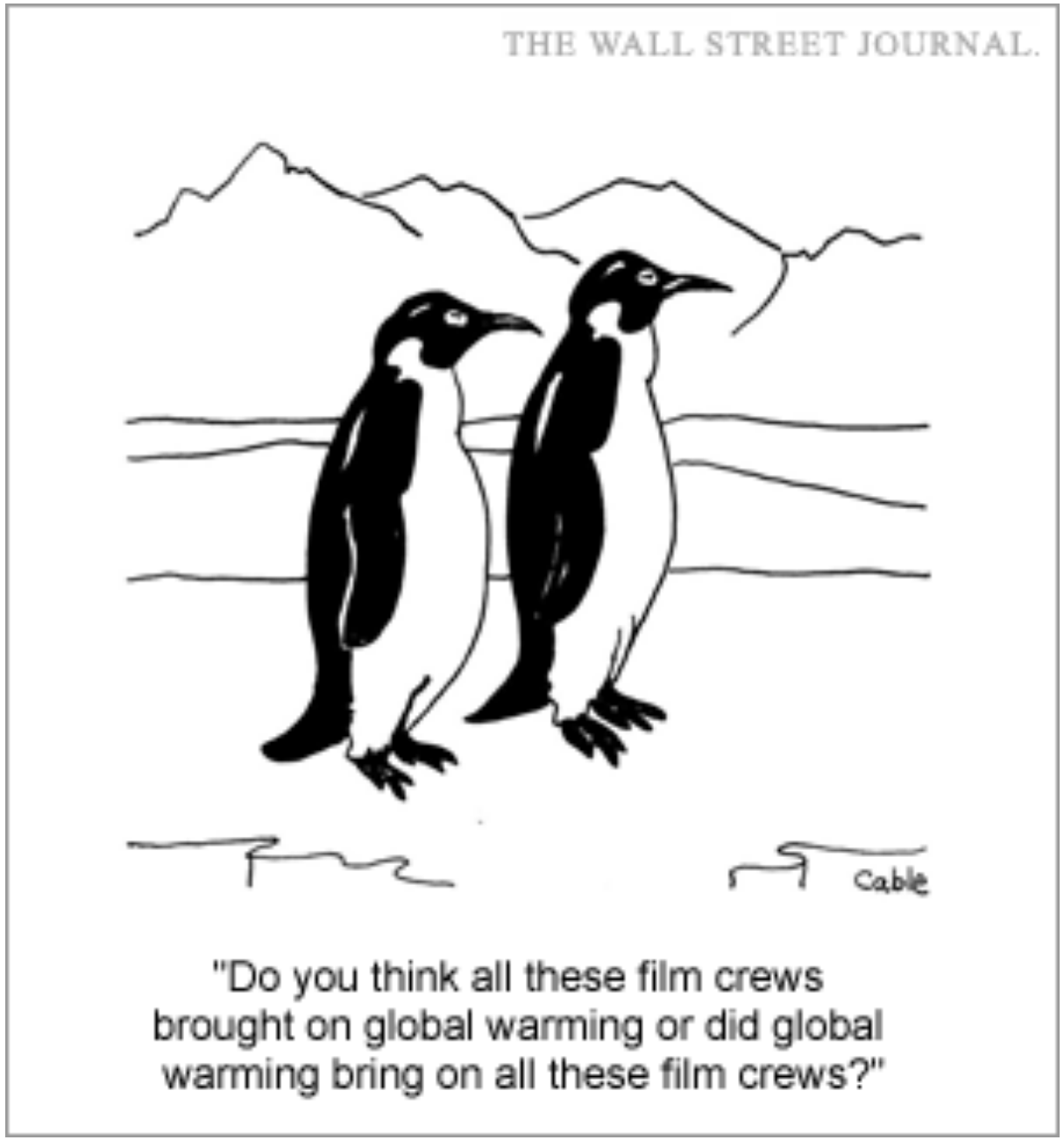
Introduction to *Urban* Data Science

Responsible Data Science

(EPA1316A)

Lecture 14

Trivik Verma



Source: Wall Street Journal

Last Time

- The *point* of points
- Point patterns
- Visualization of point patterns
- Identifying clusters of points

Today

- Responsible Data Science
- Correlation Vs Causation
- Causal inference
- Why/when causality matters
- Hurdles to causal inference & strategies to overcome them

The unfortunate history of racial bias in photography. - Vox

Responsible data scientists take steps to make **data** they depend on findable, accessible, interoperable and reusable (FAIR) while ensuring the fairness, accuracy, confidentiality and transparency (FACT) of the algorithms and tools they create.

