Interpretable Data Analysis with Causality and Explanations

Sudeepa Roy

DUKE COMPUTER SCIENCE

Joint work with
Lise Getoor, Cynthia Rudin*, Dan Suciu, Alexander Volfovsky*, Babak Salimi, Boris Glavic, Harsh Parikh, Zhengjie Miao, Marco Morucci, M. Usaid Awan, Tianyu Wang, Vittorio Orlandi, Moe Kayali, Yameng Liu, Awa Dieng, Laurel Orr, Qitian Zeng, …

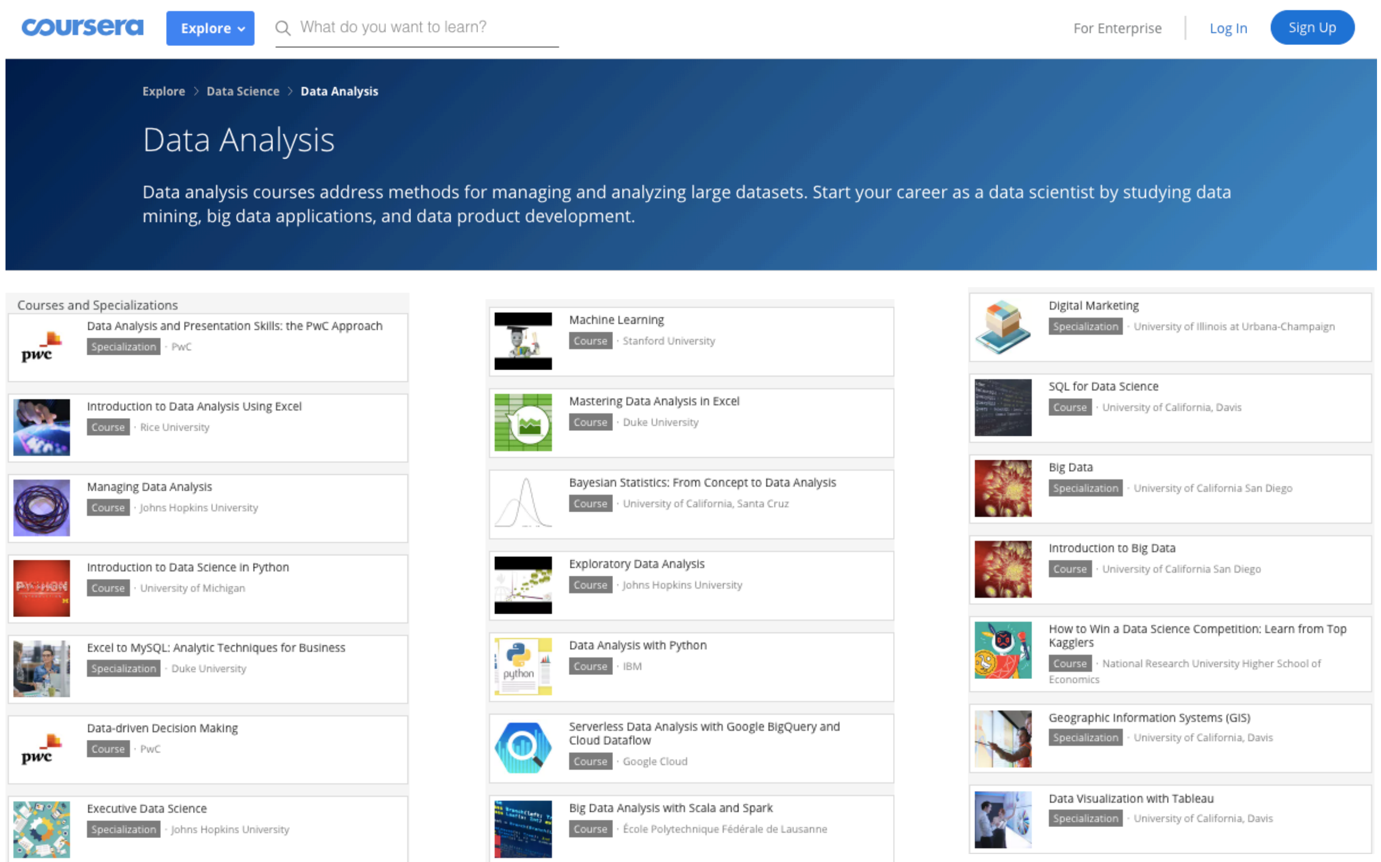
Presented at
Workshop on Social Science Modeling for Big Data in the World of Machine Learning
for the National Institute of Aging
The National Academies of Sciences, Engineering, and Medicine
October 24, 2019

* Some slides are from Cynthia and Alex!

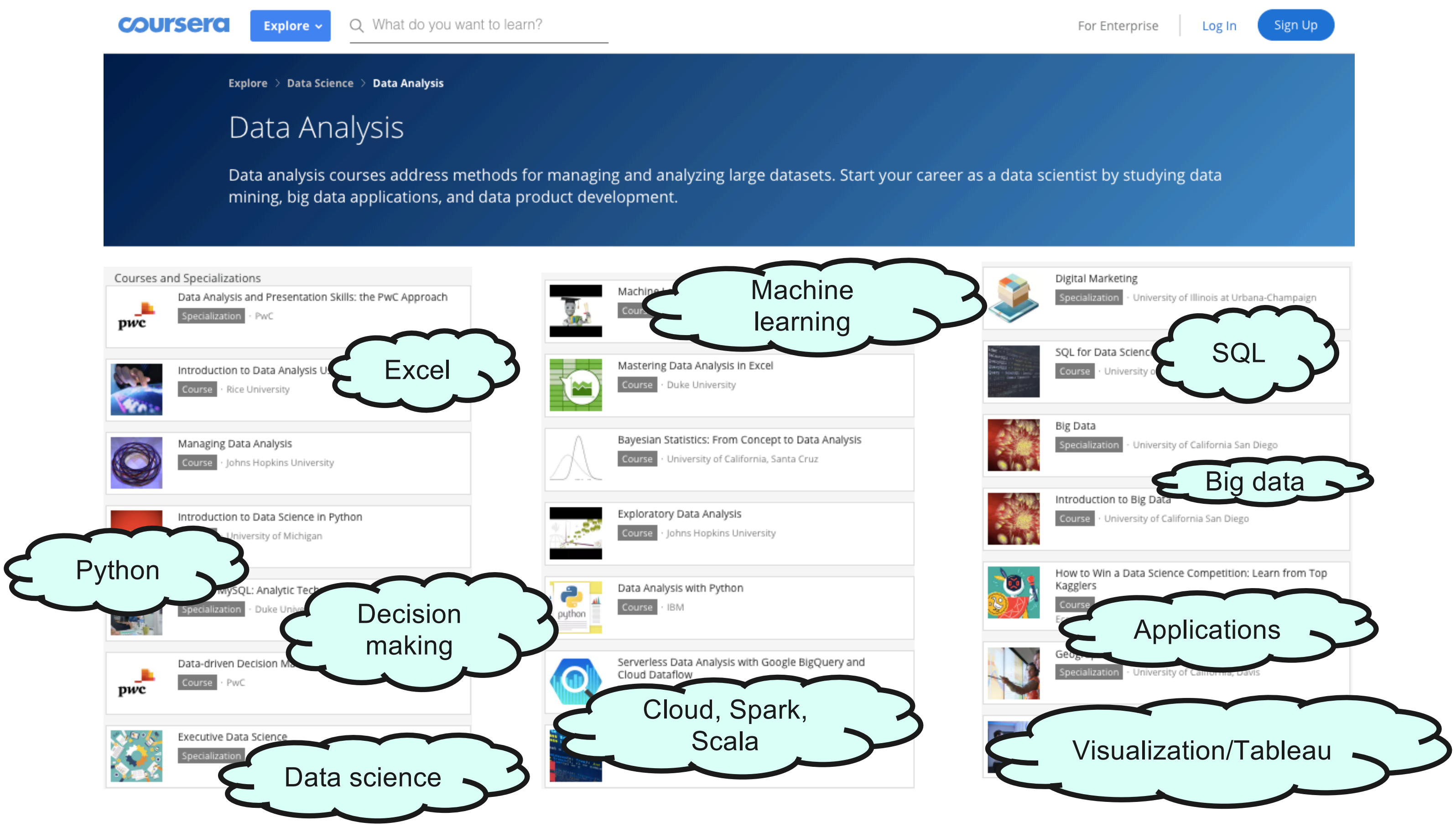
Data Analysis

* **Data**
* Advances in ML
* Computing resources
* Interests & applications
(Democratization of Data)

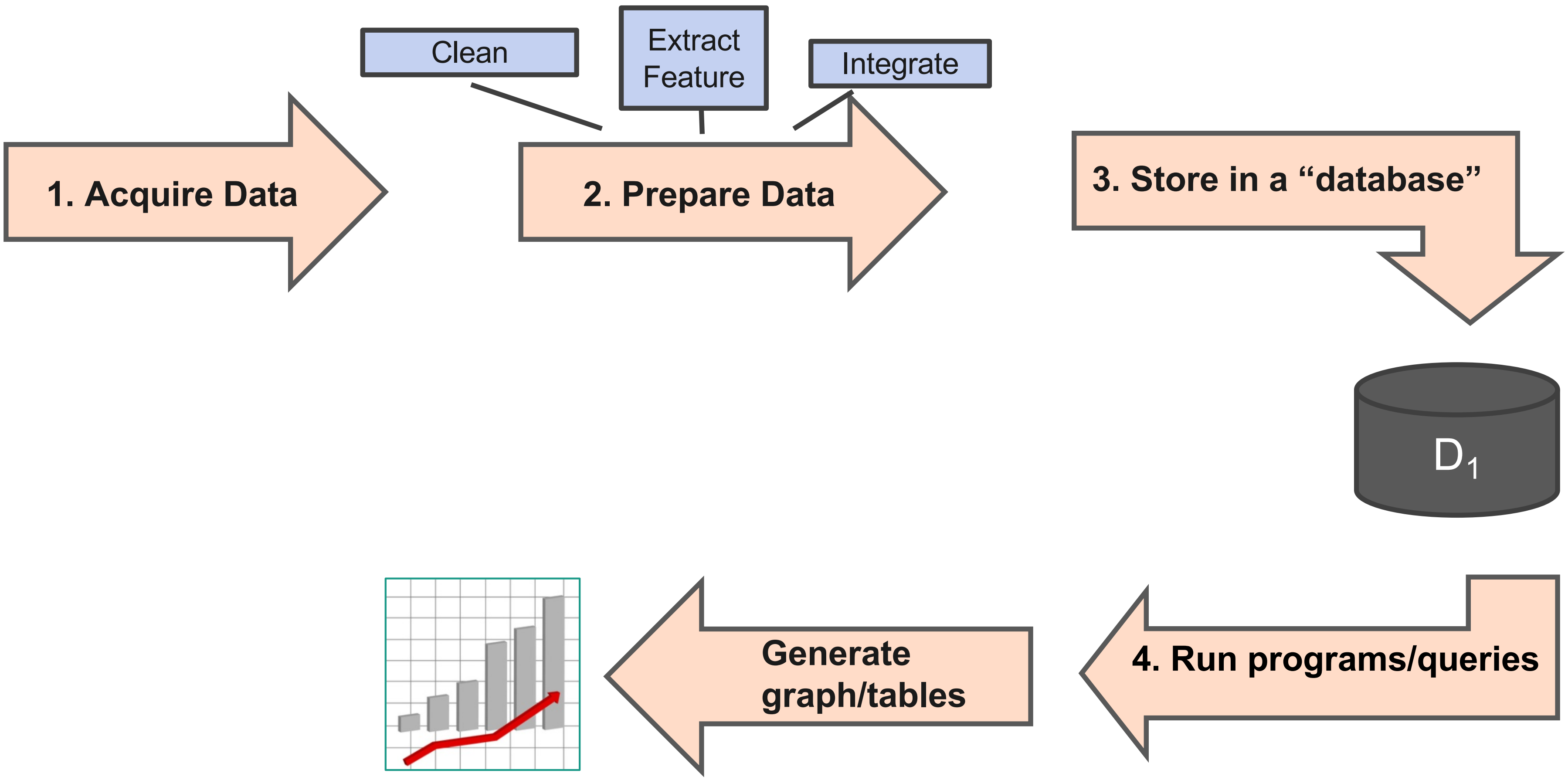
What is Data Analysis?



