Microsoft Research

Faculty Summit

2014 15TH ANNUAL

Microsoft Research
Faculty Summit
2014 15TH ANNUAL

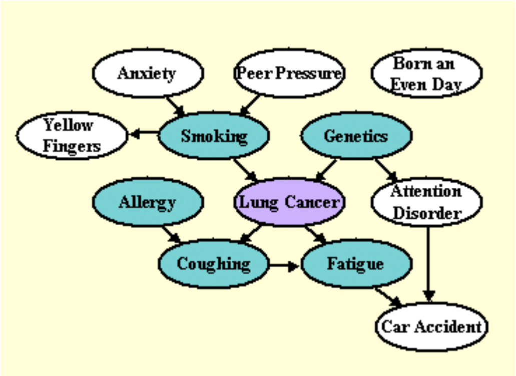
Contribution of Machine Learning Challenges to Causal Discovery

Isabelle Guyon

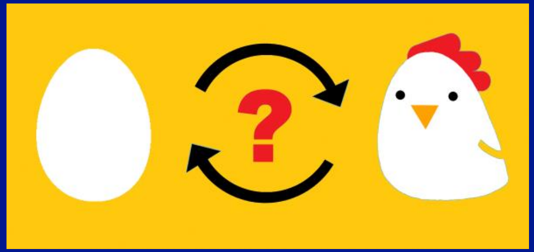
CHA LEARN

Causality challenges

Causation and Prediction (2007)

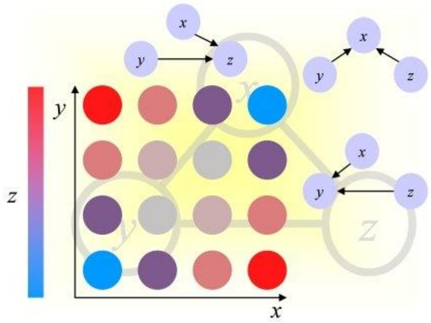


Fast causation coefficient (2014)



Neural Connectomics (2014)

Pot-luck challenge (2008)



Acknowledgements:

Initial impulse: Joris Mooij, Dominik Janzing, and Bernhard Schölkopf, from the Max Planck.

Examples of algorithms and data: Povilas Daniušis, Arthur Gretton, Patrik O. Hoyer, Dominik Janzing, Antti Kerminen, Joris Mooij, Jonas Peters, Bernhard Schölkopf, Shohei Shimizu, Oliver Stegle, and Kun Zhang, Jakob Zscheischler.

Datasets and result analysis: Isabelle Guyon + Mehreen Saeed + {Mikael Henaff, Sisi Ma, and Alexander Statnikov}, from NYU.

Website and sample code: Isabelle Guyon +

Phase 1: Ben Hamner (Kaggle) https://www.kaggle.com/c/cause-effect-pairs

Phase 2: Ivan Judson, Christophe Poulain, Evelyne Viegas, Michael Zyskowski

https://www.codalab.org/competitions/1381

Review, testing: Marc Boullé, Hugo Jair Escalante, Frederick Eberhardt, Seth Flaxman, Patrik Hoyer, Dominik Janzing, Richard Kennaway, Vincent Lemaire, Joris Mooij, Jonas Peters, Florin , Peter Spirtes, Ioannis Tsamardinos, Jianxin Yin, Kun Zhang.

Acknowledgements:

Phase 1:

1st place: Diogo Moitinho de Almeida

2nd place : José Adrián Rodríguez Fonollosa

3rd place : Spyridon Samothrakis

Phase 2:

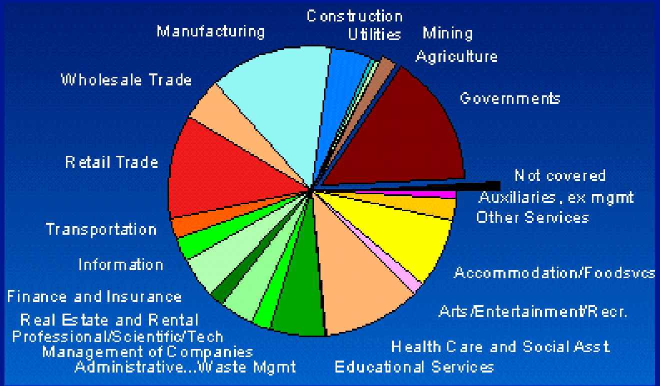
1st place: José Adrián Rodríguez Fonollosa
2nd place: Wei Zhang

3rd prize (and fastest code): David Lopez-Paz

Causal questions:

What affects…

…your health?



… the economy?

…climate changes?

Which actions will have beneficial effects?

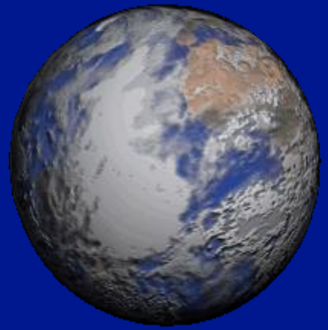
Scientific method:

1. Observe "correlations" A – B.
2. Hypothesize causal relationships:

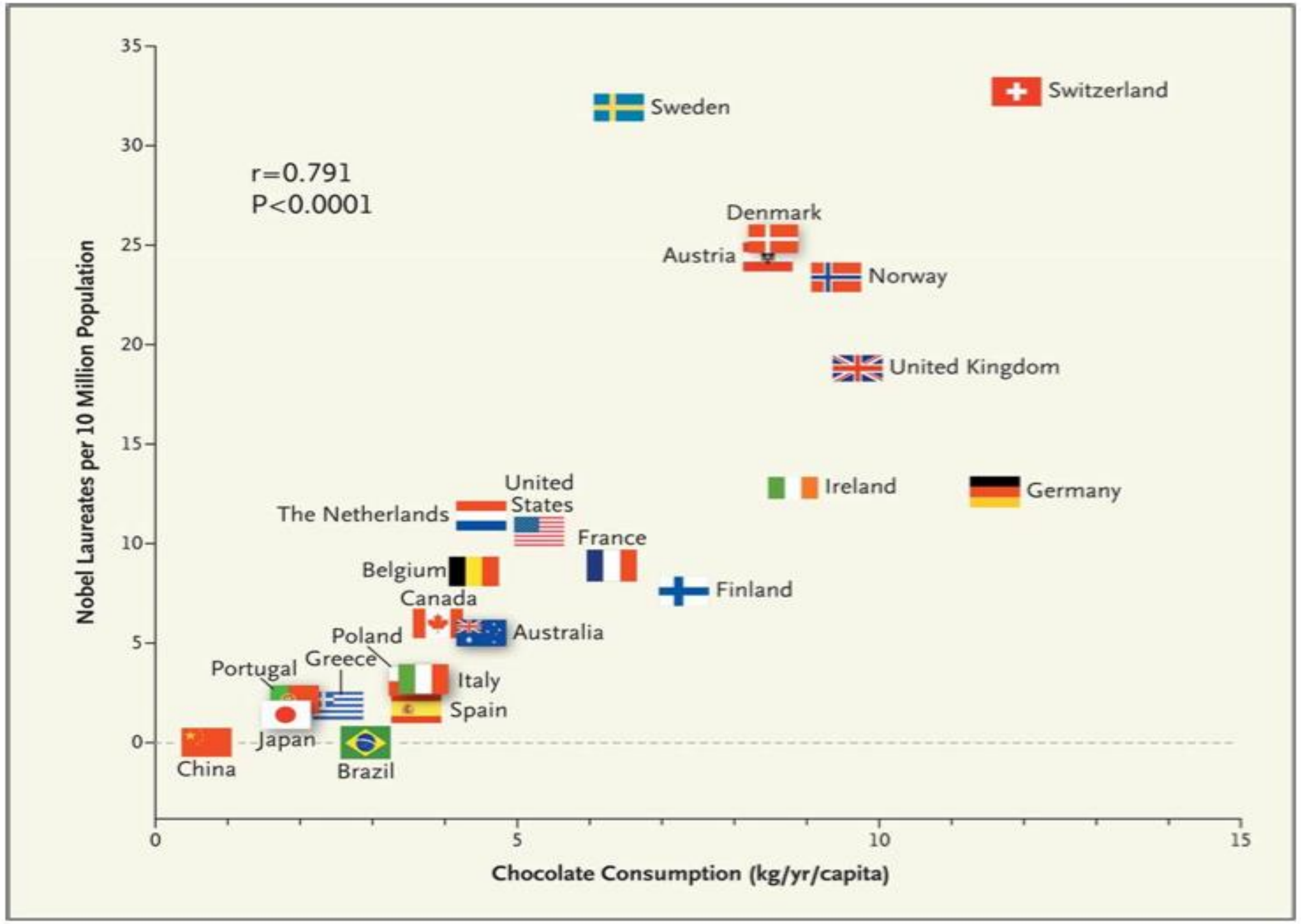
   A -> B

   B -> A

   A<- C -> B
3. Perform experiments.



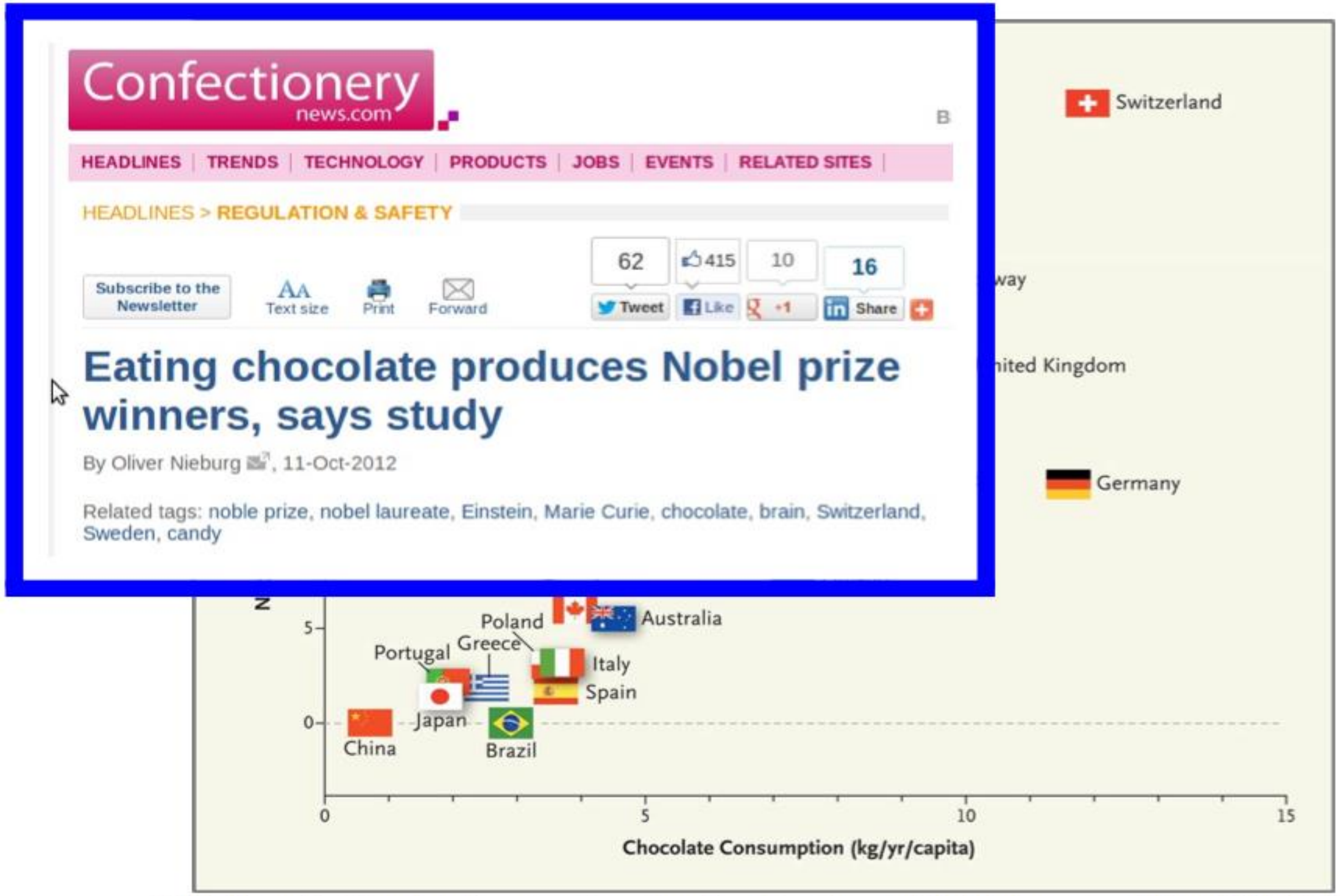
Chocolate correlates with Nobel prize



F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

Thanks to Jonas Peters for this example

Chocolate correlates with Nobel prize



F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

Thanks to Jonas Peters for this example

Chocolate correlates with Nobel prize

Confectionery

HEADLINES | T

HEADLINES >

Subscribe to the Newsletter

Eating winner

By Oliver Niebu

Related tags: Sweden, candy

F. H.

Forbes

New Posts +10 posts this hour

Most Popular Google's Driverless Car

Lis Busi

+ Follow (73)

PHARMA & HEALTHCARE | 10/10/2012 @ 5:02PM | 14,700 views

Chocolate And Nobel Prize In Study

4 comments, 2 called-out + Comment Now + Follow Comments

You don't have to be a genius to like chocolate, but geniuses are more likely to eat lots of chocolate, at least according to a new paper published in the August *New England Journal of Medicine*. Franz Messerli reports a highly

Thanks to Jonas Peters for this example

Our objective:

BIG DATA

Smoking -> Lung cancer

Pollution -> Climate changes

Education -> Crime rate

Alcohol consumption -> Car accidents

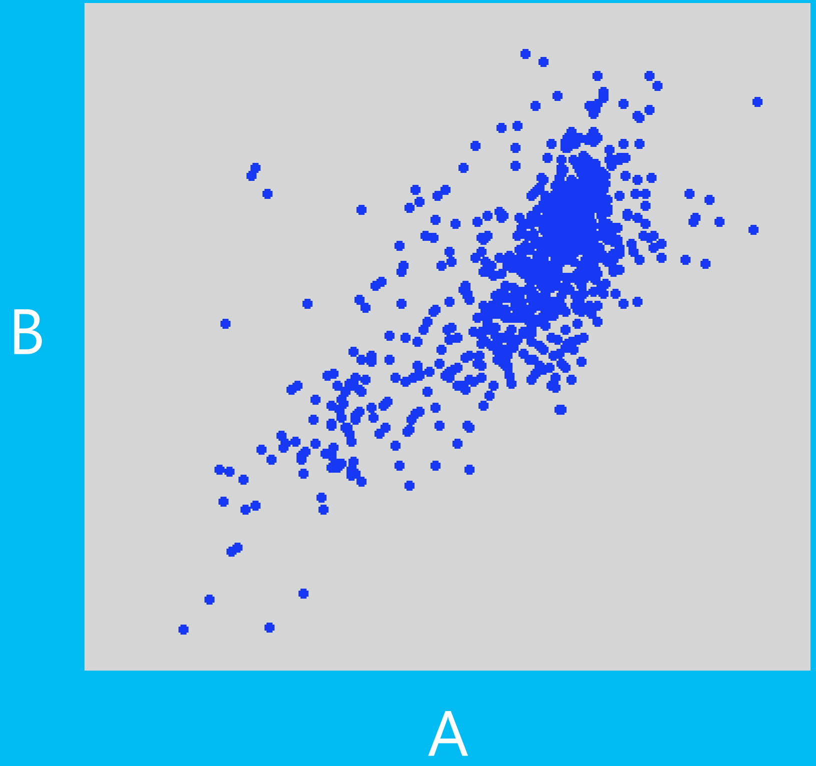
Gender -> Wages

Cholesterol -> Heart disease

Chocolate -> Nobel Prize

The problem:

Random i.i.d. samples of A and B



A -> B
B = f(A, Z)

A | B

A <- C -> B
$A = f(C, Z_1)$
$B = f(C, Z_2)$

B -> A
A = f(B, Z)

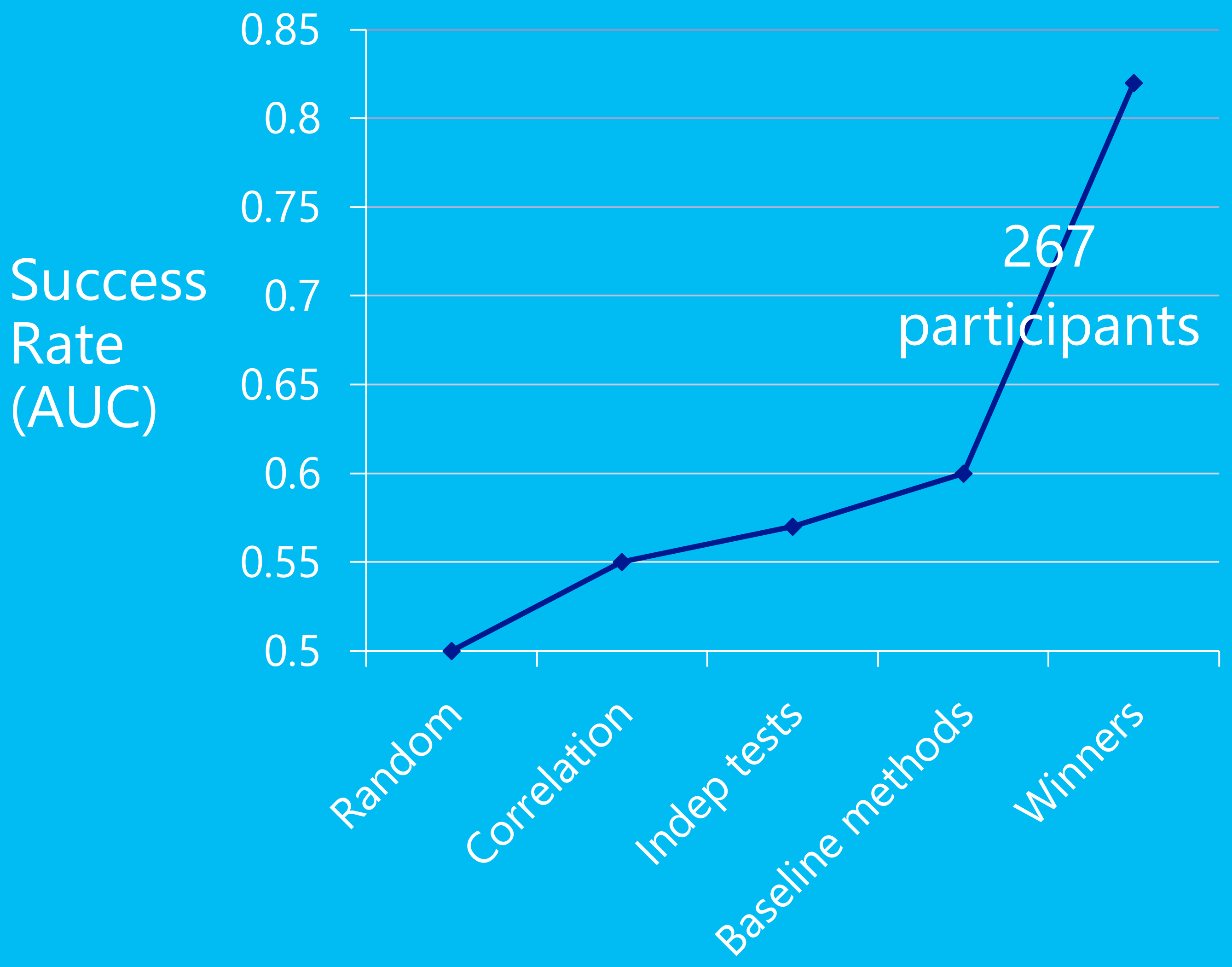
A -> B
B = f(A, Z)

A | B

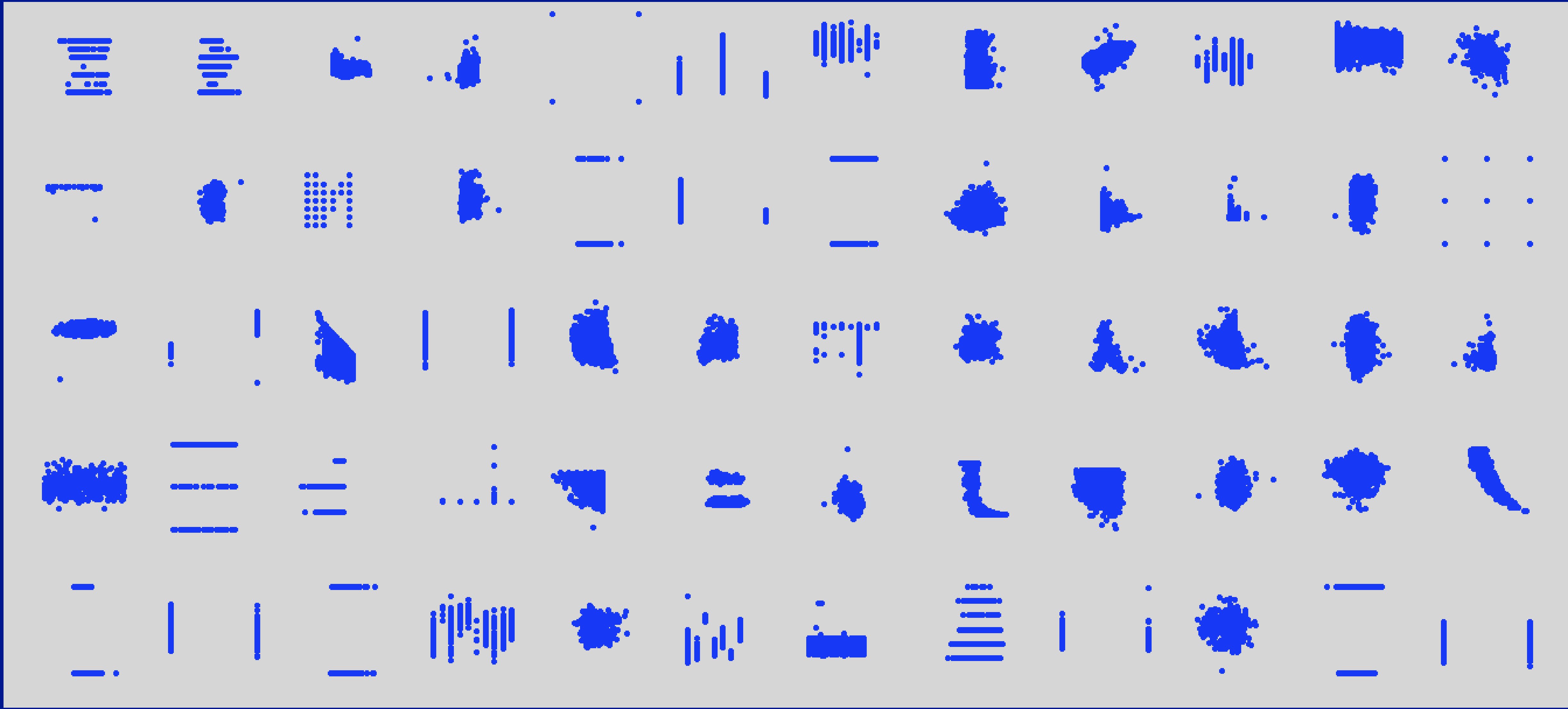
A <- C -> B
$A = f(C, Z_1)$
$B = f(C, Z_2)$

B -> A
A = f(B, Z)

The "solution":



The data:



The data:

Demographics:
Sex -> Height
Age -> Wages
Native country -> Education
Latitude -> Infant mortality

Ecology:
City elevation -> Te
Water level -> Alga
Elevation -> Vegeta
Distance to hydrology -> Fire

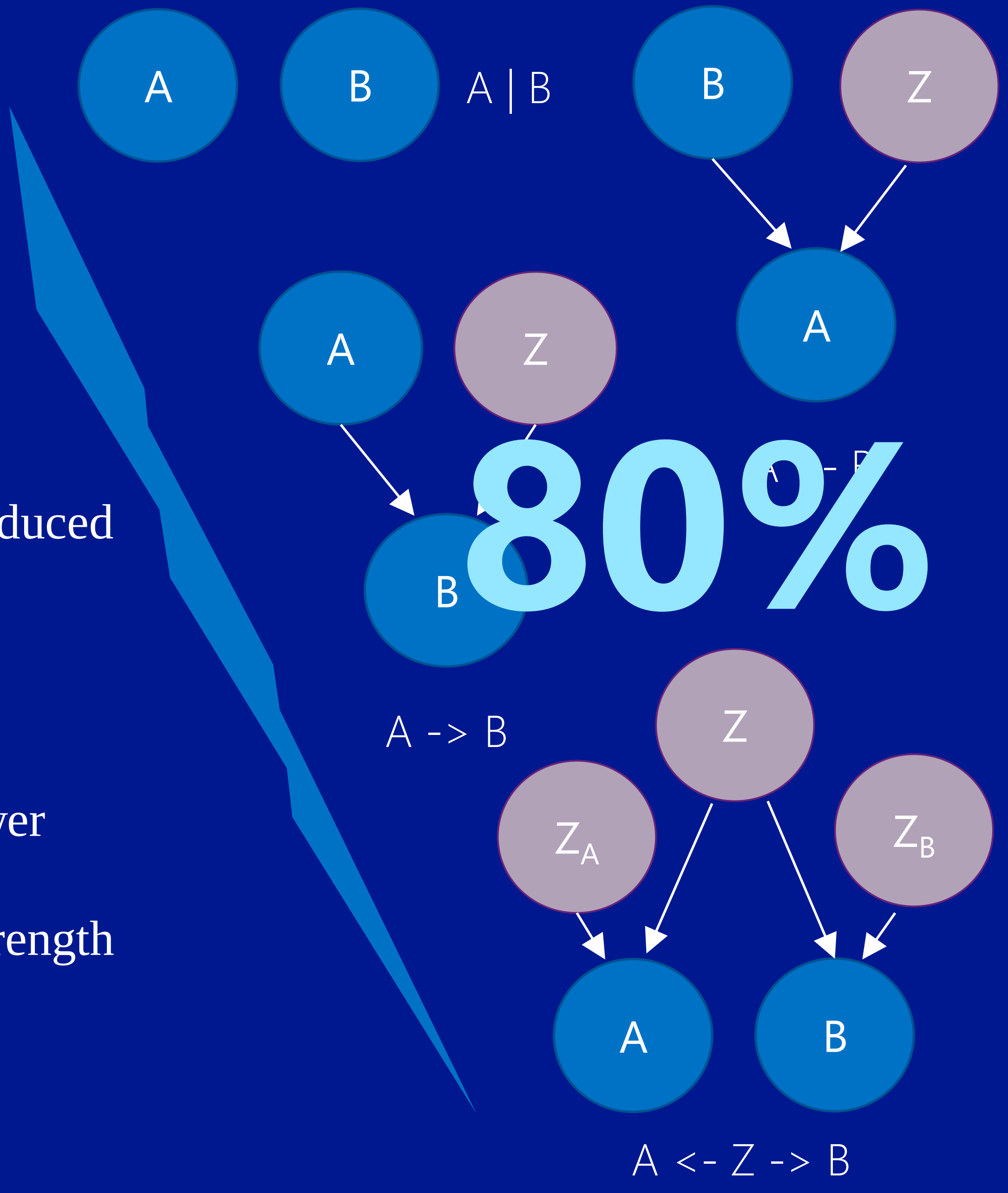
Econometrics:
Mileage -> Car resell price
Number of rooms -> House price
Trade price last day -> Trade price

Medicine:
Cancer volume -> Recurrence
Metastasis -> Prognosis
Age -> Blood pressure

RNA level):
actor -> protein induced

r -> Horsepower
Number of cylinders -> MPG
Cache memory -> Compute power
Roof area -> Heating load
Cement used -> Compressive strength

20%



Results:

Artificial data					
Rank	Team	Dependency	Confounding	Causality	Score
1	ProtoML	0.95372	0.76944	0.90946	0.84206
2	jarfo	0.98063	0.83663	0.89425	0.83499
3	HiDloN	0.94416	0.76777	0.89466	0.82883
4	FirfiD	0.97644	0.80086	0.88644	0.82249
5	mouse	0.94966	0.75831	0.86722	0.80620
6	Domcasto & Sayani	0.91789	0.72655	0.86299	0.79507

Real data					
Rank	Team	Dependency	Confounding	Causality	Score
1	ProtoML	0.88057	0.65432	0.75756	0.70420
2	jarfo	0.95721	0.70386	0.73312	0.68642
3	HiDloN	0.91476	0.69209	0.74774	0.69669
4	FirfiD	0.92352	0.69547	0.73960	0.68274
5	mouse	0.87689	0.64211	0.75008	0.69259
6	Domcasto & Sayani	0.85339	0.65786	0.78075	0.71355

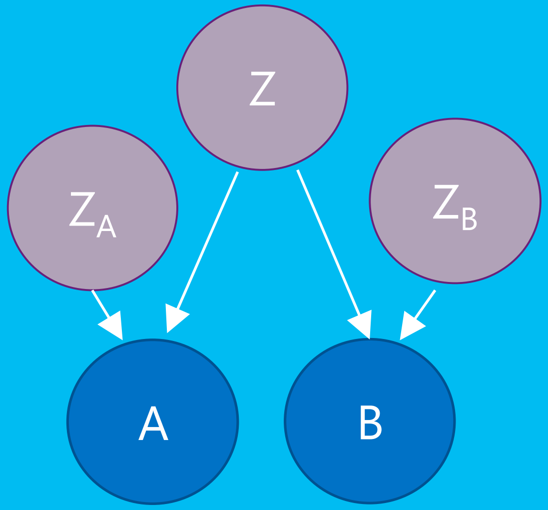
Results:

