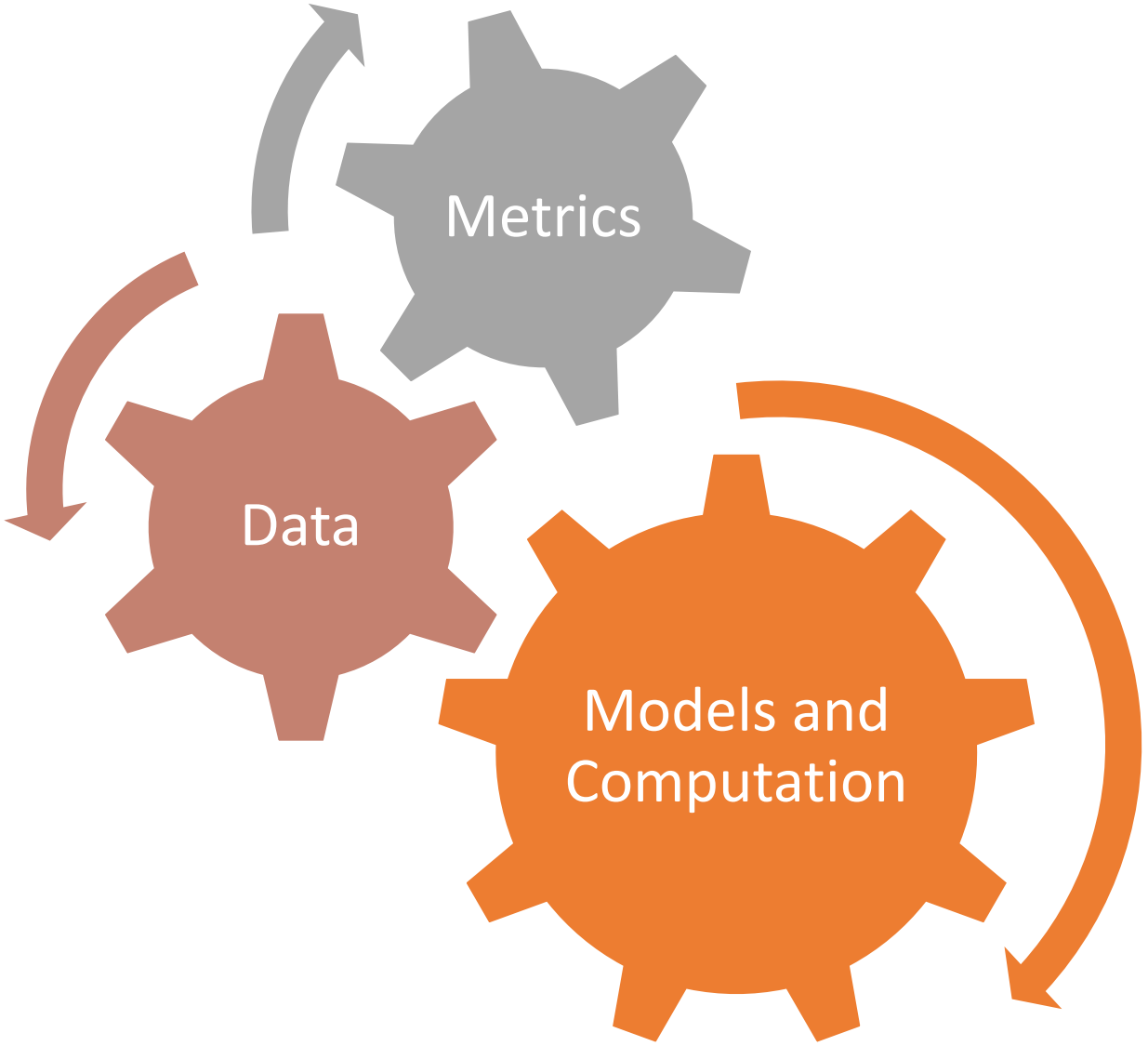
Towards Machine Learning for Personalized Healthcare

SANMI KOYEJO

CS & BECKMAN @ILLINOIS

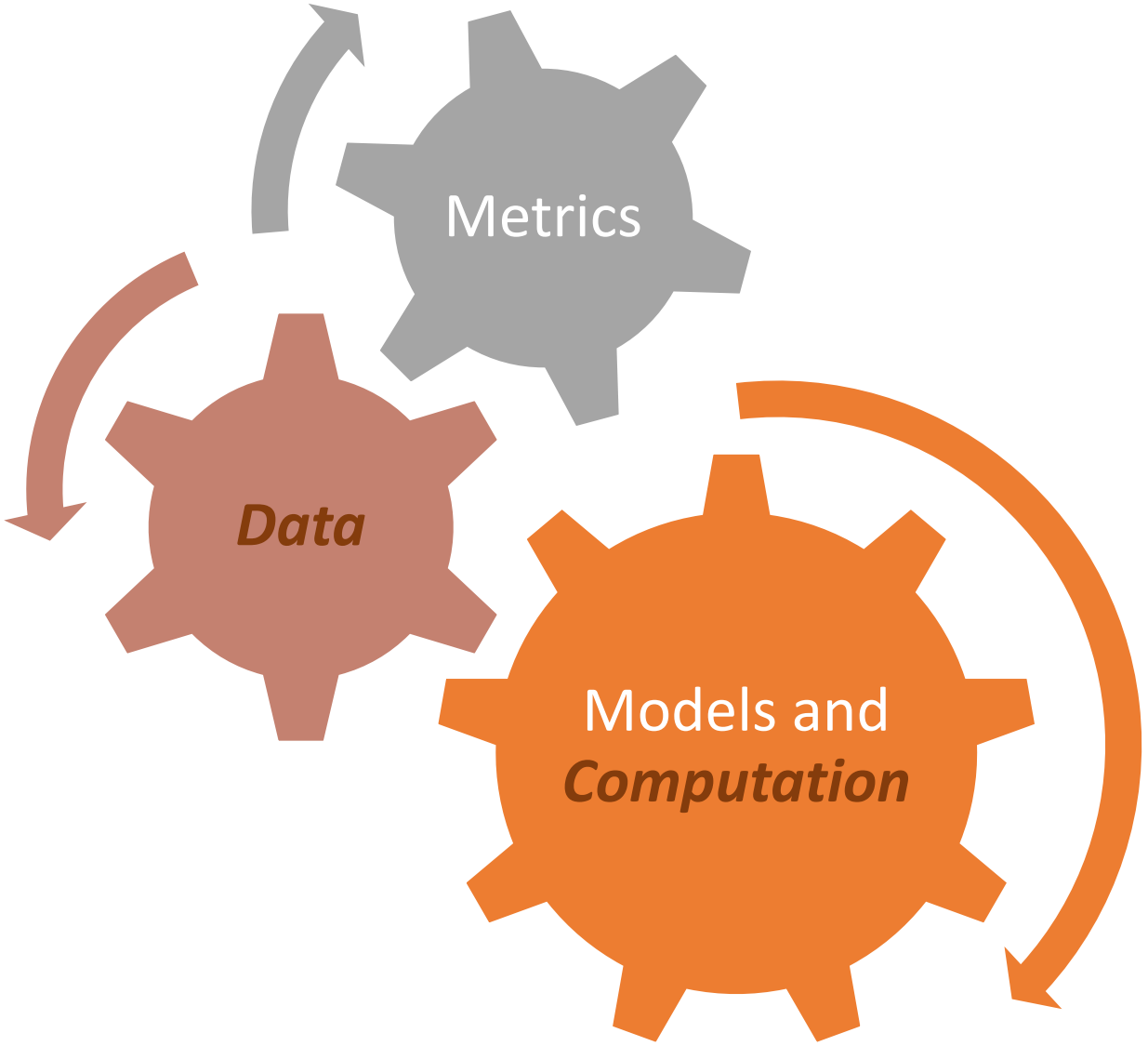
What does it take to build an effective machine learning system for healthcare?



Modeling	• (Brain) dynamics, longitudinal tracking, diagnosis • **Applications: Glioma segmentation, Cancer phylogenetics**
Evaluation	• **Selecting good metrics for machine learning** • Training models that optimize specialized metrics
Privacy	• **Data synthesis,** learning with aggregated data • **Learning on the edge**
Trust	• Explainability and interpretability using examples • **Individual recourse**

Enabling Technologies

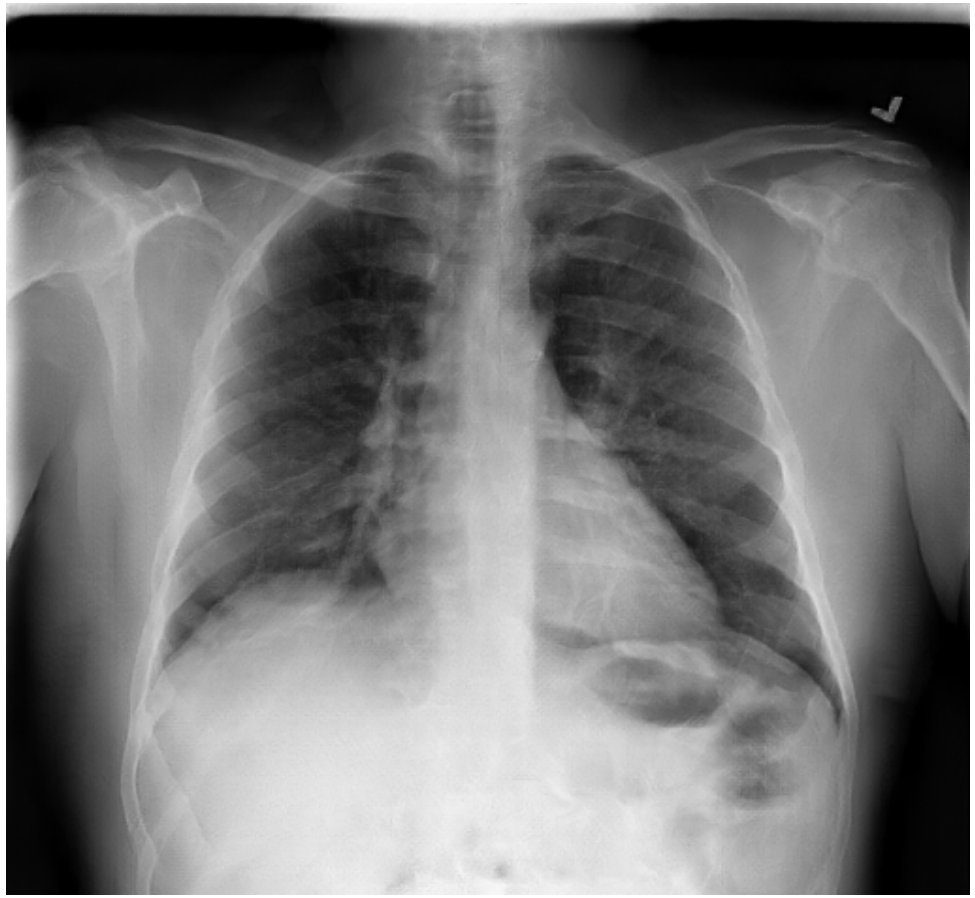
Privacy-preserving technology for healthcare ML



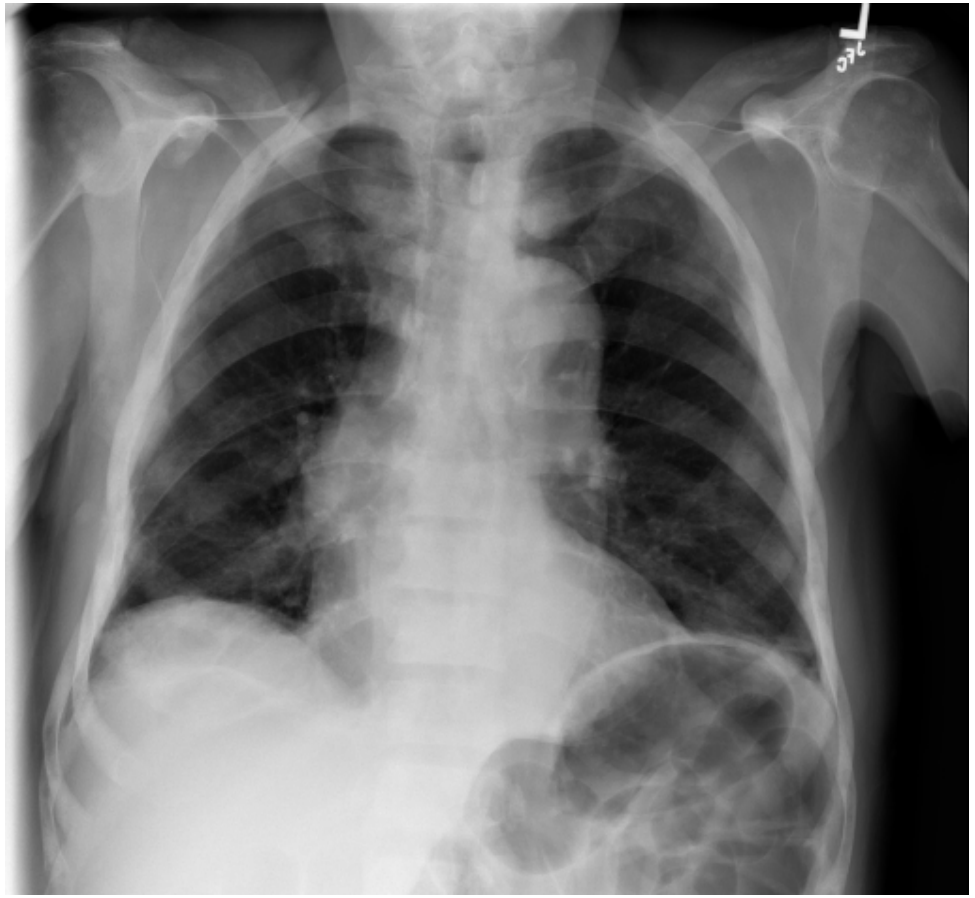
Synthesizing medical images using generative adversarial networks

