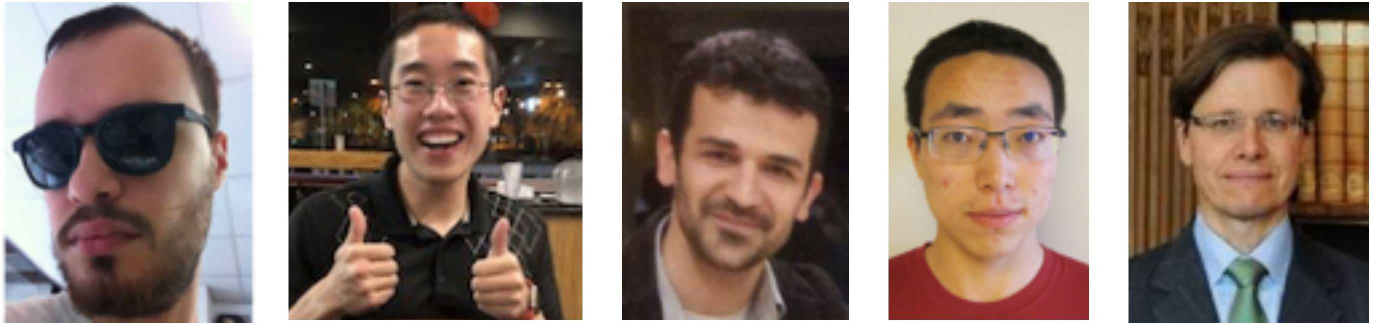
XYZ studies

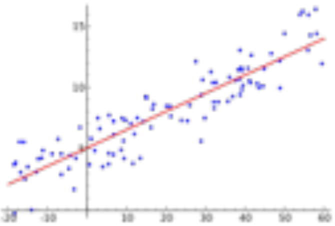
a paradigm for research in data science

Vardan Papyan

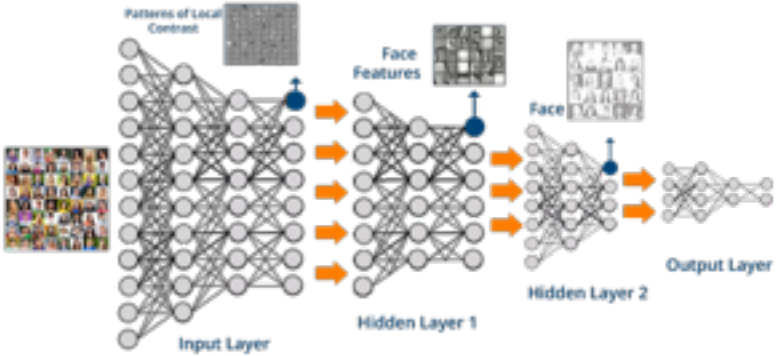


Nobody knows what data science is

Statistics:



Machine learning:



We are proposing to show you what data science is…

XYZ studies

X — all relevant methods

Y — datasets considered canonical for certain task

Z — control parameters

W — observables of interest

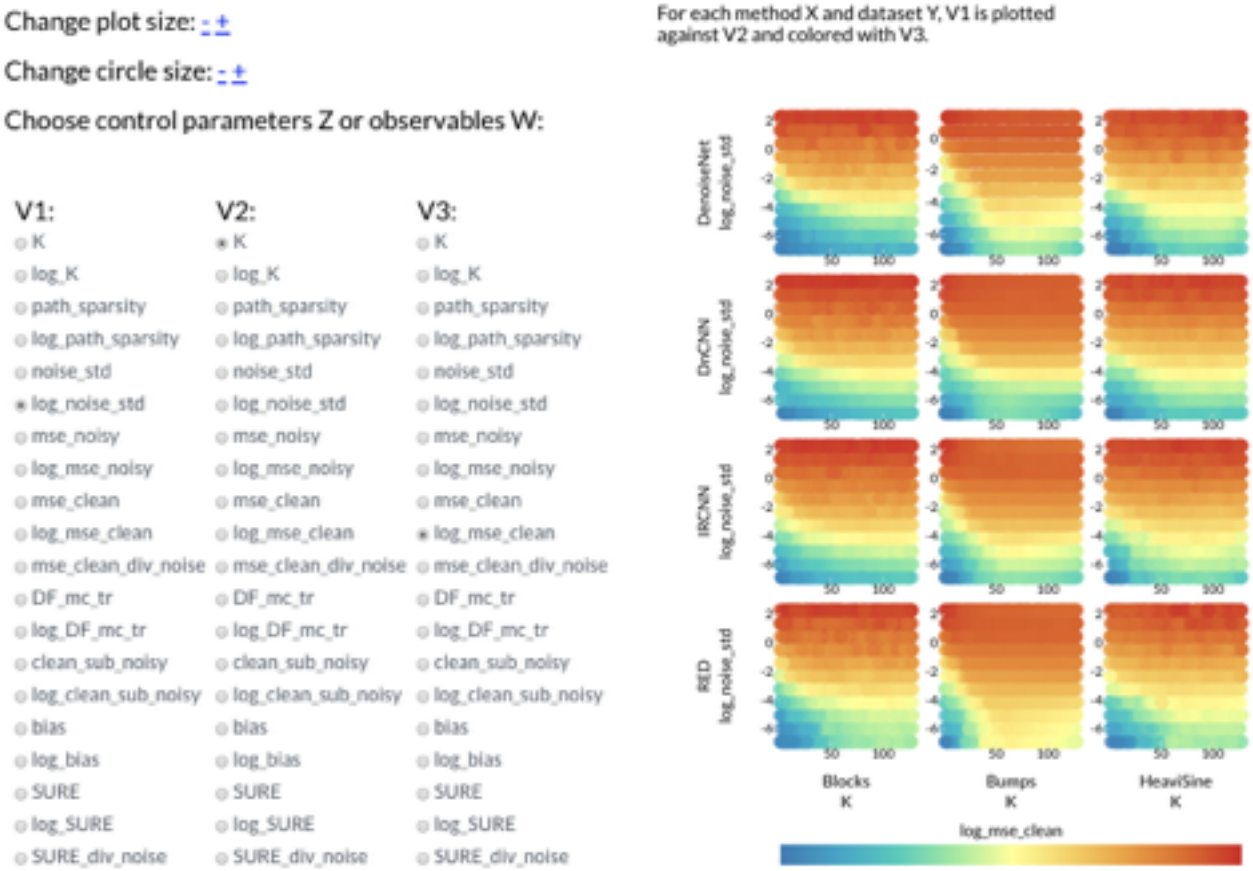
Algorithm 1: Description of XYZ experiment

```
Input  : methods X, datasets Y, control parameters Z
Output: observables W

foreach method $x \in X$ do
    foreach dataset $y \in Y$ do
        foreach control parameter $z \in Z$ do
            /* run experiment and collect observables      */
            $W(x, y, z) = \text{Experiment}(x, y, z)$
        end
    end
end
```

Finding

Navigating the space of finding



Change plot size: - +

Change circle size: - +

Choose control parameters Z or observables W:

V1:	V2:	V3:
○ K	● K	○ K
○ log_K	○ log_K	○ log_K
○ path_sparsity	○ path_sparsity	○ path_sparsity
○ log_path_sparsity	○ log_path_sparsity	○ log_path_sparsity
○ noise_std	○ noise_std	○ noise_std
● log_noise_std	○ log_noise_std	○ log_noise_std
○ mse_noisy	○ mse_noisy	○ mse_noisy
○ log_mse_noisy	○ log_mse_noisy	○ log_mse_noisy
○ mse_clean	○ mse_clean	○ mse_clean
○ log_mse_clean	○ log_mse_clean	● log_mse_clean
○ mse_clean_div_noise	○ mse_clean_div_noise	○ mse_clean_div_noise
○ DF_mc_tr	○ DF_mc_tr	○ DF_mc_tr
○ log_DF_mc_tr	○ log_DF_mc_tr	○ log_DF_mc_tr
○ clean_sub_noisy	○ clean_sub_noisy	○ clean_sub_noisy
○ log_clean_sub_noisy	○ log_clean_sub_noisy	○ log_clean_sub_noisy
○ bias	○ bias	○ bias
○ log_bias	○ log_bias	○ log_bias
○ SURE	○ SURE	○ SURE
○ log_SURE	○ log_SURE	○ log_SURE
○ SURE_div_noise	○ SURE_div_noise	○ SURE_div_noise

For each method X and dataset Y, V1 is plotted against V2 and colored with V3.

