Business Analytics (IM561)

Module1. Video Analytics and Social Science Research

Intelligent Video Analytics with Deep Learning

Jongho Kim

December & Company Inc.

Spring, 2019

KAIST COLLEGE OF BUSINESS

Module1. Video Analytics and Social Science Research

Social Science Research with Video Analytics

- Social Science + Video Data ≠ Computer Science + Video Data

- 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.
KAIST Summer Session Organized by Jiyong Park(2018)

Social Science Research with Video Analytics

- Social Science + Video Data $\neq$ Computer Science + Video Data

- 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.

Causation vs. Prediction Example

- 만약 전자레인지 폭발에 의해 불이나 난 상황에서
    1. Causation: 불이 일어난 원인은 전자레인지 폭발에 의한 것입니다.
    2. Correlation: 소방관이 몰려 있는 곳에는 불이 났을 수 있습니다. 하지만, 소방관이 불을 일으키진 않았고, 둘 사이에는 상관관계가 높을 뿐 입니다.
    3. Prediction: 하지만 소방관이 몰려있다는 것으로 부터, 불이 났을 거라는 예측을 할 수는 있습니다.

- 따라서, Prediction에 있어서는 Causation과 관련 없이 Correlation이 중요 합니다.
- Prediction과 Causation은 목표로 하는 역할이 다르며, 저희가 Machine Learning을 통해 풀고자 하는 문제의 목표는 Prediction Level과 관련이 있습니다.

Social Science Research with Video Analytics

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."

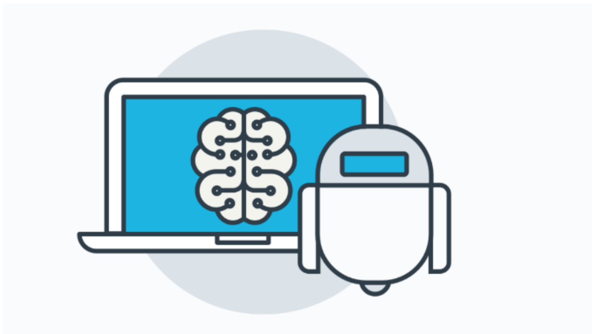
  "Big data alone is insufficient to make valid causal inferences; however, having more data certainly can improve causal inferences in large-scale datasets."

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.

How Do Video Analytics Transform Social Science?

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 Video Analytics?

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."

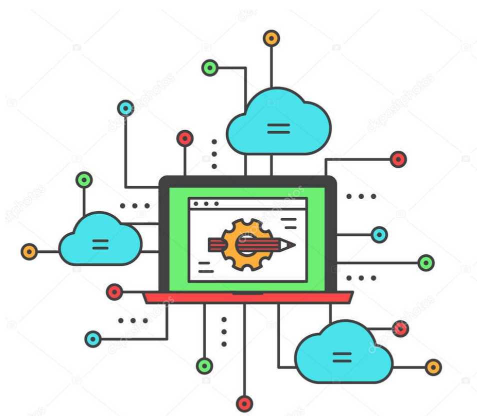
Grimmer, J., 2015. We are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together. *PS: Political Science & Politics*, 48(1), pp.80-83.

Economining = Econometrics + Data Mining

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

Inputs → Mechanism (Theory) → Outputs

New Data



New Data

Data Mining

Machine Learning

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?

Example: Social Science Research

Video Analytics 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?

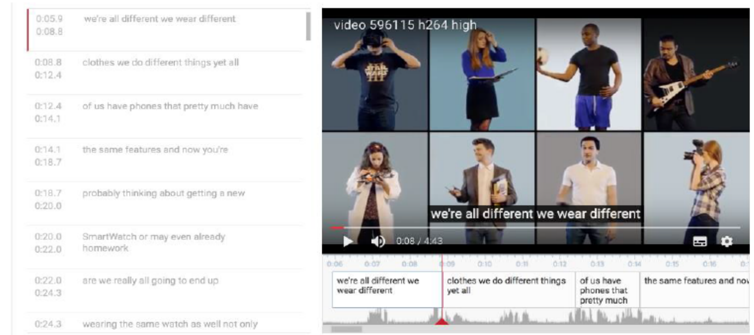
  - ➢ Economining approach

    (1) [Labeled data] Labeling personality traits from video data

    (2) [Supervised learning] Deep learning for extracting 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.*

