Synthesizing fMRI using generative adversarial networks: cognitive neuroscience applications, promises and pitfalls

SANMI KOYEJO
CS @ ILLINOIS

Page intentionally left blank

Page intentionally left blank

Synthetic

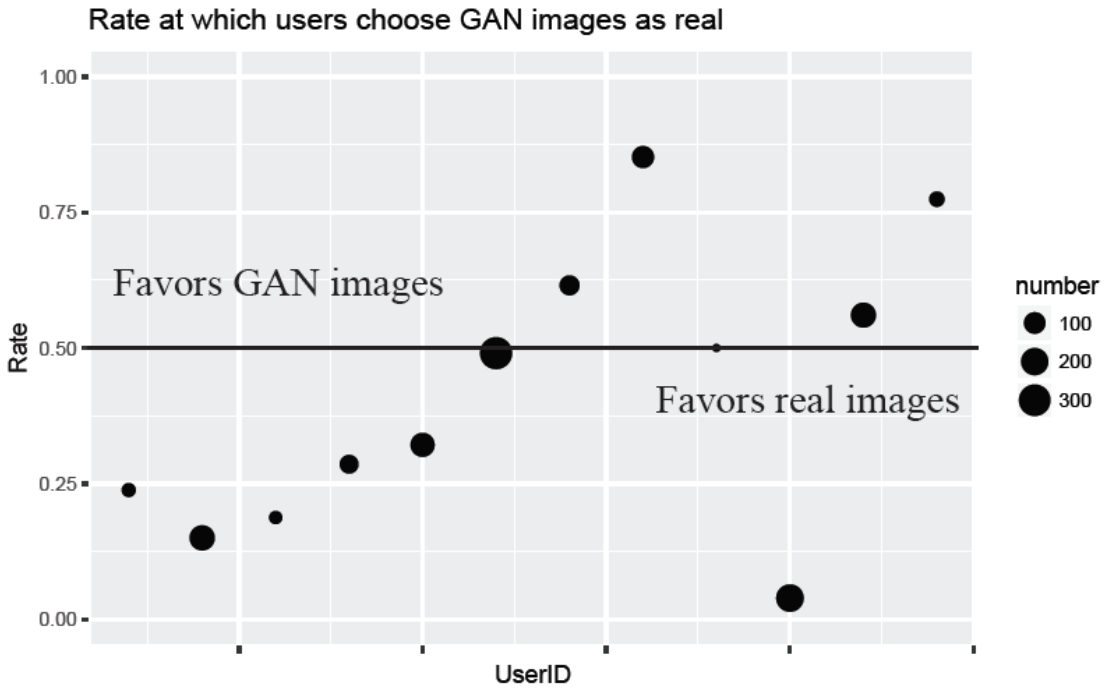
Real

Karras, T., et al. “Progressive growing of gans for improved quality, stability, and variation.” ICLR 2018.

Synthetic

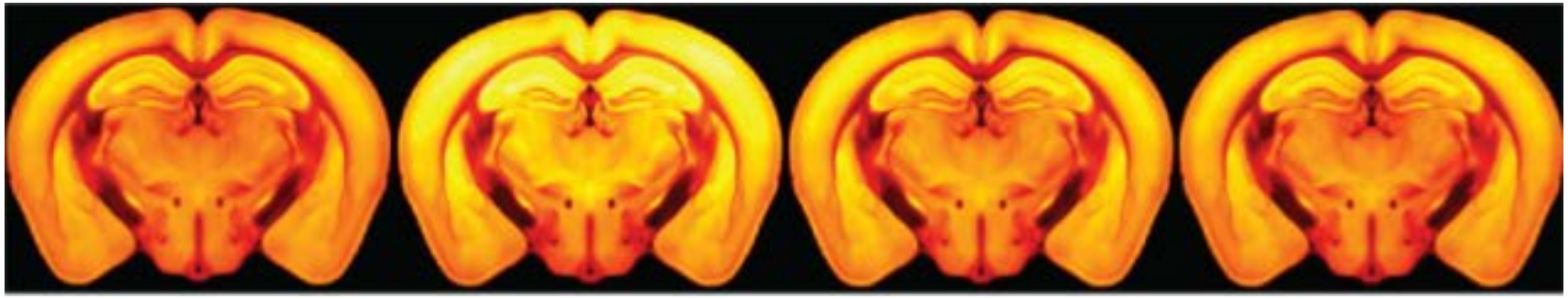
Real

Joint work with Ishan Deshpande, Alex Schwing, Ayis Pyrros, Nasir Siddiqui, RSNA 2018

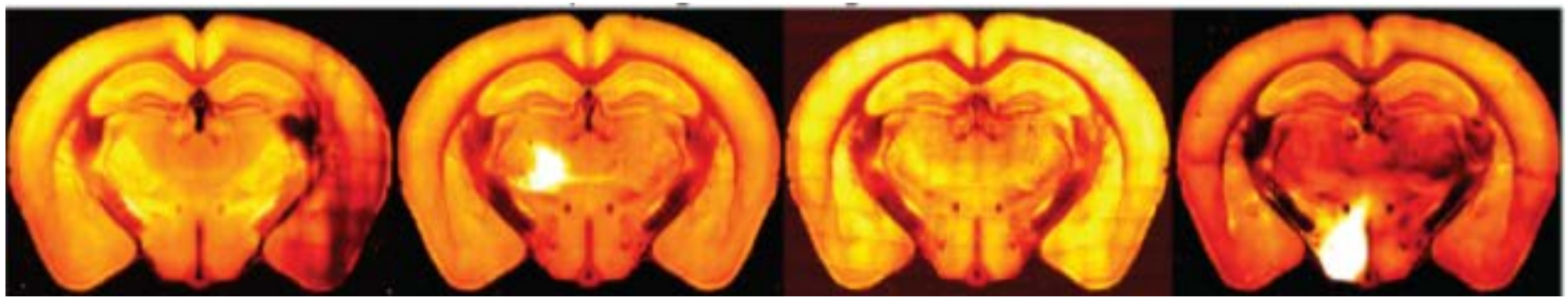


Experienced radiologists were asked to choose which of a real lung x-ray and a GAN generated image were real. Subjects favored real images slightly (on average GAN images were identified as real 39% of the time) but subject behavior varied widely. Size of blob identifies number of pairs viewed; note one subject preferred GAN images over 80% of the time, another could identify real images nearly exactly.

Joint work with Ishan Deshpande, Alex Schwing, Ayis Pyrros, Nasir Siddiqui, RSNA 2018

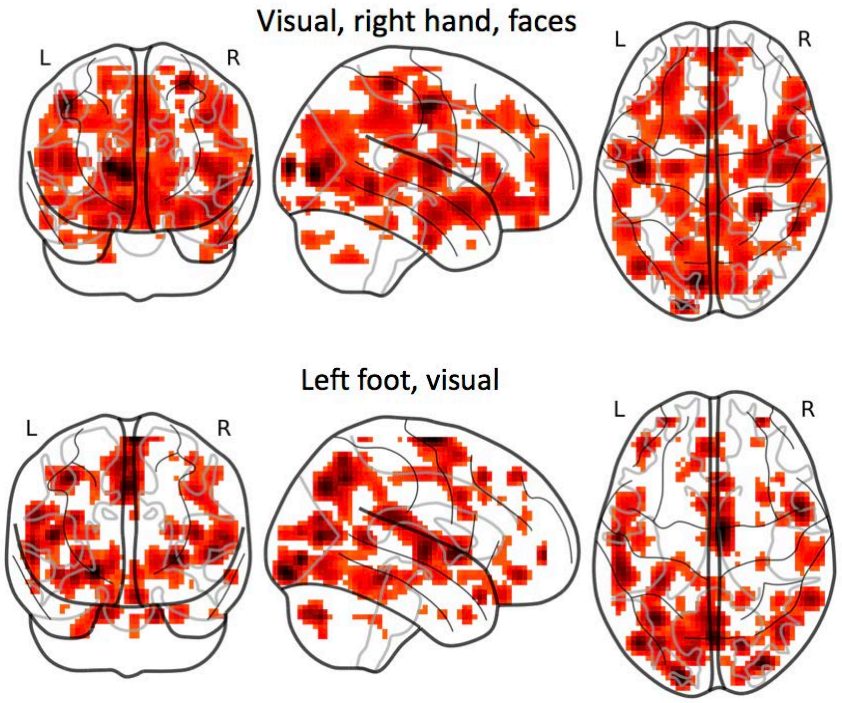


Synthetic

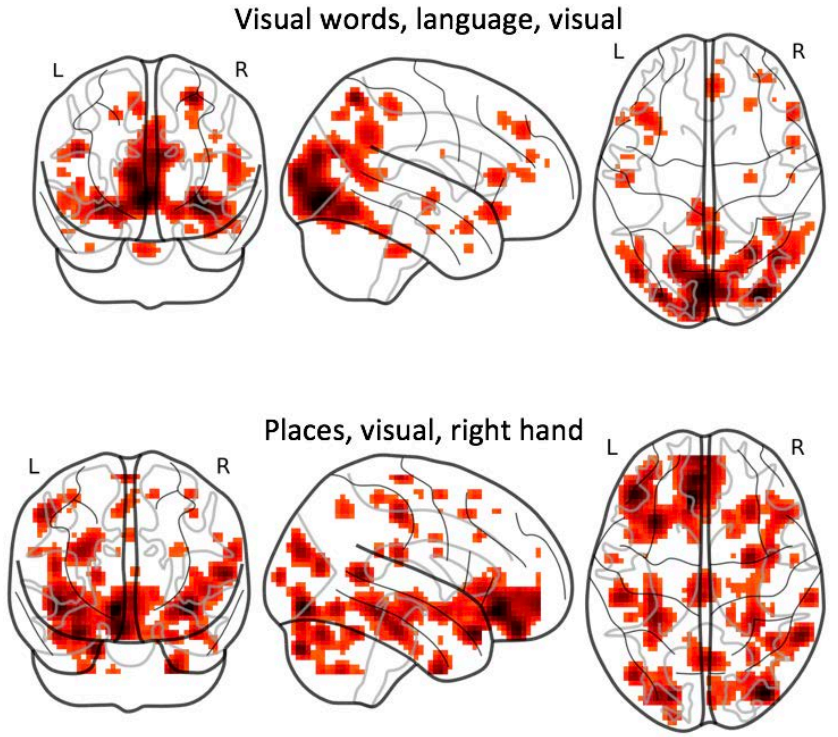


Real

Joint work with Cem Subakan, Maitham Naeemi, Julie A. Harris and Eva L. Dyer, Under review



Real



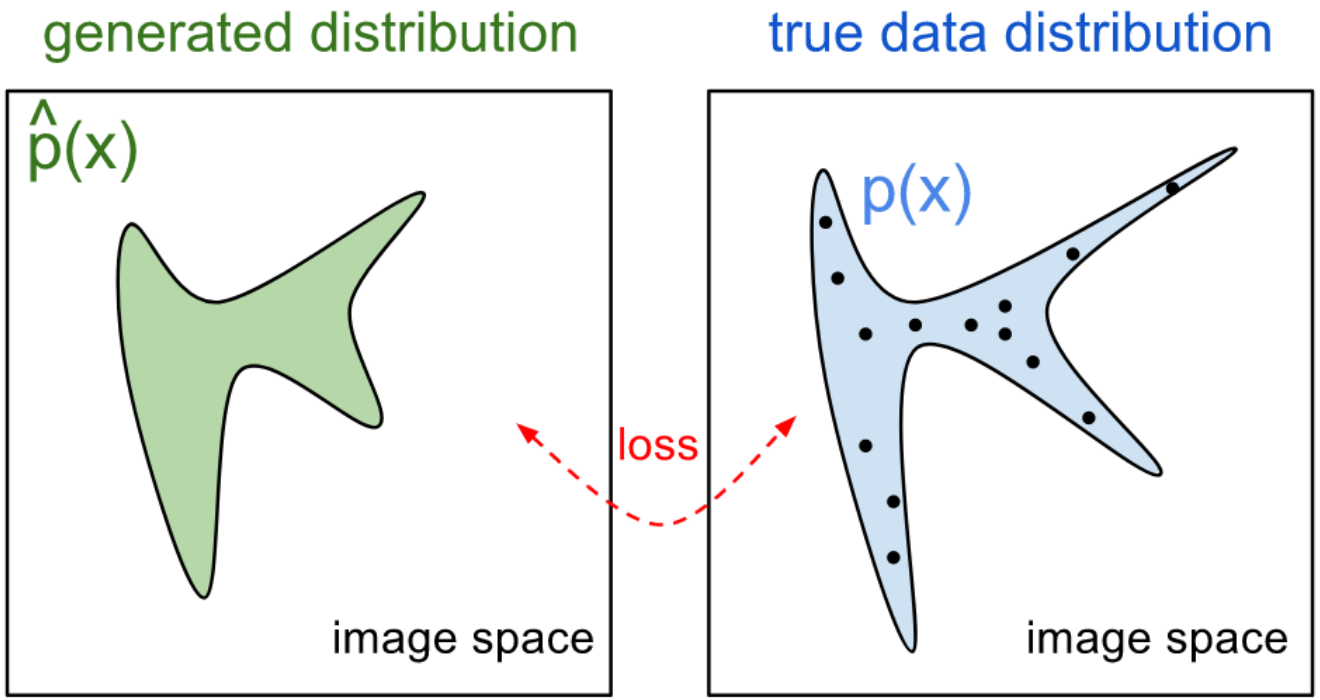
Synthetic

Joint work with Peiye Zhuang, Bliss Chapman, Ran Li, Alex Schwing, Under review

Outline

- Implicit generative models
- Cognitive neuroscience applications (synthesizing fMRI)
- Anticipated pitfalls
- Optional: Synthesizing structural images