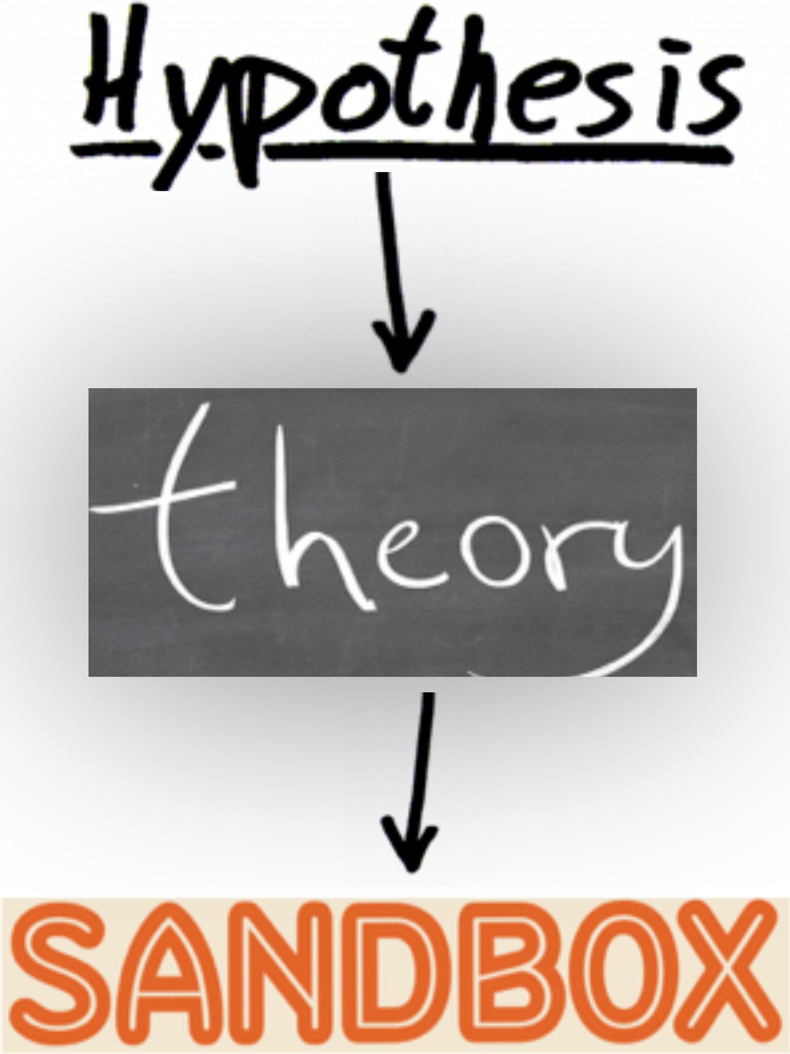
to pdf

download reproducible code

download models

download xyz array

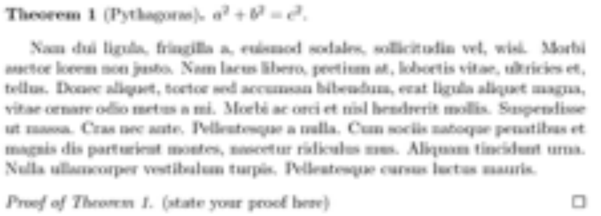
add data


Hypothesis
theory
SANDBOX

Data science **needs** to be…

- Practical **findings** that explain reality,
  NOT theorems!
- Reliable comprehensive **insights,**
  NOT poetry,
  NOT cherry picking,
  NOT inadequate experimentation.

