KAIST Summer Session 2018

Module 1. Research Design for Data Analytics

Computational Social Science

KAIST College of Business

Jiyong Park

16 July, 2018

What is Computational Social Science?

Computer Science + Social Science

- Computational social science is a new field that leverages the capacity to collect and analyze data with an unprecedented breadth and depth and scale.

SOCIAL SCIENCE

Computational Social Science

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M. and Jebara, T., 2009. Computational Social Science. *Science*, 323(5915), pp.721-723.

Computer Science + Social Data?

DOI:10.1145/3132698

Hanna Wallach

Senior Researcher at Microsoft Research
Adjunct Associate Professor at the University of Massachusetts Amherst

Viewpoint

Computational Social Science ≠ Computer Science + Social Data

The important intersection of computer science and social science.

- Computer scientist's viewpoint

  "In many prediction tasks, causality plays no role... we do not care why a model makes good predictions; we just care that it does."

- Social scientist's viewpoint

  "Explanation tasks are fundamentally concerned with causality. Here, the goal is to use observed data to provide evidence in support or opposition of causal explanations."

Wallach, H., 2018. Computational Social Science ≠ Computer Science + Social Data. *Communications of the ACM*, 61(3), pp.42-44.

Computer Science + Social Data?


SYMPOSIUM

We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together

Justin Grimmer, *Stanford University*

- Social scientist's viewpoint

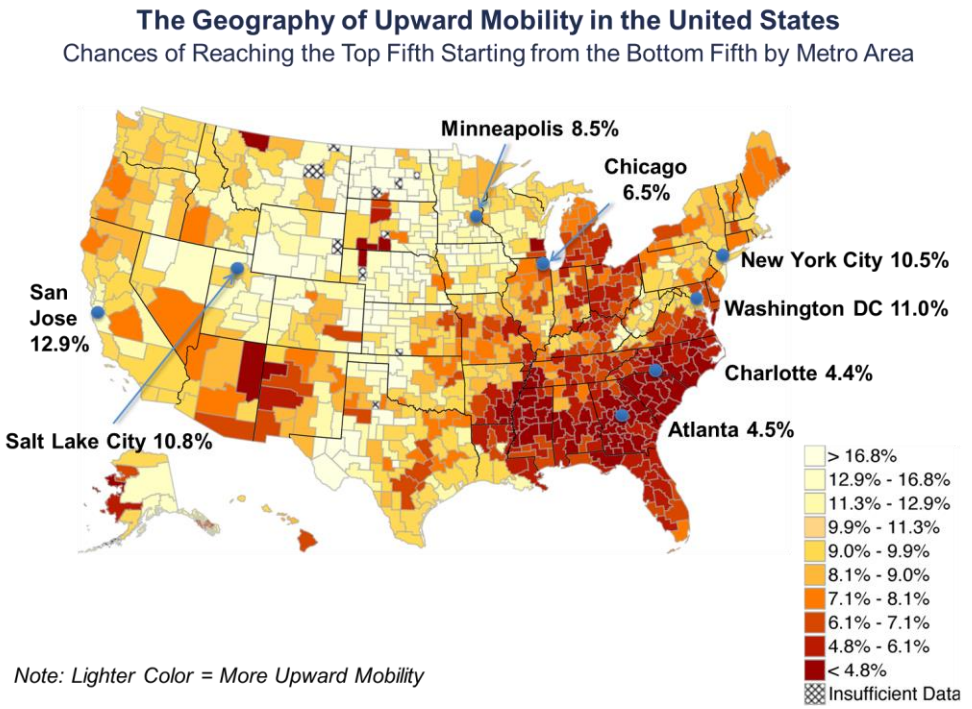
  "Of course, social scientists know that large amounts of data will not overcome the selection problems that make causal inference so difficult."

  "Big data alone is insufficient to make valid causal inferences; however, having more data certainly can improve causal inferences in laree-scale datasets."

Grimmer, J., 2015. We are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together. *PS: Political Science & Politics*, 48(1), pp.80-83.

Computational Technique + (Big Data) + Social Science!

1890 US Census



Big data of tax records enables to zoom in the entire population. (e.g., upward mobility; Chetty et al. 2017)

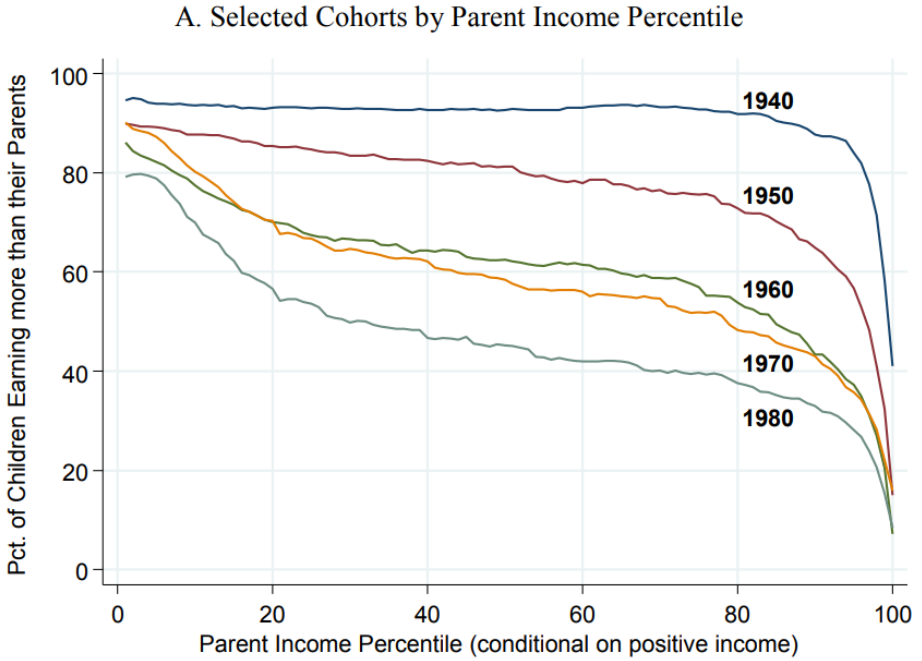
https://en.wikipedia.org/wiki/1890_United_States_Census

Chetty, R., Grusky, D., Hell, M., Hendren, N., Manduca, R. and Narang, J., 2017. The Fading American Dream: Trends in Absolute Income Mobility Since 1940. Science, 356(6336), pp.398-406.