Generative Models



Source: https://blog.openai.com/generative-models/

Low dimensions

Select a parametric family e.g. Gaussians

Find the parameters of the distribution e.g. using maximum likelihood

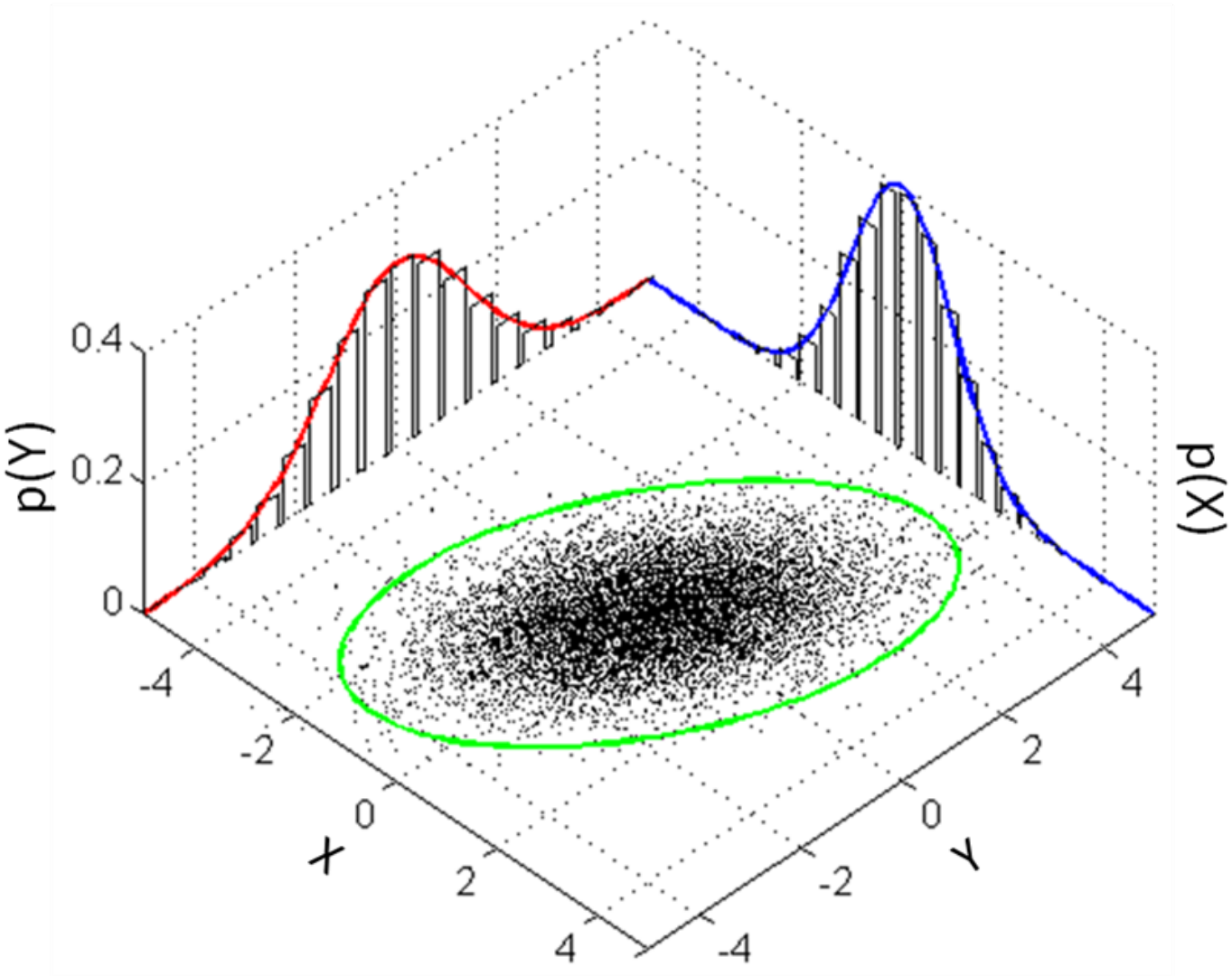
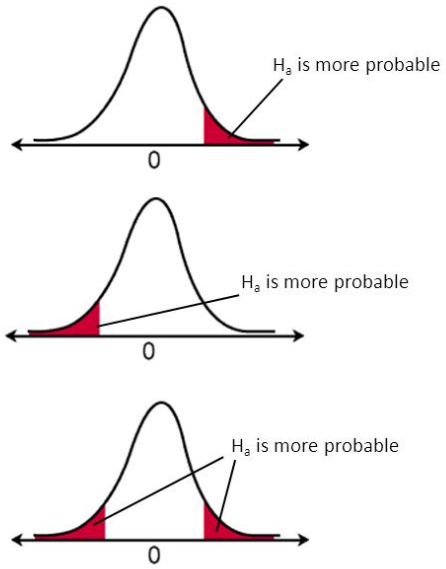


Image Source: Wikipedia

Generative models are everywhere

HYPOTHESIS TESTING



Right-tail test

$H_a: \mu > \text{value}$

Left-tail test

$H_a: \mu < \text{value}$

Two-tail test

$H_a: \mu \neq \text{value}$

*ALL OF STATISTICS

Estimation / variable selection

Prediction

Exploratory Modeling

Visualization

Image source: http://slideplayer.com/slide/9325599/

* with a few (model-free) exceptions

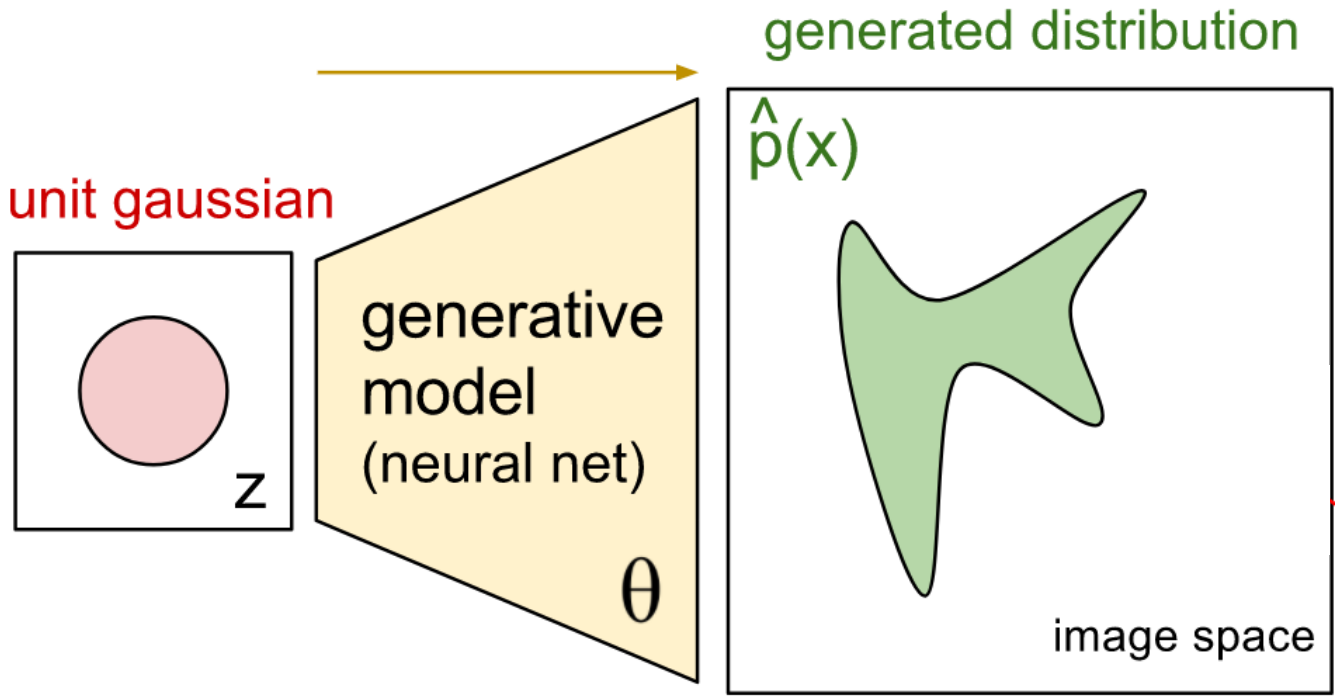
Generative models in high dimensions

None of the standard distributions / fitting methods work for high dimensional complex data.

Modern strategies:

- Variational Autoencoders (VAE): Kingma, Diederik P., and Max Welling. "Auto-encoding variational Bayes." *NIPS 2014*.
- Generative Adversarial Networks (GAN): Goodfellow, Ian, et al. "Generative adversarial nets." *NIPS* 2014.
- and many others…

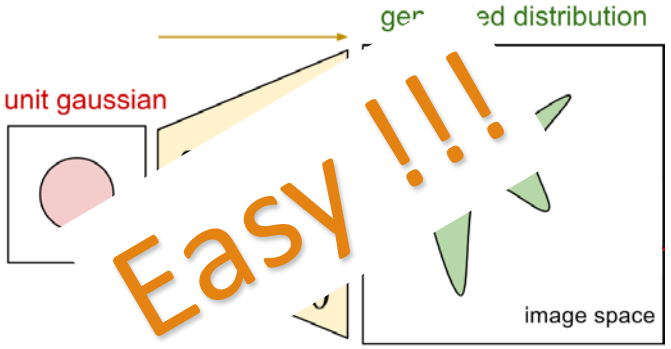
Implicit Generative Models



Source: https://blog.openai.com/generative-models/

Using the Implicit Generative Model

Synthesis
(Sampling)



$$z \sim \mathcal{N}(0,1);\ x = f_\theta(z)$$

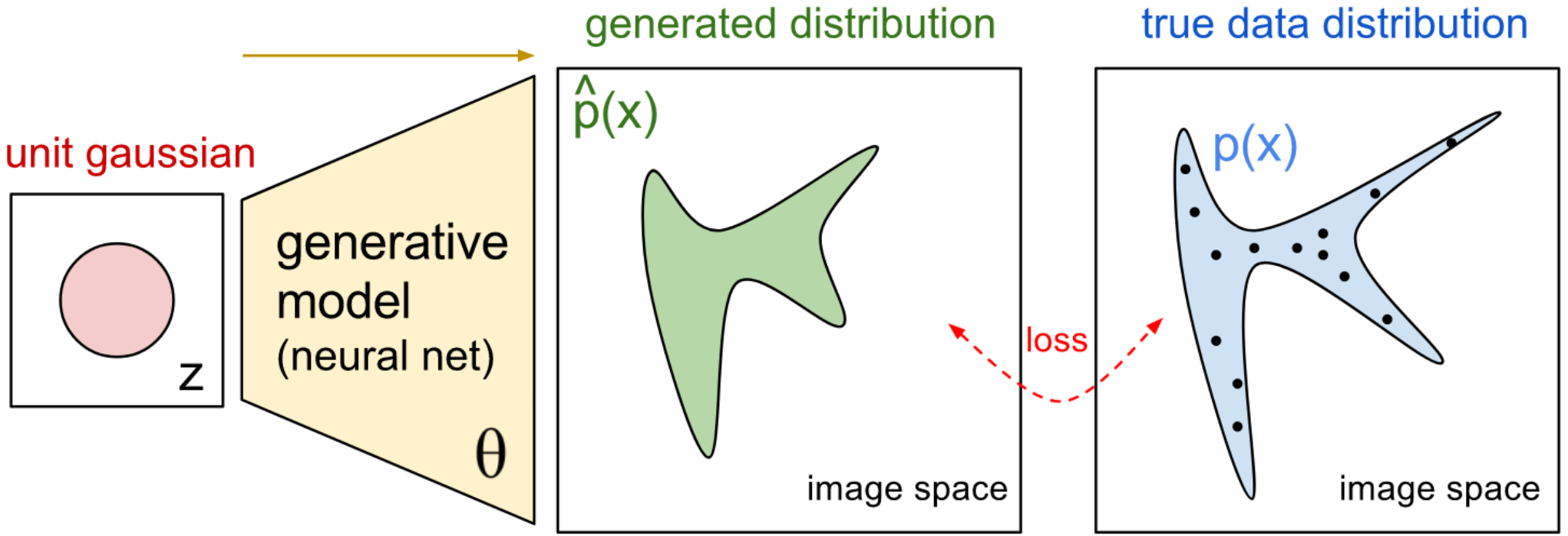
Scoring / fitting
(likelihood)

Hard!!!

$$\hat{p}(x) = \int_z \mathbf{1}_{[x=f_\theta(z)]}\mathcal{N}(0,1)dz$$

Source: https://blog.openai.com/generative-models/

Estimating the model (i.e. neural net)



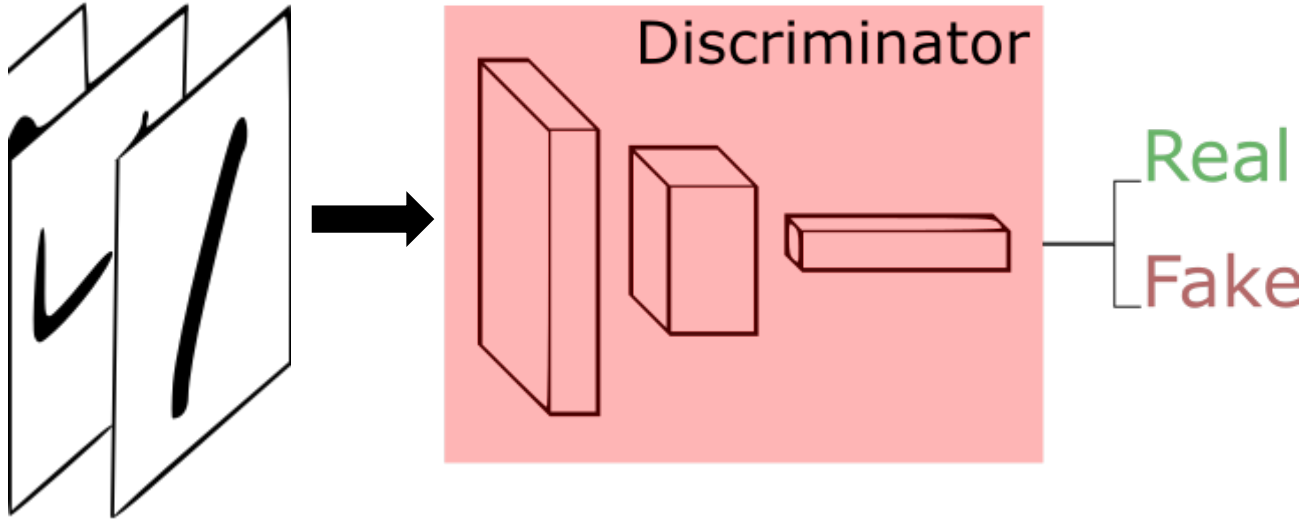
Source: https://blog.openai.com/generative-models/

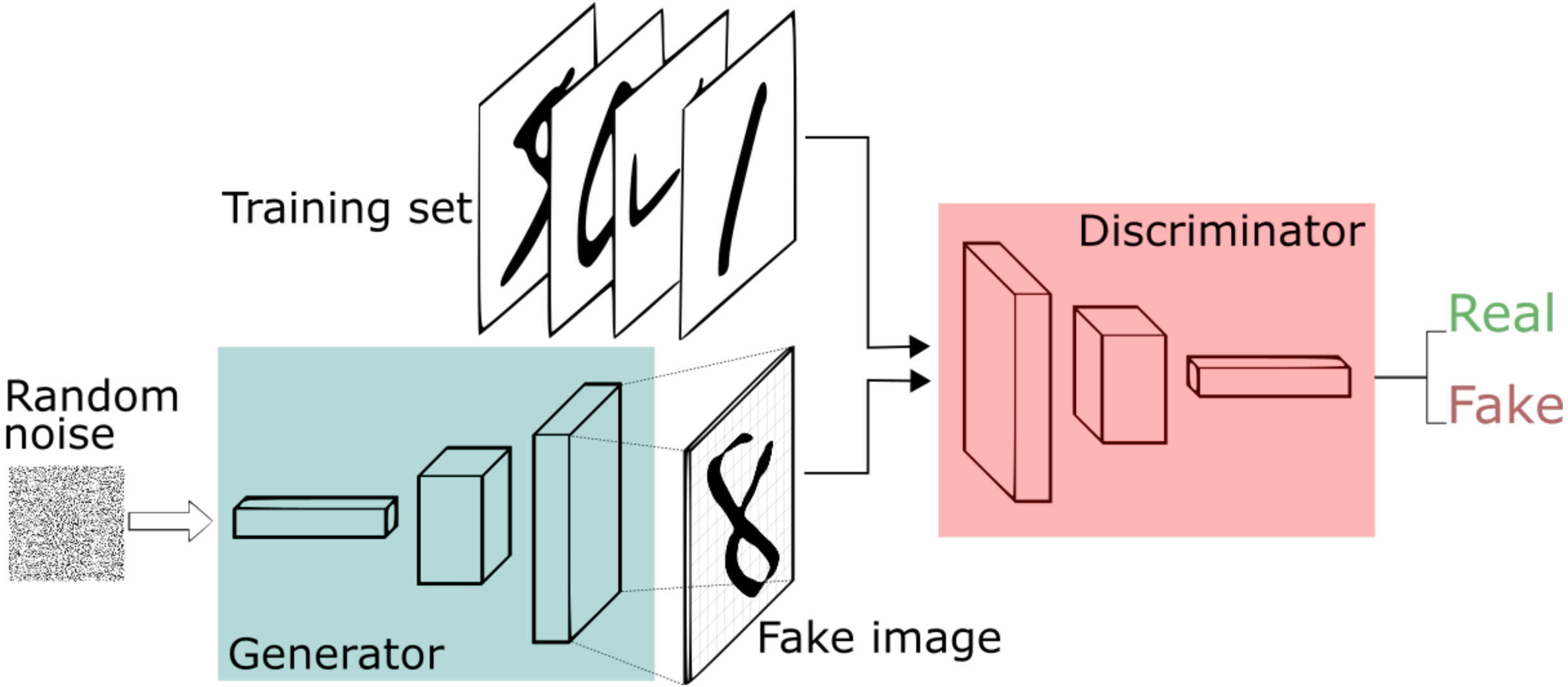
Use a classifier to “learn” a loss function

Instead of a standard loss function, the GAN uses a classifier to determine if the generated data “looks” real or fake

Interpreted as a two-player game:

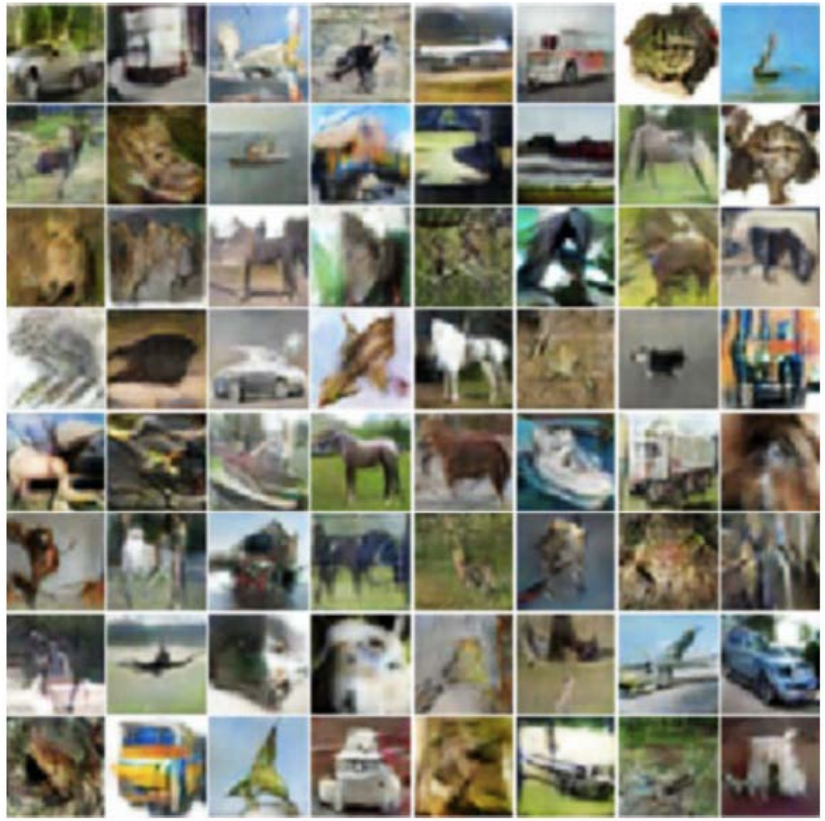
- Generator tries to fool the classifier
- Discriminator tries to avoid getting fooled





https://deeplearning4j.org/generative-adversarial-network

The strategy works!



