[#4] Finance and Data

Part II: Financing and Insurance

Johan Hombert

Road map

Introduction

Financial inclusion

Winner's curse

Lucas critique

Hirshleifer effect

Discrimination

Introduction

- Prediction is central in finance
  - Corporate valuation (PE, M&A): predict cash flow
  - Credit: predict default
  - Insurance: predict damages
  - Asset management, trading: cf. previous class
- Data are key. New data/AI $\Rightarrow$ new business applications

Credit scoring with alternative data

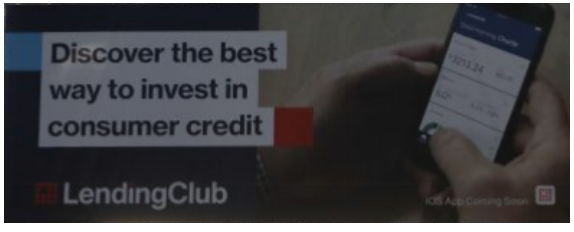
- Credit score providers: use data on people and businesses to calculate credit scores sold to banks

"Social medial insight program that extracts data from Yelp, Facebook, Twitter, and Four Square is offered for use by private lenders and traditional banks. Credit scores for over 1 billion people & businesses, including 235 million individuals US consumers and over 25 million US businesses."

Also done internally by banks

P2P lending: a brief history

1. PROSPER From crowdlending to marketplace lender
   - Crowdlending: match borrowers and lenders; lender sets financing terms
   - Marketplace: score borrowers, set interest rate, match borrowers and lenders
2. LendingClub From marketplace to (shadow) bank



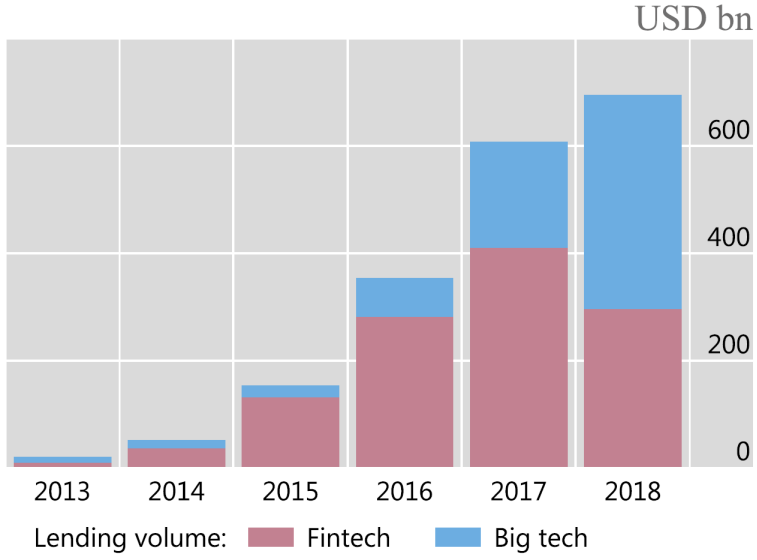
LendingClub Closing Down Their Platform for Retail Investors

Peter Renton · Peer to Peer Lending · Oct. 7, 2020 · 5 min read

3. (Shadow) Bank: score borrowers, set interest rate, lend on balance sheet

Big tech in lending markets

- Tech firms are well positioned in lending markets: they have access to consumers + have lots of data on them



Business lending

amazon Start Grow Learn Pricing Blog

Amazon Lending

Convenient financing options for your business

Consumer lending

Goldman Sachs, Apple Team Up on New Credit Card

By Tripp Mickle and Liz Hoffman

Updated May 10, 2018 11:38 am ET | WSJ Pro

Apple to offer buy now, pay later credit in challenge to Klarna and Affirm

Tech group announces short-term loans system even as pandemic ecommerce boom begins to wane

Tim Bradshaw and **Siddharth Venkataramakrishnan** in London and **Imani Moise** in New York JUNE 6 2022

InsurTech

- Use alternative data and AI to underwrite and price insurance

Data machine: the insurers using AI to reshape the industry

Groups are building detailed customer profiles to inform pricing and try to influence behaviour

AI allows insurers such as Ping An to produce highly individualised profiles of customer risk that evolve in real time © FT montage; Alamy, Dreamstime

Ian Smith in London AUGUST 16 2021

50

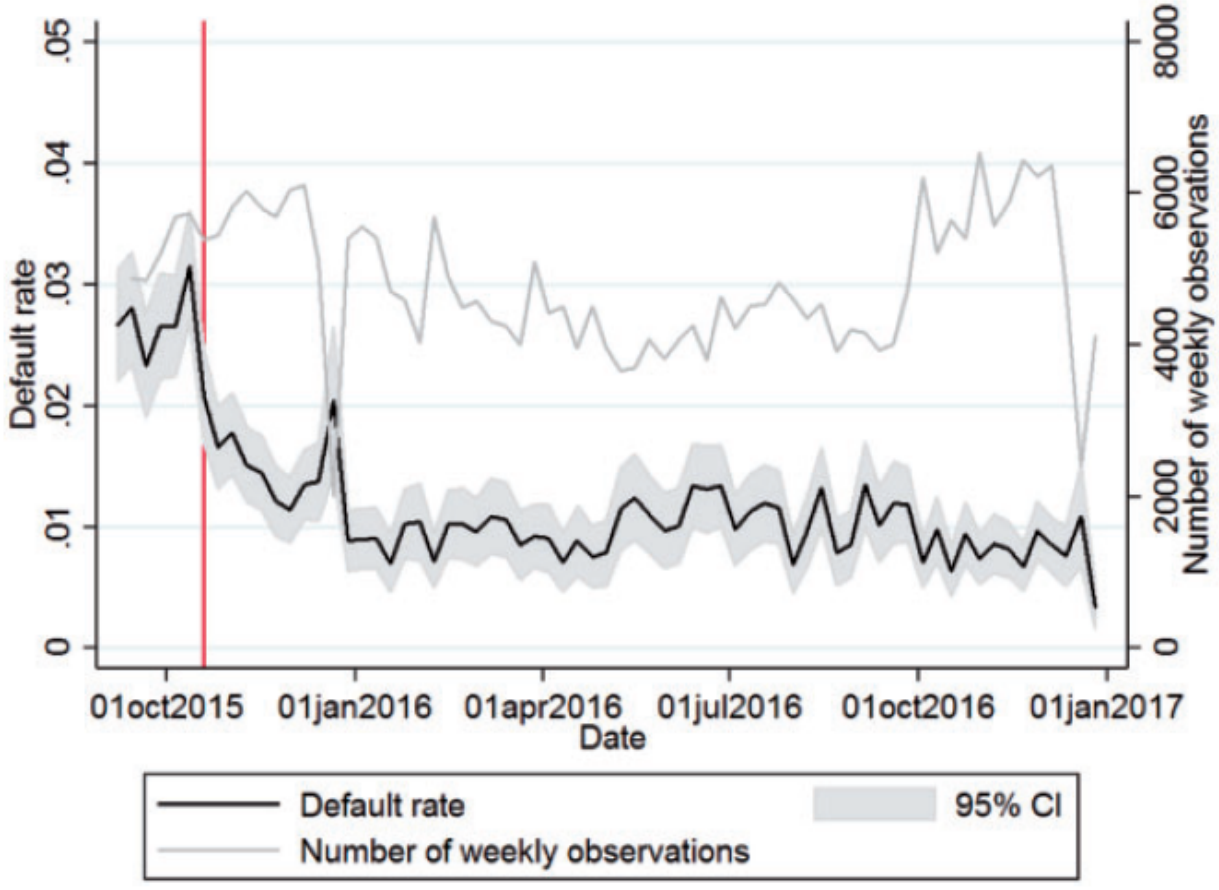
Case study

- Scoring with digital footprints at an E-commerce company[1]
  - Goods sent first, paid for later → need to assess buyer's creditworthiness
  - Credit score based on credit history, sociodemographics, past transactions, etc.
  - After Oct 2015: also collected digital footprints (OS, email, etc.)
  - Does this improve prediction of default?