Video Analytics for Empirical Research

- Example: Image Features of Advertisments (So and Oh 2018)
  - ➢ Research question

    How numerous image components of advertisements designed for fashion products affect consumers' search and purchase intentions
  - ➢ Economining approach
    1. (1) [Labeled data] Extracting various stimuli from image data
    2. (2) [Supervised learning] (1) facial expressions, (2) degrees of stimulus content, (3) image classification, (4) color properties, and (5) texts contained within the images
    3. (3) [Prediction] 13,040 products

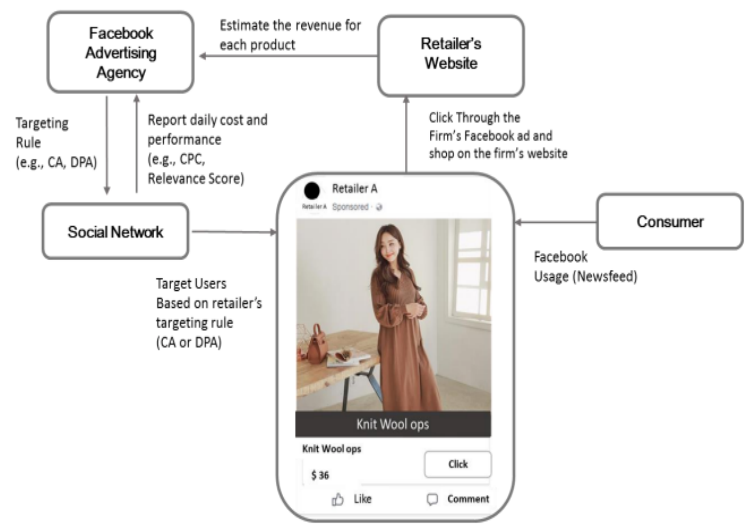


Figure 1. Facebook Targeting Mechanisms

So, H., and Oh, W., 2018. Picture Perfect: An Image Mining of Advertising Content and Its Effects on Social Targeting, *ICIS2018*

Video 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

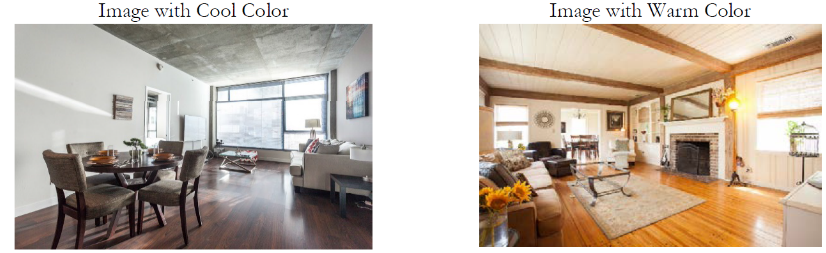
    (3) [Prediction] 380,000 images

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.*

Video 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.

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:
  - “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.”