Chetty, R. and Hendren, N., 2018. The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects. *Quarterly Journal of Economics*, forthcoming

Computational Technique + (Big Data) + Social Science!

1890 US Census



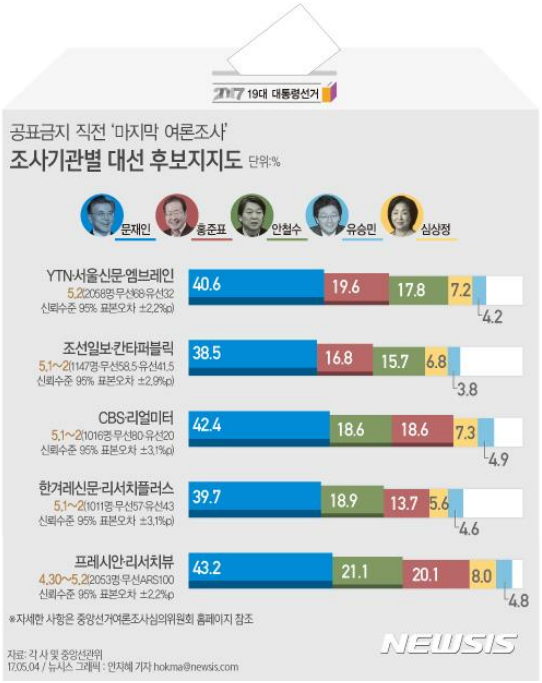
Big data of tax records enables to zoom in the entire population. (e.g., upward mobility; Chetty et al. 2017)

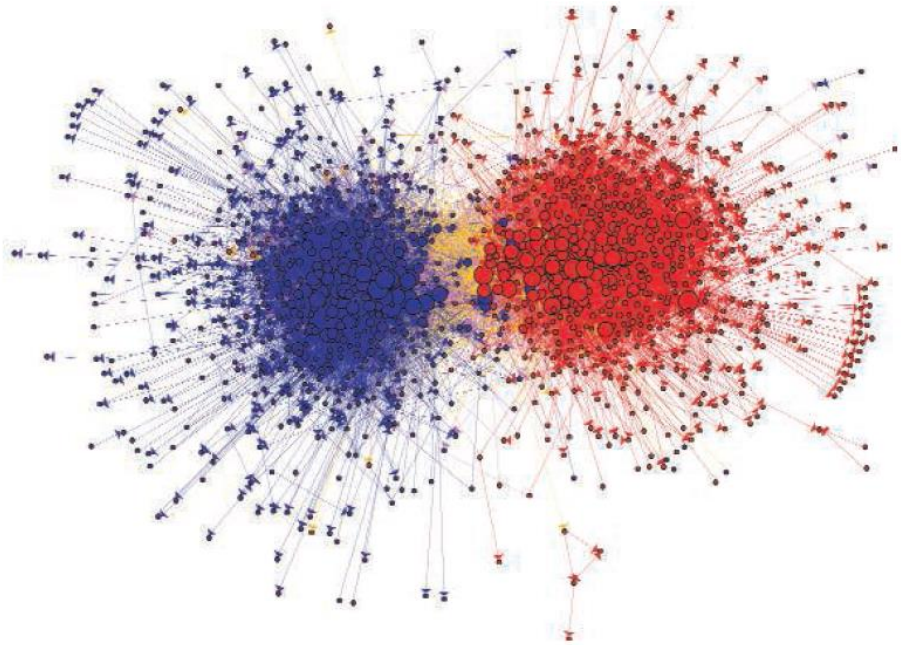
https://en.wikipedia.org/wiki/1890_United_States_Census

Chetty, R., Grusky, D., Hell, M., Hendren, N., Manduca, R. and Narang, J., 2017. The Fading American Dream: Trends in Absolute Income Mobility Since 1940. Science, 356(6336), pp.398-406.

Chetty, R. and Hendren, N., 2018. The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects. *Quarterly Journal of Economics*, forthcoming

Computational Technique + (Big Data) + Social Science!





Data from the blogosphere. Shown is a link structure within a community of political blogs (from 2004), where red nodes indicate conservative blogs, and blue liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it. [Reproduced from (*8*) with permission from the Association for Computing Machinery]