I will ask you some questions **“Imagine your employer asks you to…”**

*** Select one option from A-E**

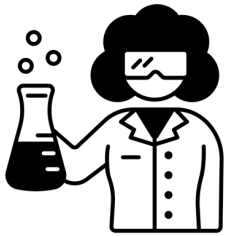
Rules:

- There is no right answer
- Up for debate
- Be respectful of all choices
- If you don’t want to answer, that is okay

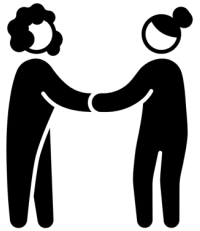
Break

CHILL

WALK



COFFEE OR TEA



MAKE FRIENDS

Icons by Iconathon on The Noun Project

Correlation Vs Causation

Correlation Vs Causation

Two fundamental ways to look at the relationship between two (or more) variables:

Correlation

Two variables have co-movement. If we know the value of one, we know something about the value of the other one.

Causation

There is a “cause-effect” link between the two and, as a result, they display co-movement.

Correlation Vs Causation

- Both are useful, but for different purposes
- Causation *implies* correlation but **not** the other way around
- It is vital to keep this distinction in mind for meaningful and credible analysis

Examples

Temperature and ice-cream consumption

Sign correlation (P or N)? Causal link (P or N)?

A. Positive Positive (PP)
B. Positive Negative (PN)
C. Negative Positive (NP)
D. Negative Negative (NN)

Non-commercial space launches & Sociology PhDs awarded

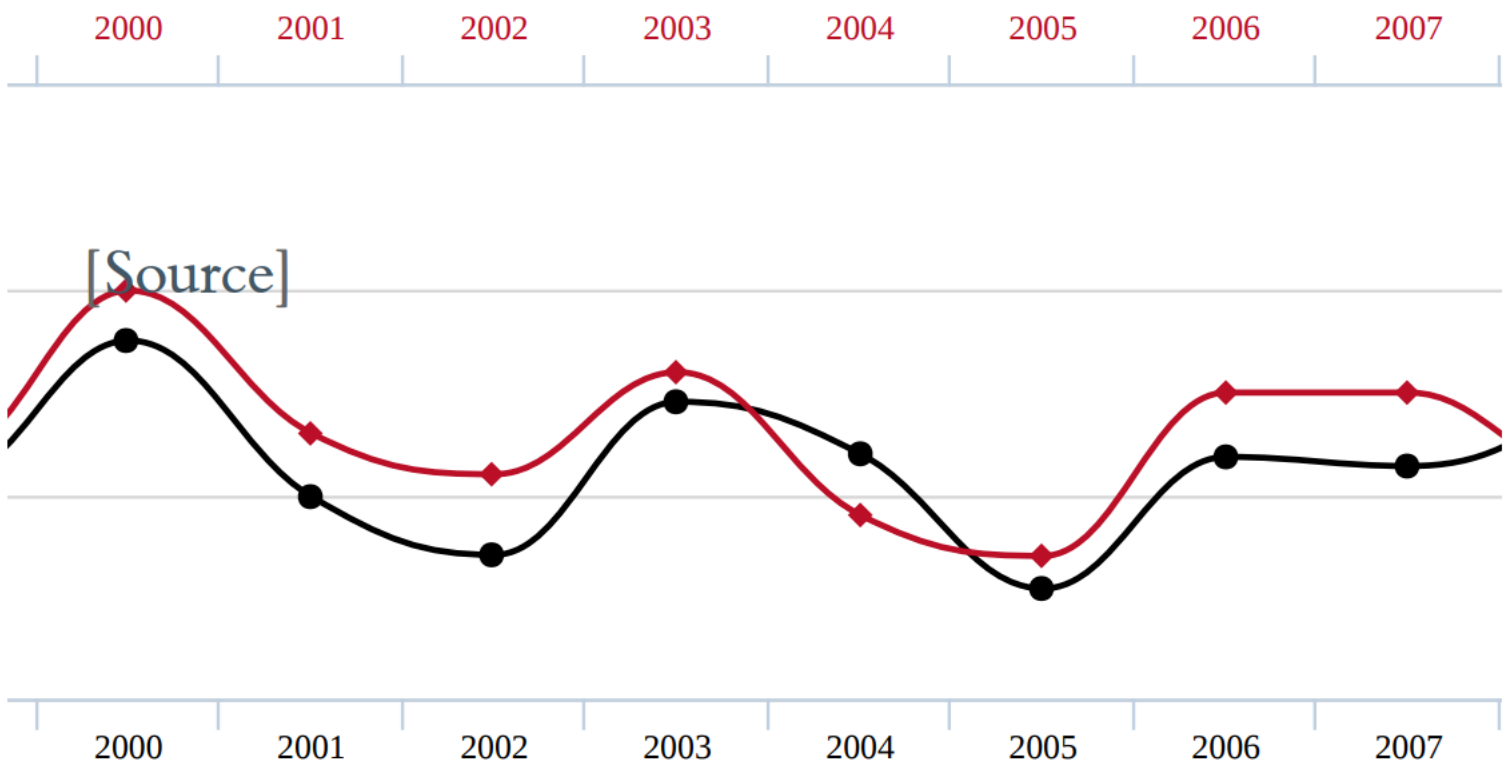
Sign correlation (P or N)? Causal link (P or N)?

