KAIST Summer Session 2018

Module 1. Research Design for Data Analytics

Economining

KAIST College of Business

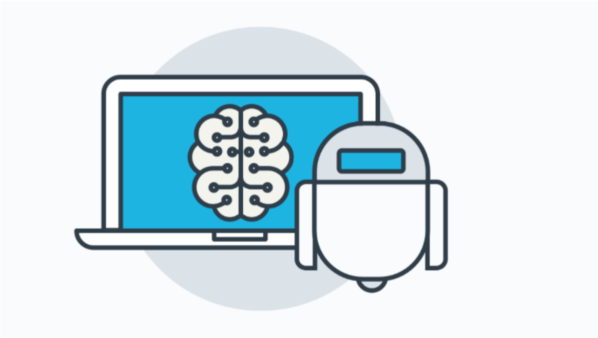
Jiyong Park

19 July, 2018

Machine Learning and Empirical Research

Empirical Research in the Age of Big Data and Machine Learning

- Big data and machine learning will revolutionize how we do research.



Machine Learning

Econometrics

"In the long run, new empirical tools have also served to expand the kinds of problems we work on. The increased use of randomized control trials has also changed the kinds of questions empirical researchers work on. Ultimately, machine learning tools may also increase the scope of our work – **not just by delivering new data or new methods but by focusing us on new questions**." (Mullainathan and Spiess 2017, p. 104)

Mullainathan, S. and Spiess, J., 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), pp.87-106.

How Can Social Science Leverage Machine Learning?

SYMPOSIUM

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

Justin Grimmer, *Stanford University*

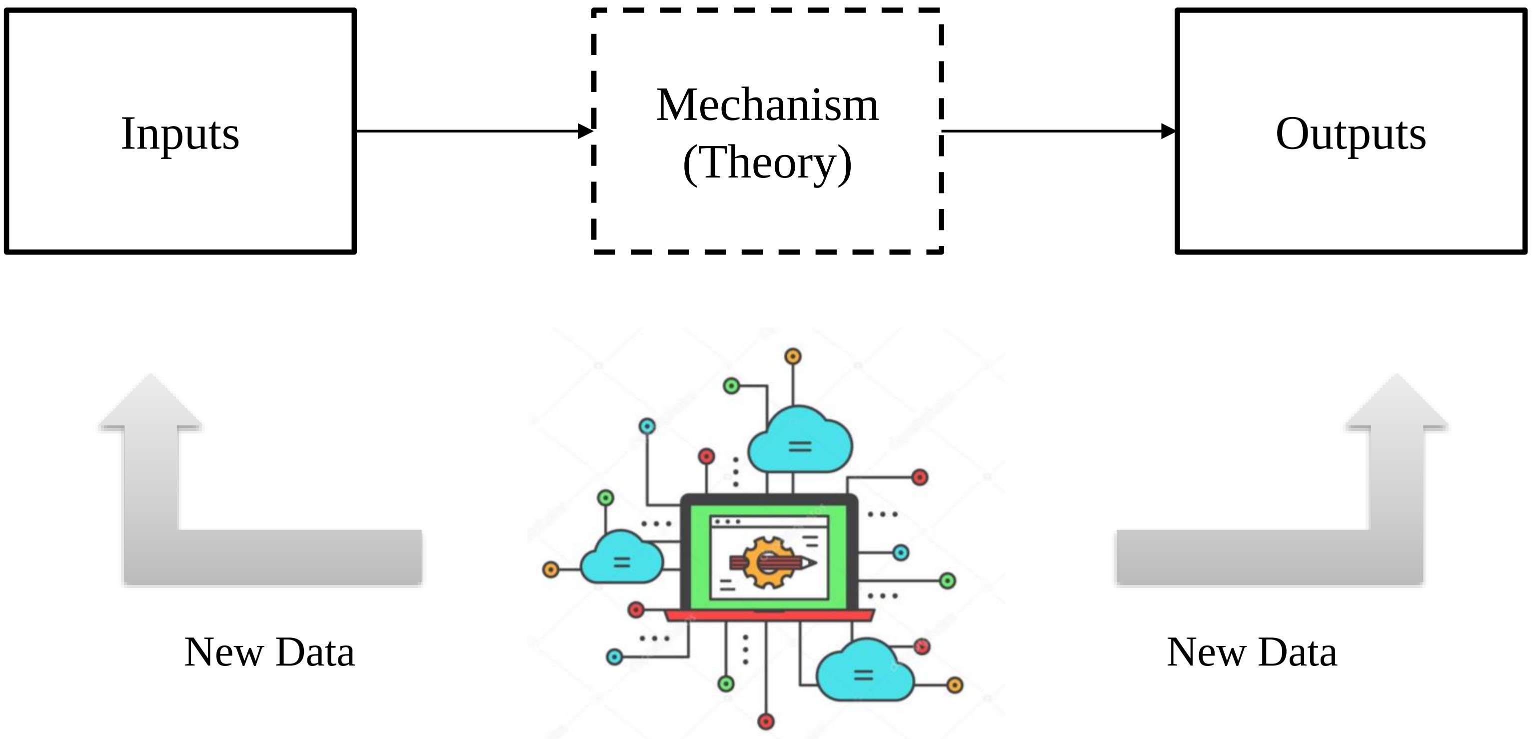
"Social scientists typically use machine-learning techniques to measure a certain characteristics or latent quantity in the world — a qualitatively different goal than computer scientists, who use the measures for prediction."

"…example of how the analysis of big data is best viewed as a subfield of the social sciences."

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.

Economining = Econometrics + Data Mining

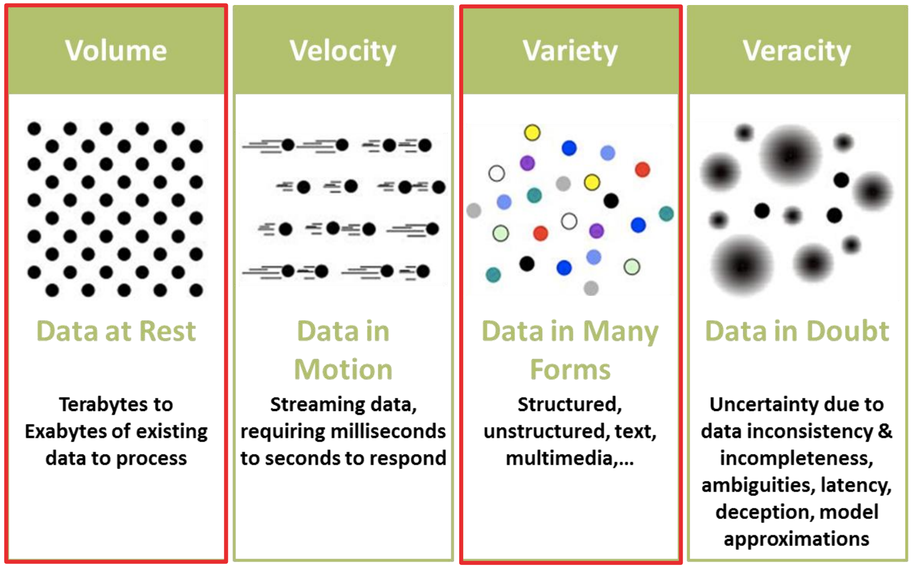
- Data itself is a central ingredient in the "empirical" research.



