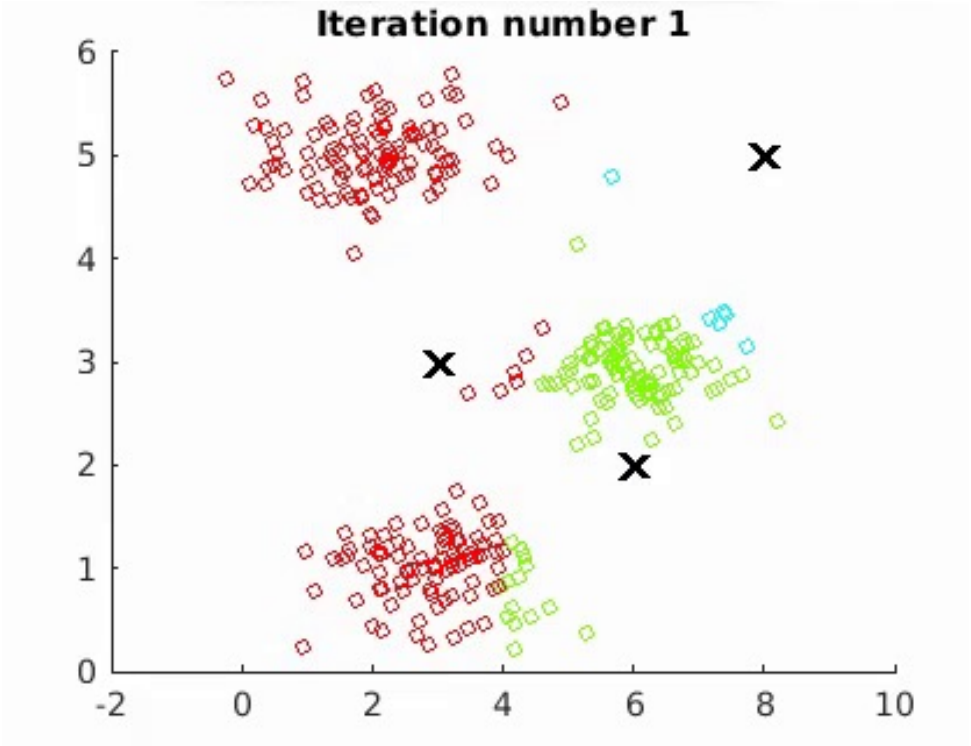


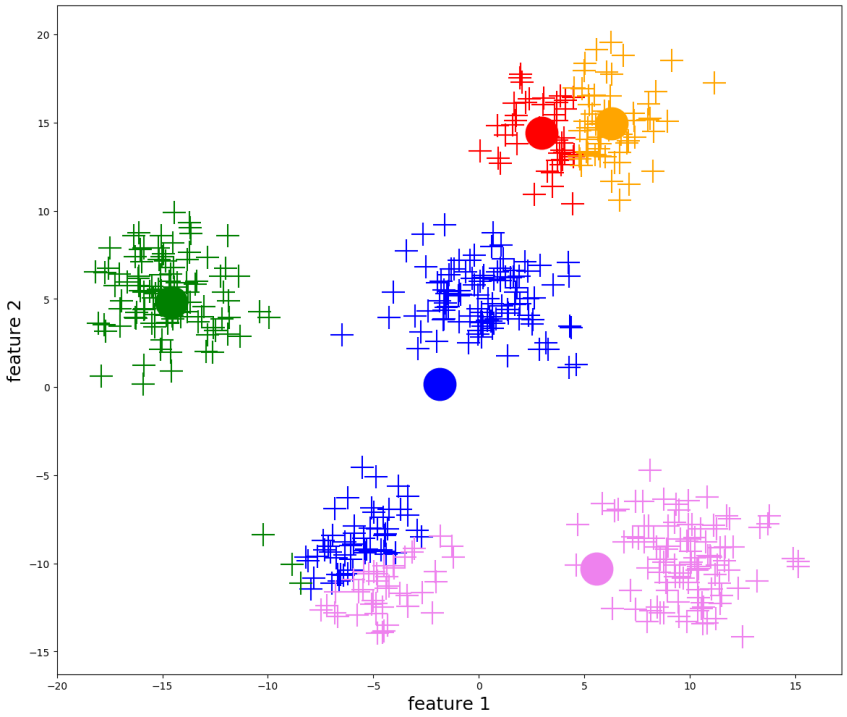
Object	1	2	3	4	5
1					
2	2				
3	6	5			
4	10	9	4		
5	9	8	5	3	