Applications to private data release and rare-event simulation

Collaborators
@Illinois: Ishan Deshpande, Alex Schwing, Peiye Zhuang, David Forsyth
@Dupage: Nasir A. Siddiqui, Ayis T. Pyrros

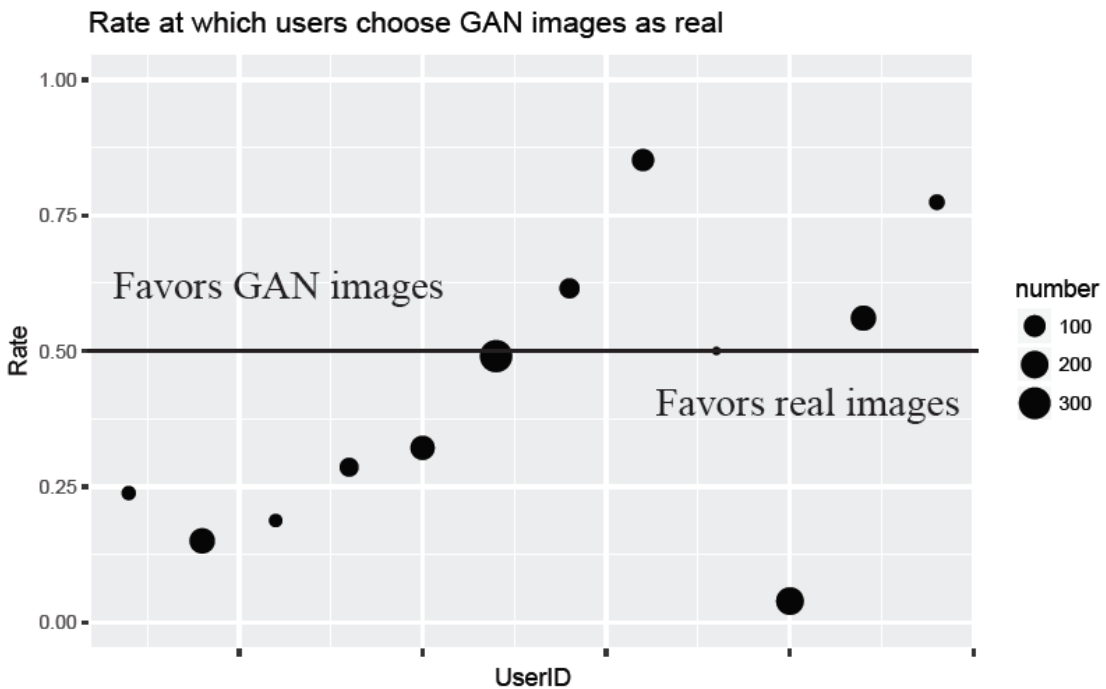


Synthetic



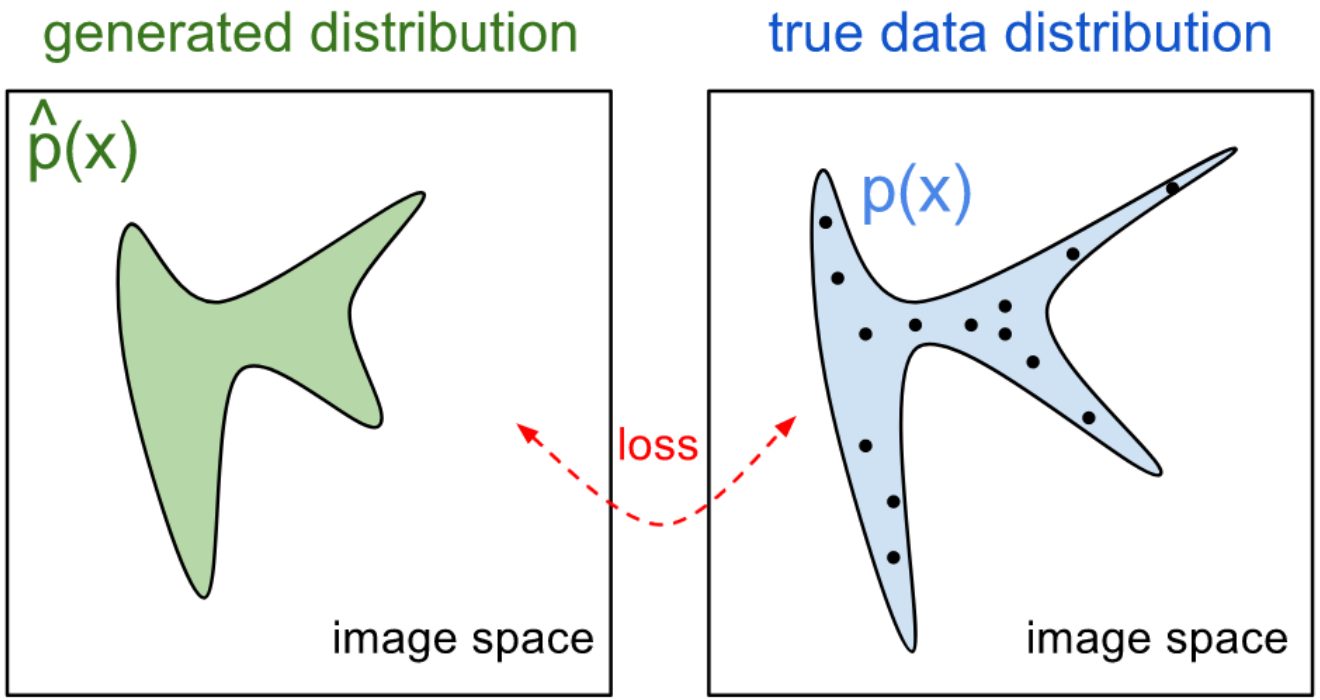
Real

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

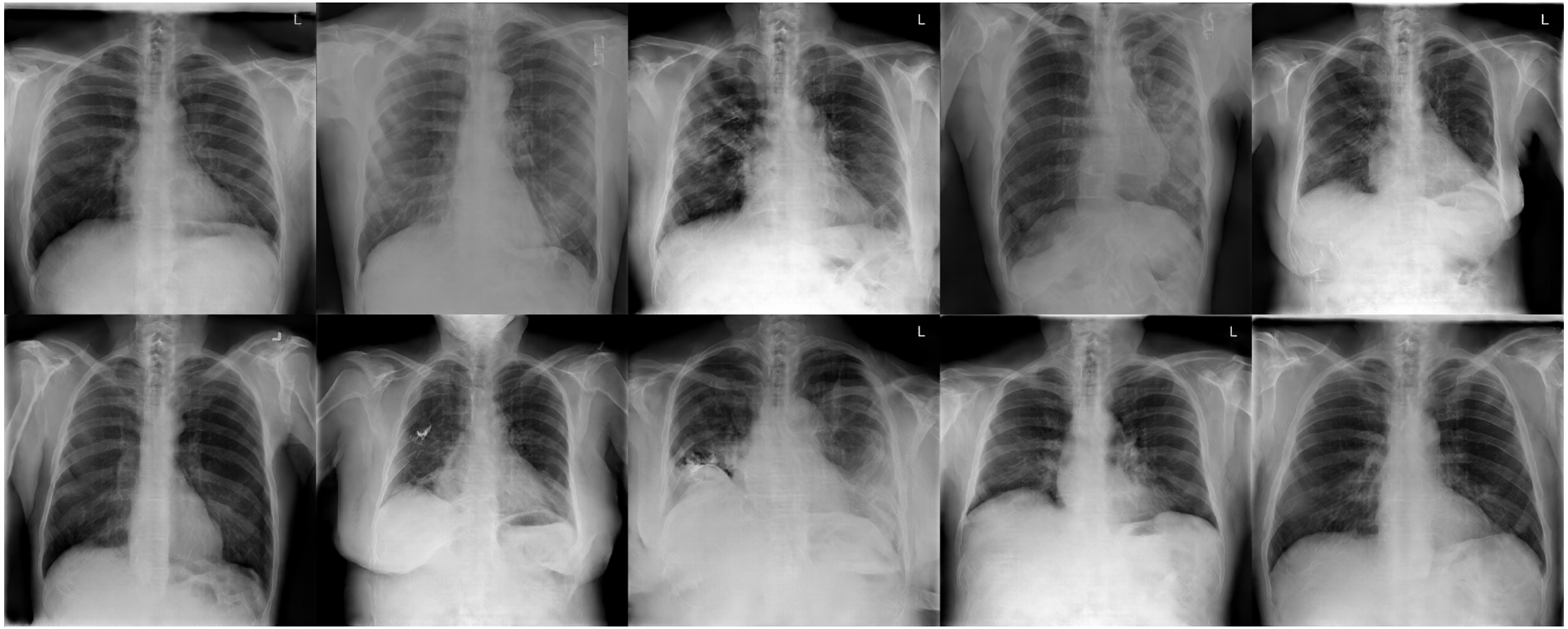


Experienced radiologists were asked to choose which of a real lung x-ray and a GAN generated image were real. Subjects favored real images slightly (on average GAN images were identified as real 39% of the time) but subject behavior varied widely. Size of blob identifies number of pairs viewed; note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