A B A | B

Artificial data					
Rank	Team	Dependency	Confounding	Causality	Score
1	ProtoML	0.95372	0.76944	0.90946	0.84206
2	jarfo	0.98063	0.83663	0.89425	0.83499
3	HiDloN	0.94416	0.76777	0.89466	0.82883
4	FirfiD	0.97644	0.80086	0.88644	0.82249
5	mouse	0.94966	0.75831	0.86722	0.80620
6	Domcasto & Sayani	0.91789	0.72655	0.86299	0.79507

Real data					
Rank	Team	Dependency	Confounding	Causality	Score
1	ProtoML	0.88057	0.65432	0.75756	0.70420
2	jarfo	0.95721	0.70386	0.73312	0.68642
3	HiDloN	0.91476	0.69209	0.74774	0.69669
4	FirfiD	0.92352	0.69547	0.73960	0.68274
5	mouse	0.87689	0.64211	0.75008	0.69259
6	Domcasto & Sayani	0.85339	0.65786	0.78075	0.71355

Results:



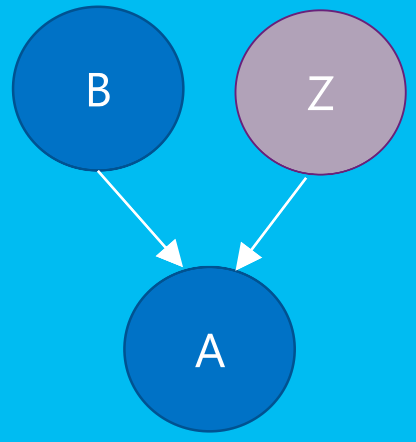
Artificial data					
Rank	Team	Dependency	Confounding	Causality	Score
1	ProtoML	0.95372	0.76944	0.90946	0.84206
2	jarfo	0.98063	0.83663	0.89425	0.83499
3	HiDloN	0.94416	0.76777	0.89466	0.82883
4	FirfiD	0.97644	0.80086	0.88644	0.82249
5	mouse	0.94966	0.75831	0.86722	0.80620
6	Domcasto & Sayani	0.91789	0.72655	0.86299	0.79507

Real data					
Rank	Team	Dependency	Confounding	Causality	Score
1	ProtoML	0.88057	0.65432	0.75756	0.70420
2	jarfo	0.95721	0.70386	0.73312	0.68642
3	HiDloN	0.91476	0.69209	0.74774	0.69669
4	FirfiD	0.92352	0.69547	0.73960	0.68274
5	mouse	0.87689	0.64211	0.75008	0.69259
6	Domcasto & Sayani	0.85339	0.65786	0.78075	0.71355

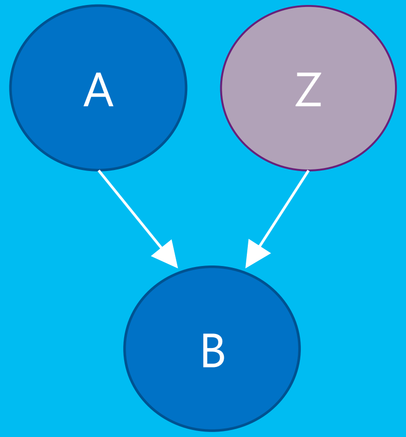
Results:

Artificial data					
Rank	Team	Dependency	Confounding	Causality	Score
1	ProtoML	0.95372	0.76944	0.90946	0.84206
2	jarfo	0.98063	0.83663	0.89425	0.83499
3	HiDloN	0.94416	0.76777	0.89466	0.82883
4	FirfiD	0.97644	0.80086	0.88644	0.82249
5	mouse	0.94966	0.75831	0.86722	0.80620
6	Domcasto & Sayani	0.91789	0.72655	0.86299	0.79507

Real data					
Rank	Team	Dependency	Confounding	Causality	Score
1	ProtoML	0.88057	0.65432	0.75756	0.70420
2	jarfo	0.95721	0.70386	0.73312	0.68642
3	HiDloN	0.91476	0.69209	0.74774	0.69669
4	FirfiD	0.92352	0.69547	0.73960	0.68274
5	mouse	0.87689	0.64211	0.75008	0.69259
6	Domcasto & Sayani	0.85339	0.65786	0.78075	0.71355

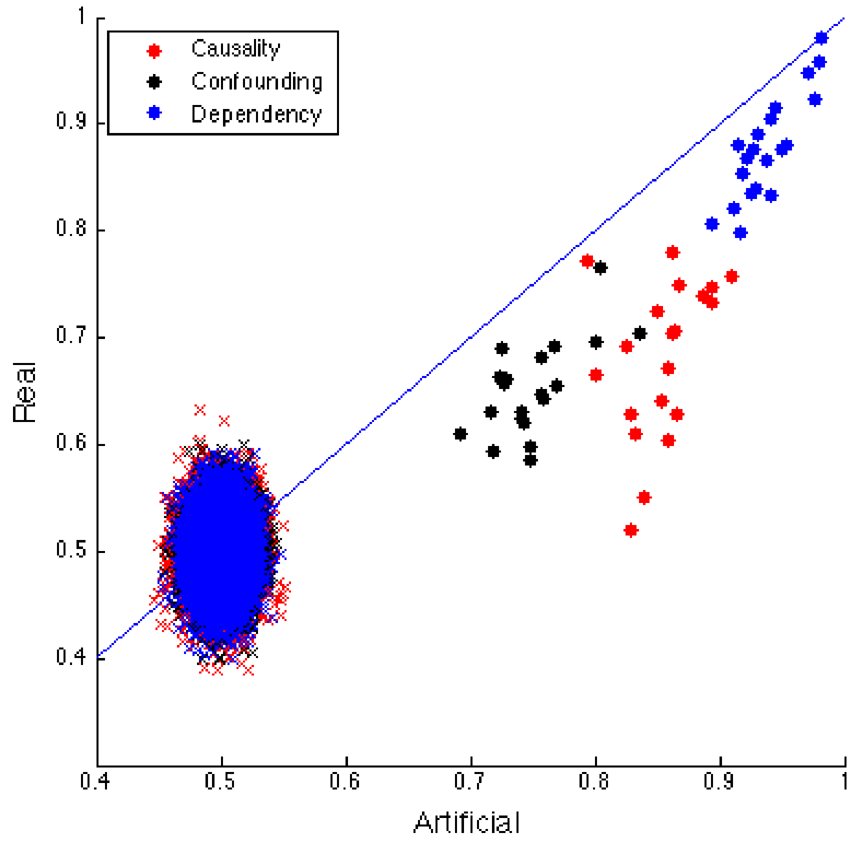


A <- B



A -> B

Results:

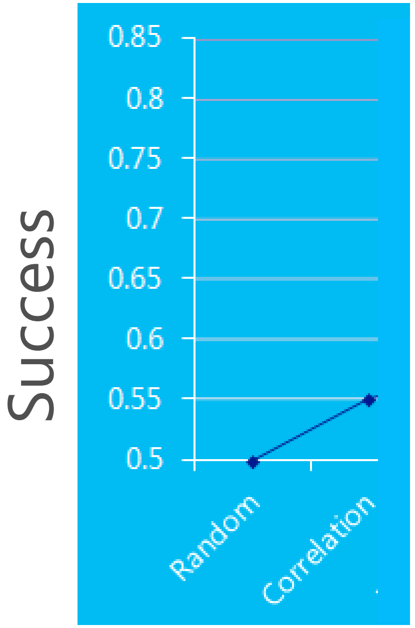


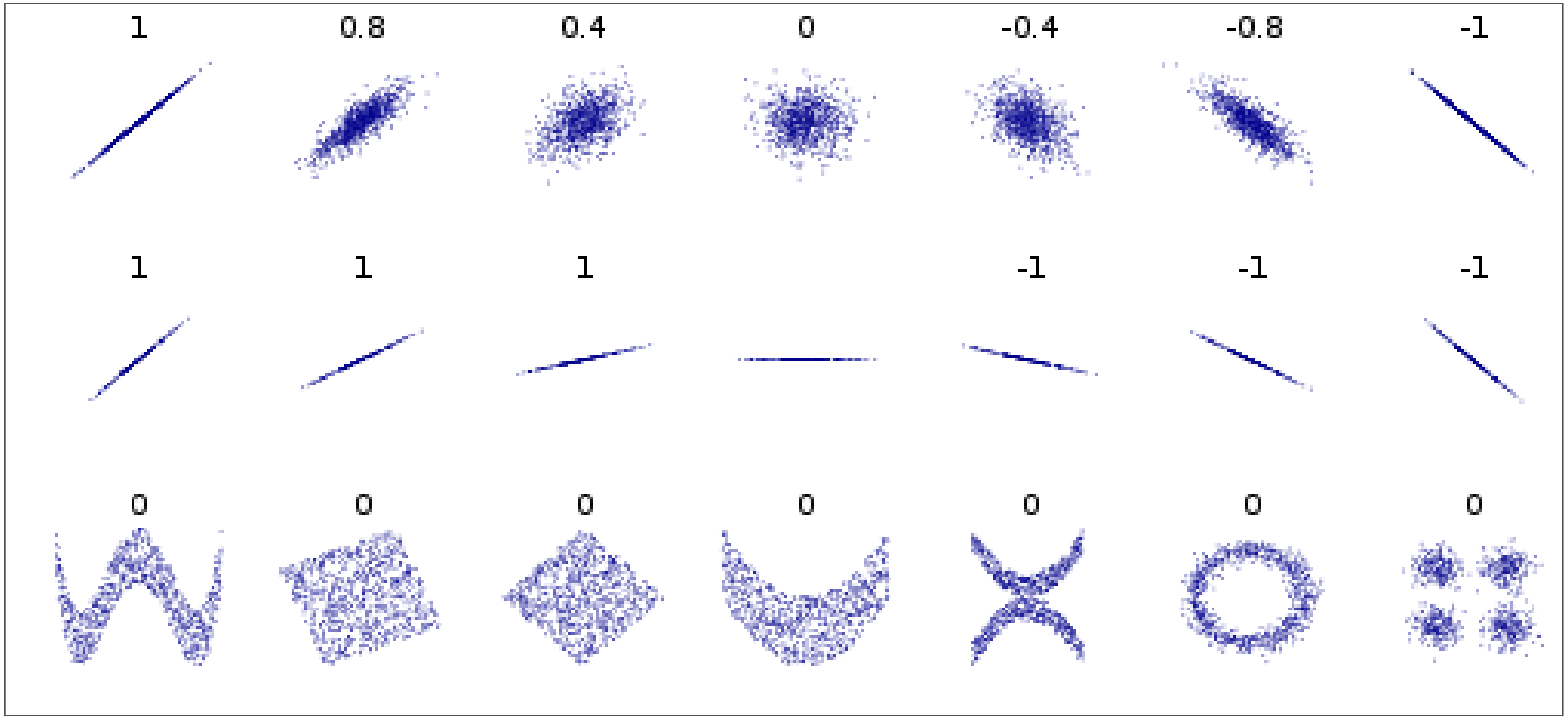
Score = Area under ROC curve (random=0.5, perfect=1)

- **Causality**: separate A->B vs. B->A
- **Counfounding**: separate A – B vs. (A->B or B->A)
- **Dependency**: separate A | B vs. (A->B or B->A)

Correlation Coefficient

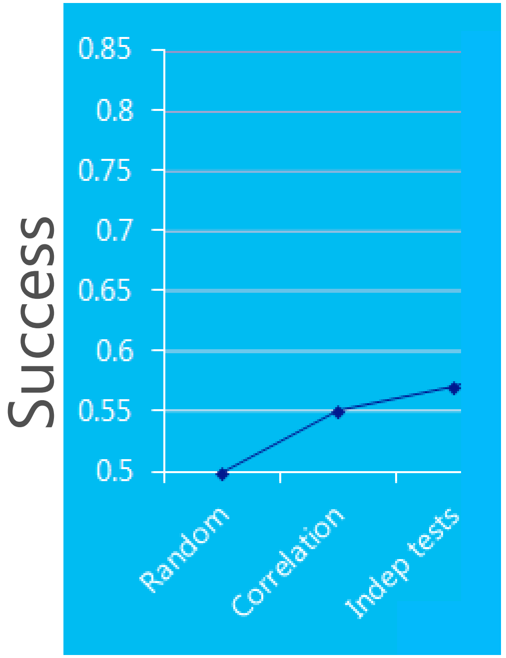
Pearson Correlation Coefficient: $C(A,B) = cov(A,B)/(\sigma_A \sigma_B)$