[1] "On the Rise of FinTechs: Credit Scoring Using Digital Footprints," 2019, Berg, Burg, Gombovic & Puri, *Review of Financial Studies* [pdf]

Default regressions (scorable customers)

	(1) Credit bureau bureau score		(2) Digital footprint		(3) Credit bureau score & digital footprint	
Variables	Coef.	z-stat	Coef.	z-stat	Coef.	z-stat
Credit bureau score	−0.17***	(−7.89)			−0.15***	(−6.67)
Device type & operating system[a]						
Desktop/Windows			Baseline		Baseline	
Desktop/Macintosh			−0.07	(−0.53)	−0.13	(−1.03)
Tablet/Android			0.29***	(3.19)	0.29***	(3.06)
Tablet/iOS			0.08	(1.05)	0.08	(0.97)
Mobile/Android			1.05***	(17.25)	0.95***	(15.34)
Mobile/iOS			0.72***	(9.07)	0.57***	(6.73)
E-mail Host [a]						
Gmx (partly paid)			Baseline		Baseline	
Web (partly paid)			0.00	(0.00)	−0.02	(−0.22)
T-Online (affluent customers)			−0.40***	(−3.90)	−0.35***	(−3.35)
Gmail (free)			0.34***	(3.81)	0.29***	(3.09)
Yahoo (free, older service)			0.75***	(9.19)	0.72***	(8.98)
Hotmail (free, older service)			0.35***	(3.70)	0.28***	(2.72)
Channel						
Paid			Baseline		Baseline	
Affiliate			−0.49***	(−5.35)	−0.54***	(−5.58)
Direct			−0.27***	(−4.25)	−0.28***	(−4.44)
Organic			−0.15*	(−1.79)	−0.15*	(−1.74)
Other			−0.47***	(−4.50)	−0.48***	(−4.36)
Checkout time						
Evening (6 p.m.-midnight)			Baseline		Baseline	
Morning (6 a.m.-noon)			0.28***	(4.50)	0.28***	(4.60)
Afternoon (noon-6 p.m.)			0.08	(1.42)	0.08	(1.47)
Night (midnight-6 a.m.)			0.79***	(7.73)	0.75***	(7.09)
Do-not-track setting			−0.02	(−0.25)	−0.07	(−0.91)
Name in e-mail			−0.28***	(−5.67)	−0.29***	(−5.70)
Number in e-mail			0.26***	(4.50)	0.23***	(3.91)
Is lowercase			0.76***	(13.10)	0.74***	(13.20)
E-mail error			1.66***	(20.00)	1.67***	(20.36)
Constant	12.42***	(5.76)	−4.92***	(−62.87)	9.97***	(4.48)
Control for *Age*, *Gender*, *Item category*, *Loan amount*, and month and region fixed effects	No		No		No	
Observations	254,819		254,819		254,819	
Pseudo R^2	**.0244**		**.0524**		**.0717**	
AUC	**0.683**		**0.696**		**0.736**	
(SE)	**(0.006)**		**(0.006)**		**(0.005)**	
Difference to AUC=50%	**0.183*** **		**0.196*** **		**0.236*** **	
Difference AUC to (1)			**0.013***		**0.053*** **	


Default rate
0
.01
.02
.03
.04
.05
Number of weekly observations
0
2000
4000
6000
8000
01oct2015
01jan2016
01apr2016
01jul2016
01oct2016
01jan2017
Date
Default rate
95% CI
Number of weekly observations

New data and AI in finance

- Many opportunities but also pitfalls to avoid for fintech entrepreneurs and for society

1. Financial inclusion
2. Winner's curve
3. Lucas critique
4. Privacy
5. Hirshleifer effect
6. Discrimination

Road map

Introduction

Financial inclusion

Winner's curse

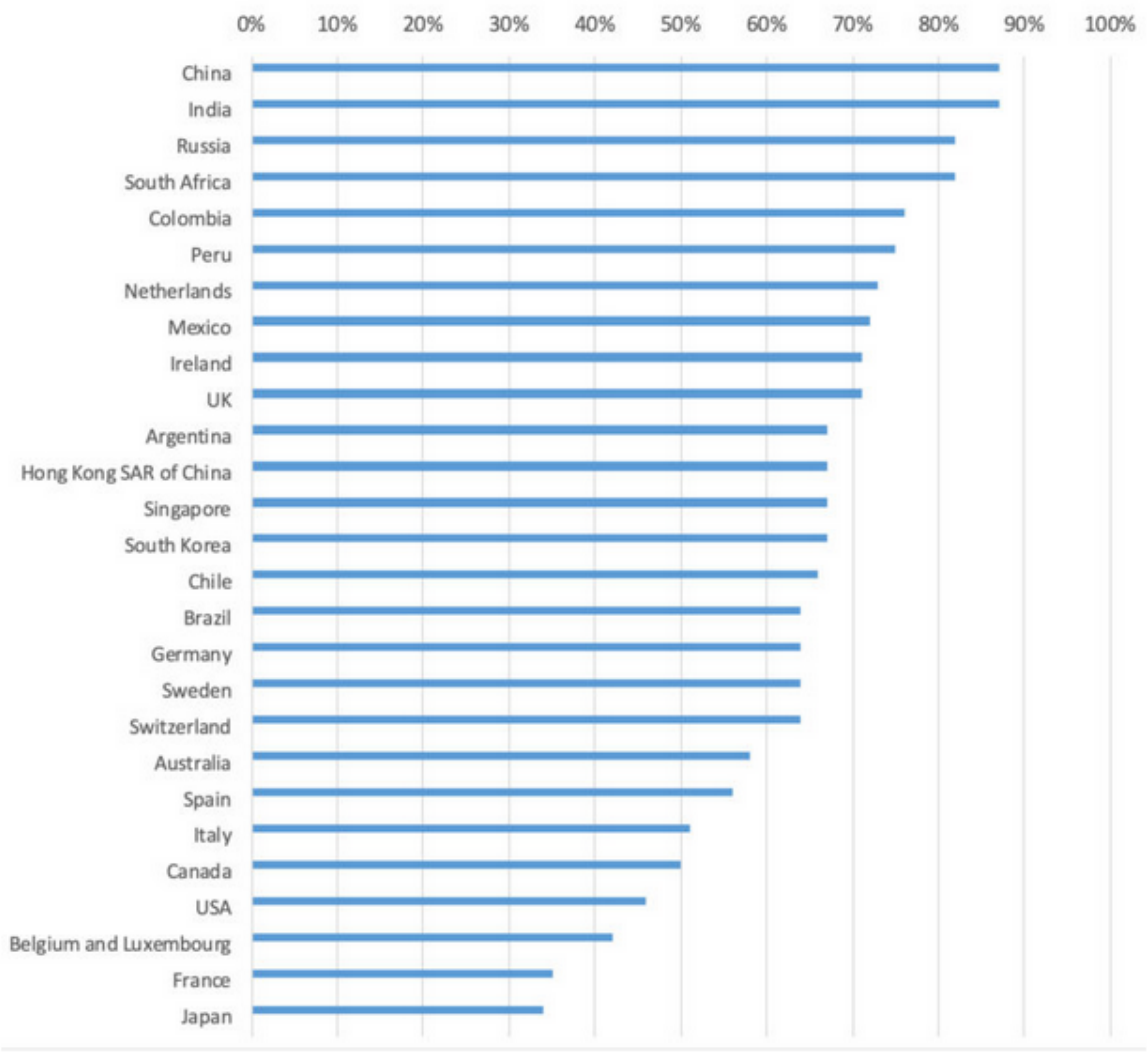
Lucas critique

Hirshleifer effect

Discrimination

Fintech and financial inclusion

Fintech adopters in percentage of digitally active population (source: EY)