2014

2016

Salimans, Tim, et al. "Improved techniques for training gans." *NIPS* 2016.

State of the art: GANs in 2018

Karras, T., et al. “Progressive growing of gans for improved quality, stability, and variation.” ICLR 2018.

GANs are popular in ML and AI

What are some recent and potentially upcoming breakthroughs in deep learning?

Yann LeCun, Director of AI Research at Facebook and Professor at NYU

Answered Jul 28, 2016 · Upvoted by Joaquin Quiñonero Candela, studied Machine Learning and Gokul Krishnan, M.Sc Computer Science & Machine Learning, ETH Zurich (2018)

There are many interesting recent development in deep learning, probably too many for me to describe them all here. But there are a few ideas that caught my attention enough for me to get personally involved in research projects.

The most important one, in my opinion, is adversarial training (also called GAN for Generative Adversarial Networks). This is an idea that was originally proposed by Ian Goodfellow when he was a student with Yoshua Bengio at the University of Montreal (he since moved to Google Brain and recently to OpenAI).

35 Innovators Under 35

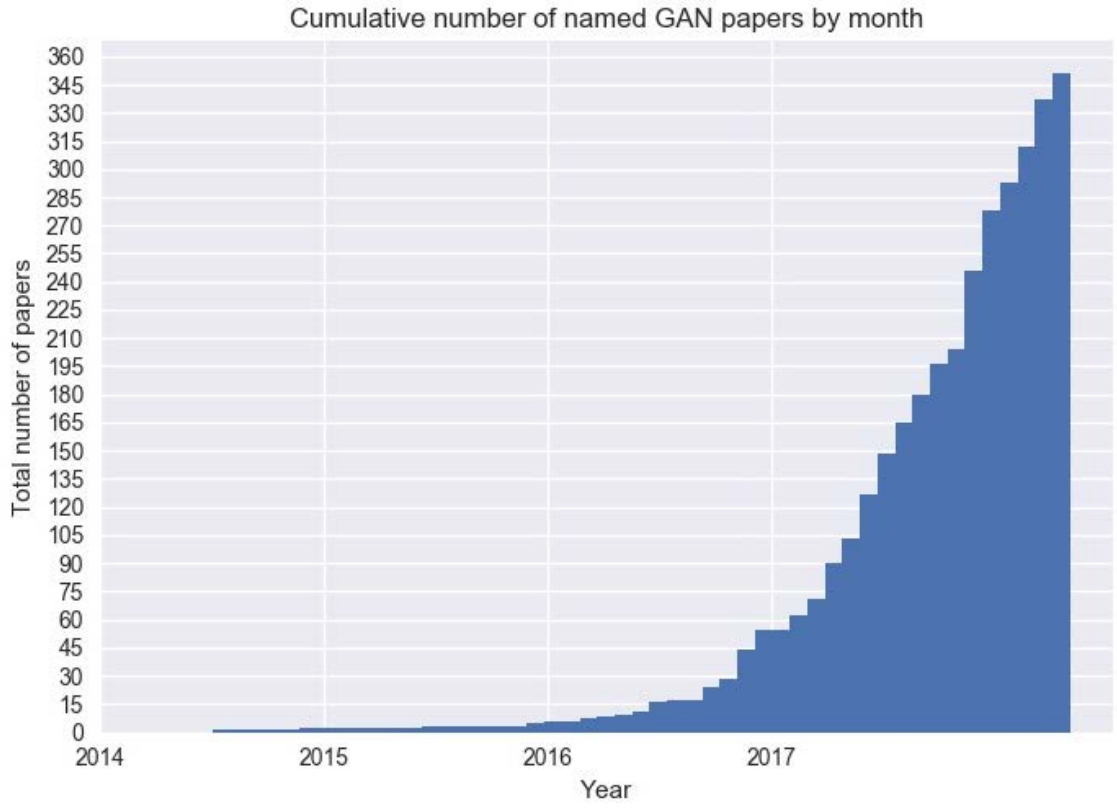
2017

Ian Goodfellow, 31

Google Brain Team

Invented a way for neural networks to get better by working together.

Lots of innovation



- 3D-ED-GAN—Shape Inpainting using 3D Generative Adversarial Network and Recurrent Convolutional Networks
- 3D-GAN—Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling (github)
- 3D-IWGAN—Improved Adversarial Systems for 3D Object Generation and Reconstruction (github)
- 3D-RecGAN—3D Object Reconstruction from a Single Depth View with Adversarial Learning (github)
- ABC-GAN—ABC-GAN: Adaptive Blur and Control for improved training stability of Generative Adversarial Networks(github)
- ABC-GAN—GANs for LIFE: Generative Adversarial Networks for Likelihood Free Inference
- AC-GAN—Conditional Image Synthesis With Auxiliary Classifier GANs

Source: https://deephunt.in/the-gan-zoo-79597dc8c347

GANs for fMRI

Joint work with Peiye Zhuang, Bliss Chapman, Ran Li, Alex Schwing, Under review

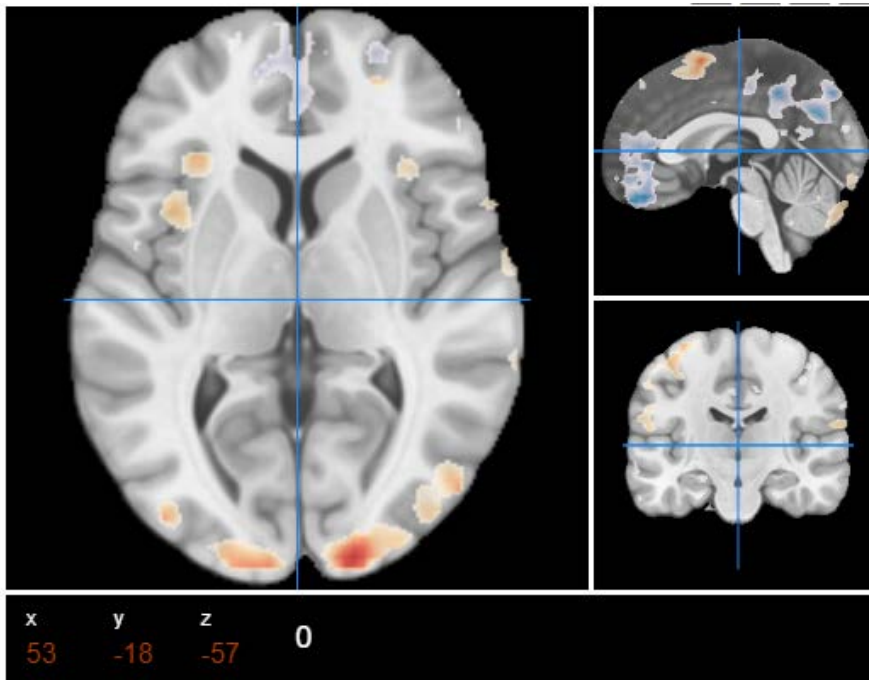
BrainPedia

Contributed by bthirion

NEUROVAULT

Show 10 entries Search:

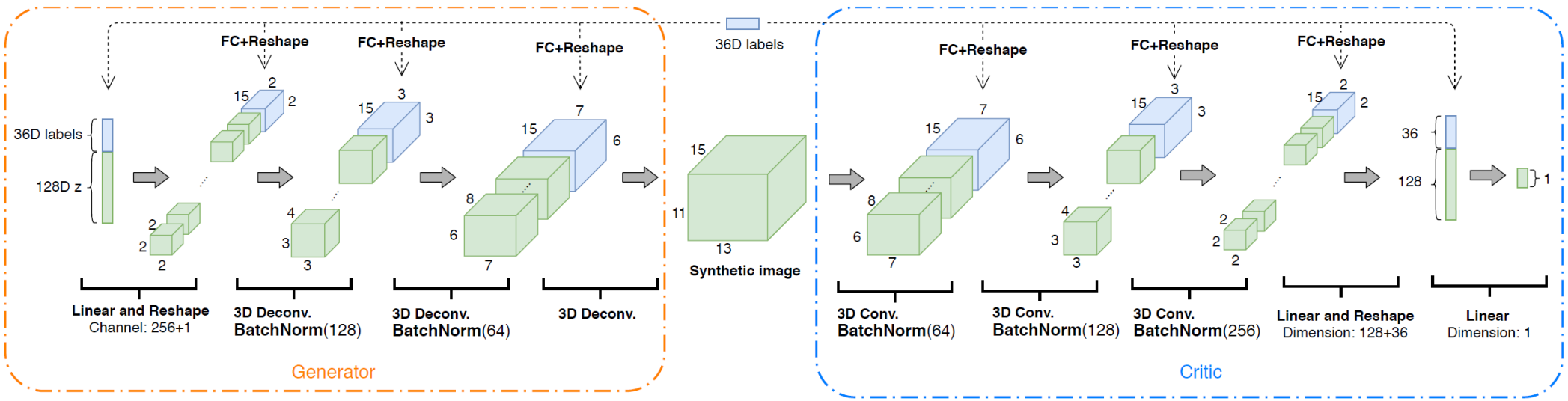
Field	Value
Add Date	Oct. 26, 2016, 7:31 p.m.
Authors	None
Contributors	
Description	BrainPedia is a collection of SPMs obtained from about 30 protocoles from OpenfMRI, the Human Connectome Project and Neurospin research center that map a wide set of cognitive functions.
DOI	None
Field Strength	None
id	1952
Journal	None



6573 z-score task contrasts

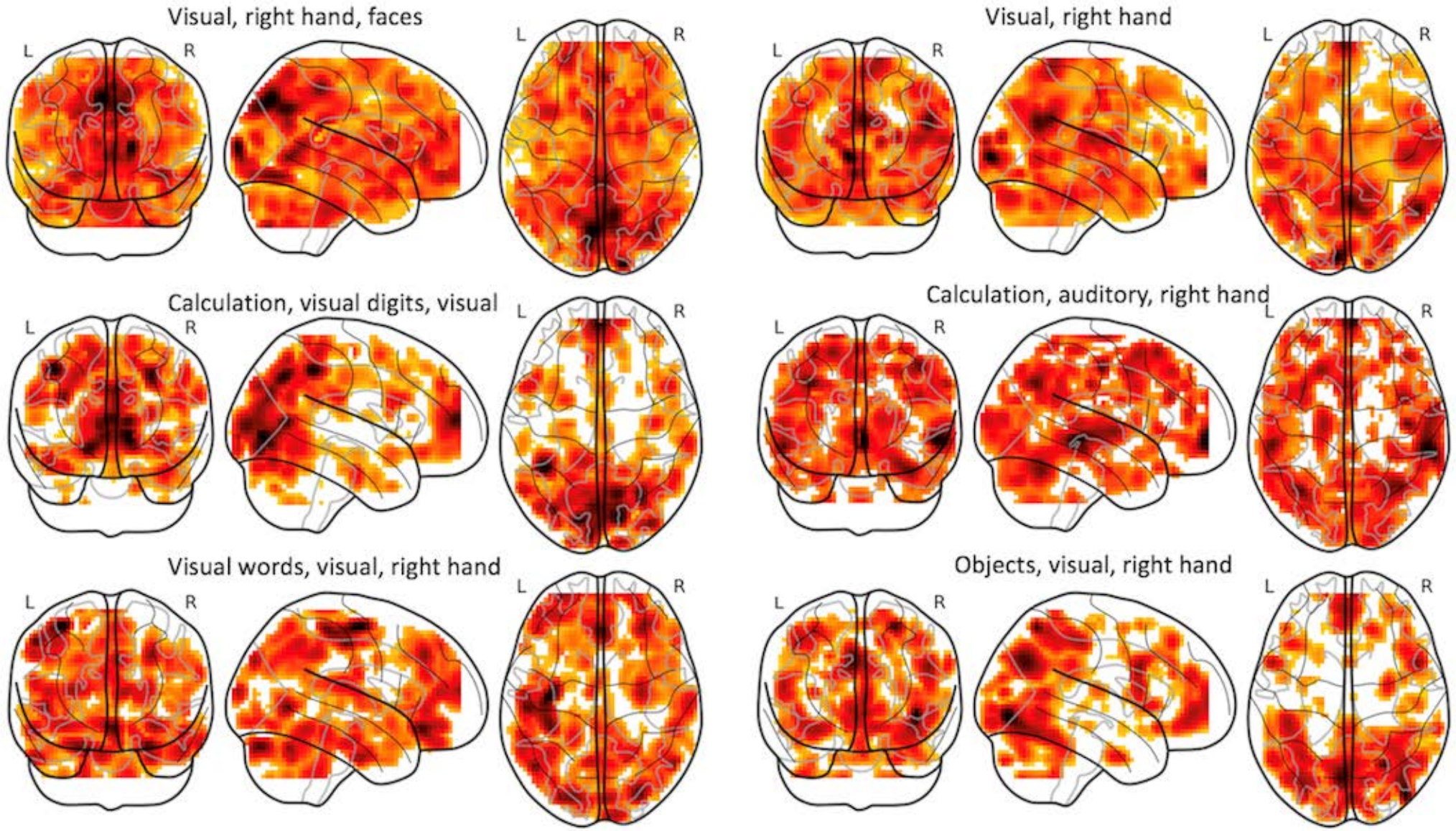
45 total tasks

19 cognitive labels

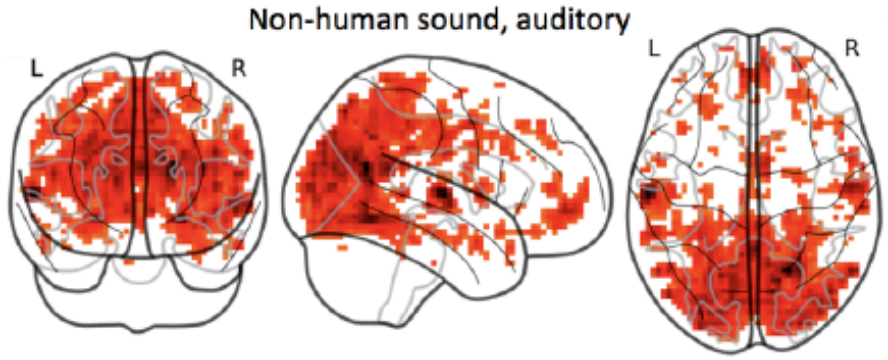


$$z \sim \mathcal{N}(0,1);\ x|k = f_\theta(z,k) \qquad D : \hat{x}, k \mapsto \{\text{real},\ \text{fake}\}$$

