

# ECON-C5100 Digital Markets

---

Iivo Vehviläinen

January 12, 2022

Aalto University

*[iivo.vehvilainen@aalto.fi](mailto:iivo.vehvilainen@aalto.fi)*

## Lecture 2: Preferences

---

- The square and the tower
  - Market institutions matter
- Digital marketplaces increase efficiency
  - Reduction of search costs and other frictions

- Preferences online
- Estimation and prediction

## Recap: Market demand builds on preferences



Figure. Demand of a good depends on the preferences of the individuals who participate to the market.

# Preferences online

Ads · Shop headphones

				
Apple AirPods mit LadeCase (...)	Essager - Kabellose TW...	Digitaler Stereo-Funk-Kopfhör...	Audeze LCD-5	Apple AirPods Pro
CHF 139.00	CHF 44.90	CHF 89.95	CHF 5'050.00	CHF 279.00
Apple	apfelkiste.ch	Pearl Schweiz	Thomann CH	Apple
Free shipping By Google	Free shipping By Smarketer	By Google	Free shipping By BiddingLab	Free shipping By Google

Ads · Shop headphones

				
In-ear Headphone...	Mifo O5 Plus Gen 2 Smar...	AirPods (3rd generation) ...	PX7 Carbon Edition...	AH-D1200 Headphone...
\$199.99	\$89.99	\$179.00	\$399.00	\$99.99
Grell Audio	Mifo US Store	Apple	Bowers & W...	Denon
Free shipping	Special offer	Special offer	Free shipping	★★★★★ (35)

Figure. Two identical Google searches, one done from Zurich, Switzerland (top) and another from Mexico City, Mexico (bottom)

## Experiment: Search engine

### Google *Beats headphones*

- Report the price from the first ad that is shown to you at *Search* chat at [presemo.aalto.fi/digimar](https://presemo.aalto.fi/digimar)
- You can use your preferred other search engine, if Google is out of the question

- What you want to do
  - For example buy a phone, watch a movie, connect with friends
- What a firm wants to do: find out what you want to do by tracking your actions. . . and try to influence behavior:
  - Design a mechanism (market place, auction, platform) to discover your preferences
  - Collect and use data to improve performance

## Tracking user data

- Big tech companies and others collect data of their users and use it to advance their business
- Ownership of this data may not be clearly defined or understood; nor is its value
- People exchange data of varying value against payment in the form of service(s) that they use

- Browser cookies are an early invention to track users
- They are still widely used, but their benefit has been reduced
  - Regulation made companies to disclose the use of cookies in the EU ("This website uses cookies to...")
  - Browsers have started to compete on privacy dimension, you can opt-out from cookies
- The tracking industry is responding with new technologies
  - Merging user information from device models to browsing behavior and keystroke or mouse signatures
  - Keeping people signed in to their devices
- With combined data from all users, it is possible to predict your behavior, even if you don't disclose it yourself

- There are two key motivations to find out your preferences:
  1. Understand the drivers behind your choices
  2. Predict how you will behave in a given situation
- The first part, causal inference, is a traditional field of economics (microeconometrics)
- The second part, prediction, is what the big tech firms are doing with the availability of big data

To pinpoint the differences:

- Let's consider a simple model

$$Y_i = f(X_i) + \epsilon_i$$

- $X_i$  are data (e.g. location, past clicks).
- $Y_i$  are variables of interest (e.g. click to an ad).
- Function  $f$  maps values of  $X$  to  $Y$  and  $\epsilon_i$  are the random noise (in measurement, unobserved variables, etc.).

The goals in estimation and prediction are different:

1. The goal of regression estimation is unbiasedness:

$$E[\hat{f}] = f.$$

2. The goal of prediction is to minimize prediction error:

$$\hat{f} = \min_{f \in \mathcal{F}} L(f) = \min_{f \in \mathcal{F}} E[\ell(f(x), y)],$$

where e.g.  $\ell(f(x), y) = (y - f(x))^2$ .