Distance matrix

Distance	Cluster
0	1,2,3,4,5
2	(1, 2), 3, 4, 5
3	(1, 2), 3, (4, 5)
4	(1, 2), (3, 4, 5)
5	(1, 2, 3, 4, 5)

Clustering – non-hierarchical, e.g., kmeans

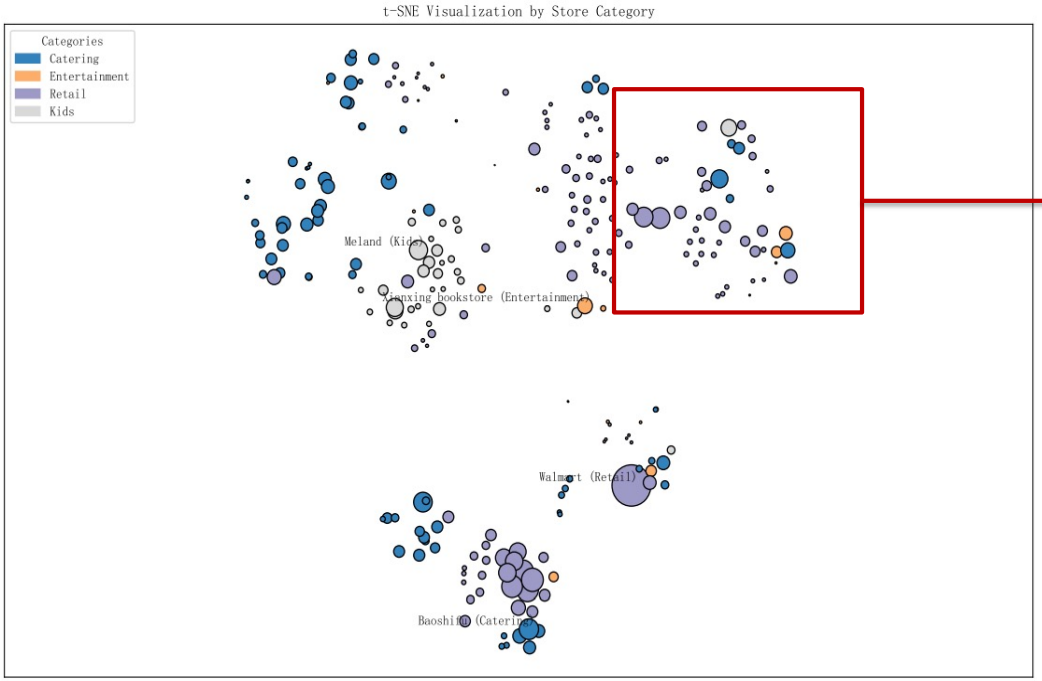




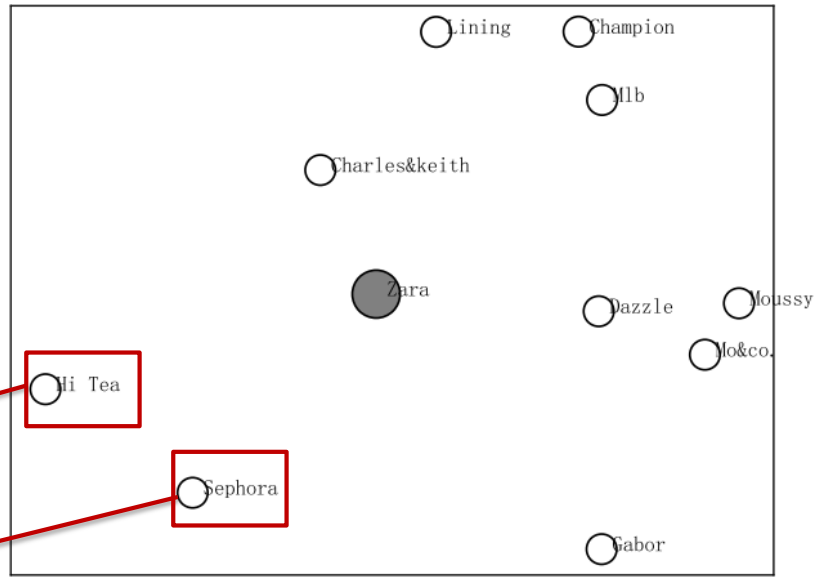
Digitalization of Retailing Spaces and Shopping Trajectories