Data Mining
Machine Learning

Machine Learning and Big Data

- Machine learning and deep learning enable researchers to investigate a variety of data on a large scale.
  - ➢ "Considering each of the four Vs of Big Data characteristics, i.e., Volume, Variety, Velocity, and Veracity, **deep learning** algorithms and architecture are more aptly suited to address issues related to **Volume and Variety of big data analytics.**" (Najafabadi et al. 2015, p. 8)



Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R. and Muharemagic, E., 2015. Deep Learning Applications and Challenges in Big Data Analytics. *Journal of Big Data*, 2(1), p.1.

Steps for Economining

- First, define the constructs you are theoretically interested in.
  - Different researchers might extract different information from same data.

- Second, build a labeled dataset (only for supervised learning).
  - Supervised learning requires labeled training data (e.g., positive vs. negative).

- Third, implement the predictive modeling.
  - Researchers should determine which technique is well-suited for their research purpose (e.g., neural networks vs. support vector machine vs. random forest).

- Finally, conduct the empirical research using new data.
  - Let's return to the basics – what's your research goal?

Identification Strategy Still Matters

- Even if machine learning is used to measure new input variables, identification strategy for causal inference is still important.

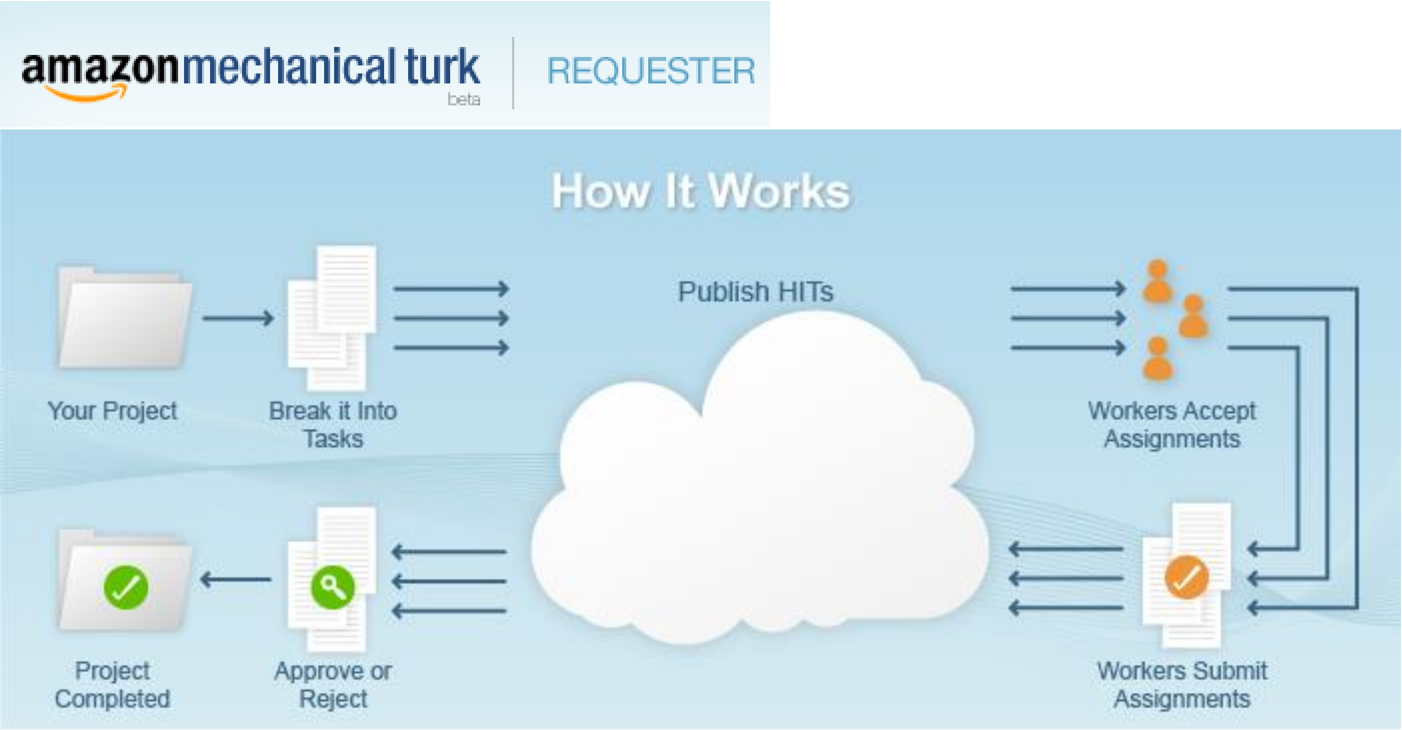
- A review I received for Park et al. (2018)
  - “I read the paper with great interest, since video is a big "black box" in crowdfunding that is probably not sufficiently looked at, or sufficiently controlled for, in the literature.”
  - “From our perspective, the biggest concern at the moment is that the authors ultimately are demonstrating that the mined covariates are predictive of fundraising success; **they do not necessarily have a causal impact**. For example, it may be the case that speech patterns are correlated with other things that have been shown to influence campaign success; e.g., race, geography, etc.”

Park, J., Kim, J., Cho, D. and Lee, B., 2018. Pitching with Style: The Role of the Entrepreneur's Video Pitch on Online Crowdfunding. *KAIST Working Paper.*

Building a Labeled Dataset

How to Label Your Data

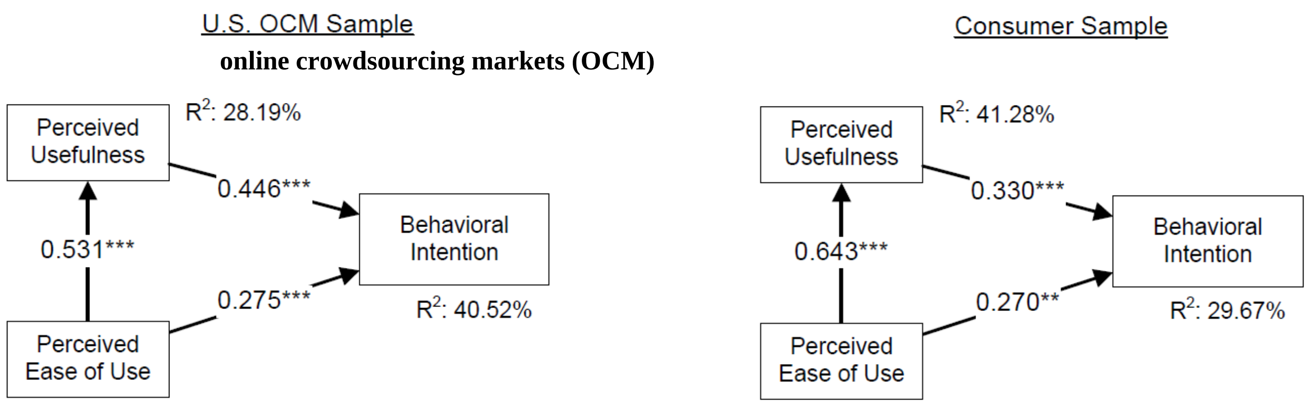
- Amazon Mechanical Turk (Mturk) has been very popular in academic research for the goal of data labeling or tagging.



https://requester.mturk.com/

Comparison of Mturk Sample and Consumer Panel

- Mturk samples have been suggested to lead to similar statistical conclusions as both students and consumer panels at a considerably lower cost.
  - ➢ (Example) Technology Acceptance Model (TAM) (Steelman et al. 2014)



- MTurk shares similar features with online crowdfunding platforms, such as crowd-based organizing, enhancing the external validity of the experiments

Steelman, Z.R., Hammer, B.I. and Limayem, M., 2014. Data Collection in the Digital Age: Innovative Alternatives to Student Samples. *MIS Quarterly*, 38(2), pp.355-378.

