CS 295 Causal Reasoning Paper Presentation
[ACM] Towards Causal Machine Learning (Bernhard Scholkopf, 2019)

Sakshi Agarwal

Department of Computer Science

University of California-Irvine

2021

OUTLINE

- BACKGROUND
- MACHINE LEARNING FOR CAUSAL DISCOVERY
- CAUSALITY FOR MACHINE LEARNING

OUTLINE

- **BACKGROUND**
- MACHINE LEARNING FOR CAUSAL DISCOVERY
- CAUSALITY FOR MACHINE LEARNING

Chandoo.org
BECOME AWESOME IN EXCEL

Welcome | Excel Tips | Charting | Advanced Excel | VBA | Excel Dashboards | Project Mgmt. | Formulas | Downloads

Amazon's recommendation system – is it crazy?

Posted on January 12th, 2008 in business , Humor , technology , wonder why - 6 comments

We have a saying in *Telugu* that goes like this, "*thaadu vundhi kada ani eddu kontama?*" which means, "just because you have a rope you dont buy a bullock to tie". Amazon's recommendation system must have been coded by someone with a skewed view of reality. How else can you explain this?

"imitate the superficial exterior of a process or system without having any understanding of the underlying substance".
(source: http://philosophyisfashionable.blogspot.com/)

Thanks to P. Laskov.

Dependence vs. Causation

Storks Deliver Babies (p= 0.008)

Robert Matthews

Article first published online: 25 DEC 2001

DOI: 10.1111/1467-9639.00013

Issue

Teaching Statistics

Volume 22, Issue 2, 38, June 2000

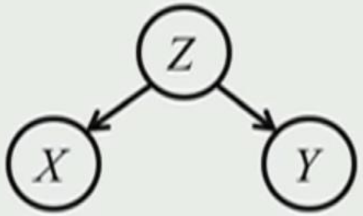
Country	Area (km^2)	Storks (pairs)	Humans (10^6)	Birth rate (10^3/yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15	188
Hungary	93,000	5000	11	124
Italy	301,280	5	57	551
Poland	312,680	30,000	mailto:rajm@compuserve.com	
Portugal	92,390	1500	10	120
Romania	237,500	5000	23	367
Spain	504,750	8000	39	439
Switzerland	41,290	150	6.7	82
Turkey	779,450	25,000	56	1576

Table 1. Geographic, human and stork data for 17 European countries

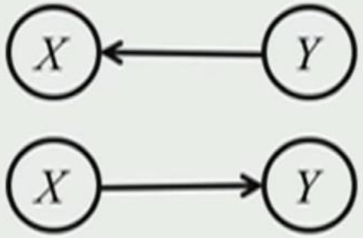
Statistical Implications of Causality

Reichenbach's
Common Cause Principle
links **causality** and **probability:**

(i) if X and Y are statistically dependent, then there is a Z causally influencing both;

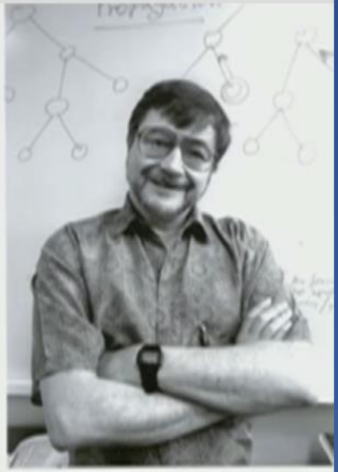


special cases:

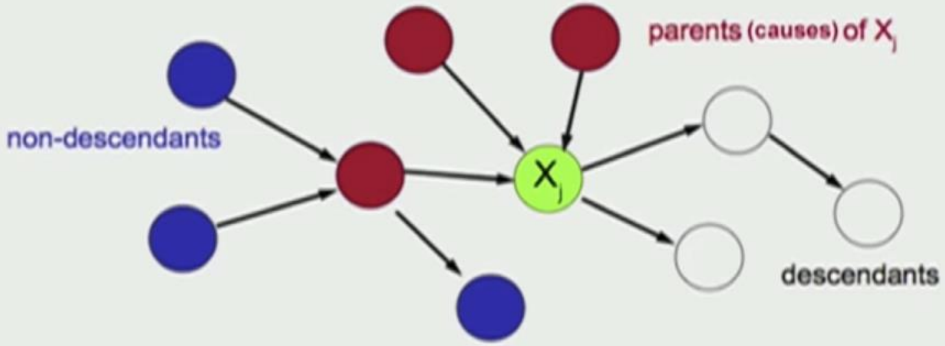


(ii) Z screens X and Y from each other (given Z, the observables X and Y become independent)

Functional Causal Model *(Pearl et al.)*



- Set of observables $X_1, \ldots, X_n$ on a directed acyclic graph (DAG) G
- arrows represent direct causal links
- $X_i := f_i(\mathrm{PA}_i, U_i)$ with independent RVs $U_1, \ldots, U_n$.



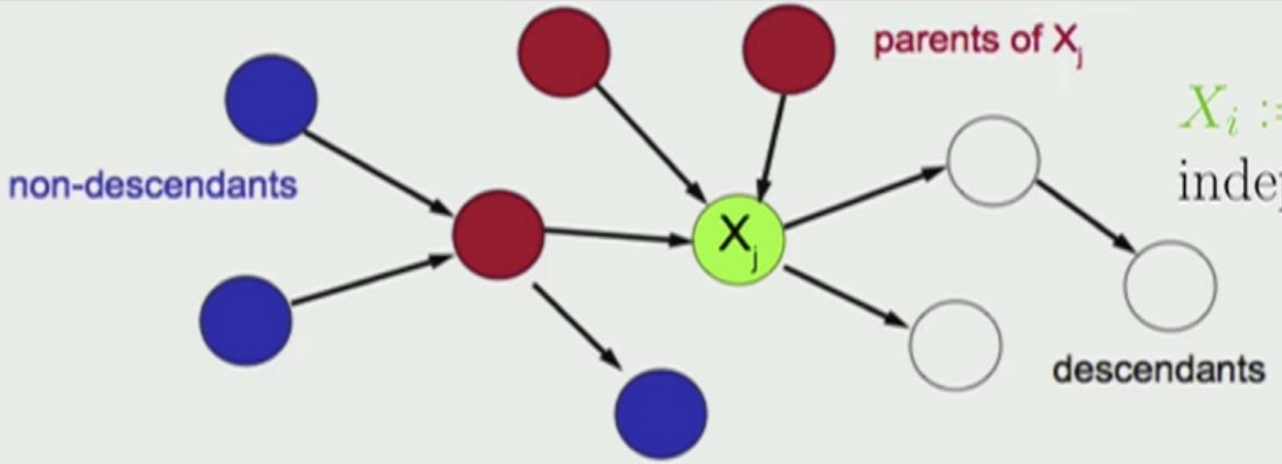
- entails $p(X_1, \ldots, X_n)$ with particular conditional independences, in particular the *causal Markov condition*:

$$\boxed{X_j \text{ independent of non-descendants, given parents}}$$

- this is a directed "graphical model"

OUTLINE

- BACKGROUND
- **MACHINE LEARNING FOR CAUSAL DISCOVERY**
- CAUSALITY FOR MACHINE LEARNING



$X_i := f_i(\mathrm{PA}_i, U_i)$ with independent RVs $U_1, \ldots, U_n$.