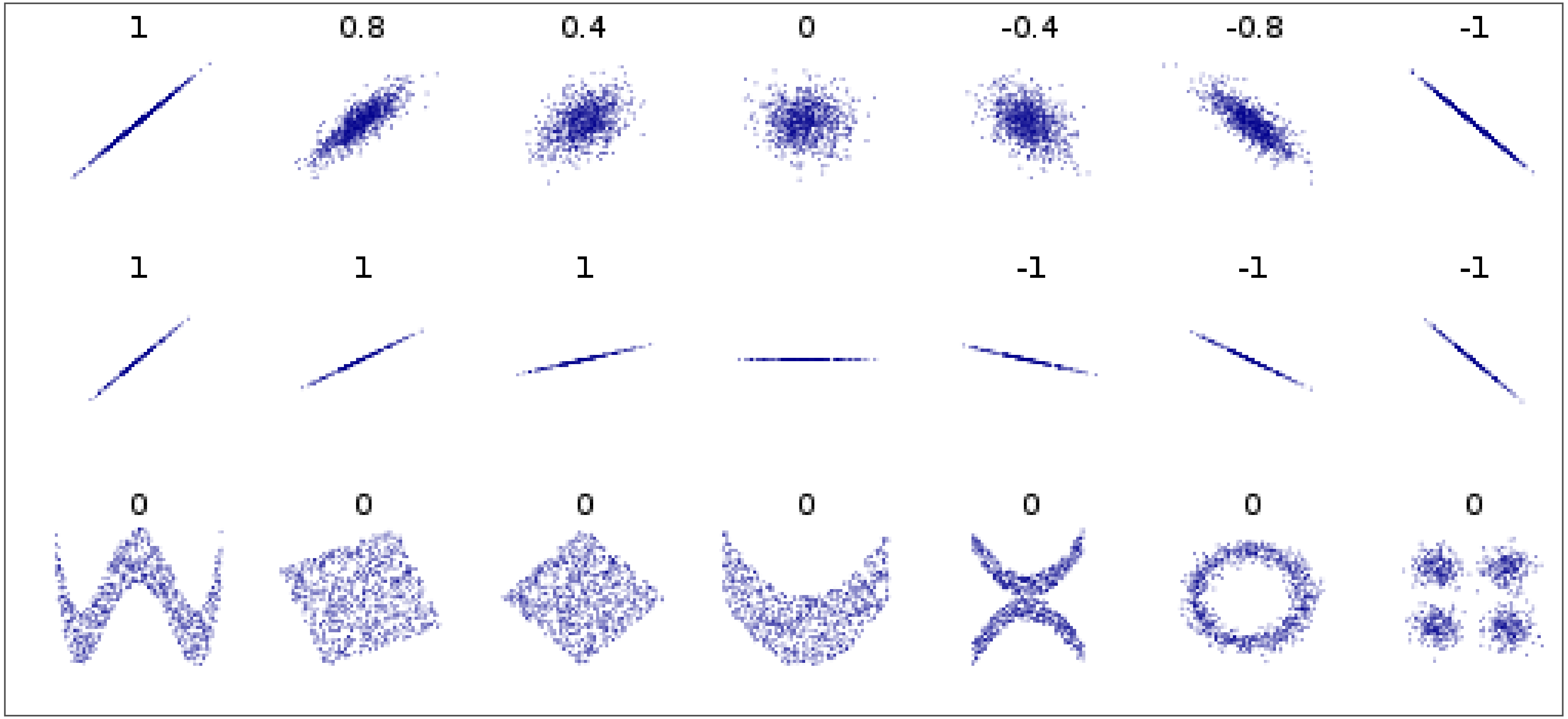




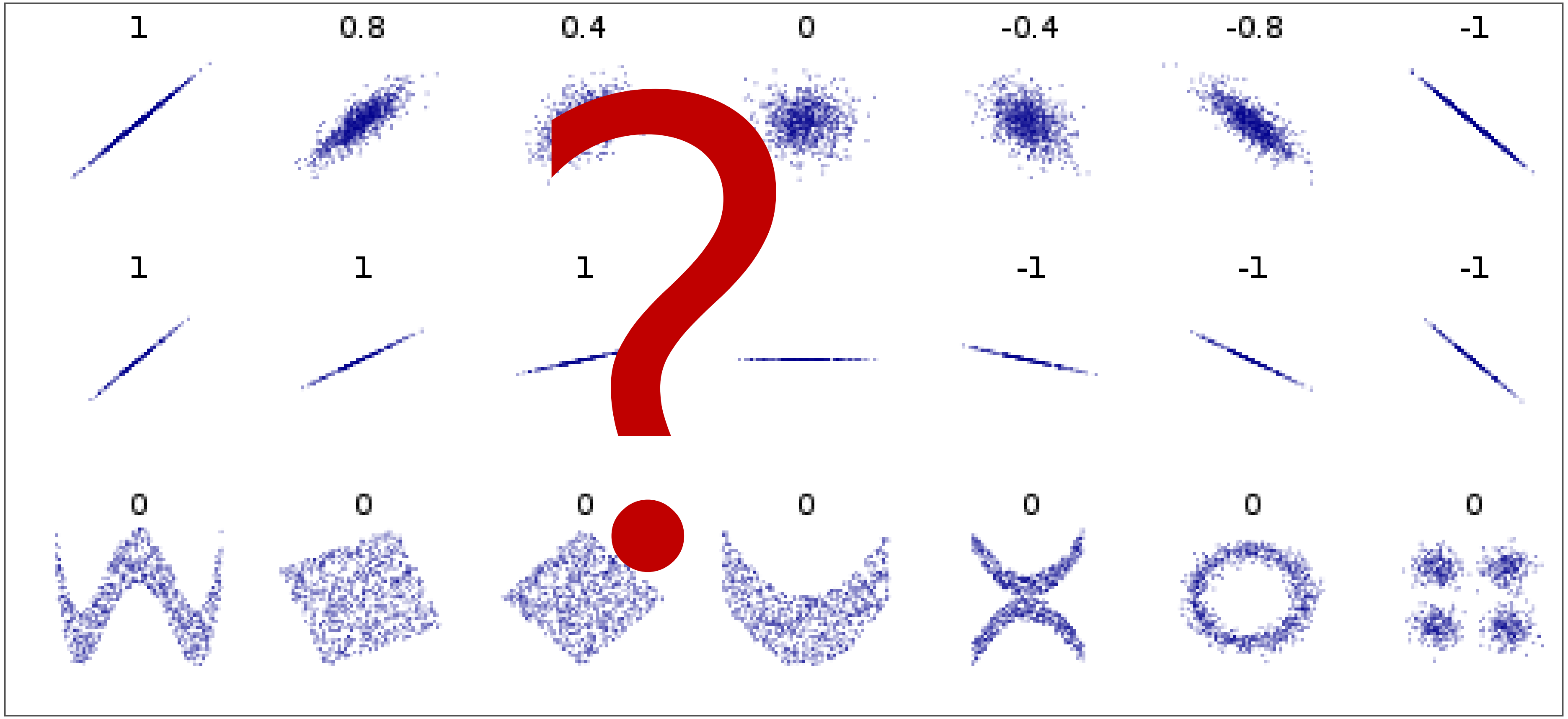
Denis Boigelot, Université libre de Bruxelles, Belgium, Wikipedia

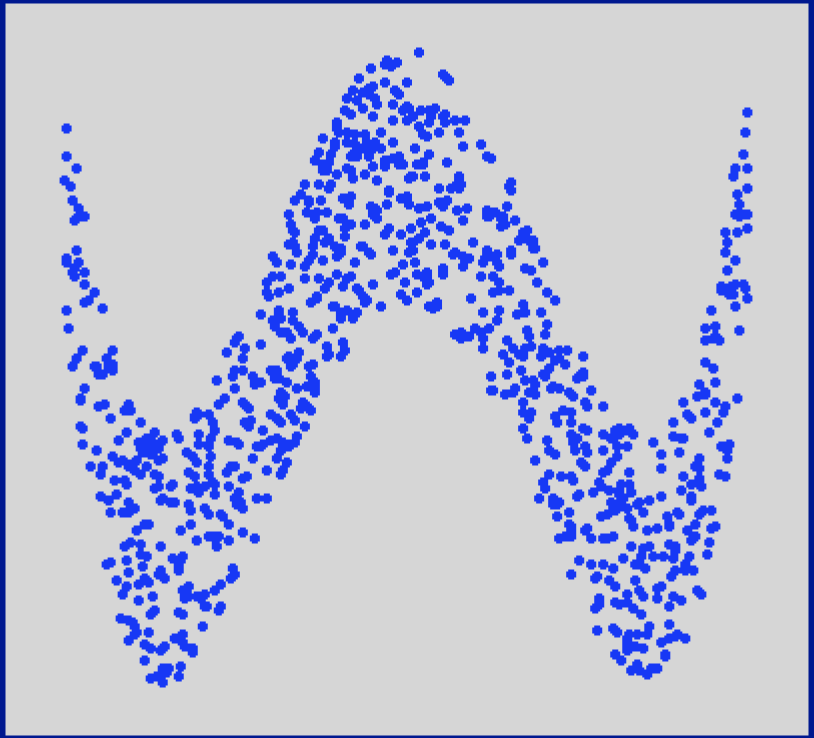
Independence tests (MI, HSIC)





Causation Coefficient





B

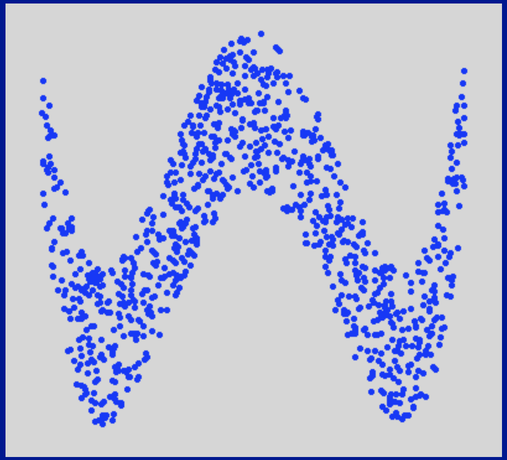
A

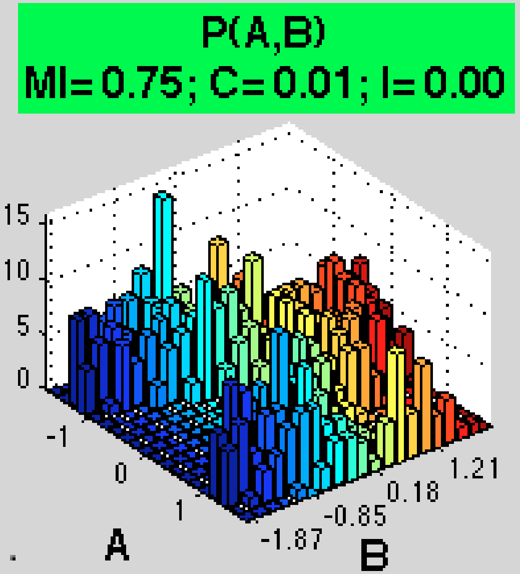
n=1000

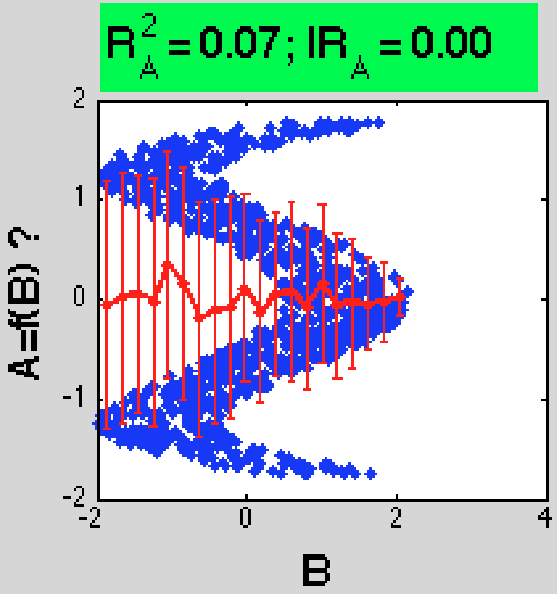
A = unif(n, -1, 1)

noise = unif(n, -1, 1)/3

$$B = 4\,(A^2 - 1/2)^2 + \text{noise}$$







Correlation:

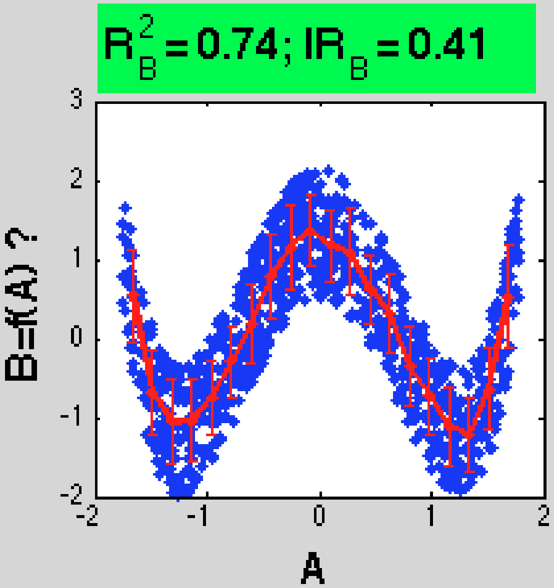
$C(A,B) = cov(A,B)/(\sigma_A \sigma_B)$

Mutual information:

$$MI(A,B) = H(A) + H(B) - H(A,B)$$
$$= KL[p(A,B) \,||\, p(A)p(B)]$$

HSIC:

$I(A,B) = pval(||C_{AB}||_{HS}^2)$



Residual:

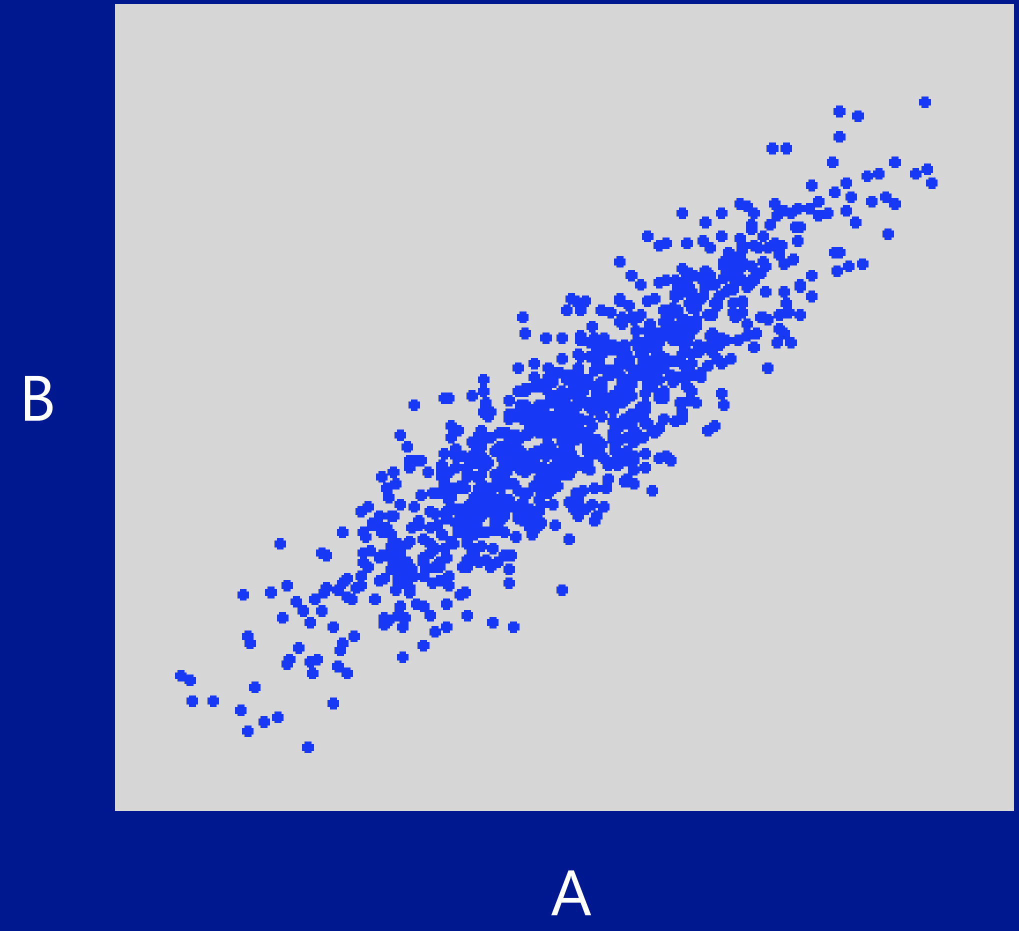
$res(B)^2=(1/n)\Sigma_i (f(A_i) - B_i)^2$

Coefficient of determination:

$R_B^2 = 1 - res(B)^2 / \sigma_B^2$

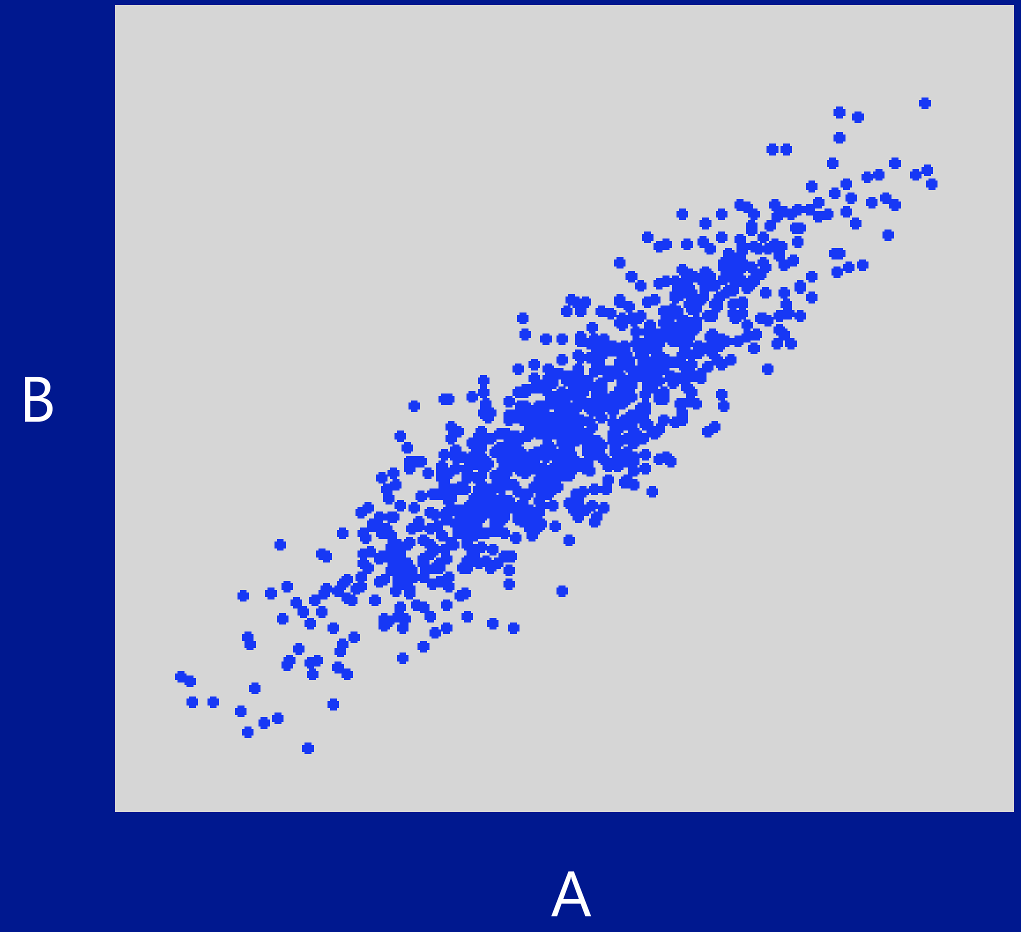
Independence(Input, Residual):

$IR_B=I(A,res(B))$


B
A

?

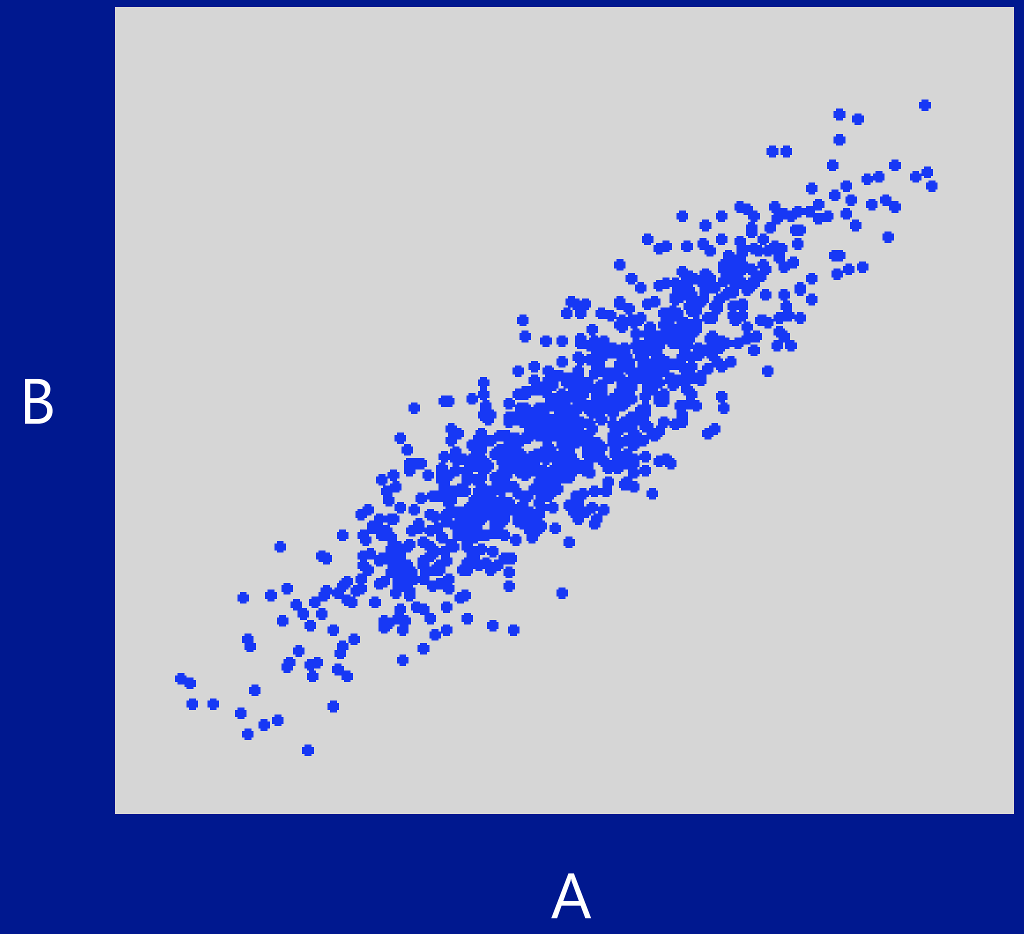
A -> B



n=1000
A = normal(n, 0, 1)
noise = normal(n, 0, 1)

B = 2 A + noise

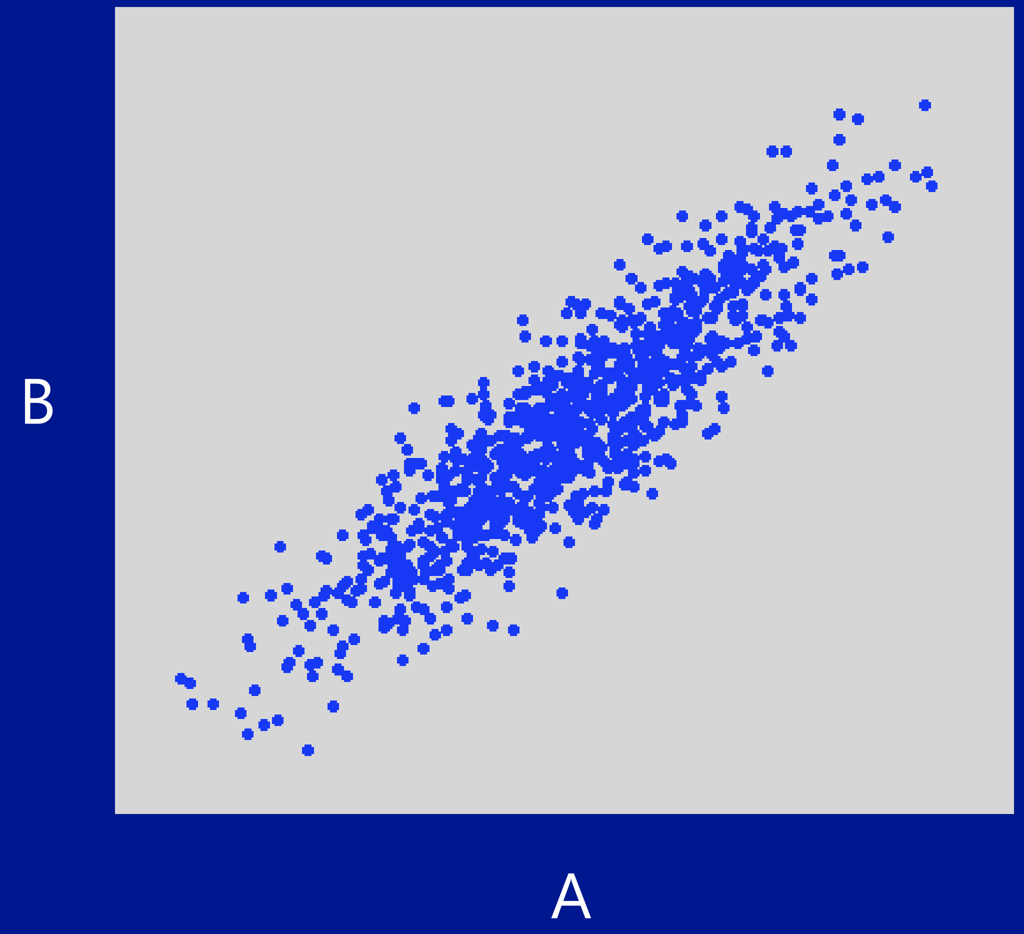
B -> A



n=1000
B = normal(n, 0, 1)
noise = normal(n, 0, 1)

A = 2 B + noise

A <- C -> B



n=1000

C = normal(n, 0, 1)

noiseA = normal(n, 0, $1/\sqrt{2}$)

noiseB = normal(n, 0, $1/\sqrt{2}$)