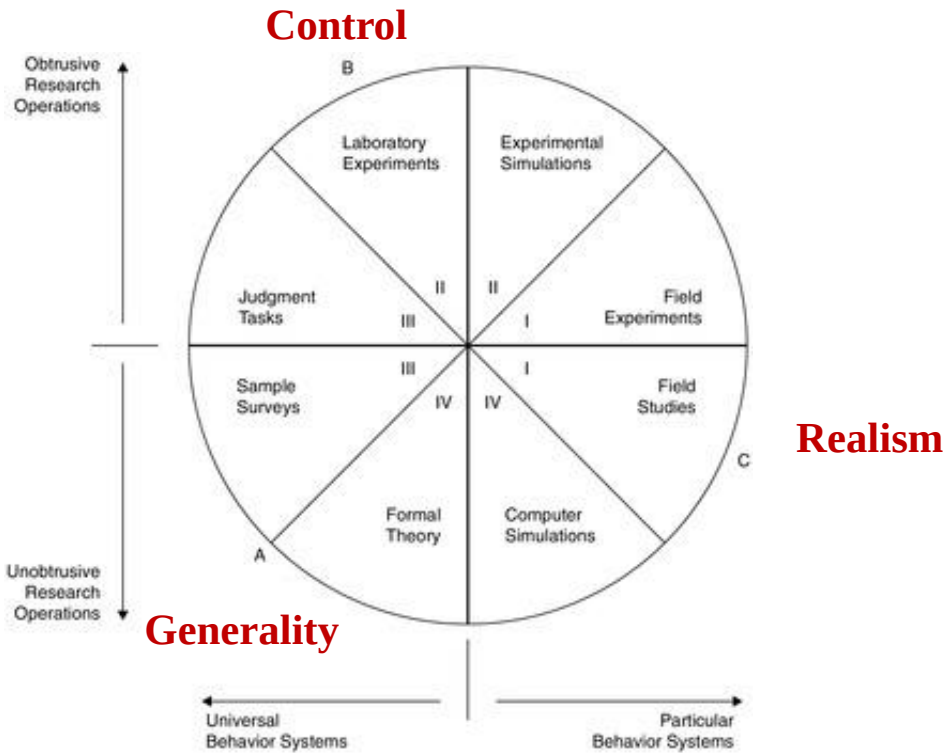
Big data of social networks enables to capture the opinions of larger population in a real-time base. Furthermore, it can reveal the network structure of political opinions.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M. and Jebara, T., 2009. Computational Social Science. *Science*, 323(5915), pp.721-723.

How Do Big Data and Machine Learning Transform Social Science?

Three-Horned Dilemma for Research Methods

- Three goals of research methods: generality, control, and realism
  - Runkel and McGrath (1972) argue that these goals cannot be maximized simultaneously with any single method. *(it has not been able to… but now?)*

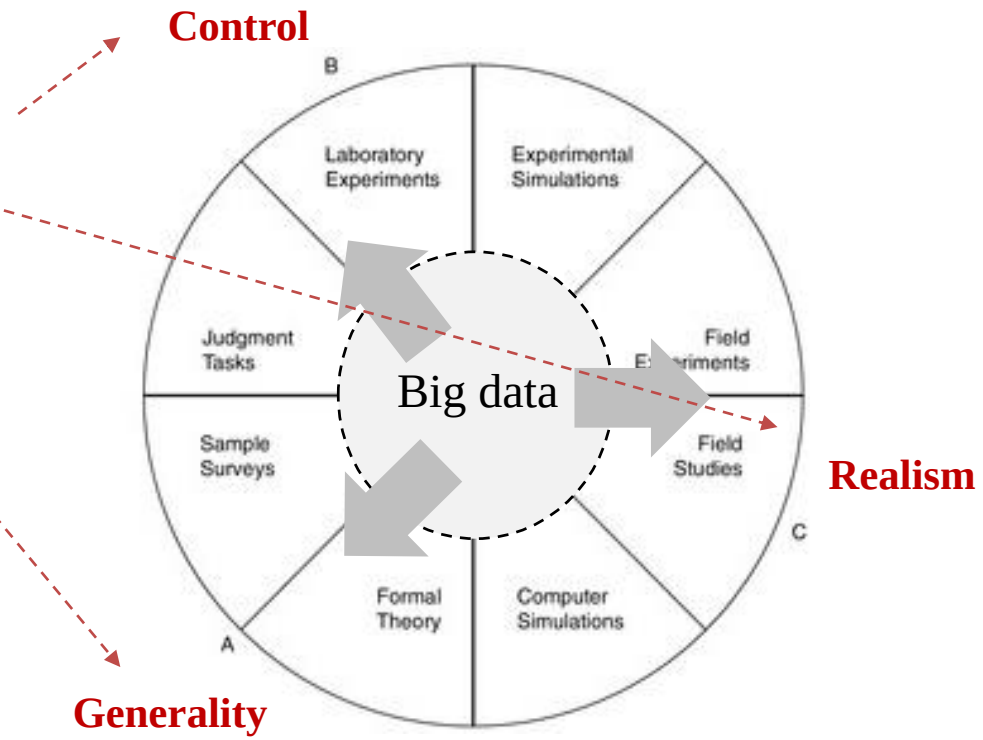


Runkel, P.J. and McGrath, J.E., 1972. Research on Human Behavior: A Systematic Guide to Method. Holt, Rinehart & Winston of Canada Ltd.

(1) Reconciling Internal and External Validities

- Big data and online platforms allow researchers to achieve generality, control, and realism at the same time, to a certain degree.

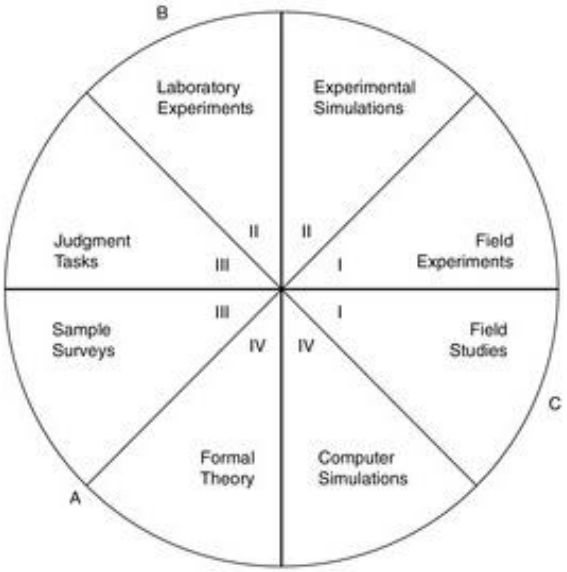
  - (Example) Aral and Walker (2012) conducted a randomized field experiment on Facebook for a representative sample of 1.3 million users, in order to investigate peer influence on social networks.