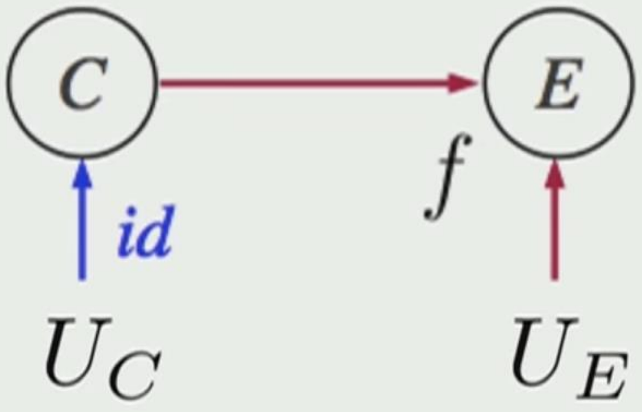
Can we recover G from p?

approach	assumptions	method	intuition
graphical approach (constraint-based methods) *(Pearl, Spirtes, Glymour, Scheines)*	noises jointly independent; faithfulness	conditional independence testing ($n \geq 3$)	track how the noises spread
independent mechanisms *(Daniušis et al., UAI 2010; Shajarisales et al., ICML 2015)*	e.g.: noises and f_i independent; f_i learnable	customized tests	noises pick up footprints of the functions
additive noise model *(Peters, Mooij, Janzing, Schölkopf, JMLR 2014)*	$X_i = f_i(\mathrm{PA}_i) + U_i$ with learnable f_i	regression & unconditional independence testing	restriction of function class

Independence of cause and mechanism

Causal structure:

C cause
E effect
U noise
f mechanism

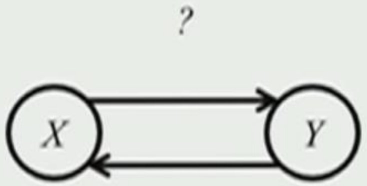


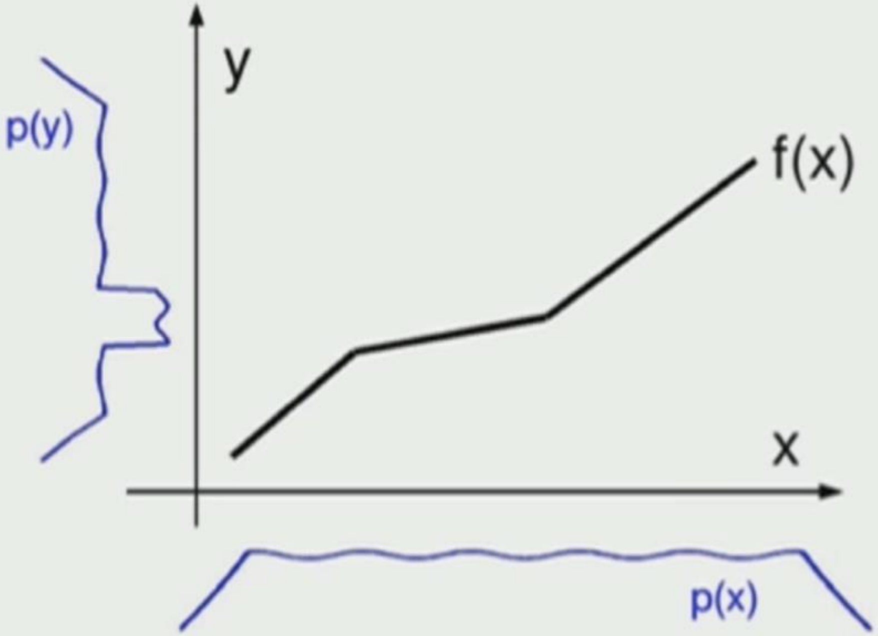
Assumption:
$p(C)$ and $p(E|C)$ are "independent"

Janzing & Schölkopf, IEEE Trans. Inf. Theory, 2010; cf. also Lemeire & Dirkx, 2007

Independence of input and mechanism, III

- No added noise
- Assumption: $y = f(x)$ with invertible f





Daniusis, Janzing, Mooij, Zscheischler, Steudel, Zhang, Schölkopf:
Inferring deterministic causal relations, *UAI* 2010

Causal independence implies anticausal dependence

Assume that f is a monotonically increasing bijection of $[0,1]$.
View p_x and $\log f'$ as RVs on the prob. space $[0,1]$ w. Lebesgue measure.

Postulate (independence of mechanism and input):

$$\mathrm{Cov}\left(\log f', p_x\right) = 0$$

Note: this is equivalent to

$$\int_0^1 \log f'(x) p(x) dx = \int_0^1 \log f'(x) dx,$$

since $\mathrm{Cov}\left(\log f', p_x\right) = E\left[\log f' \cdot p_x\right] - E\left[\log f'\right] E\left[p_x\right] = E\left[\log f' \cdot p_x\right] - E\left[\log f'\right]$.

Proposition: If $f \neq Id$,

$$\mathrm{Cov}\left(\log f^{-1'}, p_y\right) > 0.$$

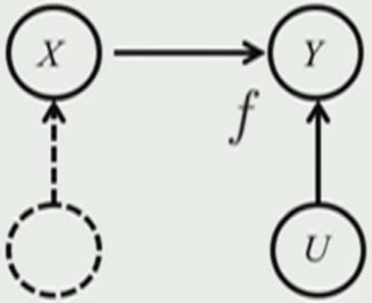
Information Geometric Causal Method (IGCI)

Causal Inference method (IGCI): *Given* $C_{X\to Y}$, *infer that* X *causes* Y *if* $C_{X\to Y} < 0$, *or that* Y *causes* X *if* $C_{X\to Y} > 0$.

$$C_{X\to Y} := D(p_X \,||\, \mathcal{E}_X) - D(p_Y \,||\, \mathcal{E}_Y) \leq 0\,.$$

$$C_{Y\to X} := D(p_Y \,||\, \mathcal{E}_Y) - D(p_X \,||\, \mathcal{E}_X) \leq 0\,.$$

Restricting the Functional Model



$X \perp\!\!\!\perp U$

- consider the graph $X \to Y$
- general functional model

$$Y = f(X, U)$$

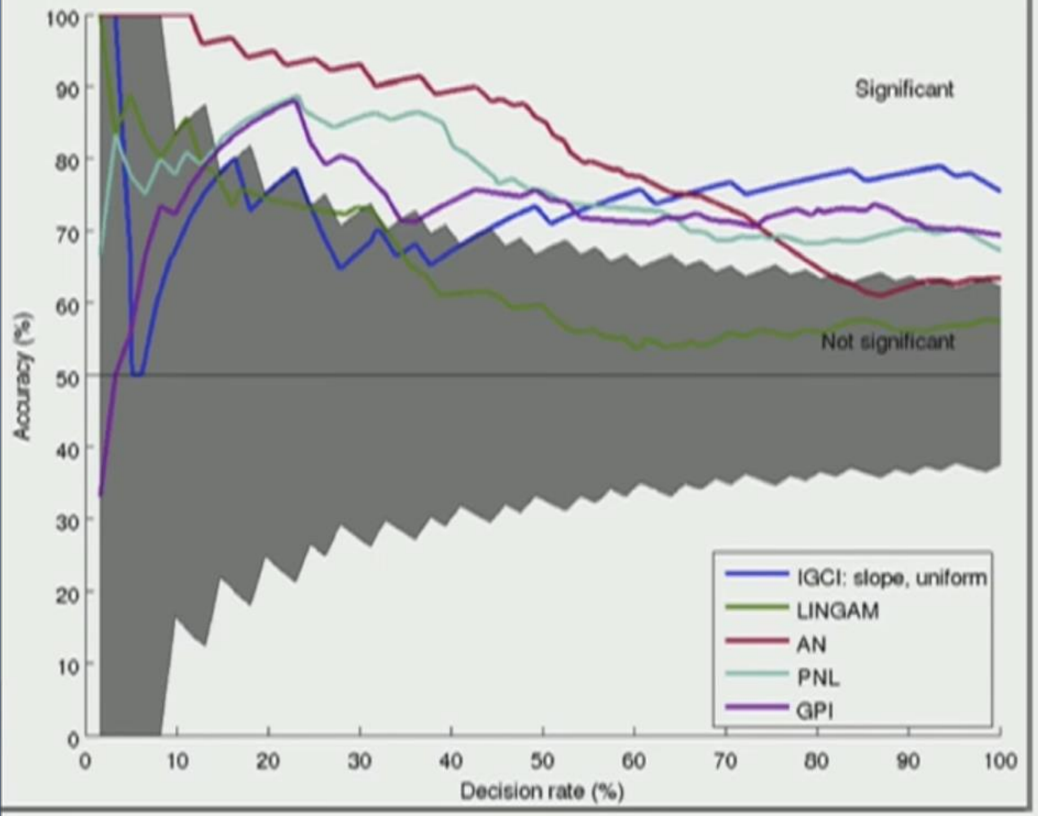
Note: if U can take d different values, it could switch randomly between mechanisms $f^1(X), \ldots, f^d(X)$