3-D Conditional GAN i.e. using known labels


Visual, right hand, faces
Visual, right hand
Calculation, visual digits, visual
Calculation, auditory, right hand
Visual words, visual, right hand
Objects, visual, right hand
L
R

Evaluation using Neurosynth decoder



Analysis	Correlation
auditory cortex	0.207
auditory	0.198
heschi	0.195
heschl gyrus	0.194
sounds	0.190
fractional anisotropy	0.189
fa	0.188
pitch	0.187
anisotropy fa	0.186
anisotropy	0.185

Cognitive Neuroscience Applications

Data augmentation for decoding

Two-sample testing: Determine sample size / power for new experiments

One sample testing: Determine if an individual differs from the population

Realistic data source for fitting complex biologically plausible models

…

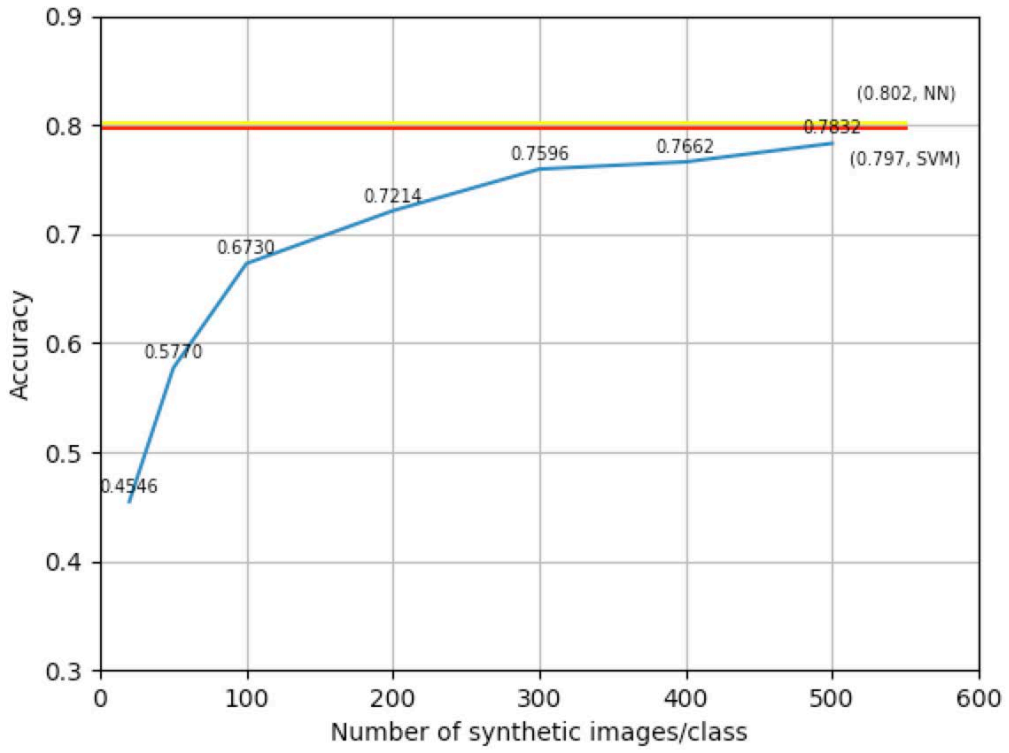
Application: Classifier data augmentation

Downsampling	Input	Classifier	Accuracy	Macro F1	Precision	Recall
4.0×	Real	SVM	0.797	0.797	0.813	0.797
	Real	NN	0.802	0.802	0.817	0.802
	Real+Synth.	SVM	0.806	0.803	0.823	0.807
	Real+Synth.	NN	**0.819**	**0.817**	**0.830**	**0.819**
2.0×	Real	SVM	0.855	0.857	0.867	0.857
	Real	NN	0.863	0.863	0.872	0.863
	Real+Synth.	SVM	0.860	0.863	0.860	0.857
	Real+Synth.	NN	**0.891**	**0.894**	**0.906**	**0.891**

Comparison to traditional generative models

Training data	Accuracy	F1	Precision	Recall
Synth. data from GMM (20 images/class)	0.203	0.309	0.309	0.202
Synth. data from GMM (500 images/class)	0.720	0.725	0.765	0.720
Real+Synth. (from GMM)	0.793	0.798	0.824	0.793
Synth. data from ICW-GAN (20 images/class)	**0.458**	**0.433**	**0.537**	**0.458**
Synth. data from ICW-GAN (500 images/class)	**0.783**	**0.776**	**0.805**	**0.783**
Real+Synth. (from ICW-GAN)	**0.819**	**0.817**	**0.830**	**0.819**

Construct classifier using only generated data



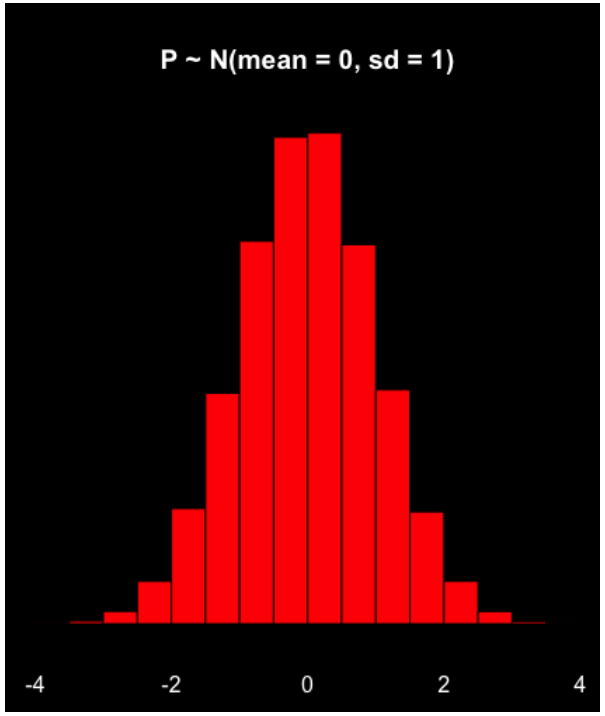
Subset of data used to train generative model

Synthesized data is used to train a classifier

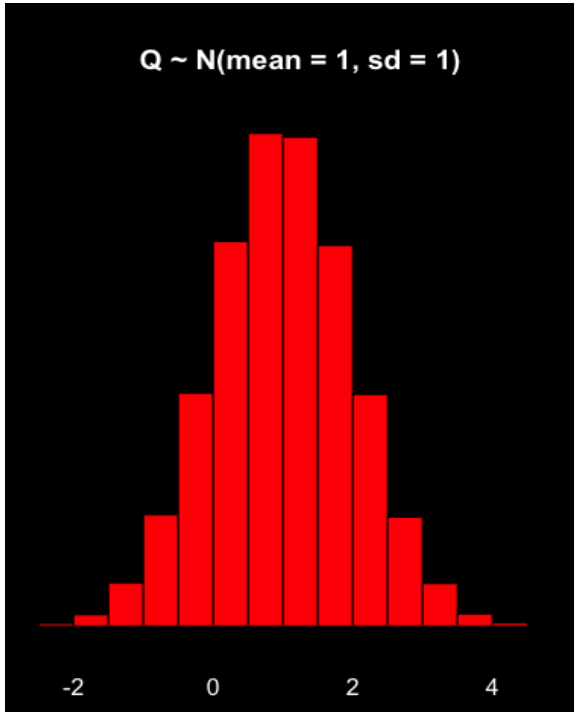
Classifier is applied to test set

Hypothesis Testing with Synthetic Data

Review: Two-Sample Testing



[1.197979366 -0.238607756 0.401953120 0.725501957 …
1.235171756 0.045692089 0.314043529 0.825958931]



[0.155670610 2.924219170 1.011259692 2.104092724 …
2.287141599 1.936105587 -0.548633905 1.711112503]

$$H_0: P = Q$$
$$H_1: P \neq Q$$

$$\bar{x} = \frac{1}{N}\sum_{i=1} x_i$$

$$s^2 = \frac{\sum_{i=1} (x_i - \bar{x})^2}{N - 1}$$

$$t = \frac{\overline{x_P} - \overline{x_Q}}{\sqrt{\frac{s_P^2}{N_P} + \frac{s_Q^2}{N_Q}}}$$

P

[1.197979366 -0.238607756 … 0.314043529 0.825958931]

$= 0.00272$

$s_P^2 = 1.00054$

Q

[0.155670610 2.104092724 … 2.287141599 1.711112503]

$= 1.00212$

$s_Q^2 = 1.00288$

$$t = -224.27$$
$$p = P(t < -224.27)$$

$p < \alpha = 0.05$ ➡ $P \neq Q$

Estimating power for new experiments

Higher power =>

more likely to correctly reject the null hypothesis =>

“more likely to detect an effect when it exists”

Data collection is expensive !!!

Want to avoid underpowered studies -- wasting time/money on low power experiments

Also want to avoid wasting resources to collect more data than required

Usually aim for smallest sample size needed to guarantee say 80% power

Current approach for power estimation

Power estimation can be challenging!

- requires estimates of null and alternative distributions
- These are often estimated from the literature, previous experiments, pilot study, or guessed

Pilot study:

- Statistically designed to decide if an experiment is feasible (not for power estimation)
- Usually very small sample sizes, extremely noisy estimate

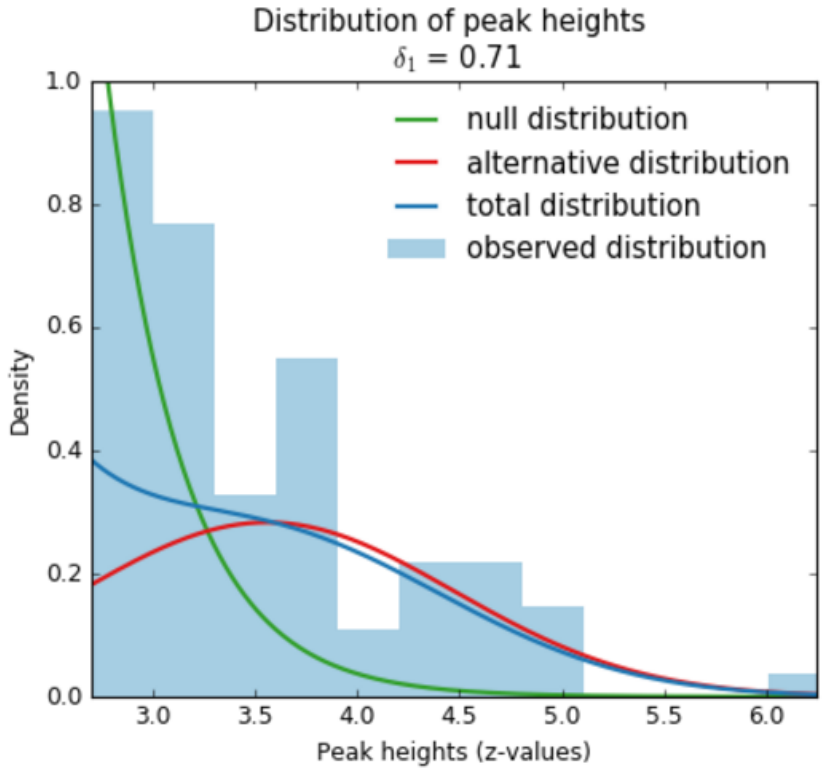
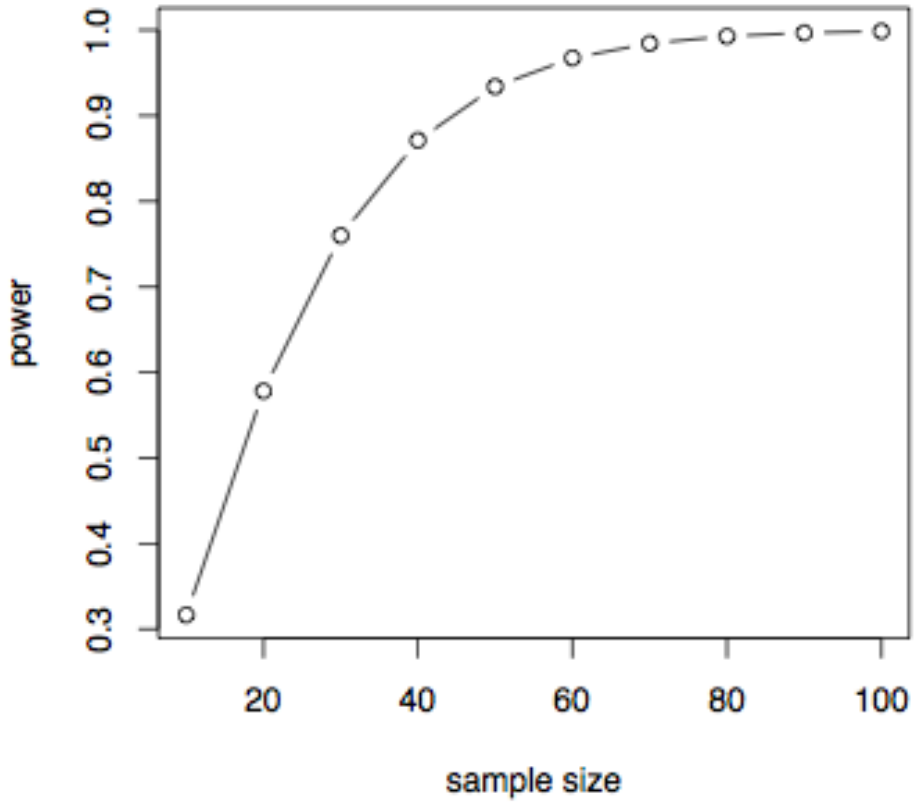


Figure : http://neuropowertools.org/

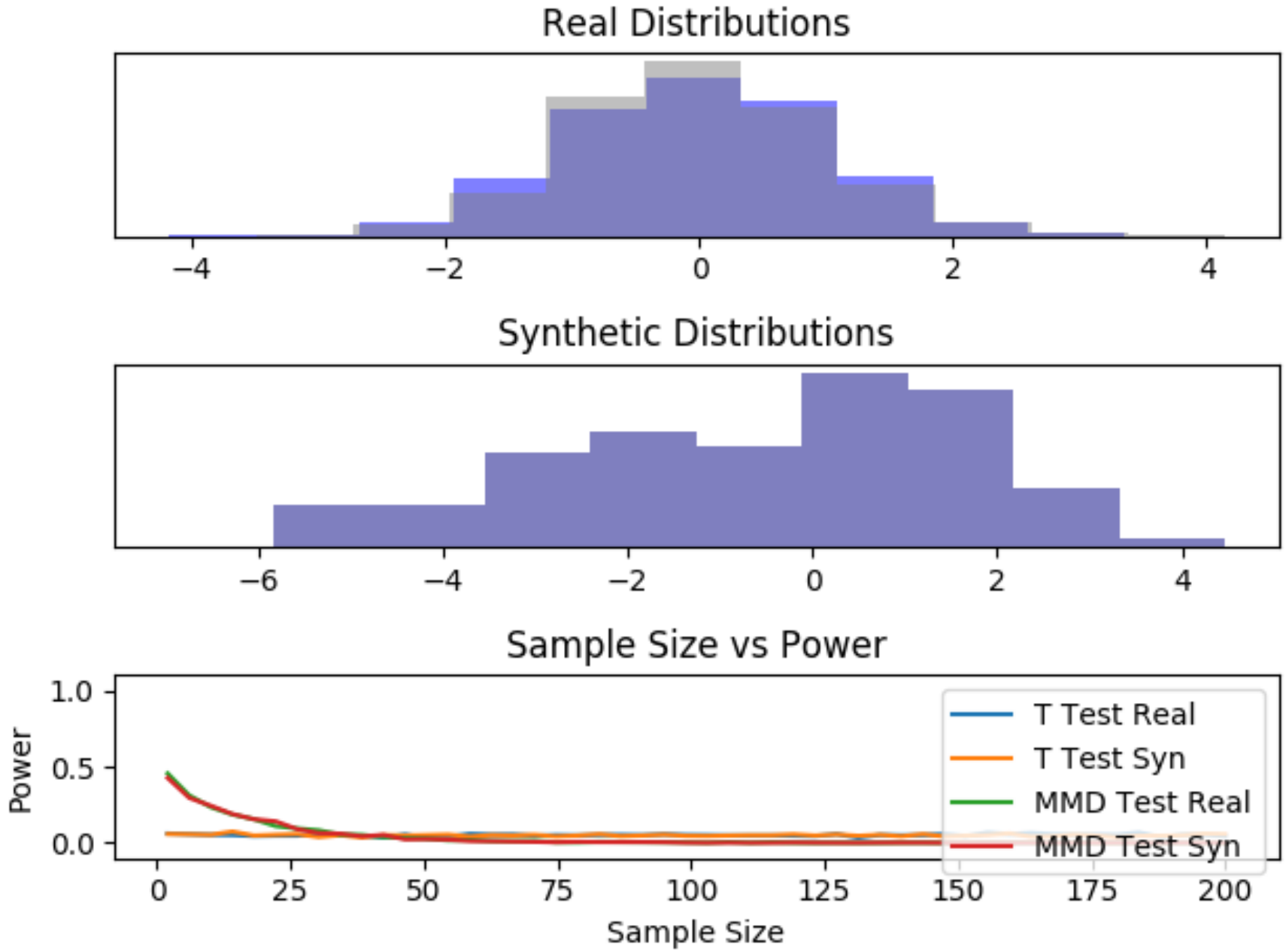
Can synthetic data be used to conduct power analyses?