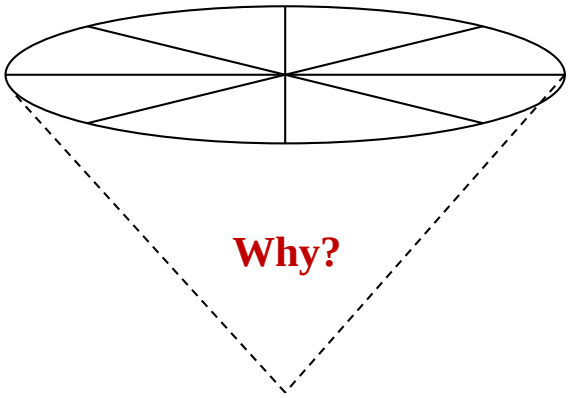
Aral, S. and Walker, D., 2012. Identifying Influential and Susceptible Members of Social Networks. *Science 337*, p. 337-341

(2) Zooming in and Digging into the Phenomena

- Big data and machine learning allow researchers to zoom in and look into the underlying mechanisms behind the observed patterns.
  - ➢ Building on a theoretical background, big data enables to tease out the underlying mechanisms.

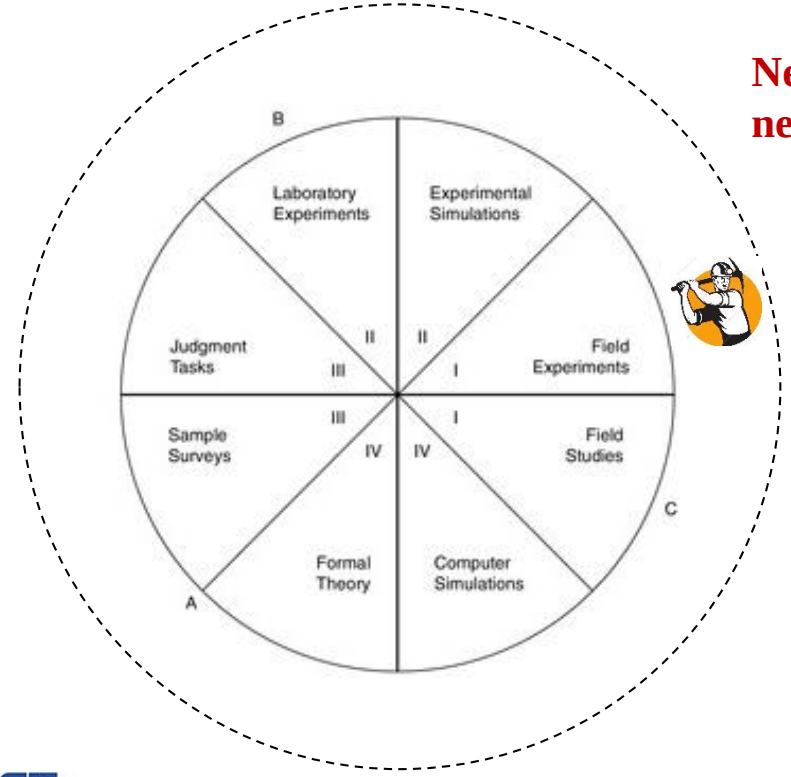


Phenomenon of interest



(3) New Ingredients for Empirical Research

- Big data and machine learning broaden the scope of enquiry in empirical research, by quantifying unstructured data which remains largely unexplored due to difficulty of coding on a large scale.