## The bias-variance tradeoff

If  $E[\epsilon] = 0$ , the expected prediction error is:

$$\begin{aligned}\text{error} &= E[(y - \hat{f}(x))^2] \\ &= E[(y - f(x))^2] + (E[\hat{f}(x)] - f(x))^2 \\ &\quad + E[(\hat{f}(x) - E[\hat{f}(x)])^2]\end{aligned}$$

for some  $x$  out of sample. Using  $\epsilon = y - f(x)$ , we have

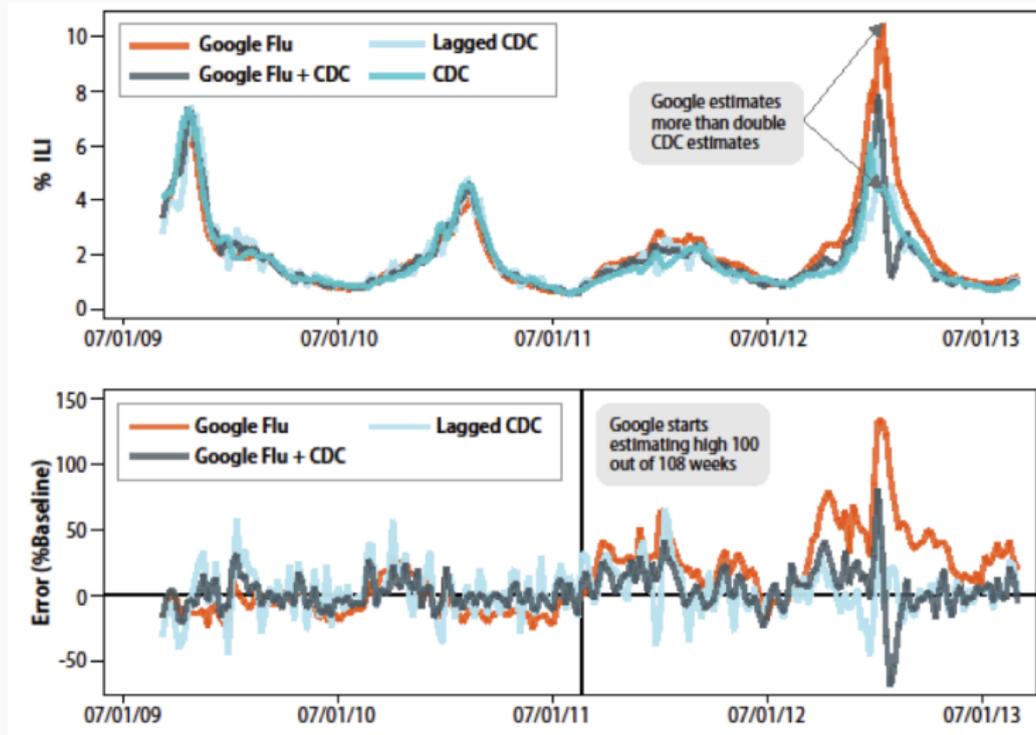
$$\text{error} = \text{Var}(\epsilon) + \text{Bias}^2 + \text{Var}(\hat{f}(x))$$

i.e. a combination of noise or variance in data ( $\text{Var}(\epsilon)$ ), bias in estimation, and variance from the fact that a sample is used to estimate  $\hat{f}$

Basic idea:

- Data has signal and noise.
- More “expressive” functions
  - Capture the signal better
  - But also pick up the noise
- Prediction algorithms allow tuning of parameters so that the signal/noise ratio is as desired

## Example: Prediction gone wrong



## Why can predictions go wrong?

- Let's go back to the simple model, and note that it most often is a vector valued (see e.g. Facebook example below)

$$Y_i = f(\mathbf{X}_i) + \epsilon_i$$

- Machine learning algorithms choose what parameters of  $\mathbf{X}_i$  are used, automatically and based on past data
- With big data, manual analysis of data would be impossible/too costly/too slow
- In complex environments with big data, no one knows why a certain set of parameters gets chosen

## Experiments and A/B testing

- A/B testing is one method to extract more data
- The idea is to take a random sample of users who are shown a version B of the website while the rest use version A
- Comparing results (e.g. clicks, purchases) between A and B can *predict* which version is better
- These are very similar to controlled experiments used in economics, medicine etc., just the objective is different
- For a website it may be enough to see which version works better (for now), but for other decisions it can be crucial to understand what are the reasons for the differences