A = 2 C+ noiseA

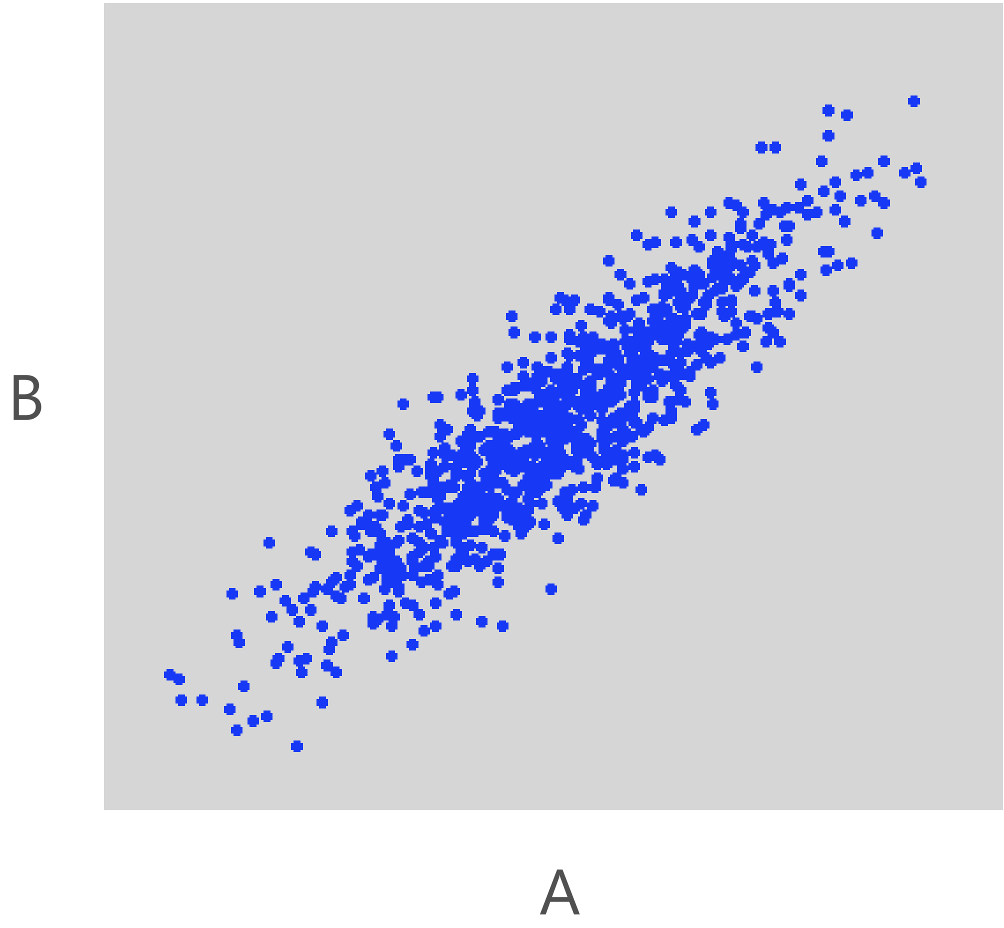
B = 2 C+ noiseB

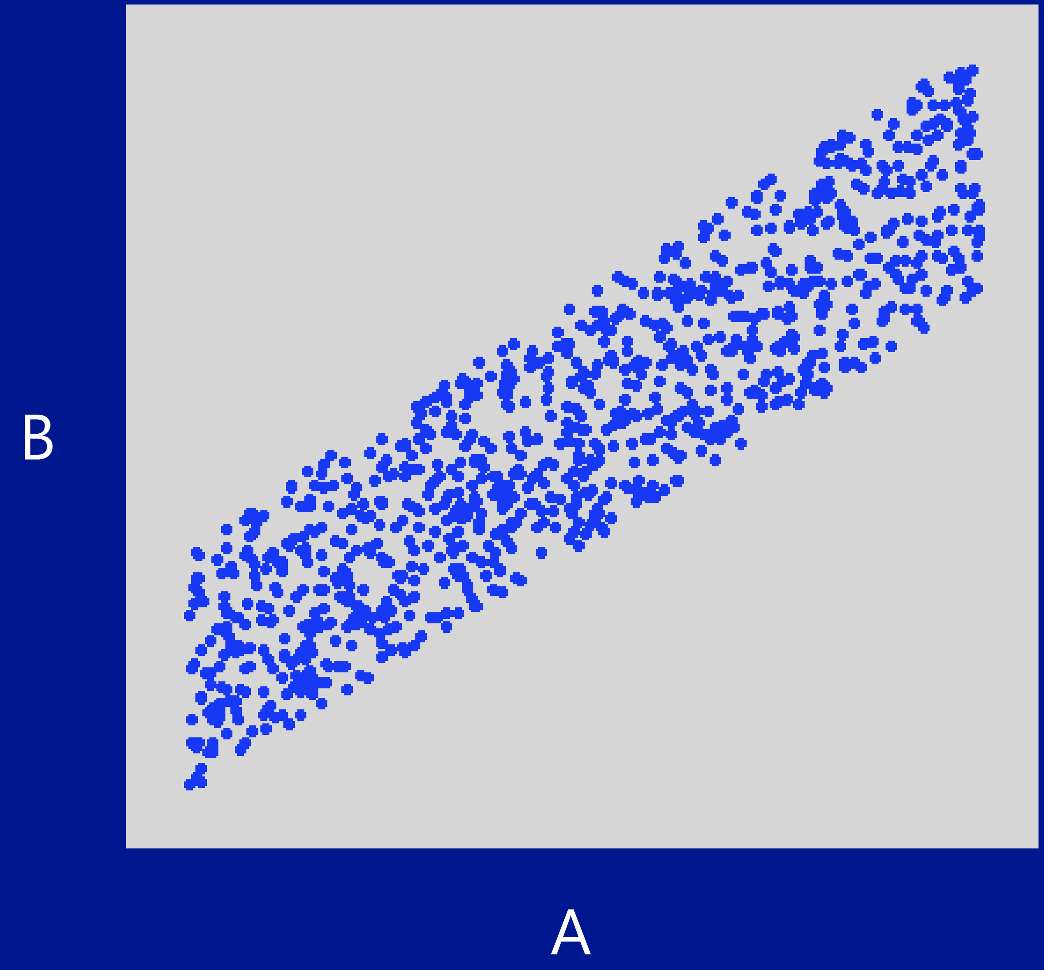
Non identifiable case

Linear

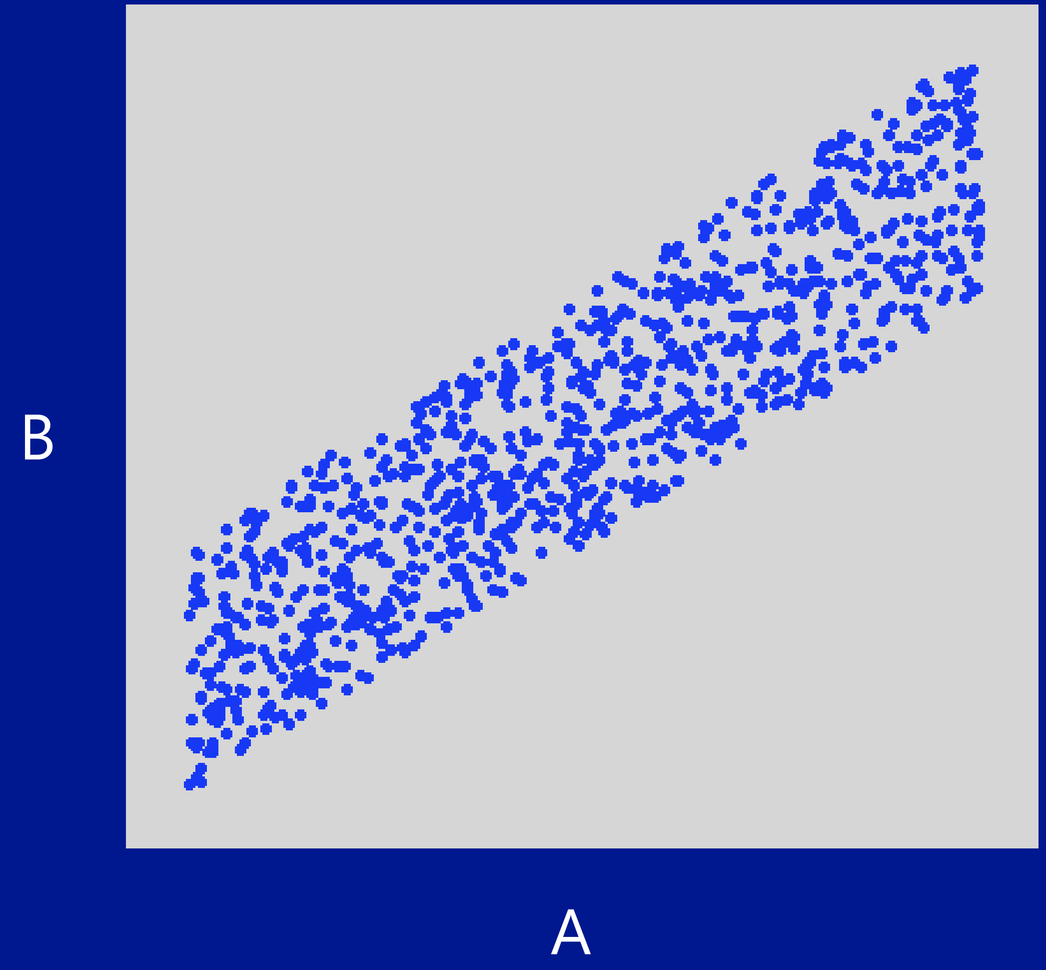
Gaussian input

Gaussian noise




B
A

?

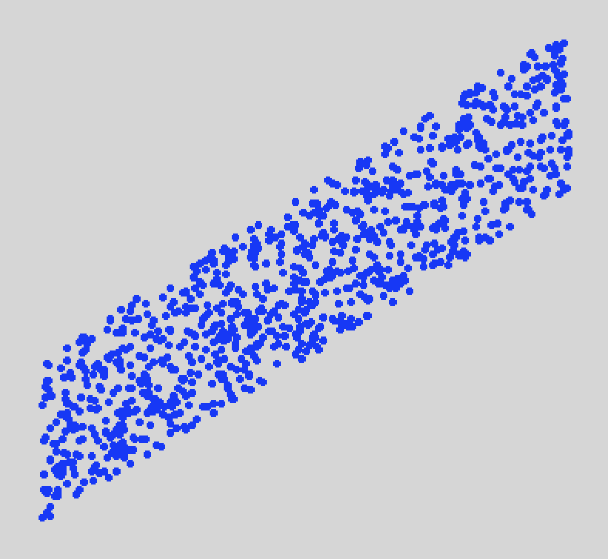


n=1000

A = unif(n, -1, 1)

noise = unif(n, -1, 1)/2

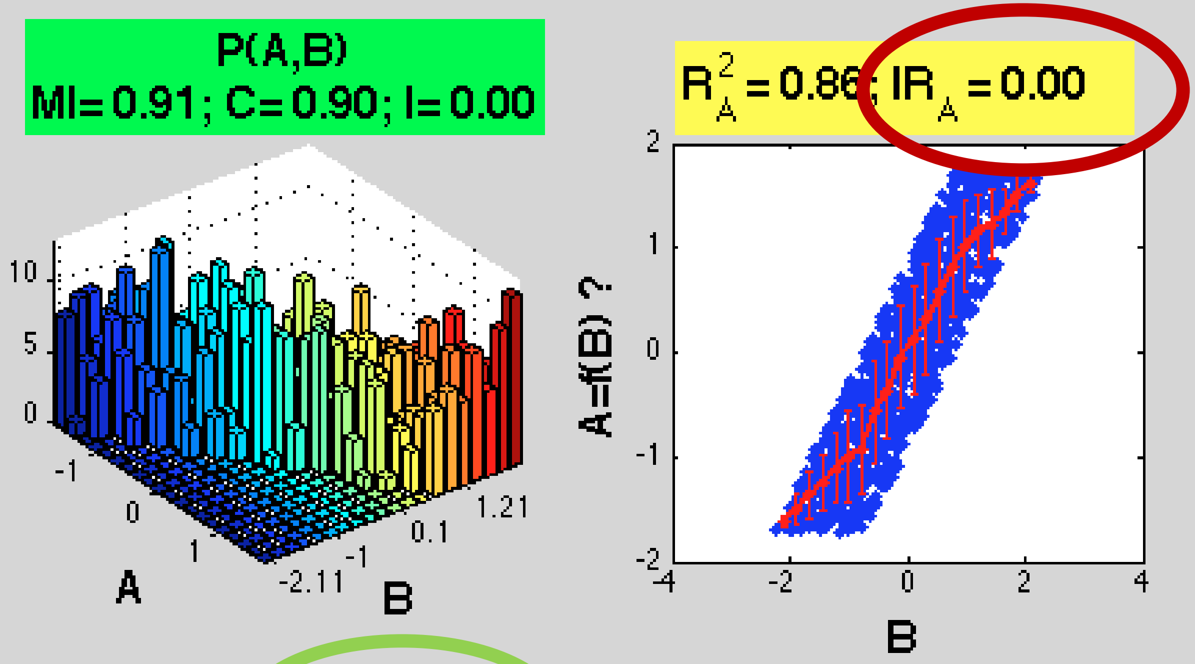
B = A + noise

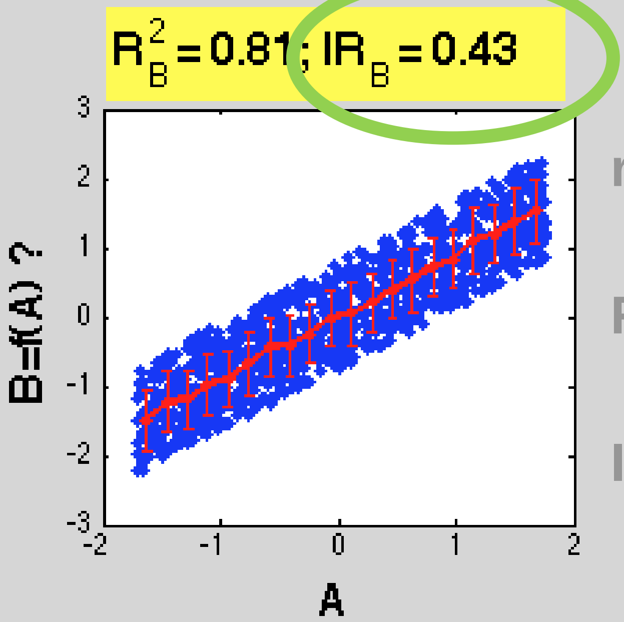


$$C(A,B) = \mathrm{cov}(A,B)/(\sigma_A\sigma_B)$$

$$MI(A,B) = H(A) + H(B) - H(A,B)$$
$$= KL[p(A,B) \,||\, p(A)p(B)]$$

$$I(A,B) = \mathrm{pval}(\ ||C_{AB}||_{HS}^2\ )$$



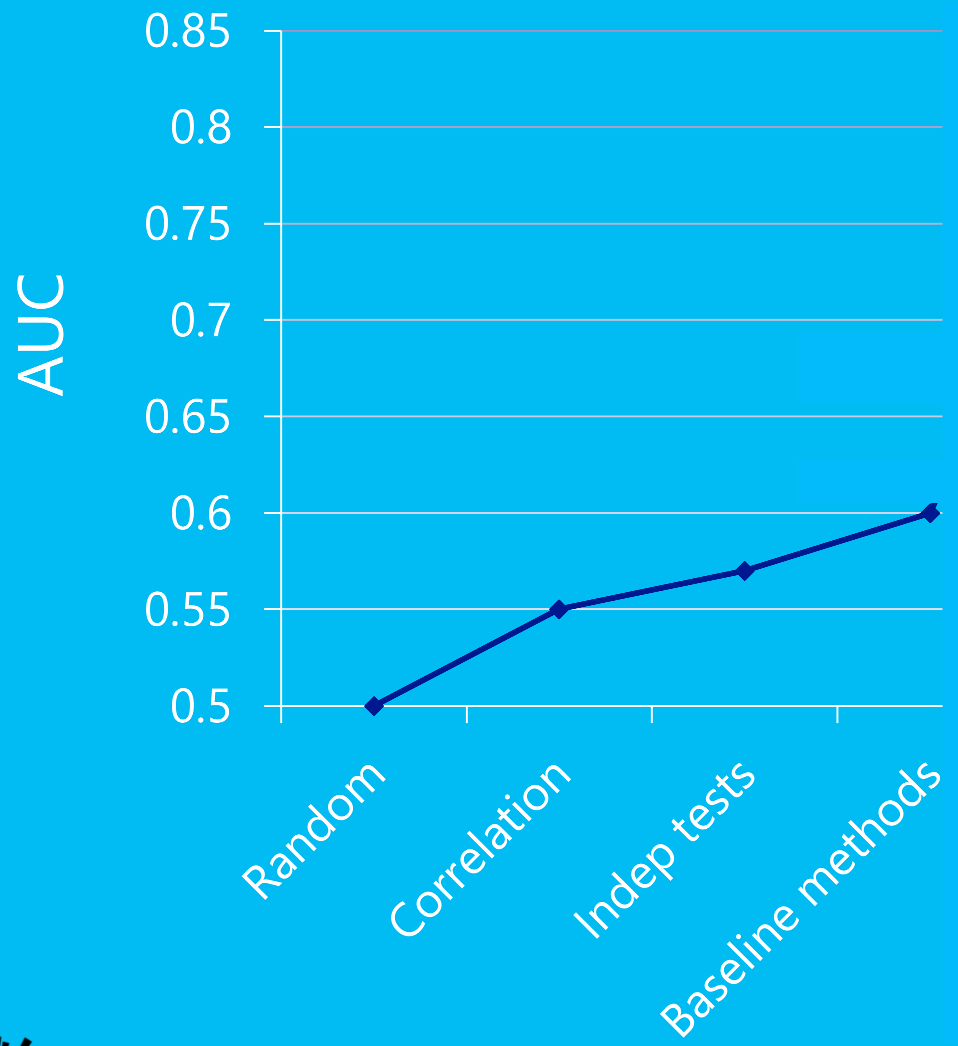


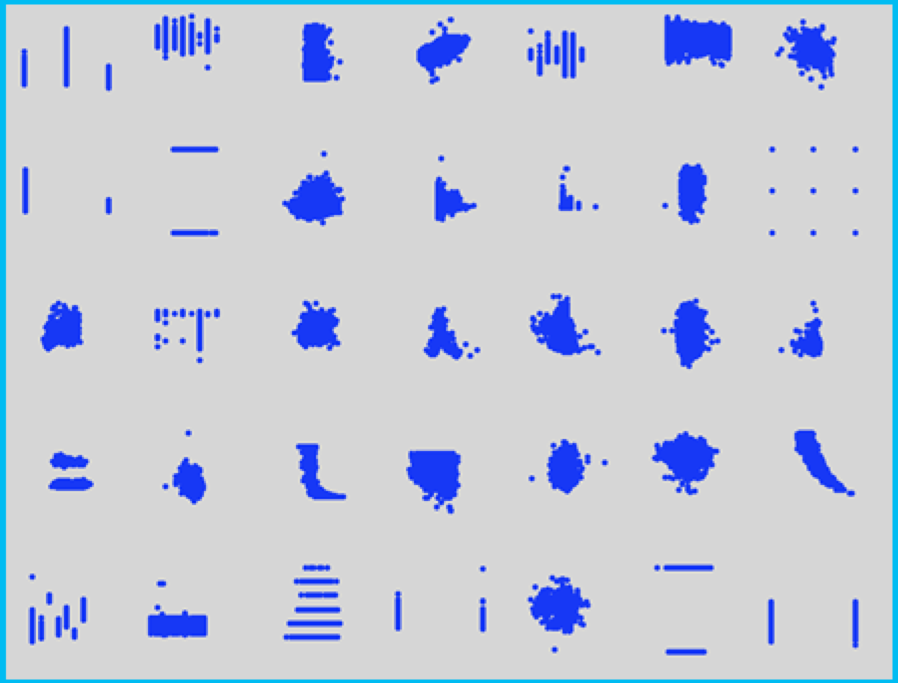
$$res(B)^2=(1/n)\Sigma_i\,(f(A_i)-B_i)^2$$