New phenomena and new causal factors

Econometrics + Data Mining

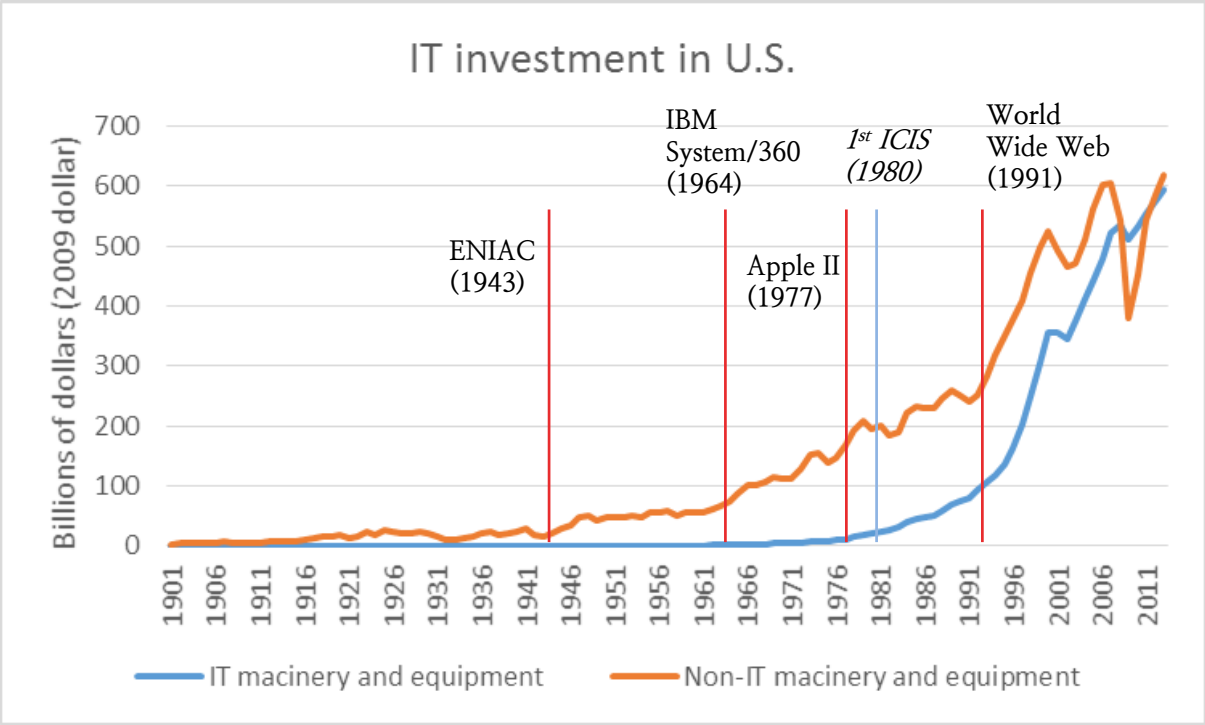
= **Economining**

This topic will be covered in the next class

Example: IT Productivity

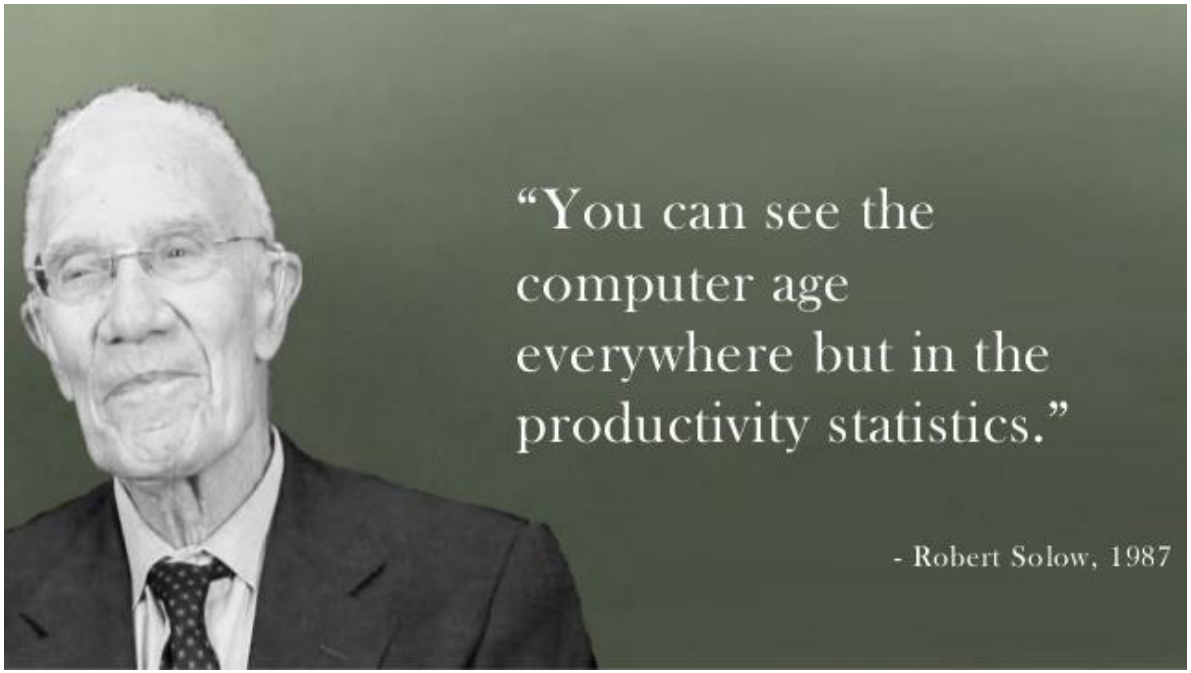
Trend in IT Investment

- IT investment has begun since the 1980s, and soared since the mid-1990s.
  - The 1st International Conference on Information Systems (ICIS) was held in 1980.



Productivity Paradox (Solow Paradox)

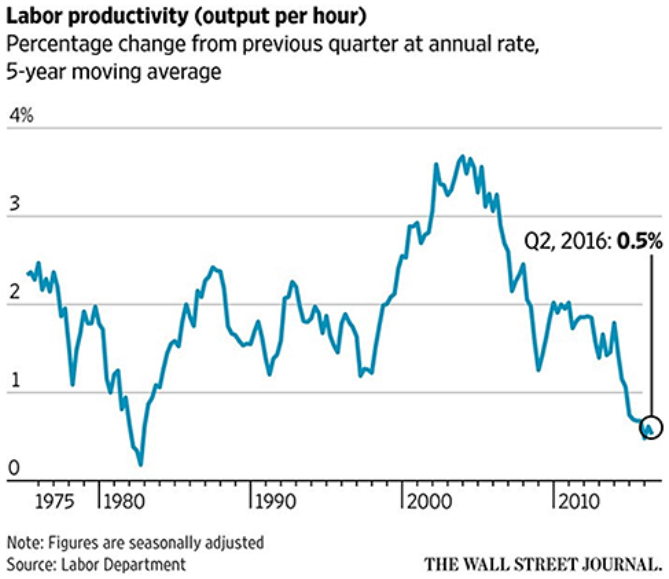
- Earlier studies reported that investments in IT showed no net contribution to total output (Loveman 1994).

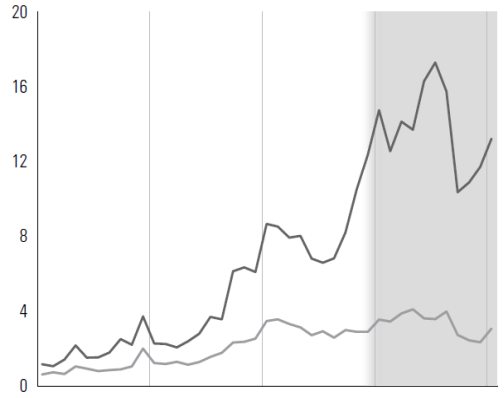


Loveman, G.W., 1994. An Assessment of the Productivity Impact of Information Technologies. In *Information Technology and the Corporation of the 1990s: Research Studies*, MIT Press

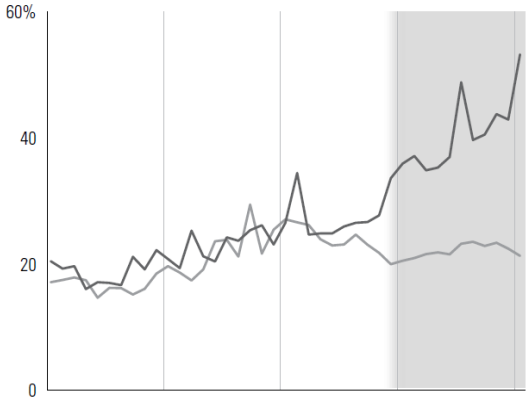
The Advent of New Economy (1995-2005)

- However, the productivity paradox seems to contradict the coincide between rapid growth of IT investment and labor productivity since the mid-1990s.
  - ➢ Industries have experienced the change of competitive dynamics, though research does not find the evidence of IT productivity.