Digital banking + Add to myFT

Banks use fintech to make up for lost time on financial inclusion

Institutions are investing in a bid to reach 'unbanked' groups in Africa and the Middle East

Laura Noonan APRIL 24 2019

Fintech + Add to myFT

How developing nations use tech to reach the 'underbanked'

Lenders in Africa and the Middle East circumvent weak digital infrastructure to make progress

Sarah Murray APRIL 24 2019

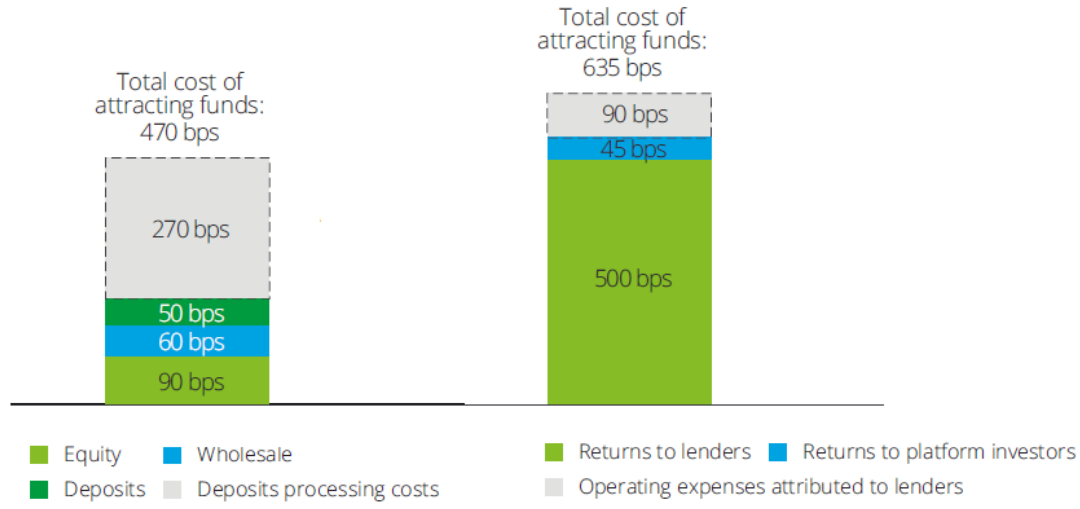
Fintech and financial inclusion

- Can fintech lenders improve access to credit?
- Fintech lenders' competitive edge over traditional banks:
  - A. Lower fixed cost of underwriting loans
  - B. More (or different) information on borrowers

A. Cost structure

- Fintech lender has:
  - lower fixed cost of underwriting a loan
    - automatized process
    - no physical premises, no loan officer
  - higher marginal cost of underwriting a loan