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

t-SNE visualization of the shopping mall market structure



Stores from different categories are also positioned closely

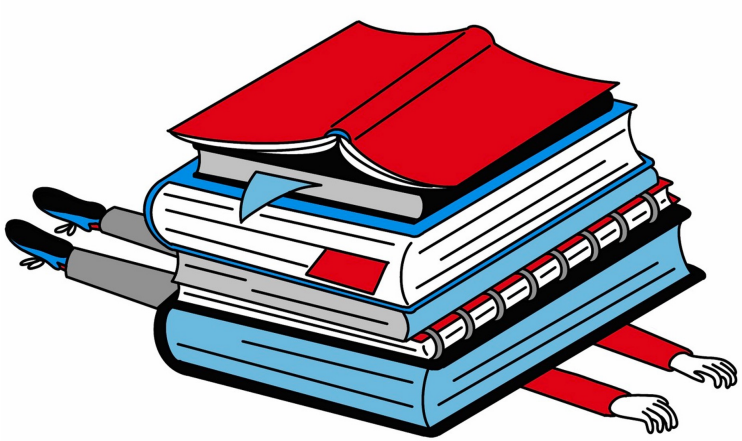


A milk tea brand

A personal care retailer

Logistics before next class (3 Dec 2024)

Very important!

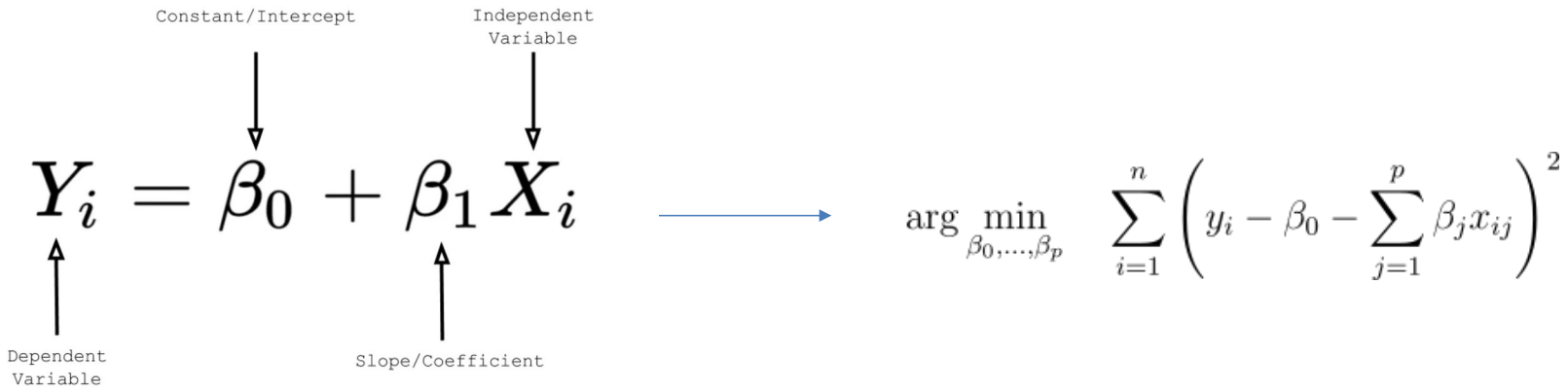


Submit your answers by **11:59pm, 2 Dec 2024**

Think about the implications (we ’ ll discuss some in the last section in next class)

Background	One Entity	Multi Entities	Relationships	Implications
• Prevalence of data or even big data • Urgency to make fast and good decisions • Limited cognitive resource and attention	• Basic summary statistics • Bias in observability of estimates and forecasts • Bias in observed estimates and forecasts (e.g., fat tails)	• Dimension reduction • Market structure analysis with structured data • Market structure analysis with unstructured data	• Variable selection in one linear regression • Variable selection in multiple linear regressions	• Optimal aggregation rules • Impact of reviews on product sales • Impact of AIGS on review activities and content consumption