Example: Computational Techniques Potentially Harvestable in Social Science

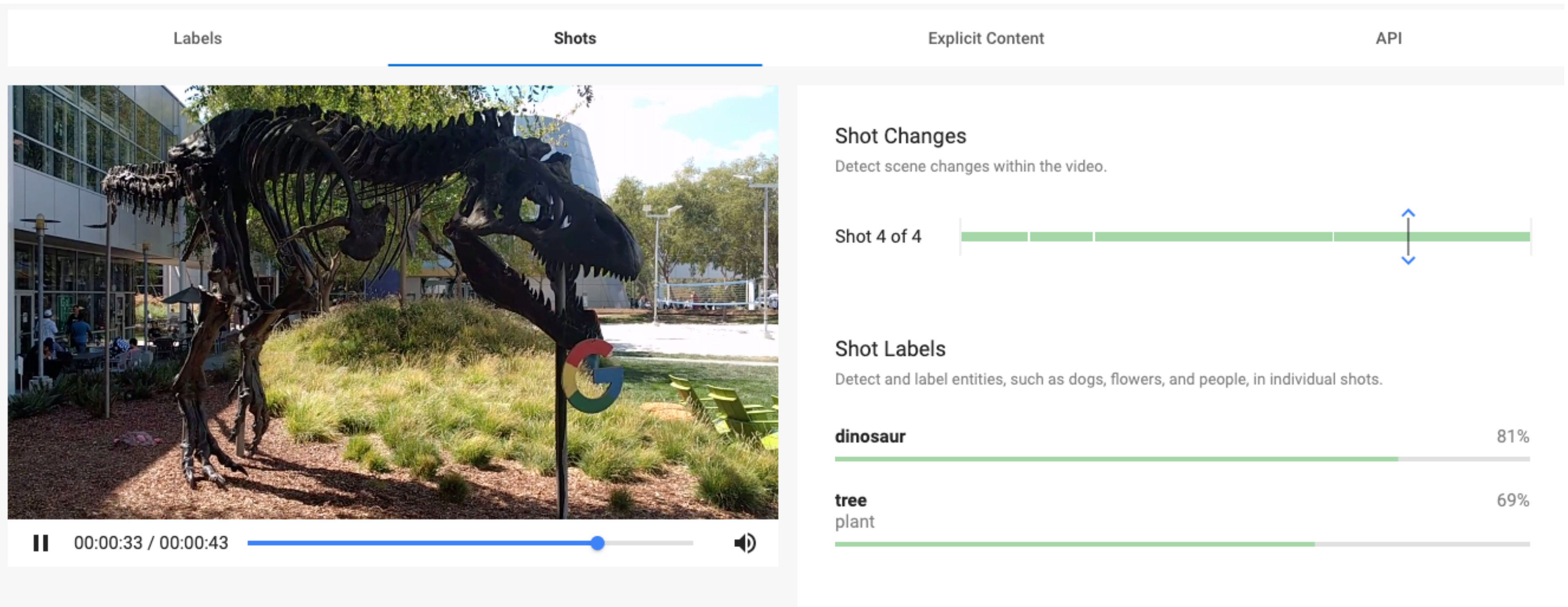
OpenPose: Detect Human Body, Hand, Facial, and Foot

- Potentially Harvestable: We did n’t fully explore human pose data

https://github.com/CMU-Perceptual-Computing-Lab/openpose

Video Scene Detection

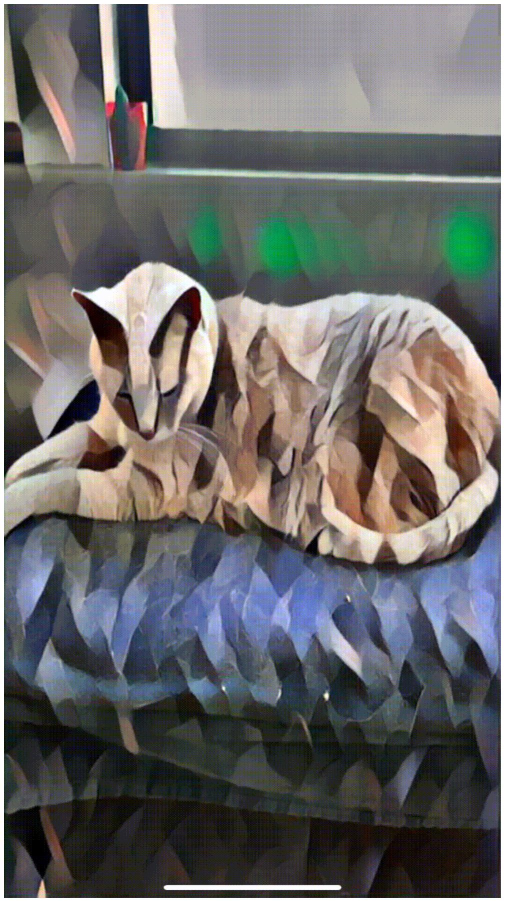
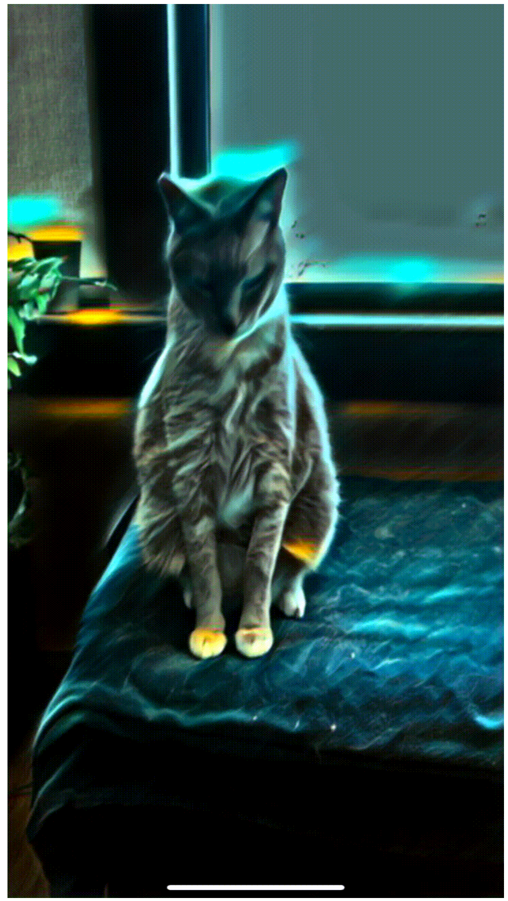
- Potentially Harvestable: Video Scene can be used for video contents control