Generative Models

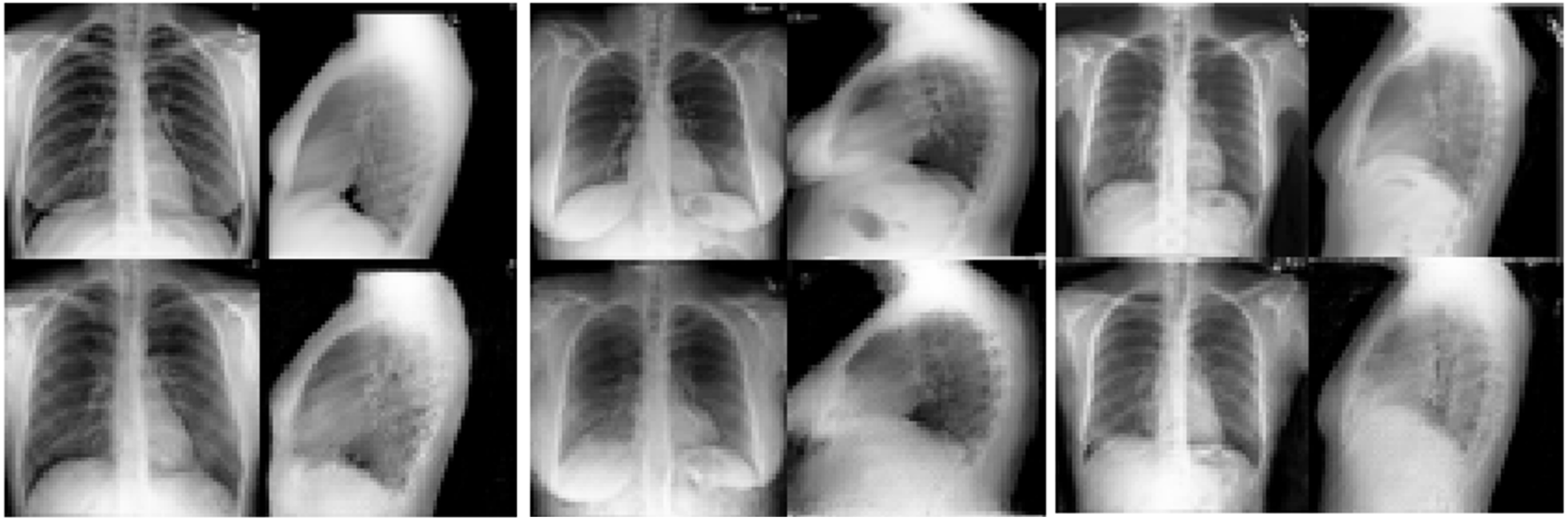


Source: https://blog.openai.com/generative-models/

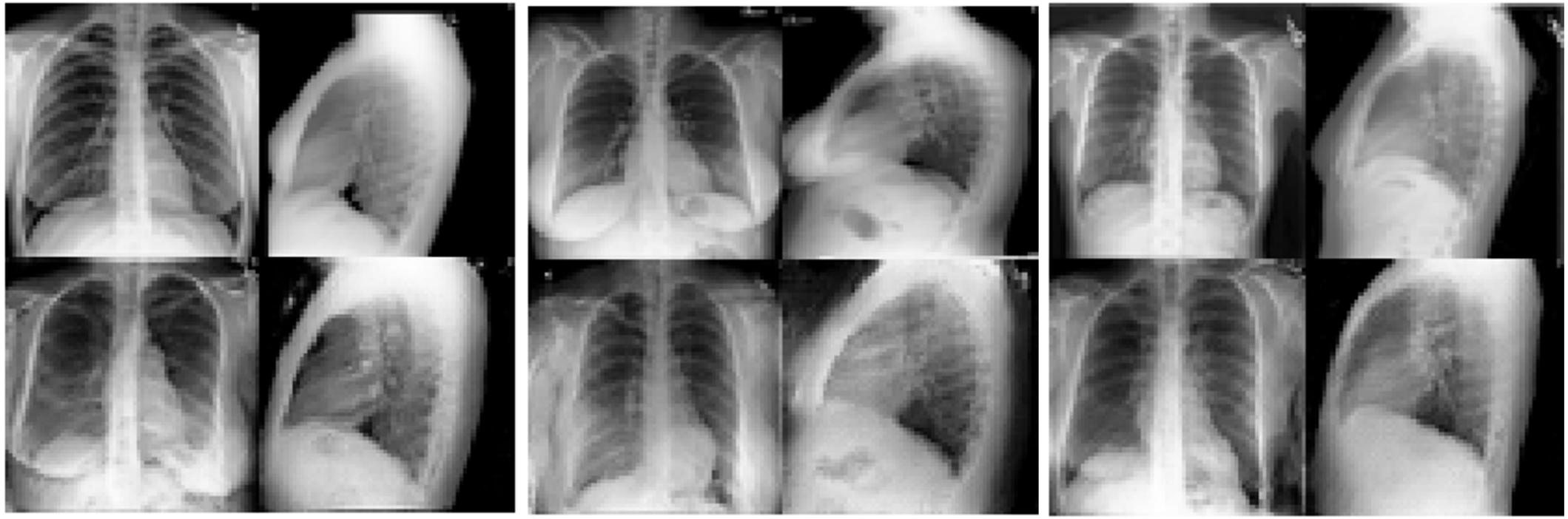


Synthesis at native resolution ~ 1024^2 pixels

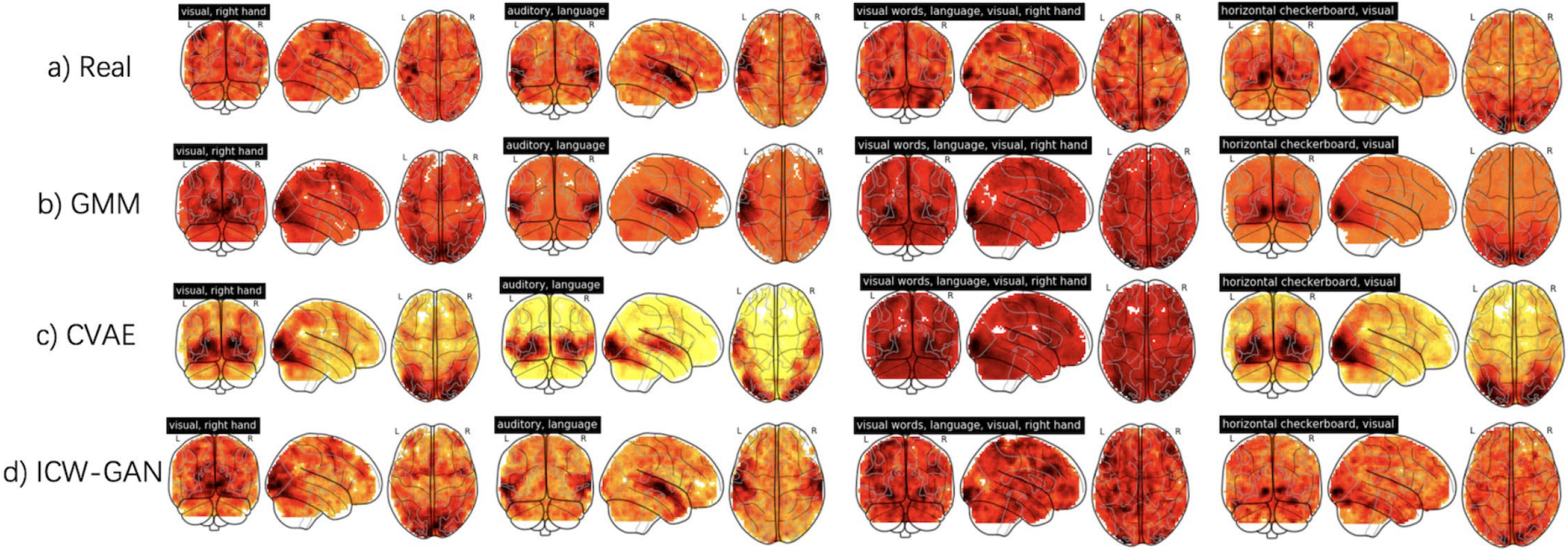
Synthesizing front and side X-rays



Co-generation (Front => Side, <=)

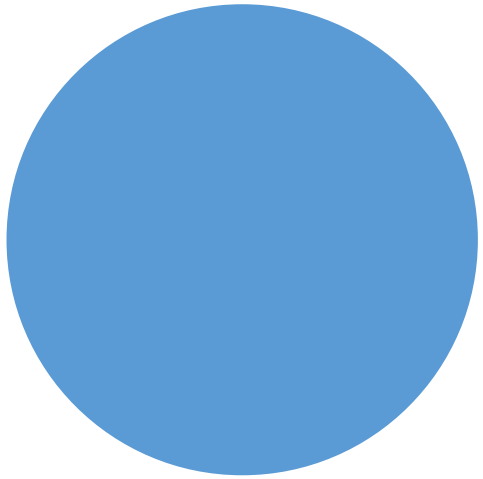
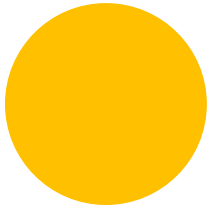


Synthesizing functional MRI



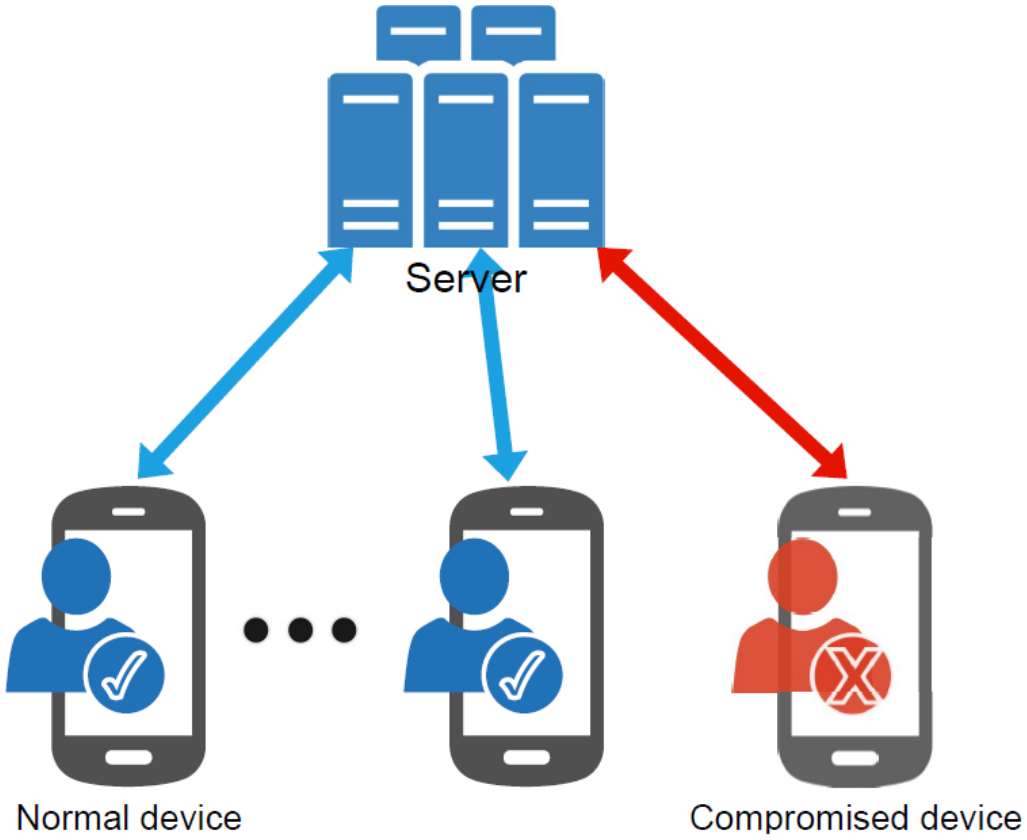
Application: classifier data augmentation

Input	Gen. model	Classifier	Accuracy	Macro F1	Precision	Recall
Real	-	SVM	0.8181	0.82	0.8333	0.8133
Real+noise	-	SVM	0.8185	0.82	0.8367	0.82
Real+Synth.	GMM	SVM	0.8188	0.82	0.8366	0.82
Real+Synth.	CVAE	SVM	0.8248	0.8267	0.8367	0.8233
Real+Synth.	ICW-GAN	SVM	**0.8311**	**0.83**	**0.8433**	**0.8333**
Real	-	DNN	0.852	0.857	0.872	0.8523
Real+noise	-	DNN	0.8581	0.856	0.8719	0.8579
Real+Synth.	GMM	DNN	0.8604	0.8631	0.8749	0.8604
Real+Synth.	CVAE	DNN	0.8684	0.869	0.8827	0.8683
Real+Synth.	ICW-GAN	DNN	**0.8799**	**0.8825**	**0.8933**	**0.88**

Privacy Preserving Federated ML

Collaborators
@Illinois:
Cong Xie,
Indy Gupta



Federated ML

- ML models can be trained and deployed in distributed settings without transferring data
- Distributed learning amortizes training costs, learns without data sharing
- When implemented correctly, distributed learning preserves privacy and is robust to failures

What are the properties of ML with distributed data?

