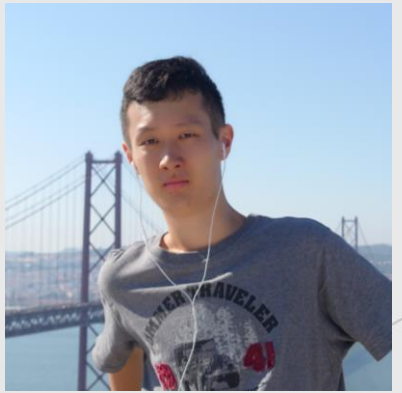
2020 Hackathon

Jiajun He, Zelin Li

01

Who are we?

Team Spaghetti Vector Monster (SVM)



Jiajun He **Zelin Li**

Major in Bioinformatics, at UCPH

02

What we did?

1

VERIFICATION

Verify whether the synthetic data can be used for data analysis

2

CLASSIFICATION

Train classifier to distinguish the two sets of data

3

ANALYSIS

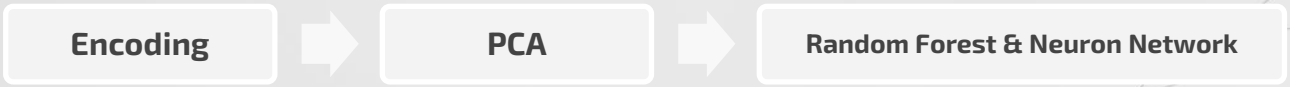
Find which dimension(s) makes two set of data different

01 VERIFICATION

IDEA

- Using synthetic data, train a model to predict readmission.
- Check the performance of this model on synthetic data and on real data.

MODEL



01 VERIFICATION

RESULT

Synthetic data (Train set)	Synthetic data (Val set)	Real data
71%	68%	53%

- There is a **significant difference** between 2 datasets.
- Real data is **harder** to make prediction. Using this set of synthetic data to do data analysis is **risky**.