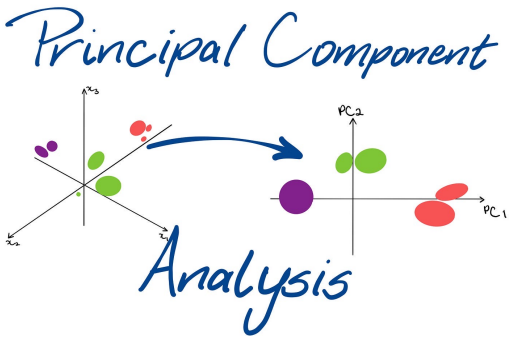
Aggregation of information about some relationships



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

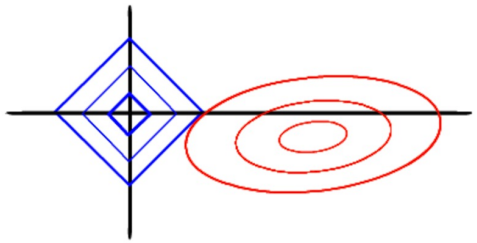
Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

Different variable selection methods

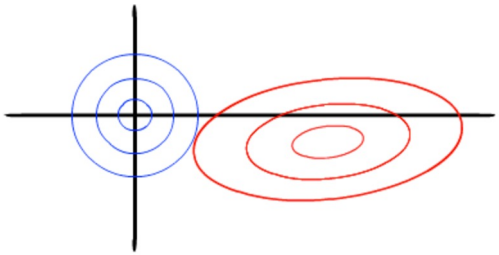


LASSO: least absolute shrinkage and selection operator

$$\arg\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

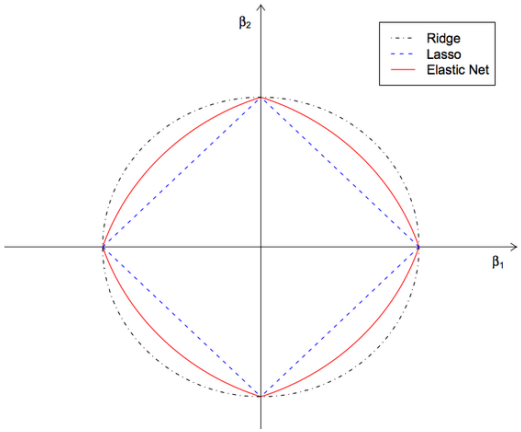


Ridge regression



The Elastic Net

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmin}} \left\|\boldsymbol{y} - \mathbf{X}\boldsymbol{w}\right\|_2^2 + \lambda_2\left\|\boldsymbol{w}\right\|_2^2 + \lambda_1\left\|\boldsymbol{w}\right\|_1$$



Applications of data aggregation in relationships

A single regression

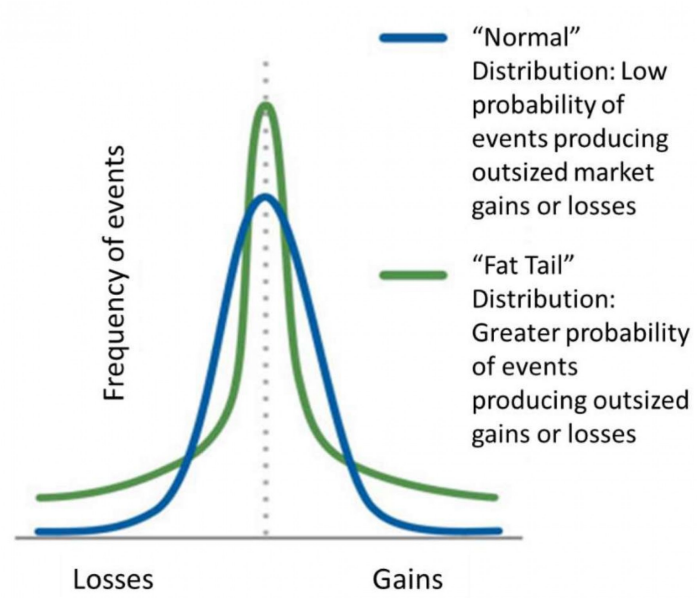
Ryzhov, Han, and Bradic (MNSC 2016)

Multiple regressions

Li, Netessine, and Koulayev (MNSC 2018)
Gu, and Kannan (JMR 2025)