Figure 8. Costs of funding an unsecured personal loan: banks and MPLs



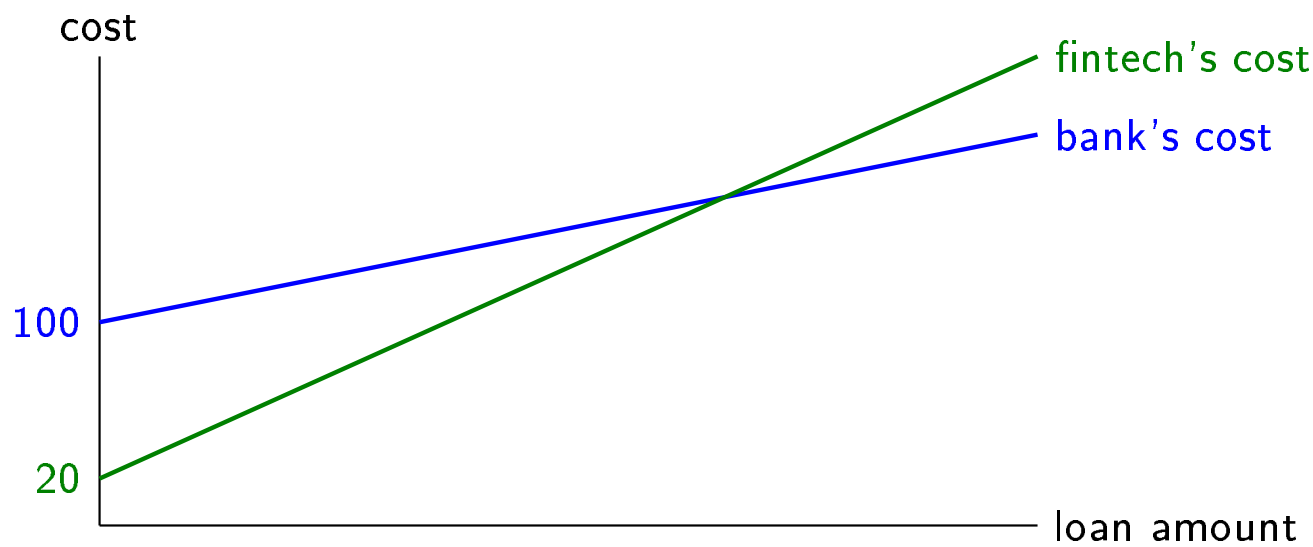
Source: Deloitte analysis

*MPL = Market-Place Lender

Cost structure

Bank: underwriting a X € loan costs a fixed 100€ plus 4% of X

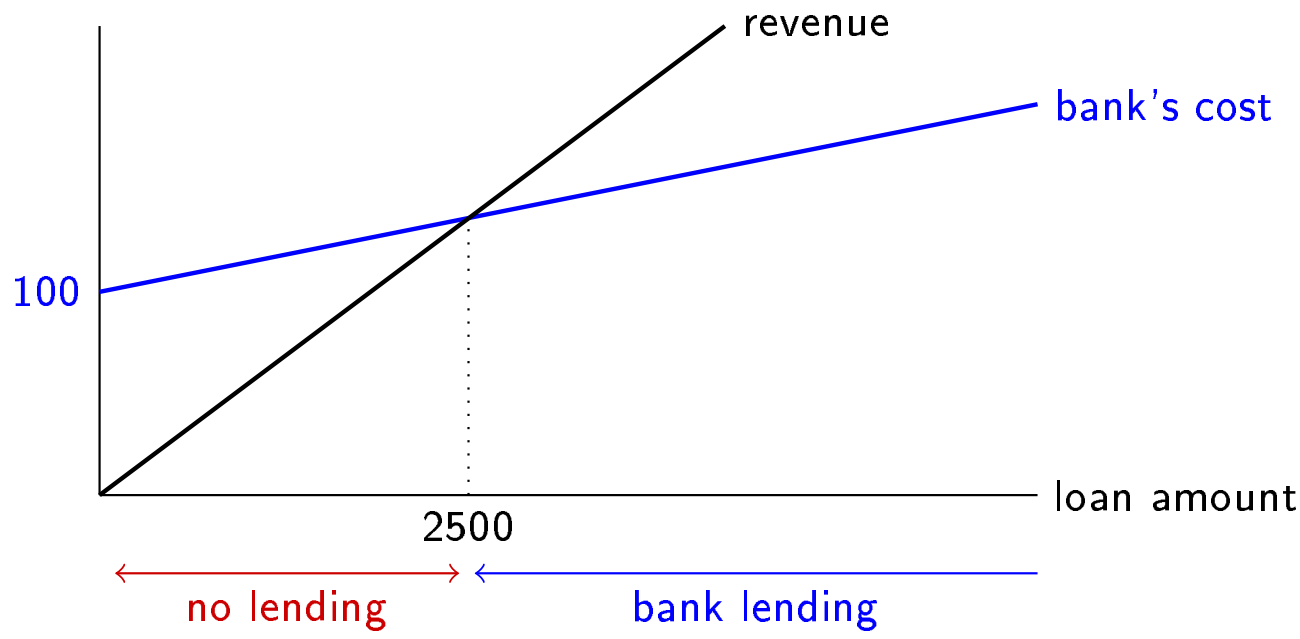
Fintech: underwriting a X € loan costs a fixed 20€ plus 6% of X



Cost structure

- Suppose different borrowers want different loan amounts X. All are willing to pay a 8% interest rate
- Before fintech entry: the bank lends if

$$\text{revenue} > \text{cost} \quad \Leftrightarrow \quad 0.08X > 0.04X + 100 \quad \Leftrightarrow \quad X > 2500\ €$$



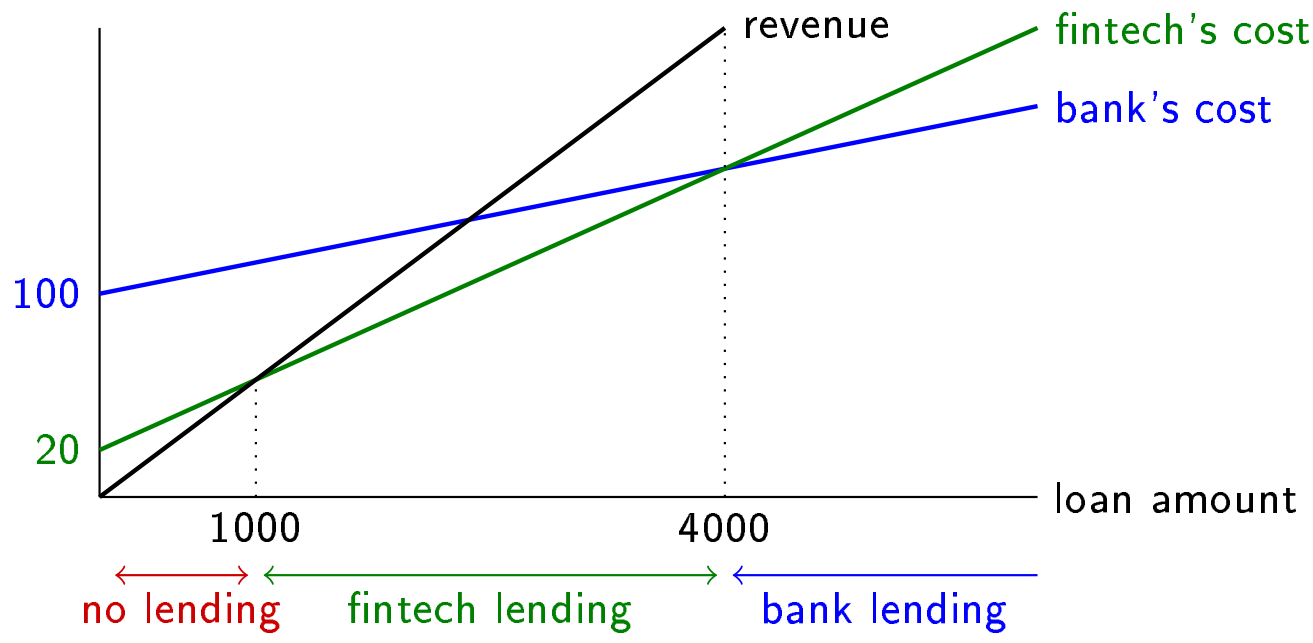
- After fintech entry: what happens?

Q1. Some borrowers gain access to credit: true or false

Q2. Some borrowers lose access to credit: true or false

Cost structure

- After fintech entry:



1. Small borrowers gain access to credit from fintech (fin. inclusion ↑)
2. Fintech gains market shares on the intermediate segment
3. Traditional lenders still dominate the large loans segment

B. New data

- Fintech lenders have alternative data on borrowers
  - For ex. digital footprints
- Example
  - Loan amount 100 with interest rate R
  - If borrower repays: $100 \times (1 + R)$ to the lender
  - If borrower defaults: 0 to the lender
  - Lender's funding cost + fixed cost of loan underwriting: 0

Calculate the interest rate at which the lender breaks even if the probability of default is 10%

9.9% 10.0% or 11.1% ?

B. New data

- Fintech lenders have alternative data on borrowers
  - For ex. digital footprints
- Example
  - Loan amount 100 with interest rate R
  - If borrower repays: $100 \times (1 + R)$ to the lender
  - If borrower defaults: 0 to the lender
  - Lender's funding cost + fixed cost of loan underwriting: 0

Calculate the interest rate at which the lender breaks even if the probability of default is 10%

$$-100 + 0.90 \times 100 \times (1 + R) \geq 0 \quad \Rightarrow \quad R \geq \frac{0.10}{0.90} = 11.1\%$$

New data

- Consider small entrepreneurs with no credit history
  - Ex.: new businesses, informal businesses
  - Entrepreneurs are willing to pay a 8% interest rate
  - Traditional lenders' best estimate of Probability of Default (PD) is 20%

Do traditional lenders lend to these entrepreneurs?

Yes No ?

New data

- Consider small entrepreneurs with no credit history
  - Ex.: new businesses, informal businesses
  - Entrepreneurs are willing to pay a 8% interest rate
  - Traditional lenders' best estimate of Probability of Default (PD) is 20%

Do traditional lenders lend to these entrepreneurs?

No: lender breaks even only if $R \geq \frac{0.20}{0.80} = 25\%$ but the entrepreneur takes the loan only if $R \leq 8\%$

What is the maximum PD above which traditional lenders stop lending?

7.4% 8.4% ?

New data

- Consider small entrepreneurs with no credit history
  - Ex.: new businesses, informal businesses
  - Entrepreneurs are willing to pay a 8% interest rate
  - Traditional lenders' best estimate of Probability of Default (PD) is 20%

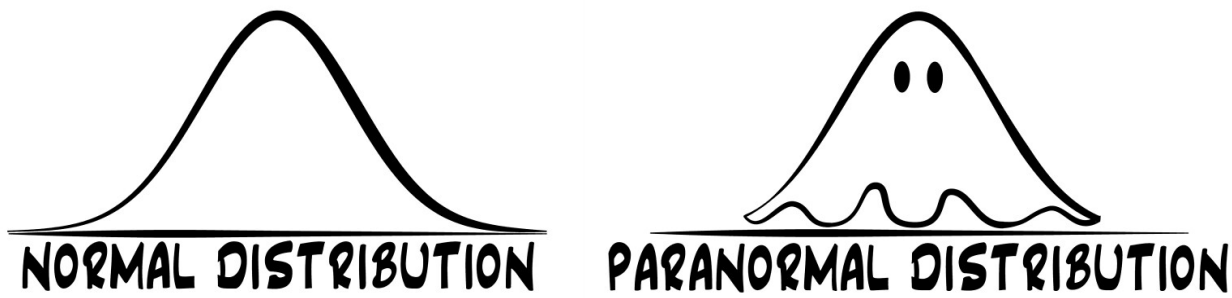
Do traditional lenders lend to these entrepreneurs?