A. Positive Positive (PP)
B. Positive Negative (PN)
C. Negative Positive (NP)
D. Negative Negative (NN)

Worldwide non-commercial space launches

correlates with

Sociology doctorates awarded (US)



Crime & Policing

Sign correlation (P or N)? Causal link (P or N)?

A. Positive Positive (PP)
B. Positive Negative (PN)
C. Negative Positive (NP)
D. Negative Negative (NN)

Causal Inference

Why/When to get Causal?

Why

- Most often, we are interested in understanding the **processes** that *generate* the world, not only in observing its outcomes
- Many of these processes are only **indirectly observable** through **outcomes**
- Example:
  - Heart attacks
  - Accidents
  - ...
- The only way to link both is through causal channels

When

Essentially when the **core interest** is to find out if something **causes** something else

- Policy interventions
- Medical trials
- Business decisions (product/feature development…)
- Empirical (Social) Sciences
- …

When not (necessarily)

Exploratory analysis

Distracting, if not enough, knowledge about the dataset

Predictive settings

Interest not in understanding the underlying mechanisms but want to obtain *best possible estimates* of a variable you do not have by combining others you do have

Hurdles to Causal Inference

Hurdles to Causal Inference

Causation *implies* Correlation

Correlation *does **not** imply* Causation

Why?

- Reverse causality
- Confounding factors/endogeneity

Reverse Causality

There **is** a causal link between the two variables but it either runs the opposite direction as we think, or runs in both

E.g. Education and income

Confounding Factors

Two variables are correlated because they are ***both*** determined by other, unobserved, variables (factors) that ***confound*** the effect

E.g. Ice cream and cold beverages consumption

Strategies

Is there any way to overcome reverse causality and confounding factors to recover causal effects?

The key is to get an **“exogenous source of variation”**

Strategies

Randomized Control Trials

Treated Vs control groups. Probability of treatment is independent of everything else