# Further example: Causal inference with prediction

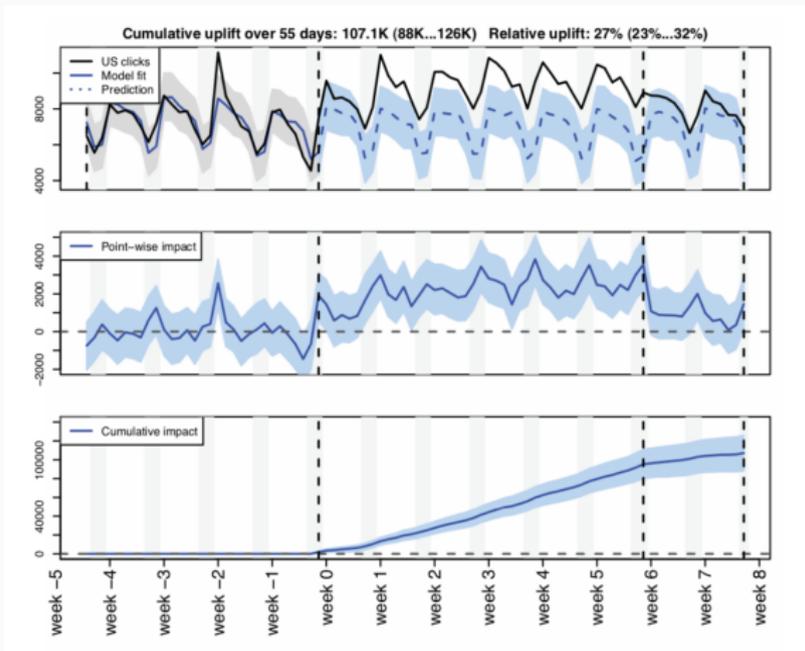
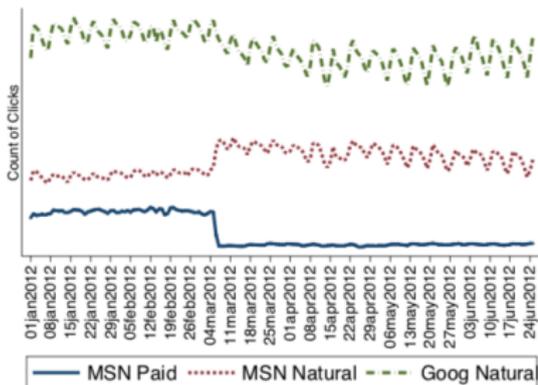


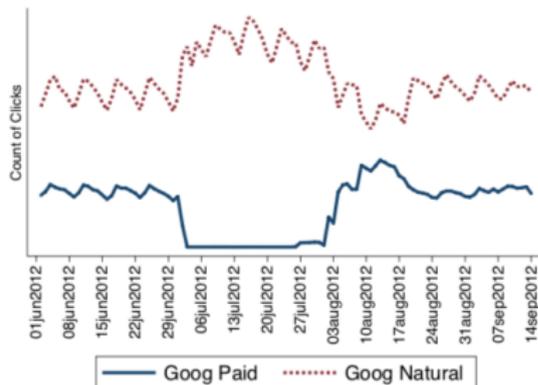
Figure. Actual clicks vs. predicted clicks and 95 % confidence interval.

# Care needed in interpretation

Figure 2: Brand Keyword Click Substitution



(a) MSN Test



(b) Google Test

Note: MSN and Google click traffic is shown for two events where paid search was suspended (Left) and suspended and resumed (Right).

*Figure.* Regular search is a substitute for eBay brand ads. One explanation: selection of consumer who want to go to eBay.

## Takeaways from today

- Firms extract data because it is of value to them
- Collecting data online easier than ever
- Prediction and causal inference aim for different goals
  - Firms mostly interested in prediction
  - Care needed to understand what are the best tools

# Materials for this week

## Reading assignment 1:

- Athey, Susan and Michael Luca (2019) “Economists (and Economics) in Tech Companies”, Journal of Economic Perspectives. Read the whole article.
- Varian, H. (2012) “Revealed Preferences and its Applications”, Economic Journal. Read the Introduction and Section 2 for now.