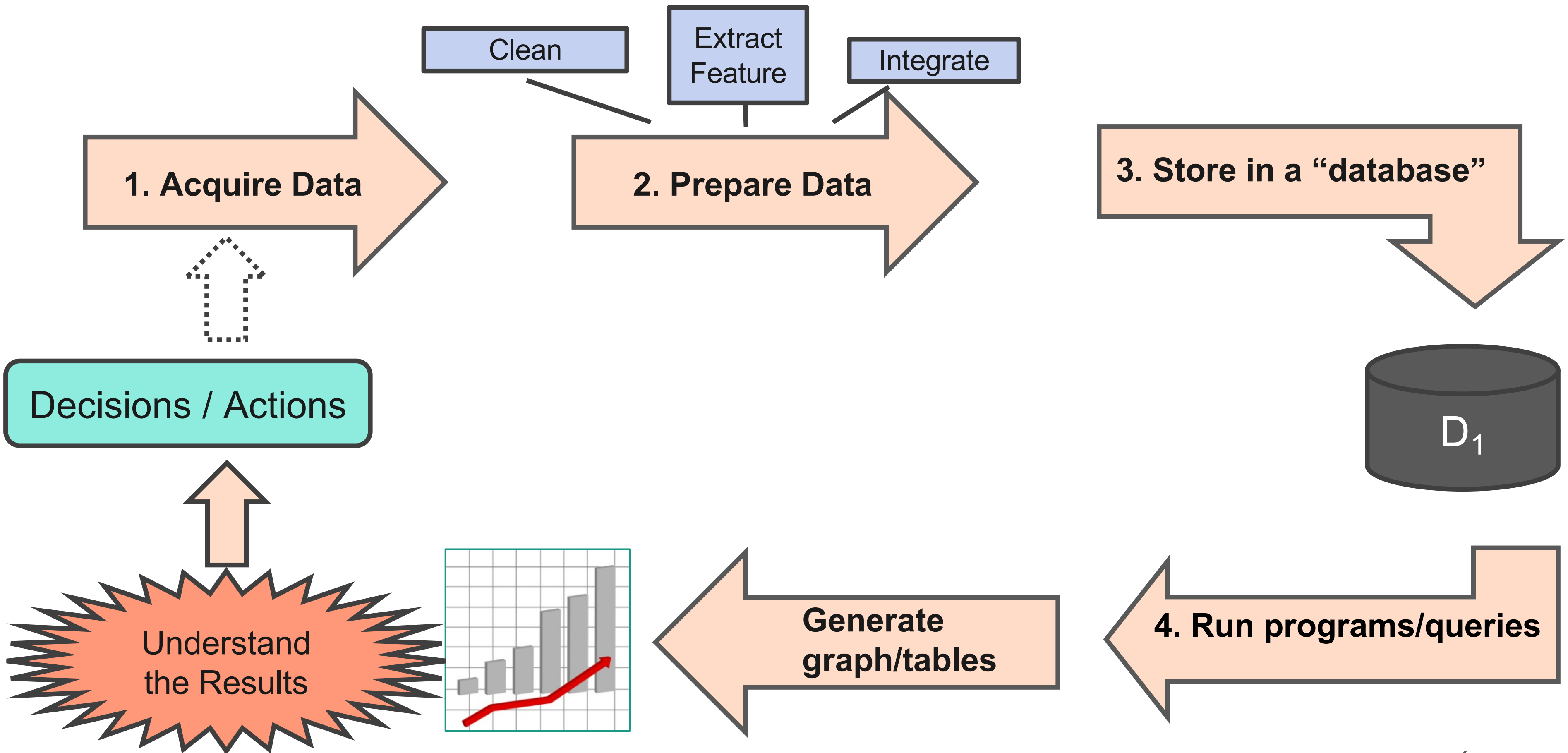
What is Data Analysis?



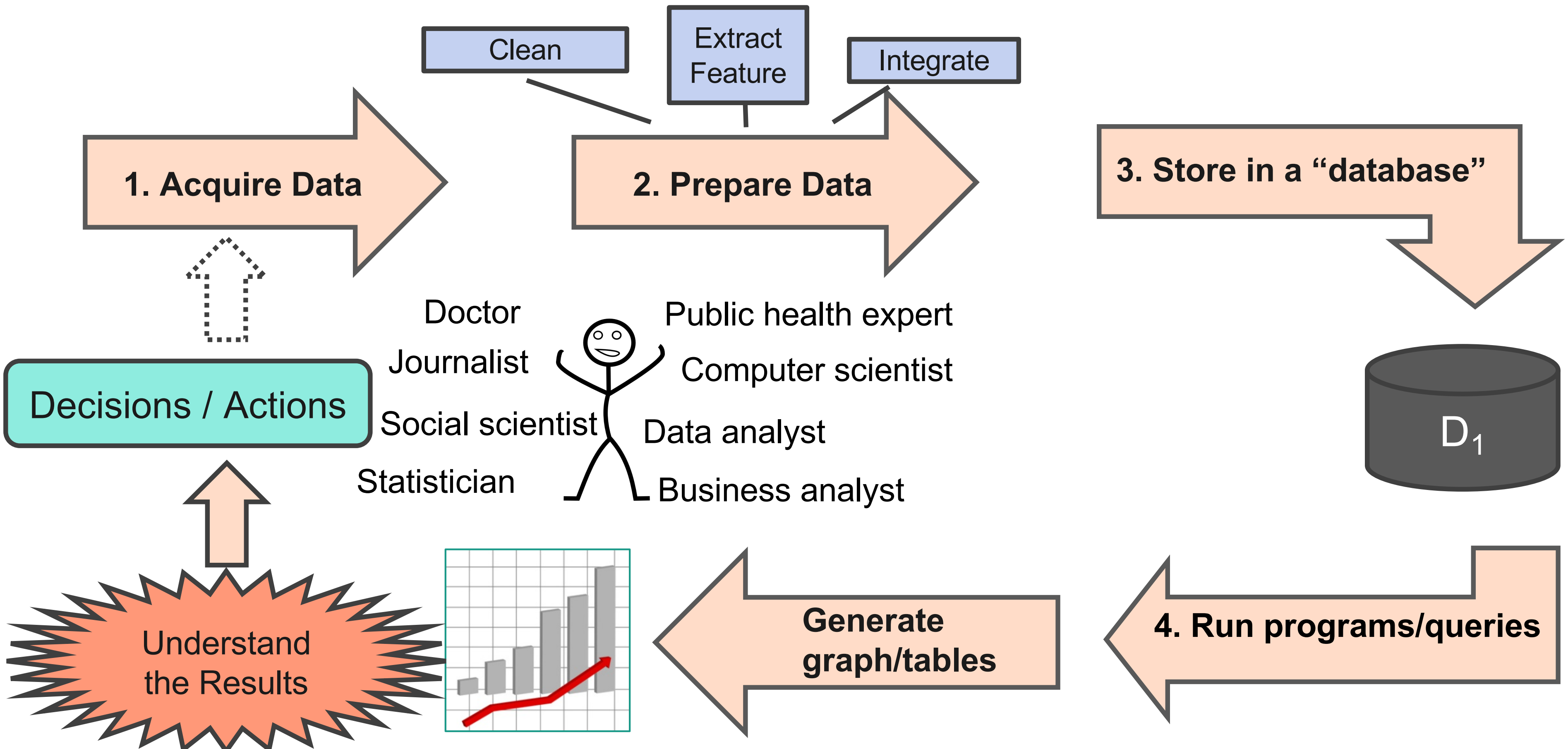
Data Analysis Loop



Data Analysis Loop

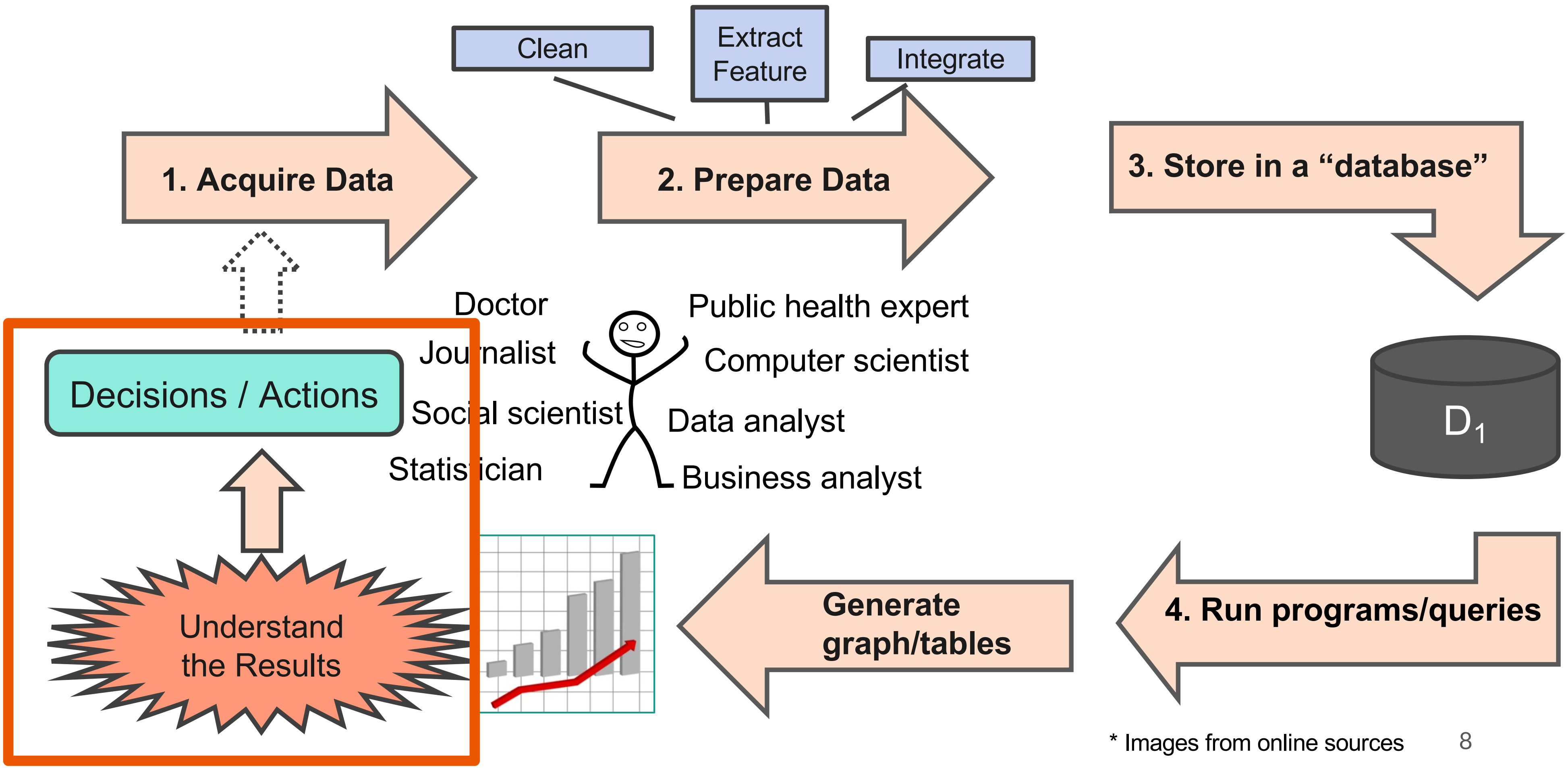


Data Analysis Loop



* Images from online sources

Data Analysis Loop

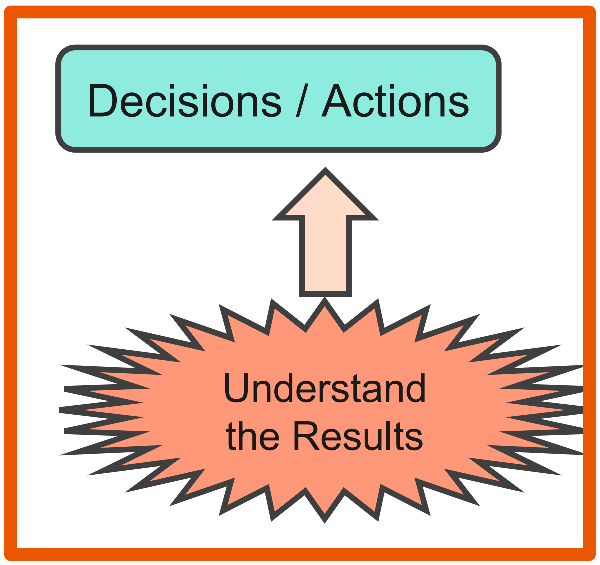


Results should be understandable

"Why do I see this output?"

"Why do I see an outlier?"

"Why is one value higher than the other?"



Results should be understandable

"Why do I see this output?"

"Why do I see an outlier?"

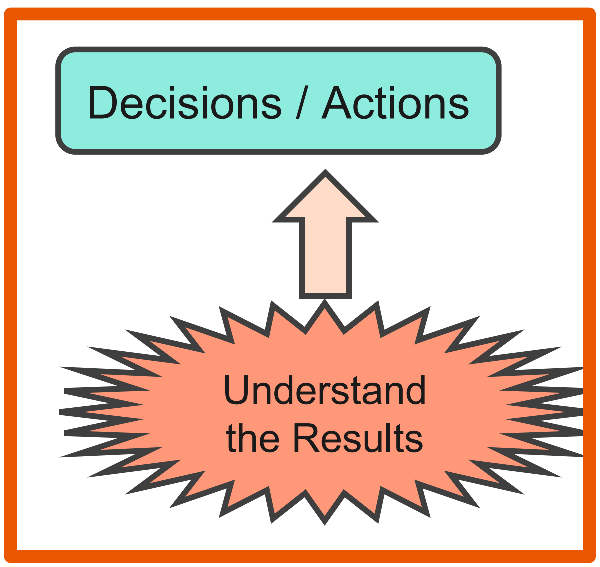
"Why is one value higher than the other?"

