The Blessings of Multiple Causes

David M. Blei
Departments of Computer Science and Statistics
Data Science Institute
Columbia University
We have *complicated data*; we want to *make sense* of it.
What is *complicated data*?

- many data points; many dimensions
- unstructured (e.g. text)
- multimodal and interconnected (e.g., images, links, text, clicks)
What is *making sense of data*?

- make predictions about the future
- identify interpretable patterns
- do science: confirm, elaborate, form causal theories
PROBABILISTIC MACHINE LEARNING

- ML methods that connect domain knowledge to data.
- Provides a computational methodology for scalable modeling
- Goal: A methodology that is expressive, scalable, easy to develop
Communities discovered in a 3.7M node network of U.S. Patents

[Gopalan and Blei PNAS 2013]
Neuroscience analysis of 220 million fMRI measurements

[Manning+ PLOS ONE 2014]
| 1 | Game  
    | Season  
    | Team  
    | Coach  
    | Play  
    | Points  
    | Games  
    | Giants  
    | Second  
    | Players  |
|---|---|
| 2 | Life  
    | Know  
    | School  
    | Street  
    | Man  
    | Family  
    | Says  
    | House  
    | Children  
    | Night  |
| 3 | Film  
    | Movie  
    | Show  
    | Life  
    | Television  
    | Films  
    | Director  
    | Man  
    | Story  
    | Says  |
| 4 | Book  
    | Life  
    | Books  
    | Novel  
    | Story  
    | Man  
    | Author  
    | House  
    | War  
    | Children  |
| 5 | Wine  
    | Street  
    | Hotel  
    | House  
    | Room  
    | Night  
    | Place  
    | Restaurant  
    | Park  
    | Garden  |
| 6 | Bush  
    | Campaign  
    | Clinton  
    | Republican  
    | House  
    | Party  
    | Democratic  
    | Political  
    | Democrats  
    | Senator  |
| 7 | Building  
    | Street  
    | Square  
    | Housing  
    | House  
    | Buildings  
    | Development  
    | Space  
    | Percent  
    | Real  |
| 8 | Won  
    | Team  
    | Second  
    | Race  
    | Round  
    | Cup  
    | Open  
    | Game  
    | Play  
    | Win  |
| 9 | Yankees  
    | Game  
    | Mets  
    | Season  
    | Run  
    | League  
    | Baseball  
    | Team  
    | Games  
    | Hit  |
| 10 | Government  
    | War  
    | Military  
    | Officials  
    | Iraq  
    | Forces  
    | Iraqi  
    | Army  
    | Troops  
    | Soldiers  |
| 11 | Children  
    | School  
    | Women  
    | Family  
    | Parents  
    | Child  
    | Life  
    | Says  
    | Help  
    | Mother  |
| 12 | Stock  
    | Percent  
    | Companies  
    | Fund  
    | Market  
    | Bank  
    | Investors  
    | Funds  
    | Financial  
    | Business  |
| 13 | Church  
    | War  
    | Women  
    | Life  
    | Black  
    | Political  
    | Catholic  
    | Government  
    | Jewish  
    | Pope  |
| 14 | Art  
    | Museum  
    | Show  
    | Gallery  
    | Works  
    | Artists  
    | Street  
    | Artist  
    | Paintings  
    | Exhibition  |
| 15 | Police  
    | Yesterday  
    | Man  
    | Officer  
    | Officers  
    | Case  
    | Found  
    | Charged  
    | Street  
    | Shot  |
Population analysis of 2 billion genetic measurements

[Gopalan+ Nature Genetics 2016]
(Fancy) discrete choice analysis of 5.7M purchases [Ruiz+ 2017]
Customized data analysis is important to many fields.