Mturk Enables to Conduct a Large-Scale Survey

- Comparing with traditional surveys, researchers can recruit survey participants easier and cheaper, enabling to conduct a large-scale survey.
  - ➢ (Example) Park et al. (2018) recruited about 1,300 participants for online experiments of 2 (two types of pitch style) × 2 (with/without a video pitch) × 2 (with/without project updates) between-subjects design.

Table 4. Partial Least Squares Results (Structural Model)

Independent variable → *Dependent variable*	Analytical and Conscientious Style (Pitch A)				Open-to-Experience and Extroverted Style (Pitch O)			
	With Updates		Without Updates		With Updates		Without Updates	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Pitch Treatment → PC		0.036 (0.406)		0.104*** (0.009)		-0.024 (0.576)		-0.044 (0.315)
PEP → PC	0.648*** (0.000)	0.651*** (0.000)	0.696*** (0.000)	0.701*** (0.000)	0.617*** (0.000)	0.616*** (0.000)	0.686*** (0.000)	0.684*** (0.000)
Pitch Treatment → PEP		-0.076 (0.189)		-0.046 (0.441)		-0.011 (0.839)		-0.050 (0.390)
Pitch Treatment → WtB		-0.126 (0.764)		-0.187 (0.544)		-0.425* (0.056)		-0.189 (0.552)
PC → WtB	0.175*** (0.005)	0.144 (0.109)	0.288*** (0.000)	0.321*** (0.000)	0.214*** (0.000)	0.231*** (0.005)	0.233*** (0.000)	0.210** (0.033)
PEP → WtB	0.241*** (0.000)	0.240** (0.023)	0.098 (0.188)	0.009 (0.929)	0.347*** (0.000)	0.209*** (0.009)	0.278*** (0.000)	0.255** (0.016)
PC x Treatment → WtB		0.203 (0.547)		-0.252 (0.534)		-0.064 (0.810)		0.057 (0.868)
PEP x Treatment → WtB		-0.154 (0.817)		0.520 (0.246)		0.617** (0.016)		0.078 (0..793)
Observations	306	306	297	297	327	327	291	291

Notes: After 1,000 replications of bootstrapping, p-values are in parentheses; WtB, PC, and PEP denote Willingness-to-Back, Perceived Competence, and Perceived Entrepreneurial Passion, respectively; * p<0.1, ** p<0.05, *** p<0.01

Park, J., Kim, J., Cho, D. and Lee, B., 2018. Pitching with Style: The Role of the Entrepreneur's Video Pitch on Online Crowdfunding. *KAIST Working Paper.*

Quality of Mturk Language Annotation (Labeling)

- High agreement between Mturk non-expert annotations and existing gold standard labels provided by expert labelers
  - For an affect recognition task, Snow et al. (2008) suggest that it requires an average of 4 non-expert labels per item in order to emulate expert-level label quality.

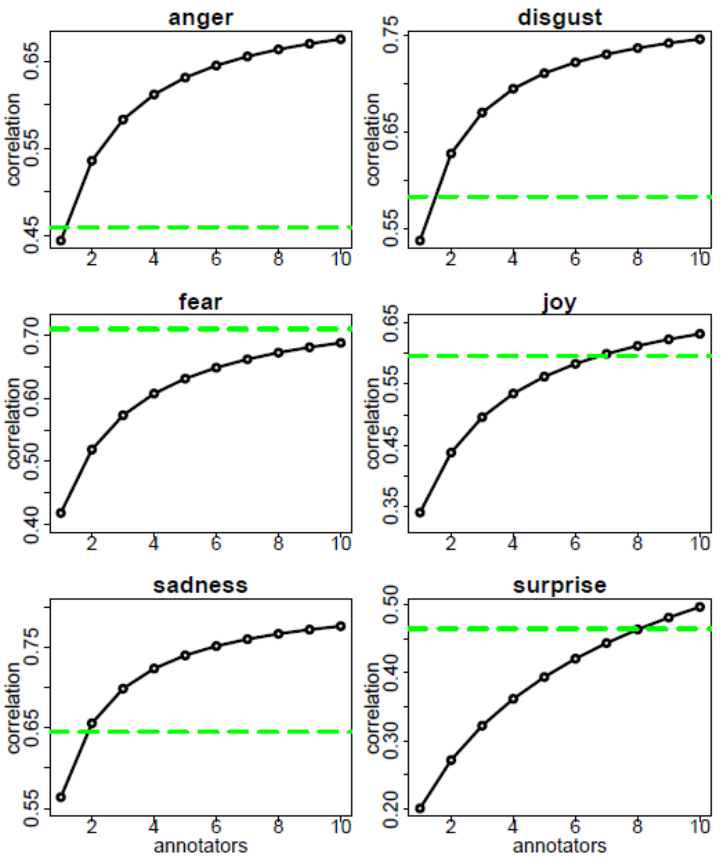
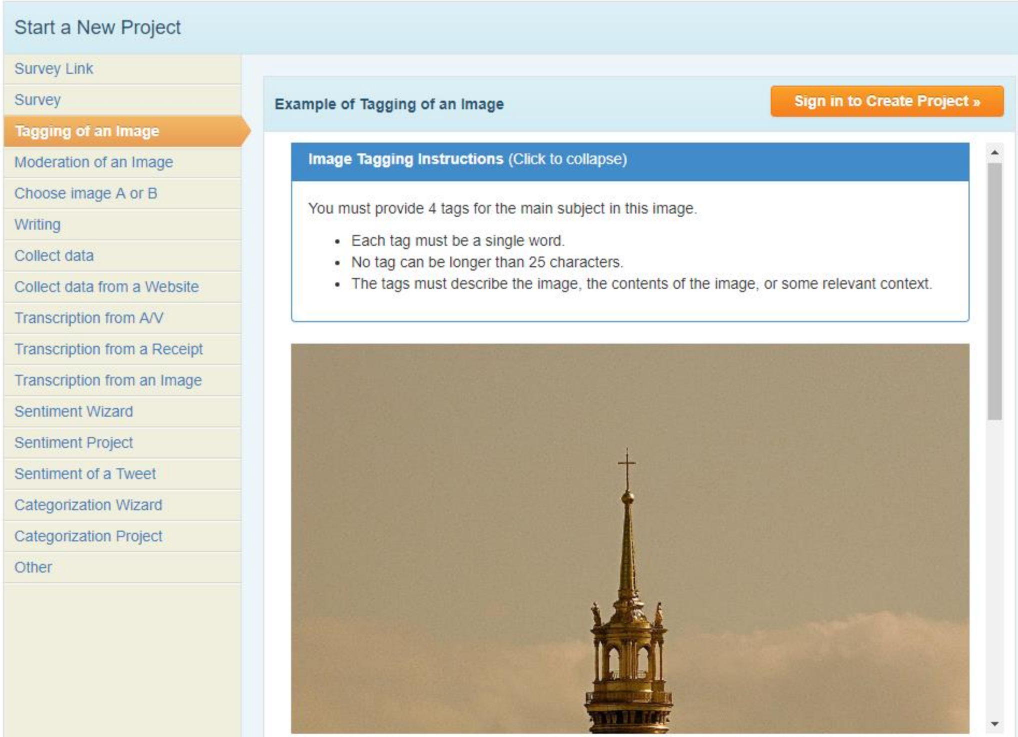