- additive noise model

$$Y = f(X) + U$$

80 Cause-Effect Pairs − Examples

	var 1	var 2	dataset	ground truth
pair0001	Altitude	Temperature	DWD	→
pair0005	Age (Rings)	Length	Abalone	→
pair0012	Age	Wage per hour	census income	→
pair0025	cement	compressive strength	concrete_data	→
pair0033	daily alcohol consumption	mcv mean corpuscular volume	liver disorders	→
pair0040	Age	diastolic blood pressure	pima indian	→
pair0042	day	temperature	B. Janzing	→
pair0047	#cars/24h	specific days	traffic	←
pair0064	drinking water access	infant mortality rate	UNdata	→
pair0068	bytes sent	open http connections	P. Daniusis	←
pair0069	inside room temperature	outside temperature	J. M. Mooij	←
pair0070	parameter	sex	Bülthoff	→
pair0072	sunspot area	global mean temperature	sunspot data	→
pair0074	GNI per capita	life expectancy at birth	UNdata	→
pair0078	PPFD (Photosynth. Photon Flux)	NEP (Net Ecosystem Productivity)	Moffat A. M.	→



Percentage of pairs for which the decision was made

IGCI: Information Geometric Method

AN: Additive Noise Model (nonlinear)

LINGAM: Shimizu et al., 2006

PNL: AN with post-nonlinearity

GPI: Mooij et al., 2010

Used the same methods to classify the direction of time:

time series *(Peters et al., ICML 2009)*

videos *(Pickup et al., CVPR 2014)*

OUTLINE

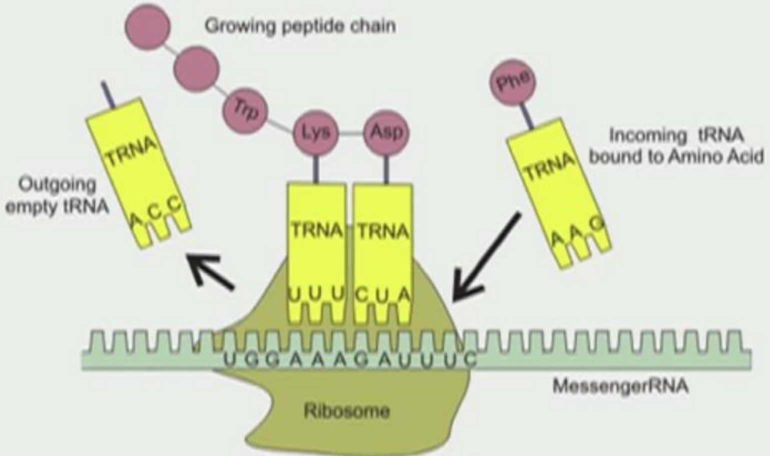
- BACKGROUND
- MACHINE LEARNING FOR CAUSAL DISCOVERY
- **CAUSALITY FOR MACHINE LEARNING**

How can causal knowledge help machine learning?

- ICML 2012: semi-supervised learning and changing distributions
- ICML 2015: modeling systematic errors for exoplanet detection

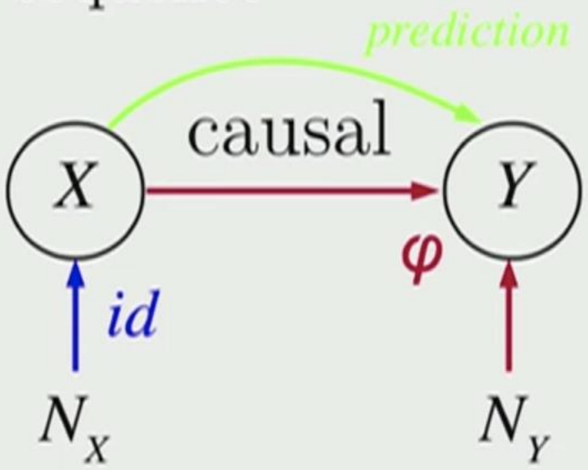
Using cause-effect knowledge

- example 1: predict protein from mRNA sequence



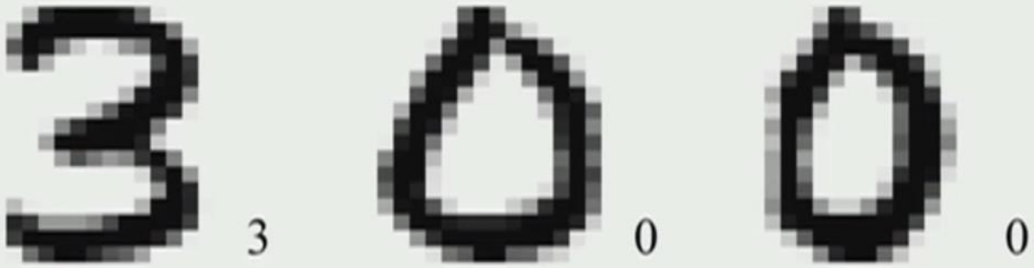
Peptide Synthesis

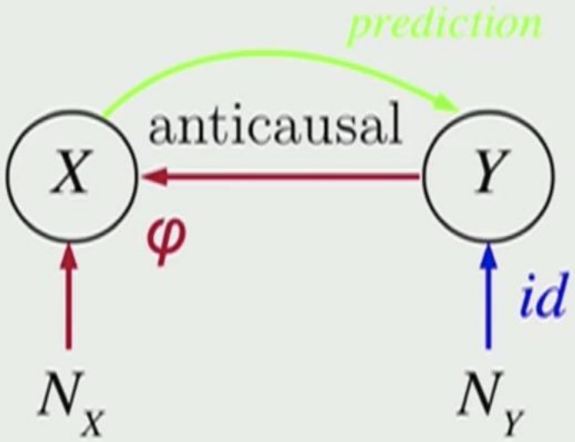
Source: http://commons.wikimedia.org/wiki/File:Peptide_syn.png



causal mechanism φ

- example 2: predict class membership from handwritten digit





Covariate Shift and Semi-Supervised Learning

Goal: learn $X \mapsto Y$, i.e., estimate (properties of) $p(Y|X)$

Semi-supervised learning: improve estimate by more data from $p(X)$

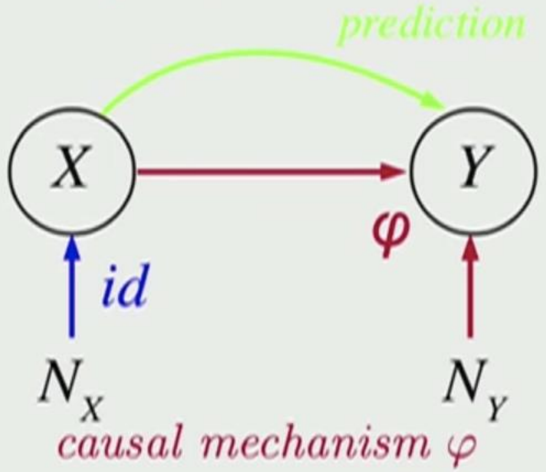
Covariate shift: $p(X)$ changes between training and test

Causal assumption: $p(C)$ and mechanism $p(E|C)$ "independent"

Causal learning

$p(X)$ and $p(Y|X)$ independent

1. *semi-supervised learning impossible*
2. *$p(Y|X)$ invariant under change in $p(X)$*

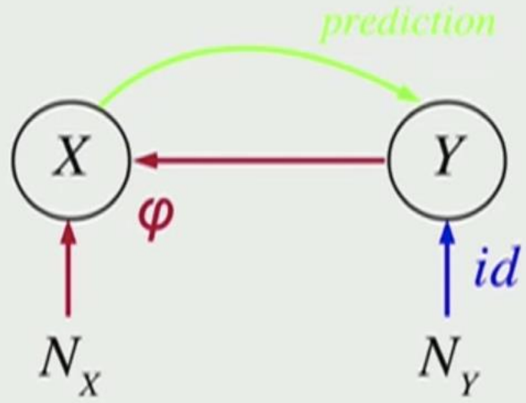


Anticausal learning

$p(Y)$ and $p(X|Y)$ independent

hence $p(X)$ and $p(Y|X)$ dependent

1. *semi-supervised learning possible*
2. *$p(Y|X)$ changes with $p(X)$*



Schölkopf, Janzing, Peters, Sgouritsa, Zhang, Mooij, 2012, cf. Storkey, 2009; Bareinboim & Pearl, 2012

Compares 11 SSL methods to the base classifiers 1-NN and SVM.