Data science **needs** to be **XYZ!**

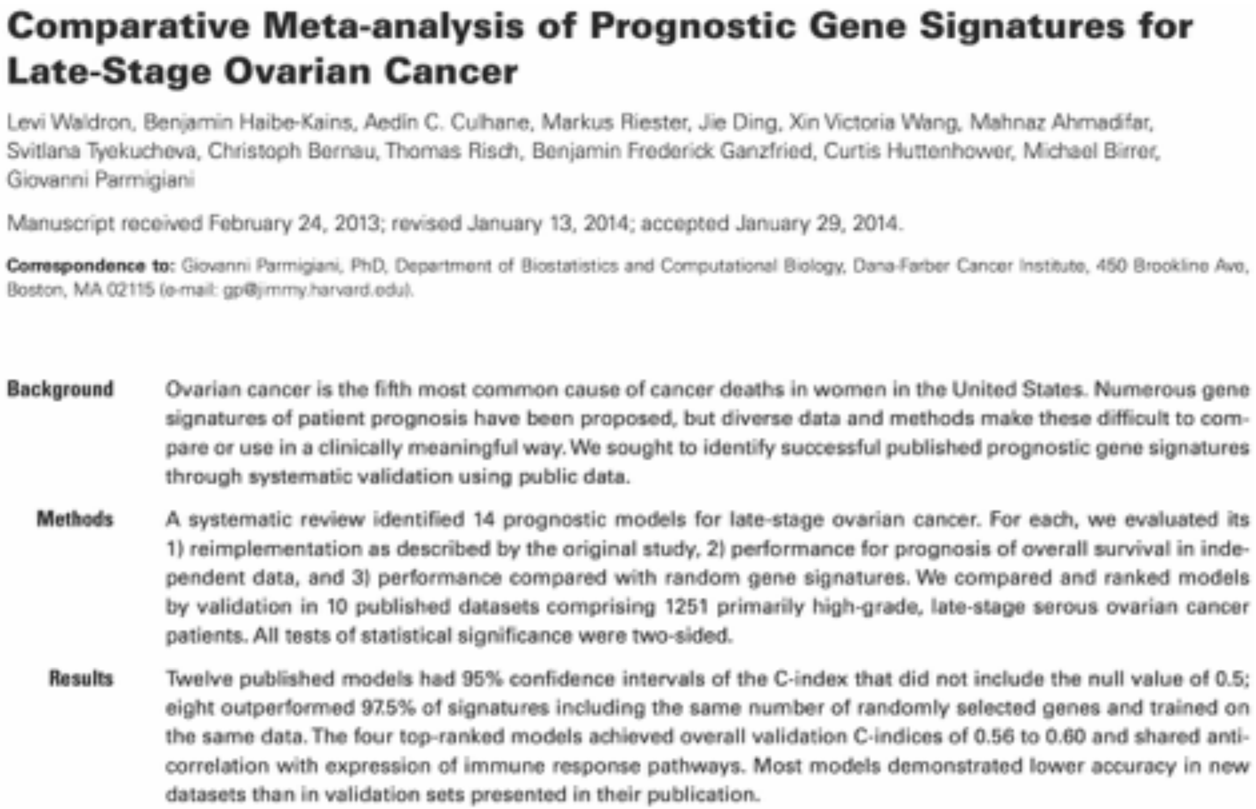


Theorem 1 (Pythagoras). $a^2 + b^2 = c^2$.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Proof of Theorem 1. (state your proof here) □

People are groping for this



Comparative Meta-analysis of Prognostic Gene Signatures for Late-Stage Ovarian Cancer

Levi Waldron, Benjamin Haibe-Kains, Aedín C. Culhane, Markus Riester, Jie Ding, Xin Victoria Wang, Mahnaz Ahmadifar, Svitlana Tyekucheva, Christoph Bernau, Thomas Risch, Benjamin Frederick Ganzfried, Curtis Huttenhower, Michael Birrer, Giovanni Parmigiani

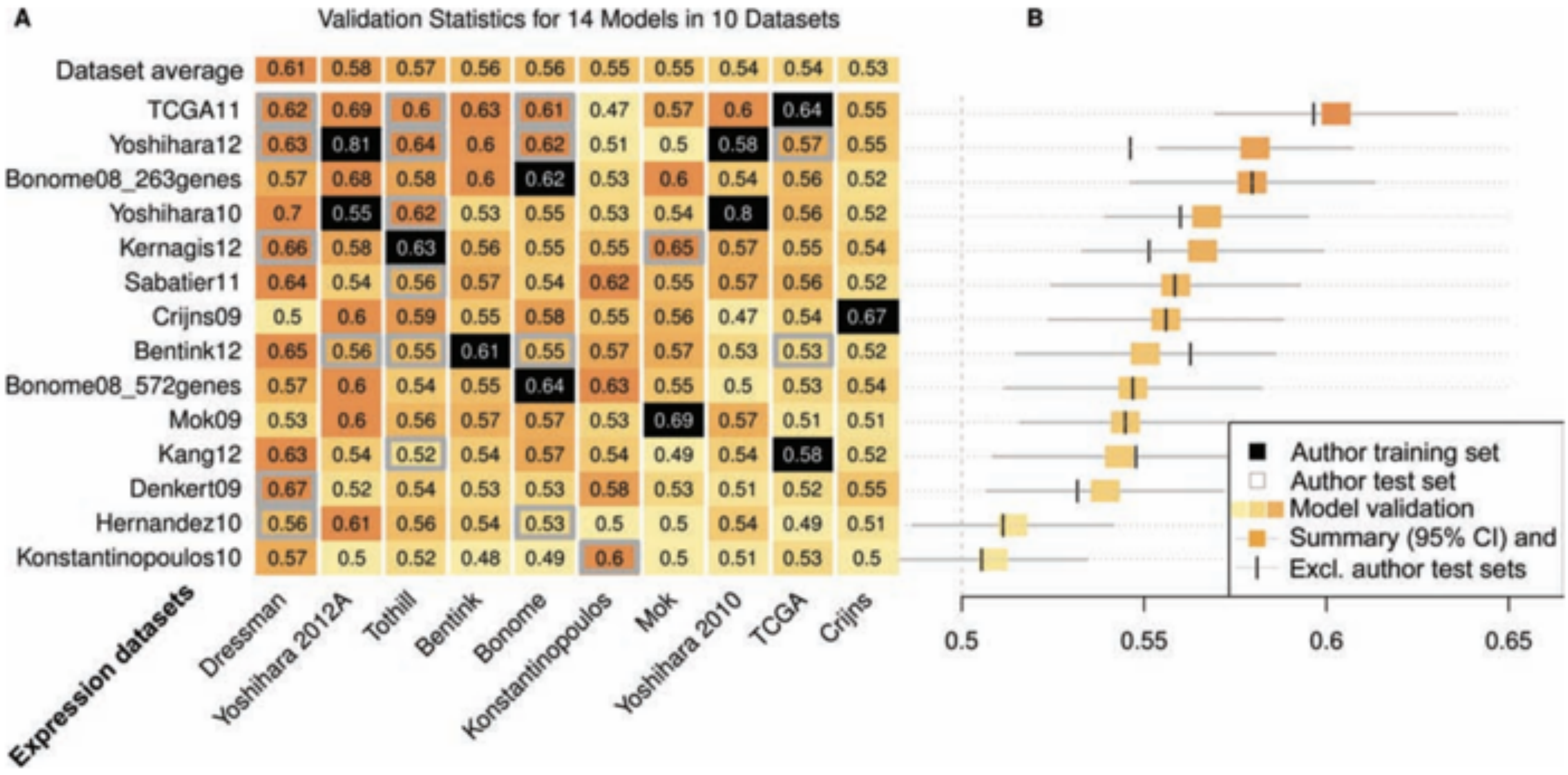
Manuscript received February 24, 2013; revised January 13, 2014; accepted January 29, 2014.

