The next challenges for deep learning in biology and medicine

Perelman
School of Medicine
University of Pennsylvania

CHILDHOOD CANCER DATA LAB
Powered by: Alex's Lemonade Stand Foundation

@GreeneScientist

Deep learning can work with images from health care.

MIT Technology Review

Artificial Intelligence Nov 16, 2017

A New Algorithm Can Spot Pneumonia Better Than a Radiologist

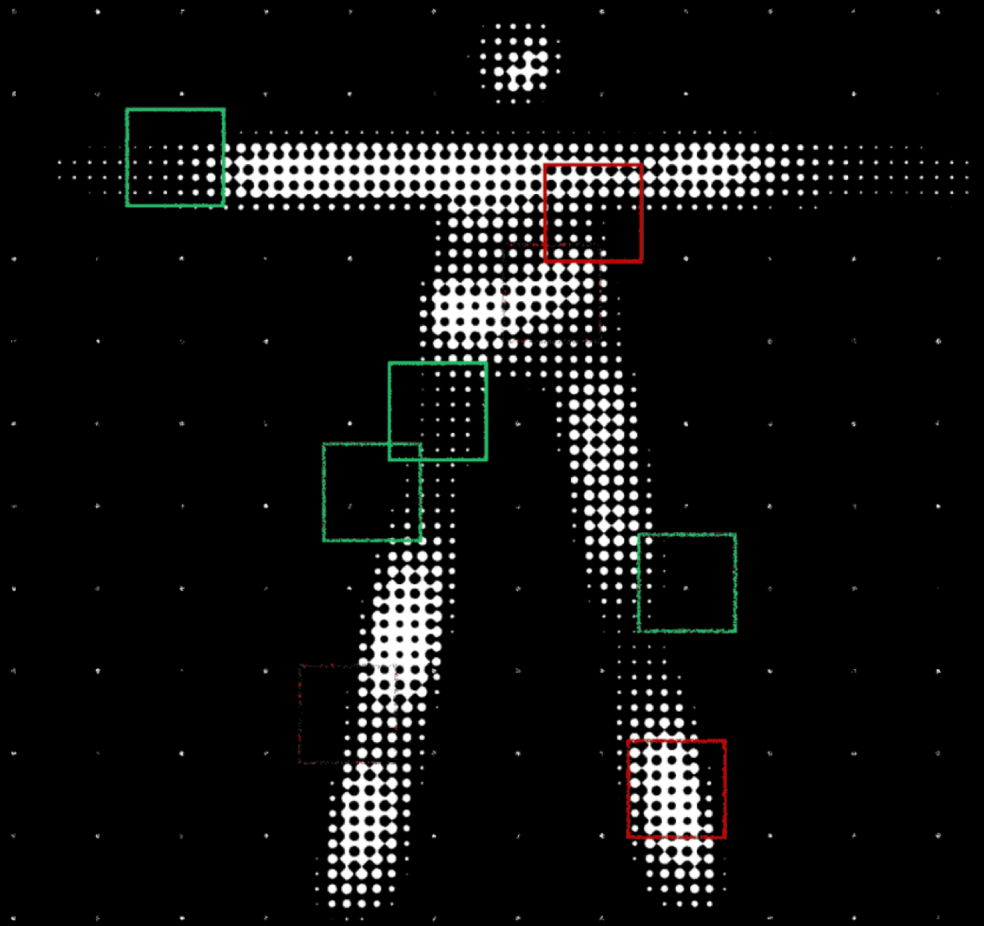
Add diagnosing dangerous lung diseases to the growing list of things artificial intelligence can do better than humans.

ANNALS OF MEDICINE

A.I. VERSUS M.D.

What happens when diagnosis is automated?

By Siddhartha Mukherjee March 27, 2017



ANNALS OF MEDICINE

A.I. VERSUS M.D.

What happens when diagnosis is automated?

By Siddhartha Mukherjee March 27, 2017

“They should stop training radiologists right now.”

- Geoffrey Hinton

Deep learning can work with images from health care.

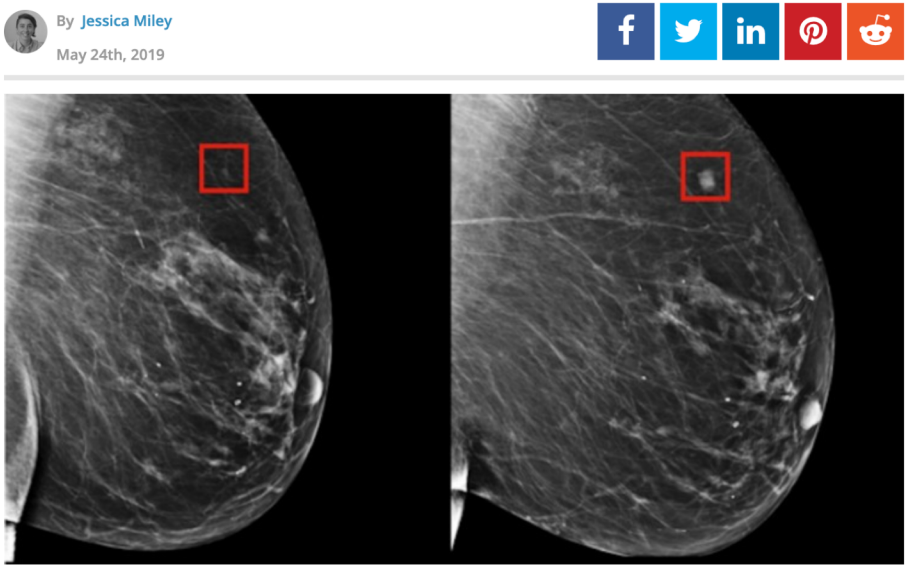
AI-powered detection system IDs ACL tears as well as radiologists

May 14, 2019 | *Michael Walter* | *Artificial Intelligence*

Deep learning can work with images from health care.

Deep Learning Model Can Predict Breast Cancer up to Five Years in Advance

The system will help develop individual risk management plans.



By Jessica Miley

May 24th, 2019

MIT

Deep learning can work with images from health care.