https://cloud.google.com/video-intelligence/

Real Time Video Style Transfer

- Potentially Harvestable: We may find economically meaningful style


https://towardsdatascience.com/real-time-video-neural-style-transfer-9f6f84590832

Lip Reading

- Potentially Harvestable: Text feature extraction from Lip

https://grail.cs.washington.edu/projects/AudioToObama/

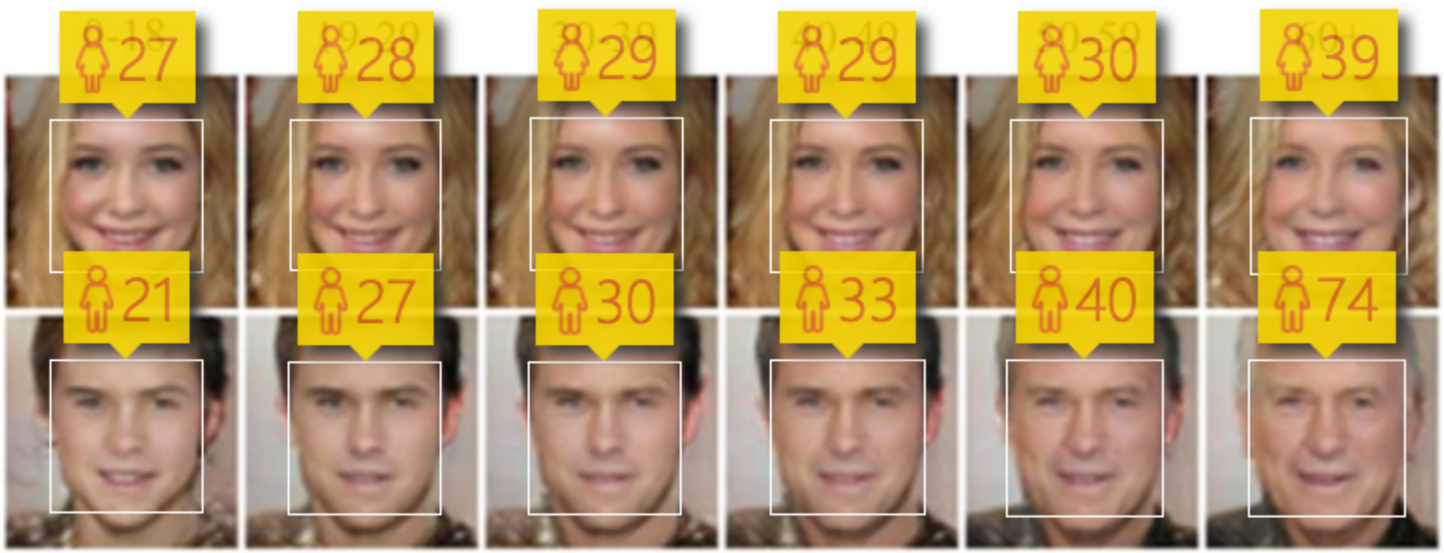
Fake Lip-Sync Videos

- Potentially Harvestable: Generate controlled videos with Specific Features
  - Ex. High tone / pitch, big mouth, moving eyebrow, etc.

https://grail.cs.washington.edu/projects/AudioToObama/

Changing face age with GANs

- Potentially Harvestable: Generate controlled videos by adjusting age features



https://grail.cs.washington.edu/projects/AudioToObama/

Concluding Remarks

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.