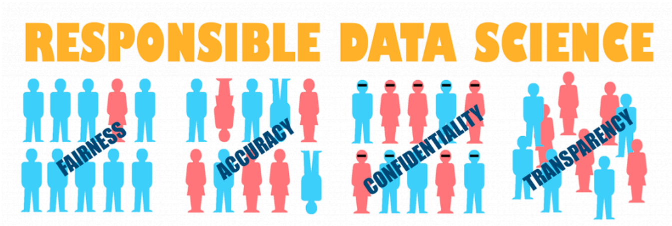
Actions should be interpretable

"How much the prestige of authors matter in the outcome of a single blind review ?"

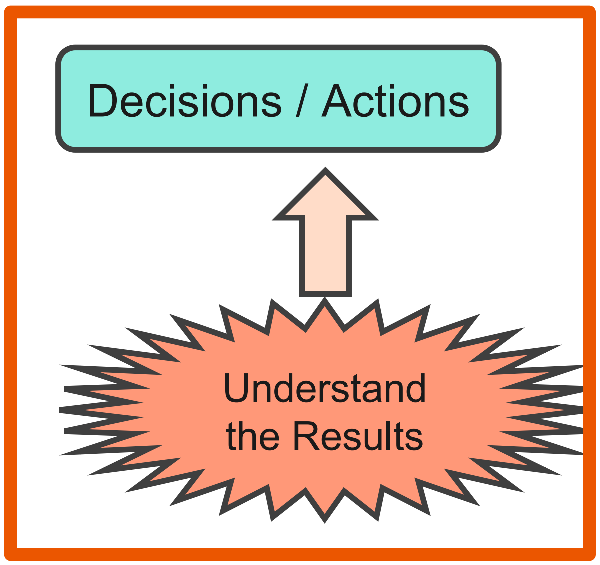
"How much drug A has an effect on disease B?"

"How much reducing housing tax encourage people to buy houses?





+
Ethics
Debugging
Accountability



Results should be understandable

"Why do I see this output?"

"Why do I see an outlier?"

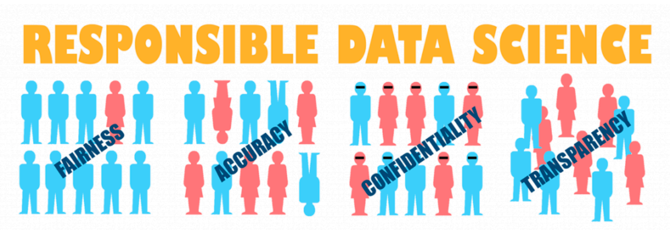
"Why is one value higher than the other?"

Actions should be interpretable

"How much the prestige of authors matter in the outcome of a single blind review ?"

"How much drug A has an effect on disease B?"

"How much reducing housing tax encourage people to buy houses?



\+
Ethics
Debugging
Accountability

Results should be understandable

"Why do I see this output?"

"Why do I see an outlier?"

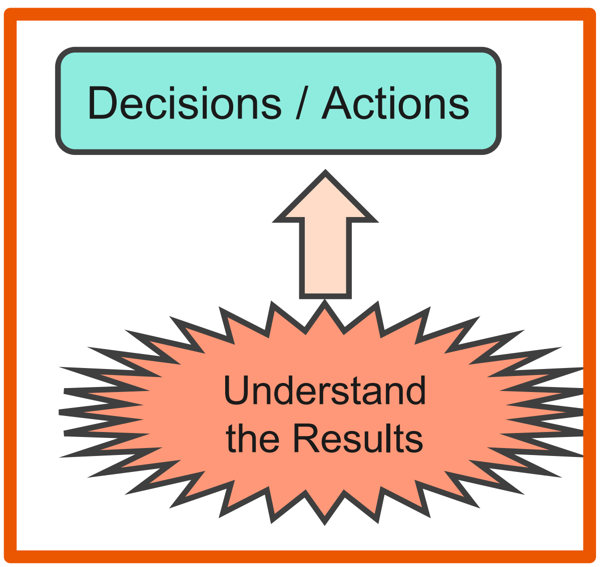
"Why is one value higher than the other?"

Actions should be interpretable

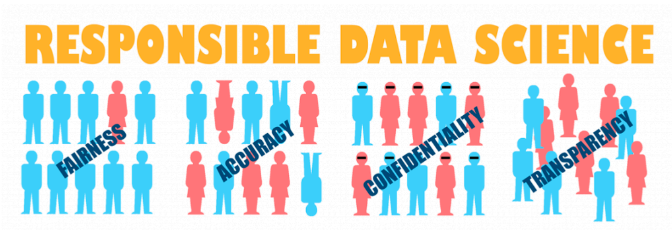
"How much the prestige of authors matter in the outcome of a single blind review ?"

"How much drug A has an effect on disease B?"

"How much reducing housing tax encourage people to buy houses?"



Causality



\+
Ethics
Debugging
Accountability

Results should be understandable

"Why do I see this output?"

"Why do I see an outlier?"

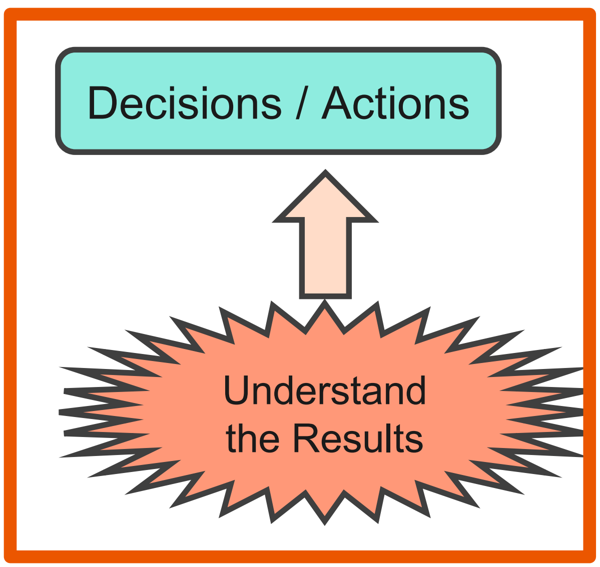
"Why is one value higher than the other?"

Actions should be interpretable

"How much the prestige of authors matter in the outcome of a single blind review ?"

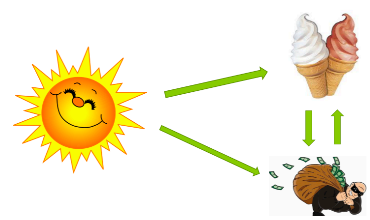
"How much drug A has an effect on disease B?"

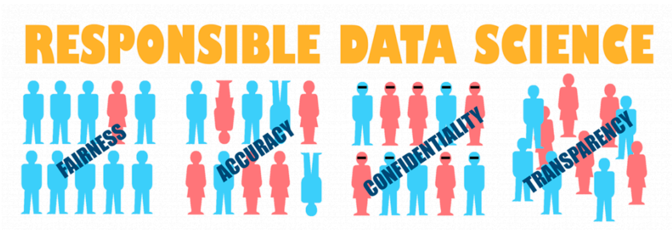
"How much reducing housing tax encourage people to buy houses?



Causality

"Correlation is not causation!"





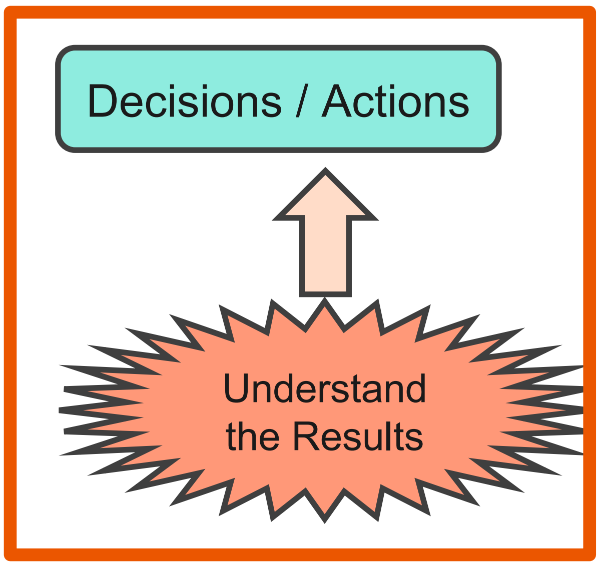
+
Ethics
Debugging
Accountability

Results should be understandable

Explanations

"Why do I see this output?"
"Why do I see an outlier?"
"Why is one value higher than the other?"

Actions should be interpretable



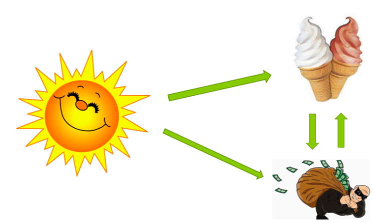
"How much the prestige of authors matter in the outcome of a single blind review ?"

"How much drug A has an effect on disease B?"

"How much reducing housing tax encourage people to buy houses?"

Causality

"Correlation is not causation!"



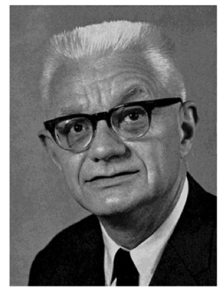
Causal Analysis on “Observational Data”

Causal Analysis

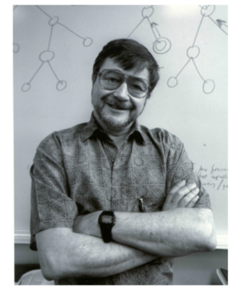
Aristotle
(384-322 BC)
Metaphysics

David Hume
(1738)
A Treatise of Human Nature

Karl Pearson
(1911)
The Grammar of Science



Carl Gustav Hempel
(1965)
Aspects of Scientific Explanation and Other Essays



Judea Pearl
Graphical Causal Models

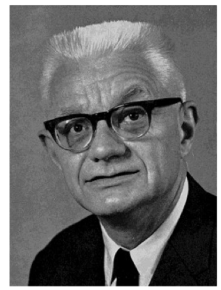
Donald Rubin
Potential Outcome Framework

Causal Analysis

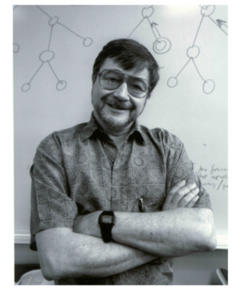
Aristotle
(384-322 BC)
Metaphysics

David Hume
(1738)
A Treatise of Human Nature

Karl Pearson
(1911)
The Grammar of Science



Carl Gustav Hempel
(1965)
Aspects of Scientific Explanation and Other Essays

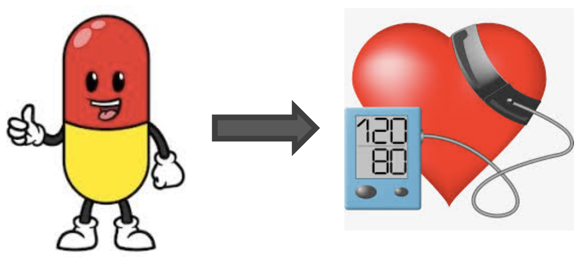


Judea Pearl
Graphical Causal Models

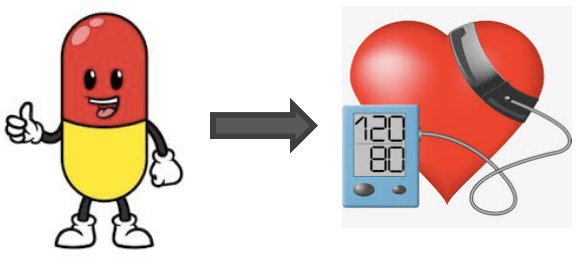
Donald Rubin
Potential Outcome Framework