Emergent Tech ▸ Artificial Intelligence

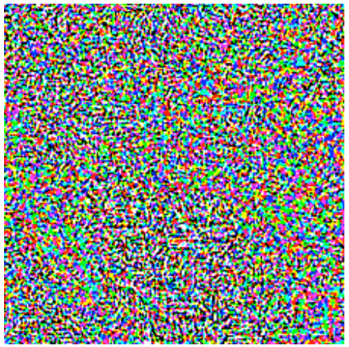
Ahem, ahem... AI engine said to be good as human docs at spotting lung cancer developing

New convolutional neural network did just as well as radiologists in clinical settings

By Katyanna Quach 21 May 2019 at 07:03 23 SHARE ▼

Oddities: Adversarial Examples

$+ .007 \times$



$\boldsymbol{x}$

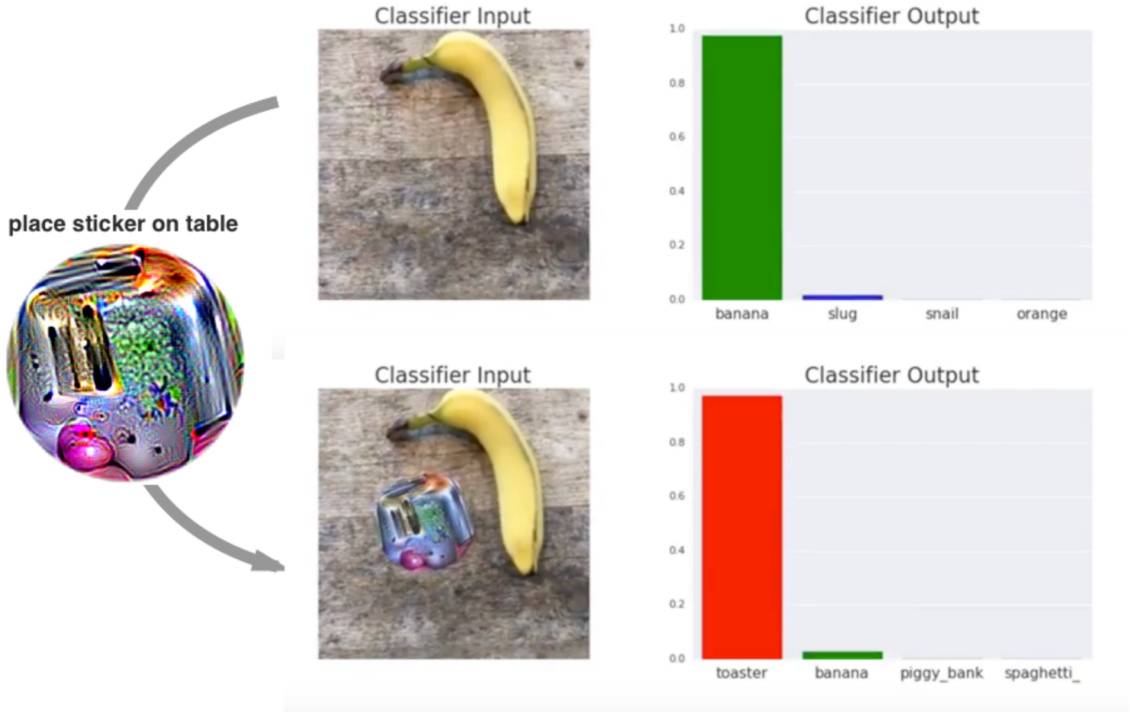
"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