No: lender breaks even only if $R \geq \frac{0.20}{0.80} = 25\%$ but the entrepreneur takes the loan only if $R \leq 8\%$

What is the maximum PD above which traditional lenders stop lending?

Lending can happen only if $\frac{PD}{1-PD} \leq 8\%$ i.e. $PD \leq 7.4\%$

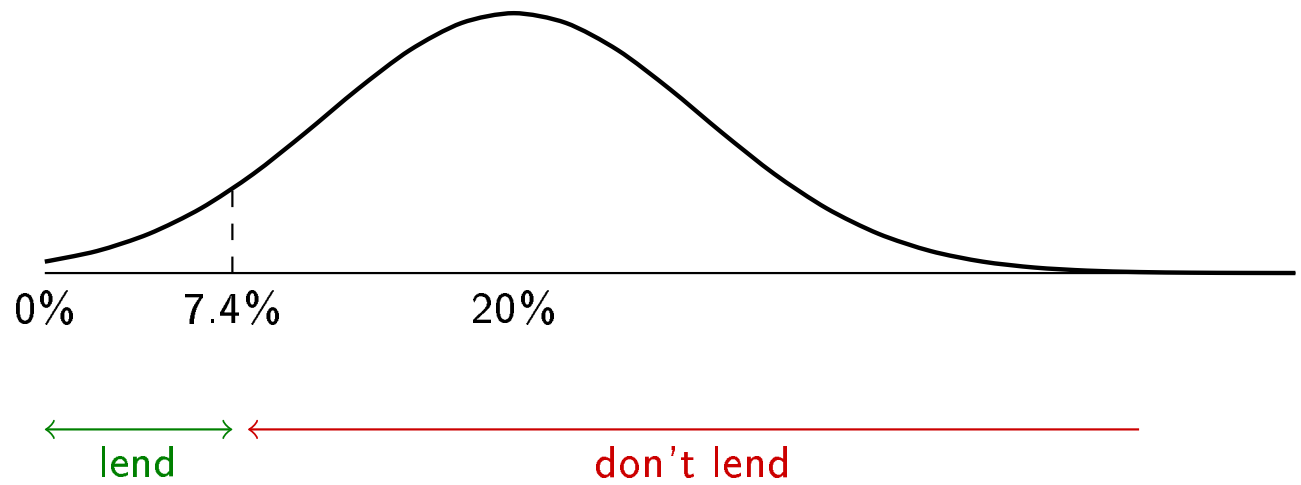
Alert: Statistics coming up!



New data

- Fintech lenders have access to entrepreneurs' digital footprints
  - While traditional lender's best estimate of PD is 20% for all entrepreneurs, fintech can better identify talented entrepreneurs and estimate a more precise PD for each entrepreneur

$\Rightarrow$ Distribution of fintech's best estimate of PD



New data

- Implications of more data

1. Talented entrepreneurs with no credit history gain financing

2. Business opportunity for fintech: can charge 8% to entrepreneurs with low PD and earn profits

$\Rightarrow$ Win-win

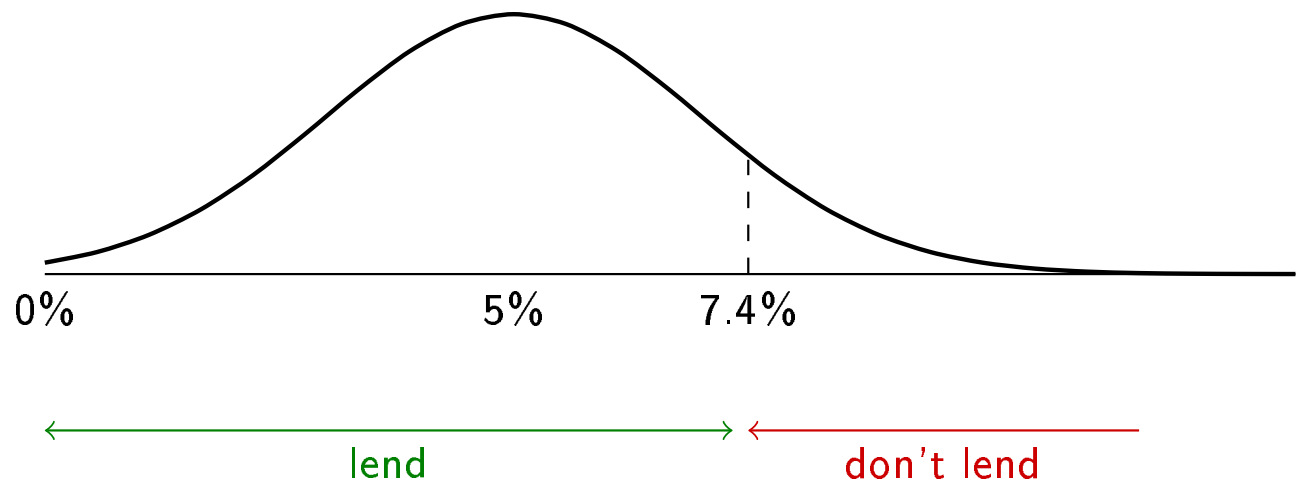
New data

- Can entrepreneurs lose financing when lenders have more data?
- Example: Entrepreneurs with good credit history
  - Traditional lenders' best estimate is $PD = 5\%$
  - Entrepreneurs are willing to pay up to 8%

  $\Rightarrow$ All these entrepreneurs are financed

New data

- Fintech lenders have access to entrepreneurs' digital footprints
  - More precise estimate of PD of each entrepreneur
  - Distribution of estimated PD



New data

- Implications of more data (cont'd)

  3. Some entrepreneurs lose financing: those with high (previously undetected) default risk

  4. Allocation of credit is better

     - Note the parallel with the previous class: better information leads to better allocation of capital

- Problem 1 in problem set

- Potential risk: discrimination (more on this later)