Pipeline separates **assumptions, computation, application**

Eases collaborative solutions to statistics/ML problems
Causal inference from observational data

- How can we understand the world through observation?
- Important to genetics, economics, physics, medicine, finance, ...
- Today: Use probabilistic machine learning for causal inference
This is joint work with Yixin Wang (Statistics)
Credit → Yixin
(Blame → Dave)
An introduction to the deconfounder
A frivolous causal inference problem

- Data about movies: casts and revenue
- Goal: Understand the causal effect of putting an actor in a movie
- Causal: “What will the revenue be if we make a movie with a particular cast?”
The naive solution

<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>
</tbody>
</table>

- Naive solution: Fit a regression (or use deep learning)
- Actors are features; revenue is the response
- Estimates revenue as a function of which actors are cast
The naive solution

<table>
<thead>
<tr>
<th>Title</th>
<th>Cast</th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avatar</td>
<td>{Sam Worthington, Zoë 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>

But standard ML does not (necessarily) provide causal inferences

- Whether an actor was cast is different from casting an actor

- Causal inference is about **prediction under intervention**
- James Bond movies are about James Bond, a British spy
- Cast James Bond, M, Q, Ms. Moneypenny
- M, Q, Ms. Moneypenny only appear in Bond movies
- Bond movies always do well at the box office
The naive solution

- James Bond-ness is an **unobserved confounder**.
- Confounders affect both the cast (“causes”) and the revenue (“effect”)
- Unobserved confounders bias ML (if used for intervention)
  - M is in the movie $\rightarrow$ James Bond-ness $\rightarrow$ big profits
  - Thus some actors overestimated; others underestimated
Unobserved confounders are everywhere.
What is causal inference?

- Causal inference is about **prediction under intervention**.
- “What will the revenue be if we make a movie with a particular cast?”
- Challenge: Unobserved confounders (like James Bond-ness)
The classical solution

**Think about confounders**

```
\text{James Bond''}
```

**Measure confounders**

\[ \{w_1, \ldots, w_n\} \]

**Estimate causal effects**

\[ \mathbb{E}[Y \mid \text{do}(a)] = \mathbb{E}\left[\mathbb{E}[Y \mid W, A=a]\right] \]

**Data**

\[
\begin{align*}
\{\text{actors}_1, \text{revenue}_1\} \\
\{\text{actors}_2, \text{revenue}_2\} \\
\vdots \\
\{\text{actors}_n, \text{revenue}_n\}
\end{align*}
\]
This approach requires that we find and measure **sufficient confounders**.

But whether we included sufficient confounders is **untestable**.

The classical solution rests on **hope**. (And it makes us **worry**.)
Multiple causal inference

<table>
<thead>
<tr>
<th>Title</th>
<th>Cast</th>
<th>Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Avatar</em></td>
<td>{Sam Worthington, Zoe Saldana, Sigourney Weaver, Stephen Lang, ...}</td>
<td>$2788M</td>
</tr>
<tr>
<td><em>Titanic</em></td>
<td>{Kate Winslet, Leonardo DiCaprio, Frances Fisher, Billy Zane, ...}</td>
<td>$1845M</td>
</tr>
<tr>
<td><em>The Avengers</em></td>
<td>{Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, ...}</td>
<td>$1520M</td>
</tr>
<tr>
<td><em>Jurassic World</em></td>
<td>{Chris Pratt, Bryce Dallas Howard, Irrfan Khan, Vincent D’Onofrio, ...}</td>
<td>$1514M</td>
</tr>
<tr>
<td><em>Furious 7</em></td>
<td>{Vin Diesel, Paul Walker, Dwayne Johnson, Michelle Rodriguez, ...}</td>
<td>$1506M</td>
</tr>
<tr>
<td><em>Avengers: Age of Ultron</em></td>
<td>{Robert Downey Jr., Chris Hemsworth, Mark Ruffalo, Chris Evans, ...}</td>
<td>$1405M</td>
</tr>
<tr>
<td><em>Frozen</em></td>
<td>{Kristen Bell, Idina Menzel, Jonathan Groff, Josh Gad, ...}</td>
<td>$1274M</td>
</tr>
<tr>
<td><em>Iron Man 3</em></td>
<td>{Robert Downey Jr., Gwyneth Paltrow, Don Cheadle, Guy Pearce, ...}</td>
<td>$1215M</td>
</tr>
<tr>
<td><em>Minions</em></td>
<td>{Sandra Bullock, Jon Hamm, Michael Keaton, Allison Janney, ...}</td>
<td>$1157M</td>
</tr>
<tr>
<td><em>Captain America: Civil War</em></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>

- But our problem is not classical.
- There are many causes (one per actor)—**multiple causal inference**
- **Multiple causes helps construct a variable that contains confounders.**
The deconfounder

