Welcome

“Data for Good: Ensuring the Responsible Use of Data to Benefit Society”

Jeannette Wing

Twitter Hashtag: #ACMLearning

Tweet questions & comments to: @ACMeducation

Post-Talk Discourse: https://on.acm.org

Additional Info:

- Talk begins at the top of the hour and lasts 60 minutes
- On the bottom panel you’ll find a number of widgets, including Twitter and Sharing apps
- For volume control, use your master volume controls and try headphones if it’s too low
- If you are experiencing any issues, try refreshing your browser or relaunching your session
- At the end of the presentation, you will help us out if you take the experience survey
- This session is being recorded and will be archived for on-demand viewing. You’ll receive an email when it’s available.
Data for Good: Ensuring the Responsible Use of Data to Benefit Society

Speaker: Jeannette Wing

Moderator: Paul Leidig
ACM.org Highlights

For Scientists, Programmers, Designers, and Managers:

- **Learning Center** - [https://learning.acm.org](https://learning.acm.org)
  - View past TechTalks & Podcasts with top inventors, innovators, entrepreneurs, & award winners
  - Access to O’Reilly Learning Platform – technical books, courses, videos, tutorials & case studies
  - Access to Skillsoft Training & ScienceDirect – vendor certification prep, technical books & courses

- **Ethical Responsibility** – [https://ethics.acm.org](https://ethics.acm.org)

**By the Numbers**

- 2,200,000+ content readers
- 1,800,000+ DL research citations
- $1,000,000 Turing Award prize
- 100,000+ global members
- 1160+ Fellows
- 700+ chapters globally
- 170+ yearly conferences globally
- 100+ yearly awards
- 70+ Turing Award Laureates

**Popular Publications & Research Papers**

- Communications of the ACM - [http://cacm.acm.org](http://cacm.acm.org)
- Queue Magazine - [http://queue.acm.org](http://queue.acm.org)
- Digital Library - [http://dl.acm.org](http://dl.acm.org)

**Major Conferences, Events, & Recognition**

- [https://www.acm.org/conferences](https://www.acm.org/conferences)
- [https://www.acm.org/chapters](https://www.acm.org/chapters)
- [https://awards.acm.org](https://awards.acm.org)
Welcome

“Data for Good: Ensuring the Responsible Use of Data to Benefit Society”

Jeannette Wing

Twitter Hashtag: #ACMLearning

Tweet questions & comments to: @ACMeducation

Post-Talk Discourse: https://on.acm.org

Additional Info:

- Talk begins at the top of the hour and lasts 60 minutes
- On the bottom panel you’ll find a number of widgets, including Twitter and Sharing apps
- For volume control, use your master volume controls and try headphones if it’s too low
- If you are experiencing any issues, try refreshing your browser or relaunching your session
- At the end of the presentation, you will help us out if you take the experience survey
- This session is being recorded and will be archived for on-demand viewing. You’ll receive an email when it’s available.
Data For Good:
Ensuring the Responsible Use of Data to Benefit Society

Jeannette M. Wing
Avanessians Director of the Data Science Institute and Professor of Computer Science
Columbia University
Adjunct Professor of Computer Science
Carnegie Mellon University

ACM Tech Talk
April 30, 2020
Data Life Cycle

- generation
- collection
- processing
- storage
- management
- analysis
- visualization
- interpretation

privacy and ethical concerns throughout
What is Data Science?

Definition:

Data science is the study of extracting value from data.
Mission

Advance the state of the art in data science

Transform all fields, professions, and sectors through the application of data science

Ensure the responsible use of data to benefit society
Tagline

Data for Good
17 Schools, Colleges, and Institutes

Graduate School of Architecture, Planning and Preservation
School of the Arts
Graduate School of Arts and Sciences
Barnard College
Columbia Business School
College of Dental Medicine
The Earth Institute
Columbia Engineering
School of International and Public Affairs
Columbia Journalism School
Columbia Law School
School of Nursing
Vagelos College of Physicians and Surgeons
Mailman School of Public Health
School of Social Work
Teachers College
Zuckerman Institute
Cross-Cutting Centers

data.science.columbia.edu/data-science-centers

Foundations
Computing Systems
Cybersecurity
Data, Media, and Society

Financial Analytics
Health Analytics
Smart Cities
Sense, Collect, and Move

Computational Social Science
Education
Materials Discovery Analytics
50% of all Columbia Business School students graduate with some data science knowledge.