Goodfellow et al. ICLR 2015

Oddities: Adversarial Patches



Brown et al. arXiv 1712.09665

Oddities: #aiweirdness

Let's go back to that automated radiologist finding pneumonia

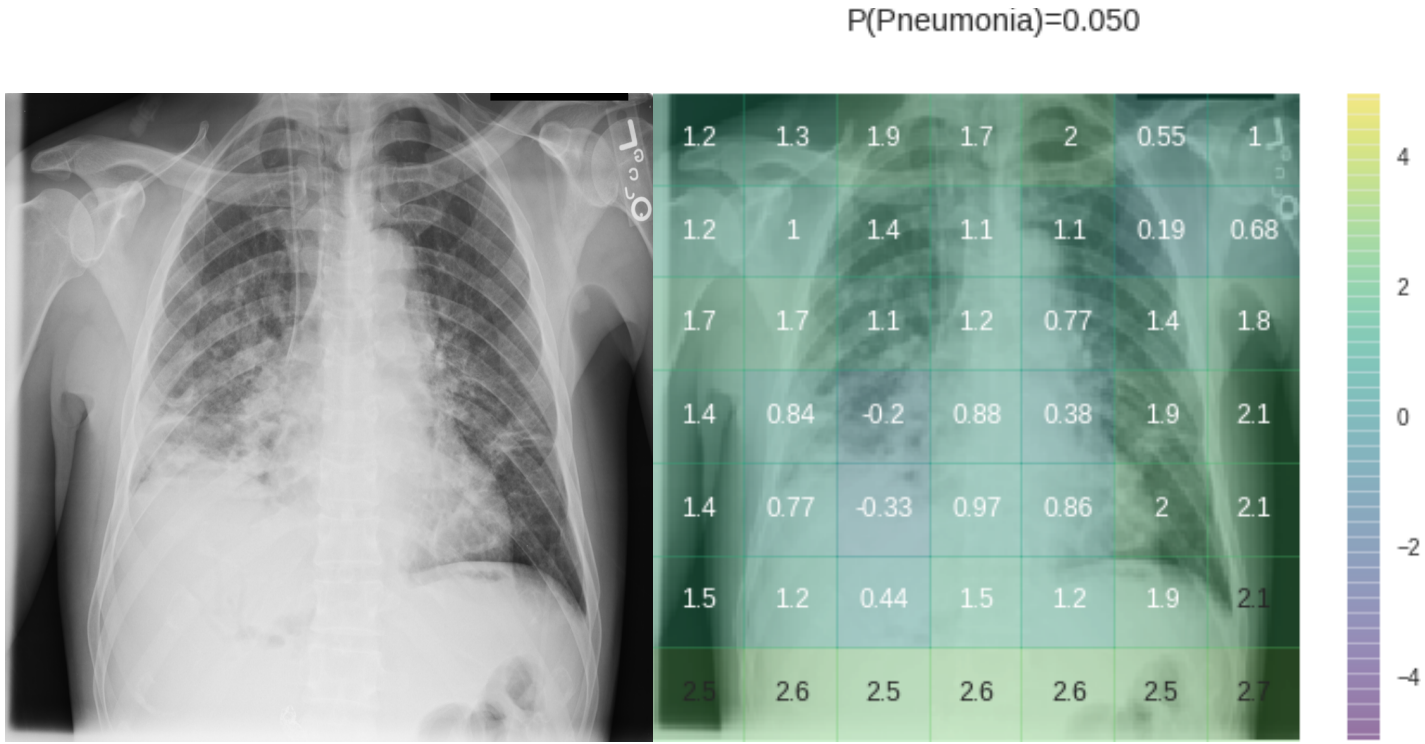
MIT Technology Review

Artificial Intelligence Nov 16, 2017

A New Algorithm Can Spot Pneumonia Better Than a Radiologist

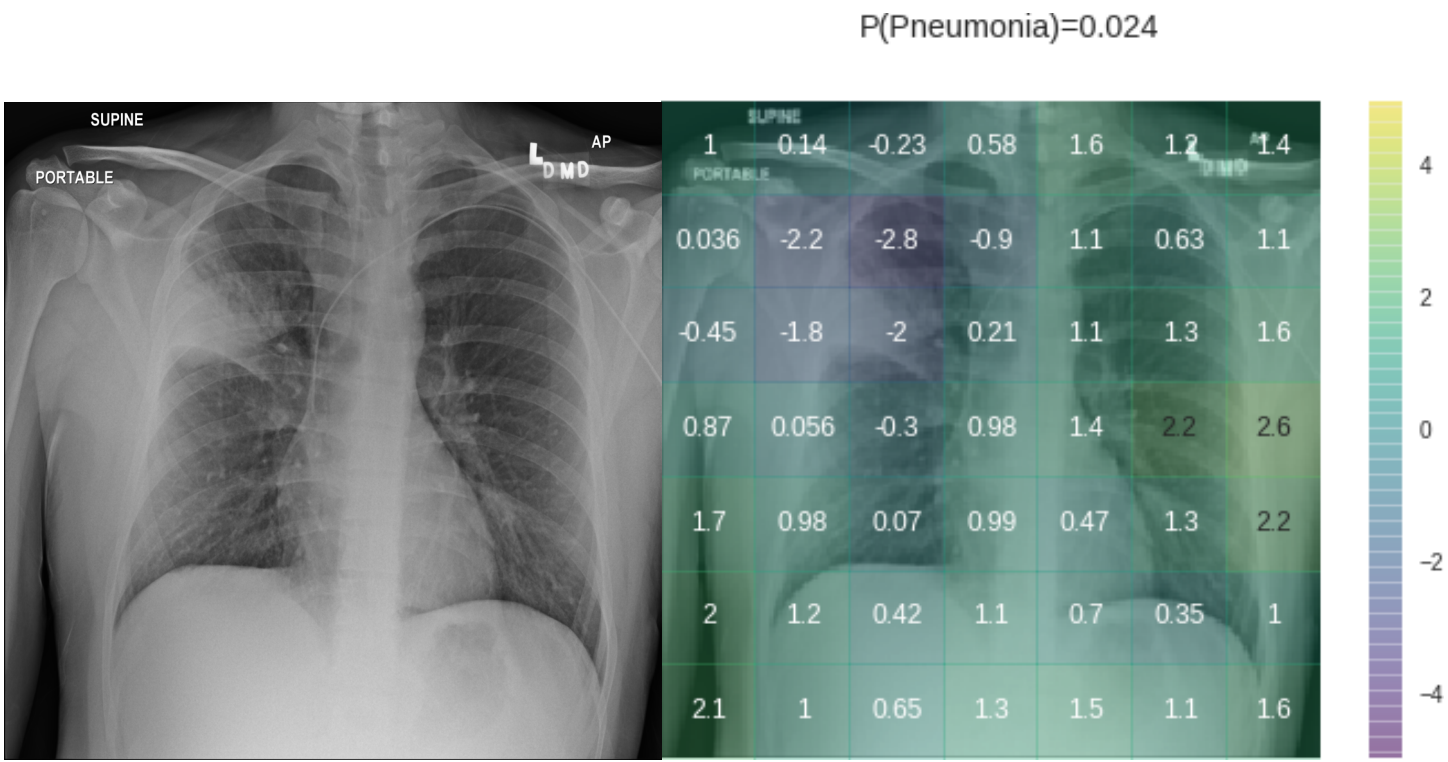
Add diagnosing dangerous lung diseases to the growing list of things artificial intelligence can do better than humans.

Why is the model making good predictions?



What are radiological deep learning models actually learning? John Zech (medium.com)

PORTABLE is a giveaway, but it's probably not the one you want.



What are radiological deep learning models actually learning? John Zech (medium.com)

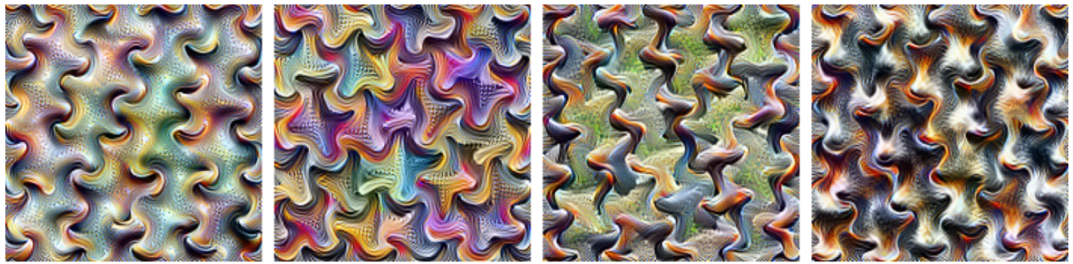
On big data, data collection biases are always larger than statistical uncertainty…

- Daniel Himmelstein

Don’t romanticize deep NNs.

Low-level Neuron

Simple Optimization



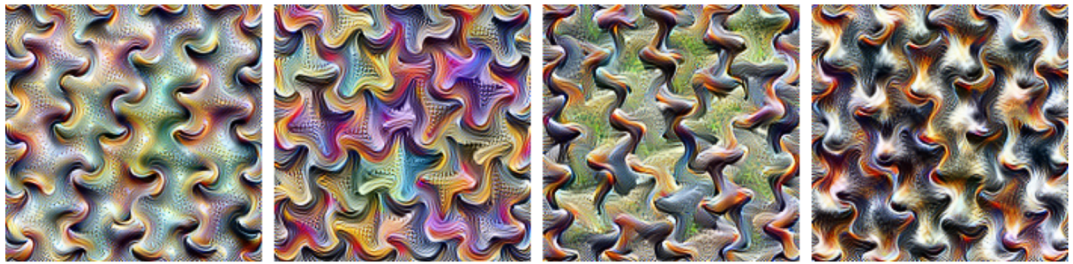
Dataset examples

Olah et al. Feature Visualization. *Distill*, 2017

Don’t romanticize deep NNs.