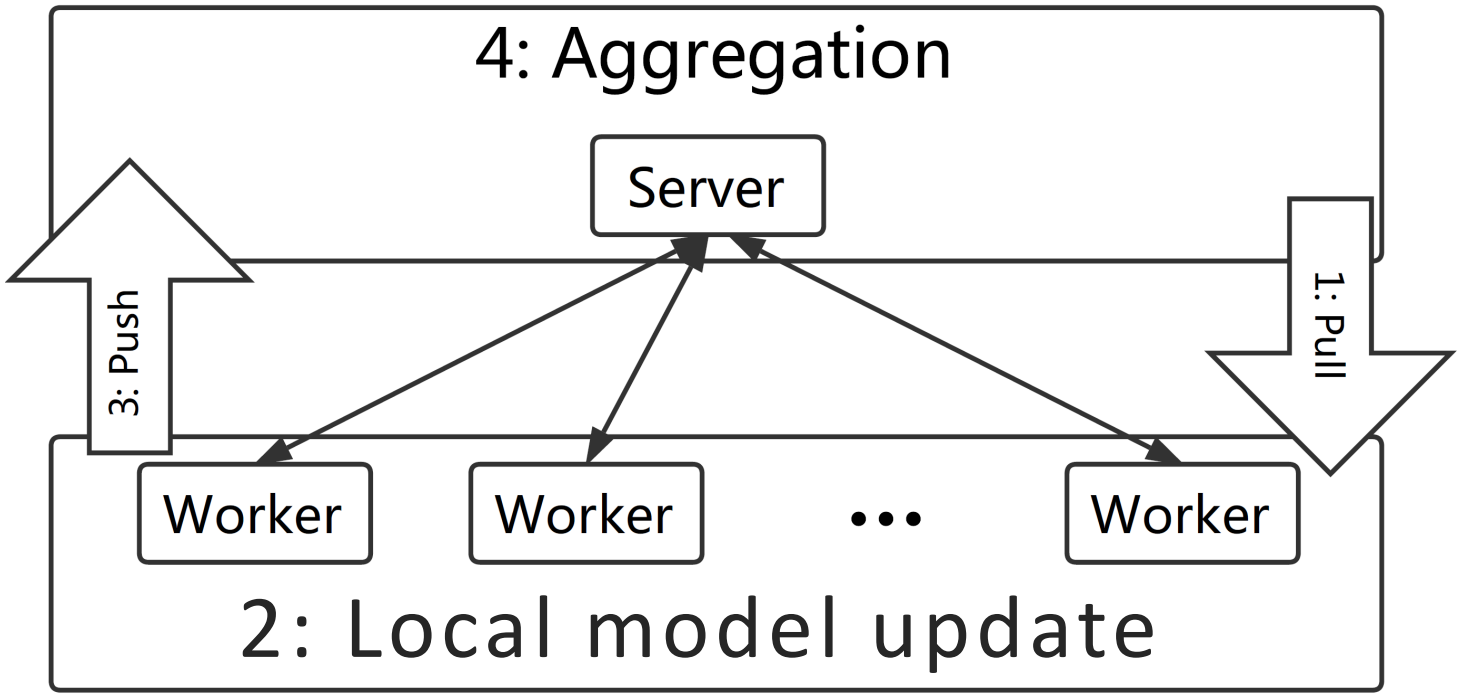
unbalanced, non-IID device data

limited, heterogeneous device computation

infrequent task scheduling

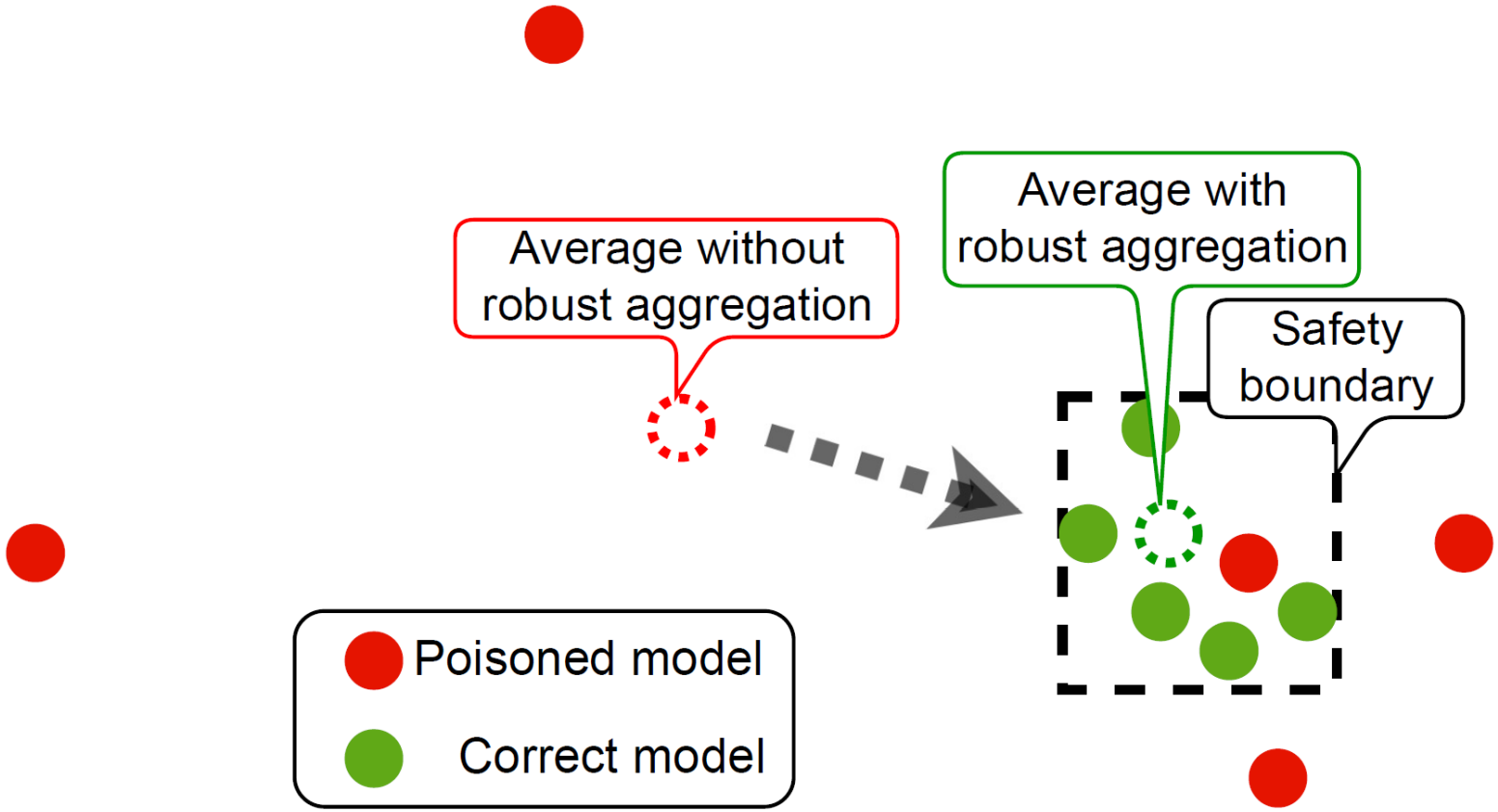
limited, infrequent communication, congestion

untrusted devices and data poisoning

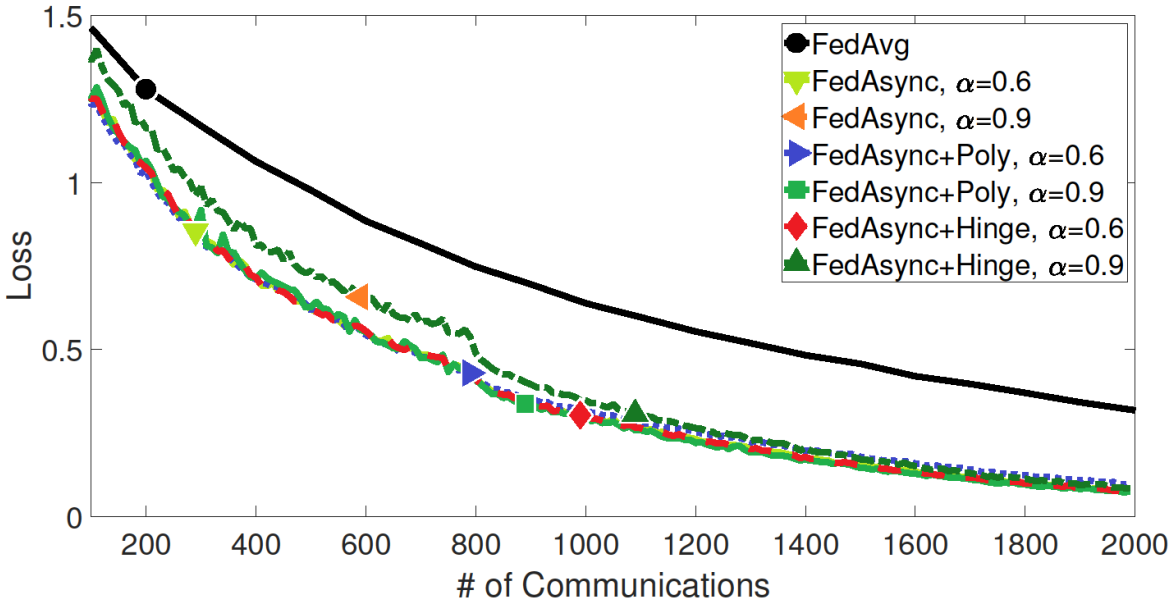


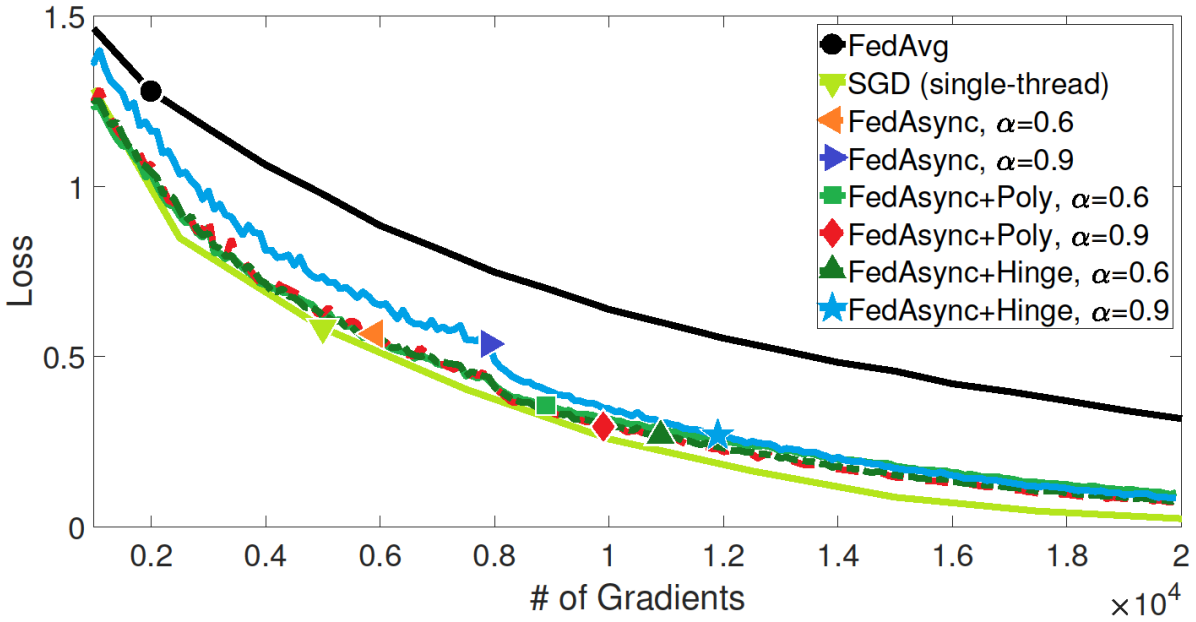
Workers compute updated local model parameters, no need to share data

Robust aggregation enables learning with failures



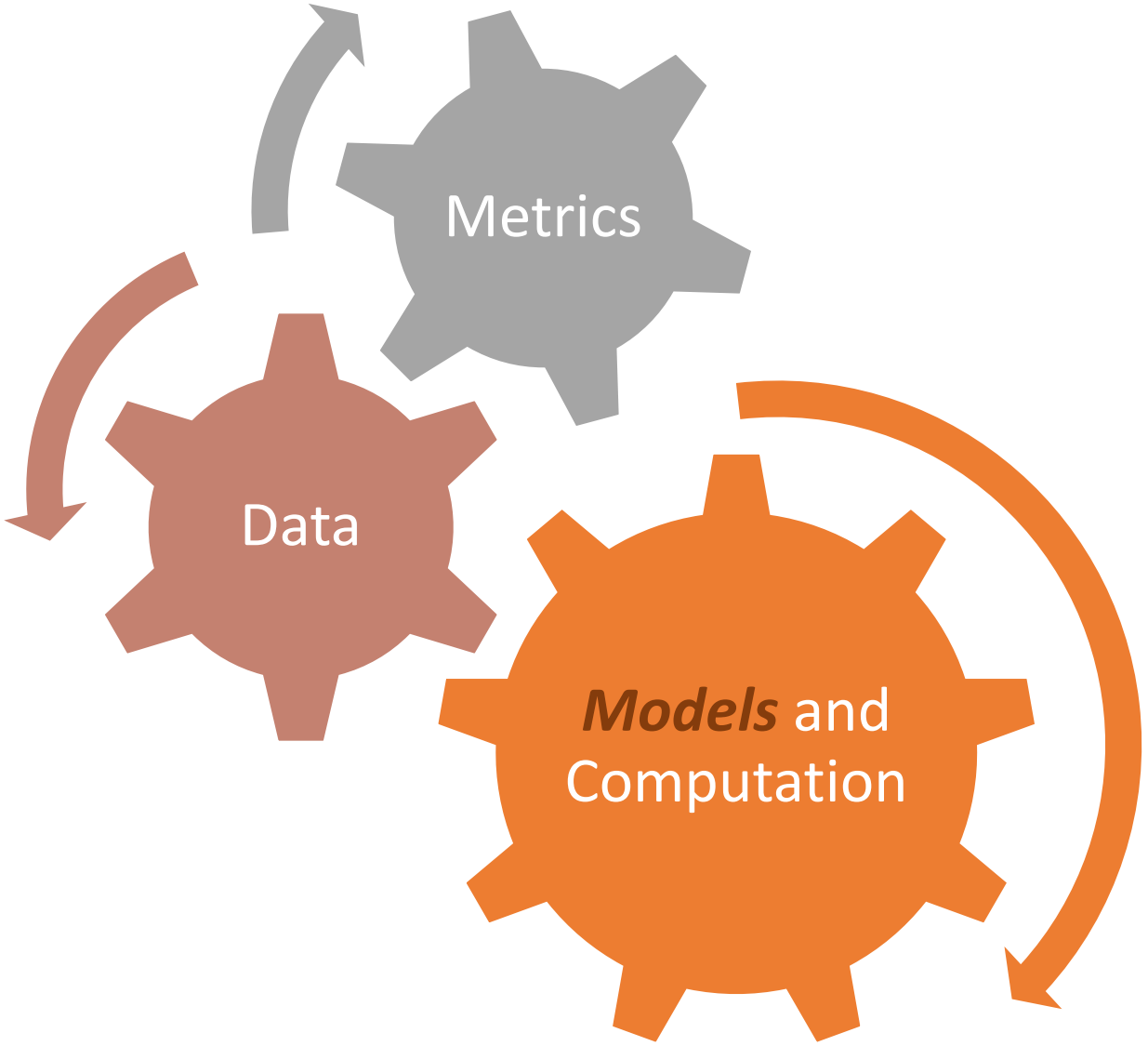
5-layer CNN, Unbalanced data, 100 devices