Figure 1: Non-expert correlation for affect recognition

Snow, R., O'Connor, B., Jurafsky, D. and Ng, A.Y., 2008. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 254-263).

Let's Test on the Mturk Sandbox

- The Mechanical Turk Sandbox is a simulated environment that lets you test your applications and Human Intelligence Tasks (HITs) prior to publication in the marketplace.



https://requestersandbox.mturk.com/

Jiyong Park (jiyong.park@kaist.ac.kr)

Exploiting Unstructured Data

Image Analytics for Empirical Research

- Example: Image features on Airbnb (Zhang et al. 2017)
  - ➢ Research question

    Does the verified photo in Airbnb influence property demand? If so, what kinds of photo are effective in Airbnb?

  - ➢ Economining approach

    (1) [Labeled data] Human-coding through Amazon Mechanical Turk for 3,000 Airbnb property images

    (2) [Supervised learning] Support vector machine for image classification based on Fisher Vector

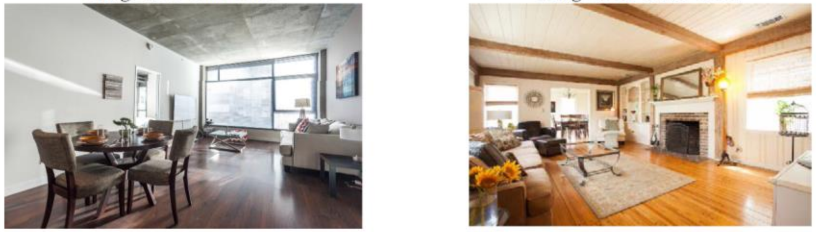
    (3) [Prediction] 380,000 images

Figure 3 Compare Images on Rule of Thirds

Image (Relatively) Doesn't Follow Rule of Thirds — Image (Relatively) Follows Rule of Thirds

Figure 5 Compare Images on Hue (Cool Color vs. Warm Color)

Image with Cool Color — Image with Warm Color



Zhang, S., Lee, D., Singh, P. V., and Srinivasan, K., 2017, How Much is an Image Worth? Airbnb Property Demand Estimation Leveraging Large Scale Image Analytics. *CMU Working Paper.*

Image Analytics for Empirical Research

- Example: Hotel location characteristics (Ghose et al. 2012)
  - ➢ Research question

    Does the location-based hotel characteristics (e.g., near the beach, near downtown) affect hotel demand?
  - ➢ Economining approach

    (1) [Labeled data] (i) locations tagged by users on a social tagging site such as Geonames.org or (ii) locations annotated by users on Amazon Mturk

    (2) [Supervised learning] Support vector machine for image classification

Beach

Downtown

Table 5(a) Extended Model (II) Mean Coefficients

	Mean coefficients (Std. error)
PRICE (log)	-0.145^{***} (0.003)
CHARACTERS (log)	0.009^{***} (0.002)
COMPLEXITY	-0.012^{***} (0.003)
SYLLABLES (log)	-0.045^{***} (0.008)
SMOG	0.083^{**} (0.029)
SPELLERR (log)	-0.129^{***} (0.003)
SUB	-0.138^{***} (0.007)
SUBDEV	-0.403^{***} (0.016)
ID	0.055^{*} (0.030)
CLASS	0.037^{***} (0.008)
CRIME (log)	-0.025^{*} (0.016)
AMENITYCNT (log)	0.005^{**} (0.002)
EXTAMENITY (log)	0.007^{***} (0.001)
BEACH	0.158^{***} (0.005)
LAKE	-0.111^{***} (0.021)
TRANS	0.159^{***} (0.003)
HIGHWAY	0.064^{*} (0.030)
DOWNTOWN	0.045^{***} (0.002)
TA_RATING	0.033^{**} (0.012)
TL_RATING	0.031^{**} (0.011)
TA_REVIEWCNT (log)	0.180^{***} (0.046)
TA_REVIEWCNT^2 (log)	-0.055^{***} (0.007)
TL_REVIEWCNT (log)	0.014^{***} (0.003)
TL_REVIEWCNT^2 (log)	-0.021^{**} (0.008)
Constant	0.037^{**} (0.017)

$^{***}p \le 0.001$; $^{**}p \le 0.01$; $^{*}p \le 0.05$; $^{\dagger}p \le 0.1$.

Ghose, A., Ipeirotis, P.G. and Li, B., 2012. Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. *Marketing Science*, 31(3), pp.493-520.

Image Analytics for Empirical Research

- Example: Image complexity on social media (Shin et al. 2018)
  - Research question

    Does the (image) content complexity influence consumer engagement on social media?
  - Economining approach

    (1) [Labeled data] Flickr dataset of more than 1.2 million images with 1,700 object categories

    (2) [Supervised learning] Convolutional neural networks (CNN) for object classification in images

    (3) [Prediction] 35,651 posts in Tumblr

Based on Yahoo API

$$ImageObjectComplexity = -\sum_{i=1}^{d} \mathrm{p}_i \log(\mathrm{p}_i).$$

(a) Low object complexity (0.262)

(b) High object complexity (3.584)

Shin, D., He, S., Lee, G., Whinston, A., Cetintas, S. and Lee, K. 2018. Enhancing Social Media Analysis with Visual Analytics: A Deep Learning Approach. *UT Austin Working Paper*.