Low-level Neuron

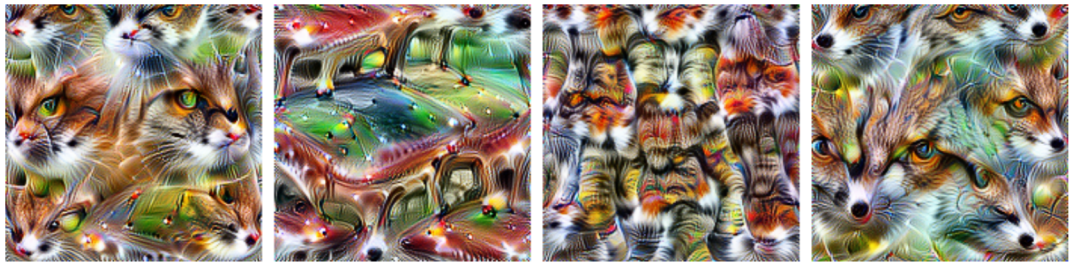
Simple Optimization



Dataset examples

High-level Neuron

Simple Optimization



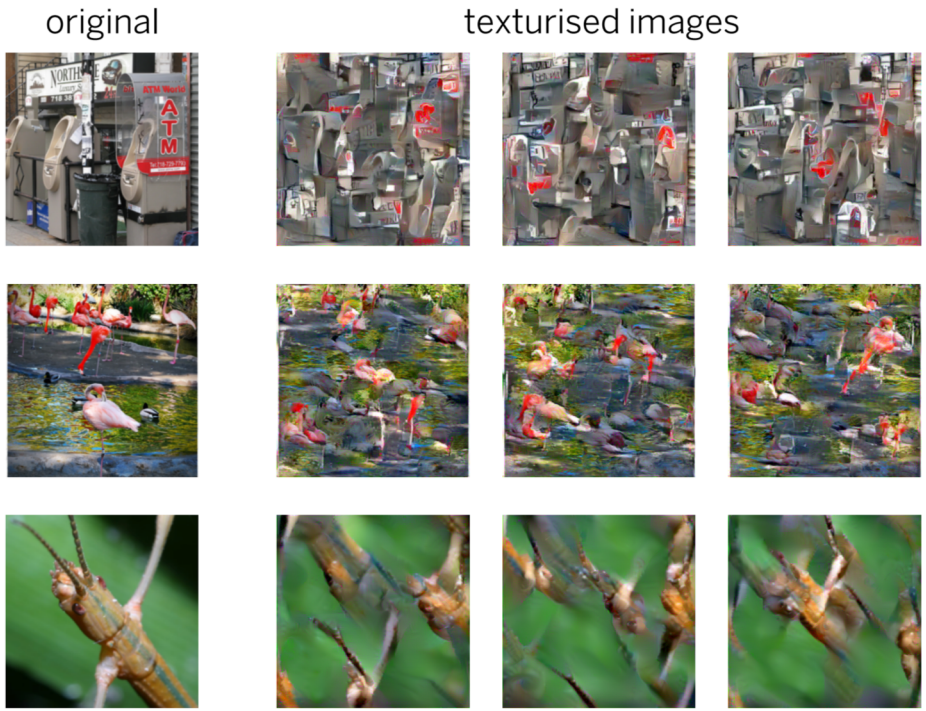
Dataset examples

Olah et al. Feature Visualization. *Distill*, 2017

Don’t romanticize deep NNs.

Brendel & Bethge. *ICLR*, 2019

Don’t romanticize deep NNs.

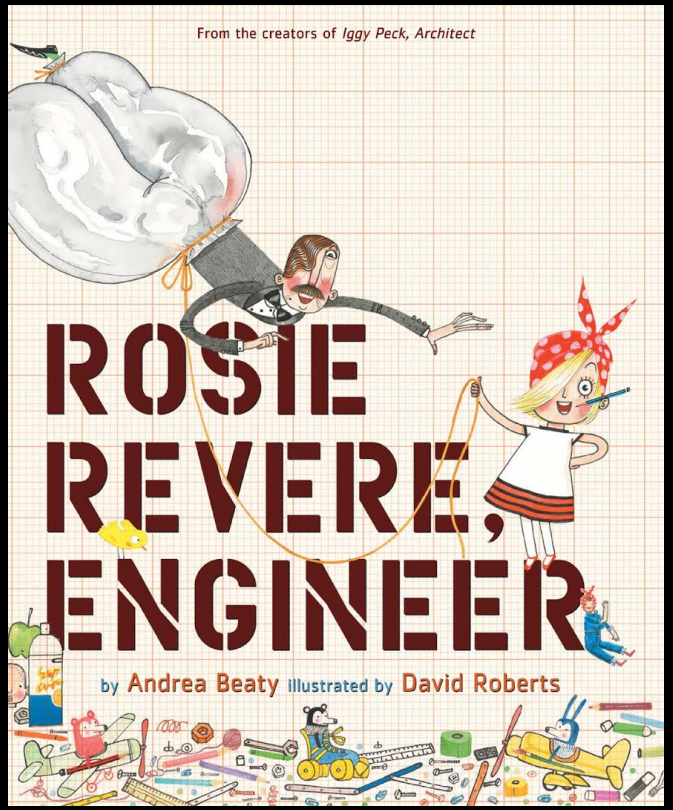


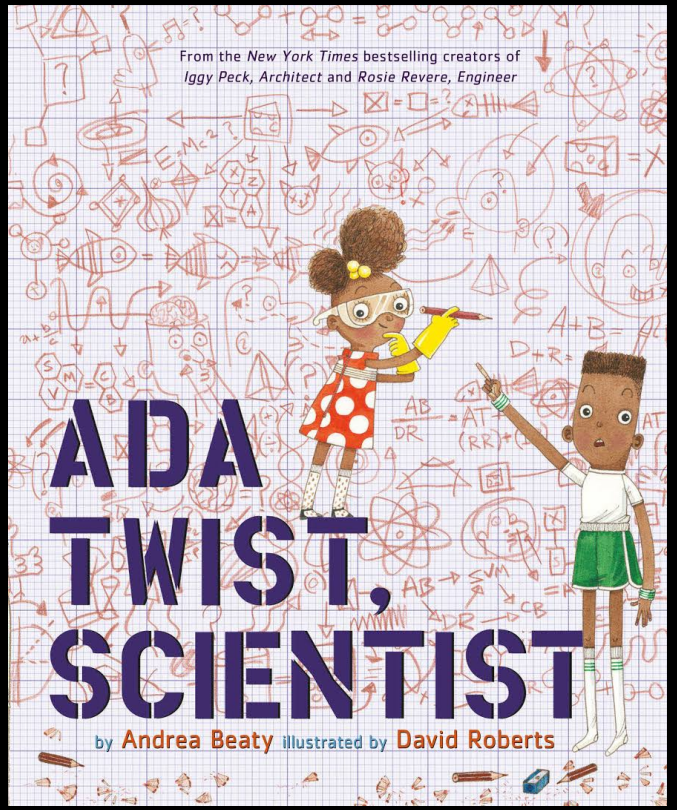
Brendel & Bethge. *ICLR*, 2019

Challenges for Deep NNs in Bio/Med (especially genomics/health records)

- There are likely to be very few examples relative to other fields (self driving cars, object recognition).
- The examples that you do have are not IID.
- Nothing is really static over time (new drugs, treatments, compensation rates).
- Deep NNs are unlikely to learn what you think they are learning and it probably matters that they don't.