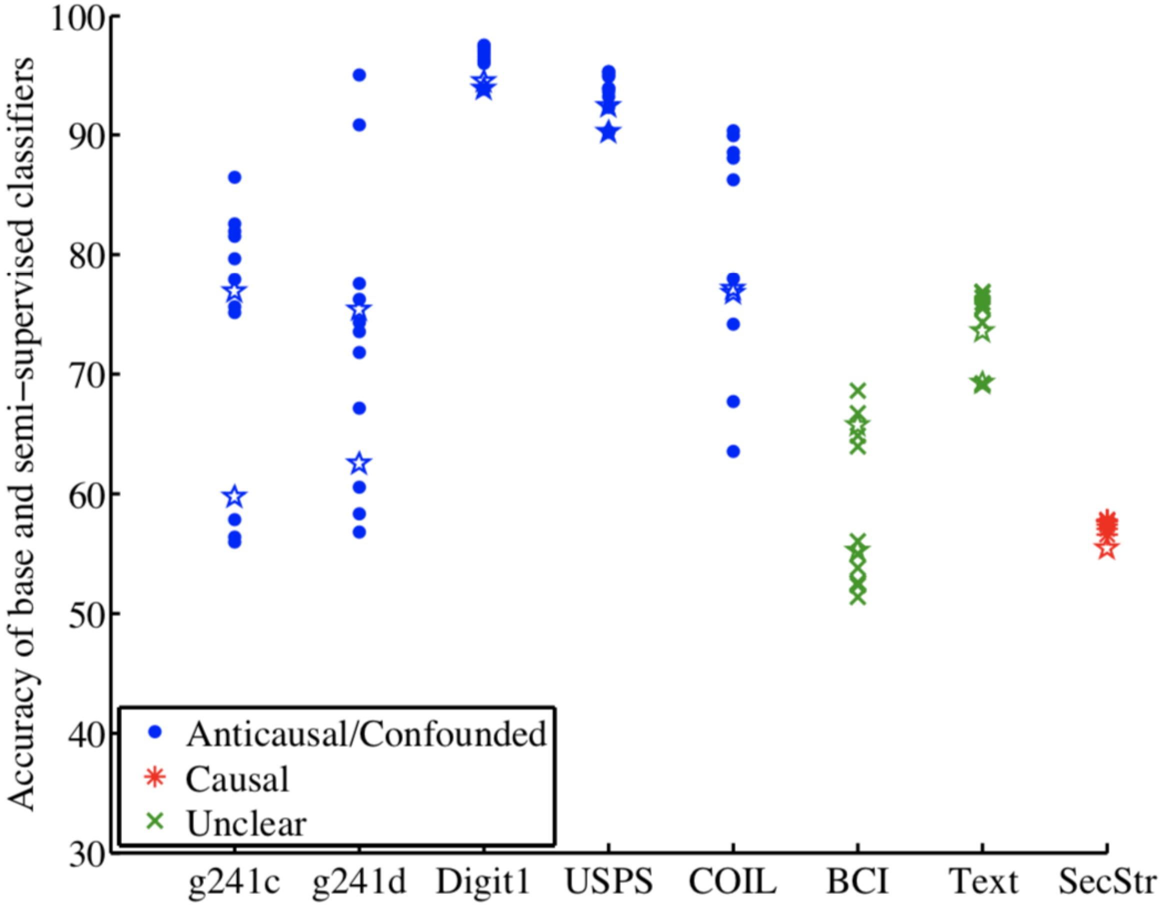


Figure 5. Accuracy of base classifiers (star shape) and different SSL methods on eight benchmark datasets.

Difficult to draw conclusions from this small dataset

Extend supervise learning to self-training with unlabeled data

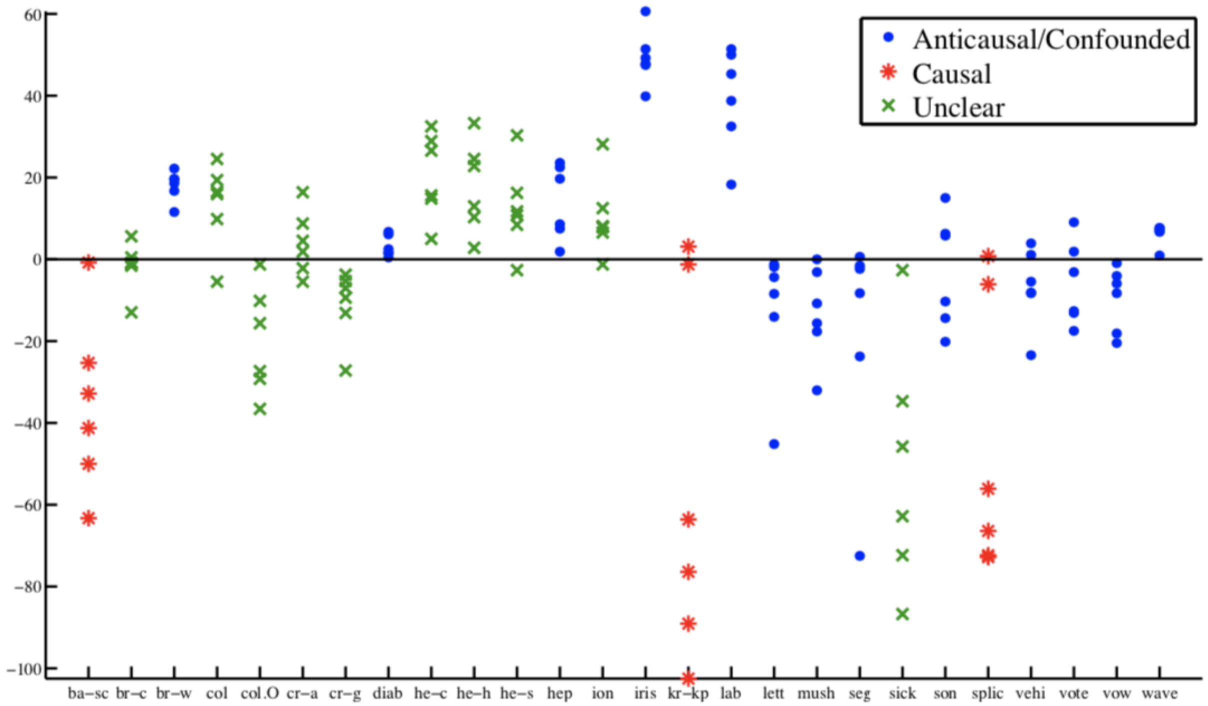
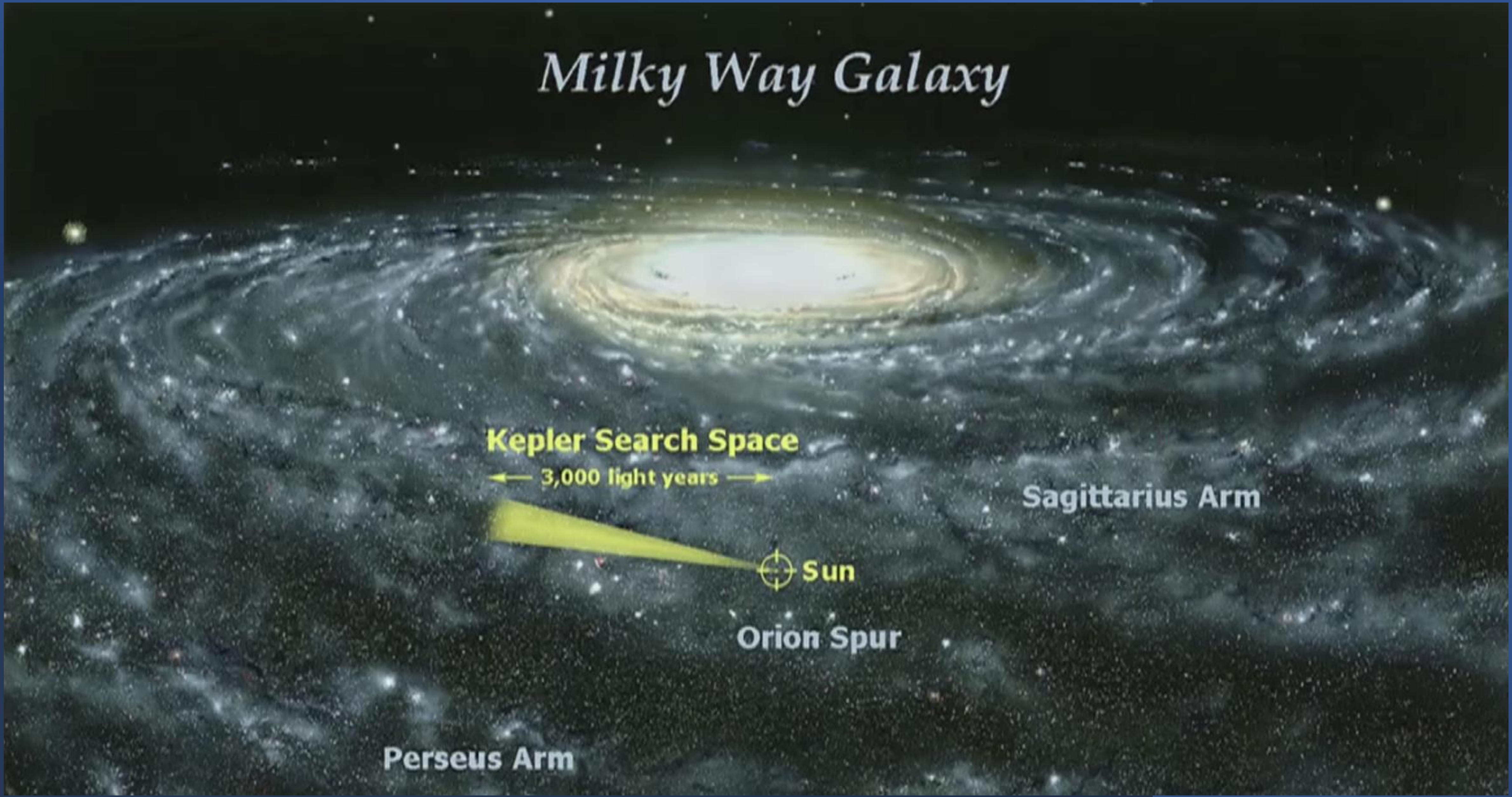