Turbulence



Performance Spread

McAfee, A. and Brynjolfsson, E., 2008. Investing in the IT that Makes a Competitive Difference. *Harvard Business Review*, 86(7/8), p.98.

Evidence on IT Productivity (Brynjolfsson and Hitt 1996)

- "Shortfall of evidence is not necessarily evidence of a shortfall."
- Four explanations for the paradox (Brynjolfsson 1993)
  - Mismeasurement of outputs (e.g., product variety and quality)
  - Lags due to learning and adjustment
  - Redistribution and dissipation of profits
  - Mismanagement of IT
- Brynjolfsson and Hitt (1996) find the evidence of positive returns to IT investment using a new dataset.
  - Later time period (1987-1991)
  - Different and more detailed firm-level data (Fortune 500 firms)

Brynjolfsson, E., 1993. The Productivity Paradox of Information Technology. *Communications of the ACM*, 36(12), pp.66-77.
Brynjolfsson, E. and Hitt, L., 1996. Paradox Lost? Firm-Level Evidence on the Returns to Information Systems Spending. *Management Science*, 42(4), pp.541-558.

Ensuring Internal Validity (Tambe and Hitt 2012)

- Long-standing empirical limitations in the IT productivity literature
  - "A persistent concern in the IT value literature has been establishing how much of the excess rate of return observe for IT investment is because of reverse causality or the endogenetiy of IT investment." (Tambe and Hitt 2012, p. 599)

- How did Tambe and Hitt (2012) address this endogeneity issue to ensure the internal validity?
  - Using more comprehensive, long panel data (about 600,000 IT workers data)
  - Employing advanced econometric techniques for panel data (system GMM and Levinsohn-Petrin GMM)

  : Panel data is mostly preferred to cross-sectional data for addressing endogeneity.

Tambe, P. and Hitt, L., 2012. The Productivity of Information Technology Investments: New Evidence from IT Labor Data. *Information Systems Research*, 23(3-part-1), pp.599-617.

Diving into IT Productivity: Job Hopping (Tambe and Hitt 2013)

- Where does IT productivity come from?
  - "The primary goal of this study is to test the hypothesis that firms benefit from the IT investments of other firms because the flow of specialized technical know-how among organizations facilitates the implementation of new IT innovation" (Tambe and Hitt 2013, p. 339)
  - "Identification and measurement of the economic impact of these spillovers of technical know-how has implications for understanding productivity differences and heterogeneity in IT returns." (Tambe and Hitt 2013, p. 338)

- How did Tambe and Hitt (2013) investigate the underlying mechanism?
  - Using more comprehensive, long panel data (about 10 million users who posted or modified their career histories online)

Tambe, P. and Hitt, L., 2013. Job Hopping, Information Technology Spillovers, and Productivity Growth. *Management Science*, 60(2), pp.338-355.

Diving into IT Productivity: Big Data Investment (Tambe 2014)

- Where does IT productivity come from?
  - "Differences in the supply of workers with the skills complementary to the new information technologies… may explain differences in the rates at which firms in different labor markets are able to unlock value from new IT innovations." (Tambe 2014, p. 1452)
  - "This paper examines how labor markets have shaped early returns on investment in a key big data technology—Hadoop-based systems." (Tambe 2014, p. 1453)

- How did Tambe (2014) investigate the underlying mechanism?
  - Using a new data source including technical skill decsriptions (about 175 million LinkedIn users)

Tambe, P., 2014. Big Data Investment, Skills, and Firm Value. *Management Science*, 60(6), pp.1452-1469.

Example: Word-of-Mouth and Social Network

Word-of-Mouth

- One of the most widely accepted notions in consumer behavior is that word-of-mouth (WOM) communication plays an important role in shaping consumers' attitudes and behaviors.

- Electronic word-of-mouth (EWOM)
  - WOM has been applied to online or technology-enabled information exchange between individuals (e.g., social media).

Social Ties and Word-of-Mouth (Brown and Reingen 1987)

- Brown and Reingen (1987) examine which interpersonal ties are more influential in consumers' decision-making.

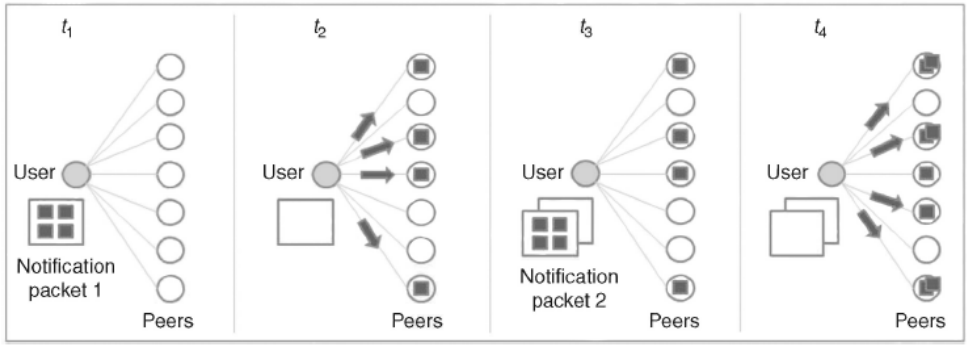
- Interview-based data collection
  - "The respondents (118 individuals) were asked how they first learned about their piano teacher. When a subject mentioned another person, an address and telephone number were requested… These mentioned individuals where then notified by mail and telephoned by the same interviewer." (Brown and Reingen 1987, p. 355)
  - "With regard to methodology, a limitation is that serveral WOM paths could not be completed due to non-response or incomplete retrieval of WOM instances from memory." (Brown and Reingen 1987, p. 361)

Brown, J.J. and Reingen, P.H., 1987. Social Ties and Word-of-Mouth Referral Behavior. *Journal of Consumer Research*, 14(3), pp.350-362.