Performance vs # Gradients Max staleness of 4,with *Poly* and *Hinge* temporal smoothing

Modeling complex high-dimensional data

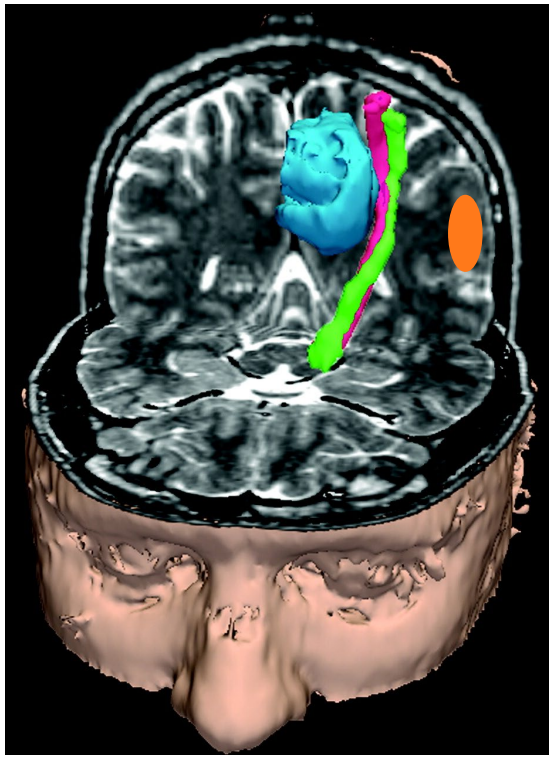


Glioma Segmentation

Collaborators
@Illinois: Chase Duncan, Peiye Zhuang, Brad Sutton
@Jump: Matt Bramlet
@OSF: Deepak Nair

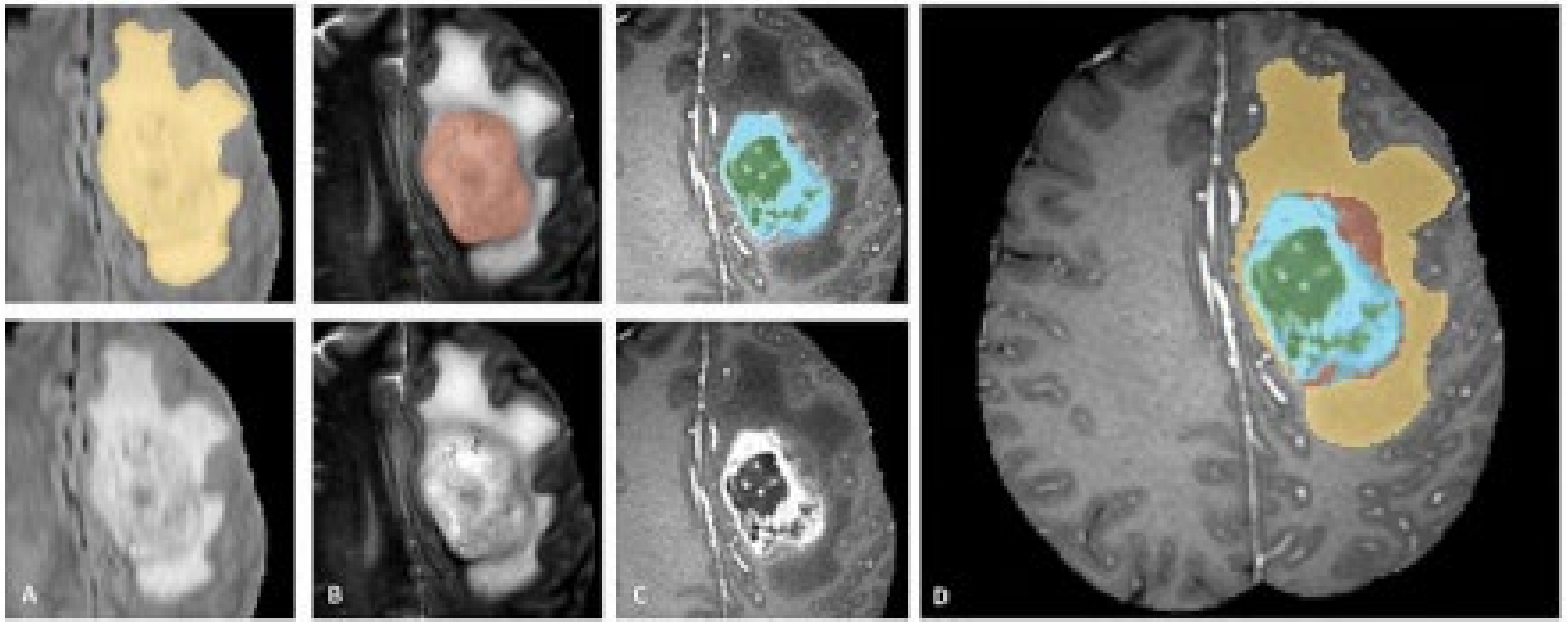
Glioma Segmentation Workflow

INPUTS	PROCESS	OUTPUTS
Standard Brain, T1/T2 with contrast (DICOM)	**Machine Learning Code** Autoseg tumor (enhance region, necrotic, edema, non-tumor) Gray/white matter	**Labeled Tumor** Enhance/necrotic/edema Gray/White In 3D STL's
Functional MRI DICOM and jpeg (processed in NordicNeuro?) Language, Motor	**3D Activation maps** Pull out activations from subject and normative data. Use DICOM to reassemble volume. Image registration required. Visualization overlay only	**Activation regions** 3D map of activation overlays: Subject and population average
DTI DICOM and streamlines (Processed in BrainPath?)	**3D Streamlines for Tracts** Group sets of tracks into tubes for visualizing large fiber bundles. Image registration required. Visualization only.	**3D Tubes** Main fiber pathways, subject specific.



Merge to Enduvo VR

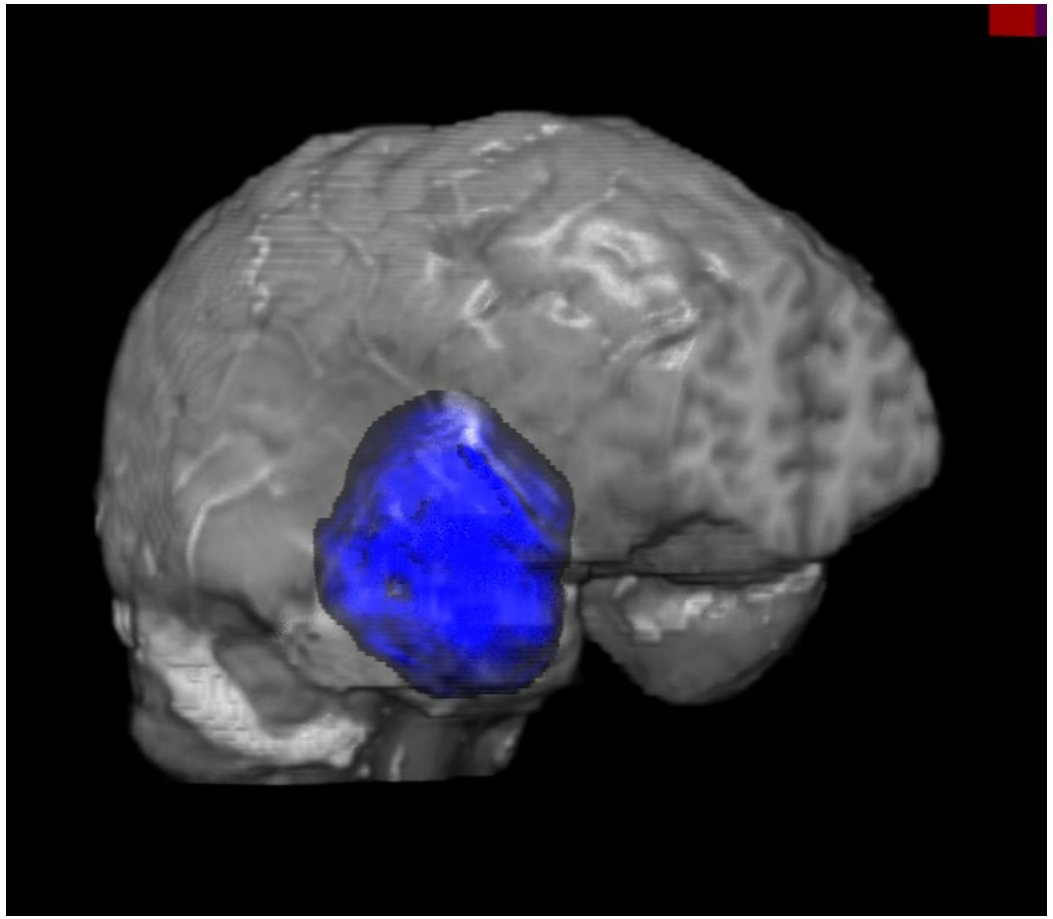
Picture modified from: Mukherjee, Berman, Chung, Hess, Henry. AJNR, 2008, 29(4): 632-41