Road map

Introduction

Financial inclusion

Winner's curse

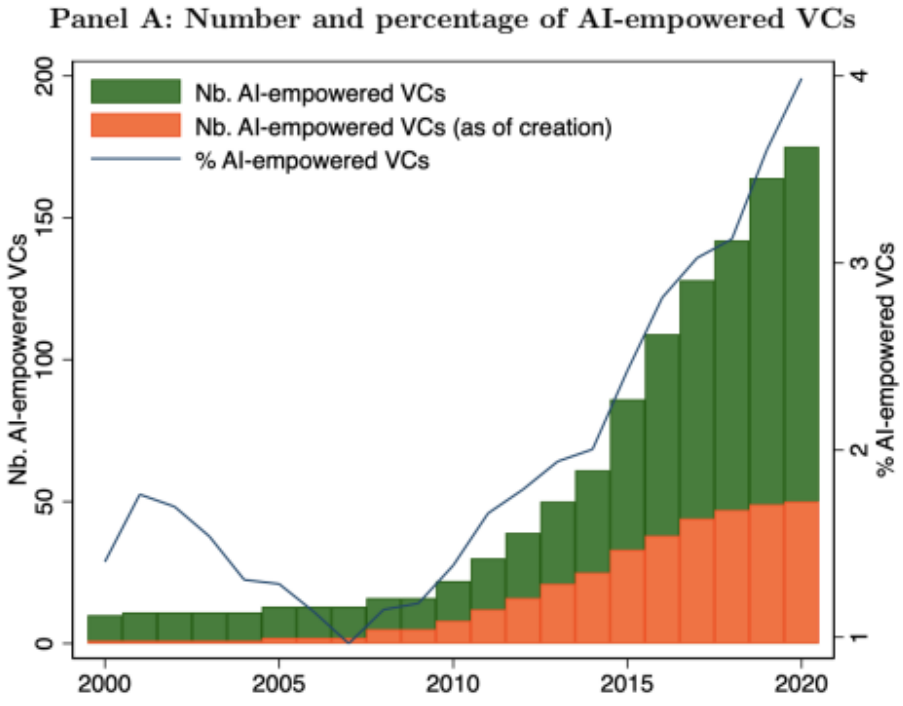
Lucas critique

Hirshleifer effect

Discrimination

Many potential applications of AI

- Credit: consumer credit, business lending...
- Equity: private equity, venture capital...



- ...but there's a curse to avoid!

VC game

- You manage a venture capital (VC) fund
  - A startup seeks financing
  - Potential deal: you receive a given share of the company's equity in exchange for cash payment P
- Valuation
  - Value of the company's shares when you cash out (V) is uncertain
  - Company risk is idiosyncratic and risk-free rate is zero
  - A priori estimate of present value = $E[V]$

VC game

- Information
  - You analyze the company to form a more precise estimate of V
  - Your best estimate is: check your private signal
  - Your best estimate is unbiased: I generated it as the true V (which I know but you don't) + a random noise with mean 0 and s.d. 20
- Competition
  - You are in competition with other VC funds (other students in the classroom) who also have their own best estimate of V
  - Each VC makes an offer to acquire the company's shares. The highest offer wins the deal

VC game

- Profits
  - If you win the deal at price P and the true company value is V, your profit (or loss if negative) is $V - P$
  - If you don't win the deal, your profit is zero
- Submit your bid at the link in your mailbox

Winner's curse

- Estimates are unbiased
  - Some are above, some are below, the average is V
  - Naturally, the highest estimate is above V
- If everyone bids their best estimate, the winning price is the highest among all best estimates $\rightarrow$ it is above V $\rightarrow$ the winner overpays
- This is the **winner's curse**

Winner's curse in...

- Lending/financing
  - Borrowers take your loan offer when other lenders don't want to lend to them
- Trading
  - See previous class
- Insurance
  - Customers buy insurance from you when other insurers estimate the risk is high and charge high premiums
  - ...or when customers know their risk is higher than you think

Winner's curse in...

- Real estate

Zillow: Machine learning and data disrupt real estate

Learn how big data and the Zillow Zestimate changed and disrupted real estate. It's an important case study on the power of machine learning models and digital innovation.

Written by **Michael Krigsman**, Contributor
Posted in **Beyond IT Failure** on July 30, 2017

Interview with Zillow's Chief Analytics Officer Stan Humphries in 2017

ZD: How accurate is the Zestimate?

S.H.: Our models are trained such that half of the Earth will be positive and half will be negative; meaning that on any given day, half of [all] homes are going to transact above the Zestimate value and half are going to transact below.

Winner's curse in. . .

- Real estate

Zillow: Machine learning and data disrupt real estate

Learn how big data and the Zillow Zestimate changed and disrupted real estate. It's an important case study on the power of machine learning models and digital innovation.

Written by **Michael Krigsman**, Contributor
Posted in **Beyond IT Failure** on July 30, 2017

Interview with Zillow's Chief Analytics Officer Stan Humphries in 2017

ZD: How accurate is the Zestimate?

S.H.: Our models are trained such that half of the Earth will be positive and half will be negative; meaning that on any given day, half of [all] homes are going to transact above the Zestimate value and half are going to transact below.

CHRIS STOKEL-WALKER BUSINESS NOV 11, 2021 :00 AM

WIRED

Why Zillow Couldn't Make Algorithmic House Pricing Work

Winner's curse

- Winner's curse is similar to adverse selection (cf. previous class)
- Ask yourself which info you DON'T have and others may have
- How to avoid the winner's curse?

1. Be more conservative than your best estimate (margin of safety, "bid shading")
2. Even better: run experiments
   - Experiment pricing on small random sample + backtest
   - Done routinely by tech firms; can be done in consumer credit markets

Backward-looking AI

- Another limitation of AI in a VC context: AI is backward-looking in nature whereas VC is about identifying novel ideas

"The Adoption of Artificial Intelligence by Venture Capitalists"
Maxime Bonelli (HEC PhD 2023) [pdf]

- Main result of research

  VC funds using AI to screen startups tend to select startups:

  a. that survive and receive follow-up funding ⇒ AI good at avoiding mistakes

  b. but that are less likely to file patents and IPO ⇒ AI less good at identify breakthrough ideas

Road map

Introduction

Financial inclusion

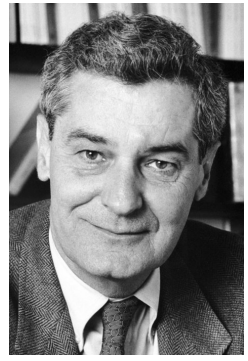
Winner's curse

Lucas critique

Hirshleifer effect

Discrimination

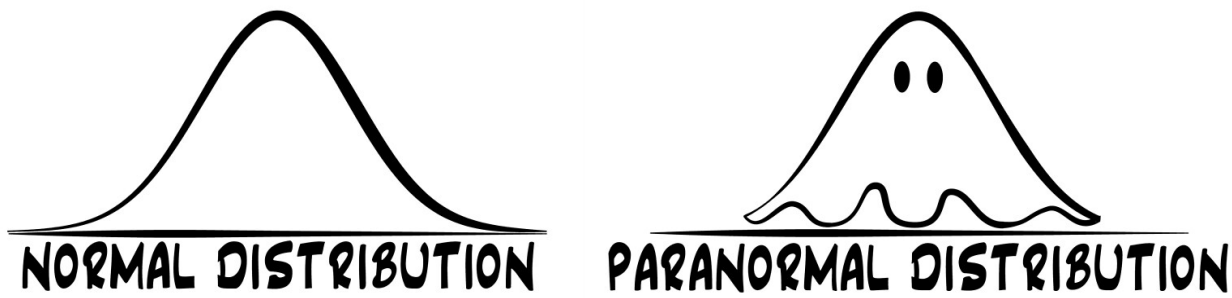
Lucas critique



Robert E. Lucas Jr.
(Nobel prize 1995)

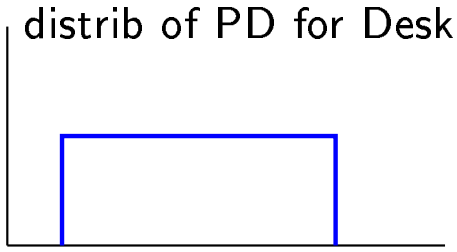
- Basic idea: People adapt their behavior to companies/governments' policies

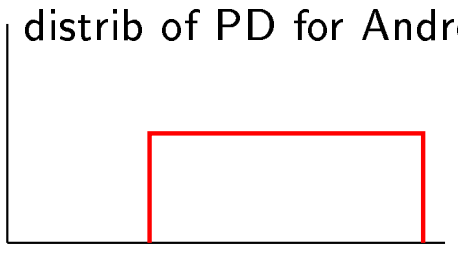
Alert: Statistics coming up!