Identifying (Causal) Peer Influence (Aral and Walker 2014)

- Empirical challenges in identifying causal peer influence
  - Statistical challenges (from endogeneity) make it difficult to distinguish causal peer influence from other confounds that create behavioral clustering in network.
  - Aral and Walker (2014) conduct vivo randomized experiments (1.3 million users) to examine the role of tie strength and embeddedness in causal peer influence.

Figure 1 **Randomized Targeting of Influence-Mediating Messages**



Notes. A diagram depicting the message target randomization employed in the experiment is shown. Notification packets are generated when an application user takes a packet-generating action within the Facebook application. For each packet that is generated, the notifications in the packet are distributed to a randomly selected subset of the application user's peers. The figure displays two sequential packet distributions. Different recipient targets are randomly chosen at the time of distribution for each packet.

Aral, S. and Walker, D., 2014. Tie Strength, Embeddedness, and Social Influence: A Large-Scale Networked Experiment. *Management Science*, 60(6), pp.1352-1370.

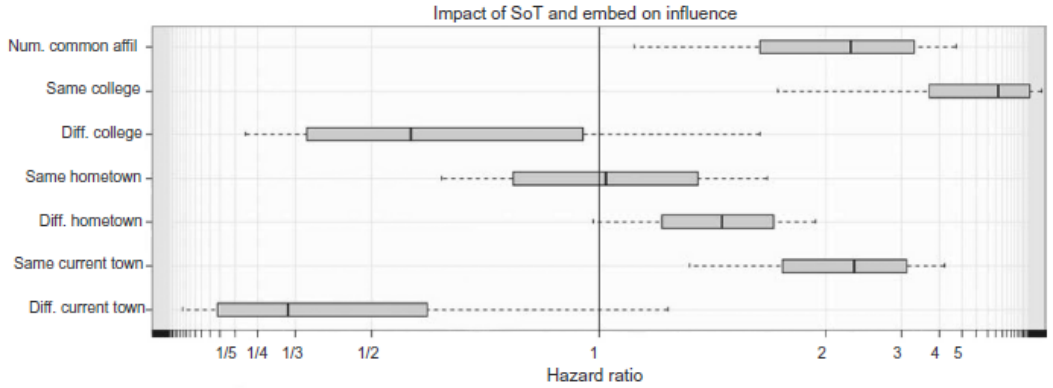
Identifying (Causal) Peer Influence (Aral and Walker 2014)

- The power of big data
  - ➢ "The real power lies in the granularity of the data (not just its scale) combined with a new ability to engineer and randomize social settings to…" (Aral and Walker 2014, p. 1353)

    (a) robustly estimate the causal effects of different policy alternatives

    (b) explore the heterogeneity in these causal effects across subpopulations

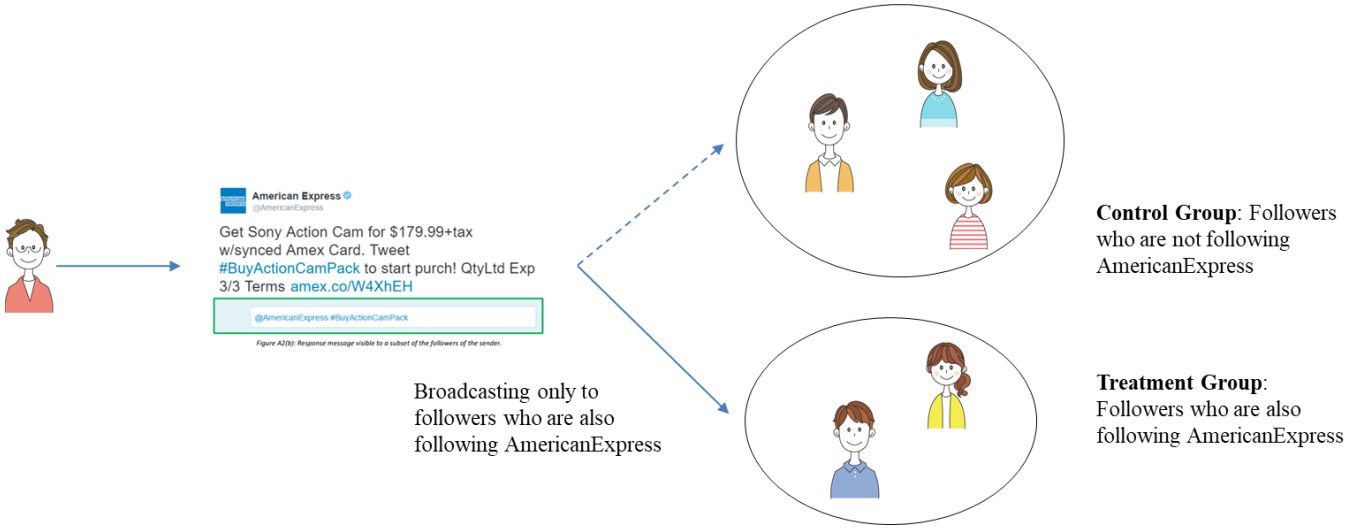
    (c) unpack the nuanced mechanisms that underlie the causal outcomes



Aral, S. and Walker, D., 2014. Tie Strength, Embeddedness, and Social Influence: A Large-Scale Networked Experiment. *Management Science*, 60(6), pp.1352-1370.