Correspondence to: Giovanni Parmigiani, PhD, Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 450 Brookline Ave, Boston, MA 02115 (e-mail: gp@jimmy.harvard.edu).

Background Ovarian cancer is the fifth most common cause of cancer deaths in women in the United States. Numerous gene signatures of patient prognosis have been proposed, but diverse data and methods make these difficult to compare or use in a clinically meaningful way. We sought to identify successful published prognostic gene signatures through systematic validation using public data.

Methods A systematic review identified 14 prognostic models for late-stage ovarian cancer. For each, we evaluated its 1) reimplementation as described by the original study, 2) performance for prognosis of overall survival in independent data, and 3) performance compared with random gene signatures. We compared and ranked models by validation in 10 published datasets comprising 1251 primarily high-grade, late-stage serous ovarian cancer patients. All tests of statistical significance were two-sided.

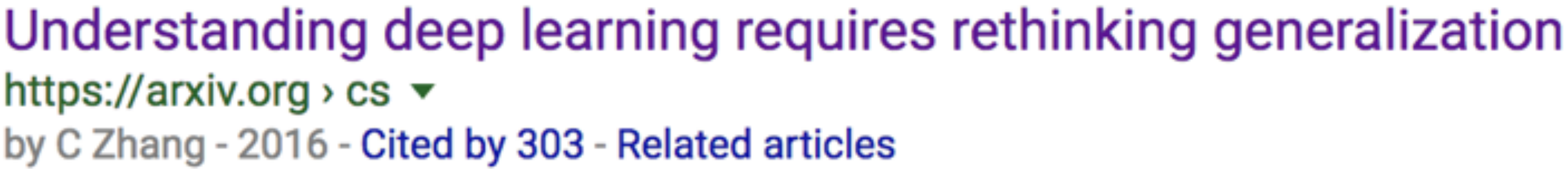
Results Twelve published models had 95% confidence intervals of the C-index that did not include the null value of 0.5; eight outperformed 97.5% of signatures including the same number of randomly selected genes and trained on the same data. The four top-ranked models achieved overall validation C-indices of 0.56 to 0.60 and shared anticorrelation with expression of immune response pathways. Most models demonstrated lower accuracy in new datasets than in validation sets presented in their publication.