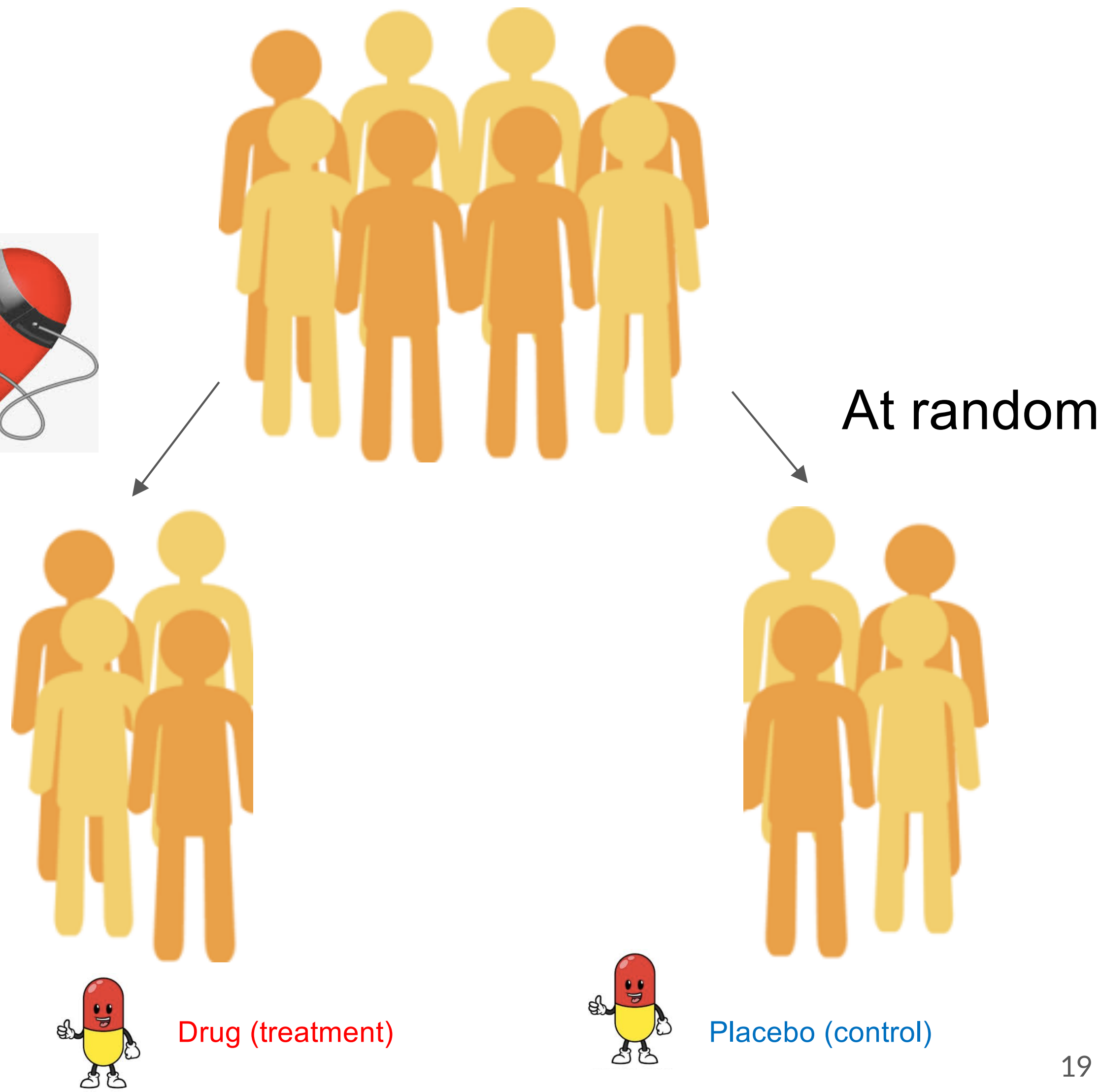
Gold standard: A randomized controlled experiment!
(e.g. Clinical Trials)

Controlled Experiments

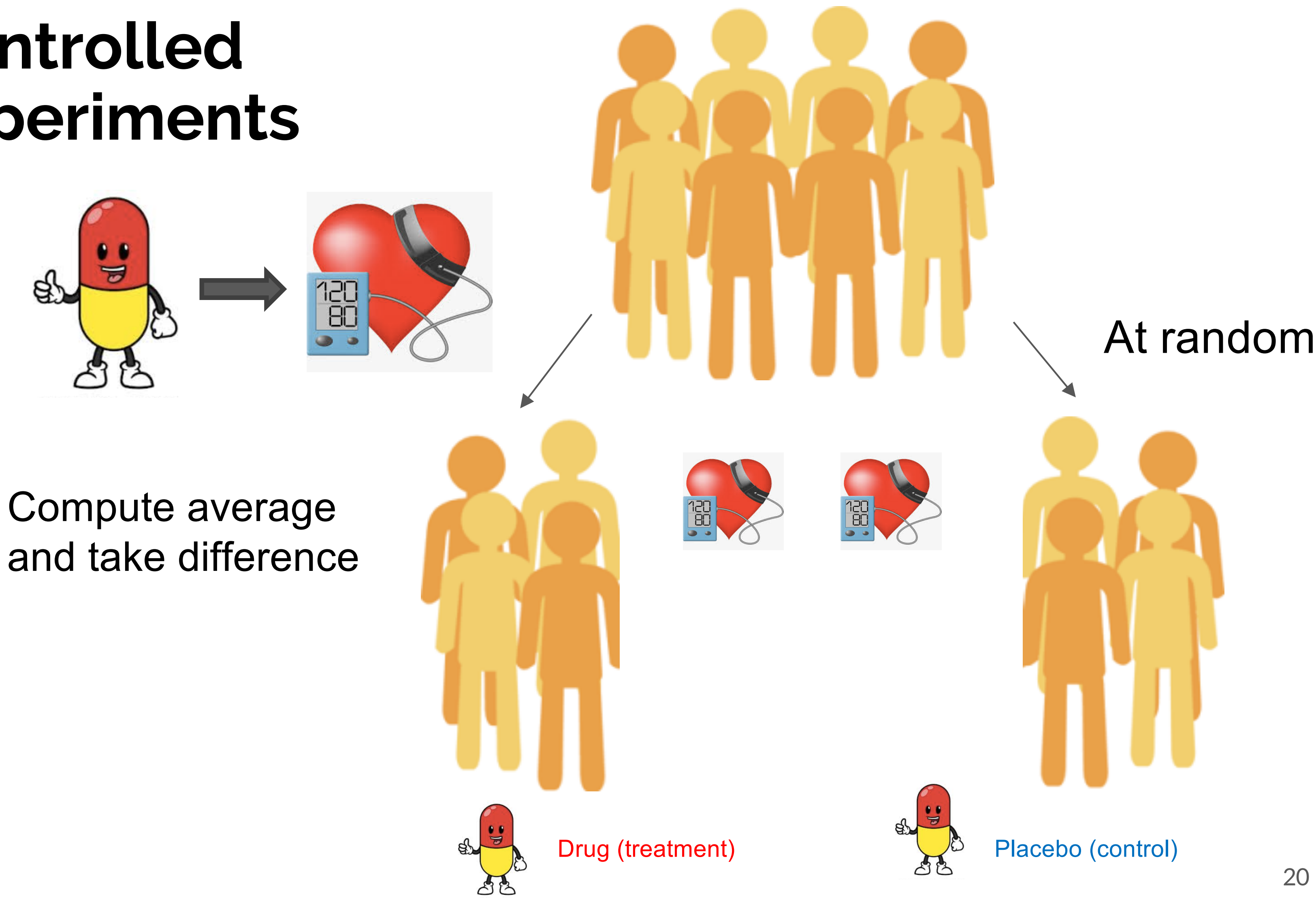


Controlled Experiments

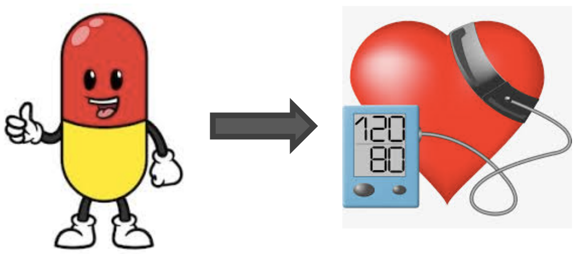


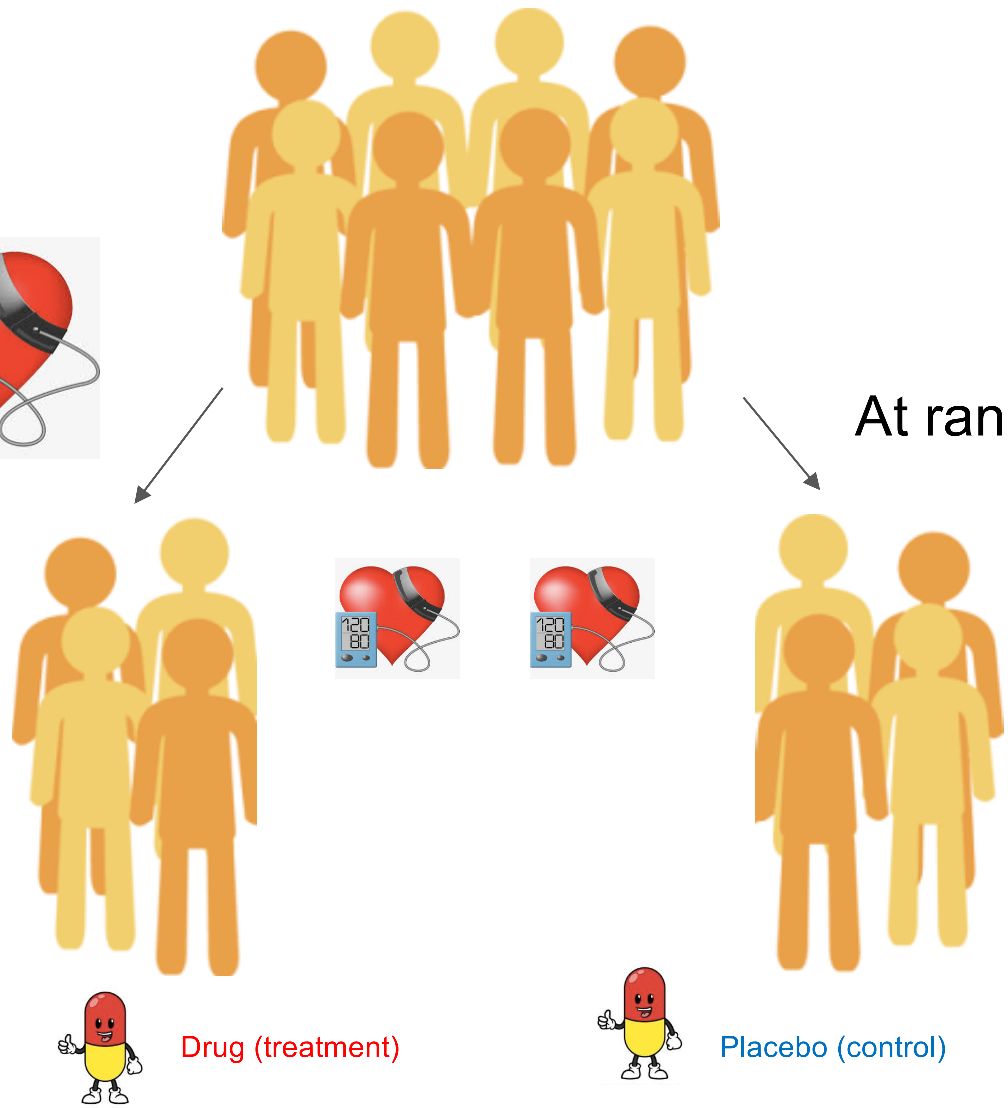


Controlled Experiments



Controlled Experiments





Randomization is crucial to estimate causal effect without bias

What if we cannot do randomized controlled experiments?

Due to ethical, time, or cost constraints

- *"Does smoking cause lung cancer?"*
- *"Does growing up in a poor neighborhood make a child earn less as an adult?"*
- *"Does smoking during pregnancy affect newborn's health?"*

What if we cannot do randomized controlled experiments?

Due to ethical, time, or cost constraints

- *“Does smoking cause lung cancer?”*
- *“Does growing up in a poor neighborhood make a child earn less as an adult?”*
- *“Does smoking during pregnancy affect newborn’s health?”*

Fortunately, we can do

“Observational Causal Studies”

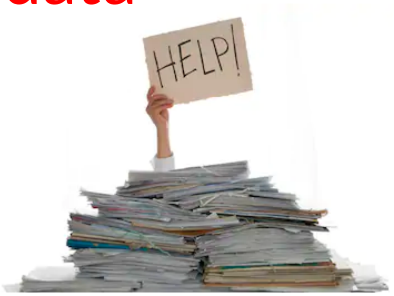
Under certain assumptions

Our work: Observational causal studies for "Big Data"

Existing causal studies work for small, simple data

Our work: Observational causal studies for "Big Data"

Existing causal studies work for small, simple data



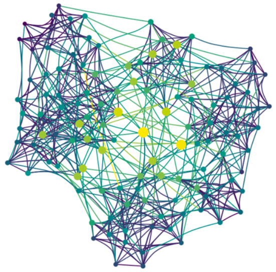
Large scale data:

- Large number of "units" (n)
- Large number of "features/covariates" (p)

Our work: Observational causal studies for "Big Data"

Existing causal studies work for small, simple data

Large scale data:
- Large number of "units" (n)
- Large number of "features/covariates" (p)



Complex data:
- Network effect on homogenous units
- Relational effect on heterogenous units

Observational Causal Study setup

Rubin'74
Rosenbaum-Rubin'83

X,	Y,	T
n x p	n x 1	n x 1 {0,1}

Y = Stroke

T = Drug S for migraine

Average Treatment Effect ATE = E[Y(1) – Y(0)]

Assumptions for observational studies:

1. SUTVA: Stable Unit Treatment Value Assumption
   T_1 does not affect Y_2
   Single treatment

2. Strong Ignorability: $Y(0), Y(1) \perp T \mid X$

“Matching” in Observational Data

Ideally…

(1) Find “units” (e.g. patients) with same/similar “confounding covariates”

- e.g., of same age, gender, height, ethnicity, …

(2) Make sure all groups have both treated and control units

(3) Estimate the causal effect within each group and take average

Rosenbaum-Rubin'83

Exact Matching = Interpretability

There are other methods like "Propensity Score Matching"

- "Match" on $e(X) = Pr(T = 1 | X)$: need a model, hard to interpret

Go model free - Exact matching to the rescue!