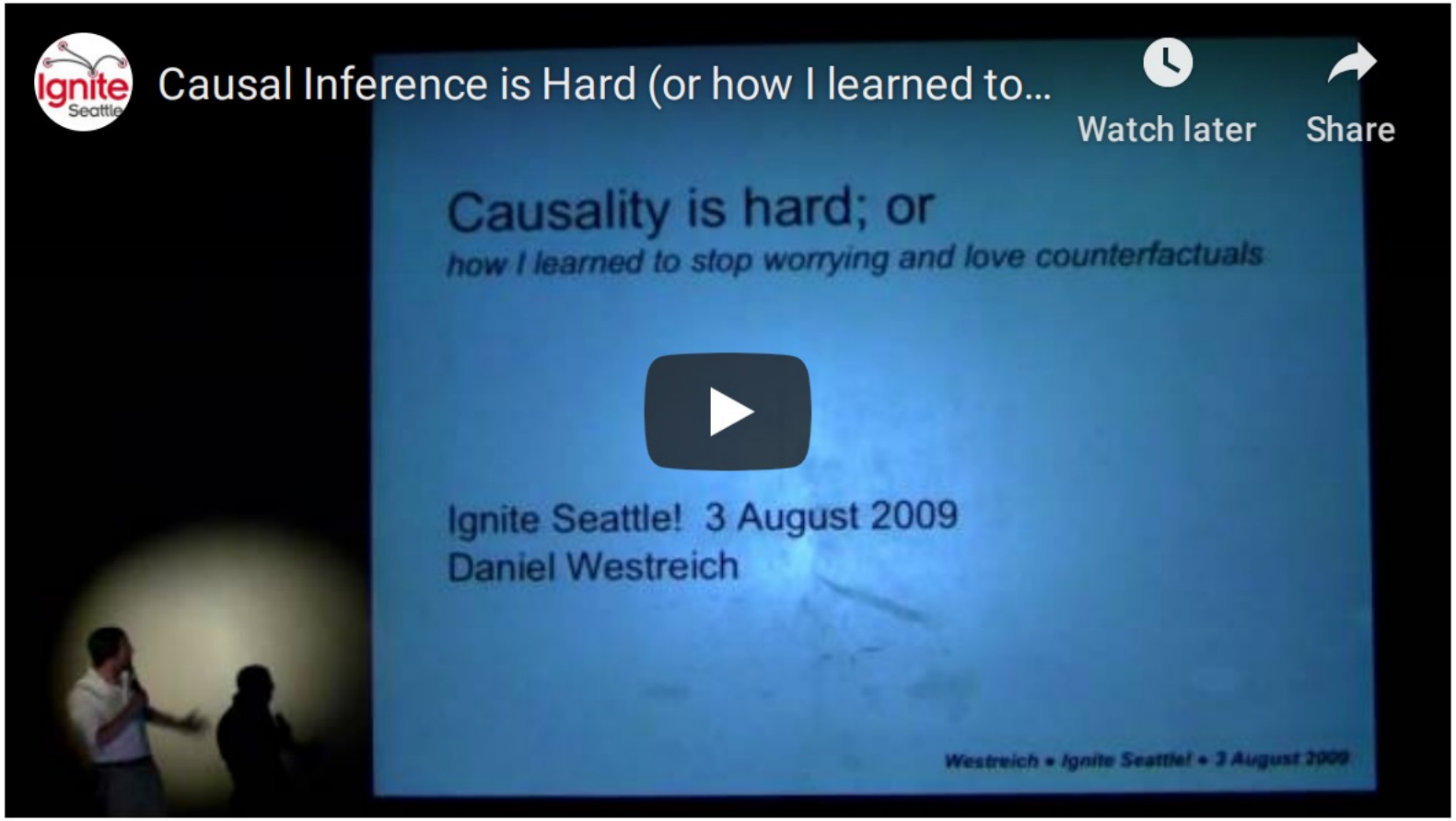
Quasi-natural experiments

Like a RCT, but that just “happen to occur naturally” (natural disasters, exogenous law changes…)

Econometric techniques

For the interested reader: space-time regression, instrumental variables, propensity score matching, differences-in-differences, regression discontinuity…

Causal Inference



Source: Causal Inference is hard @ Ignite Seattle

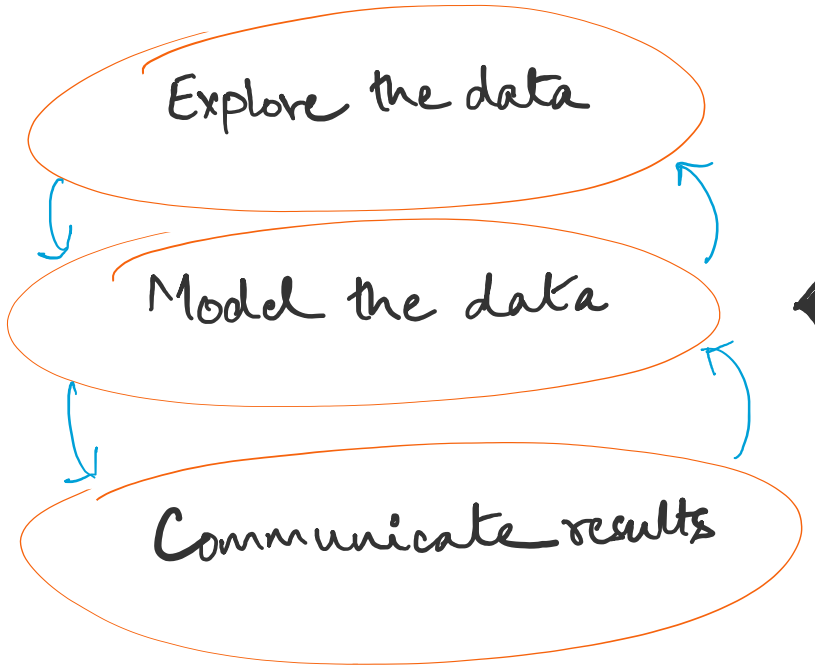
That’s it! The course is done.

After this course

You will be able to…

- **Obtain**: Obtaining data from multiple **open** data sources.
- **Scrub**: Data cleaning, munging, sampling to consolidate all information into a dataset that is manageable, informative and relates to your problem.
- **Explore**: Exploratory data analysis to make sense of what your data is trying to say.
- **Model**: Estimation and modelling based on statistical tools such as regression and clustering.
- **Interpret**: Communicating results and reflections through visualisation, storytelling and interpretable summaries.

For Q2, 3 and 4



DATA

Agent based modelling — Igor Nikolic

Network Science — Piet van Mieghem

Machine Learning — CS Masters @ EWI

Survey methodology — TPM

System Dynamics — Willem Auping + Jan Kwakkel

Simulation — Alexander Verbraeck

DATA

Visualisation Q1

Uncertainty Jan Kwakkel Q4

Thank you! 😇

Trivik Verma

Assistant Professor

Policy Analysis, MAS

@TrivikV

Internships at
https://cusp.tbm.tudelft.nl/opportunities/

Den Haag, NL

Image by Bichi Zhang