Lucas critique

- Suppose lender observes whether borrowers connect from Android or Desktop and did not use this info in the past





- Lender starts setting lower rate for Desktop users and higher rates for Android users. Some Android users find out and switch to Desktop

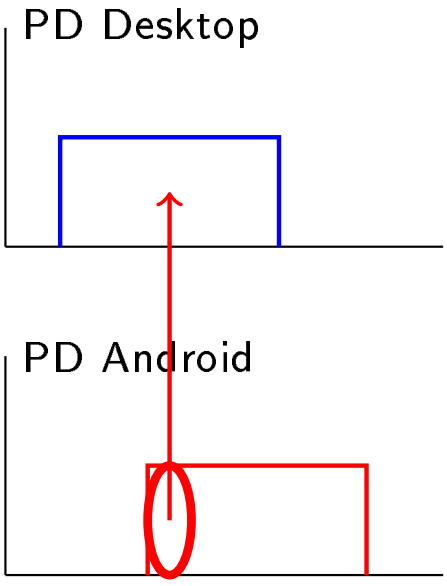
Q. How does this affect the average PD of Desktop users?

a. increases b. decreases c. may increase or decrease

Lucas critique

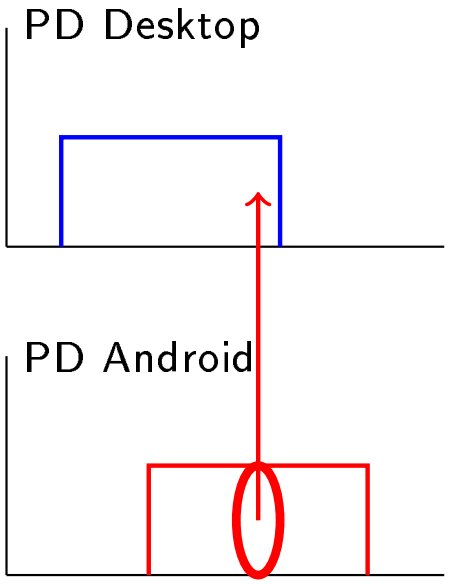
- It depends on which Android users switch to Desktop

If safe ones switch



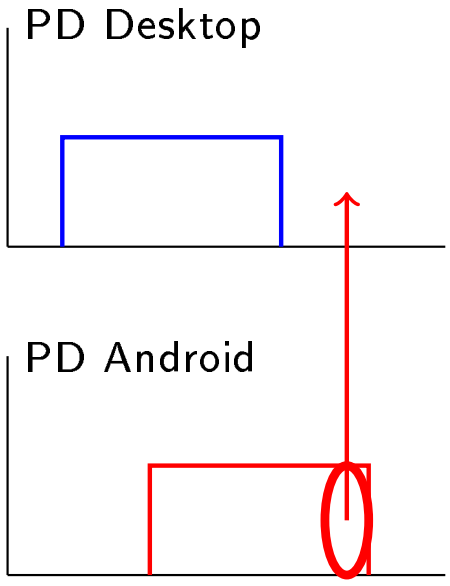
PD of Desktop users =
PD of Android users ↑

If intermediate ones switch



PD of Desktop users ↑
PD of Android users =

If risky ones switch

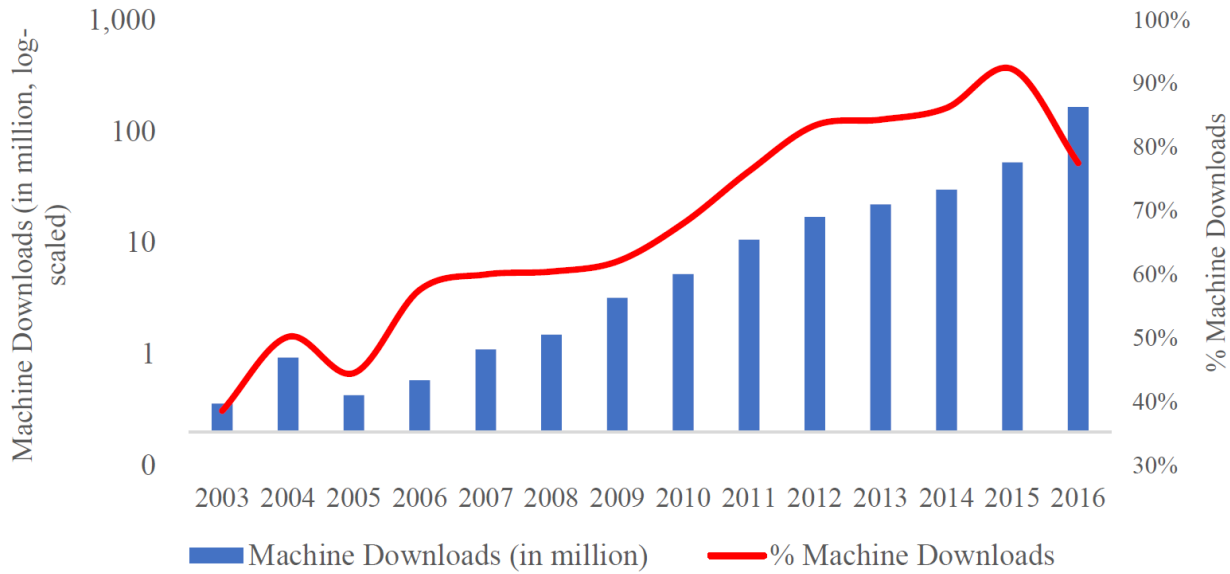


PD of Desktop users ↑
PD of Android users ↓

Case study: Corporate reporting[2]