Personality Similarity and Word-of-Mouth (Adamopoulos et al. 2018)

- Heterogeneity in word-of-mouth (WOM) effectiveness
  - ➢ Do personality traits of online users affect the WOM in social media?
  - ➢ Using "a novel combination of text-mining techniques with econometric methods and a quasi-experiment" (200,000 subjects), Adamopoulos et al. (2018) find that exposure to WOM messages from similar users in terms of personality, rather than dissimilar users, increases the likelihood of a post-purchase.



Adamopoulos, P., Ghose, A. and Todri, V., 2018. The Impact of User Personality Traits on Word of Mouth: Text-Mining Social Media Platforms. *Information Systems Research*. forthcoming

Personality Similarity and Word-of-Mouth (Adamopoulos et al. 2018)

- The power of deep learning and machine learning
  - "Thanks to the recent advances in deep learning and machine learning, both firms and researchers have the unique opportunity to leverage the abundance of unstructured data in social media to identify users' latent characteristics and traits that can impact the effectiveness of WOM" (Adamopoulos et al. 2018, p. 2).

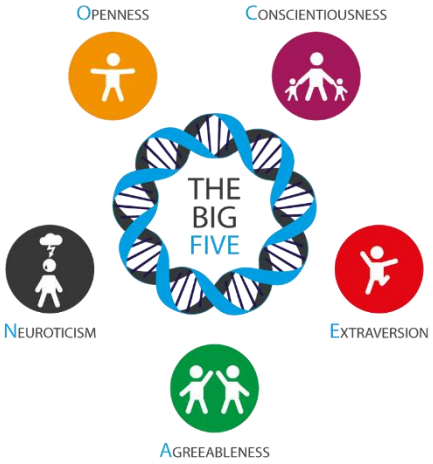
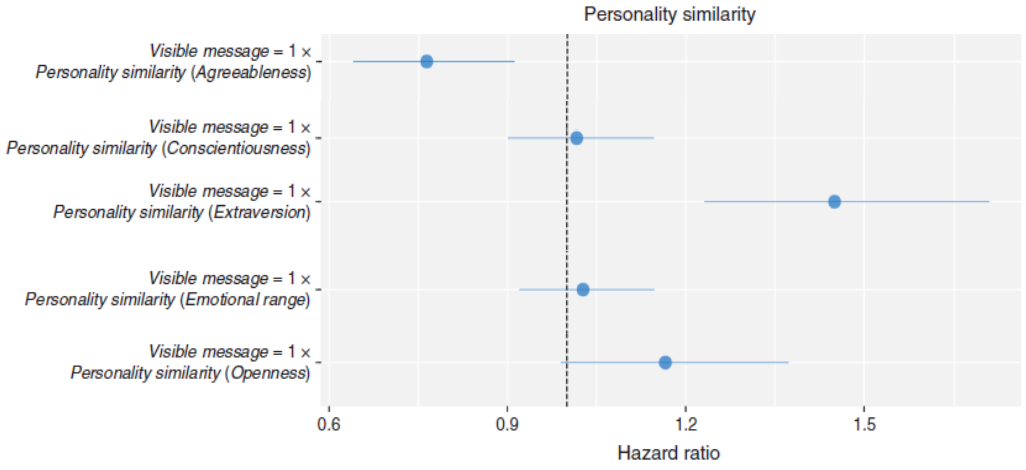


Figure 1. (Color online) Effects of Similarity of User Personality on Dyadic WOM Effectiveness

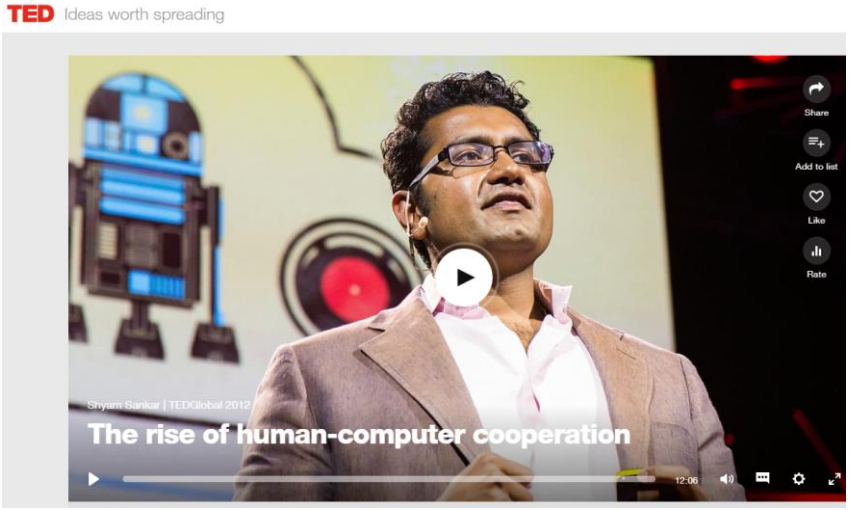


Adamopoulos, P., Ghose, A. and Todri, V., 2018. The Impact of User Personality Traits on Word of Mouth: Text-Mining Social Media Platforms. *Information Systems Research*. forthcoming

Conclusion

Let's Cooperate with Machine (Learning)

- Analytic capability consists of human, computer, and human-computer cooperation.



https://www.ted.com/talks/shyam_sankar_the_rise_of_human_computer_cooperation

"In a freestyle chess tournament in 2005… The surprise came at the end. Who won? Not a grandmaster with a supercomputer, but actually two American amateurs using three relatively weak laptops …

This is an astonishing result: average men, average machines beating the best man, the best machine. And anyways, isn't it supposed to be man versus machine? **Instead, it's about cooperation, and the right type of cooperation.**"

- Shyam Sankar *at* TEDGlobal 2012

End of Document