**MODEL**
- **ASSIGNED CAUSES**

**ESTIMATE**
- **SUBSTITUTE CONFOUNDERS**

\[
\{ \hat{Z}_i, \ldots, \hat{Z}_n \}
\]

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

**ESTIMATE CAUSAL EFFECTS**

\[
\mathbb{E}[Y | \text{do}(a)] = \mathbb{E}[\mathbb{E}[Y | Z, A=a]]
\]

**DATA**

\[
\begin{align*}
\{ \text{actors}_1, & \text{revenue}_1 \\
\text{actors}_2, & \text{revenue}_2 \\
\vdots, & \vdots \\
\text{actors}_n, & \text{revenue}_n 
\end{align*}
\]

actors

revenue

2

\ldots

actors
Find, fit, and check a factor model of the assigned causes.

Use the model to form substitute confounders for each individual.

Use the substitute confounders in a causal model of the outcome.
The deconfounder

- Blends **unsupervised learning**, **model checking**, **causality**
- Provides a **checkable** approach to multiple causal inference
- Requires **weaker assumptions** than classical causal inference.
The deconfounder

- Find, fit, and check a **probabilistic matrix factorization** of movie casts.
- Use the model to infer the **per-movie variables** in the matrix factorization.
- Use these variables in a **regression from casts to earnings**.
How do genes affect a trait?

- The causes are genes (or SNPs).
- The effect is a trait (or phenotype).
- Confounder: Each person’s ancestry (induces correlation in *multiple genes*)
How do sports players affect how well the team is doing?

- The causes are who is in the game.
- The effect is the points scored in the game.
- Confounder: The coach (uses multiple players together)
How do prices of items affect how much money is spent?

- The causes are the prices of each item for sale.
- The effect is how much money is spent.
- Confounder: Holidays (affect the prices and demand of multiple items)
**Intuition and assumptions**

- **Definitions**
  - *multi-cause confounder*: affects multiple causes and outcome
  - *single-cause confounder*: affects one cause and outcome

- **Assumption: No unobserved single-cause confounders**
  - But this is weaker than “no unobserved confounders”
Intuition and assumptions

- Intuition: Multi-cause confounders induce dependence among the causes.
- We observe that dependence, which is “data” about the confounders.
- We can capture that data with a factor model.
Case study: Actors

- “Overestimated”:
- “Underestimated”:
- Most “corrected”: 
The deconfounder in more detail
Multiple causal inference

<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>
</tbody>
</table>

\[ \mathcal{D} = \{(a_1, y_1), \ldots, (a_n, y_n)\} \]

- assigned causes \( a_i = \{a_{i1}, \ldots, a_{im}\} \)
- outcome \( y_i \)

\[ \mathbb{E}[Y \mid \text{do}(a)] \]

- Observed dataset \( \mathcal{D} = \{(a_1, y_1), \ldots, (a_n, y_n)\} \)
### Multiple causal inference

<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>
</tbody>
</table>

- If there are unobserved confounders then

\[
\mathbb{E}[Y \mid \text{do}(a)] \neq \mathbb{E}[Y \mid A = a].
\]

- We can calculate the right term from data, but it’s not equal to the left term.
Find, fit, and check a factor model of the movie casts.
Use the factor model to form substitute confounders for each movie.
Use the substitute confounders in a causal model of movie revenue.
Fit a probabilistic factor model

A probabilistic factor model has the following form,

\[ \beta_j \sim p(\beta_j) \quad j = 1, \ldots, m \]
\[ z_i \sim p(z_i) \quad i = 1, \ldots, n \]
\[ a_{ij} \sim p(a_{ij} \mid z_i, \beta_j). \]

E.g., mixtures, matrix factorization, deep generative models, topic models, ...
Poisson factorization [Gopalan+ 2015]

\[ \begin{align*}
\beta_{jk} & \sim \text{Gam}(a, b) & i \in \{1, \ldots, n\} \\
z_{ik} & \sim \text{Gam}(a, b) & j \in \{1, \ldots, m\} \\
a_{ij} & \sim \text{Poi}(z_i^\top \beta_j) & k \in \{1, \ldots, d\}
\end{align*} \]

- Provides a generative model of the assigned causes \( a_{ij} \).
- Can be fit to large datasets with variational methods
- A Bayesian form of non-negative matrix factorization [Lee and Seung 1999]
Consider the dataset of casts \( a_{1:n} \).