## Online resources (make sure you know these before you take Exercise 1):

- Make sure you know the basics of consumer choice: e.g. [mru.org](http://mru.org): [Consumer Choice](#) and/or [www.core-econ.org](http://www.core-econ.org) 3.2-3.5, 3.7.1.
- and supply–demand equilibrium e.g. [mru.org](http://mru.org): [Supply, Demand, and Equilibrium](#) and/or [www.core-econ.org](http://www.core-econ.org) 8.1, 8.2.

### Competition

- Perfect competition
- Monopoly
- Oligopoly

# Appendix

---

## Example: Netflix

- Netflix has a 60 PB+ of data collected from all its operations, with 100+ million subscribers (around 600 MB per customer).
- Company policy is to only do data driven business decisions:
  - Recommend films
  - Choose which content to purchase
  - Create content (House of Cards)
  - Improve user experience through A/B testing
  - Give users information on which Internet provider to use
- But where should the limits on data use be?
  - Facebook was caught providing Netflix access to its user data without the consent of the users (NYT, 18 Dec 2018)

# Example: Facebook

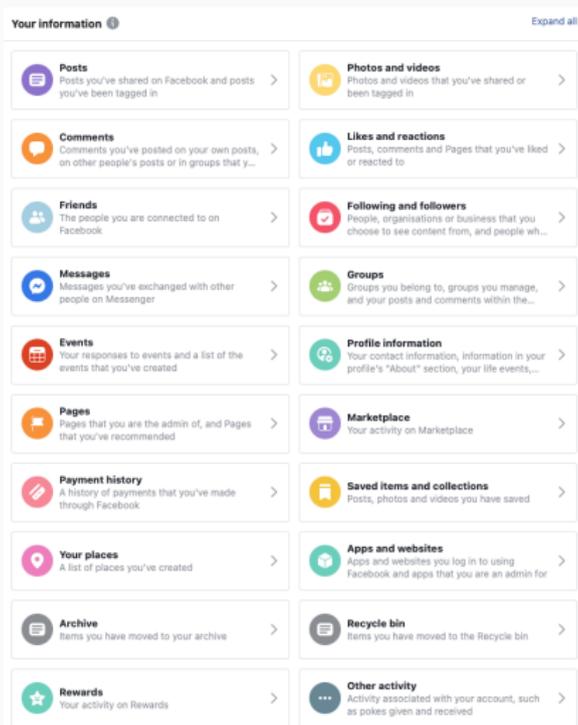
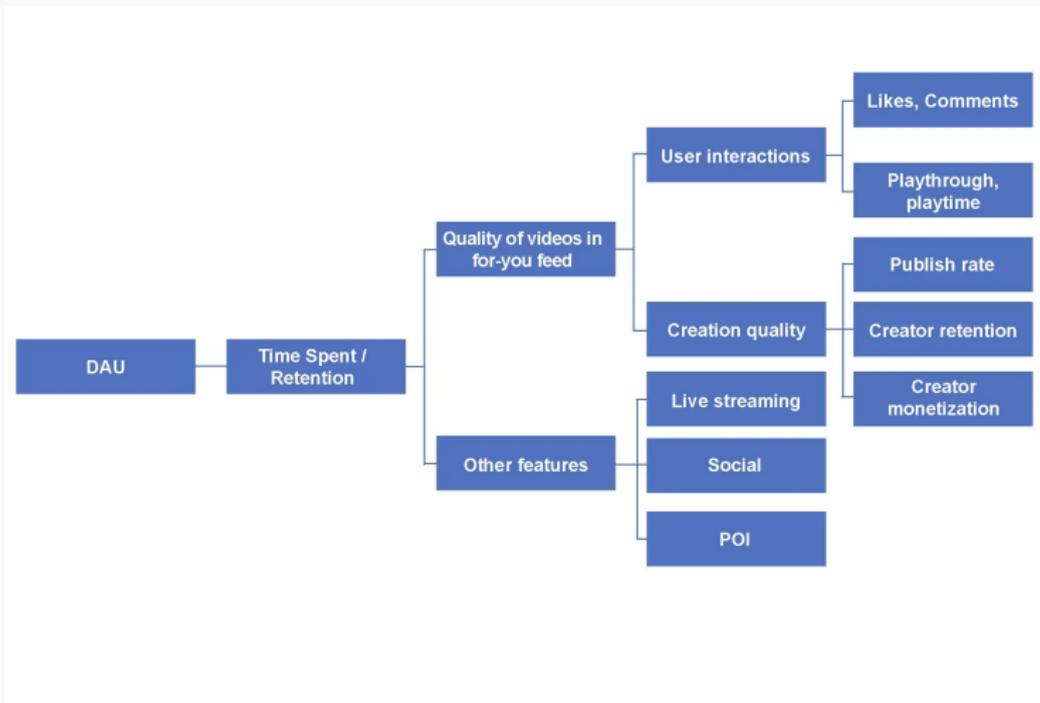