Estimating statistical power

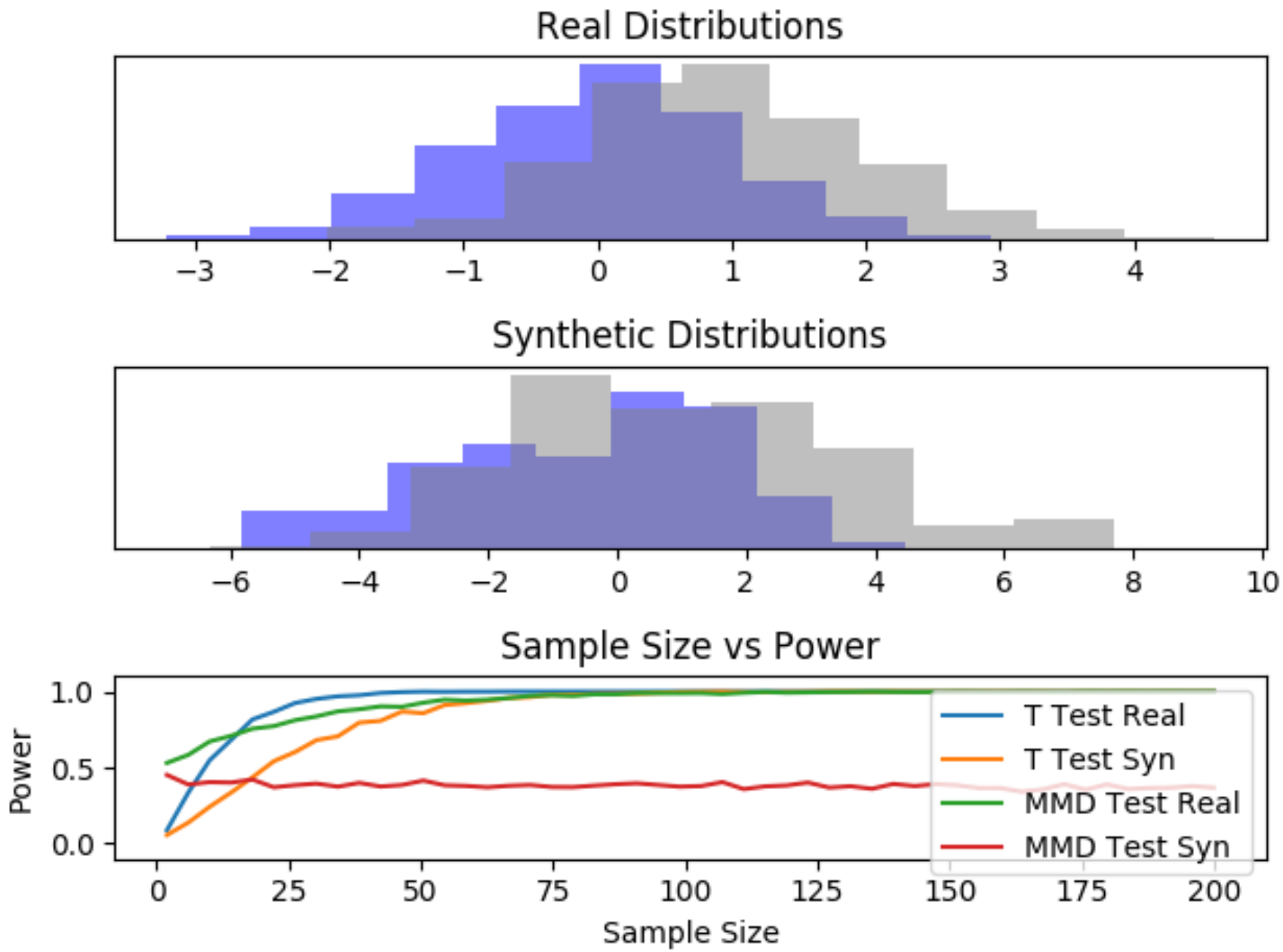
1. Synthesize data for fixed sample size
2. Compute p-value
3. Repeat N-times
4. $$\boldsymbol{Power} \approx \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i}=\mathbf{1}} \mathbf{1}(\boldsymbol{p_i} < \boldsymbol{\alpha})$$



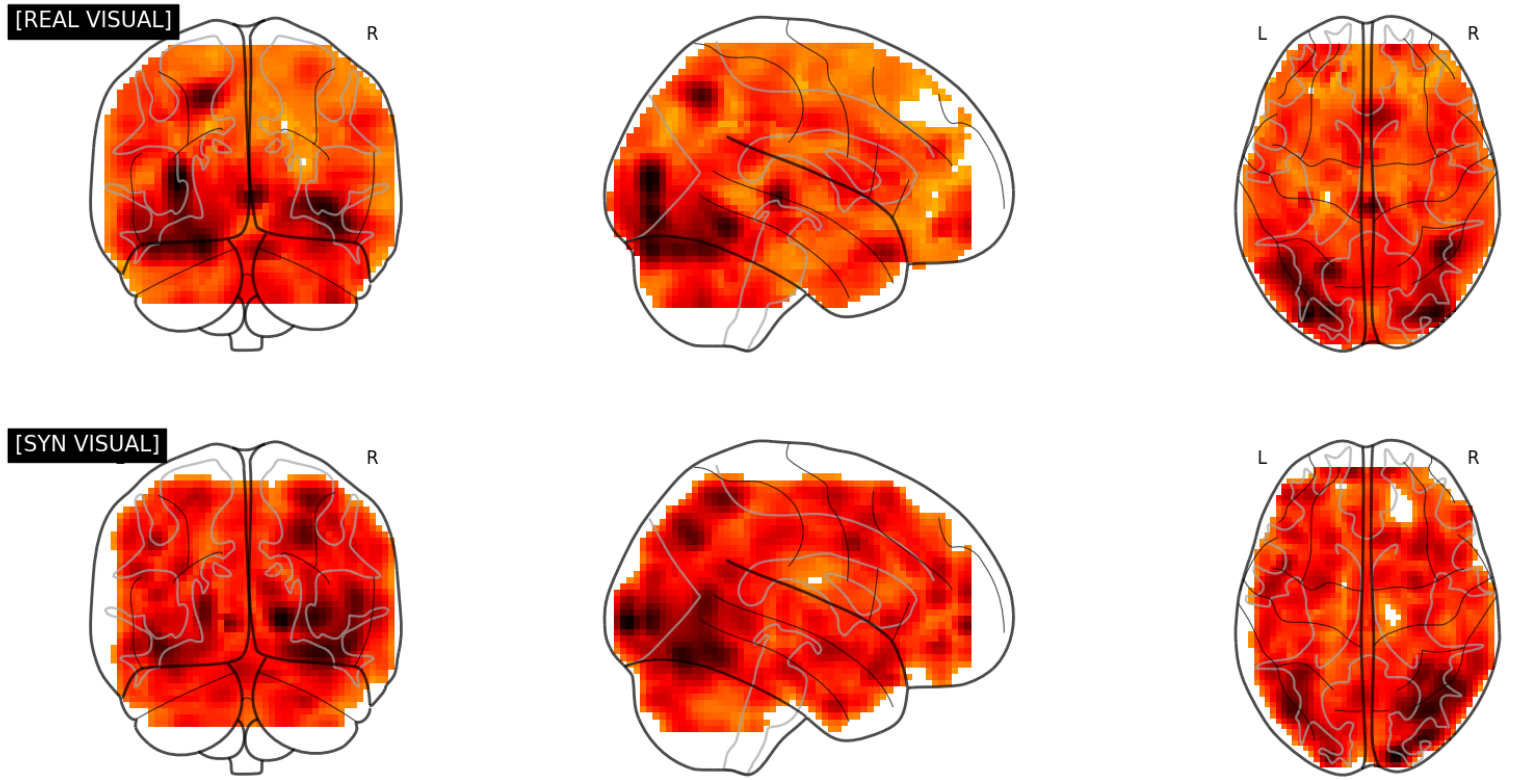
No effect i.e.
Null is True

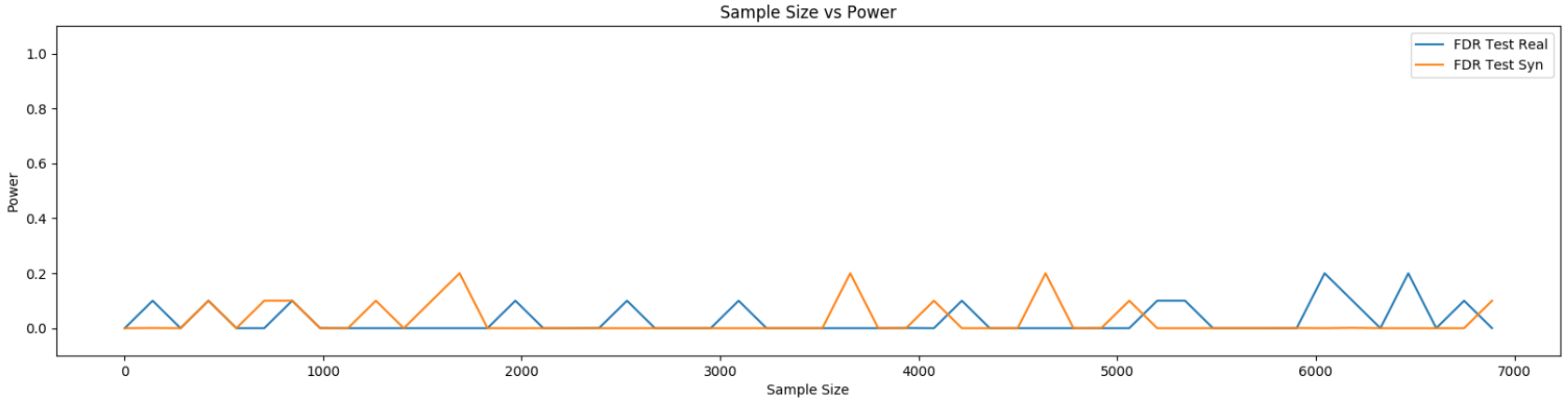


Yes effect i.e.
Null is False



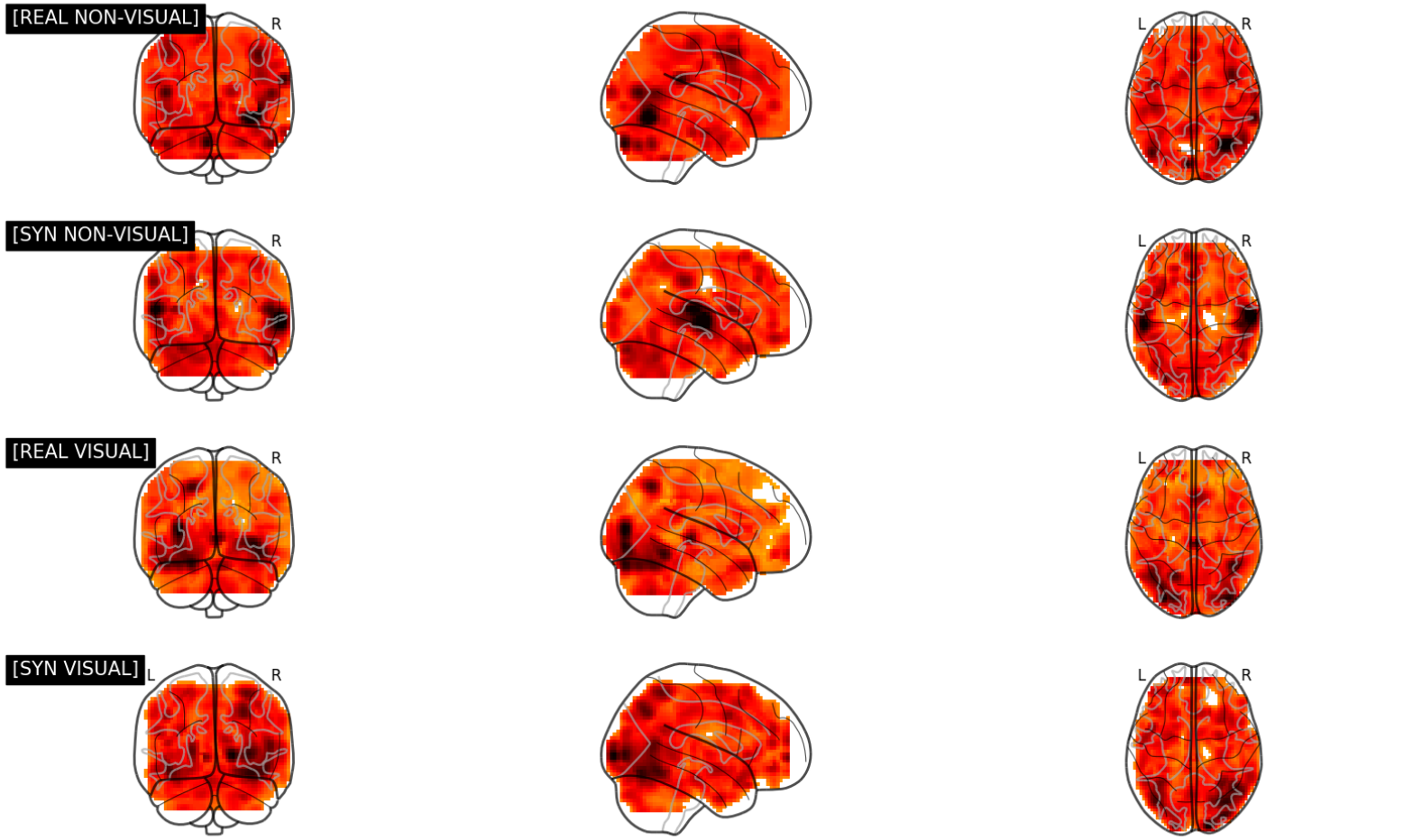
No effect i.e. Null is True

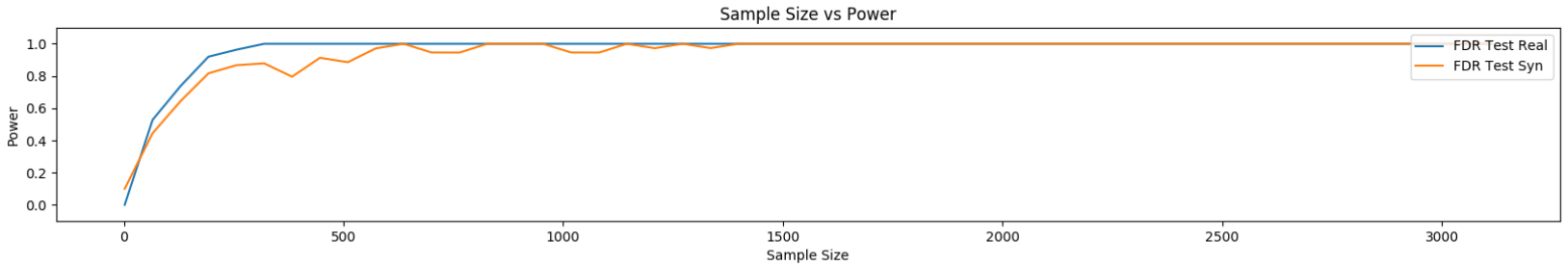




**Power computed voxel-wise after FDR correction

Yes effect i.e. Null is False





**Power computed voxel-wise after FDR correction

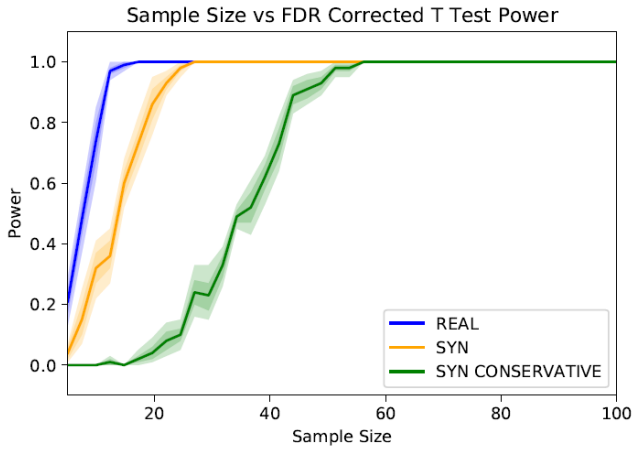


Figure 6: 'Visual' vs NO 'Visual'