Who are you?


From the creators of *Iggy Peck, Architect*
ROSIE REVERE, ENGINEER
by Andrea Beaty illustrated by David Roberts


From the *New York Times* bestselling creators of *Iggy Peck, Architect* and *Rosie Revere, Engineer*
ADA TWIST, SCIENTIST
by Andrea Beaty illustrated by David Roberts

MultiPLIER: a transfer learning framework for transcriptomics reveals systemic features of rare disease

GitHub: http://github.com/greenelab/multi-plier

Taroni, Grayson, et al. *Cell Systems.* 2019

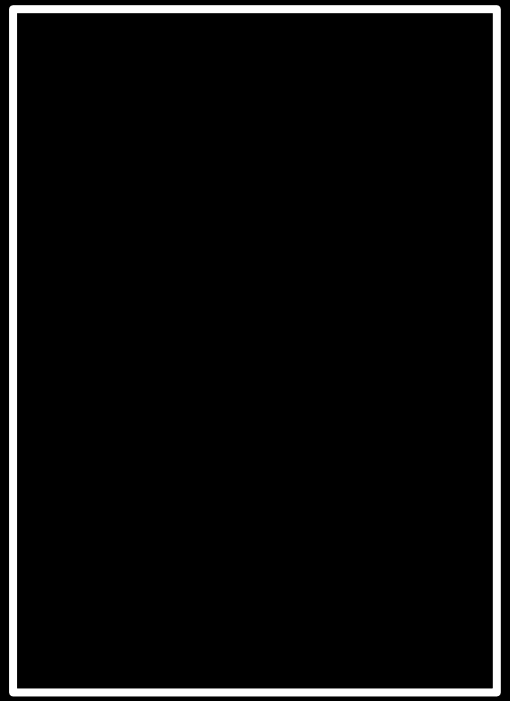
http://bit.ly/MultiPLIER

Machine learning isn’t magic.

Image by Hayden Beaumont (flickr)

Machine learning needs many examples.

We can know some things



About many people

Rare diseases have few examples.

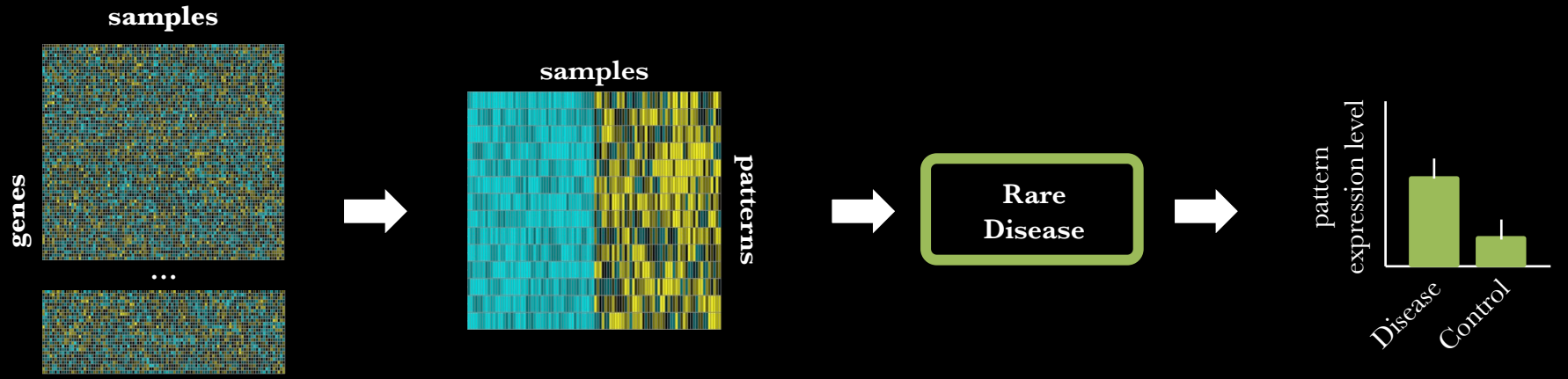
We can know many things

About few people

We need to make ML models for *many* biological contexts.

Can we sidestep the challenge?

HYPOTHESIS: Biologically meaningful patterns, including cell type information, can be learned from heterogeneous gene expression data and then *transferred* to the dataset(s) of interest.

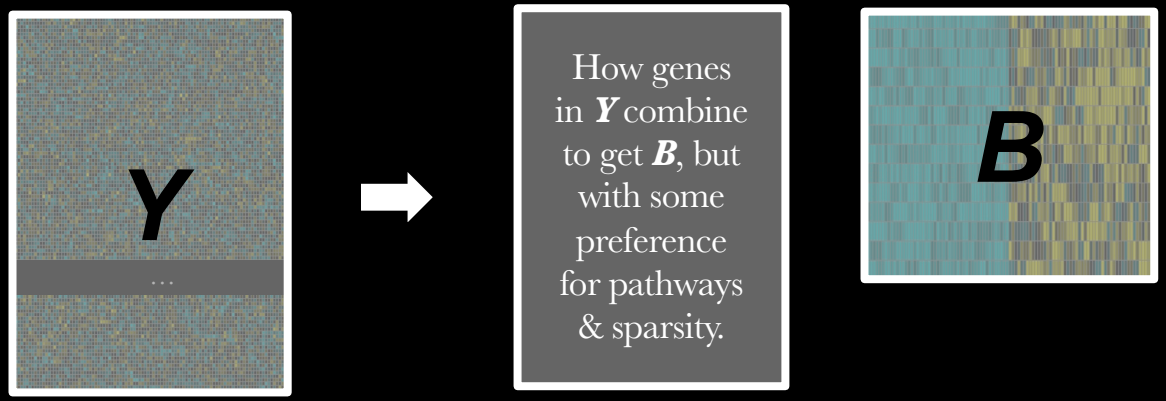


Hypothetical data

recount2: Collado-Torres et al.

70,603 RNA-seq samples

PLIER: Pathway Level Information ExtractoR



Mao, et al. *bioRxiv*. 2017.