Co-taught by Applied Math and History professors.
Industry Affiliates Program

industry.datascience.columbia.edu
Columbia-IBM Center on Blockchain and Data Transparency
Mission

Advance the state of the art in data science

Transform all fields, professions, and sectors through the application of data science

Ensure the responsible use of data to benefit society
Multiple Causal Inference

Understanding Causal Effect

What happens to movie revenue if we place an actor in a movie?

Goal: $E[Y_i(a)]$, $E[Y_i \mid do(a)]$

<table>
<thead>
<tr>
<th>Title</th>
<th>Cast</th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avatar</td>
<td>{Sam Worthington, Zoe Saldana, Sigourney Weaver, Stephen Lang, ...}</td>
<td>$2788M</td>
</tr>
<tr>
<td>Titanic</td>
<td>{Kate Winslet, Leonardo DiCaprio, Frances Fisher, Billy Zane, ...}</td>
<td>$1845M</td>
</tr>
<tr>
<td>The Avengers</td>
<td>{Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, ...}</td>
<td>$1520M</td>
</tr>
<tr>
<td>Jurassic World</td>
<td>{Chris Pratt, Bryce Dallas Howard, Irrfan Khan, Vincent D’Onofrio, ...}</td>
<td>$1514M</td>
</tr>
<tr>
<td>Furious 7</td>
<td>{Vin Diesel, Paul Walker, Dwayne Johnson, Michelle Rodriguez, ...}</td>
<td>$1506M</td>
</tr>
<tr>
<td>Avengers: Age of Ultron</td>
<td>{Robert Downey Jr., Chris Hemsworth, Mark Ruffalo, Chris Evans, ...}</td>
<td>$1405M</td>
</tr>
<tr>
<td>Frozen</td>
<td>{Kristen Bell, Idina Menzel, Jonathan Groff, Josh Gad, ...}</td>
<td>$1274M</td>
</tr>
<tr>
<td>Iron Man 3</td>
<td>{Robert Downey Jr., Gwyneth Paltrow, Don Cheadle, Guy Pearce, ...}</td>
<td>$1215M</td>
</tr>
<tr>
<td>Minions</td>
<td>{Sandra Bullock, Jon Hamm, Michael Keaton, Allison Janney, ...}</td>
<td>$1157M</td>
</tr>
<tr>
<td>Captain America: Civil War</td>
<td>{Chris Evans, Robert Downey Jr., Scarlett Johansson, Sebastian Stan, ...}</td>
<td>$1153M</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Many Applications
- **Confounders** affect both the causes and the outcomes.
- We should correct for all confounders in causal inference, which requires in theory to measure **all confounders**.
- But, whether we have measured all confounders is (famously) **untestable**.
New Idea: The Deconfounder

1. **Fit** a “local latent-variable model” of the assigned causes (e.g., Factor Analysis).
2. **Infer** the latent variable for each data point; it is a substitute confounder.
3. **Correct** for the substitute confounder in a causal inference.

\[
\hat{Z}_i = \mathbb{E}[Z_i | A_i = a_i]
\]

\[
\mathbb{E}[Y_i(a)] = \mathbb{E}[\mathbb{E}[Y_i(A) | A_i = a, Z_i]]
\]
New Idea: The Deconfounder

**MODEL**

Assigned Causes

**ESTIMATE**

Substitute Confounders

**ESTIMATE**

Causal Effects

Weaker assumptions: No unobserved single-cause confounder. (But no need to measure all confounders.)

Checkable procedure: We can check if the substitute confounder is good.

Unbiased inference: We prove the deconfounder gives unbiased causal inference.

\[
\begin{align*}
\{{\zhat_1, \ldots, \zhat_n}\} \\
\zhat_i &= E[Z_i \mid A_i = a_i] \\
E-Y(a) &= E[E[Y(A) \mid A = a, Z]]
\end{align*}
\]

Assumption: No unobserved single-cause confounder.
With the deconfounder,
(1) Sean Connery’s (James Bond) value goes up.
(2) Bernard Lee’s (M) and Desmond Llewelyn’s (Q) values go down.