$$R_B^2 = 1 - res(B)^2 / \sigma_B^2$$

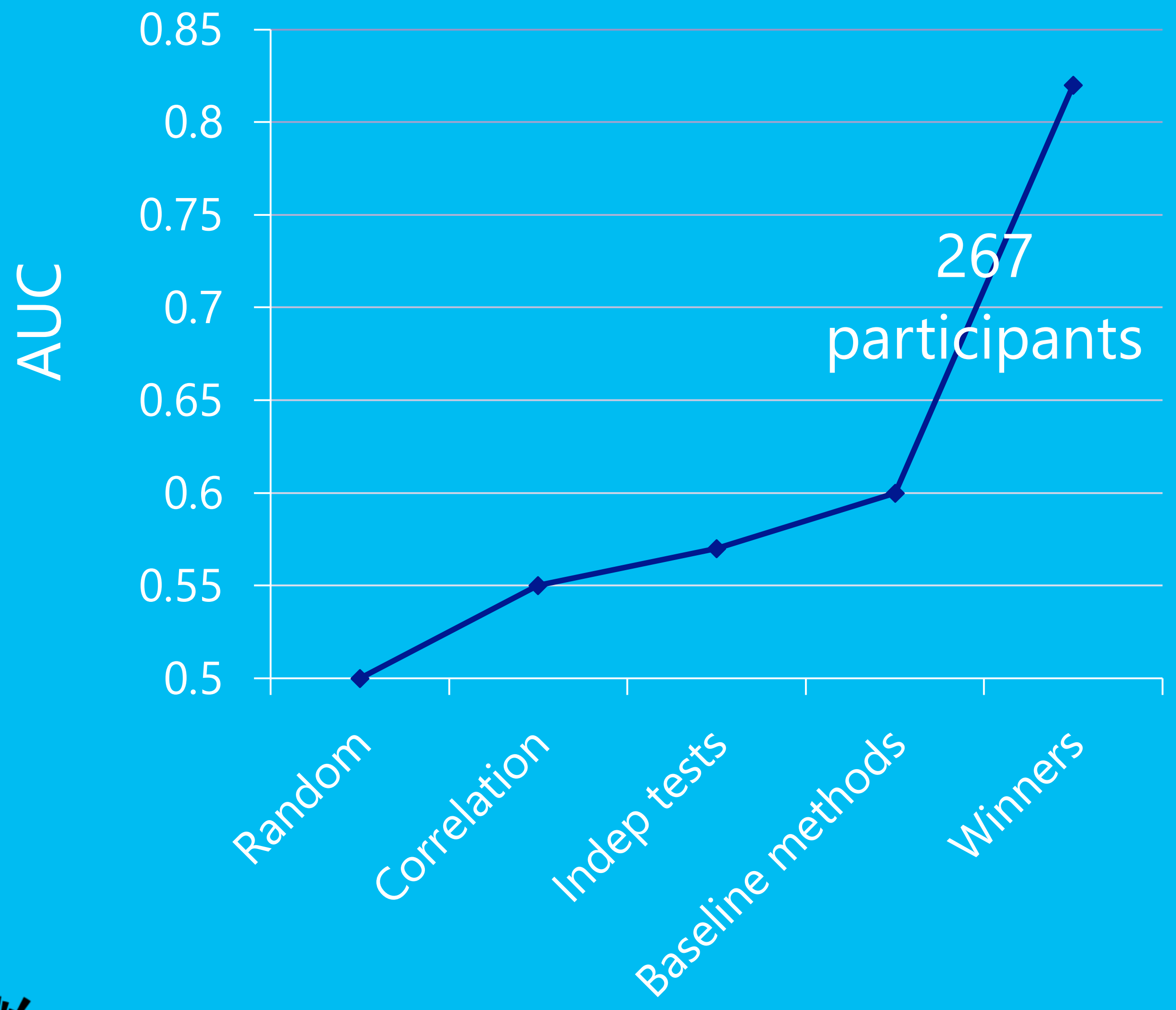
$$IR_B=I(A,res(B))$$

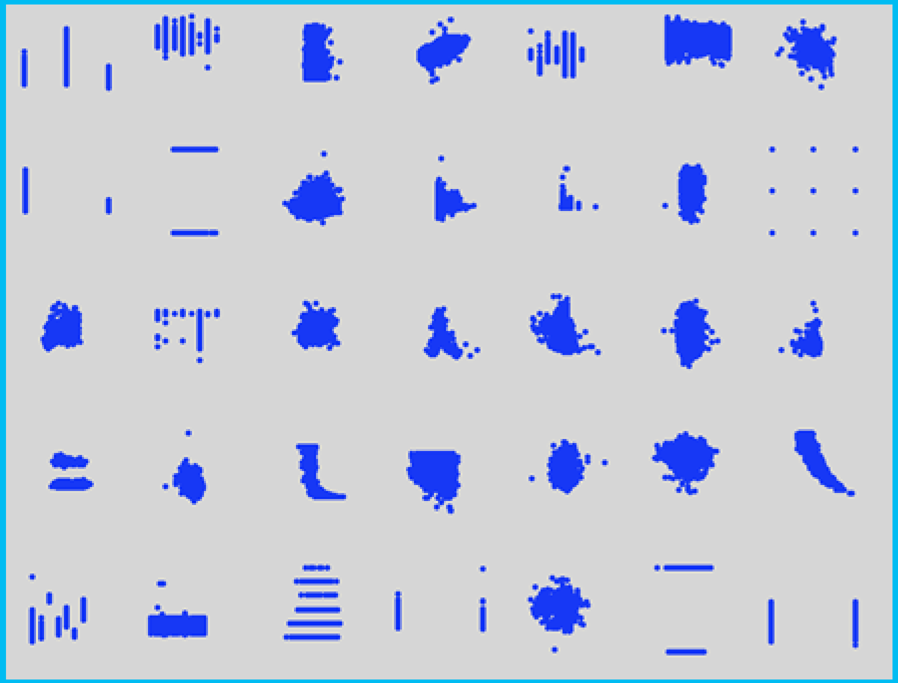
Baseline methods:





Winning methods:





Typical research methods

Restrict setting

- Only continuous variables
- Non Gaussian noise
- Non linear, no noise

Study a particular model

- ANM
- PNL
- GP

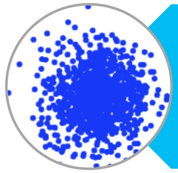
Propose a statistical test

- Identifiability
- Consistency, bias, variance, power
- Computational properties

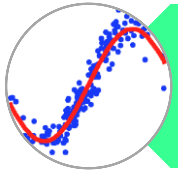
Typical challenge methods

Preprocess data	• Standardize • Discretize • Re-label categorical variables
Compute a lot of features	• ProtoML: 20,000 (324 after selection) • Jarfo: 43 features • FirfID: 30 features
Use non-linear classifier (ensemble of decision trees).	• RF • GTB

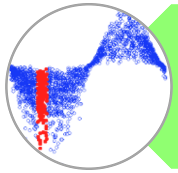
Typical features



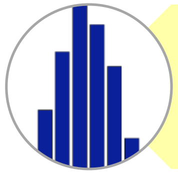
Independence tests: Pearson correlation, Spearman correlation, HSIC, Mutual Information.



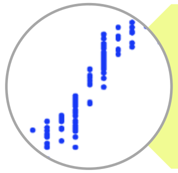
Curve fitting: Residuals of fits with various models and loss functions; model complexity; independence of residual and input.



Discretized conditional distribution: Variance, and other moments and statistics.

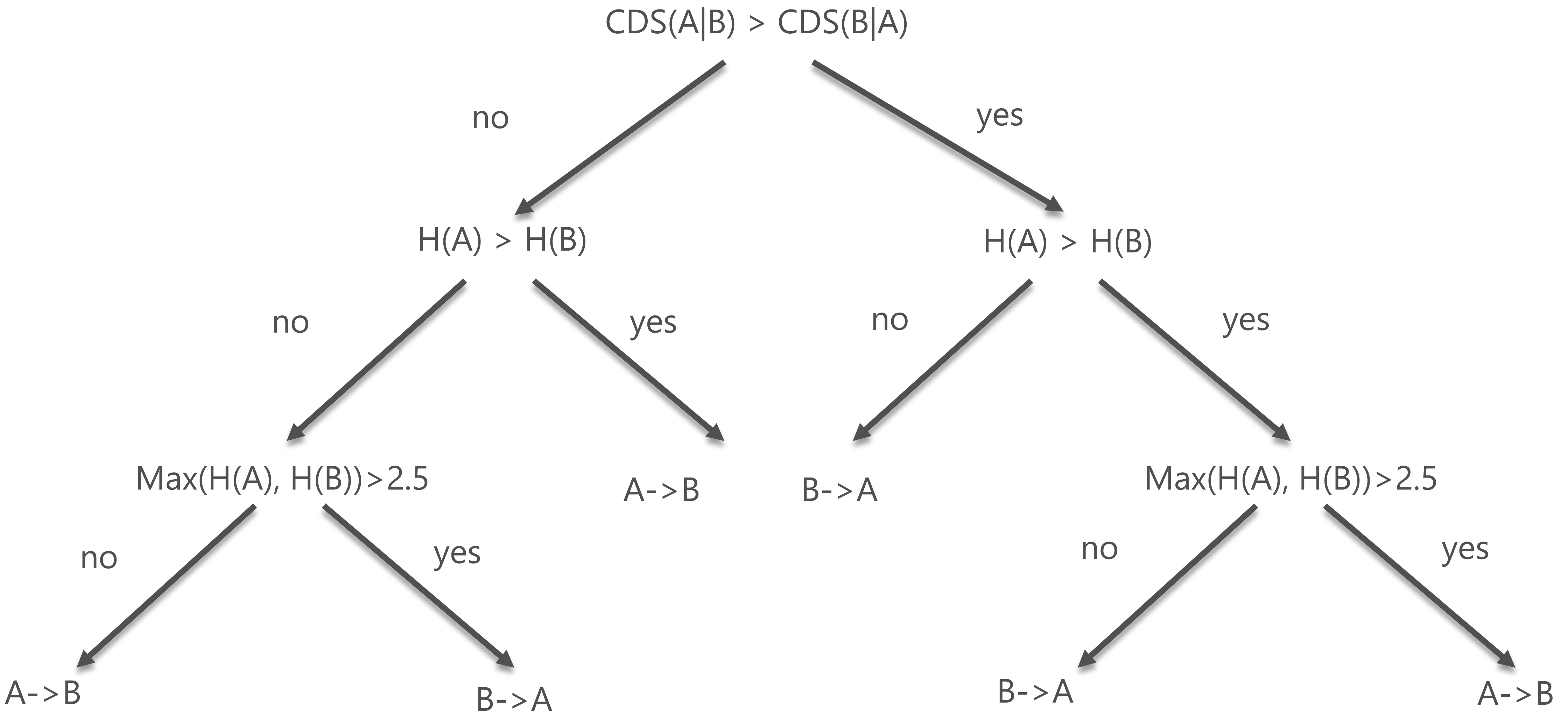


Information theoretic features: Entropy, KL divergence to Gaussian or Uniform distrib. conditional entropy, IGCI.



Data statistics: Num. data points, num. unique values, variable type.

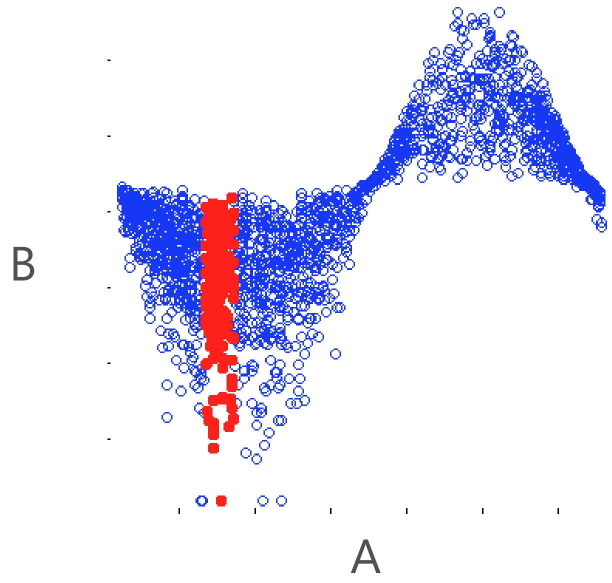
Informative features

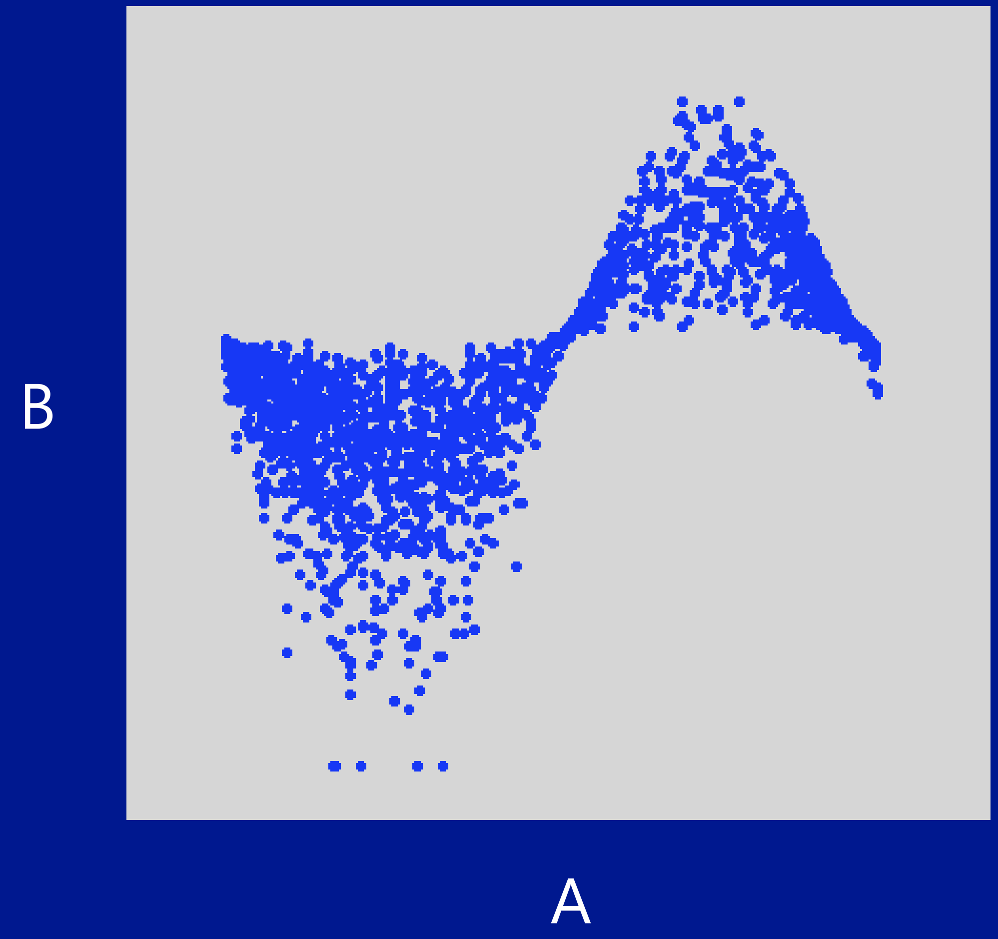


What is CDS?