PLIER: Pathway Level Information ExtractoR

$$||Y - ZB||^2_F + \lambda_1||Z - CU||^2_F + \lambda_2||B||^2_F + \lambda_3||U||_{L1}$$

Y = gene expression matrix

Z = loadings

B = latent variables matrix

"Minimize the error of representing the gene expression with the latent variables"

Slides from Ben Heil

PLIER: Pathway Level Information ExtractoR

$$||Y - ZB||^2_F + \lambda_1||Z - CU||^2_F + \lambda_2||B||^2_F + \lambda_3||U||_{L1}$$

Z = Loadings

C = Gene Sets

U = Gene Set Coefficients

"Ensure that the loadings for the latent variables look like gene sets"

Slides from Ben Heil

PLIER: Pathway Level Information ExtractoR

$$||Y - ZB||^2_F + \lambda_1||Z - CU||^2_F + \lambda_2||B||^2_F + \lambda_3||U||_{L1}$$

B = Latent variables

“Regularize the latent variables with an L2 Norm”

Slides from Ben Heil

PLIER: Pathway Level Information ExtractoR

$$||Y - ZB||^2_F + \lambda_1||Z - CU||^2_F + \lambda_2||B||^2_F + \lambda_3||U||_{L1}$$

U = Gene Set Coefficients

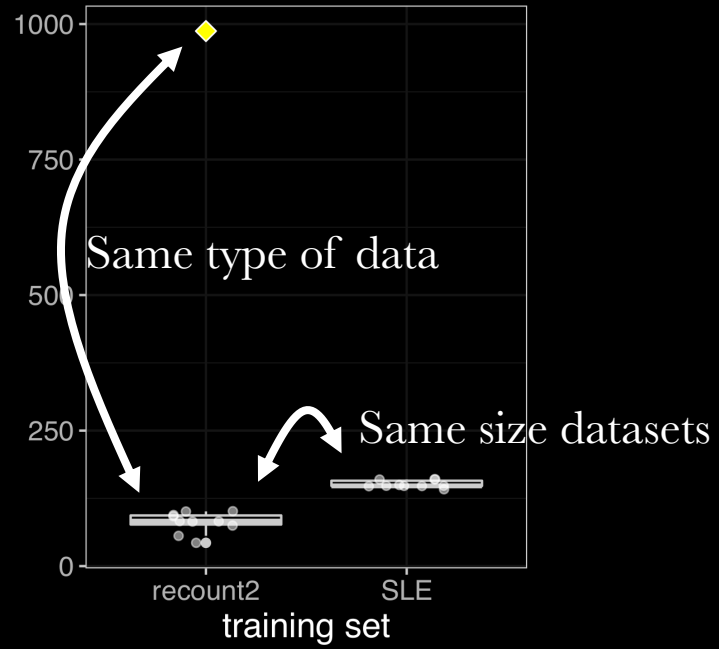
"Regularize the gene set coefficients with L1 Norm"

Slides from Ben Heil

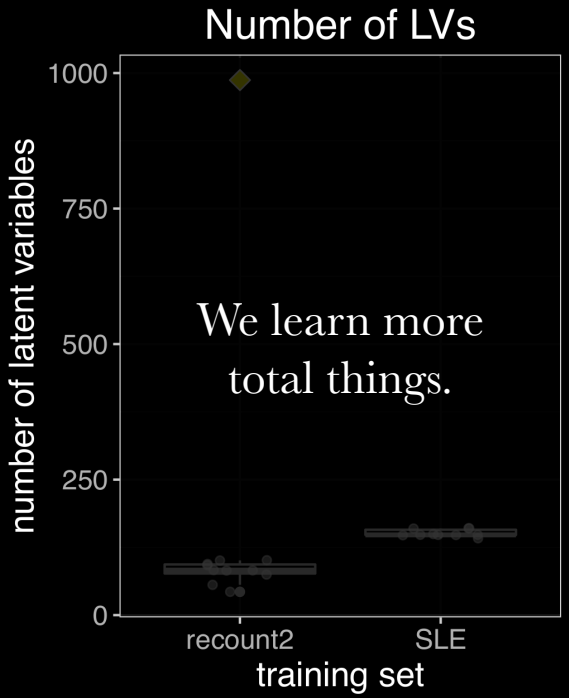
We need to make ML models for *many* biological contexts.

recount2 + PLIER = MultiPLIER.

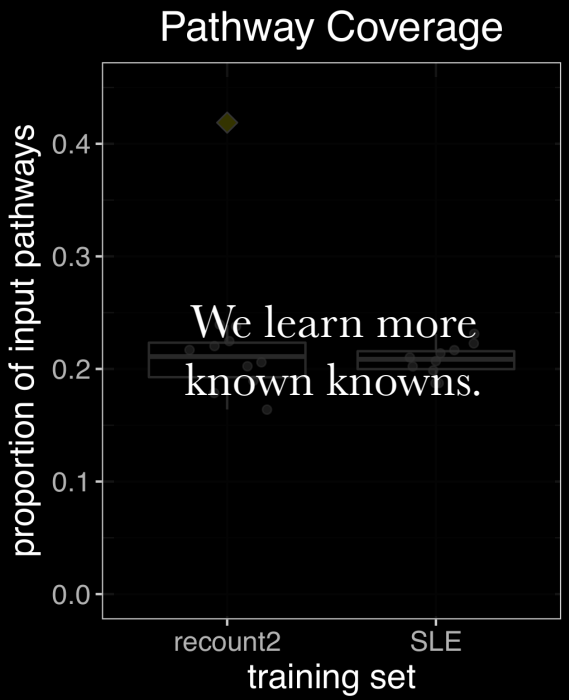
The MultiPLIER model captures more known knowns and unknown unknowns.



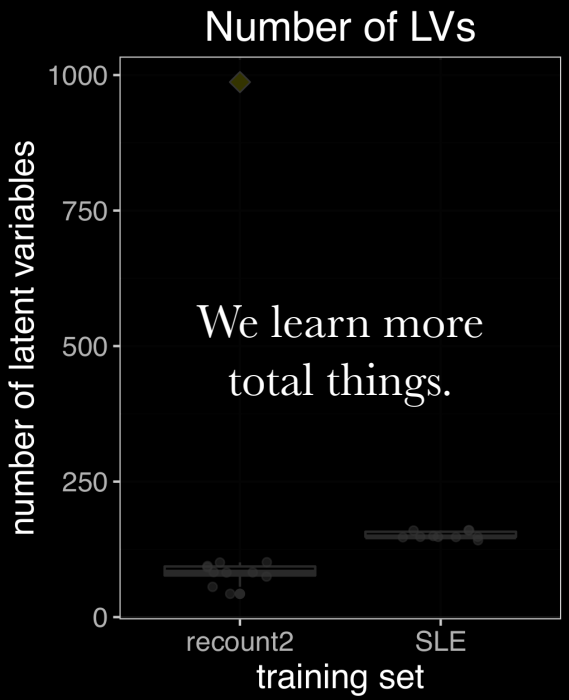
The MultiPLIER model captures more known knowns and unknown unknowns.