Approximate the posterior distribution \( p(z_{1:n}, \beta_{1:p} \mid a_{1:n}) \).

We only model the actors \( a_i \); the outcome is not involved.
Check the factor model

- Estimate the local latent variable \( \hat{z}_i = \mathbb{E}_{\text{model}}[Z | a_i, \beta] \).
- Check how well \( \hat{z}_i \) captures the distribution of the actors.
- E.g., use a **predictive check** on actors. (No need for exact inference.)
Check the factor model

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictive score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probabilistic PCA</td>
<td>0.14</td>
</tr>
<tr>
<td>Poisson factorization</td>
<td>0.16</td>
</tr>
<tr>
<td>Mixtures</td>
<td>0.01</td>
</tr>
<tr>
<td>Deep exponential families</td>
<td>0.19</td>
</tr>
</tbody>
</table>
The estimated local variables $\hat{z}_i$ are substitute confounders.

They are latent attributes of movie casts that the factorization has discovered.

Form an augmented dataset of triplets $(a_i, y_i, \hat{z}_i)$. 

---

Form substitute confounders

<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>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>
</tbody>
</table>
Do causal inference

<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>
</tbody>
</table>

- The estimated local variables $\hat{z}_i$ are **substitute confounders**.
- They are latent attributes of movie casts that the factorization has discovered.
- Form an **augmented dataset** of triplets $(a_i, y_i, \hat{z}_i)$.  

Table 1: Top revenue movies

Table 2: Bottom revenue movies
Do causal inference

<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>
</tbody>
</table>

- Use the substitute confounders in a **causal inference**.
- E.g., fit regression from casts and confounders to revenue,

\[
\mathbb{E}[Y \mid a, \hat{z}] = \beta^\top a + \eta^\top \hat{z}.
\]

- Then use “backdoor adjustment” to perform causal inference,

\[
\mathbb{E}[Y \mid \text{do}(a)] = \mathbb{E}_{\hat{Z}}[\mathbb{E}[Y \mid a, \hat{Z}]]
\].
The deconfounder

- Find, fit, and check a **factor model** of the movie casts.
- Use the factor model to form **substitute confounders** for each movie.
- Use the substitute confounders in a **causal model** of movie revenue.
Case study: Actors

- “Overestimated”:

- “Underestimated”:

- Most “corrected”:
A little theory
Suppose we fit a **good factor model** of the assigned causes (the actors).

Then its local latent variable will contain **multi-cause confounders**.

Main assumption: No single cause confounders.
If we find a good factor model then

\[ p(a_{i1}, \ldots, a_{im} | z_i, \beta_1:m) = \prod_{j=1}^{m} p(a_{ij} | z_i, \beta_j) \]
There cannot be an unobserved \textbf{multi-cause confounder}.

Contradiction: If one existed then the independence statement would not hold.
Intuition (through graphical models)

▶ Note: there still might be a **single-cause confounder**
▶ This is a weaker assumption than “strong ignorability.”
Simulation study
We did many simulations and studies.
Example: Genome-wide association studies (GWAS)

- GWAS is a problem of multiple causal inference
- How is genetic variation causally connected to a trait?
- For each individual: a trait and many measurements of the genome (SNPs).
Example: Genome-wide association studies (GWAS)

- Multiple-cause confounding is a problem.
- Non-causal SNPs may be highly correlated to causal SNPs
- Misestimates causal effects
Algorithm 1: The Deconfounder

Input: a dataset of assigned causes and outcomes \( \{(a_i, y_i)\}, i = 1,..., n \)

Output: the average potential outcome \( E[Y(a)] \) for any causes \( a \)

repeat
  choose an assignment model from the class in Equation (4)
  fit the model to the assigned causes \( \{a_i\}, i = 1,..., n \)
  check the fitted model until the assignment check is satisfactory
foreach datapoint \( i \) do
  calculate \( \hat{z}_i = E[\hat{M}[Z|a_i]] \).
end
repeat
choose an outcome model from Equation (29)
fit the outcome model to the augmented dataset \( \{(a_i, y_i, \hat{z}_i)\}, i = 1,..., n \)
check the fitted outcome model until the outcome check is satisfactory
estimate the average potential outcome \( E[Y(a)] \) by Equation (28)

causally affect height? Here we give a brief account of how to use the deconfounder, omitting many of the details. We analyze GWAS problems extensively in Section 3.2.