We can now answer questions such as: What happens to revenue if we place Desmond Llewelyn in A Beautiful Mind? How about Sean Connery?

The deconfounder corrects for unobserved confounders: genre, sequel, etc.
Advance the state of the art in data science

Transform all fields, professions and sectors through the application of data science

Ensure the responsible use of data to benefit society
Biology and Big Data: Understanding Tumor Microbiome to Combat Cancer

Cosmology and Neural Networks

Monopsony: Economics and Machine Learning

Robo-Advising: Finance and Reinforcement Learning

Event Discovery: History and Topic Modeling

Distinguish between topics describing “business as usual” and those that deviate from such patterns.
Data for Good: responsible use of data
FAT* → Trustworthy AI

Fairness → Robustness
Accountability → Interpretability/Explainability
Transparency
Ethics
Safety → Reliability
Security → Availability
Privacy → Usability
DeepXplore: Testing Deep Learning Systems

DeepXplore

https://github.com/peikexin9/deepxplore

- Efficiently and systematically tests DNNs of hundreds of thousands of neurons without labeled data (only needs unlabeled seeds)
- Key ideas: neuron coverage (akin to code coverage), differential testing, and domain-specific constraints for focusing on realistic inputs
- Testing as a joint optimization problem (maximize both number of differences and neuron coverage)
- Found 1000s of fatal errors in 15 state-of-the-art DNNs for ImageNet, self-driving cars, and PDF/Android malware
DP and Machine Learning: PixelDP

Problem

Solution

1. Add a noise layer a la Differential Privacy

2. Provable guarantee from DP says classifier is robust to some degree of input perturbations.
Data for Good:
tackling societal grand challenges
PANGEO: Climate Science and Big Data

https://pangeo-data.github.io/

PI: Ryan Abernathey
(Dept. of Earth & Env. Sci., LDEO, Columbia University)

Co-PIs: Chiara Lepore, Michael Tippett, Naomi Henderson, Richard Seager (LDEO)
Kevin Paul, Joe Hamman, Ryan May, Davide Del Vento
(National Center for Atmospheric Research)
Matthew Rocklin (Anaconda; formerly Continuum Analytics)

Data Science and Agriculture

Main Results

- If India’s crop production continues to homogenize towards rice, food supply in the country may be more vulnerable to increasingly frequent climate shocks (e.g., droughts, extreme heat).

- Increasing the share of production contributed by coarse cereals (such as millets and sorghum) could improve the resilience of India’s food production against climatic changes, especially in the places where coarse cereal yields are already comparable to rice yields.

- More broadly, diversifying crop mixes in agriculturally important areas can help buffer against some aspects of climate change such as droughts and extreme heat.
Healthcare: Observational Health Data Sciences and Informatics (OHDSI, pronounced “Odyssey”)

Goal: 1 billion patient records for observational research
- 25 countries
- 200 researchers
- 80 databases
- 600 million patient records

Columbia University is the coordinating center

Heterogeneity of Observational Research Results
The Medical Deconfounder

- Extract EHRs from the OHDSI database
- Fit the medical deconfounder
- Analyze the causal effects of medications
- Evaluate the results by medical literature review

Treatment Effects on Hemoglobin A1c (Type 2 Diabetes)

- The unadjusted model
  \[ Y_i \sim \mathcal{N}\left(\sum_{j=1}^{D} \beta_j A_{ij}, \sigma^2\right) \]

- The medical deconfounder
  \[ Y_i \sim \mathcal{N}\left(\sum_{j=1}^{D} \beta_j A_{ij} + \sum_{k=1}^{K} \gamma_k \hat{Z}_{ik}, \sigma^2\right) \]

- The deconfounder reduces both false positive and false negative rates: acetaminophen (c2nc); amolodipine and hydrochlorothiazide (nc2c).

- It identifies effective (causal) drugs that are more consistent with the medical literature.
Data for Good
Thank you!
The Learning Continues...

TechTalk Discourse: https://on.acm.org
TechTalk Inquiries: learning@acm.org
TechTalk Archives: https://learning.acm.org/techtalks
Learning Center: https://learning.acm.org
Professional Ethics: https://ethics.acm.org
Queue Magazine: https://queue.acm.org
Data Science Task Force: http://dstf.acm.org/