Validation Statistics for 14 Models in 10 Datasets

Expression datasets	Dressman	Yoshihara 2012A	Tothill	Bentink	Bonome	Konstantinopoulos	Mok	Yoshihara 2010	TCGA	Crijns
Dataset average	0.61	0.58	0.57	0.56	0.56	0.55	0.55	0.54	0.54	0.53
TCGA11	0.62	0.69	0.6	0.63	0.61	0.47	0.57	0.6	0.64	0.55
Yoshihara12	0.63	0.81	0.64	0.6	0.62	0.51	0.5	0.58	0.57	0.55
Bonome08_263genes	0.57	0.68	0.58	0.6	0.62	0.53	0.6	0.54	0.56	0.52
Yoshihara10	0.7	0.55	0.62	0.53	0.55	0.53	0.54	0.8	0.56	0.52
Kernagis12	0.66	0.58	0.63	0.56	0.55	0.55	0.65	0.57	0.55	0.54
Sabatier11	0.64	0.54	0.56	0.57	0.54	0.62	0.55	0.57	0.56	0.52
Crijns09	0.5	0.6	0.59	0.55	0.58	0.55	0.56	0.47	0.54	0.67
Bentink12	0.65	0.56	0.55	0.61	0.55	0.57	0.57	0.53	0.53	0.52
Bonome08_572genes	0.57	0.6	0.54	0.55	0.64	0.63	0.55	0.5	0.53	0.54
Mok09	0.53	0.6	0.56	0.57	0.57	0.53	0.69	0.57	0.51	0.51
Kang12	0.63	0.54	0.52	0.54	0.57	0.54	0.49	0.54	0.58	0.52
Denkert09	0.67	0.52	0.54	0.53	0.53	0.58	0.53	0.51	0.52	0.55
Hernandez10	0.56	0.61	0.56	0.54	0.53	0.5	0.5	0.54	0.49	0.51
Konstantinopoulos10	0.57	0.5	0.52	0.48	0.49	0.6	0.5	0.51	0.53	0.5

Deep Learning ?



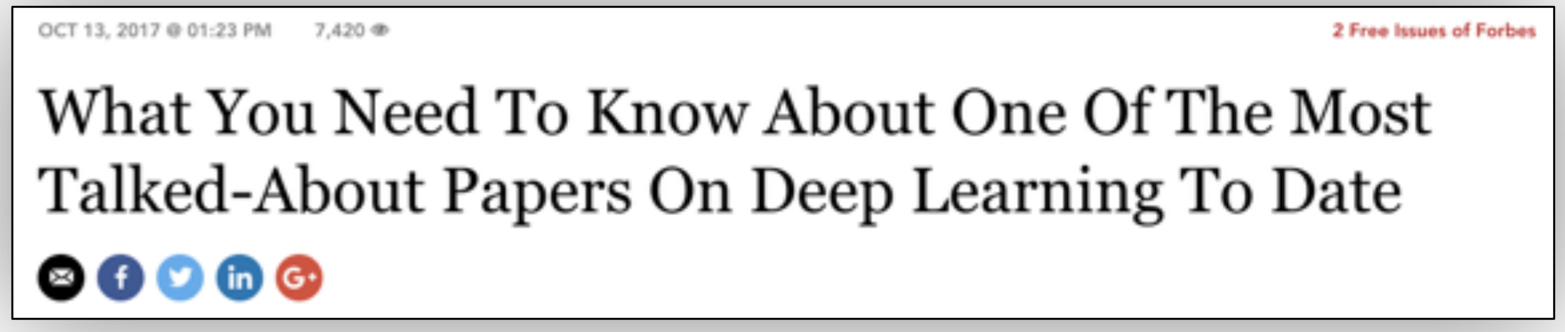
Understanding deep learning requires rethinking generalization

https://arxiv.org › cs ▾

by C Zhang - 2016 - Cited by 303 - Related articles

Perfect score on the ICLR reviews

ICLR 2017 best paper award



OCT 13, 2017 @ 01:23 PM 7,420

2 Free Issues of Forbes

What You Need To Know About One Of The Most Talked-About Papers On Deep Learning To Date

Rethinking Generalization by Zhang et. al	CIFAR10, ImageNet	MLP, AlexNet, Inception	% randomized labels	number of epochs until perfect fit, test error at epoch of perfect fit	Could be done on more datasets and methods

Rethinking Generalization by Zhang et. al	CIFAR10, ImageNet	MLP, AlexNet, Inception	% randomized labels	number of epochs until perfect fit, test error at epoch of perfect fit	Could be done on more datasets and methods
Importance of Single Direction Generalization Morcos et. al	MNIST, CIFAR10, ImageNet	MLP, 11-layer CNN, ResNet	dropout, batch normalization, % randomized labels	reliance on single neuron, class specificity	Relied on control parameters in previous work

	X	Y	Z	W	?
Rethinking Generalization by Zhang et. al	CIFAR10, ImageNet	MLP, AlexNet, Inception	% randomized labels	number of epochs until perfect fit, test error at epoch of perfect fit	Could be done on more datasets and methods
Importance of Single Direction Generalization Morcos et. al	MNIST, CIFAR10, ImageNet	MLP, 11-layer CNN, ResNet	dropout, batch normalization, % randomized labels	reliance on single neuron, class specificity	Relied on control parameters in previous work
Are GANs Created Equal? Lucic et. al	MNIST, FASHION - MNIST, CIFAR10, CELEBA	MM GAN, NS GAN, LSGAN, WGAN, WGAN GP, DRAGAN, BEGAN, VAE	seed, computational budget	precision, recall, F1	Great example!