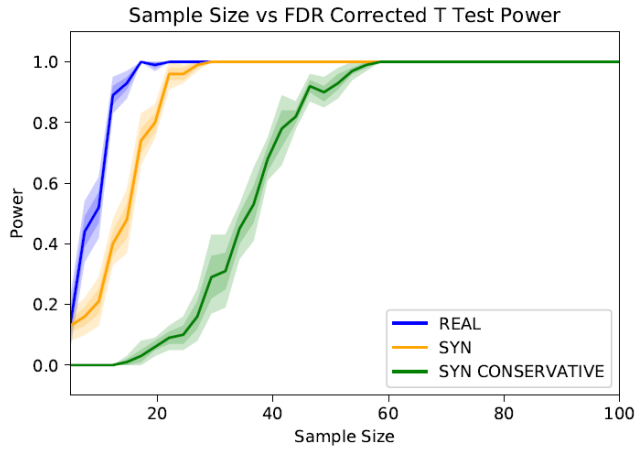


Figure 7: 'Auditory' vs NO 'Auditory'

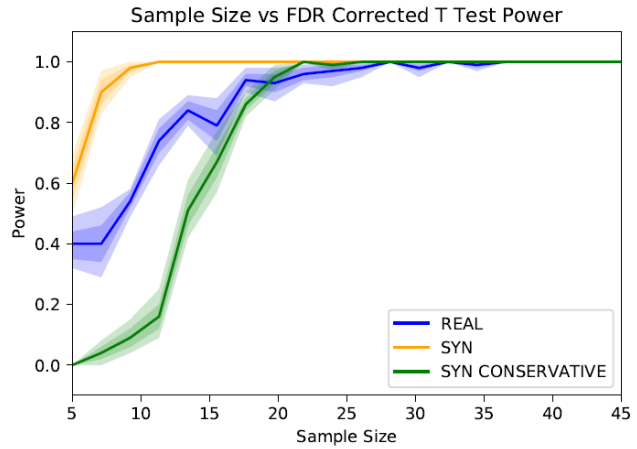


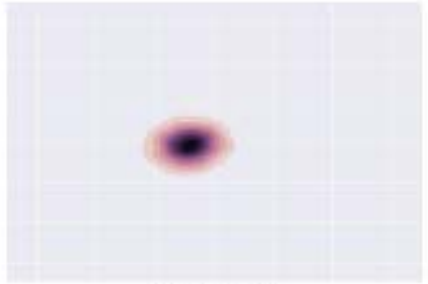
Figure 8: 'Visual, Auditory' vs NO 'Visual, Auditory'

Can use this to synthesize power curves for new experiments i.e. label combinations not observed during training

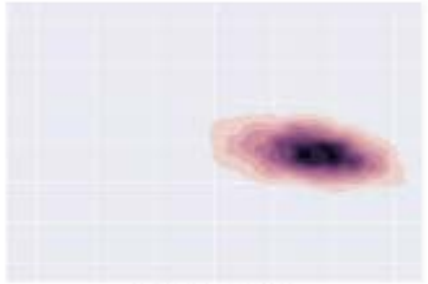
Pitfalls

GANs are susceptible to “Mode Collapse”

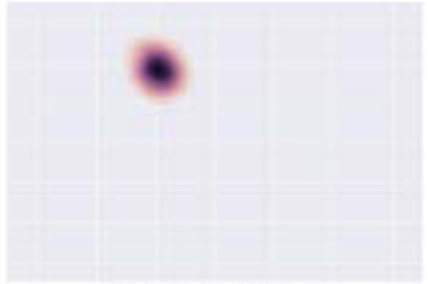
Target



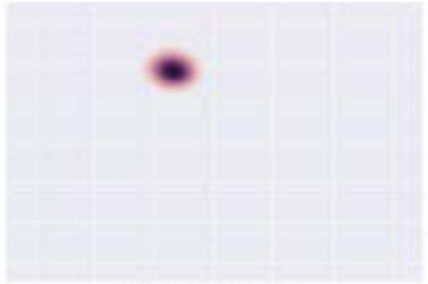
Step 0



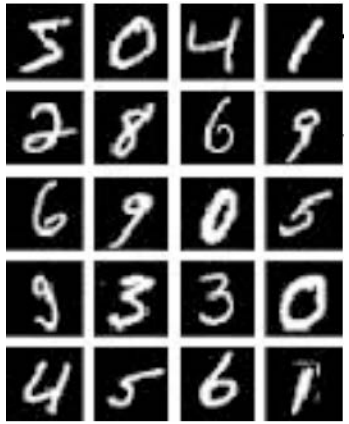
Step 5k



Step 10k



Step 15k



MNIST dataset

Implicit Generative Models are difficult to evaluate

A NOTE ON THE EVALUATION OF GENERATIVE MODELS

Lucas Theis*
University of Tübingen
72072 Tübingen, Germany
lucas@bethgelab.org

Matthias Bethge
University of Tübingen
72072 Tübingen, Germany
matthias@bethgelab.org

and other tasks. Given this wide range of applications, it is not surprising that a lot of heterogeneity exists in the way these models are formulated, trained, and evaluated. As a consequence, direct comparison between models is often difficult. This article reviews mostly known but often underappreciated properties relating to the evaluation and interpretation of generative models with a focus on image models. In particular, we show that three of the currently most commonly used criteria—average log-likelihood, Parzen window estimates, and visual fidelity of samples—are largely independent of each other when the data is high-dimensional. Good performance with respect to one criterion therefore need not imply good performance with respect to the other criteria. Our results show that extrapolation from one criterion to another is not warranted and generative models need to be evaluated directly with respect to the application(s) they were intended for. In addition, we provide examples demonstrating that Parzen window

Theis, et al. "A note on the evaluation of generative models." *ICLR 2016*

Synthesis is not always realistic

Karras, T., et al. “Progressive growing of gans for improved quality, stability, and variation.” ICLR 2018.

More failure cases

Karras, T., et al. “Progressive growing of gans for improved quality, stability, and variation.” ICLR 2018.

Summary of pitfalls

GANs are susceptible to mode collapse

High dimensional generative models are difficult to evaluate

GANs are hard to visualize and explain

Likelihood function is intractable, thus new methods are required for standard probabilistic tasks e.g. hypothesis testing

Model training is compute-intensive

Frequently Asked Questions (FAQ)

What if I don’t like GANs? (because my awesome new generative model is much better)

We use GANs, but that’s not the point. The goal is to improve fMRI generative modeling, and explore how the resulting models can be used.

Frequently Asked Questions (FAQ)

GANs are usually trained using hundreds of thousands of images, how do you get away with ~6000?

Observation and experiments suggest that (registered, healthy) fMRI are much less complex than natural images:

- Biology enforces strong constraints on the data
- Registration removes most of the spatial variability

Frequently Asked Questions (FAQ)

Why should I trust your model?

This is ongoing work (both for us, and for the larger ML community). We suggest a variety of tests. Nature of the tests should depend on proposed application.

I want to try this myself! How much computational power do I need?

Deep learning is compute heavy, I suggest a platform with modern GPU. Cloud computing (e.g. Amazon, Azure, Google) is a good option.

Frequently Asked Questions (FAQ)

I want to try this myself! Give me the code?

GitHub:

- Synthetic Statistics: https://github.com/BlissChapman/SyntheticStatistics
- fMRI GAN: https://github.com/BlissChapman/ICW-fMRI-GAN/

Synthesizing Structural fMRI

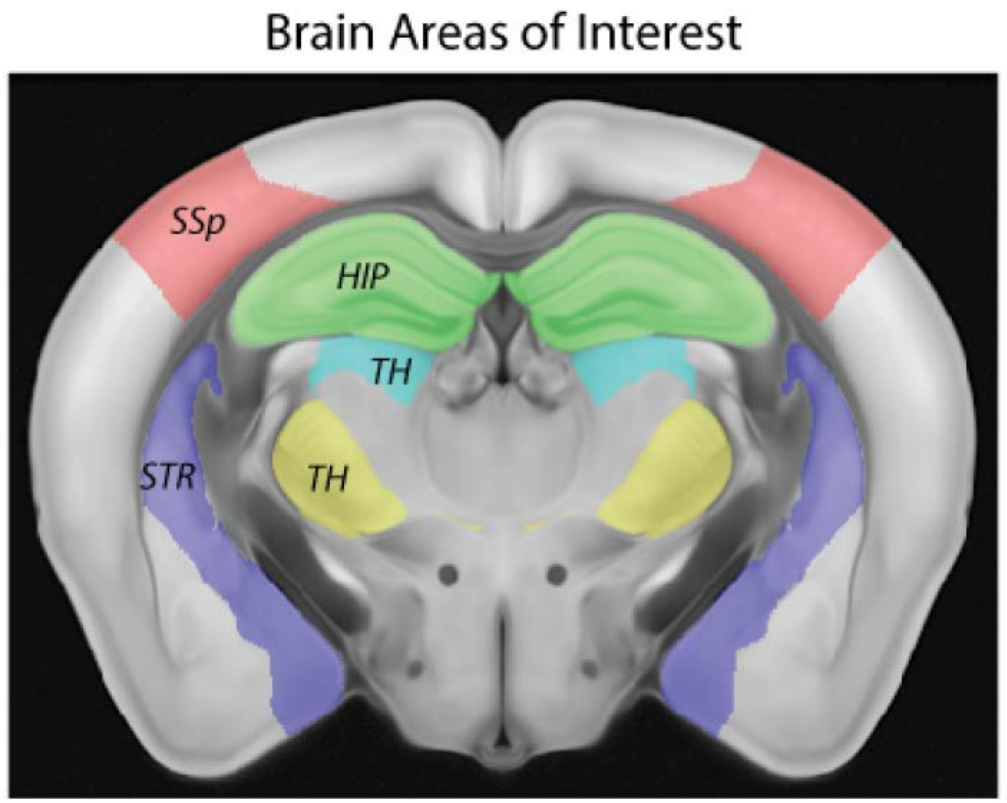
Joint work with Cem Subakan, Maitham Naeemi, Julie A. Harris and Eva L. Dyer, Under review

1,714 images from the Allen Institute for Brain Science's (AIBS)

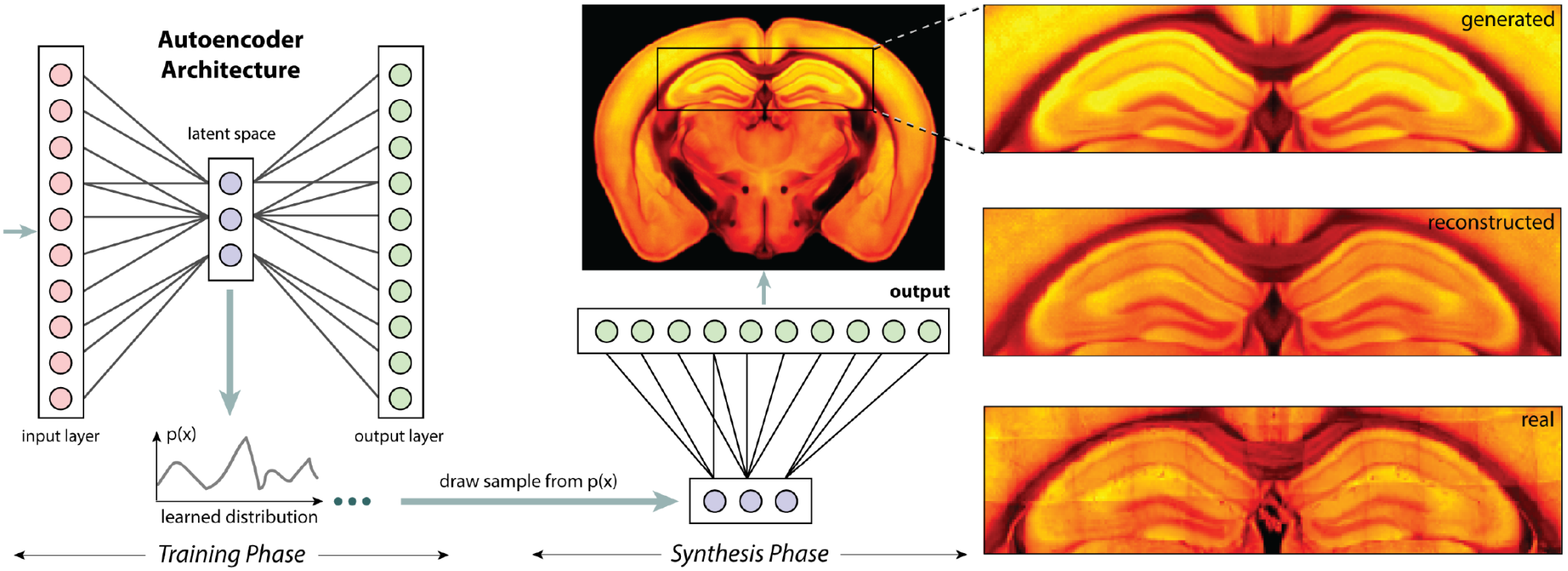
3D image volumes acquired using serial 2-photon tomography, downsampled the data to 25 microns (320 x 456)

Selected 2D slice (#284)

Mouse Connectivity Atlas dataset is available to the public at http://connectivity.brain-map.org/