- Highlights overlap between treatment and control populations
- Helps us to find uncertainty and determine what type of additional data must be collected
- Interpret causal estimates within matched populations as "conditional average treatment effects (CATE)" in addition to ATE

Rosenbaum-Rubin’83

Exact Matching: Good but challenging

“As a method of multivariate adjustment, subclassification has the advantage that it involves direct comparisons of ostensibly comparable groups of units within each subclass and therefore can be both understandable and persuasive to an audience with limited statistical training... ”

Rosenbaum-Rubin’83

Exact Matching: Good but challenging

“As a method of multivariate adjustment, subclassification has the advantage that it involves direct comparisons of ostensibly comparable groups of units within each subclass and therefore can be both understandable and persuasive to an audience with limited statistical training... ”

- Subclassification = exact matching
- Direct comparisons = individualized effects
- Persuasive = intuitive, uncomplicated, reproducible

Rosenbaum-Rubin'83

Exact Matching: Good but challenging

"As a method of multivariate adjustment, subclassification has the advantage that it involves direct comparisons of ostensibly comparable groups of units within each subclass and therefore can be both understandable and persuasive to an audience with limited statistical training... "

- Subclassification = exact matching
- Direct comparisons = individualized effects
- Persuasive = intuitive, uncomplicated, reproducible

"A major problem with subclassification .. is that as the number of confounding variables increases, the number of sublcasses grows dramatically, so that even with only two categories per variable, yielding 2^P classes for P variables, most subclasses will not contain both treated and control units."

Rosenbaum-Rubin’83

Exact Matching: Good but challenging

“As a method of multivariate adjustment, subclassification has the advantage that it involves direct comparisons of ostensibly comparable groups of units within each subclass and therefore can be both understandable and persuasive to an audience with limited statistical training... ”

- Subclassification = exact matching
- Direct comparisons = individualized effects
- Persuasive = intuitive, uncomplicated, reproducible

“A major problem with subclassification .. is that as the number of confounding variables increases, the number of sublcasses grows dramatically, so that even with only two categories per variable, yielding 2^P classes for P variables, most subclasses will not contain both treated and control units.”

- Confounders = variables of potential interest
- Number of subclasses = types of individualized effects
- Empty subclasses = impossible to draw causal conclusions

Wang-Morucci-Awan-Liu-Roy-Rudin-Volfovsky'19

FLAME: Fast Large Almost Matching Exactly

	Important Covariates							Unimportant Covariates						
covariates:	age, gender, heart conditions, blood pressure,							toenail length, eyeball width, etc.						
treated patient Marietta	[50	F	1	0	1	1	68	1.5cm	2cm	1	0	3	0	.....]
control patient Lee Ann	[50	F	1	0	1	1	68	14cm	1cm	4	1	5	6	.....]

Wang-Morucci-Awan-Liu-Roy-Rudin-Volfovsky’19

FLAME: Fast Large Almost Matching Exactly

	Important Covariates	Unimportant Covariates
covariates:	age, gender, heart conditions, blood pressure,	toenail length, eyeball width, etc.

treated patient
Marietta [50 F 1 0 1 1 68 1.5cm 2cm 1 0 3 0 .....]

control patient
Lee Ann [50 F 1 0 1 1 68 14cm 1cm 4 1 5 6 .....]

- Match treatment and control units using as many *important* covariates as possible
- Handle large datasets

Wang-Morucci-Awan-Liu-Roy-Rudin-Volfovsky'19

FLAME: Fast Large Almost Matching Exactly

	Important Covariates	Unimportant Covariates
covariates:	age, gender, heart conditions, blood pressure,	toenail length, eyeball width, etc.

treated patient
Marietta [50 F 1 0 1 1 68 1.5cm 2cm 1 0 3 0 .....]

control patient
Lee Ann [50 F 1 0 1 1 68 14cm 1cm 4 1 5 6 .....]

- Match treatment and control units using as many *important* covariates as possible
- Handle large datasets

From learning

Wang-Morucci-Awan-Liu-Roy-Rudin-Volfovsky'19

FLAME: Fast Large Almost Matching Exactly

Important Covariates **Unimportant Covariates**

covariates: age, gender, heart conditions, blood pressure, toenail length, eyeball width, etc.

treated patient
Marietta [50 F 1 0 1 1 68 1.5cm 2cm 1 0 3 0 .....]

control patient
Lee Ann [50 F 1 0 1 1 68 14cm 1cm 4 1 5 6 .....]

- Match treatment and control units using as many *important* covariates as possible
- Handle large datasets

From learning

Using techniques from data management

Optimization Problem for FLAME

Variable Selector Indicator: $\boldsymbol{\theta} \in \{0,1\}^{\text{p}}$

Matched Group for i on variables :: $\boldsymbol{\theta}$

$$\mathcal{MG}_i(\boldsymbol{\theta}, \mathcal{S}) = \{i' \in \mathcal{S} : \mathbf{x}_{i'} \circ \boldsymbol{\theta} = \mathbf{x}_i \circ \boldsymbol{\theta}\}$$

Prediction Error on training set

$$\hat{\text{PE}}_{\mathcal{F}_{\|\boldsymbol{\theta}\|_0}}(\boldsymbol{\theta}, \mathcal{S}) = \min_{f^{(1)} \in \mathcal{F}_{\|\boldsymbol{\theta}\|_0}} \frac{1}{|\mathcal{S}_1|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_1} (f^{(1)}(\mathbf{x}_i \circ \boldsymbol{\theta}) - y_i)^2 + \min_{f^{(0)} \in \mathcal{F}_{\|\boldsymbol{\theta}\|_0}} \frac{1}{|\mathcal{S}_0|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_0} (f^{(0)}(\mathbf{x}_i \circ \boldsymbol{\theta}) - y_i)^2.$$

Objective:

$$\boldsymbol{\theta}^*_{i,\mathcal{S}} \in \arg\min_{\boldsymbol{\theta}} \hat{\text{PE}}_{\mathcal{F}_{\|\boldsymbol{\theta}\|_0}}(\boldsymbol{\theta}, \mathcal{S}) \text{ s.t. } \exists \ell \in \mathcal{MG}_i(\boldsymbol{\theta}, \mathcal{S}) \text{ s.t. } t_\ell = 0$$

Optimization Problem for FLAME

Variable Selector Indicator: $\boldsymbol{\theta} \in \{0, 1\}^{p}$

Matched Group for i on variables :: $\boldsymbol{\theta}$

$$\mathcal{MG}_i(\boldsymbol{\theta}, \mathcal{S}) = \{i' \in \mathcal{S} : \mathbf{x}_{i'} \circ \boldsymbol{\theta} = \mathbf{x}_i \circ \boldsymbol{\theta}\}$$

Prediction Error on training set

$$\hat{\text{PE}}_{\mathcal{F}_{\|\boldsymbol{\theta}\|_0}}(\boldsymbol{\theta}, \mathcal{S}) = \min_{f^{(1)} \in \mathcal{F}_{\|\boldsymbol{\theta}\|_0}} \frac{1}{|\mathcal{S}_1|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_1} (f^{(1)}(\mathbf{x}_i \circ \boldsymbol{\theta}) - y_i)^2 + \min_{f^{(0)} \in \mathcal{F}_{\|\boldsymbol{\theta}\|_0}} \frac{1}{|\mathcal{S}_0|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_0} (f^{(0)}(\mathbf{x}_i \circ \boldsymbol{\theta}) - y_i)^2.$$

Objective:

$$\boldsymbol{\theta}^*_{i,\mathcal{S}} \in \arg\min_{\boldsymbol{\theta}} \hat{\text{PE}}_{\mathcal{F}_{\|\boldsymbol{\theta}\|_0}}(\boldsymbol{\theta}, \mathcal{S}) \text{ s.t. } \exists \ell \in \mathcal{MG}_i(\boldsymbol{\theta}, \mathcal{S}) \text{ s.t. } t_\ell = 0$$

For every treatment unit, find
The best possible match with at
least one control unit

Best = Low predictive error
on a holdout set

Drop least useful covariate
and continue

Efficient exact matching with database queries

```
SELECT Age, Race, Gender, State, Education,
    ((SUM(T*Y)/SUM(T)) – (SUM(1-T)*Y)/(COUNT(*)-SUM(T))) AS ATE
FROM  Population
GROUP BY Age, Race, Gender, State, Education
HAVING SUM(T)>= 1 AND SUM(T) <= COUNT(*) - 1
```

SQL “Group-by” queries:
Finds all groups of units with the same values of covariates
very efficiently

Some (insightful) experiments

$$y = \sum_{i=1}^{10} \alpha_i x_i + T \sum_{i=1}^{10} \beta_i x_i + T \cdot U \sum_{i=1\ldots5, \gamma=1..5, \gamma>i} x_i x_\gamma,$$

+20 irrelevant covariates, where $\alpha_i = \beta_i = 0$

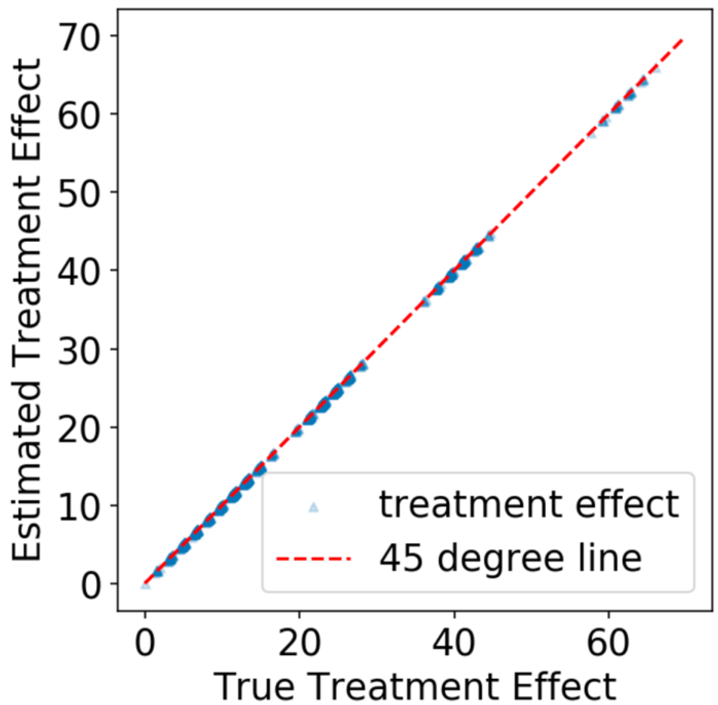
$$x_i \sim \text{Bernoulli}(0.5) \text{ for } 1 \leq i \leq 10$$

$$10 < i \leq 30, \quad x_i \sim \text{Bernoulli}(0.1) \text{ in the control group}$$

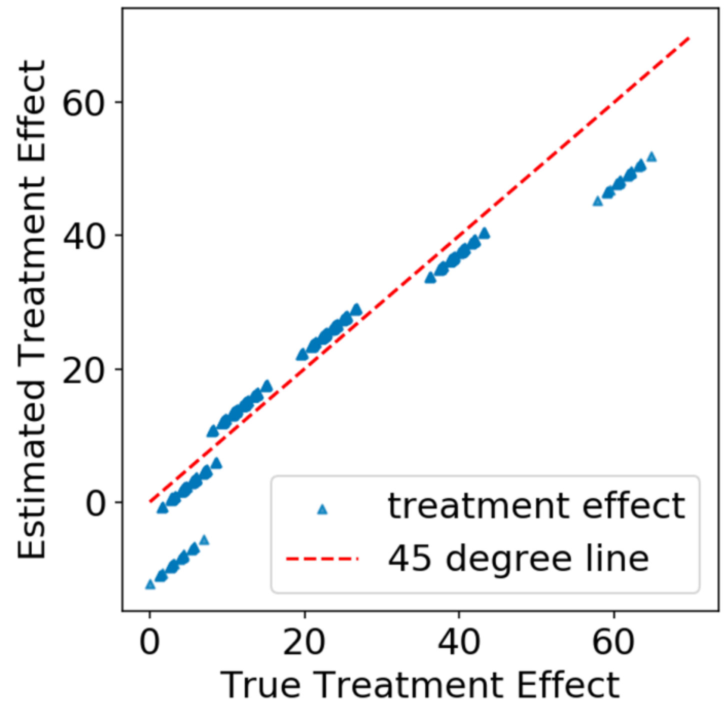
$$x_i \sim \text{Bernoulli}(0.9) \text{ in the treatment group.}$$

20K units, 10K treatment, 10K control

(no noise)