Thank you ☺

Contact Info: quantic.jh@gmail.com

Appendix. 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

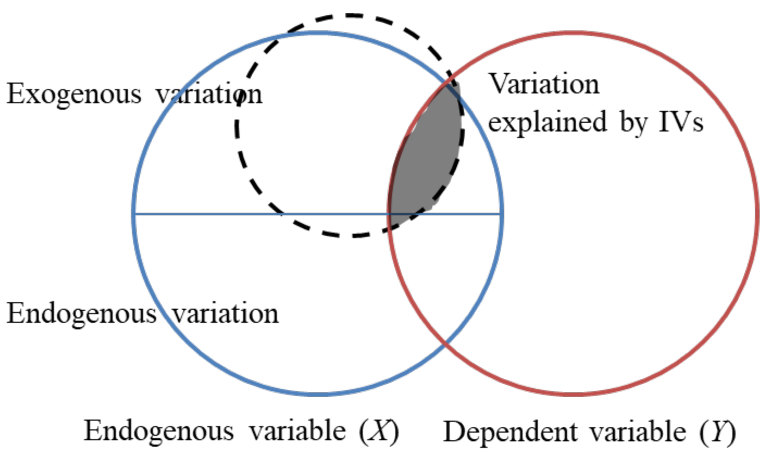
Sendhil Mullainathan and Jann Spiess (Harvard University)

- "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

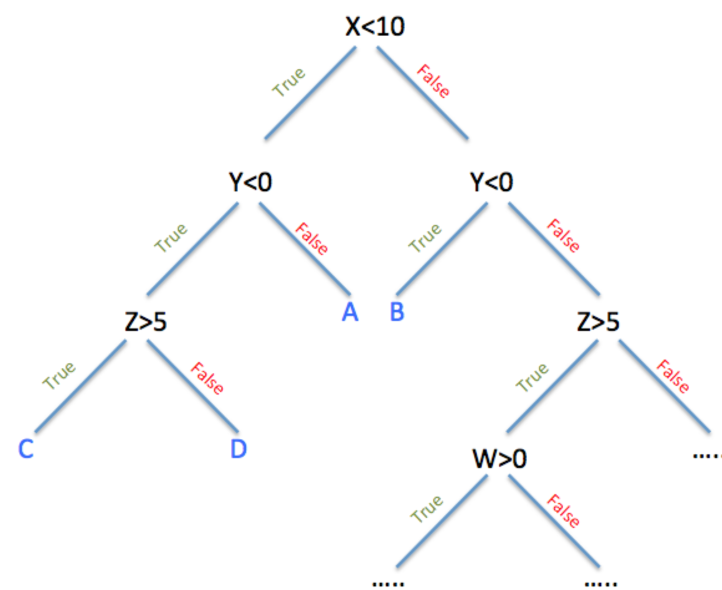
- $Y = \alpha + \beta\hat{X} + Controls + \varepsilon$

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.