Figure 6. Plot of the relative decrease of error when using self-training, for six base classifiers on 26 UCI datasets. Here, relative decrease is defined as (error(base) − error(self-train)) / error(base). Self-training, a method for SSL, overall does not help for the causal datasets, but it does help for several of the anti-causal/confounded datasets.

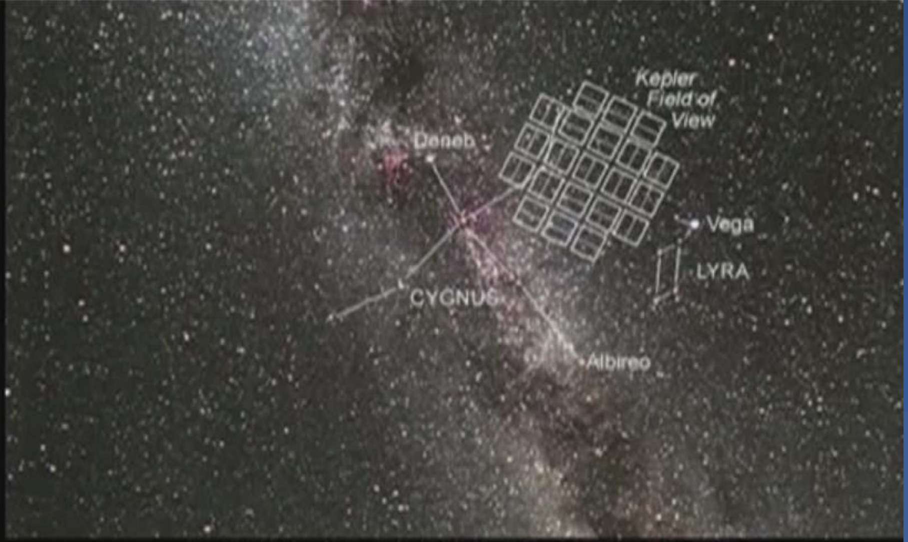

Milky Way Galaxy
Kepler Search Space
3,000 light years
Sagittarius Arm
Sun
Orion Spur
Perseus Arm