(a) FLAME

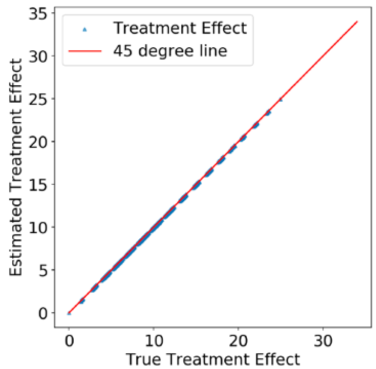


(b) Double linear regressors

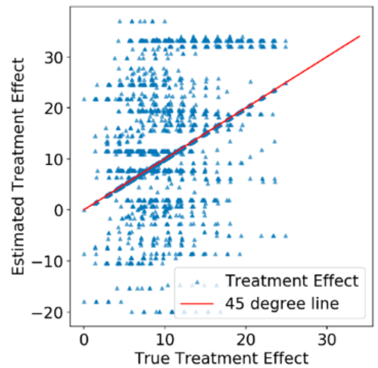
Regression cannot handle model misspecification

Accuracy: FLAME beats all other methods

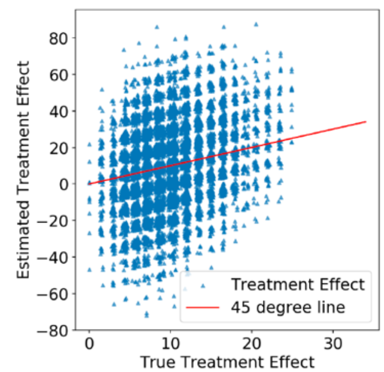
FLAME has less error



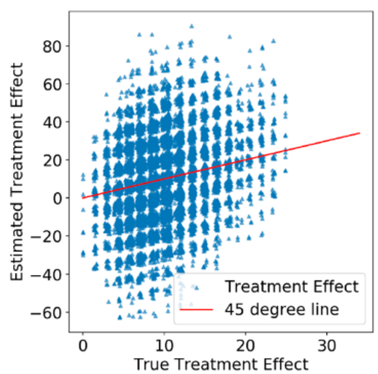
(a) FLAME (Early Stopping)



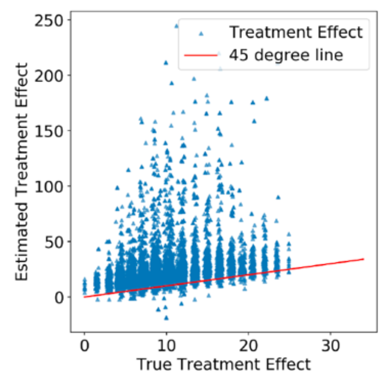
(b) FLAME (Run Until No More Matches)



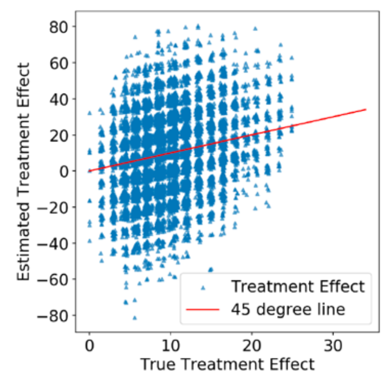
(c) 1-PSNNM



(d) GenMatch



(e) Causal Forest



(f) Mahalanobis

Time: FLAME beats all other methods on large data!

Small (er) data 30k units

Method	Time (seconds)
FLAME-bit	27.68 ± 0.80
FLAME-db	57.93 ± 0.47
Causal Forest	52.34 ± 1.82
1-PSNNM	14.78 ± 0.70
Mahalanobis	76.79 ± 0.49
GenMatch	> 150
Cardinality Match	> 150

On the census dataset with
~ 1 million tuples and ~60 covariates

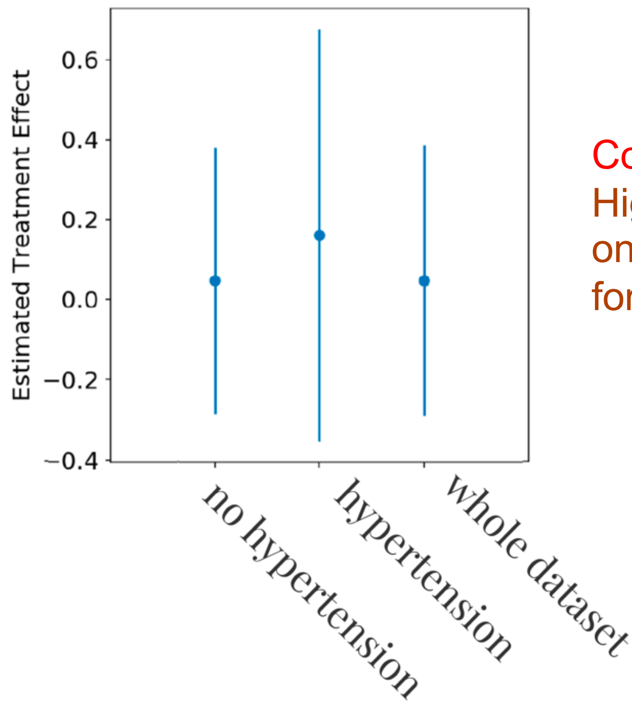
Method	Time (hours)
FLAME-bit	Crashed
FLAME-db	1.37
Causal Forest	Crashed
1-PSNNM	> 10
Mahalanobis	> 10
GenMatch	> 10
Cardinality Match	> 10

FLAME is scalable

Application: Natality data

- publicly available dataset on 2010 Natality dataset
- 86 variables includes health information of pregnant women and newborns
- causal effect of smoking on risk of child abnormal health conditions
- 204,886 treated units, 1,985,524 control. 10% used as holdout

Public data from CDC
~4 million tuples



Conditional Average Treatment Effect (CATE)
Higher causal effect
on smoking during pregnancy
for mothers with hypertension

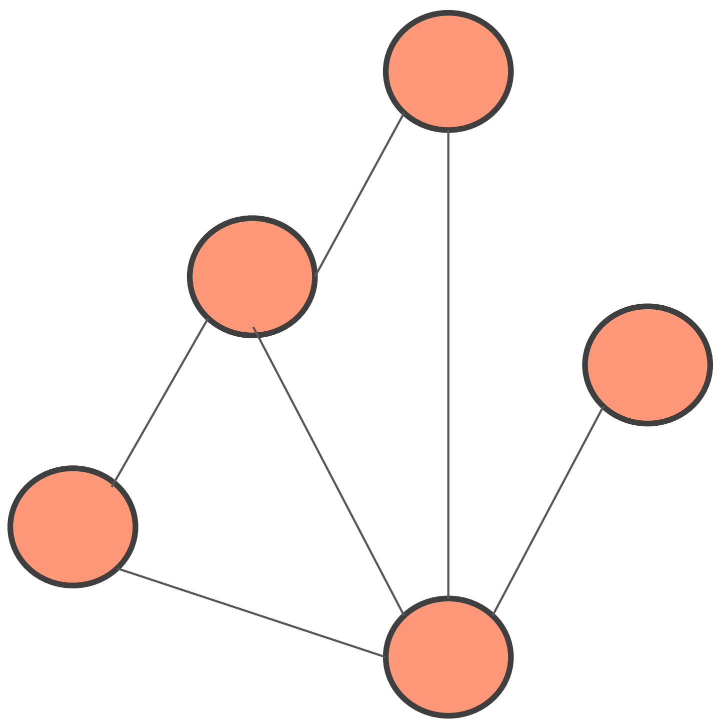
Dieng-Liu-Roy-Rudin-Volfovsky – AISTATS'19
Awan-Liu-Morucci-Roy-Rudin-Volfovsky – UAI'19

Extensions of FLAME

- FLAME is greedy, DAME (Dynamic Almost Exact Matching) finds optimal solution by an exhaustive search – but efficiently, by ideas from data mining
    - Worse running time than FLAME, but better quality matches
- Extension to instrumental variables
- Takeaway: FLAME and DAME leverage ideas from ML + databases
    - Scalable
    - Accurate
- Ongoing: continuous covariates, time series data, …

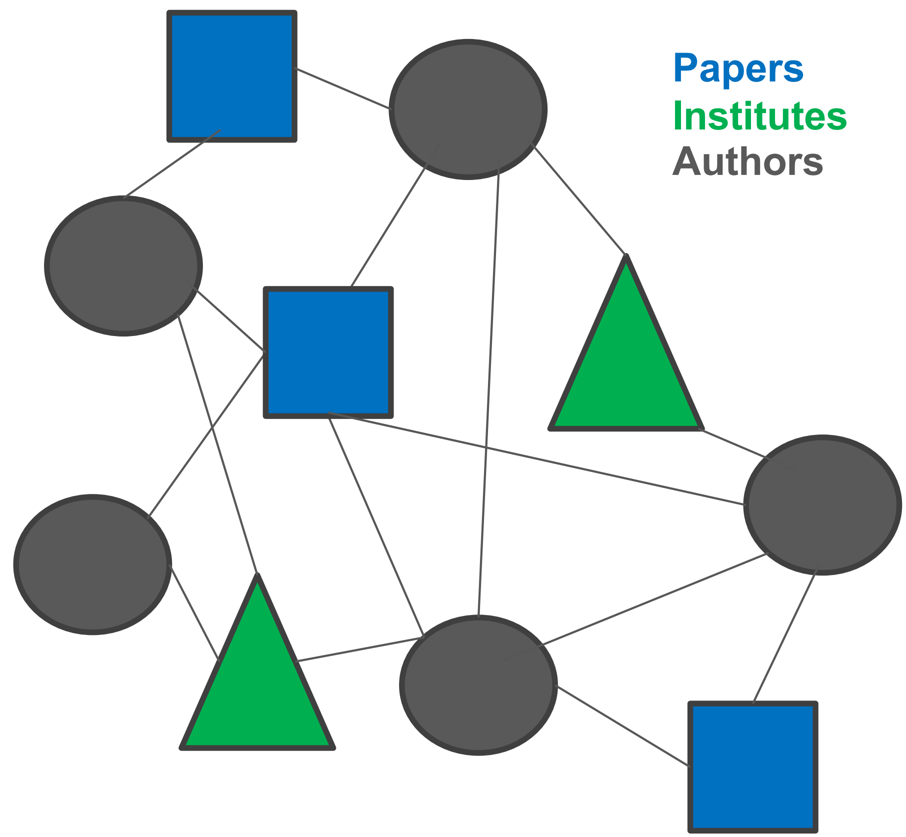
All these on a single “table” with “Independent Units”

Complex Data



Student sharing rooms in college dorms

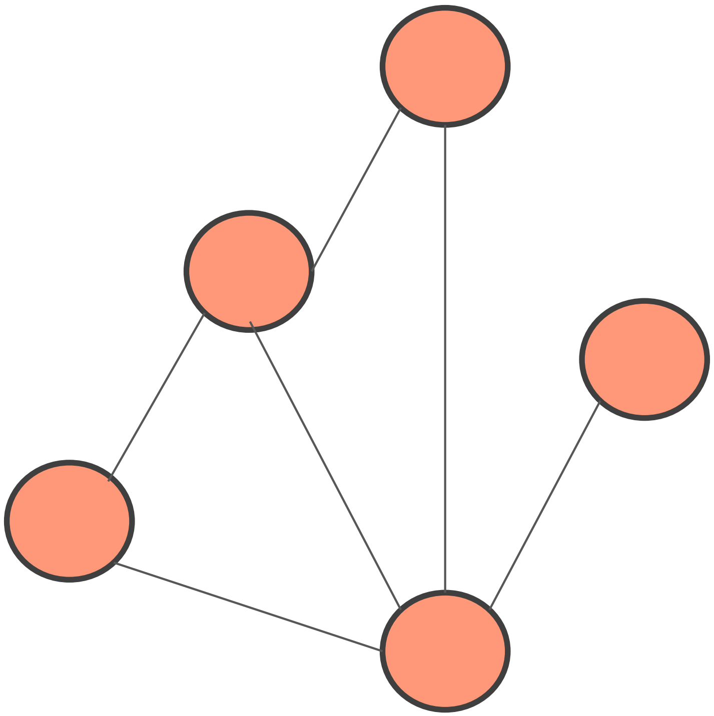
"homogenous units"



"heterogenous units"