If you want to be a data scientist…

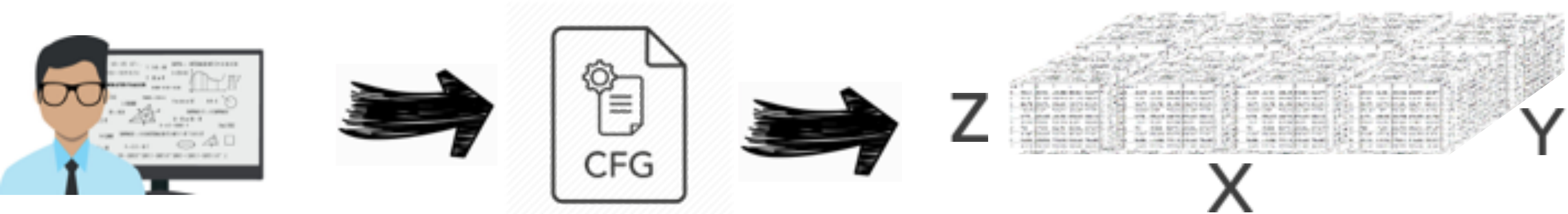
You **must** do research this way

You **must** evaluate others this way

And…

You **must** accept this is the only way,

otherwise your work will be irrelevant



hadoop

CodaLab

Apache Spark

ElastiCluster

Pywren

Caffe

Chainer

DL4J Deeplearning4j