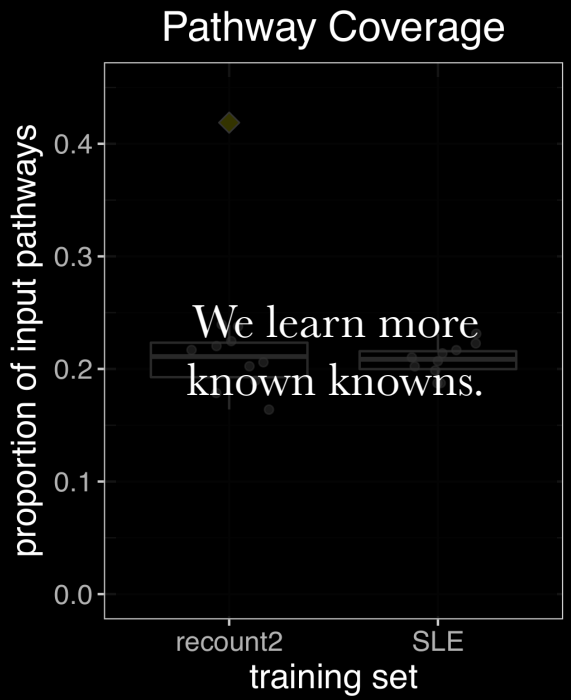


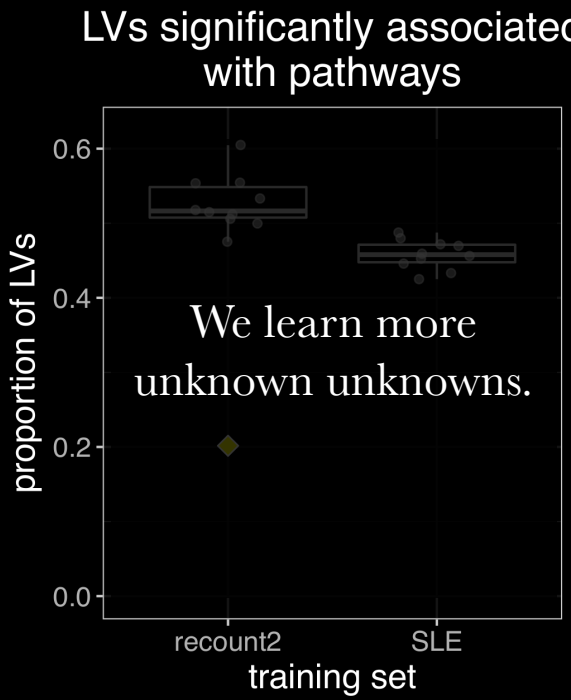
The MultiPLIER model captures more known knowns and unknown unknowns.



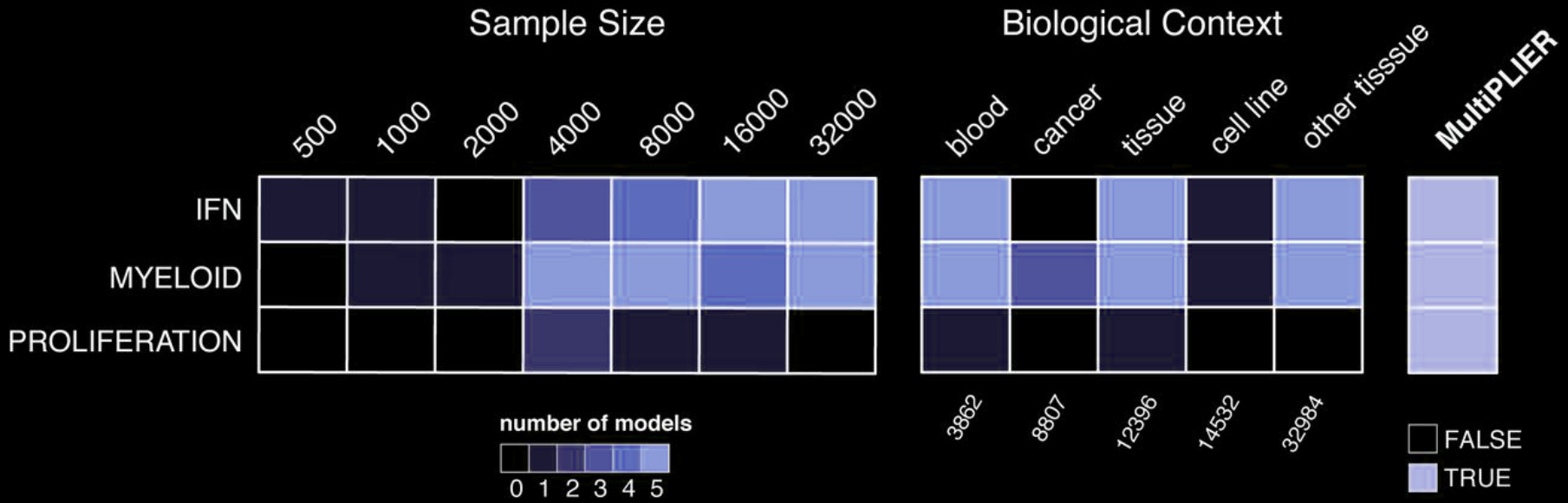
The MultiPLIER model captures more known knowns and unknown unknowns.



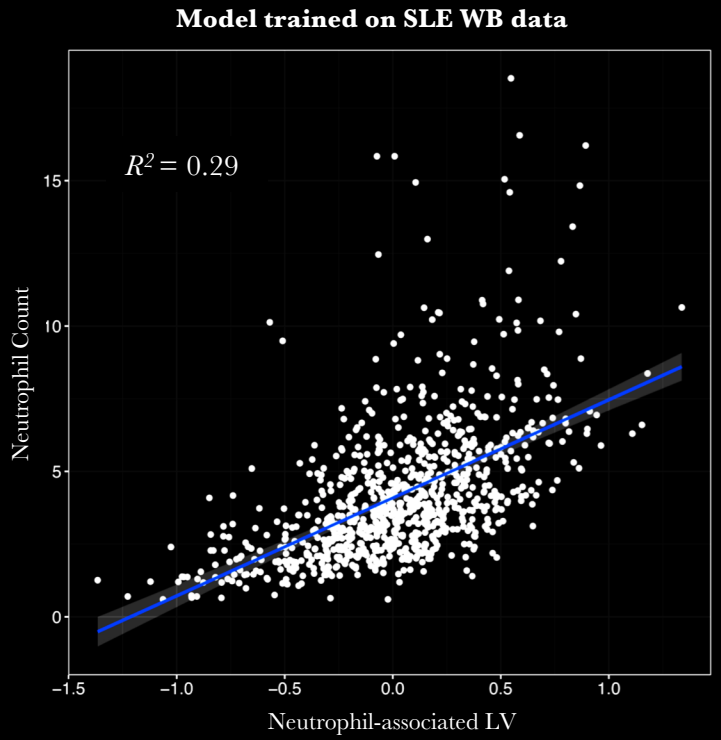




MultiPLIER disentangles related pathways & benefits from size and content.