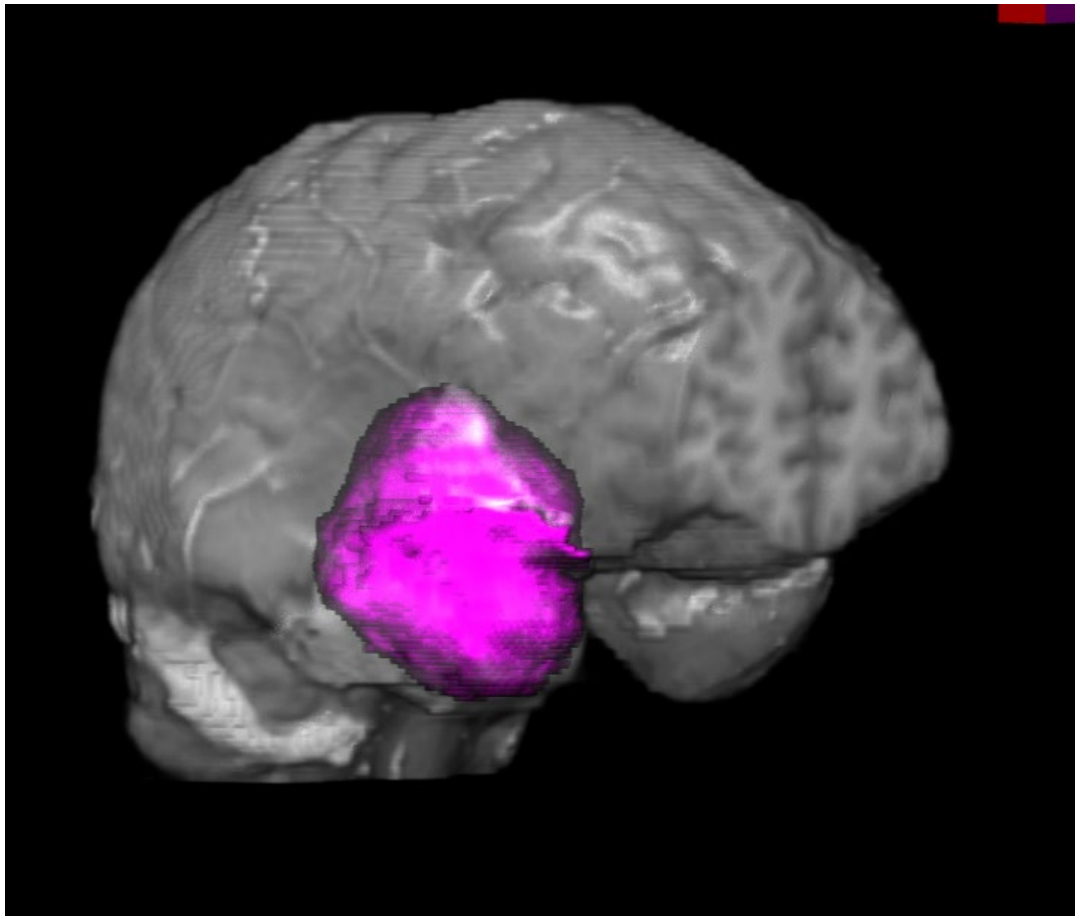


Problem formulation & approach

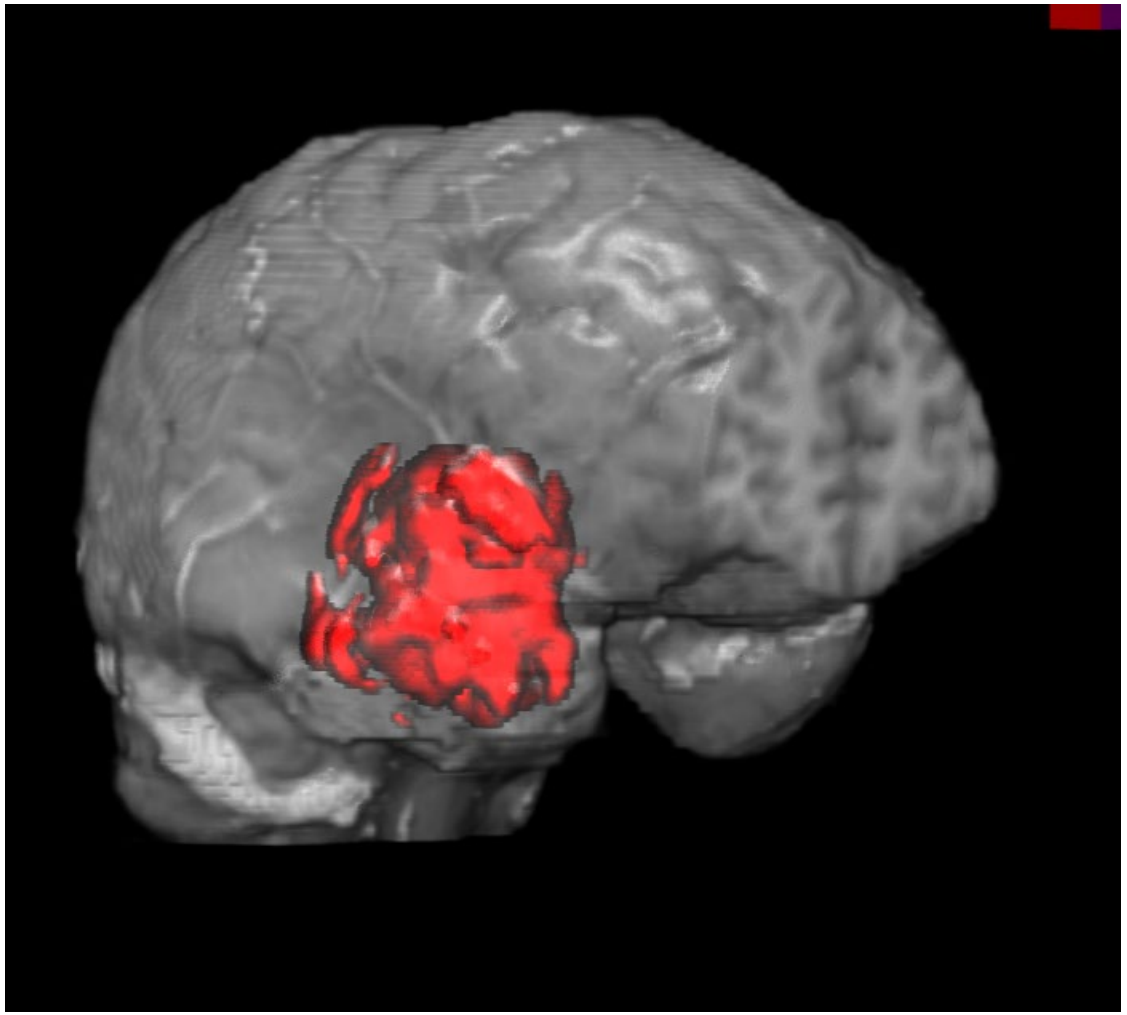
- Label each voxel as tumor vs. not tumor
- We use a variation of the U-net with improved regularization

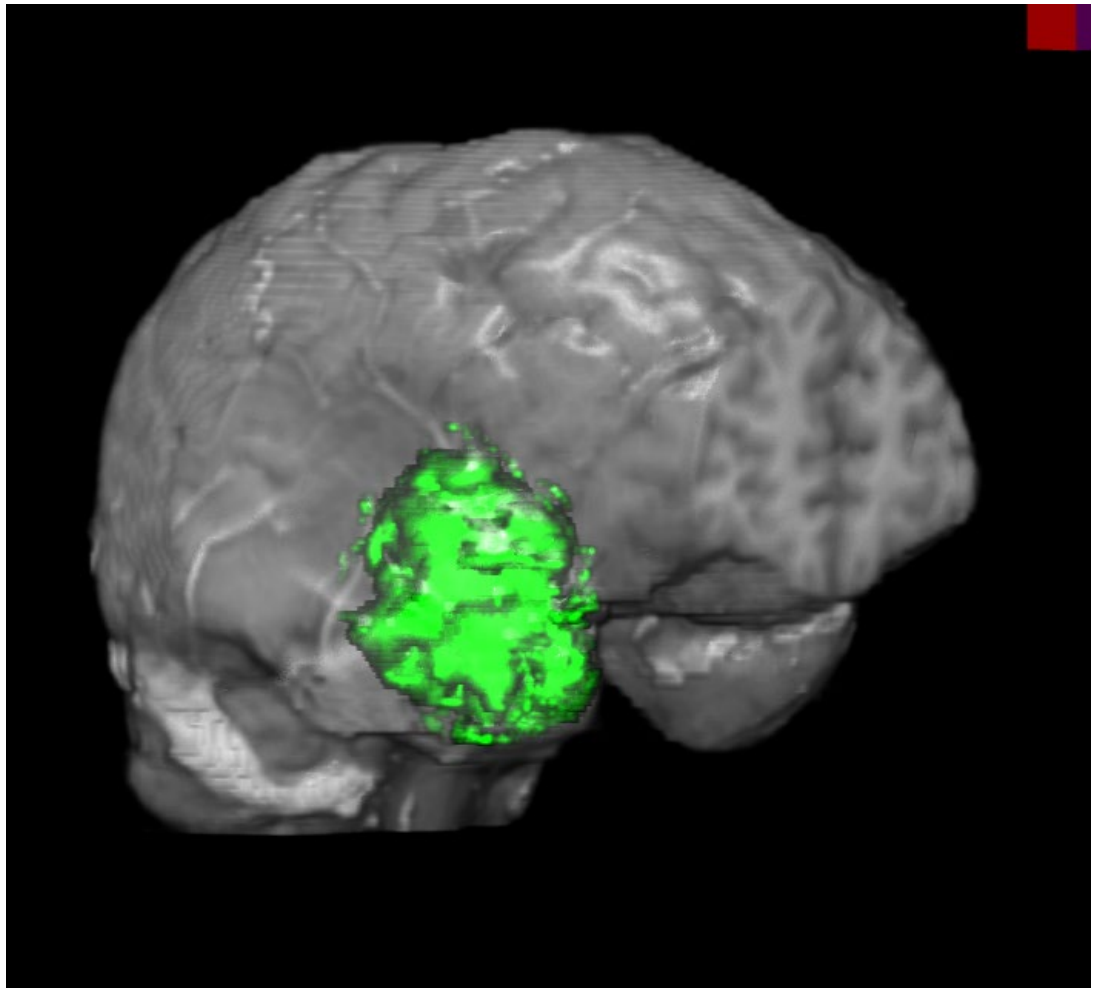
Enhancing tumor: prediction vs. ground truth





Tumor core: prediction vs. ground truth



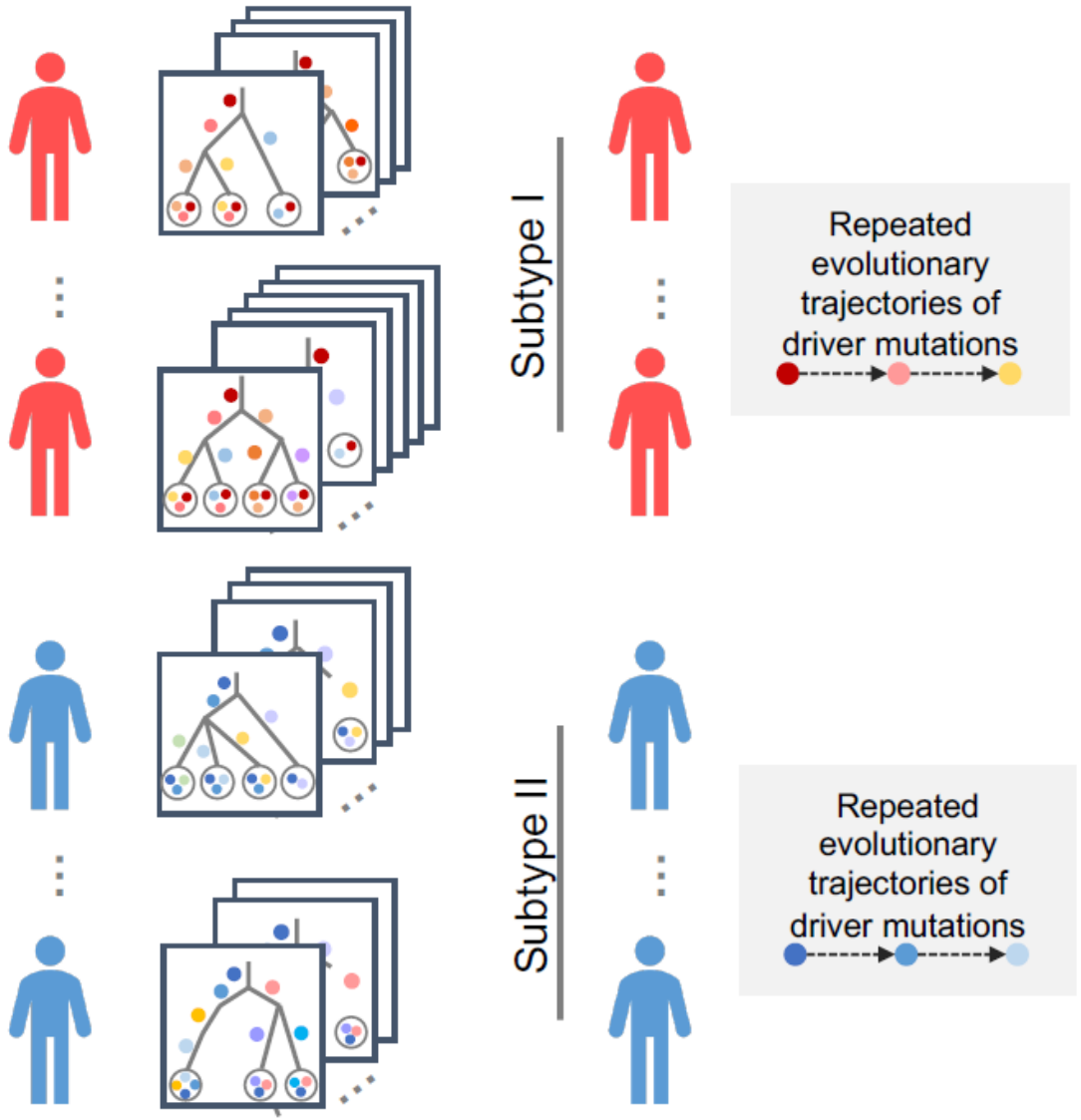


Cancer Phylogenetics

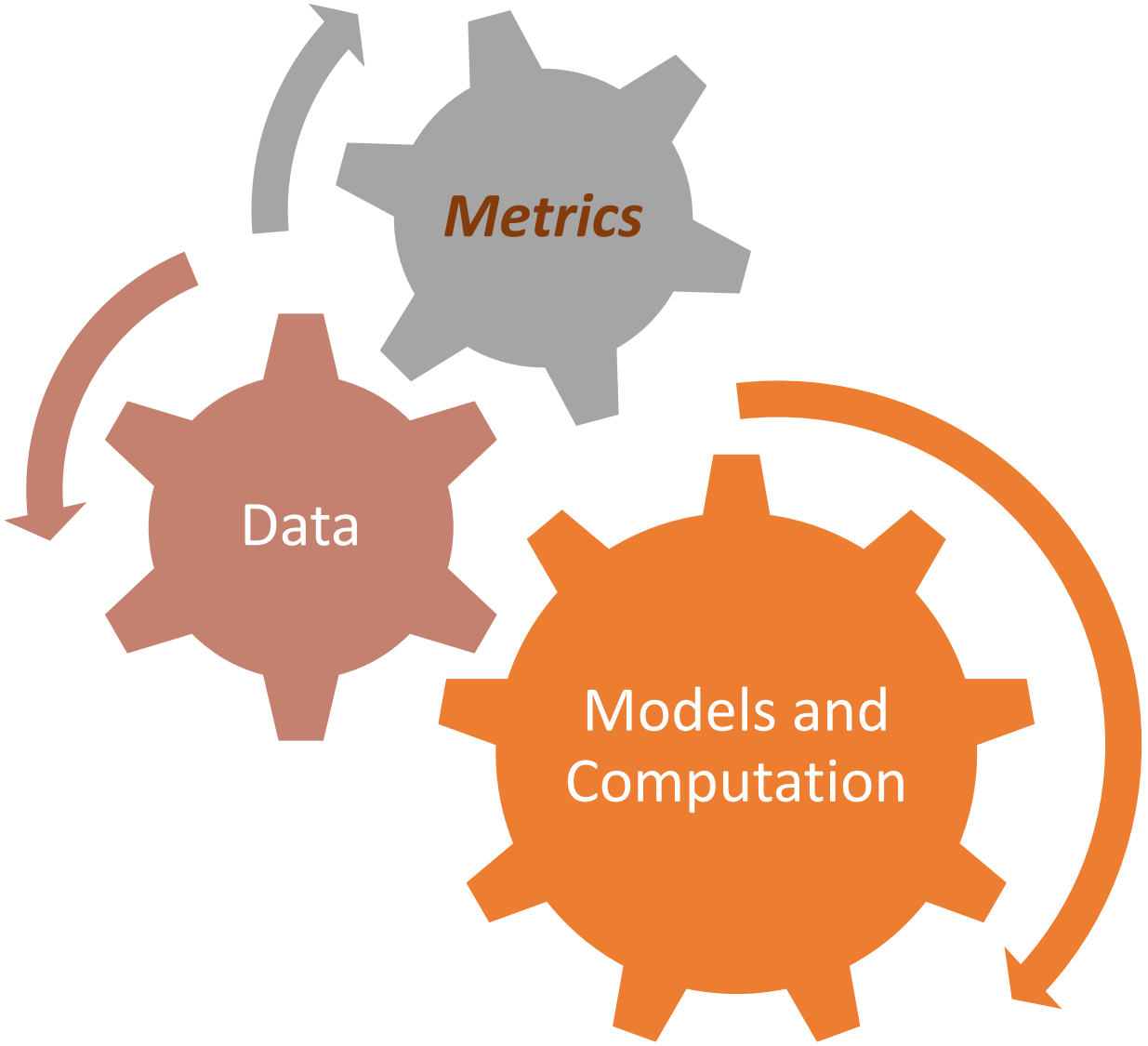
Collaborators
@Illinois: Juho Kim,
Sarah Christensen,
Mohammed El-Kebir
@Mayo: Nick Chia