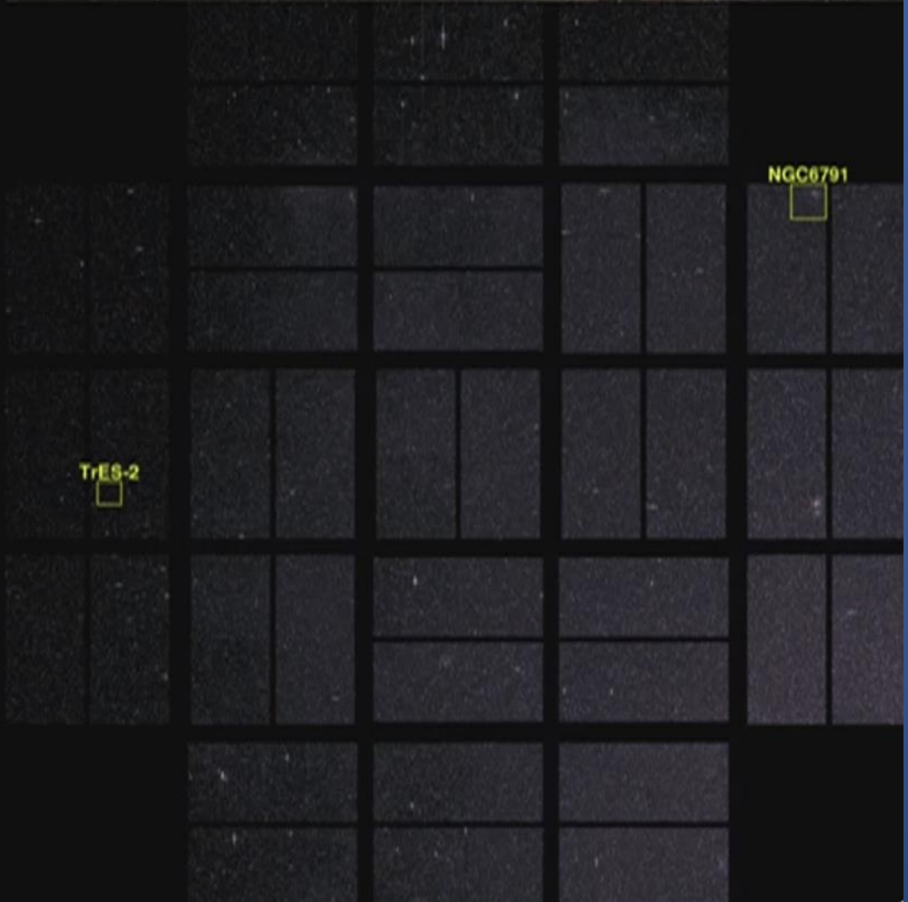
Kepler Spacecraft

4 years

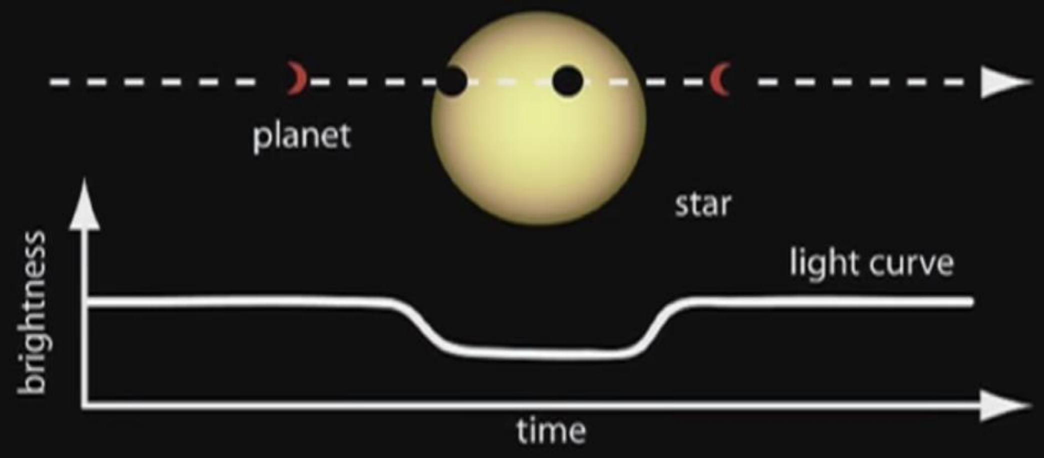
150000 stars

series of 30 min exposures





Exoplanet Transits

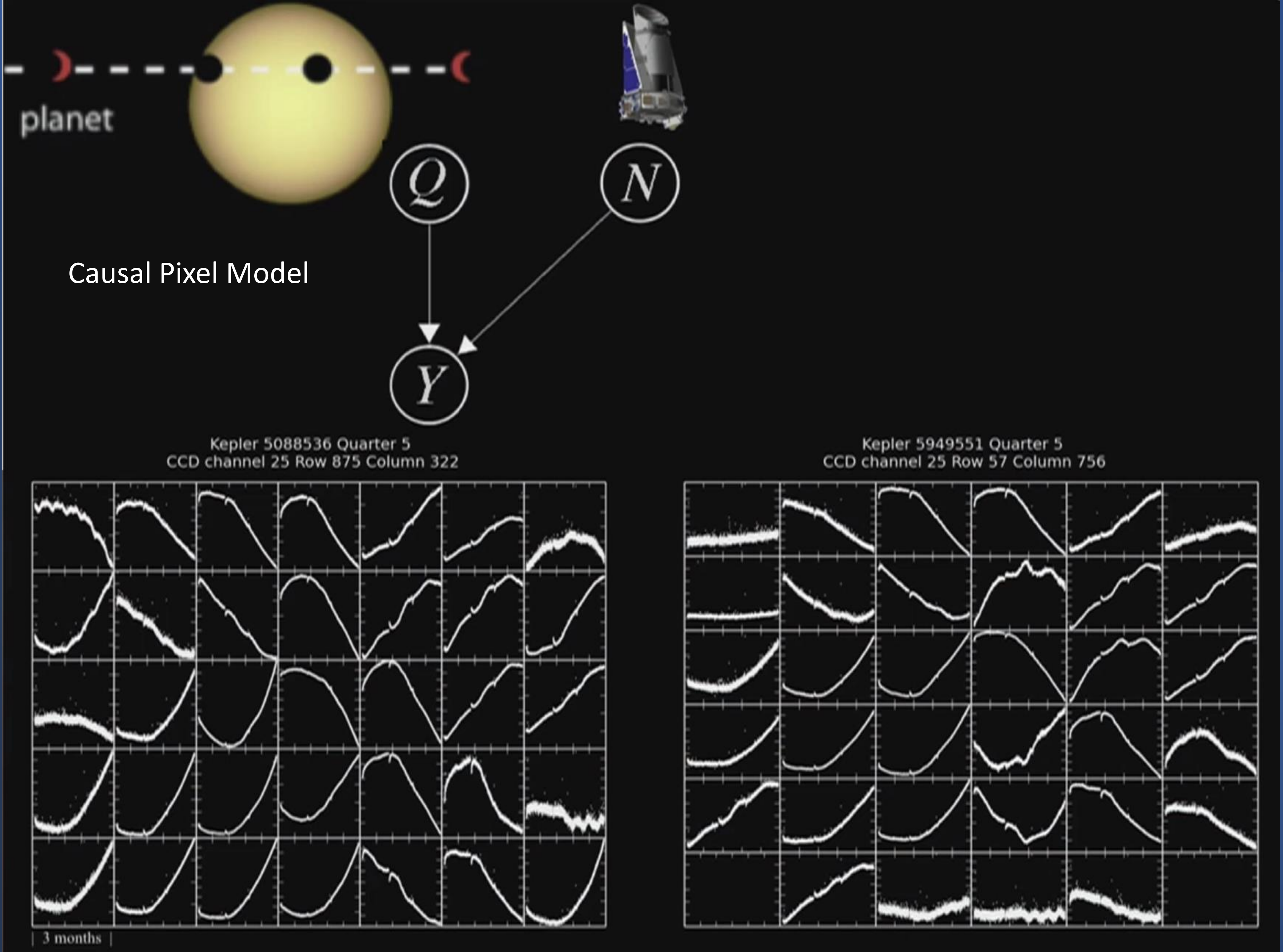


earth: annual 84ppm signal for ½ day, visible from 0.5% of all directions

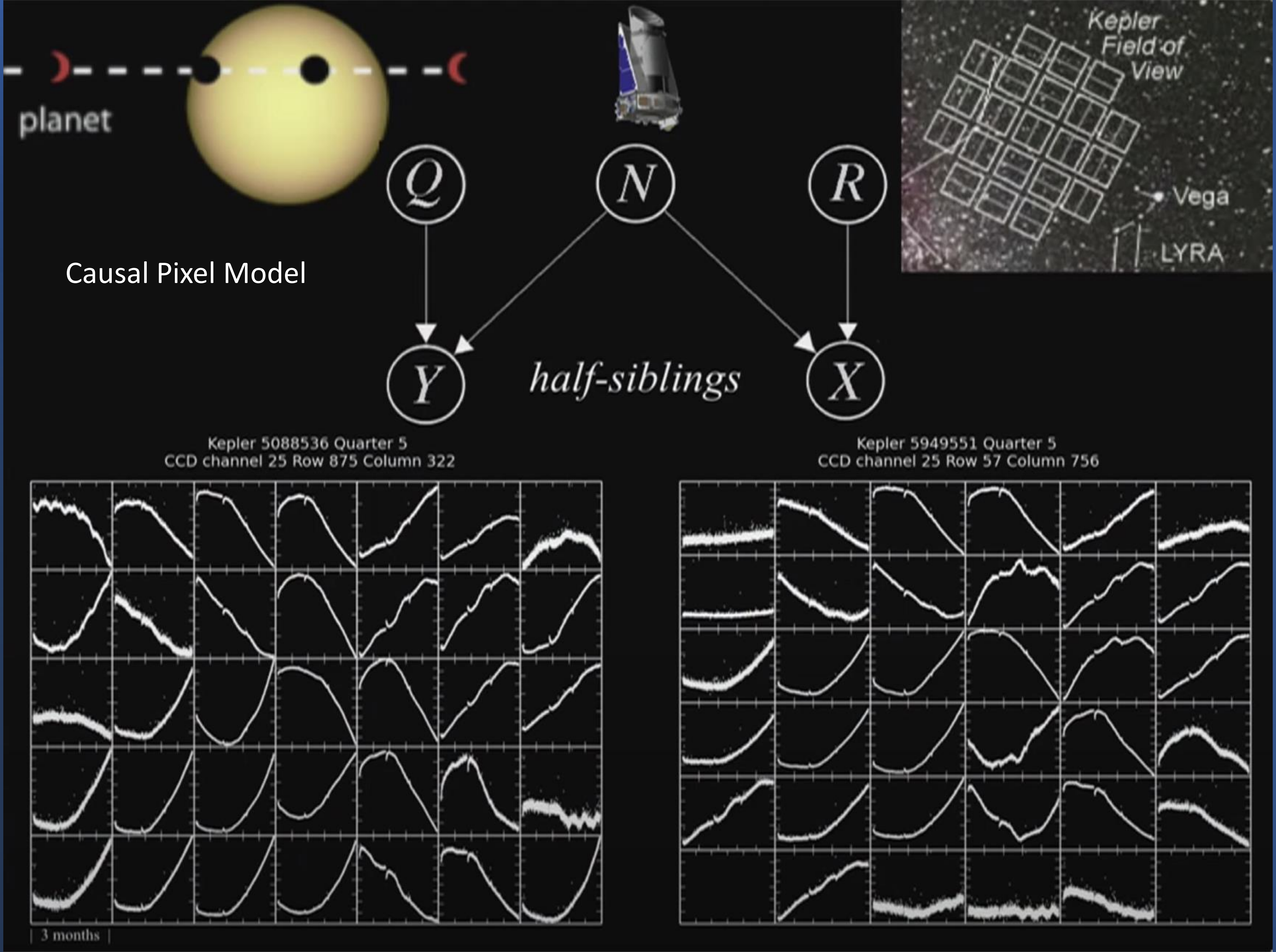
many planets found, but nothing quite like earth/sun