Background	One Entity	Multi Entities	Relationships	Implications
• Prevalence of data or even big data • Urgency to make fast and good decisions • Limited cognitive resource and attention	• Basic summary statistics • Bias in observability of estimates and forecasts • Bias in observed estimates and forecasts (e.g., fat tails)	• Dimension reduction • Market structure analysis with structured data • Market structure analysis with unstructured data	• Variable selection in one linear regression • Variable selection in multiple linear regressions	• Optimal aggregation rules • Impact of reviews on product sales • Impact of AIGS on review activities and content consumption

Optimal aggregation when estimates are fat tailed



A simple heuristic that performs quite well across all possible scenarios.

$$\text{AMA}\, x = \frac{1}{2}\left(\text{Average}\, x + \text{Median}\, x\right).$$

(a) GN_1 samples (fat tails)

Judges	Bayes GN_1	Bayes $GN_{1.5}$	Average	Trimmed	Median	AMA
3	0%	3%	6%	—	4%	0%
5	0%	4%	12%	1%	5%	2%
10	0%	6%	20%	8%	1%	4%
20	0%	7%	26%	11%	2%	6%

(b) $GN_{1.5}$ samples (intermediate tails)

Judges	Bayes GN_1	Bayes $GN_{1.5}$	Average	Trimmed	Median	AMA
3	0%	0%	1%	—	11%	2%
5	1%	0%	2%	3%	12%	2%
10	3%	0%	3%	1%	8%	1%
20	5%	0%	4%	1%	11%	1%

(c) GN_2 samples (normal tails)

Judges	Bayes GN_1	Bayes $GN_{1.5}$	Average	Trimmed	Median	AMA
3	2%	0%	0%	—	16%	4%
5	5%	1%	0%	6%	19%	5%
10	10%	2%	0%	3%	18%	5%
20	14%	3%	0%	3%	21%	6%

Lobo, Miguel, and Dai Yao (2025). Fat tails in human judgment

Impact of ratings on product sales and subsequent ratings

TABLE 4
The Effects of WOM Communication on Weekly Box Office Revenue

Week (t)	1	2	3	4	5	6	7	8
CONST	7.273 (.000)**	9.346 (.000)**	6.449 (.000)**	8.171 (.000)**	6.469 (.000)**	4.981 (.015)**	.594 (.000)**	4.067 (.317)
$LNSCRN_t$	.566 (.000)**	.401 (.006)**	.673 (.000)**	.985 (.000)**	.973 (.000)**	1.050 (.000)**	1.379 (.000)**	1.185 (.000)**
$LNMSG_{t-1}$	.592 (.000)**	.345 (.000)**	.366 (.000)**	.275 (.005)**	.387 (.000)**	.169 (.0103)	–.088 (.449)	.144 (.362)
$POSIPER_{t-1}$	.784 (.257)	–.300 (.674)	.812 (.133)	.186 (.603)	.342 (.312)	.045 (.910)	–.185 (.497)	.118 (.715)
$NEGPER_{t-1}$	.059 (.952)	–1.369 (.092)*	–.610 (.264)	.445 (.309)	.385 (.136)	–.173 (.620)	–.265 (.399)	–.605 (.655)
LNCRITIC	.759 (.018)**	1.106 (.001)**	.894 (.020)**	.343 (.323)	.421 (.137)	.657 (.192)	.358 (.228)	.561 (.369)
CRPRO	.695 (.019)**	.398 (.235)	.688 (.076)*	.728 (.047)**	.736 (.009)**	.340 (.336)	.386 (.288)	.307 (.454)
LNNEW	–.111 (.505)	.005 (.984)	.153 (.494)	–.251 (.127)	–.136 (.423)	–.086 (.738)	.073 (.711)	–.112 (.682)
LNAGE	.146 (.760)	–.339 (.516)	.022 (.972)	–1.378 (.045)**	–.681 (.140)	–.039 (.964)	–1.510 (.097)*	.060 (.963)
Model fit F (p value)	44.87 (.000)	16.86 (.000)	22.70 (.000)	27.50 (.000)	50.37 (.000)	23.58 (.000)	69.04 (.000)	22.45 (.000)
Adjusted R^2	.902	.765	.820	.855	.925	.862	.954	.905
Joint F test of WOM measures	14.01 (.000)	9.64 (.000)	10.87 (.000)	4.01 (.017)	10.48 (.000)	1.07 (.382)	.43 (.735)	.53 (.672)

*$p < .10$.
**$p < .05$.
Notes: The dependent variable is $LNREV_t$. p values are in parentheses.