KAIST COLLEGE OF BUSINESS

Text Mining for Empirical Research

- Example: Advertising textual content (Lee et al. 2018)
  - ➢ Research question

    Does the textual advertising content influence consumer engagement on social media?
  - ➢ Economining approach

    (1) [Labeled data] Human-coding through Amazon Mechanical Turk for 5,000 advertising contents on Facebook

    (2) [Supervised learning] Ensemble learning for content tagging to combine results from the multiple classifiers

    (3) [Prediction] 100,000 text messages

Sample Messages	Content Tags
Welcome to the unveiling of the Pro Staff RF97 that I co-designed with Wilson Tennis. Learn more at http://bit.ly/29JXLdA. #FromFederer	BRANDMENTION, PRODMENTION, PRODLOCATION, HTTP
Coach Seve and me. Excited to be back in Brisbane! Happy we got the 1st practice of year out of the way!	SMALLTALK, EMOTION
Hello fans from Colombia! I am very happy to see you at the exo I am playing vs Tsonga on Saturday, December 15th! Buy your tickets starting September 12th on www.tuboleta.com. I hope to see you all there!	EMOTION, SMALLTALK, TARGET, PRODAVAIL, PRODLOCATION, PRODMENTION, HTTP
The Walking Dead Season 1 DVD/Blu-ray is now available, purchase it now!!! http://blogs.amctv.com/the-walking-dead/2011/03/season-1-dvd-blu-ray.php	BRANDMENTION, PRODMENTION, PRODAVAIL, PRODLOCATION, HTTP
Daryl makes a funny. What are some of your favorite #TheWalkingDead quotes? The highest rated quote will be turned into a graphic! #tbt	SMALLTALK, EMOTION, QUESTION, BRANDMENTION, ASKCOMMENT

Lee, D., Hosanagar, K. and Nair, H.S., 2018. Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook. *Management Science*. forthcoming

Text Mining for Empirical Research

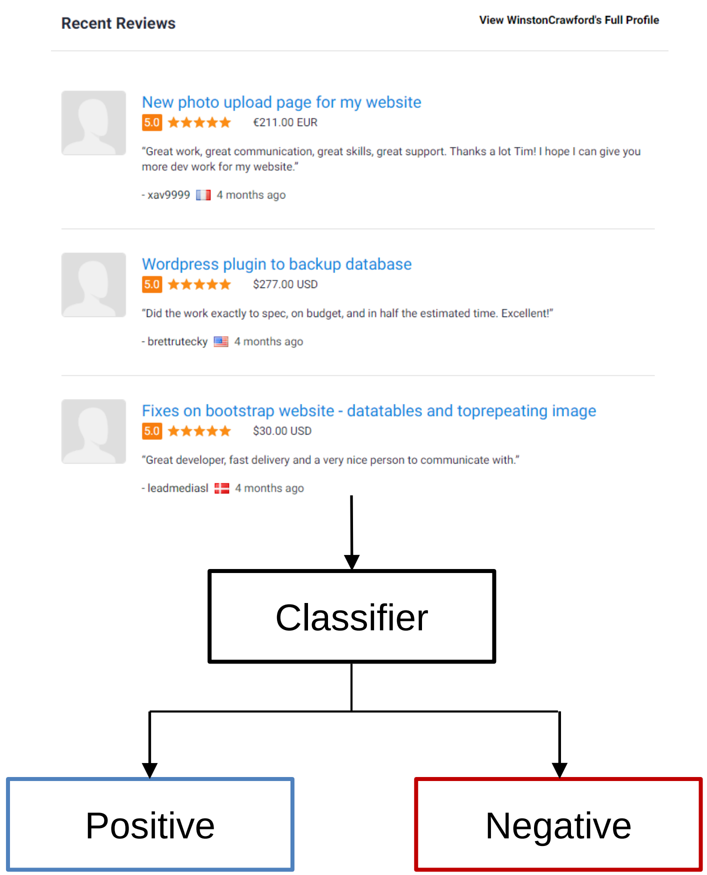
- Example: Reputation in online service marketplace (Moreno and Terwiesch 2014)
  - ➢ Research question

    How do participants in online service marketplaces react to the information tracked by the reputation system (i.e., textual review comments)?
  - ➢ Economining approach

    (1) [Labeled data] Human-coding for positive or negative, through Amazon Mechanical Turk for 2,000 review comments

    (2) [Supervised learning] Naïve Bayes and maximum entropy classifiers

    (3) [Prediction] 100,000 review comments



Moreno, A. and Terwiesch, C., 2014. Doing Business with Strangers: Reputation in Online Service Marketplaces. *Information Systems Research*, 25(4), pp.865-886.

Text Mining for Empirical Research

- Example: Linguistic style of video pitch (Park et al. 2018)
  - ➢ Research question

    Does the video pitch (speech) on online crowdfunding influence funding outcomes? If so, what kinds of pitch style are effective in Kickstarter?

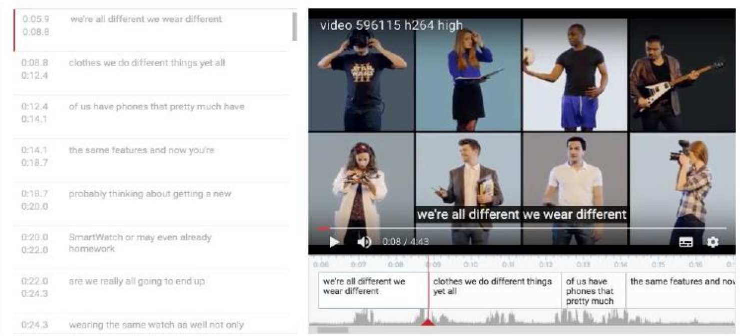
  - ➢ Economining approach

    (1) [Labeled data] Labeling personality traits for text corpus

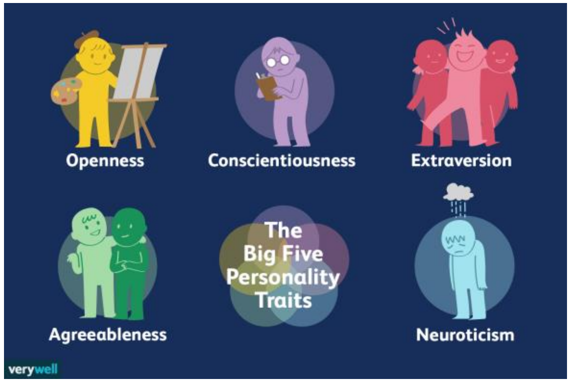
    (2) [Supervised learning] Deep learning for text classification for personality traits

    (3) [Prediction] 4,700 videos

Based on IBM API



Video transcript (Google API)



Park, J., Kim, J., Cho, D. and Lee, B., 2018. Pitching with Style: The Role of the Entrepreneur's Video Pitch on Online Crowdfunding. *KAIST Working Paper.*

Extracting Latent Information

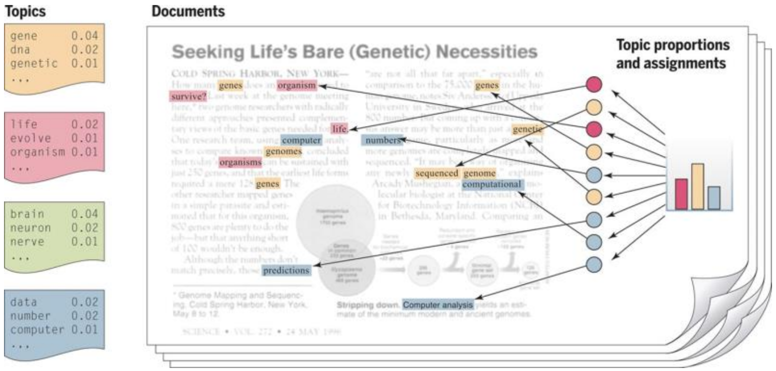
Topic Models for Revealing Latent Topics

- Example: Textual risk disclosures (Bao and Datta 2014)
  - ➢ What information to be extracted

    Different types of corporate risks from unstructured, textual risk disclosures

  - ➢ Economining approach

    In this paper, the authors develop a variation of the latent Dirichlet allocation (LDA) topic model and its learning algorithm for simultaneously discovering and quantifying risk types from textual risk disclosures.