both spacecraft and stars vary, leading to changes that are sometimes much bigger than the signal

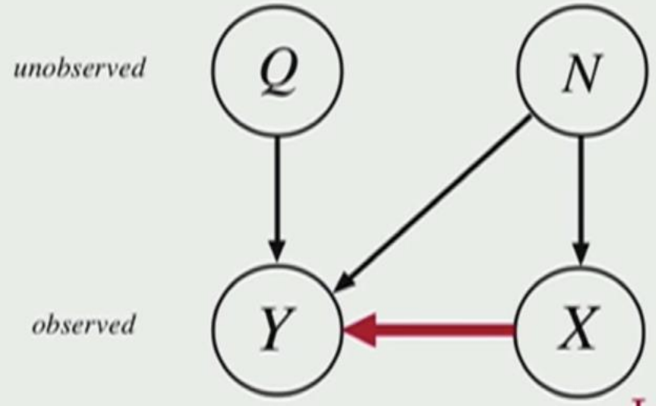
Causal Pixel Model



Causal Pixel Model



Half-Sibling Regression



Idea: remove $E[Y|X]$ from Y to reconstruct Q.

$X \perp\!\!\!\perp Q$

X and Y share information (only) through N

If we try to predict Y from X, we only pick up the part due to N

with David Hogg, Dan Foreman-Mackey, Dun Wang, Dominik Janzing, Jonas Peters, Carl-Johann Simon-Gabriel *(ICML 2015)*

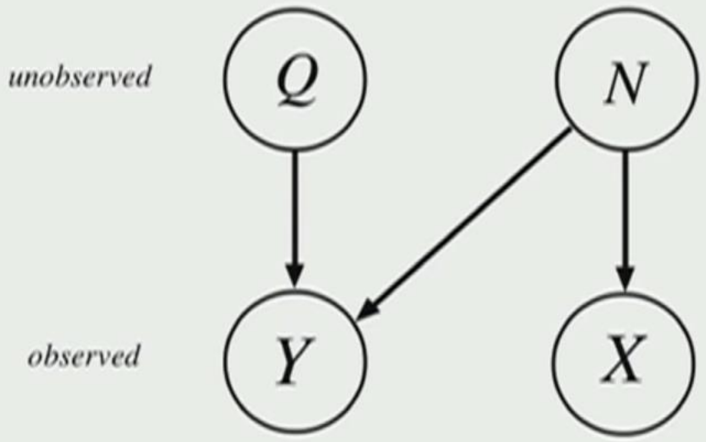
Proposition. Q, N, Y, X random variables, $X \perp\!\!\!\perp Q$, and f measurable. Suppose

- $Y = Q + f(N)$ *(additive noise model)*
- $f(N) = \psi(X)$ for some ψ *(complete information).*

Then $\hat{Q} := Y - \mathbb{E}[Y|X] = Q - \mathbb{E}[Q]$.

Device can be self-calibrated based on measured data only.

Q can be reconstructed, up to a constant offset, from Y and $\mathbb{E}[Y|X]$.

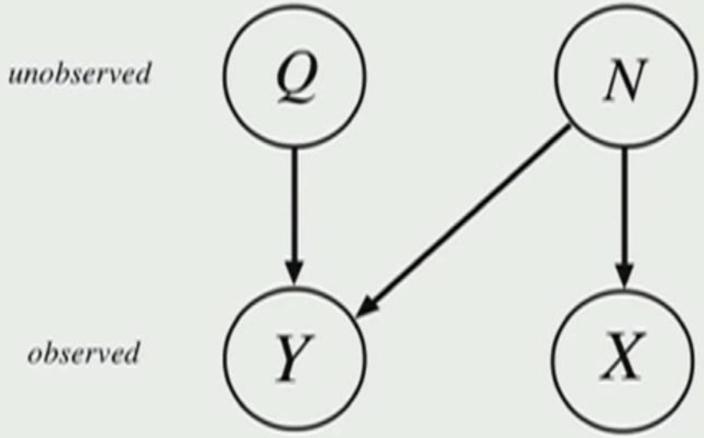


Proposition. Q, N, Y, X random variables, $X \perp\!\!\!\perp Q$, and f measurable. Suppose

- $Y = Q + f(N)$ *(additive noise model)*

Then $E[(\hat{Q} - (Q - E[Q]))^2] = E[\mathrm{Var}[f(N)|X]]$.

If $f(N)$ can (in principle) be predicted well from X, then Q can be reconstructed well.