Elucidating Patterns of Cancer Evolution

- Sequencing is used to measure mutations in patients
- **Goal:** Resolve ambiguity and recover evolutionary patterns, i.e., phylogenetic tree
- Clustering patients based on evolutionary trees resolves shared patterns, enables targeted treatments

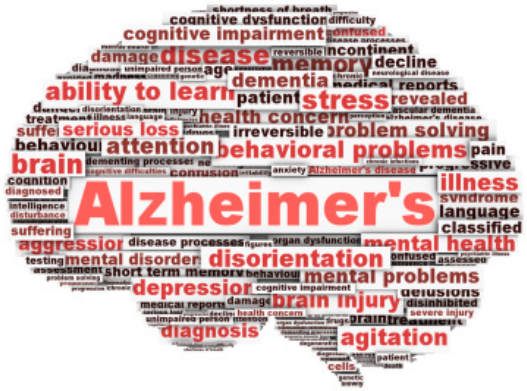


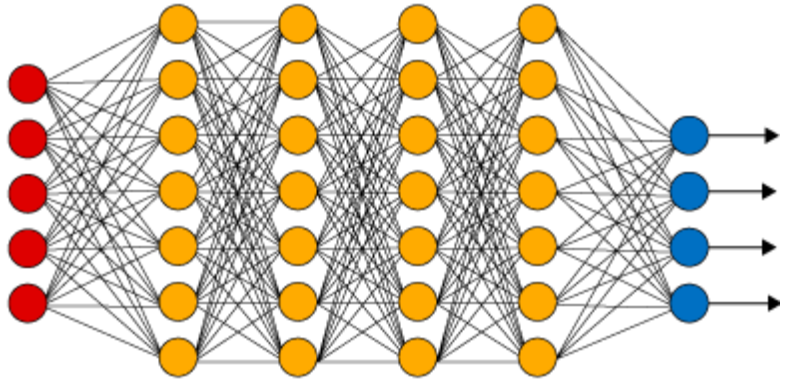
Evaluating performance and model selection



Choosing the right metrics for healthcare ML

Collaborators
@Illinois:
Gaurush Hiranandani
Shant Boodaghians
Ruta Mehta



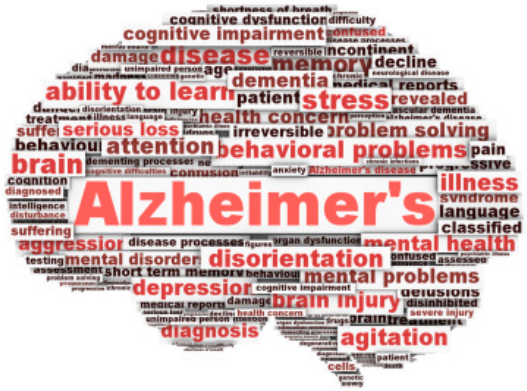


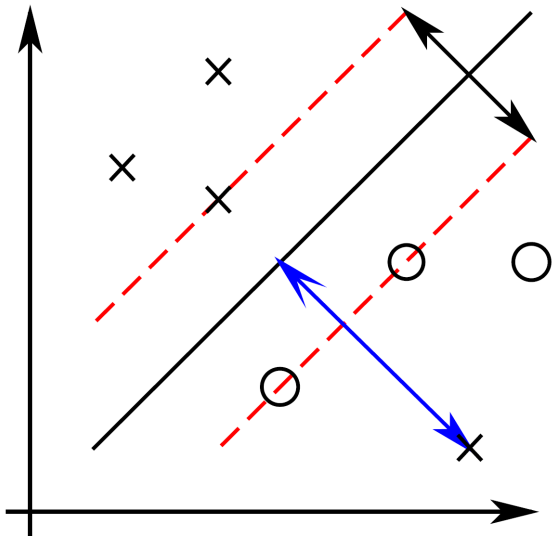
- Accuracy = 94.1%

Positive = Healthy; Negative = Alzheimer's

- False positive rate = Predict healthy when patient has Alzheimer's = 90%
- False negative rate = Predict Alzheimer's when patient is healthy = 5%

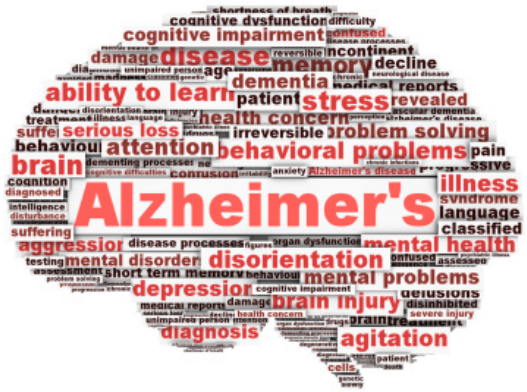
Image Source: https://becominghuman.ai/deep-learning-made-easy-with-deep-cognition-403fbe445351





- Accuracy = 80.1%

- False positive rate = Predict healthy when patient has Alzheimer's = 10%
- False negative rate = Predict Alzheimer's when patient is healthy = 20%

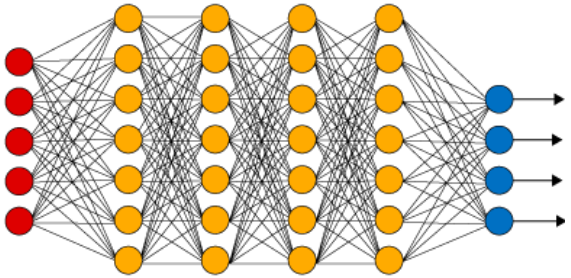


Always Predict Healthy

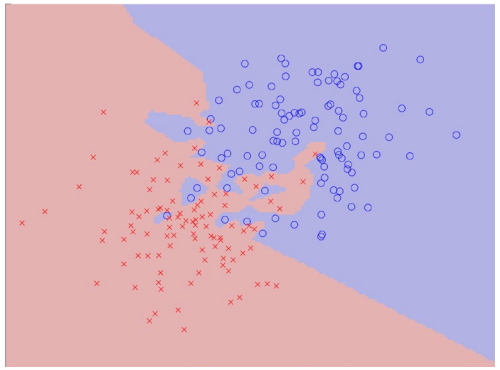
- Accuracy = 99%

- Prevalence of Alzheimer's disease is <1% of the population*
- False positive rate = Predict healthy when patient has Alzheimer’s = 100%
- False negative rate = Predict Alzheimer’s when patient is healthy = 0%

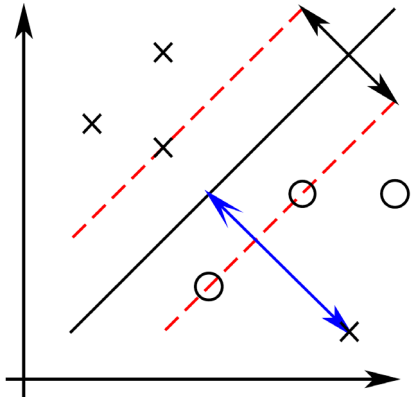
* https://www.alz.org/facts/



- 94.1% Accuracy
- 90% false positives
- 5% false negatives



- 89.6% Accuracy
- 50% false positives
- 1% false negatives



- 80.1% Accuracy
- 10% false positives
- 20% false negatives

Always Predict Healthy

- 99% Accuracy
- 100% false positives
- 0% false negatives

Which ML model should you use?

HOW SHOULD YOU MEASURE THE PERFORMANCE OF YOUR ML MODEL?

It depends… on the relative cost/benefit of different kinds of errors.

The **metric** is a quantitative description of tradeoffs -- used to compare models, or optimized directly.

sklearn.metrics : Metrics

Regression metrics

See the Regression metrics section of the user guide for further details.

metrics.explained_variance_score (y_true, y_pred)	Explained variance regres
metrics.mean_absolute_error (y_true, y_pred)	Mean absolute error regre
metrics.mean_squared_error (y_true, y_pred[, …])	Mean squared error regre
metrics.mean_squared_log_error (y_true, y_pred)	Mean squared logarithmic
metrics.median_absolute_error (y_true, y_pred)	Median absolute error reg
metrics.r2_score (y_true, y_pred[, …])	R^2 (coefficient of determ

Multilabel ranking metrics

See the Multilabel ranking metrics section of the user guide for further details

metrics.coverage_error (y_true, y_score[, …])	Coverage error mea
metrics.label_ranking_average_precision_score (…)	Compute ranking-ba
metrics.label_ranking_loss (y_true, y_score)	Compute Ranking lo

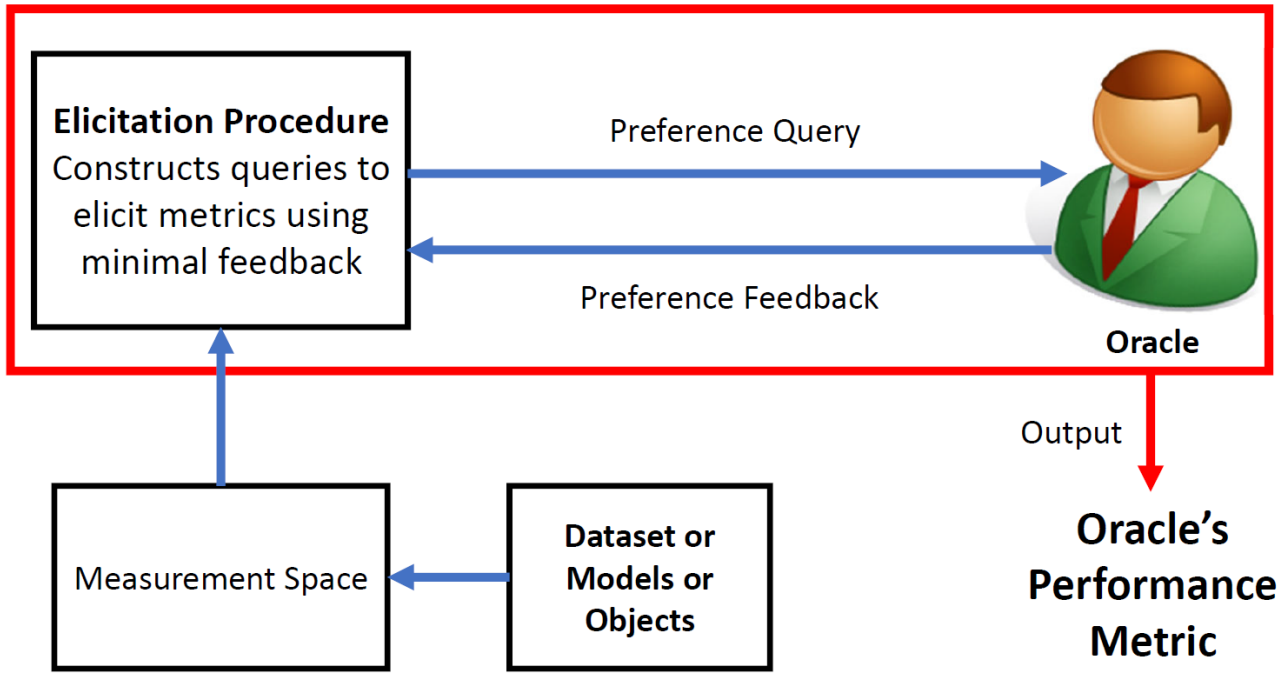
Clustering metrics

See the Clustering performance evaluation section of the user guide for further details.