CDS = Conditional Distribution Similarity

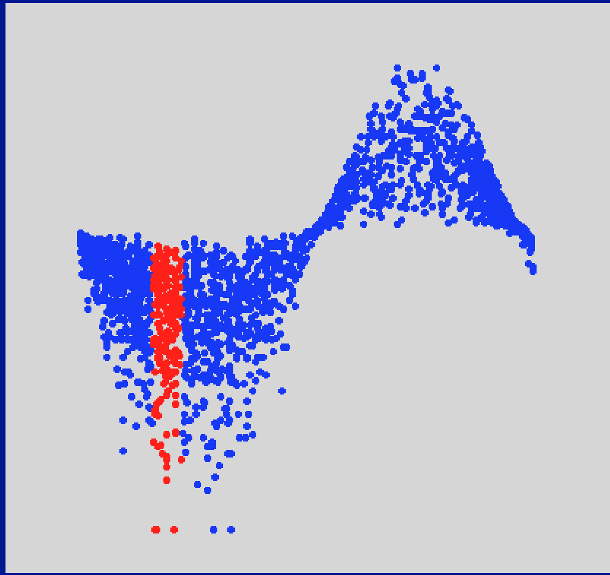
Low CDS means the "shape" of the "noise" distribution does not vary with the input.





A = Aspect

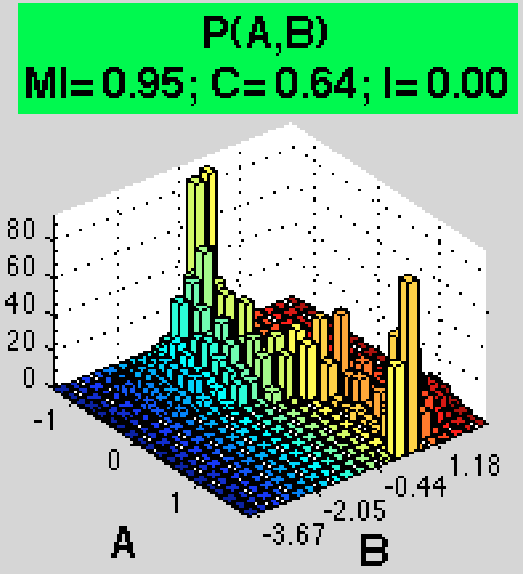
B = Hillshade 3pm

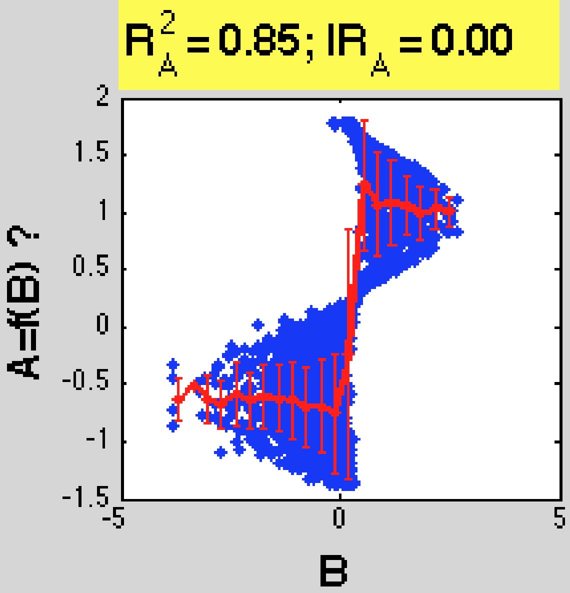


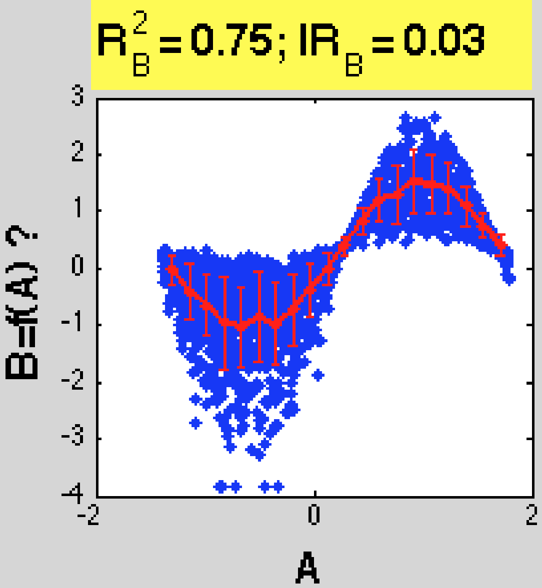
$$C(A,B) = \mathrm{cov}(A,B)/(\sigma_A \sigma_B)$$

$$MI(A,B) = H(A) + H(B) - H(A,B)$$
$$= KL[p(A,B) \,||\, p(A)p(B)]$$

$$I(A,B) = \mathrm{pval}(\,||C_{AB}||_{HS}^2\,)$$



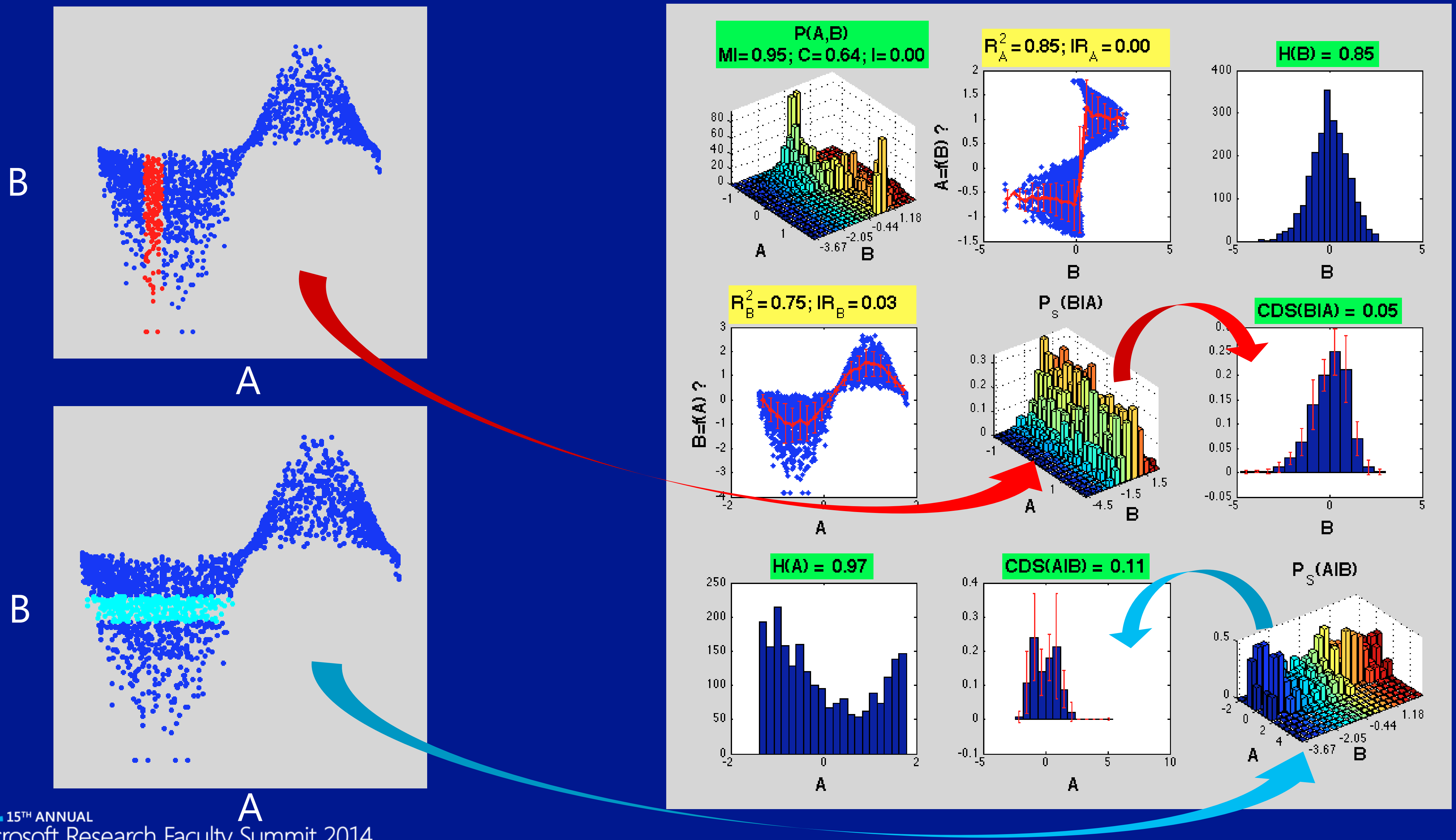




$$res(B)^2 = (1/n)\Sigma_i\,(f(A_i) - B_i)^2$$

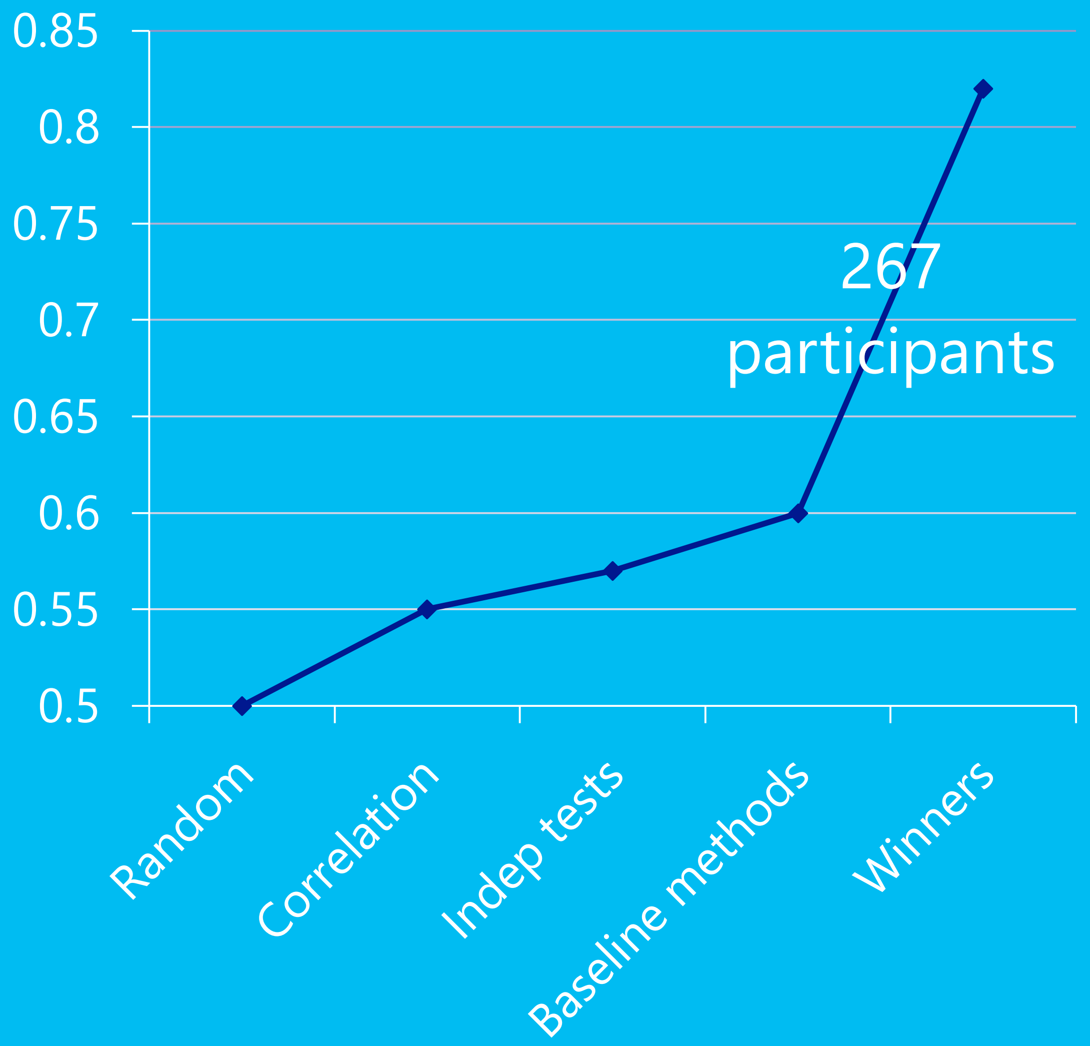
$$R_B^2 = 1 - res(B)^2 / \sigma_B^2$$

$$IR_B = I(A, res(B))$$



B
A
B
A
P(A,B)
MI=0.95; C=0.64; I=0.00
R^2_A=0.85; IR_A=0.00
A=f(B) ?
H(B) = 0.85
R^2_B=0.75; IR_B=0.03
B=f(A) ?
P_S(B|A)
CDS(B|A) = 0.05
H(A) = 0.97
CDS(A|B) = 0.11
P_S(A|B)

Conclusion:



Challenges can be a powerful tool to advance the state of the art, but…

most data scientists prefer “hacking” than producing papers.

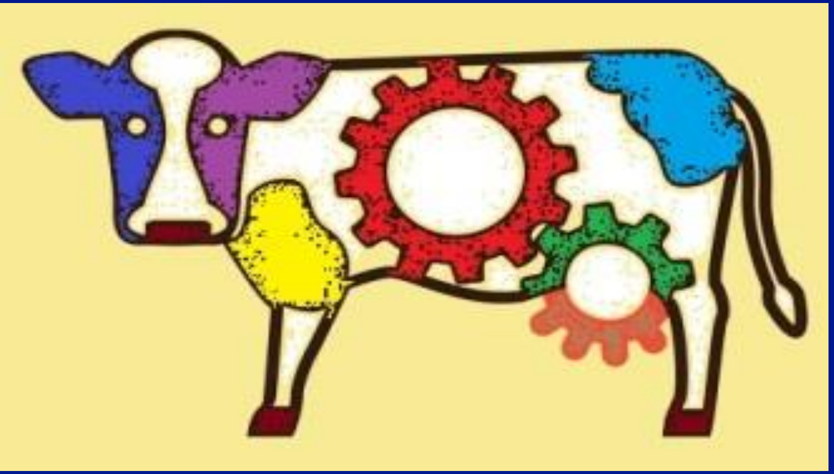
We are preparing a paper in which we will include:

- Analysis the features.
- Theoretical results.
- Test on other data.

Our next challenges will address

- causality in time series and
- entire network reconstruction.

What's next?



Google group: causalitychallenge

Save the planet and return your name badge before you leave (on Tuesday)

Microsoft