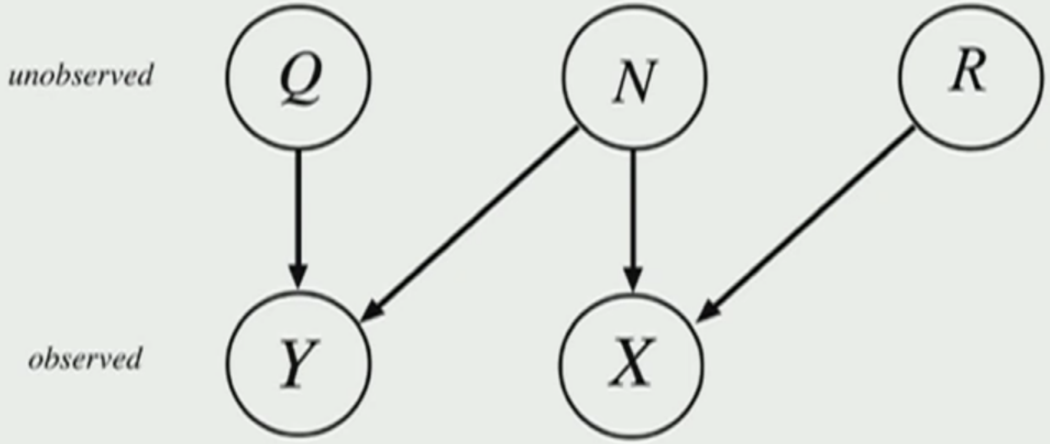


Proposition. R, N, Q jointly independent.
Suppose

$$X = g(N) + R$$

Recovery results if either

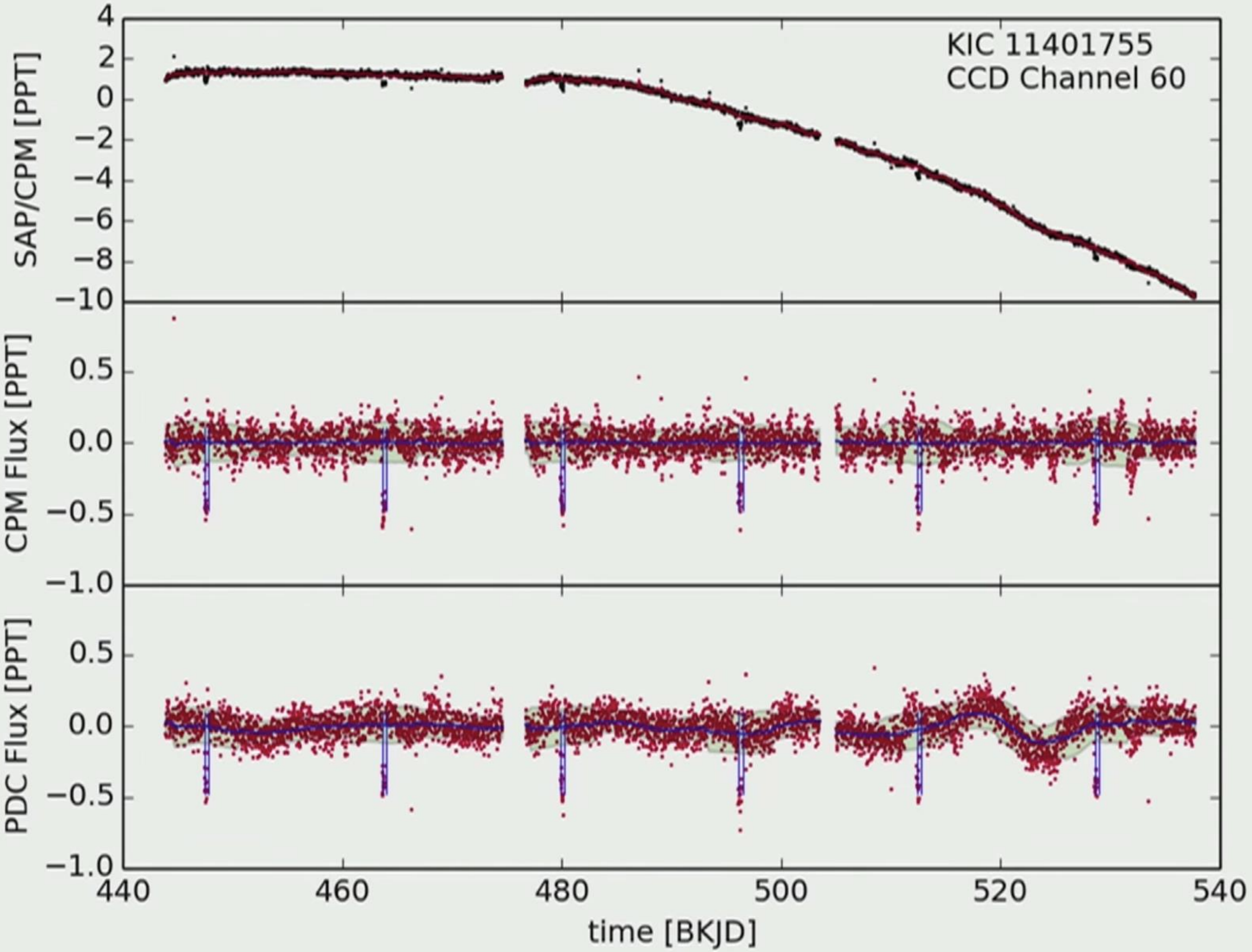
(i) magnitude of R goes to 0 (i.e., influence of stars negligible), or
(ii) R is a random vector whose components are jointly independent (i.e., many independent stars).



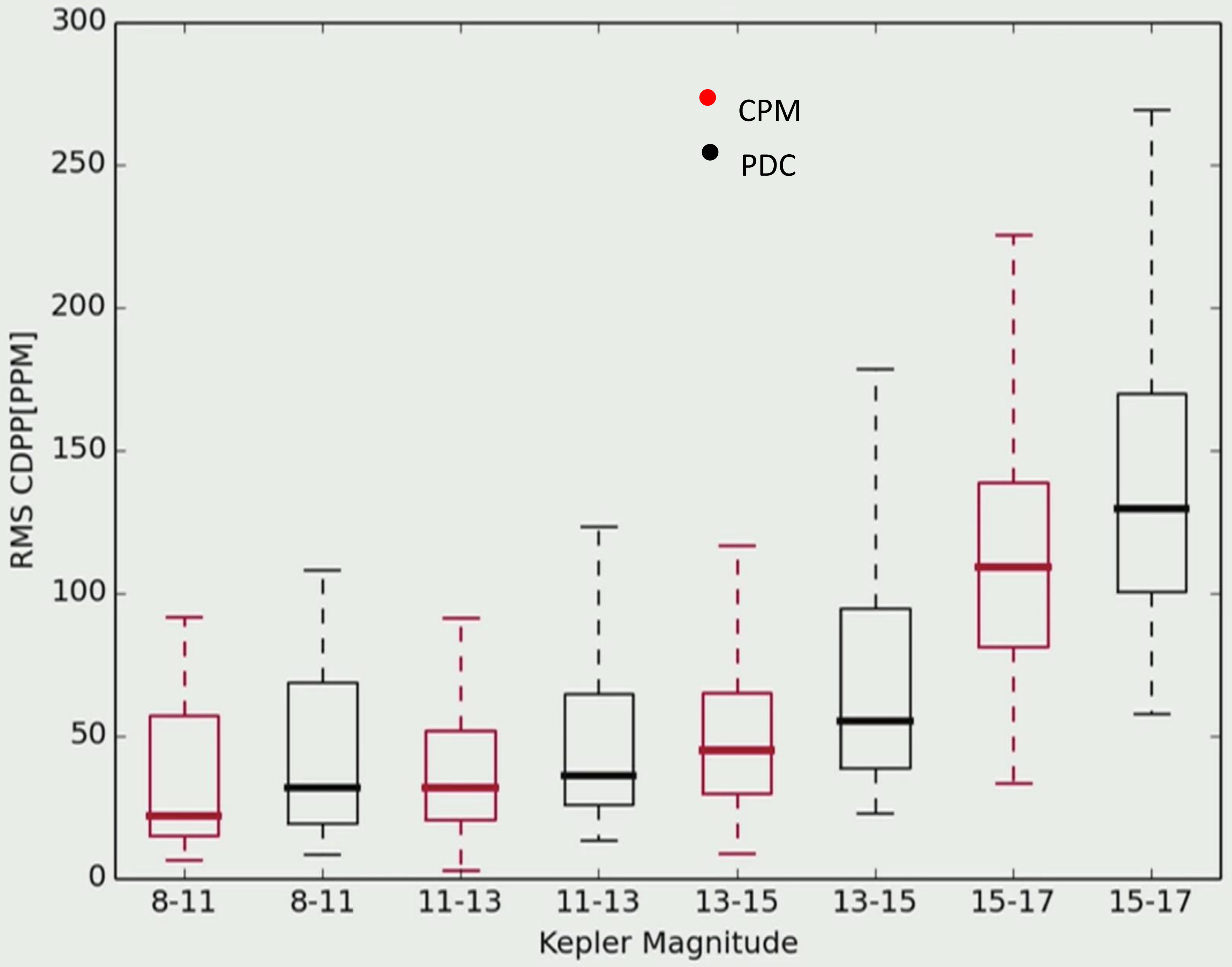
CPM : Causal Pixel Model

PDC : Pre-search Data Conditioning

SAP : Simple Aperture Photometry



CDPP: Combined Differential Photometric Precision (CDPP) – indicates the noise level seen by a transit signal in a given duration.



Conclusion

- Knowing Causal Structure might be helpful in some ML tasks

- can disregard causal structure for some applications

Thank You!