Topic models

[Topic label] risk factors
[T1: human resources risks] The company's success depends largely on its ability to attract and retain key personnel.
[T2: intellectual property risks] The company's business relies on access to patents and intellectual property obtained from third parties, and the company's future results could be adversely affected if it is alleged or found to have infringed on the intellectual property rights of others.
[T3: potential/ongoing lawsuits] Unfavorable results of legal proceedings could adversely affect the company's results of operations.
[T5: catastrophes] War, terrorism, public health issues, and other circumstances could disrupt supply, delivery, or demand of products, which could negatively affect the company's operations and performance.

Bao, Y. and Datta, A., 2014. Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures. *Management Science*, 60(6), pp.1371-1391.

KAIST COLLEGE OF BUSINESS

Similarity Measure based on Textual Descriptions

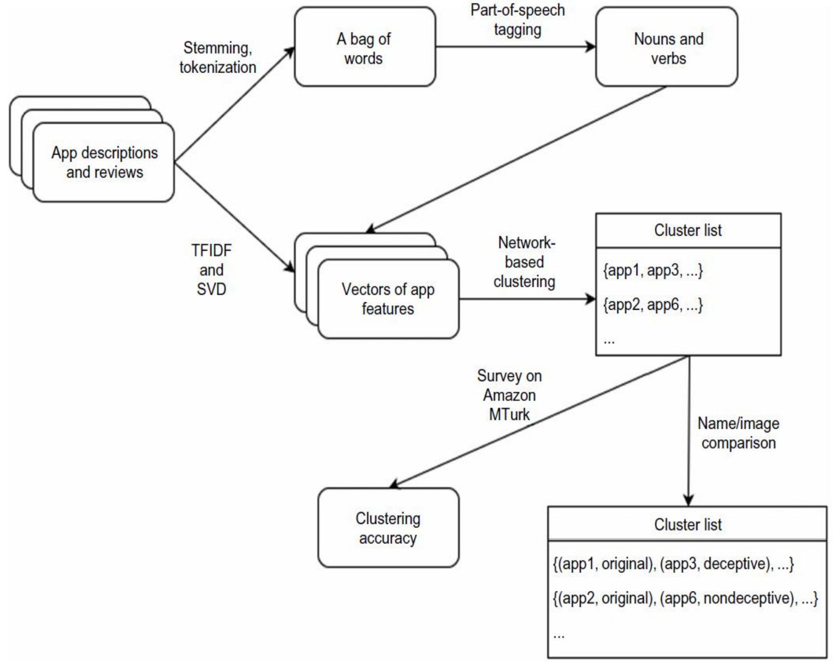
- Example: Copycats app detection (Wang et al. 2018)
  - What information to be extracted

    Relationship or similarity between mobile apps from unstructured, textual descriptions and customer reviews

  - Economining approach

    The authors measure app similarity based on cosine similarity between term frequency-inverse document frequency (TF-IDF) vectors for about 10,000 apps.

Figure 1. Flow Chart of Copycat Detection Framework



Wang, Q., Li, B. and Singh, P., 2018. Copycats vs. Original Mobile Apps: A Machine Learning Copycat Detection Method and Empirical Analysis. *Information Systems Research*. forthcoming

Dictionary-based Text Classification

- Example: User personality traits on social media (Adamopoulos et al. 2018)
  - What information to be extracted

    User personality traits from social media posts
  - Economining approach

    Counting the dictionary for pre-defined categories (e.g., positive/negative, personality)

Top correlations between the Big Five and individual words.

Trait	No. of words sig. at p <.001	Top 20 words
Neuroticism	24	awful (0.26), though (0.24), lazy (0.24), worse (0.21), depressing (0.21), irony (0.21), road (-0.2), terrible (0.2), Southern (-0.2), stressful (0.19), horrible (0.19), sort (0.19), visited (-0.19), annoying (0.19), ashamed (0.19), ground (-0.19), ban (0.18), oldest (-0.18), invited (-0.18), completed (-0.18)
Extraversion	20	bar (0.23), other (-0.22), drinks (0.21), restaurant (0.21), dancing (0.2), restaurants (0.2), cats (-0.2), grandfather (0.2), Miami (0.2), countless (0.2), drinking (0.19), shots (0.19), computer (-0.19), girls (0.19), glorious (0.19), minor (-0.19), pool (0.18), crowd (0.18), sang (0.18), grilled (0.18)
Openness	393	folk (0.32), humans (0.31), of (0.29), poet (0.29), art (0.29), by (0.28), universe (0.28), poetry (0.28), narrative (0.28), culture (0.28), giveaway (-0.28), century (0.28), sexual (0.27), films (0.27), novel (0.27), decades (0.27), ink (0.27), passage (0.27), literature (0.27), blues (0.26)
Agreeableness	110	wonderful (0.28), together (0.26), visiting (0.26), morning (0.26), spring (0.25), porn (-0.25), walked (0.23), beautiful (0.23), staying (0.23), felt (0.23), cost (-0.23), share (0.23), gray (0.22), joy (0.22), afternoon (0.22), day (0.22), moments (0.22), hug (0.22), glad (0.22), fuck (-0.22)
Conscientiousness	13	completed (0.25), adventure (0.22), stupid (-0.22), boring (-0.22), adventures (0.2), desperate (-0.2), enjoying (0.2), saying (-0.2), Hawaii (0.19), utter (-0.19), it's (-0.19), extreme (-0.19), deck (0.18)

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

Machine Learning for Causal Inference

Machine Learning for Causal Inference

Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 87–106

Machine Learning: An Applied Econometric Approach

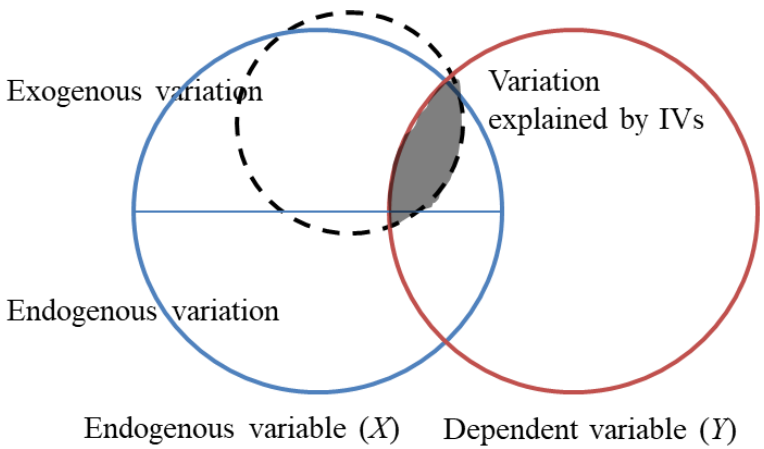
Sendhil Mullainathan and Jann Spiess (Harvard University)

- "In another category of applications, the key object of interest is actually a parameter β, but the inference procedures (often implicitly) contain a prediction task." (Mullainathan and Spiess 2017, p. 88)
  - "For example, (1) the first stage of a linear instrumental variables regression is effectively prediction. The same is true when (2) estimating heterogeneous treatment effects, testing for effects on multiple outcomes in experiments, and (3) flexibly controlling for observed confounders."