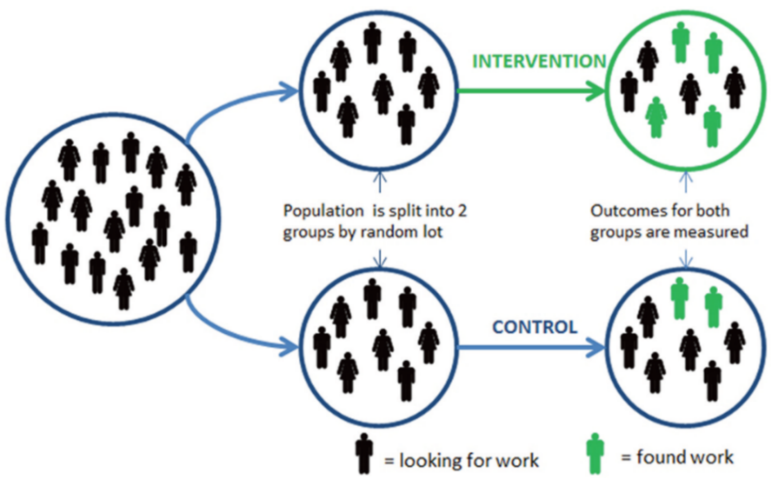
Effect of Abortion on Crime

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

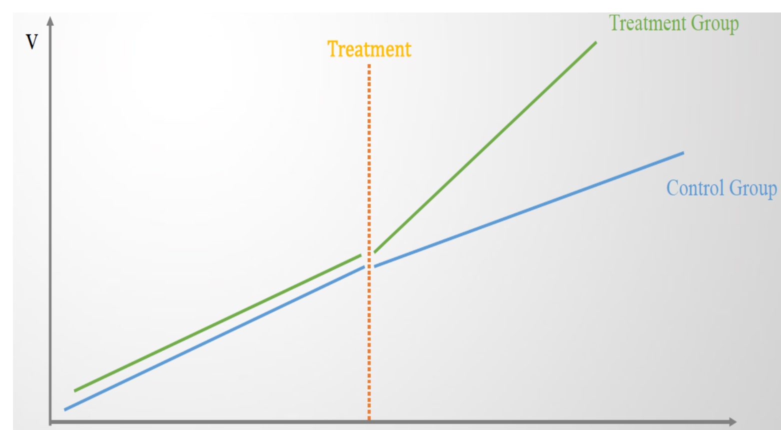
Belloni, A., Chernozhukov, V. and Hansen, C., 2014. High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2), pp.29-50.
Donohue III, J.J. and Levitt, S.D., 2008. Measurement Error, Legalized Abortion, and the Decline in Crime: A Response to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



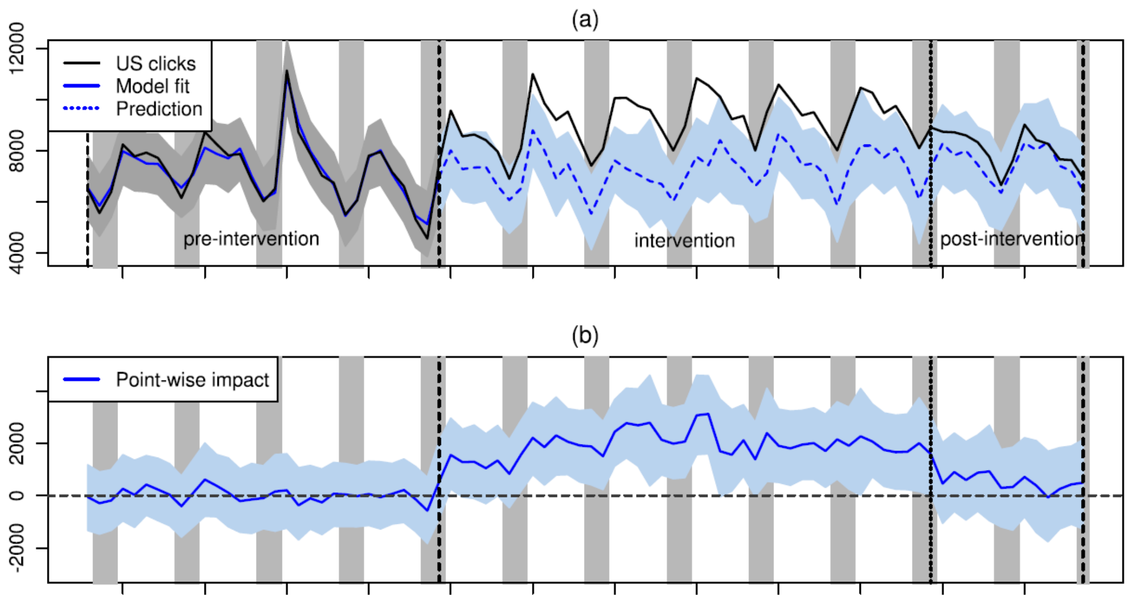
Randomized experiment



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.