Model Overview



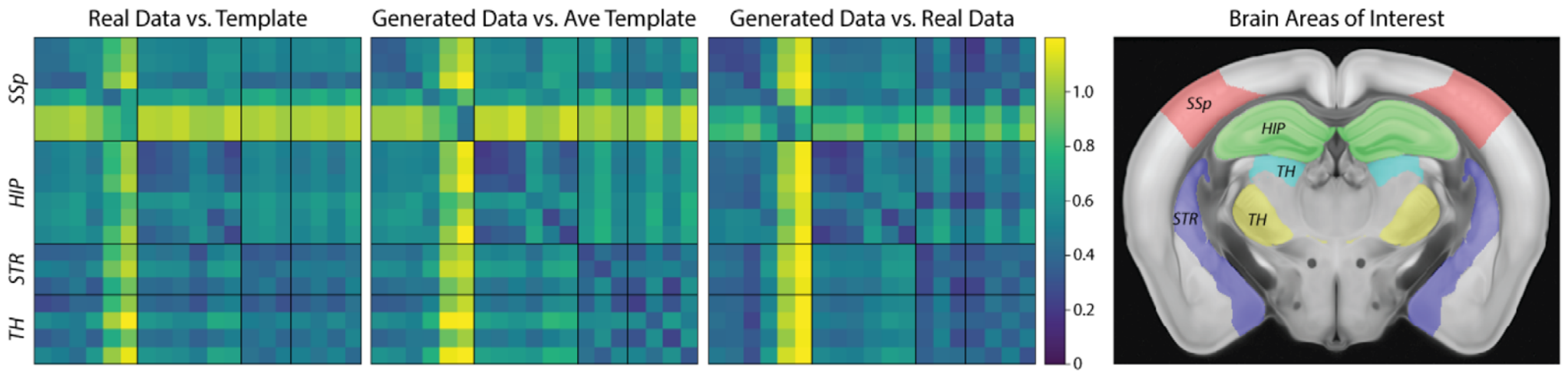
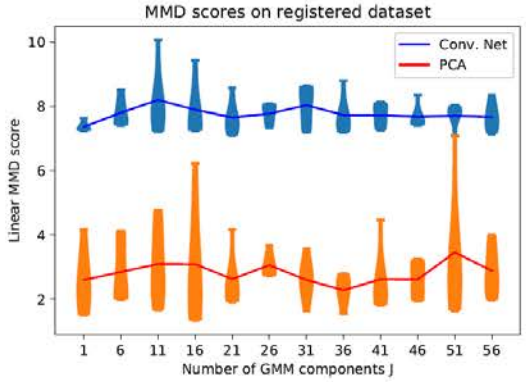
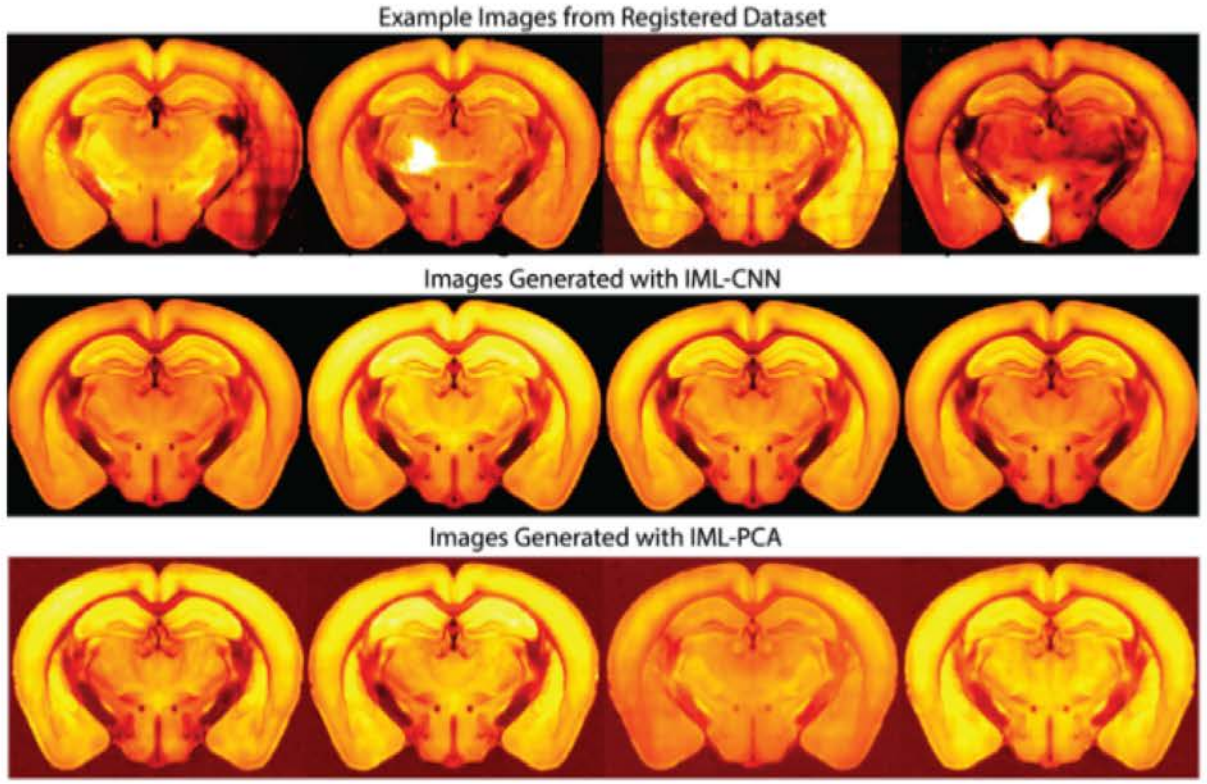


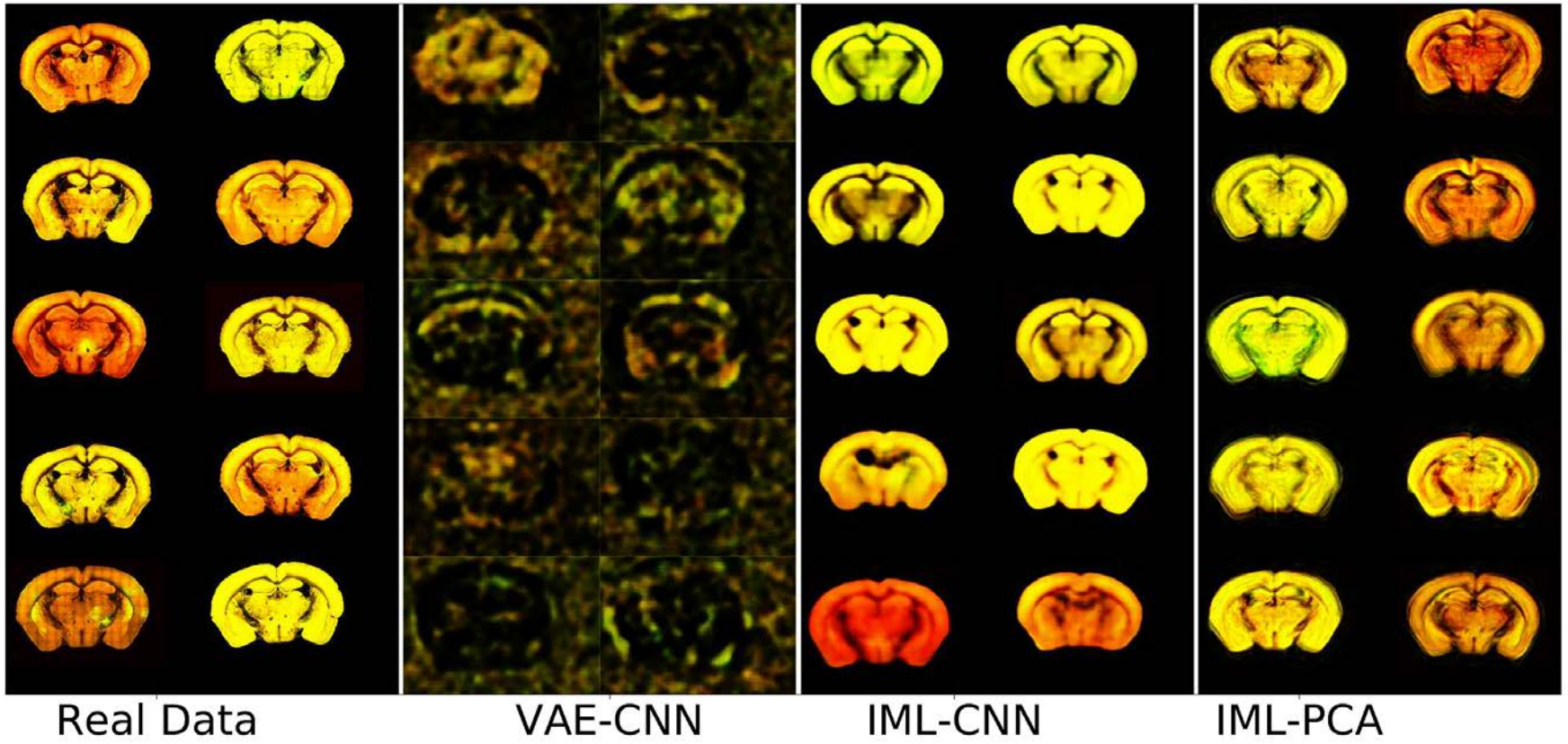
Figure 4: *Comparison of local statistics of real and generated images.* From left to right, we display the KL-divergences between the distributions of local texture features (local binary patterns) in twenty areas of interest, for the: real data vs. average template (L), generated data vs. average template (M), and the generated vs. real data (R). To the right of these KL-divergences, we display the masked brain areas used for this local texture analysis.

Evaluating model quality

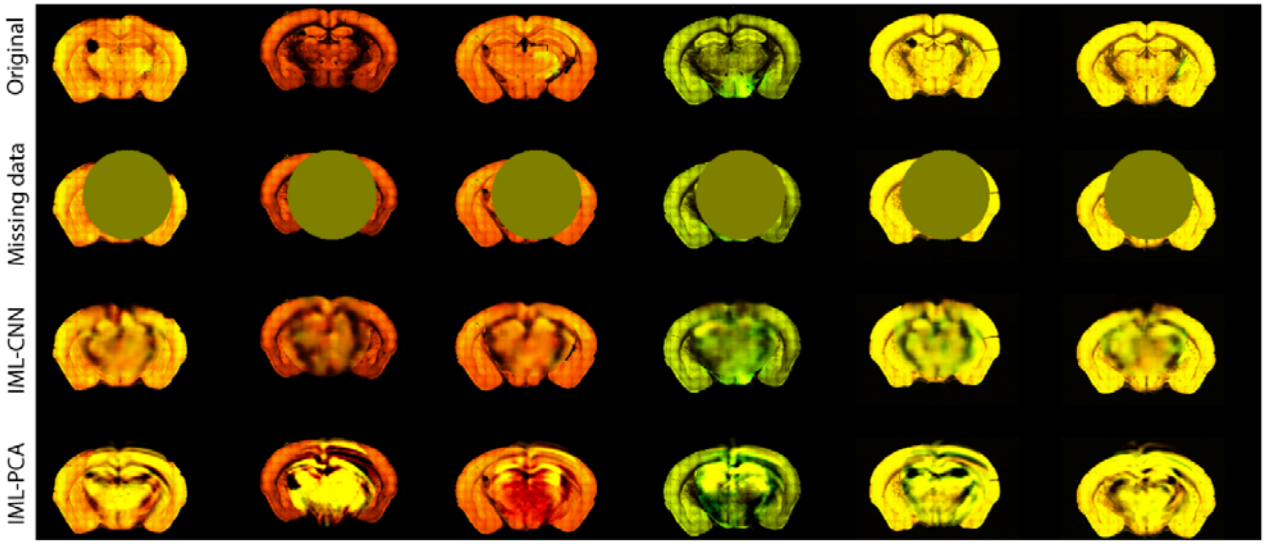




Synthesis for unregistered data



Application to missing data completion



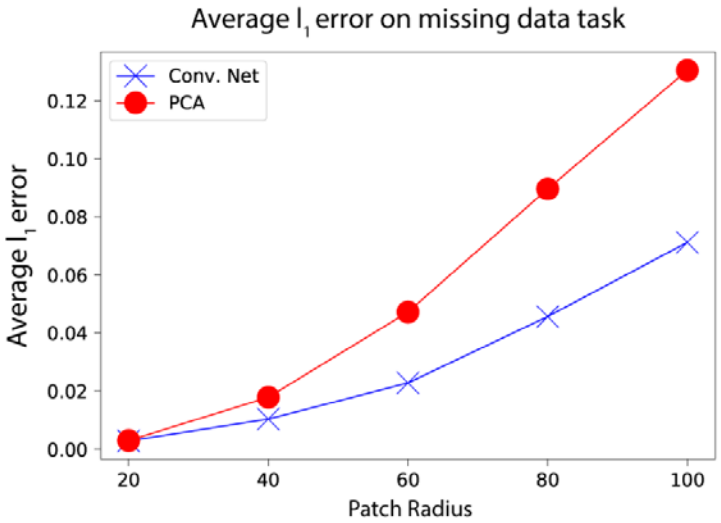


Figure 5: *Results for Missing Data Completion with IML.*(**top**) We show example images from the unregistered dataset, the same images with a chunk of data missing, and reconstructions of the missing data with IML-CNN and IML-PCA. We show examples for missing data patch of radii 100 (**bottom**) We show the l_1 error for IML-CNN and IML-PCA as the amount of missing data increases.

Application to outlier detection

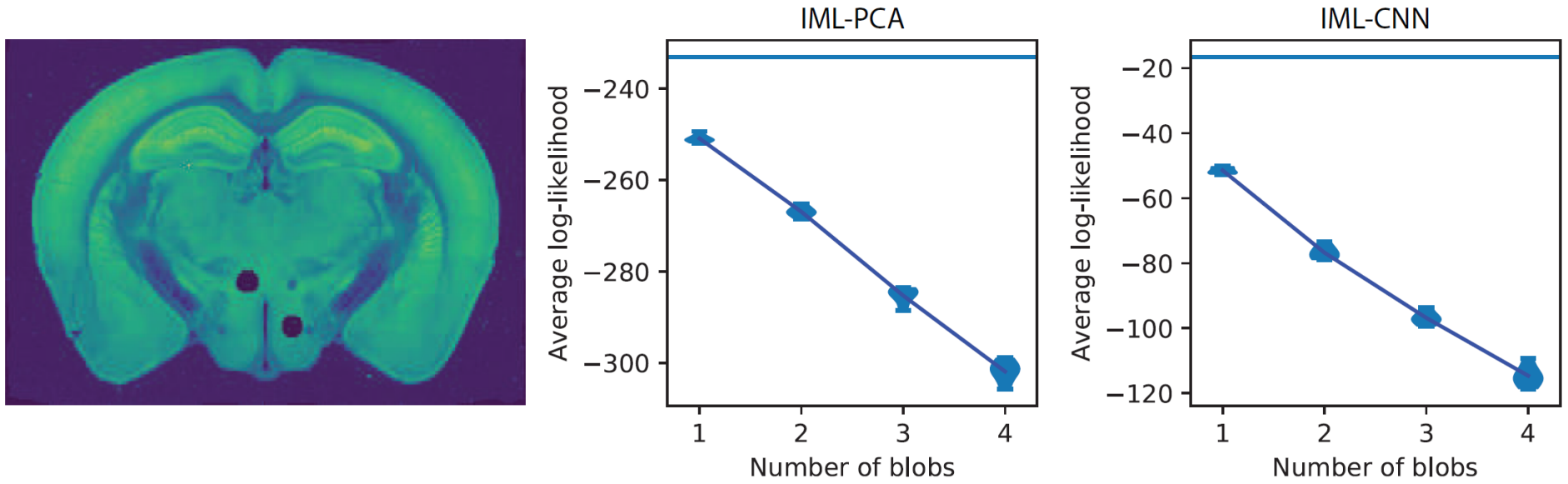


Figure 6: *Results for Outlier Detection with IML.* **(left)** An example image contaminated with two outliers. **(right)** Estimates for the model likelihood for different number of outliers for IML-PCA and IML-CNN. The likelihoods for clean images is shown with horizontal bars.