Mullainathan, S. and Spiess, J., 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), pp.87-106.

(1) Deep Instrument Variable (Deep IV)

- Instrument variables (IVs) aim at isolating the exogenous variation from endogenous independent variables. This is basically prediction.
  - ➢ Improving the predictive power of IVs could mitigate weak instrument biases, leading to efficient and unbiased estimation.
  - ➢ Hartford et al. (2017) propose the Deep IV framework, which is flexible (non-parametric) and effective for heterogeneity.



Two-Stage Least Squares (2SLS)

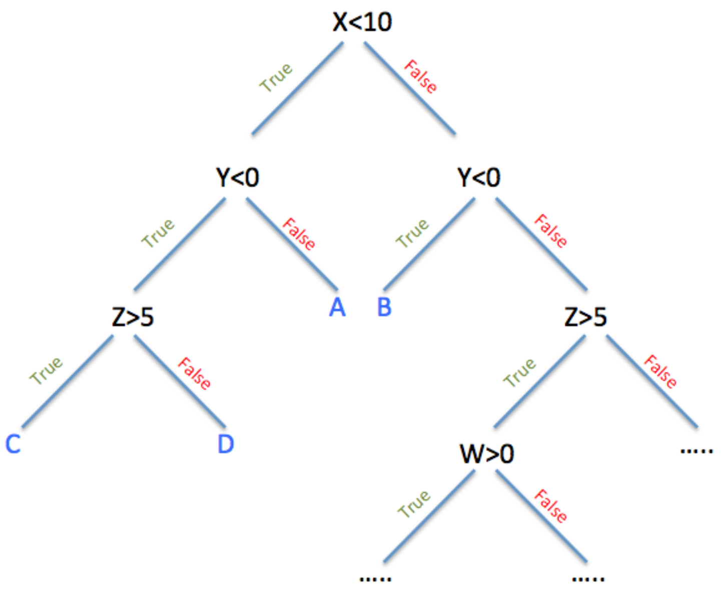
- ➢ (1) First-stage estimation
  - $X = a + bZ + Controls + \mu$
- ➢ (2) Second-stage estimation
  - $Y = \alpha + \beta\hat{X} + Controls + \varepsilon$

Predicted value from first-stage

Hartford, J., Lewis, G., Leyton-Brown, K. and Taddy, M., 2017. Deep IV: A Flexible Approach for Counterfactual Prediction. In *International Conference on Machine Learning (ICML).*

(2) Heterogeneous Causal Effects

- Machine learning techniques can be used to partition the data into subpopulations that differ in the magnitude of their treatment effects.
  - ➢ Athey and Imbens (2016) develop a regression tree method (*causal trees*), which uses a different criterion for building the tree: rather than focusing on improvements in mean-squared error of the prediction of outcomes, it focuses on mean-squared error of treatment effects.



Athey, S. and Imbens, G., 2016. Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences (PNAS)*, 113(27), pp.7353-7360.

(3) Dimensionality Reduction

- To mitigate omitted variable biases, researchers might need to include a large number of variables relative to the sample size (high-dimensional data).
  - "Researchers are thus faced with a large set of potential variables formed by different ways of interacting and transforming the underlying variables" (Belloni et al. 2014, p. 29)

- Machine learning techniques for regularization can be used to select relevant variables.
  - Belloni et al. (2014) propose a double selection procedure, where they first use a LASSO regression to select covariates that are correlated with the outcome, and then again to select covariates that are correlated with the treatment.

Belloni, A., Chernozhukov, V. and Hansen, C., 2014. High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2), pp.29-50.

(3) Dimensionality Reduction

- Example: Impact of legalized abortion on crime

QUARTERLY JOURNAL
OF ECONOMICS

Vol. CXVI May 2001 Issue 2

THE IMPACT OF LEGALIZED ABORTION ON CRIME*

JOHN J. DONOHUE III AND STEVEN D. LEVITT

"We offer evidence that legalized abortion has contributed significantly to recent crime reductions." (p. 379)

The Quarterly Journal of Economics, February 2008

THE IMPACT OF LEGALIZED ABORTION ON CRIME: COMMENT*

CHRISTOPHER L. FOOTE AND CHRISTOPHER F. GOETZ

"Their cross-state regressions, by contrast, imply a large selection effect… We argue that the cross-state results are not robust to controls for omitted variables." (p. 421)

MEASUREMENT ERROR, LEGALIZED ABORTION, AND THE DECLINE IN CRIME: A RESPONSE TO FOOTE AND GOETZ*

JOHN J. DONOHUE III AND STEVEN D. LEVITT

"Our further analysis of their claims regarding omitted variable bias as an explanation for the link between legalized abortion shows that their results are extremely sensitive to minor alterations." (p. 439)

Donohue III, J.J. and Levitt, S.D., 2001. The Impact of Legalized Abortion on Crime. *Quarterly Journal of Economics*, 116(2), pp.379-420.
Foote, C.L. and Goetz, C.F., 2008. The Impact of Legalized Abortion on Crime: Comment. *Quarterly Journal of Economics*, 123(1), pp.407-423.
Donohue III, J.J. and Levitt, S.D., 2008. Measurement Error, Legalized Abortion, and the Decline in Crime: A Response to Foote and Goetz. *Quarterly Journal of Economics*, 123(1), pp.425-440.

(3) Dimensionality Reduction

- Example: Impact of legalized abortion on crime
  - Belloni et al. (2014) apply the double selection procedure to automatically select relevant omitted variables to address Donohue and Levitt (2008)'s comment:

“The Foote and Goetz findings, however, prove to be very sensitive to minor alterations in specification. Foote and Goetz's Table II, column (5) results include Census division-year interactions. Column (4) of our Table III shows that without the division-year interactions, but including the interaction of 1970–1984 mean log per capita crime rates and a linear crime trend, the effect of abortion on crime remains highly statistically significant for violent and property crime.” (p. 436)

  - Among hypothetical 284 controls (for 600 observations), Belloni et al. (2014) select relevant variables and obtain insignificant estimations.