Figure. Categories of data that Facebook collects of you.

# Example: Tiktok



*Figure. The goals of Tiktok's algorithm*

## Example: Automated selection

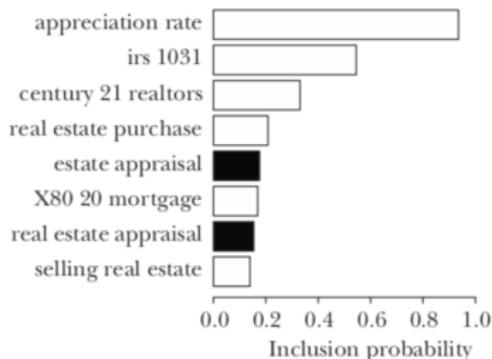
### An Example Using Bayesian Structural Time Series (BSTS)

*(finding Google queries that are predictors of new home sales)*

A: Initial Predictors



B: Final Predictors



Source: Author using HSN1FNSA data from the St. Louis Federal Reserve Economic Data.

Panel A shows uncurated Google Correlate results to predict new homes sold in the US. Choices adjusted by hand in Panel B

# Simple solutions: Change one parameter at a time

- Question: What price to charge customers?
- Direct variation of prices can be problematic, instead indirect means are used to extract data (see Figure).
- Such testing, done offline as well, can provide an understanding on how demand changes when prices change.
- Online market places differ from traditional sales in the speed and reach of testing.

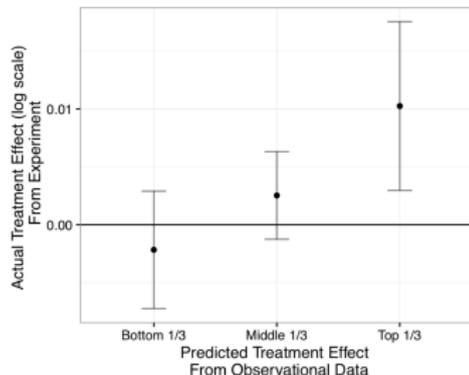
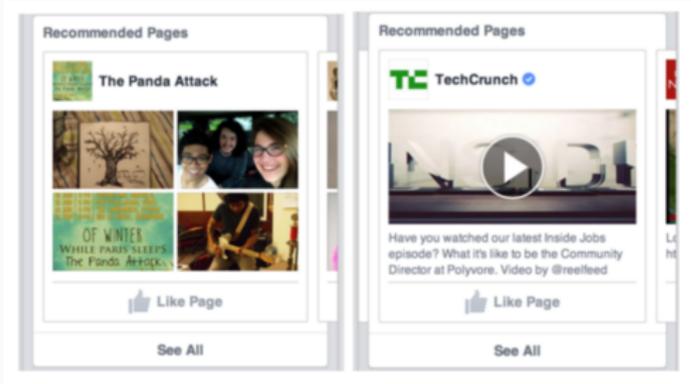
	<b>Basic</b> \$20 Monthly	<b>Premium</b> \$40 Monthly
Set-up Fee	None	None
Storage	5gb	10gb
Custom Domains	✗	✗
Secure SSL	✗	✗
	Select Plan	Select Plan

*recommended way* (with arrow pointing to Basic plan)

*v/s* (between Storage rows)

Figure: VWO.

# More complex example: A/B testing at Facebook



- Problem with simple solutions is that people respond differently.
- Advanced tools with big data are actively developed to yield better identification. Above example from Facebook lab is from an A/B test with 400,000 randomly chosen users.

Source: Peysakhovich and Lada (2016).