- Companies' reports are heavily analyzed by AI

Fig.: Machine downloads of companies' filings with the SEC

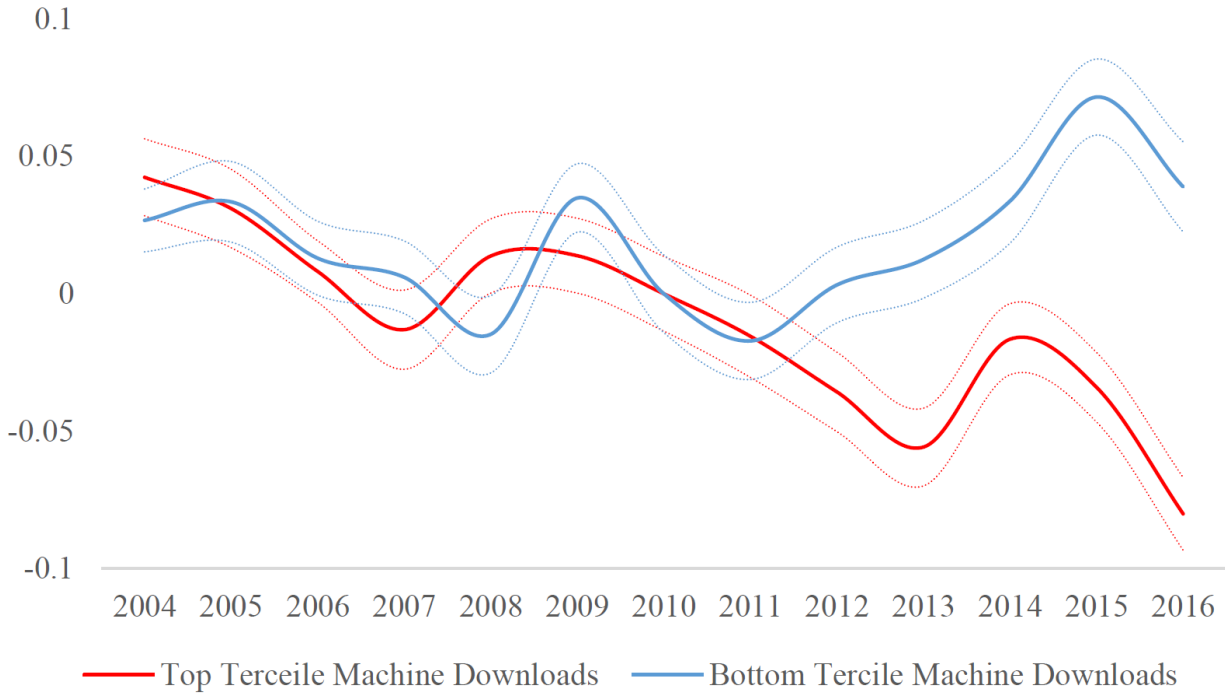


⇒ Companies have incentives to use language evaluated positively by textual analysis algorithms

[2] "How to Talk When a Machine is Listening: Corporate Disclosure in the Age of AI," Cao, Jiang, Yang & Zhang, 2020 [pdf]

Case study: Corporate reporting

Fig.: Use of words with negative connotation in the Loughran-McDonald dictionary (widely used by quant funds)



- Companies scrutinized by algos (in red) avoid words interpreted negatively by algorithms after the post-2010 rise of AI

Lucas critique: Implications

1. Beware scoring on behavior that can be strategically modified

   $\Rightarrow$ Ask yourself if data is exogenous or the outcome of a strategic choice

2. A good predictor when not used, can become a poor predictor once used (Goodhart's law)

   $\Rightarrow$ Check if the predictive power changes after data is used

Lucas critique: Implications

3. More data can make everyone (fintech AND consumers) worse off

   - Android mobile owners must connect from a desktop or buy a new mobile → they're worse off

     AND the predictive power of Android has disappeared → fintech is worse off

   - Companies avoid certain words in their filings, creating unnecessary complexity for everyone

Lucas critique: Implications

4. Commitment NOT TO use data can sometimes create value

   - Privacy can be a source of economic value
   - Customers' valuation for privacy can be estimated using A/B testing

"The Value of Privacy: Evidence from Online Borrowers"
Huan Tang (HEC PhD 2020) [pdf]

   - Credibility of commitment is key: once personal data exists, it is tempting for companies to use them
   - More on privacy:
     https://johanhombert.github.io/blog/20210418-privacy-paradox

Road map

Introduction

Financial inclusion

Winner's curse

Lucas critique

Hirshleifer effect

Discrimination

Information in insurance

- Suppose we discover how to predict perfectly who will get sick (but this foreknowledge does not help to prevent or cure diseases)

  Insurers use this information to price health insurance

Q1. Will this make people better or less-well insured?

a. better insured b. less-well insured

Q2. Will this make insurers more or less profitable?

a. more profitable b. less profitable

Information in insurance

- What will happen?
  - No insurers accept to insure people predicted to be sick
  - People predicted to be healthy don't need insurance

$\Rightarrow$ The health insurance market breaks down

  - People are worse off: they can't get insurance
  - Insurers are worse off: they can't sell insurance

- This is the Hirshleifer effect: Information can destroy insurance

How to overcome the Hirshleifer effect?

- Suppose an insurer announces it will not use the information (and suppose it is credible)

Q3. Does it overcome the Hirshleifer effect and allow the insurer to sell insurance?

a. yes b. no

How to overcome the Hirshleifer effect?

- Suppose an insurer announces it will not use the information (and suppose it is credible)

Q3. Does it overcome the Hirshleifer effect and allow the insurer to sell insurance?

a. yes b. no

- No, because of adverse selection
  - People predicted to be healthy are offered cheap insurance from other insurers (or they don't even buy insurance)
  - The insurer only gets people who will be sick, so it cannot insure them
- Problem 2 in problem set

How to overcome the Hirshleifer effect?

- Solution 1: Ensure no insurer uses the information
  - Industry self-regulation
  - Regulation (e.g., insurers banned from using genetic information)

How to overcome the Hirshleifer effect?

- Solution 2: Insure before information is revealed
  - Long-term insurance / Premium guaranteed over long period

Road map

Introduction

Financial inclusion

Winner's curse

Lucas critique

Hirshleifer effect

Discrimination

Discrimination

- Legal distinction between *direct* discrimination and *indirect* discrimination

Direct discrimination

- Treatment is based on a protected characteristic (sex, ethnicity, social origin, etc.)
  - Called "disparate treatment" in US law
- May happen for two reasons:

1. Outright prejudice
   - Ex.: job opening for white men only
2. Statistical discrimination: the protected characteristic is a predictor of risk
   - Ex.: cheaper car insurance for women because they have fewer accidents

- Illegal in both cases in EU and US

Indirect discrimination

- Treatment is not based on protected characteristics but ends up being different for people with a protected characteristic
  - Called "disparate impact" in US law
- Happens when treatment is based on variables correlated with protected characteristics
  - Ex.: interest rate based on borrower's job occupation may end up being different across people with different ethnic origins
- May be legal or illegal (e.g. legal for "business necessity" in US law)

Data and discrimination

Wells Fargo, Upstart criticized after study finds loan disparities

Published Feb. 6, 2020 • Updated Feb. 14, 2020

The request comes a week after the nonprofit Student Borrower Protection Center found that an Upstart borrower who attended historically black Howard University would pay thousands of dollars more on average for a five-year loan than a borrower with an identical credit profile who studied at New York University.

Discriminatory algorithms?

- Algos are (a priori) not subject to human prejudices but...

1. Biased data: algos are fed with human-world data, which may be contaminated by discrimination

   - May amplify or dampen human-world biases

     Ex. Suppose men and women make equally good entrepreneurs but biased venture capitalists set a higher bar for women[3]

     ⇒ In databases of VC-backed startups, the average female-founded company will be better than the average male-founded company

     ⇒ Algos fed with these data will "learn" that women are better entrepreneurs than men

[3] "Gender Stereotypes and Entrepreneur Financing," Hébert, 2020 [pdf]

Discriminatory algorithms?

2. Triangulation: algos may "triangulate" protected characteristics from other data (without intent to do so) and use them

- Solution
  - Algorithm interpretability: understand how algos make decisions
  - Ex.: feed algo with simulated data and analyze outcomes

Case study: Fintech lenders in the US mortgage market

- "Consumer Lending Discrimination in the FinTech Era," Bartlett et al., 2021, *Journal of Financial Economics* [pdf]

- For given borrower characteristics, Latin and African-American mortgage borrowers pay higher interest rates

  +8 basis points per year with traditional lenders

  +5 basis points per year with fintech lenders

Case study: Fintech lenders in the US mortgage market

- Half-full glass
  - Less discrimination by fintechs
  - Discrimination by traditional lenders has decreased over time, perhaps as a result of competition from fintechs
- Half-empty glass
  - Algos discriminate (although less so than humans)
  - Algos "learn" that minority borrowers are more likely to accept a high rate, because they are offered higher rates by traditional lenders