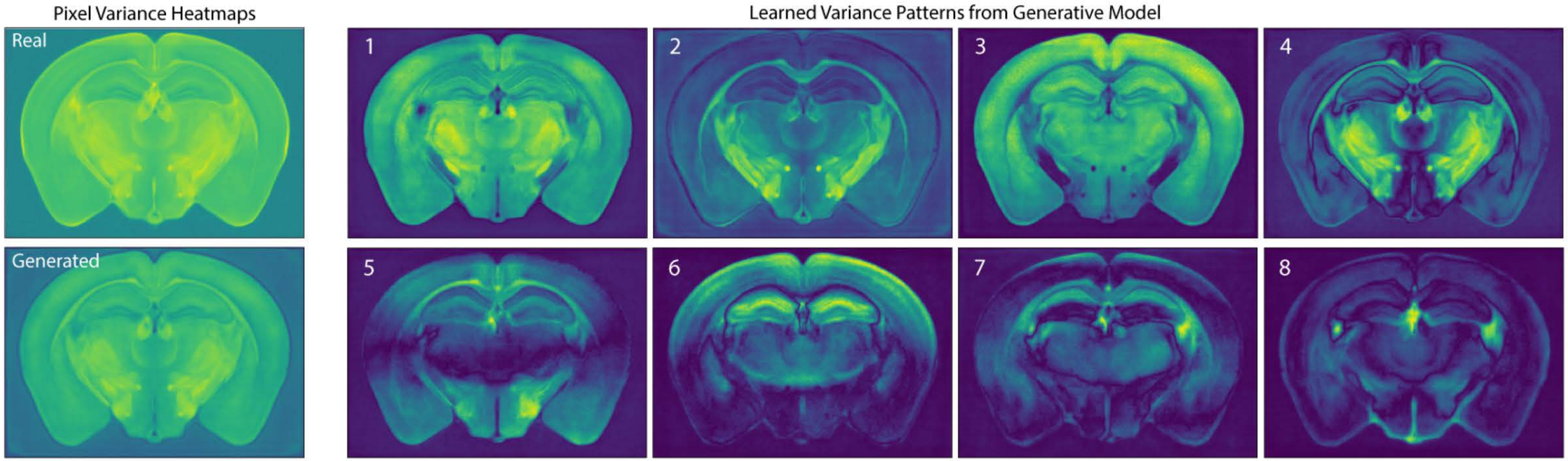


Figure 7: *Learned Signatures of Variability.* **(left)** Global variance heatmap of real (top) and generated images (bottom). **(right)** To understand different patterns of variability across subjects, we applied PCA to the latent factors in our autoencoder and then generated images that perturbed the learned average brain along the axis of each principal component. After generating 100 images for each principal component using this perturbation strategy, we then computed the pixel wise standard deviation across the perturbed images and display their standard deviation as a heatmap (principal components 1-8). These maps highlight the brain areas that exhibit the most variability for each principal component.

Code

https://github.com/ycemsubakan/mouse_brains

Conclusion & Future Work

Synthetic fMRI is an exciting and promising technology for accelerating brain data analysis.

Have discussed applications to:

- Classification
- Hypothesis testing

Initial results are encouraging…

Lots of work to do:

- Higher resolution / higher quality synthesis
- Improved evaluation
- Modeling of individual differences
- Outlier detection
- Data imputation

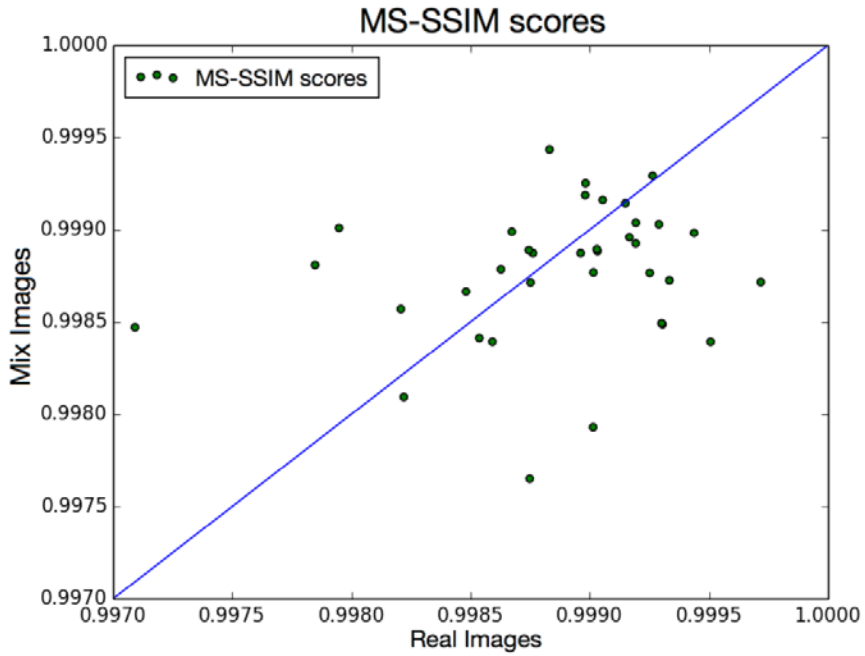
- Potential clinical applications
- Extension to time series
- …

Thank you

QUESTIONS?

SANMI@ILLINOIS.EDU

Backup Slides

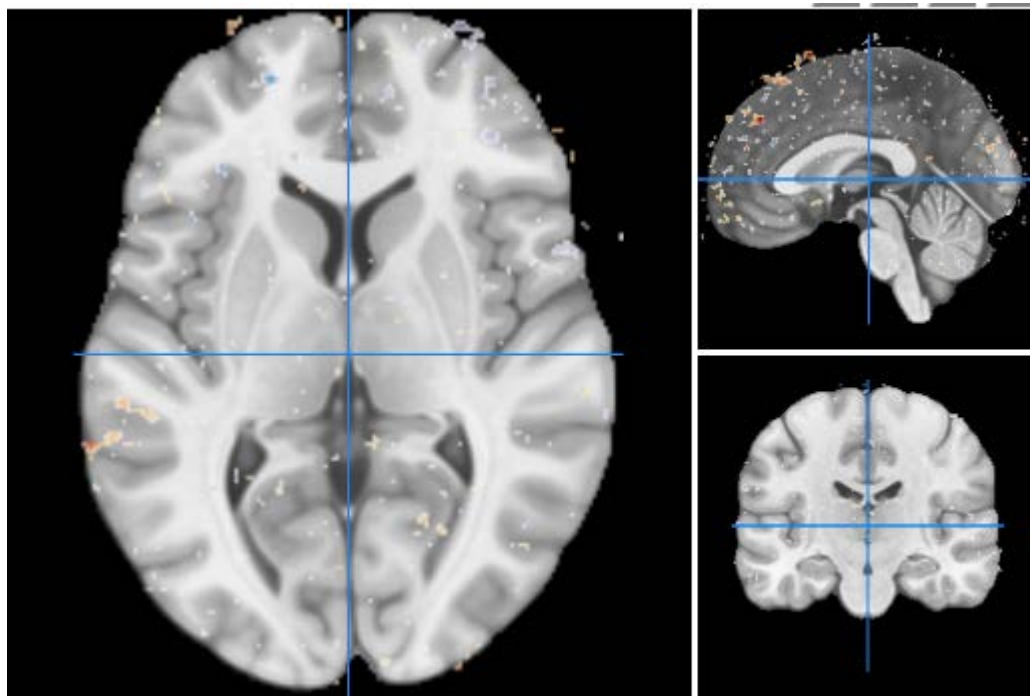


(a) Mean MS-SSIM scores between pairs of images within a given class, which were calculated on real data and synthetic data (using the ICW-GAN). Each point represents an individual class. Values in horizontal axis are MS-SSIM scores computed on real images, while values on the vertical axis are calculated using mixed images.

IBC: Individual brain maps from the Individual Brain Charting dataset

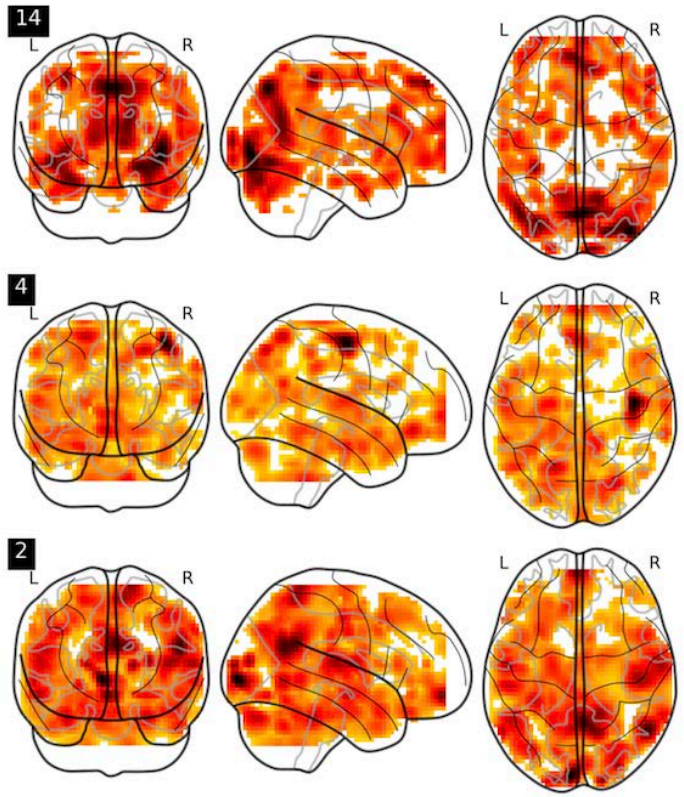
Contributed by bthirion

Add Date	Jan. 11, 2017, 4:57 p.m.
Authors	None
Contributors	
Description	The individual Brain Charting (IBC) Project is using high resolution fMRI to map 12 subjects that undergo a large number of tasks: the HCP tasks, the so-called ARCHI tasks, a specific language task, video watching, low-level visual stimulation etc. The native resolution of the data is 1.5mm isotropic. Their main value lies in the large number of contrasts probed, the level of detail and the high SNR per subject. This dataset is meant to provide the basis of a functional brain atlas. We upload here unsmoothed individual SPMs. The uploaded maps are fixed effects across maps acquired with AP and PA phase encoding directions.
DOI	None
Field Strength	None
id	2138
Journal	None



1,824 zstat brain maps

Generated Images

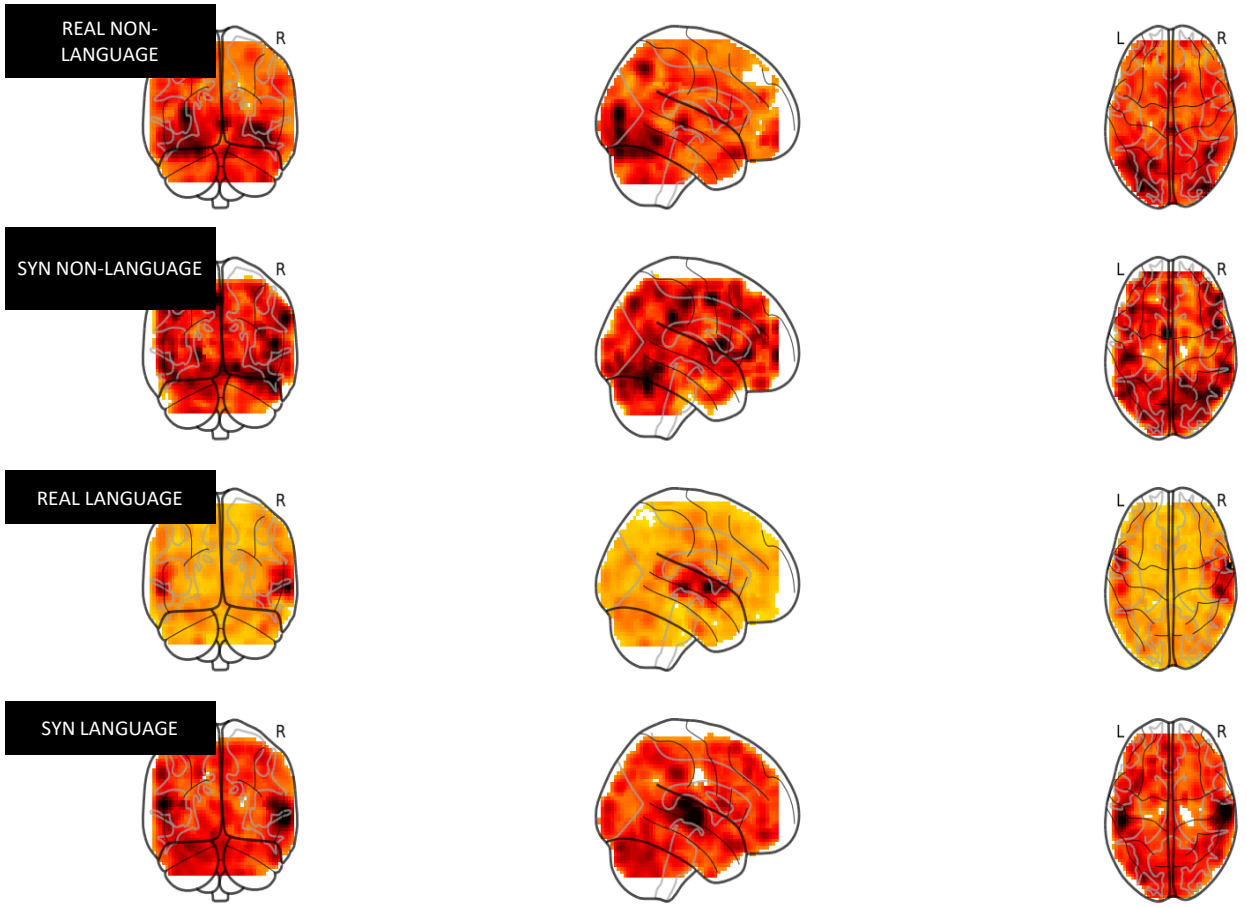


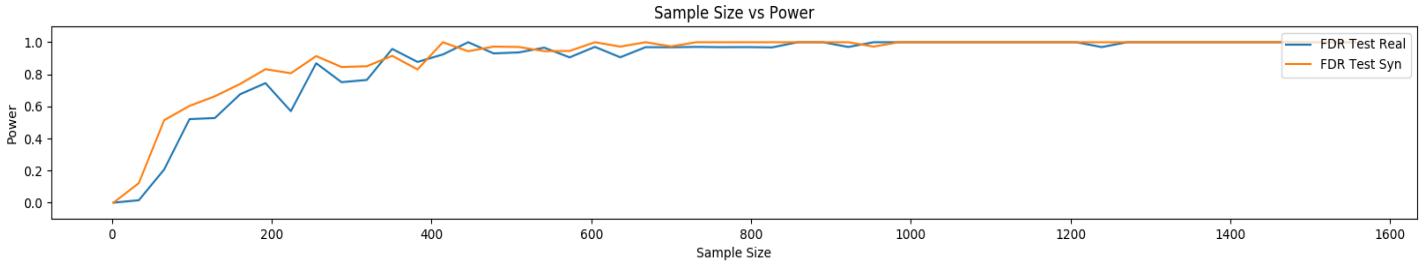
- **14:** visual form recognition, feature comparison, response selection, response execution, relational comparison, visual pattern recognition
- **22:** response execution, working memory, body maintenance, visual body recognition
- **4:** response selection, response execution, punishment processing
- **43:** motion detection
- **2:** response selection, response execution, animacy perception, animacy decision, motion detection
- **3:** response selection, response execution, motion detection

Application: Classifier data augmentation

Downsampling	Input	Classifier	Accuracy	Macro F1	Precision	Recall
8.0×	Real	SVM	0.523	0.480	0.497	0.523
	Real	NN	0.530	0.517	0.545	0.530
	Real+Synth.	SVM	0.531	0.493	0.510	0.533
	Real+Synth.	NN	**0.562**	**0.539**	**0.568**	**0.563**
4.0×	Real	SVM	0.555	0.507	0.517	0.533
	Real	NN	0.723	0.712	**0.737**	**0.723**
	Real+Synth.	SVM	0.562	0.517	0.527	0.563
	Real+Synth.	NN	**0.737**	**0.715**	0.727	**0.723**

Yes effect i.e. Null is False





**Power computed voxel-wise after FDR correction