Morucci-Awan-Orlandi-Roy-Rudin-Volfovsky'19

Homogenous units on a network



Student sharing rooms in college dorms

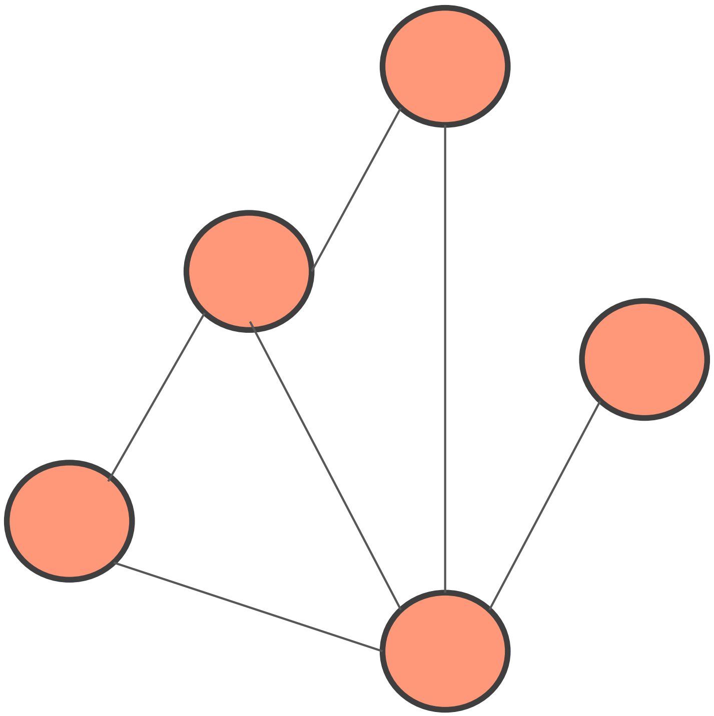
"homogenous units"

Basic assumptions like SUTVA do not hold

For two neighbors 1 and 2:
Interference T1 affects Y2
Contagion Y1 affects Y2
Entanglement T1 = T2

Morucci-Awan-Orlandi-Roy-Rudin-Volfovsky’19

Homogenous units on a network



Basic assumptions like SUTVA do not hold

For two neighbors 1 and 2:
Interference T1 affects Y2
Contagion Y1 affects Y2
Entanglement T1 = T2

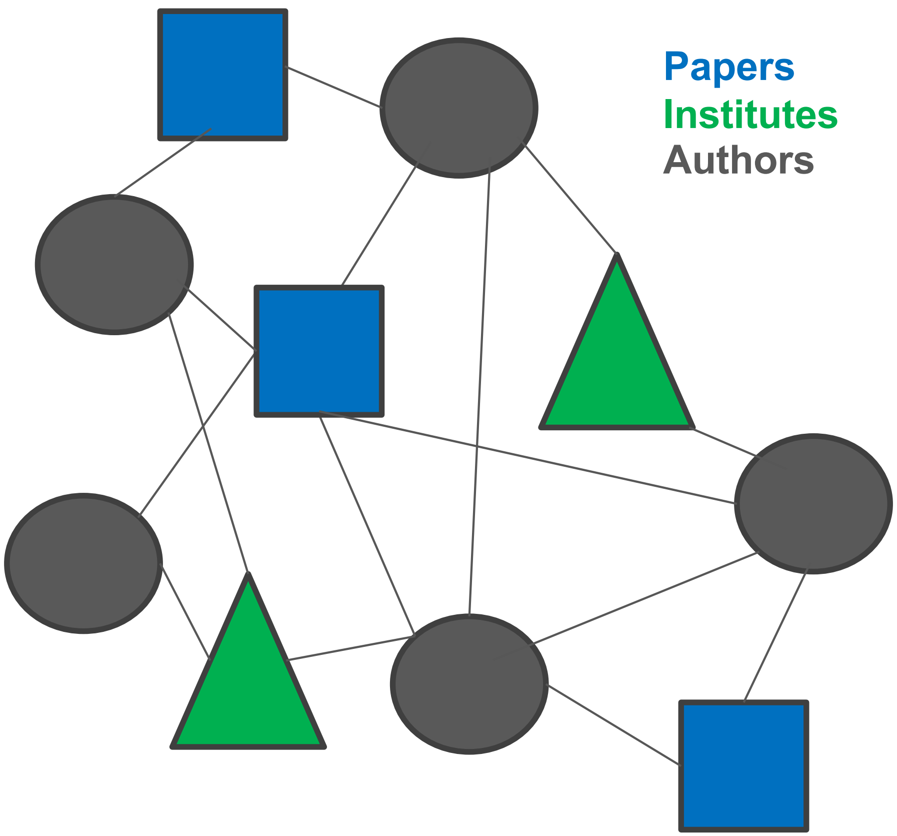
Our (initial) work:
- Matching on neighborhood structure on experimental data
- Match on all possible subgraphs, use FLAME

Student sharing rooms in college dorms

“homogenous units”

Salimi-Kayali-Parikh-Getoor-Roy-Suciu’19

Heterogenous relational data

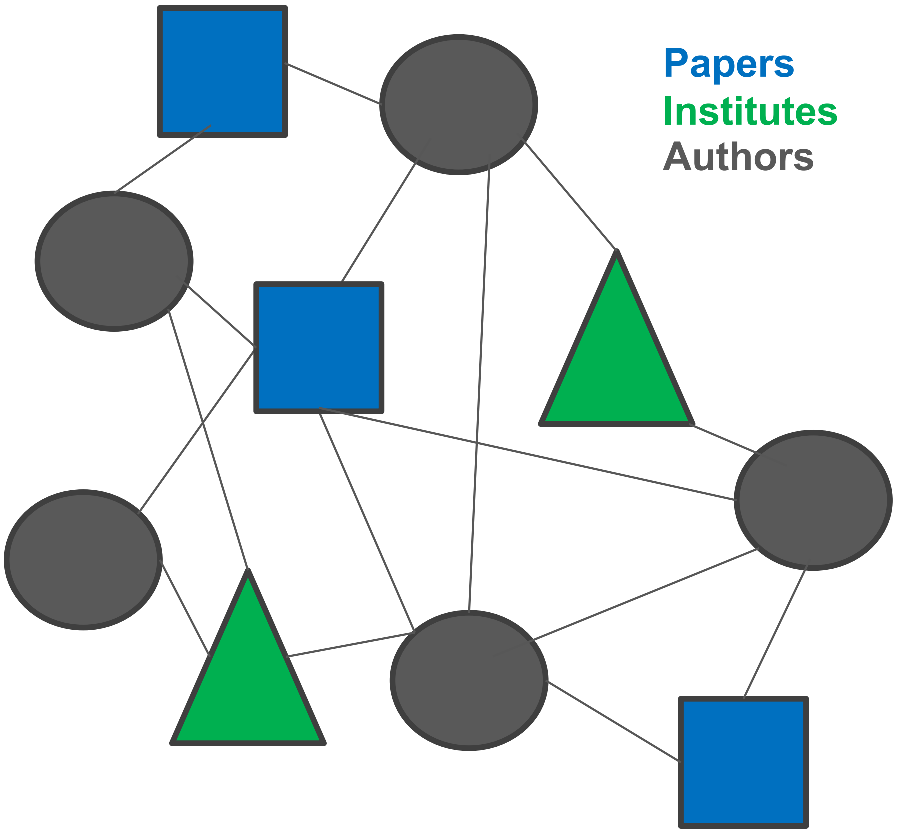


Multiple tables:
Papers(pid, venue, year, title, …)
Institute(iid, city, country, rank)
Authors(aid, name, position)
Affiliation(aid, iid)
Wrote(aid, pid)
Review(pid, rid, is-single-blind, **score**)

“heterogenous units”

Salimi-Kayali-Parikh-Getoor-Roy-Suciu'19

Heterogenous relational data



Papers
Institutes
Authors

Multiple tables:
Papers(pid, venue, year, title, …)
Institute(iid, city, country, rank)
Authors(aid, name, position)
Affiliation(aid, iid)
Wrote(aid, pid)
Review(pid, rid, is-single-blind, **score**)

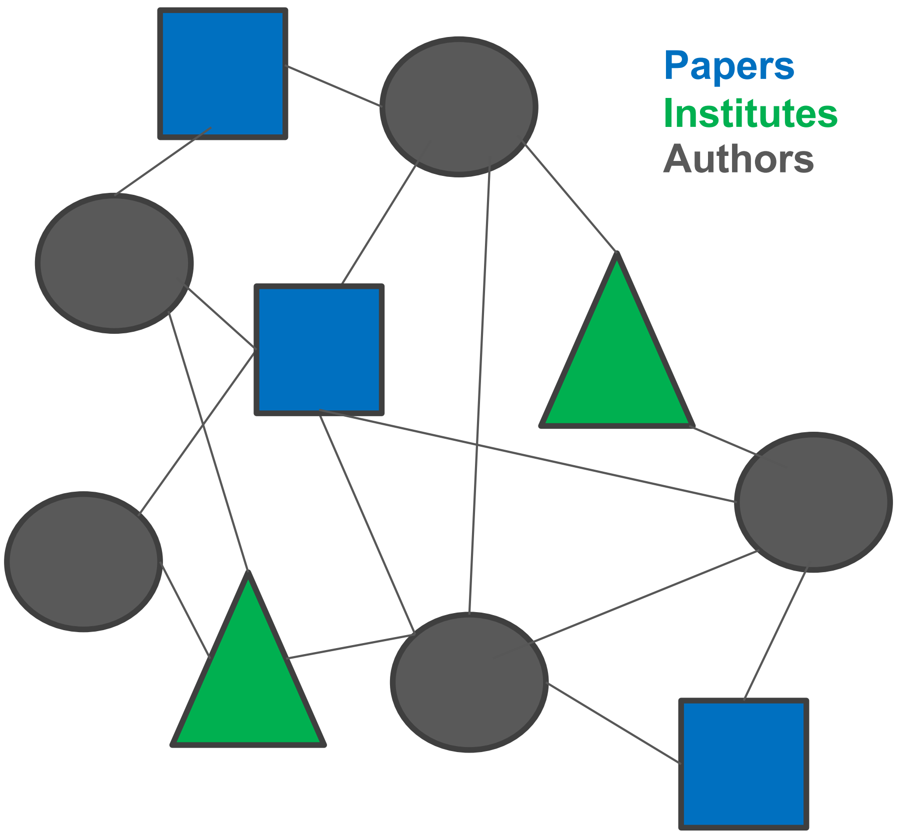
Does institutional rank (prestige) causally affect Scores received by papers in reviews?

- For single-blind reviews?
- For double-blind reviews?

"heterogenous units"

Salimi-Kayali-Parikh-Getoor-Roy-Suciu'19

Heterogenous relational data



Papers
Institutes
Authors

From two tables

T

Y

Multiple tables:
Papers(pid, venue, year, title, ...)
Institute(iid, city, country, rank)
Authors(aid, name, position)
Affiliation(aid, iid)
Wrote(aid, pid)
Review(pid, rid, is-single-blind, **score**)

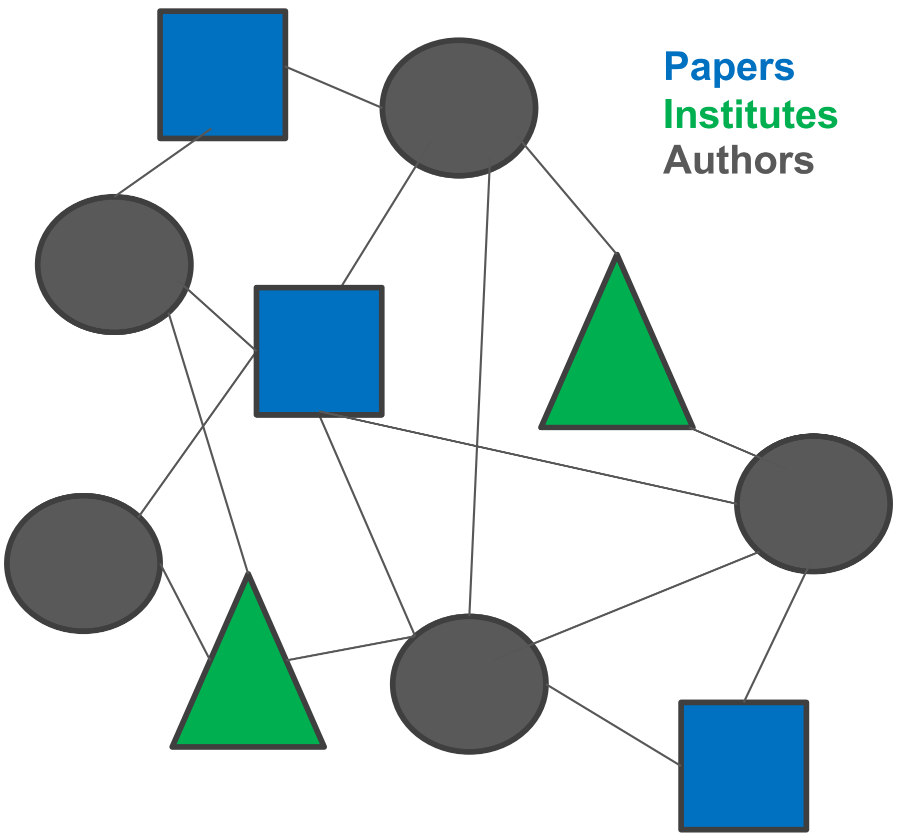
Does institutional rank (prestige) causally affect Scores received by papers in reviews?

- For single-blind reviews?
- For double-blind reviews?