Imagine we collect a dataset of \( n = 5,000 \) individuals; for each individual, we measure height and genotype, specifically the alleles at \( m = 100,000 \) locations, called the single-nucleotide polymorphisms (SNPs).

Each SNP is represented by a count of 0, 1, or 2; it encodes how many of the individual's two nucleotides differ from the most common pair of nucleotides at the location.

Table 2 illustrates a snippet of the data (10 individuals).

| ID (i) | SNP_1  | SNP_2  | SNP_3  | SNP_4  | SNP_5  | SNP_6  | SNP_7  | SNP_8  | SNP_9  | ... | SNP_100K       | Height (feet) |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|     | (a_i,100K)      |               |
|       | (a_{i,1}) | (a_{i,2}) | (a_{i,3}) | (a_{i,4}) | (a_{i,5}) | (a_{i,6}) | (a_{i,7}) | (a_{i,8}) | (a_{i,9}) | ... | (a_{i,100K})   |               |
| 1     | 1      | 0      | 0      | 1      | 0      | 0      | 1      | 2      | 0      | ... | 0              | 5.73          |
| 2     | 1      | 2      | 2      | 1      | 2      | 1      | 1      | 0      | 1      | ... | 2              | 5.26          |
| 3     | 2      | 0      | 1      | 1      | 0      | 1      | 0      | 1      | 1      | ... | 2              | 6.24          |
| 4     | 0      | 0      | 0      | 1      | 1      | 0      | 1      | 2      | 0      | ... | 0              | 5.78          |
| 5     | 1      | 2      | 1      | 1      | 1      | 0      | 1      | 0      | 0      | ... | 1              | 5.09          |
| ...   | ...    | ...    | ...    | ...    | ...    | ...    | ...    | ...    | ...    | ... | ...            | ...           |

- Generate SNPs \( a_{ij} \), where each individual belongs to a latent group \( c_i \).
- The true outcome is a trait \( y_i \), drawn from

\[
y_i = \sum_j \beta_j a_{ij} + \lambda_{c_i} + \varepsilon_i \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_{c_i}),
\]

where many \( \beta_j \) are zero, i.e., non-causal SNPs.
- Confounded: the intercept \( \lambda_{c_i} \) and error \( \varepsilon_i \) are connected to the latent group.
Simulation study

<table>
<thead>
<tr>
<th></th>
<th>pred. score</th>
<th>Real-valued outcome</th>
<th>Binary outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RMSE×10^2</td>
<td>RMSE×10^2</td>
</tr>
<tr>
<td>No control</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Control for confounders*</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>(G)LMM</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>PPCA</td>
<td>0.14</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>PF</td>
<td>0.15</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>LFA</td>
<td>0.14</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Mixture</td>
<td>0.00</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>DEF</td>
<td>0.20</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

- Predictive checks indicate downstream causal performance.
- The deconfounder provides good causal estimates.
- All computation was done in Edward [Tran+ 2018].
Simulation study

<table>
<thead>
<tr>
<th>Pred. score</th>
<th>Real-valued outcome RMSE×10²</th>
<th>Binary outcome RMSE×10²</th>
</tr>
</thead>
<tbody>
<tr>
<td>No control</td>
<td>58.82</td>
<td>29.50</td>
</tr>
<tr>
<td>Control for confounders*</td>
<td>25.32</td>
<td>25.77</td>
</tr>
<tr>
<td>(G)LMM</td>
<td>35.18</td>
<td>28.87</td>
</tr>
<tr>
<td>PPCA</td>
<td>33.32</td>
<td>26.70</td>
</tr>
<tr>
<td>PF</td>
<td>33.38</td>
<td>26.84</td>
</tr>
<tr>
<td>LFA</td>
<td>33.93</td>
<td>26.83</td>
</tr>
<tr>
<td>Mixture</td>
<td>57.59</td>
<td>29.96</td>
</tr>
<tr>
<td>DEF</td>
<td>26.47</td>
<td>25.91</td>
</tr>
</tbody>
</table>