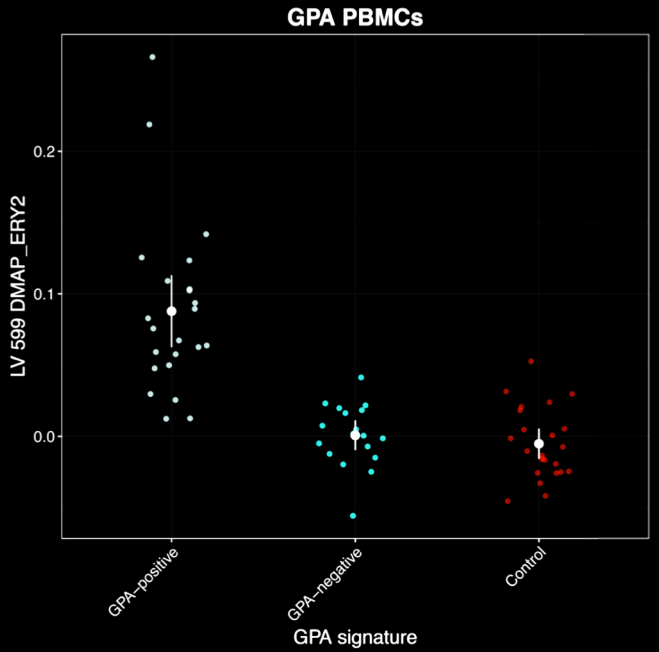


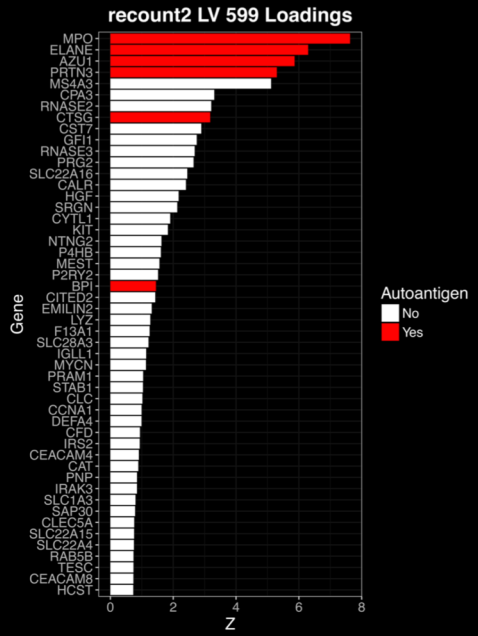
MultiPLIER LVs correlate with neutrophil counts.



Data from Banchereau, et al. *Cell*. 2016.

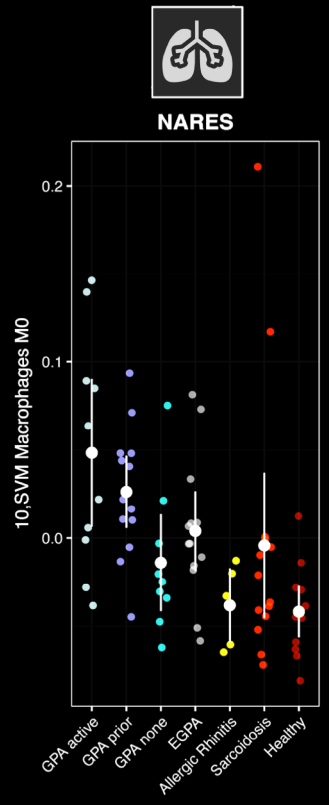
ANCA antigens are captured by MultiPLIER





Expression data from Cheadle, et al. *A&R*. 2010.

Multi-tissue comparison via MultiPLIER



http://bit.ly/MultiPLIER

ML analyses that re-use data reach a level of detail that was otherwise impossible.

Notebooks

Analysis notebooks are numbered and present in the top level directory. We've enabled Github pages for easy viewing of the notebooks. Some steps in the pipeline are R scripts rather than notebooks due to their computationally intensive nature; we exclude these from the TOC below.

- PLIER functions proof of concept
- Exploratory analysis of the recount2 PLIER model (MultiPLIER)
- MultiPLIER on isolated immune cell populations microarray data
- Reconstruction of isolated immune cell data
- Training PLIER on the SLE whole blood compendium
- Analyzing cell type-associated LVs in the SLE WB data with SLE WB PLIER model
- Analyzing cell type patterns in the SLE WB data using MultiPLIER
- Identifying interferon-related LVs in the SLE WB and MultiPLIER models
- Preparing IFN trials data for plotting
- Plotting IFN trial results
- Training a PLIER model on the NARES nasal brushing microarray dataset
- Comparing the NARES and MultiPLIER latent spaces
- Comparing the MultiPLIER neutrophil-associated LV expression values to `MCPcounter` estimates
- Evaluating PLIER models trained on subsampled recount2 compendia
- Plotting metrics for PLIER model repeats
- Identifying differentially expressed MultiPLIER LVs (DELVs) in the NARES nasal brushing dataset
- Identifying DELVs in granulomatosis with polyangiitis (GPA) peripheral blood mononuclear cells (PBMCs)
- Identifying DELVs in microdissected glomeruli cohort
- Identifying DELVs common across the 3 AAV tissues
- ANCA antigens in the GPA PBMCs
- Examining high weight genes in DELVs
- Exploring a rituximab (RTX) dataset (preliminary)
- Predicting RTX response (very preliminary, see #18)
- Describing the recount2 training set with MetaSRA predictions

Greene Lab:

Daniel Himmelstein (Postdoc)
YoSon Park (Postdoc)
Jake Crawford (Grad Student)
Ben Heil (Grad Student)
Ariel Hippen-Anderson (Grad Student)
Alexandra Lee (Grad Student)
David Nicholson (Grad Student)
Dongbo Hu (Programmer)
Vince Rubinetti (Programmer)
Michael Zietz (Undergrad)

CC Data Lab:

Jaclyn Taroni (Data Scientist)
Chante Bethell (Data Analyst)
Candace Savonen (Data Analyst)
Deepa Prasad (UX Designer)
Ariel Rodriguez-Romero (Programmer)
Kurt Wheeler (Programmer)

Collaborators/Alums:

Deep Learning Review Authors (especially Anthony Gitter)

Funding:

Gordon and Betty Moore Foundation (ML)
Alex's Lemonade Stand Foundation (CCDL)
Sloan Foundation (Manubot)
National Science Foundation (CAFA)
Chan-Zuckerberg Initiative (Single Cell + ML)
Pfizer (Hetmech)
NCI (AACES Subtypes, HGSC scRNA-seq)
NHGRI (PheWAS & Networks)
NIAMS (Rare Autoimmune Diseases)
NINDS (NF1)
NHLBI (TOPmed & Networks)

http://greenelab.com / http://ccdatalab.org