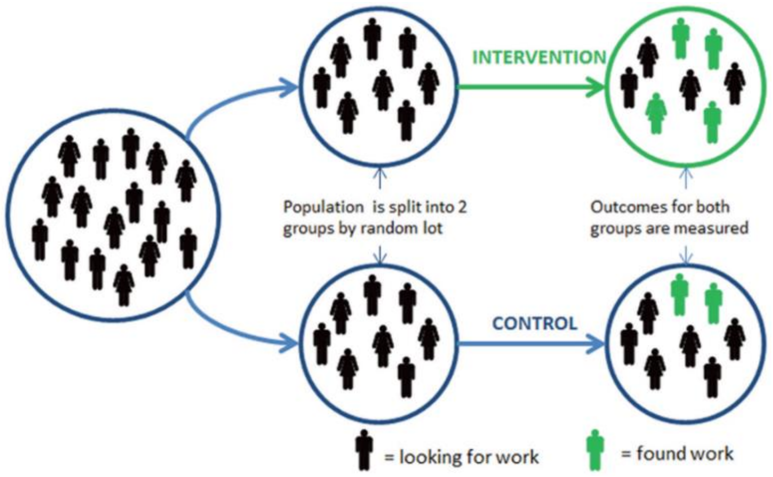
Effect of Abortion on Crime

	Violent	
Estimator	*Effect*	*Std. error*
First-difference	−.157	.034
All controls	.071	.284
Double selection	−.171	.117

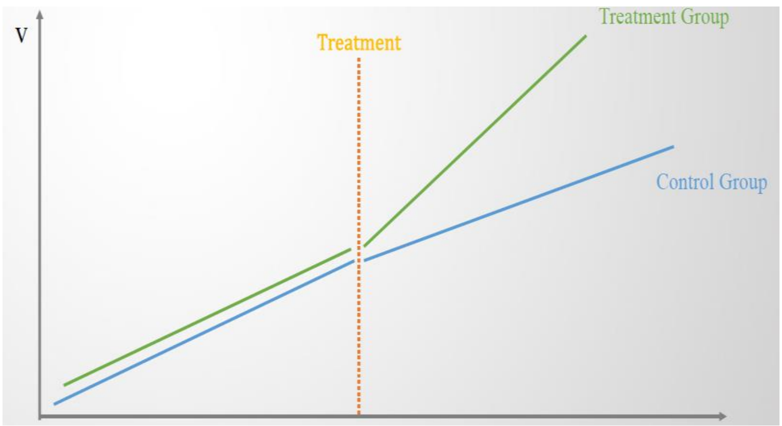
Belloni, A., Chernozhukov, V. and Hansen, C., 2014. High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2), pp.29-50.
Donohue III, J.J. and Levitt, S.D., 2008. Measurement Error, Legalized Abortion, and the Decline in Crime: A Response to Foote and Goetz. *Quarterly Journal of Economics*, 123(1), pp.425-440.

(4) Pseudo Control Group

- The control group, which provides an estimate of the counterfactual, is the gold standard for causal inference.
  - ➢ However, even if we do not have a true control group, we might be able to develop a predictive model of the counterfactual.



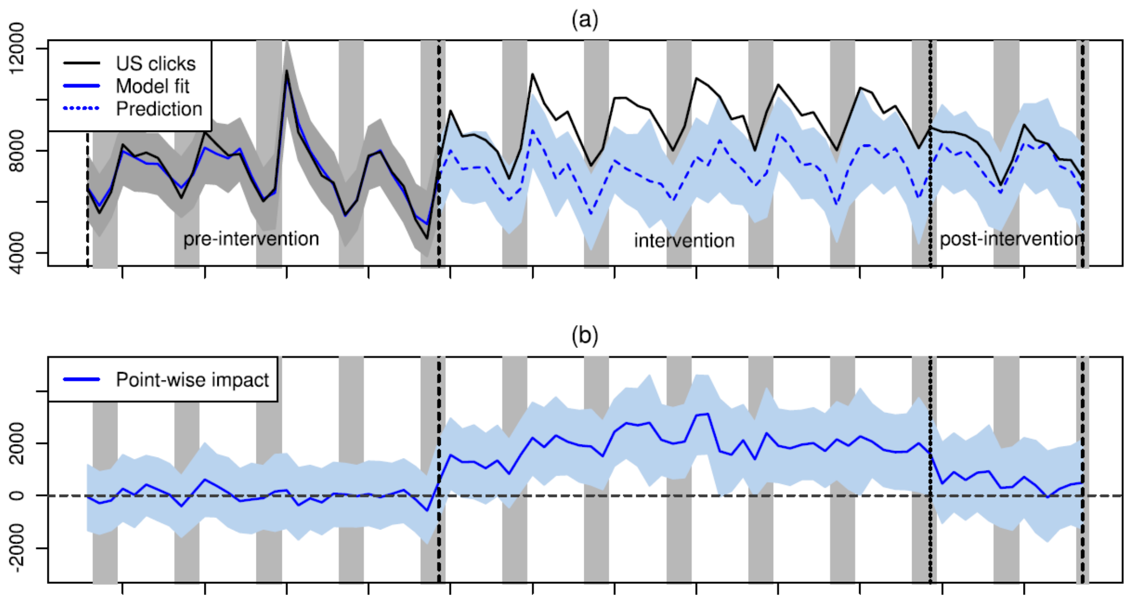
Randomized experiment



Quasi-experiment
(Difference-in-differences)

(4) Pseudo Control Group

- Example: Train-Test-Treat Compare (TTTC) model
  - ➢ Brodersen et al. (2015) compare the predicted trends of organic clicks (as a counterfactual; control group) with the realized trends treated by advertising (i.e., organic + paid clicks).

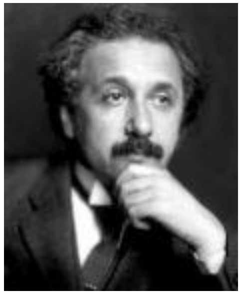


Brodersen, K.H., Gallusser, F., Koehler, J., Remy, N. and Scott, S.L., 2015. Inferring Causal Impact Using Bayesian Structural Time-Series Models. *The Annals of Applied Statistics*, 9(1), pp.247-274.

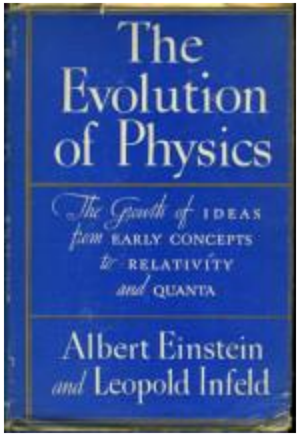
Concluding Remarks

Surviving in the Age of Big Data and Machine Learning

- Computational techniques or algorithms itself are not our major interest.
  - "Problem-formulation skills represent core skills for data scientists over the next decade." (Dhar 2013, p. 70)



Albert Einstein



"**The formulation of a problem is often more essential than its solution**, which may be merely a matter of mathematical or experimental skill. To raise new questions, new possibilities, to regard old problems from a new angle, requires creative imagination and marks real advance in science"

Dhar, V. 2013. Data Science and Prediction. *Communications of the ACM*, *56*(12), 64-73.

Methodology is Important only for Important Question

- No research methodology can save the poor research question.

"Type III errors occur when a researcher answers the wrong question using the right methods. A lot of effort may be expended, a great deal of rigor may be applied, but coming up with the right answer to the wrong question does not create value." (p. iii)

"An incomplete or imprecise answer to the right question can be a significant advance, while a complete and precise answer to the wrong question does not create value." (p. vii)

- Arun Rai, Editor-in-Chief of *MIS Quarterly*

Rai, A., 2017. Editor's Comments: Avoiding Type III Errors: Formulating IS Research Problems that Matter. *MIS Quarterly*, 41(2), pp.iii-vii.

End of Document