The sklearn.metrics.cluster submodule contains evaluation metrics for cluster analysis results. There are two forms of evaluation:

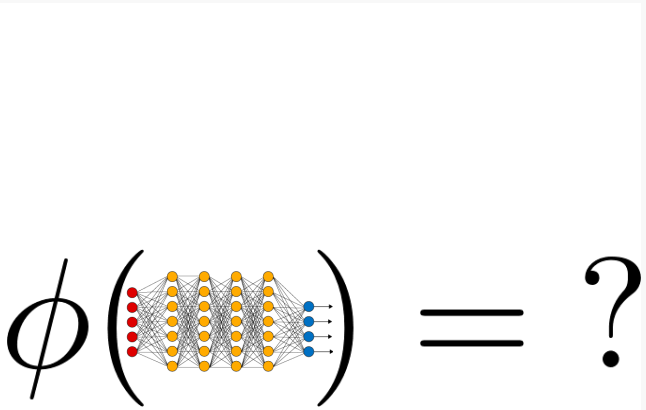
- supervised, which uses a ground truth class values for each sample.
- unsupervised, which does not and measures the 'quality' of the model itself.

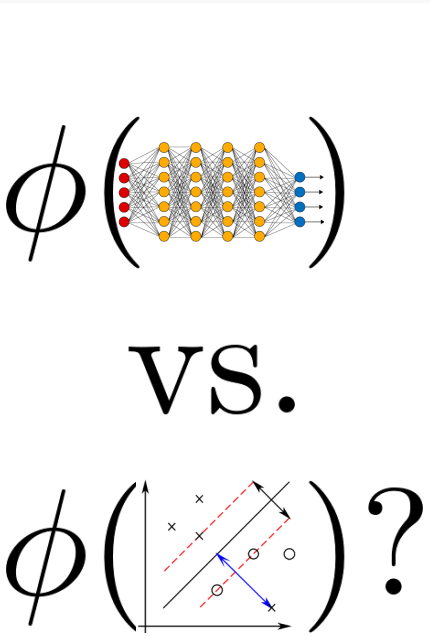
metrics.adjusted_mutual_info_score (…[, …])	Adjusted Mutual Information between two clusterings.
metrics.adjusted_rand_score (labels_true, …)	Rand index adjusted for chance.
metrics.calinski_harabaz_score (X, labels)	Compute the Calinski and Harabaz score.
metrics.davies_bouldin_score (X, labels)	Computes the Davies-Bouldin score.
metrics.completeness_score (labels_true, …)	Completeness metric of a cluster labeling given a ground truth.
metrics.cluster.contingency_matrix (…[, …])	Build a contingency matrix describing the relationship between labels.
metrics.fowlkes_mallows_score (labels_true, …)	Measure the similarity of two clusterings of a set of points.
metrics.homogeneity_completeness_v_measure (…)	Compute the homogeneity and completeness and V-Measure scores at once.
metrics.homogeneity_score (labels_true, …)	Homogeneity metric of a cluster labeling given a ground truth.
metrics.mutual_info_score (labels_true, …)	Mutual Information between two clusterings.
metrics.normalized_mutual_info_score (…[, …])	Normalized Mutual Information between two clusterings.
metrics.silhouette_score (X, labels[, …])	Compute the mean Silhouette Coefficient of all samples.
metrics.silhouette_samples (X, labels[, metric])	Compute the Silhouette Coefficient for each sample.
metrics.v_measure_score (labels_true, labels_pred)	V-measure cluster labeling given a ground truth.



Metric Elicitation

EFFICIENTLY QUERY AN EXPERT TO QUANTIFY UTILITY OF ML MODELS





Querying the expert

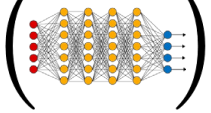
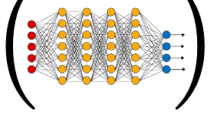
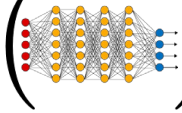
EXPERTS ARE OFTEN INACCURATE WHEN ASKED TO QUANTIFY VALUE

TOO MANY QUERIES MAY RESULT IN FATIGUE

Goal:
accurately elicit the expert's metric using a few pairwise queries

Example: Binary Classification, Linear Metrics

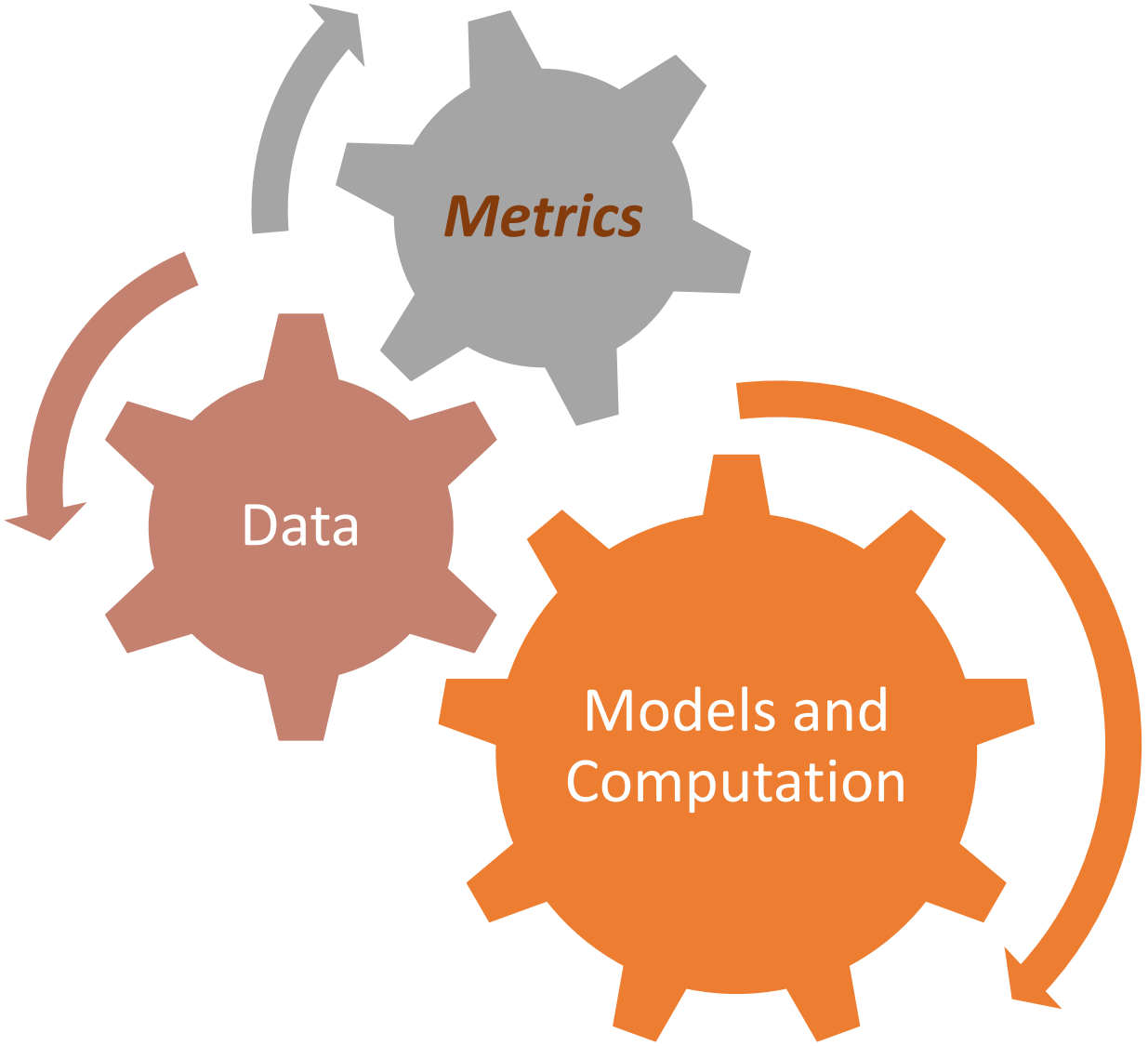
- Binary search elicitation provably recovers the expert's weighted metric:

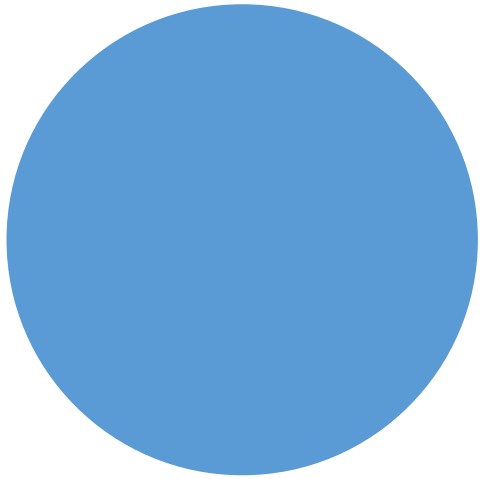
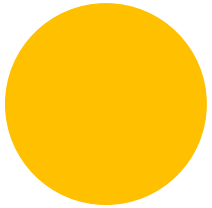
$$\phi^*(\text{  }) = 1 - (a_1^* \mathrm{FP}(\text{  }) + a_2^* \mathrm{FN}(\text{  }))$$

- Guaranteed to be ϵ accurate after $\mathcal{O}\left(\log(\frac{1}{\epsilon})\right)$ queries
- Achieves the theoretical optimal elicitation rate
- Stable to system noise e.g. noisy responses from the expert

Hiranandani, Boodaghians, Mehta, **K.** (2019) x2

Explainability and Trust



Interpreting Machine Learning Using Examples

Collaborators
@Vector: Shalmali Joshi
@Berkeley: Rajiv Khanna
@Google: Been Kim
@Texas: Joydeep Ghosh

Why do we care about transparency and interpretability in ML?

For ML experts:

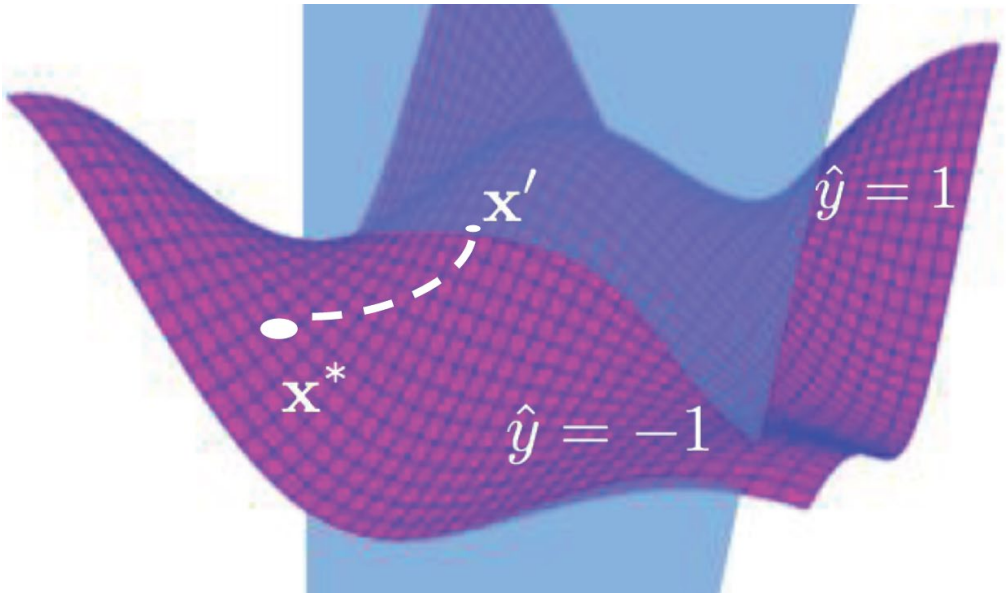
- debugging trained models.

For healthcare professionals:

- the key to discovery e.g. scientific applications,
- useful for detecting failure and corner cases.

For everyone else:

- ensure that predictions are fair, non-discriminatory,
- actionable recourse i.e. how do I change the prediction outcome?



REVISE

What is the smallest “realistic” change in input that modifies the model prediction?

- Probing healthcare ML systems for counterfactuals
- Components
  - generative model of data distribution
  - algorithmic decision, i.e., classifier
  - constrained optimization to identify recourse

Modeling	• (Brain) dynamics, longitudinal tracking, diagnosis • **Applications: Glioma segmentation, Cancer phylogenetics**
Evaluation	• **Selecting good metrics for machine learning** • Training models that optimize specialized metrics
Privacy	• **Data synthesis,** learning with aggregated data • **Learning on the edge**
Trust	• Explainability and interpretability using examples • **Individual recourse**

Enabling Technologies

Thank you

sanmi@Illinois.edu
@sanmikoyejo