"heterogenous units"

Salimi-Kayali-Parikh-Getoor-Roy-Suciu’19

Heterogenous relational data



“heterogenous units”

From two tables

T

Y

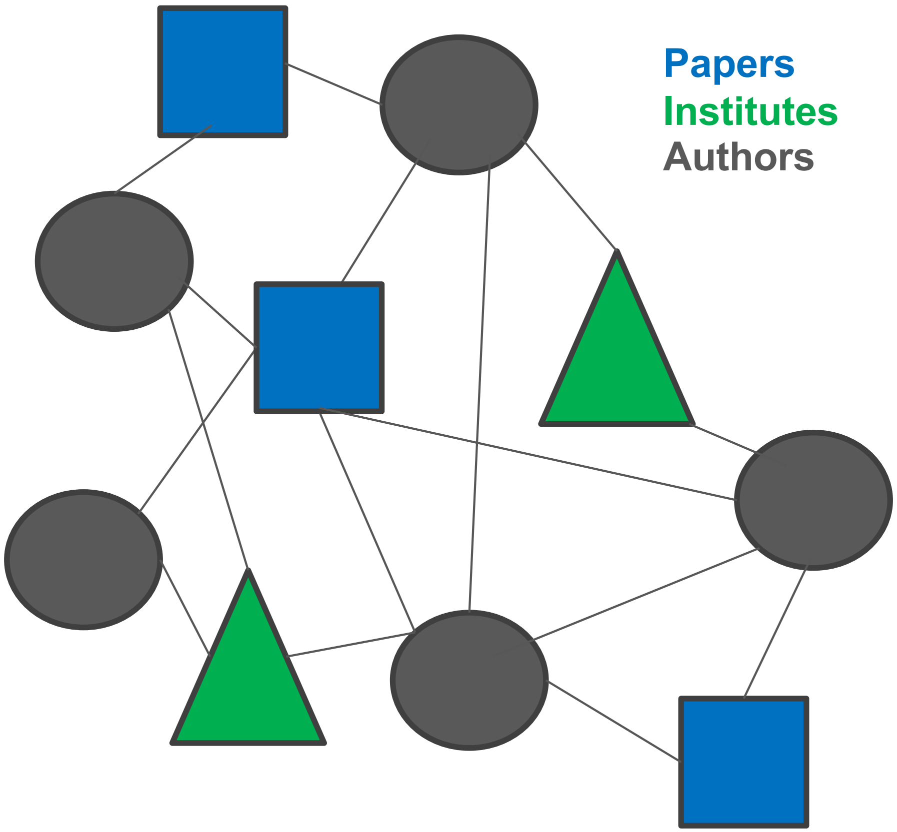
Multiple tables:
Papers(pid, venue, year, title, …)
Institute(iid, city, country, rank)
Authors(aid, name, position)
Affiliation(aid, iid)
Wrote(aid, pid)
Review(pid, rid, is-single-blind, **score**)

Does institutional rank (prestige) causally affect Scores received by papers in reviews?

- For single-blind reviews?
- For double-blind reviews?

Doctors – Patients – Disease - Treatment - Cost ..

Heterogenous relational data

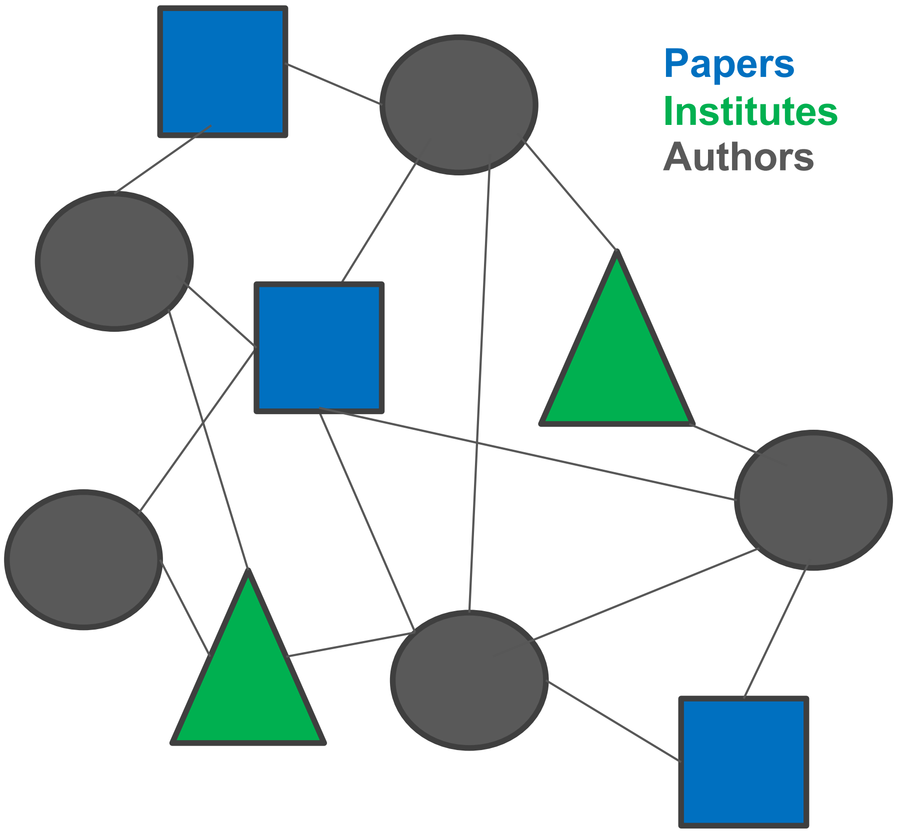


“heterogenous units”

Multiple tables:
Papers(pid, venue, year, title, …)
Institute(iid, city, country, rank)
Authors(aid, name, position)
Affiliation(aid, iid)
Wrote(aid, pid)
Review(pid, rid, is-single-blind, score)

- Need to find the right set of “unified” units
  By multiple levels of “mapping”
- Need to find the right set of covariates
  Using “causal graphs”

Heterogenous relational data



"heterogenous units"

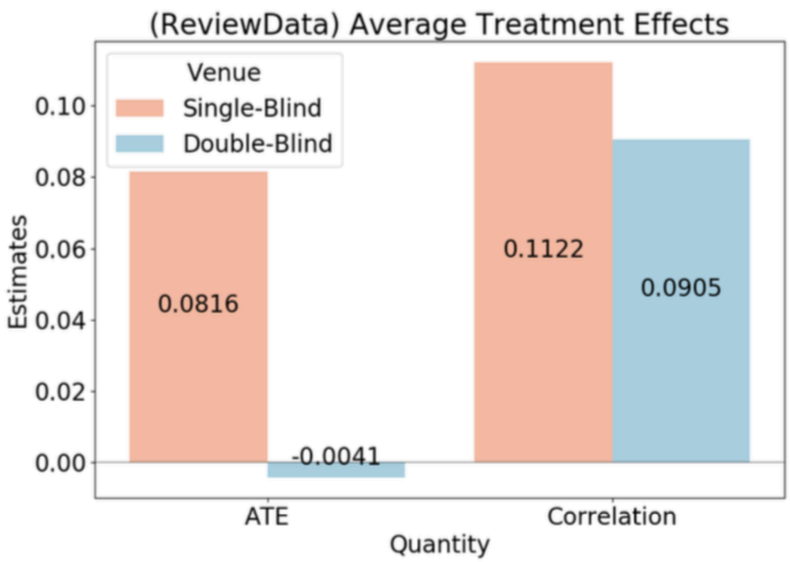
Multiple tables:
Papers(pid, venue, year, title, …)
Institute(iid, city, country, rank)
Authors(aid, name, position)
Affiliation(aid, iid)
Wrote(aid, pid)
Review(pid, rid, is-single-blind, score)

- Need to find the right set of "unified" units
  By multiple levels of "mapping"

- Need to find the right set of covariates
  Using "causal graphs"

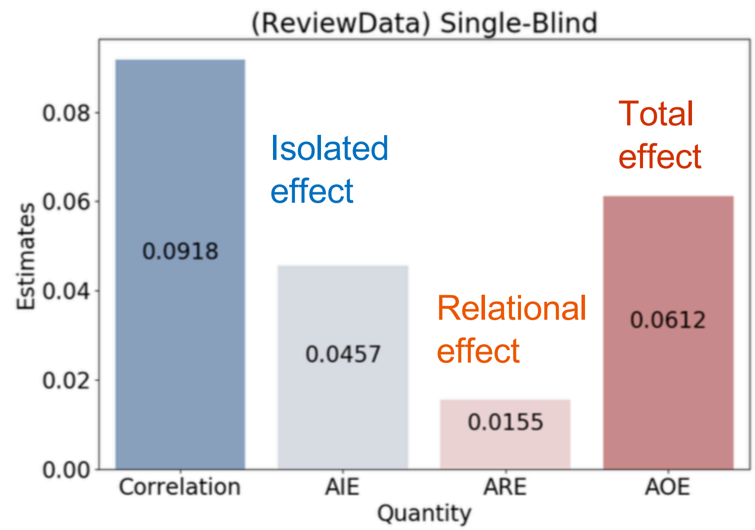
We do all these "declaratively"

Sample results



(a)

Causation vs. Correlation



(b)

Isolated, relational, and total effect

Explaining Results Motivated by Causality

Results should be understandable

Explanations

"Why do I see this output?"
"Why do I see an outlier?"
"Why is one value higher than the other?"

Y is a "cause" of Z if we can change Z by manipulating Y

Results should be understandable

Explanations

"Why do I see this output?"
"Why do I see an outlier?"
"Why is one value higher than the other?"

Y is a "cause" of Z if we can change Z by manipulating Y

A subset of input is an explanation to user's question if we can change the results by "manipulating" this subset

- and provide a compact description of the subset as the explanation (e.g., a predicate)

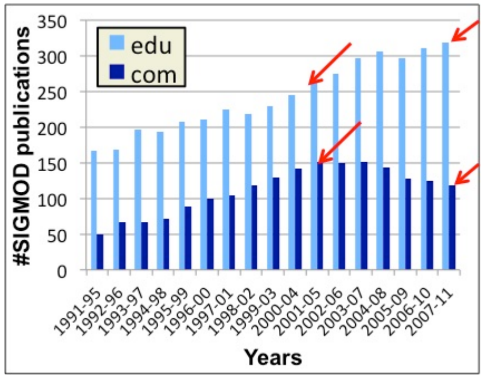
Explanations: Examples

Roy-Suciu- SIGMOD'14
Roy-Orr-Suciu – PVLDB'15
Miao-Zeng-Glavic-Roy – SIGMOD'19

Intervention — If these patterns were not there, situation would change

Counterbalance — A "low" outlier can be explained by a "high" outlier

Q. Why industry SIGMOD papers reduced compared to academia?



	Explanations
1	inst = ibm.com
2	inst = bell-labs.com
3	name = Rajeev Rastogi
4	inst = ucla.edu
5	name = Hamid Pirahesh
6	inst = asu.edu
7	name = Rakesh Agrawal

- Many papers from Bell Labs, IBM around 2000
- Either they are not active (intervention)

Or

- They shifted focus (counterbalance)

What next?

- What improvements to the research infrastructure are needed?
  - A joint research agenda in addition to helping each other's agenda
  - Platform to facilitate cross-disciplinary collaboration
  - One of the key challenges is writing our papers is finding an application and a good dataset
  - Easy access to data
  - Discussion board?
  - More frequent workshops like this

What next?

- What types of training are most important for this type of research?
  - Rigorous training in computer science, machine learning, artificial intelligence, statistics, maths, programming, algorithms, ...
  - Ability to understand problems in an application domain and communicate with domain experts
  - Back and forth contributions

    Applications ⇒ Methodology ⇒ Application ⇒ Methodology ....

    (decision making/policy?)

What next?

- What are the future research needs (methods, analyses and interventions, etc.)?
    - Model all the complexity in the data (constraints, structure, continuous/discrete features, incompleteness/uncertainty in noisy data)
    - Make data analysis interpretable … and accessible.. to a broad range of data scientists and domain experts from technical and non-technical background

Joint work with

1R01EB025021-01

CAREER
IIS-1552538
IIS-1703431


Cynthia Rudin
Duke CS


Alexander Volfovsky
Duke Statistics

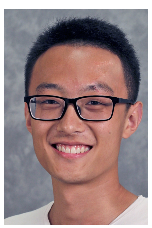

Lise Getoor
UCSC


Dan Suciu
UW


Babak Salimi
UW


Boris Glavic
IIT Chicago

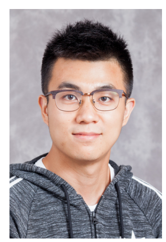

Harsh Parikh

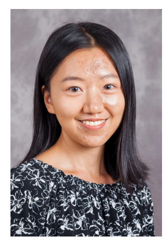

Tianyu Wang


Marco Morucci

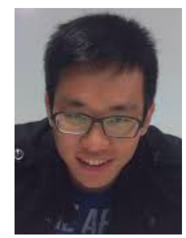

M. Usaid Awan


Vittorio Orlandi


Zhengjie Miao


Yameng Liu


Moe Kayali, UW


Qitian Zeng, IIT


Laurel Orr, UW

DUKE

And many others..

* Code available online