- Predictive checks indicate downstream causal performance.
- The deconfounder provides good causal estimates.
- All computation was done in Edward [Tran+ 2018].
Explains and justifies existing methods for GWAS

- Linear mixed models [Yu+ 2006; Kang+ 2008; etc.]
- Principal component analysis [Price+ 2006]
- Logistic factor analysis [Song+ 2015; Hao+ 2015]
- Mixed-membership models [Pritchard+ 2000a,b; Falush+ 2003; Falush+ 2007]
- Deep generative models [Tran and Blei 2018]
Discussion
Causal inference from observational data

➤ How can we understand the world through observation?
➤ Important to genetics, economics, physics, medicine, finance, ...
➤ Today: Use probabilistic machine learning for causal inference
The deconfounder

- Find, fit, and check a **factor model** of the assigned causes.
- Use the factor model to form **substitute confounders** for each individual.
- Use the substitute confounders in a **causal model** of the outcome.
Suppose we fit a good factor model of the assigned causes (the actors).

Then its local latent variable will contain multi-cause confounders.

(There are assumptions.)
The deconfounder

- Blends **unsupervised learning, model checking, causality**
- Provides **checkable** approach to multiple causal inference
- Requires **weaker assumptions** than classical causal inference.
Further reading and current research


► Other readings
  – Tran and Blei (2018), ICLR
  – Ranganath and Perotte (2018), arXiv 1805.08273

► Current research
  – SEMs: selection bias, connections between causes
  – measures of confidence
  – the bias-variance trade-off
  – the effect of model choice and overlap
  – many applications
Extra slides
Theorem and assumptions
THEOREM: THE DECONFOUNDER

Suppose $p_{\text{true}}(a)$ can be written as
$$
\int p(z) \prod_j p(a_j | z, \beta) dz.
$$

Then $Z$ blocks the backdoor path between the causes and the effect.

This implies that,
$$
\mathbb{E} [Y | \text{do}(a)] = \mathbb{E}_Z [\mathbb{E}_Y [Y | Z, a]].
$$

Thus we can estimate the interventional expectation.
Theory: It works

THE ASSUMPTIONS

- No unobserved single-cause confounders.
- DAG: Variables $a$ are ancestors of $y$; the variable $z$ is an ancestor of $a$.
- Overlap: Conditional on the substitute confounder
  $$p(a | z) > 0 \quad \text{for all } a.$$
- Identifiability of the factor model (e.g., Bayesian factor models)
Checking a factor model
Predictive checks

- Goal: Make sure the factor model fits the data well.
- Our checks are similar to classical posterior predictive checks.
- Generate replicated datasets from the fitted factor model.
- Compare replicated data to real data
Predictive checks

- Statistic: Expected predictive log-likelihood on heldout entries
  \[ t(a_{i,\text{heldout}}) = \mathbb{E}[\log p(a_{i,\text{heldout}} | Z) | a_{i,\text{obs}}] \]

- Compute predictive scores
  \[ \text{predictive-score}_{i} = p\left(t(a_{i,\text{heldout}}^{\text{rep}}) < t(a_{i,\text{heldout}})\right). \]

- A factor model passes the check if the mean of the predictive scores > 0.1. (This is a subjective criterion.)
Identification and the bias-variance trade off
The bias-variance tradeoff

- The factor model should fit the assigned causes well. It makes sure the substitute confounder captures all multi-cause confounders; hence it leads to unbiased causal estimates.

- We lose efficiency in causal-effect estimation when the factor model fits the assigned causes too well, e.g., when the latent space has too many dimensions. (The resulting causal estimates are still unbiased.)

- The latent space can theoretically have more dimensions than the number of causes, but it might result in highly variable causal-effect estimates.
The bias-variance tradeoff

- Notice the predictive check can tell when the factor model overfits. When the check returns a score very close to 1, the factor model is likely overfitting.

- For an optimal bias-variance tradeoff, we choose the smallest factor model that passes the predictive check.
Causal identification holds when the factor model is identifiable.

Identifiability of the factor model is guaranteed when the expectation of its latent variable exists, i.e., most probabilistic factor models.

Causal identification can break when both the factor model is (nonprobabilistic) PCA and the outcome model is linear.

Nonlinearity or non-Gaussianity can recover causal identification.

(See Alex D’Amour’s blog post and our paper.)