02 CLASSIFICATION

MODEL

Encoding → PCA → Neuron Network

RESULT

	Training set	Validation set
Accuracy	90%	88%

02 CLASSIFICATION

MODEL

Encoding → PCA → Neuron Network

RESULT

	Training set	Validation set
Accuracy	90%	88%

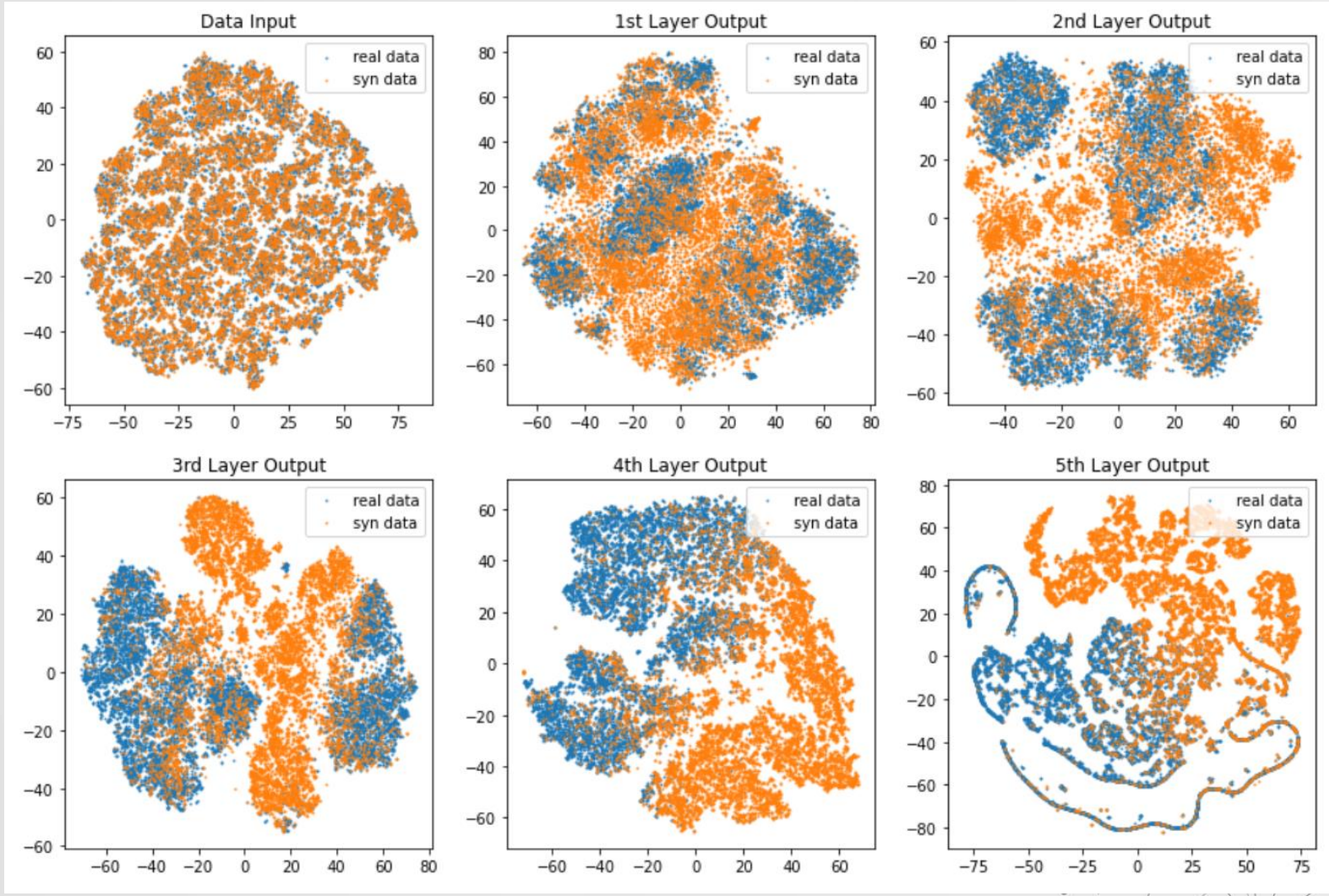
03 ANALYSIS

- **Low dimensional embedding**
  – check decoupling ability of the model
- **Check hidden-layer activation**
  – find features the decoupling layers are detecting during prediction
- **Re-train the model using combination of dimensions**
  – check the most crucial dimension
- **Analyze crucial dimensions**
  – find interpretable difference between real data and synthetic data

3.1 Low dimension embedding

- Embedding and visualization by t-SNE

- Check the decoupling of each layer in neuron network


Data Input
real data
syn data
1st Layer Output
real data
syn data
2nd Layer Output
real data
syn data
3rd Layer Output
real data
syn data
4th Layer Output
real data
syn data
5th Layer Output
real data
syn data

3.2 Check hidden layer activation

- Try to find which features the first layer detects

Find the samples that mostly activate the first layer → **Recover the input by taking pseudo-inverse** → **Compare the difference**

$$A = relu(W \cdot X + b)$$
$$X \approx W^{-1} \cdot (A - b)$$

The inspiration comes from
Zeiler M.D., Fergus R. (2014) Visualizing and Understanding Convolutional Networks.

3.2 Check hidden layer activation

- By comparing the result, we teased out dimensions not used in decoupling.
- Remain 22 dimensions.

3.3 Train model with combination of dimensions

- To check the most important dimensions among the 22 dimension.
- Find 3 most important dimensions:

“insulin”
“change”
“diabetesMed”

3.3 Train model with combination of dimensions

	All features	Only 3 features	All other features
Accuracy	88%	72%	74%

3.4 Understand the difference

- Compare the correlation coefficient of this 3 dimensions.

Real data

Corrs	insulin	change	diabetesMed
insulin	1.0	-0.14	0.26
change	-0.14	1.0	-0.51
diabetesMed	0.26	-0.51	1.0

Synthetic data

Corrs	insulin	change	diabetesMed
insulin	1.0	-0.02	0.01
change	-0.02	1.0	0.02
diabetesMed	0.01	0.02	1.0

3.5 Conclusion

- In real data, these three dimensions are **highly related**;

- while in synthetic data, these three dimensions seems to be generated **independently**.

Future Plan

- Re-sampling these 3 dimensions to generate a better synthetic data

- Use this new set of synthetic data to train a model predicting re-admission, then check if this model works well on real data

Thank you for listening