KERAS

Microsoft CNTK

MatConvNet

MINERVA

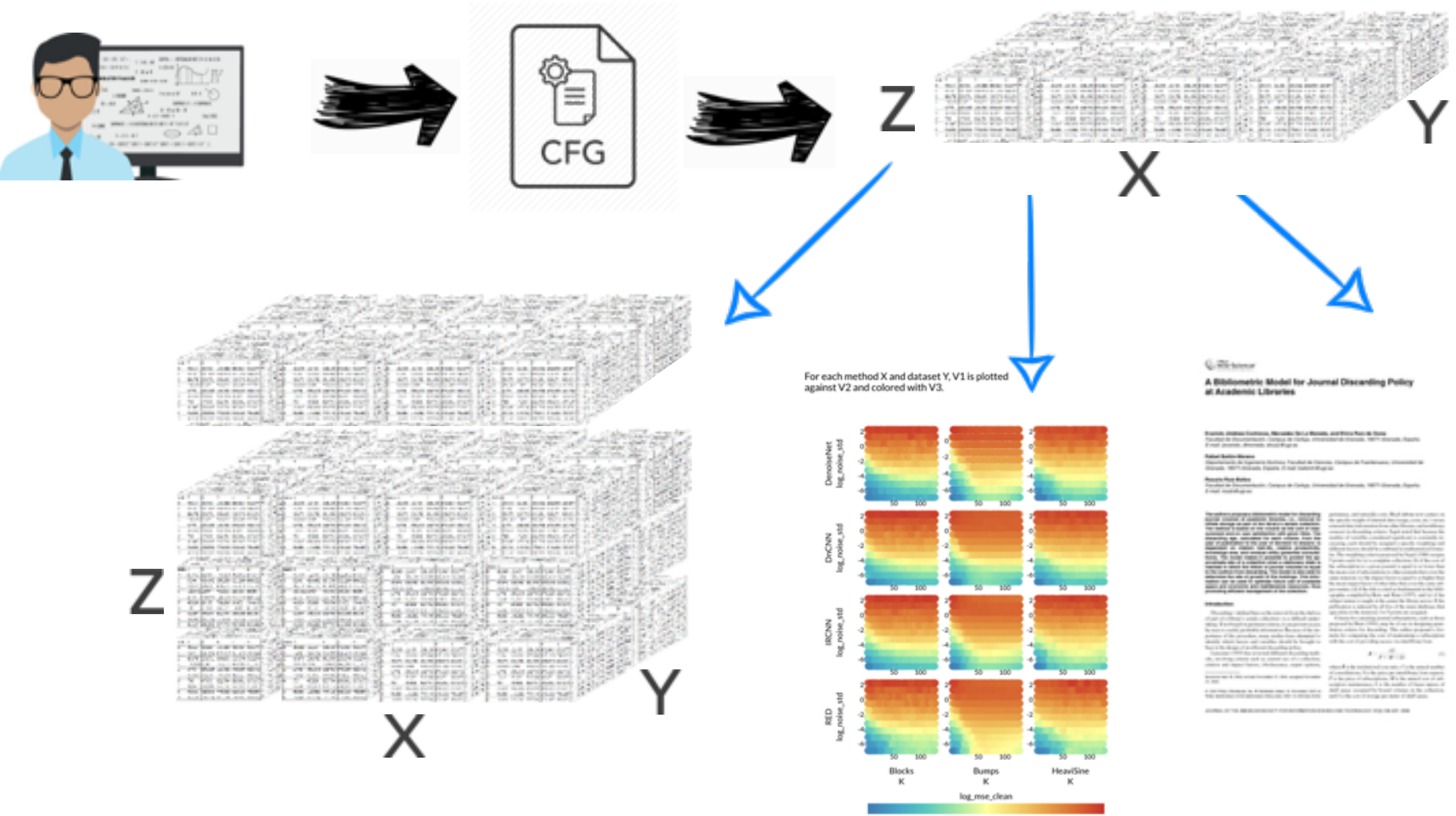
mxnet

Purine

TensorFlow

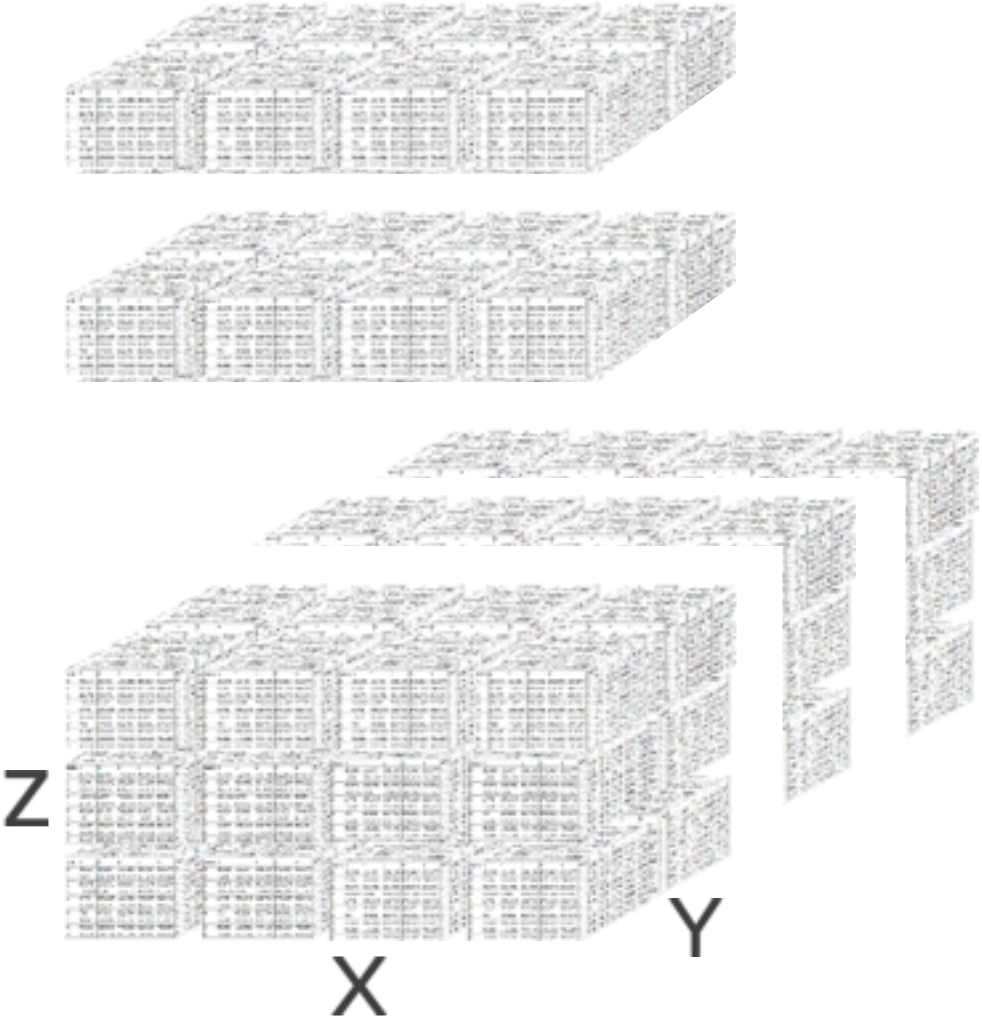
theano

torch



For each method X and dataset Y, V1 is plotted against V2 and colored with V3.

A Bibliometric Model for Journal Discarding Policy at Academic Libraries


Z
Y
X