Liu, Yong (JM 2006). Word of mouth for movies: Its dynamics and impact on box office revenue

Impact of AIGS on review activities and content consumption

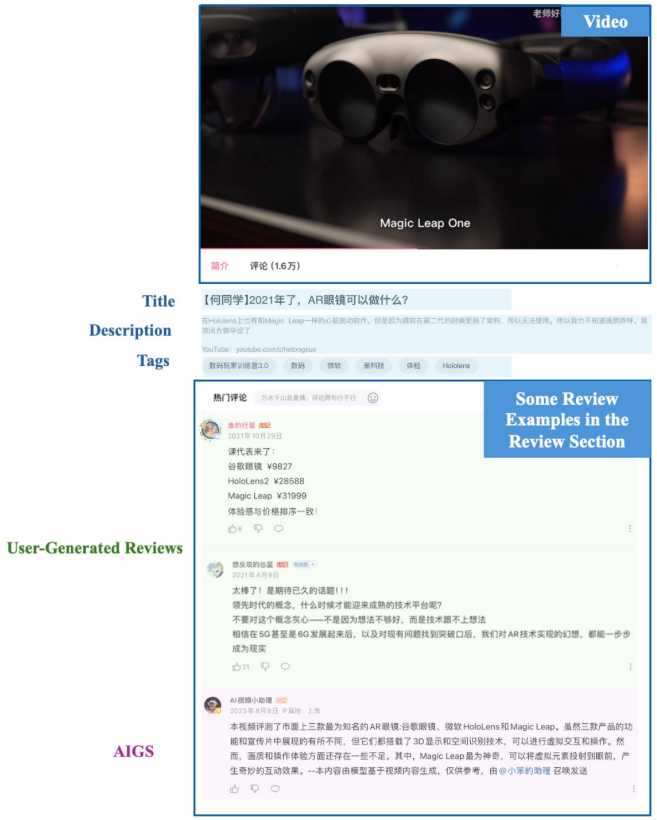


Table 3: Impact of AIGS on Online Video Consumption

	DailyViews (1)	DailyLikes (2)	DailyShares (3)	DailyReviews (4)	DailyTips (5)
$Treat_i * After_t$	.142***	.0758***	.0332***	.0294**	.0246*
	(.0257)	(.0155)	(8.51e-03)	(.0106)	(.0104)
$Followers_{it}$	1.73e-05***	-1.25e-06	-1.23e-05***	-1.29e-05***	-8.84e-06***
	(4.30e-06)	(2.59e-06)	(1.42e-06)	(1.77e-06)	(1.73e-06)
Age_{it}	-.0554***	-.0454***	-.0152***	-.0260***	-.0213***
	(4.34e-03)	(2.62e-03)	(1.42e-03)	(1.76e-03)	(1.72e-03)
Age_{it}^2	-1.26e-04**	1.52e-04***	9.63e-05***	2.07e-04***	1.57e-04***
	(4.41e-05)	(2.66e-05)	(1.46e-05)	(1.82e-05)	(1.77e-05)
Video fixed effect	Yes	Yes	Yes	Yes	Yes
Time fixed effect	Yes	Yes	Yes	Yes	Yes
N	29,363	29,363	29,363	29,363	29,363
R^2	7.4%	7.0%	3.7%	5.1%	3.6%

Notes: (1) Robust standard errors clustered at the video level in parentheses;
(2) $^{*}\,p < 0.05$, $^{**}\,p < 0.01$, $^{***}\,p < 0.001$

Qin, Rui, Yue Katherine Feng, and Dai Yao (2024). The impact of AI-generated summaries on video consumption

Background	One Entity	Multi Entities	Relationships	Implications
• Prevalence of data or even big data • Urgency to make fast and good decisions • Limited cognitive resource and attention	• Basic summary statistics • Bias in observability of estimates and forecasts • Bias in observed estimates and forecasts (e.g., fat tails)	• Dimension reduction • Market structure analysis with structured data • Market structure analysis with unstructured data	• Variable selection in one linear regression • Variable selection in multiple linear regressions	• Optimal aggregation rules • Impact of reviews on product sales • Impact of AIGS on review activities and content consumption