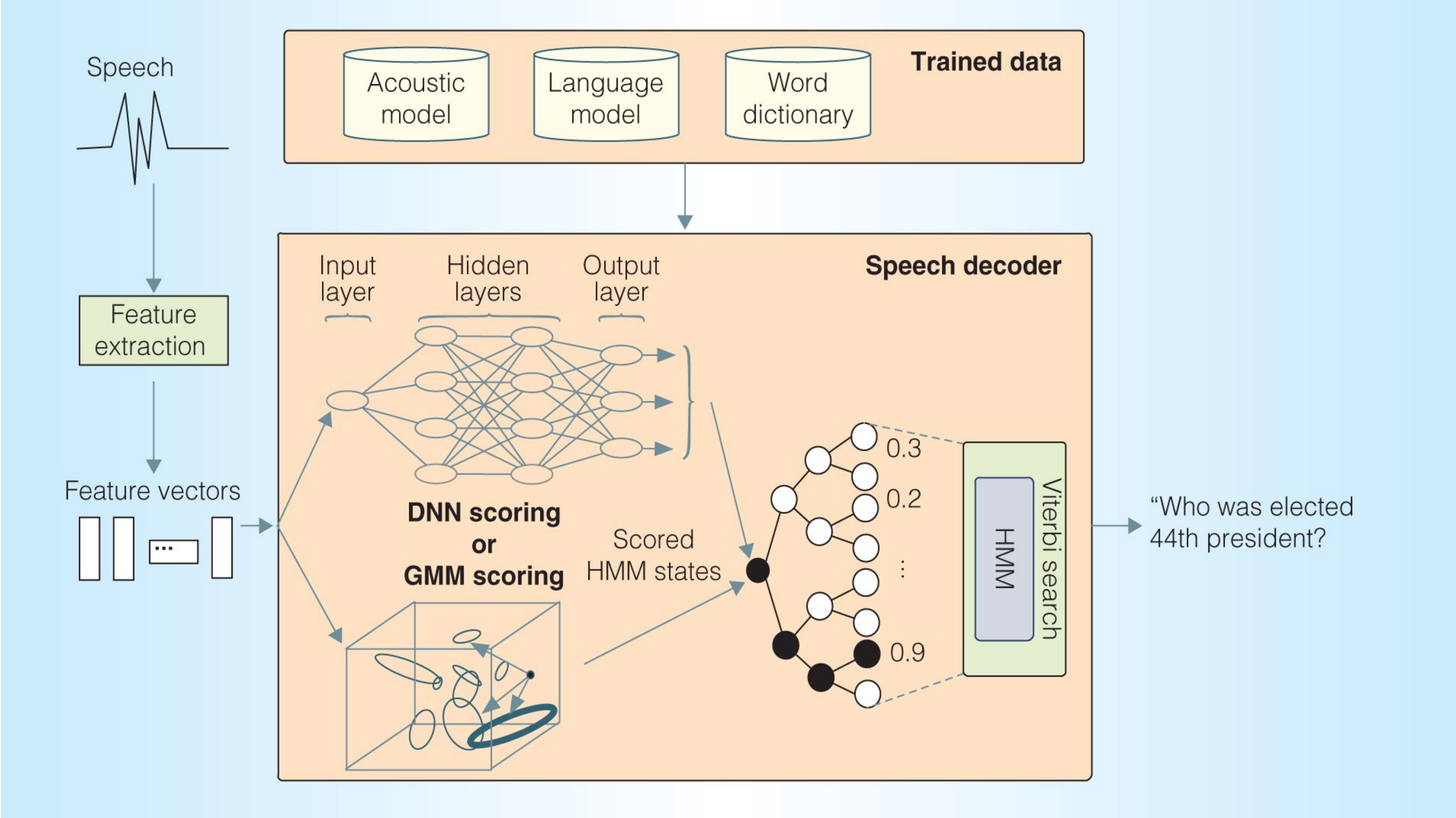
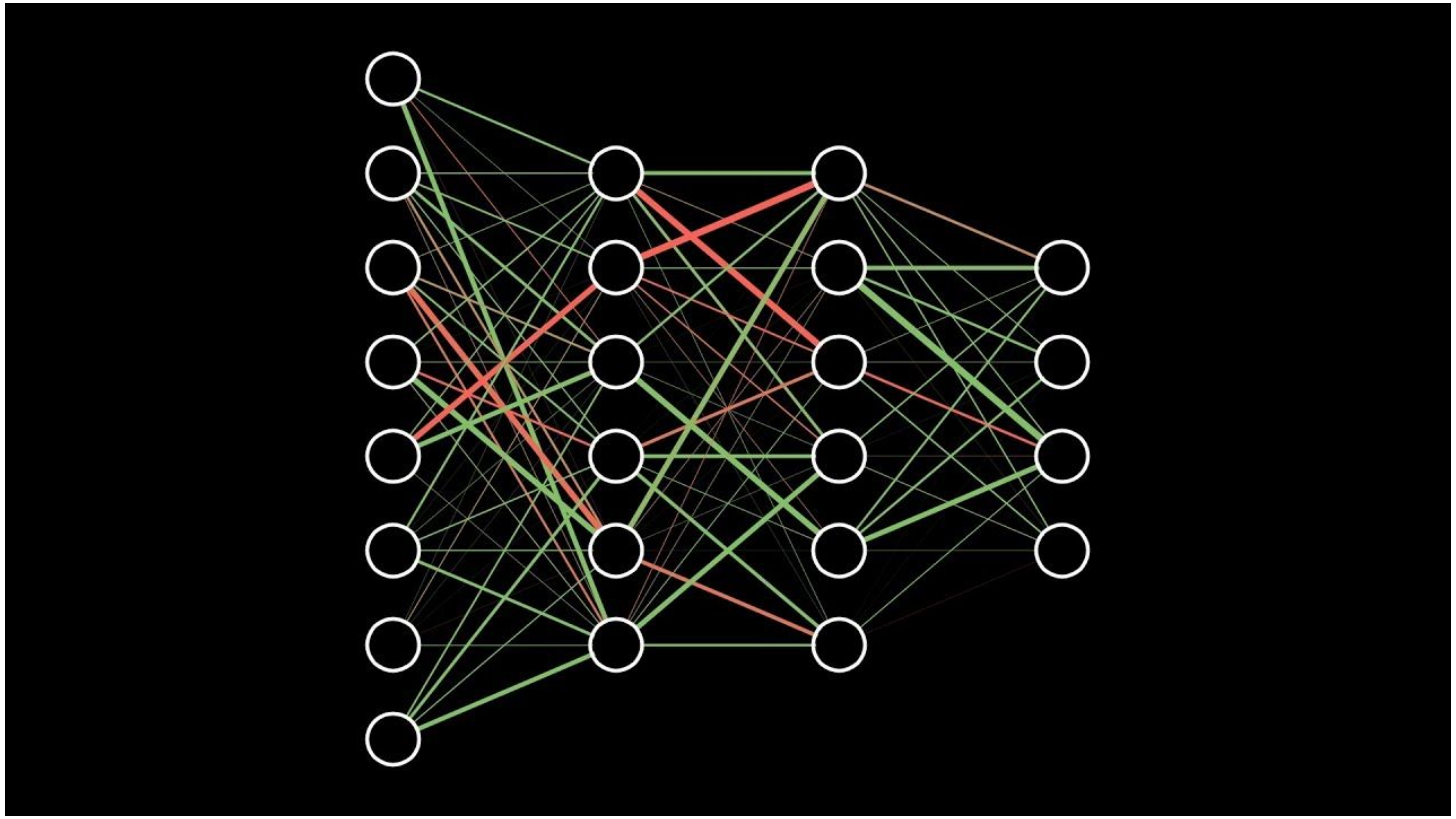
A New Method for the Exploitation of Speech Recognition Systems

Suha Hussain
Grade 11
Queens High School for the Sciences at York College
Mentors: Professor Ramesh Karri and Zahra Ghodsi at New York University

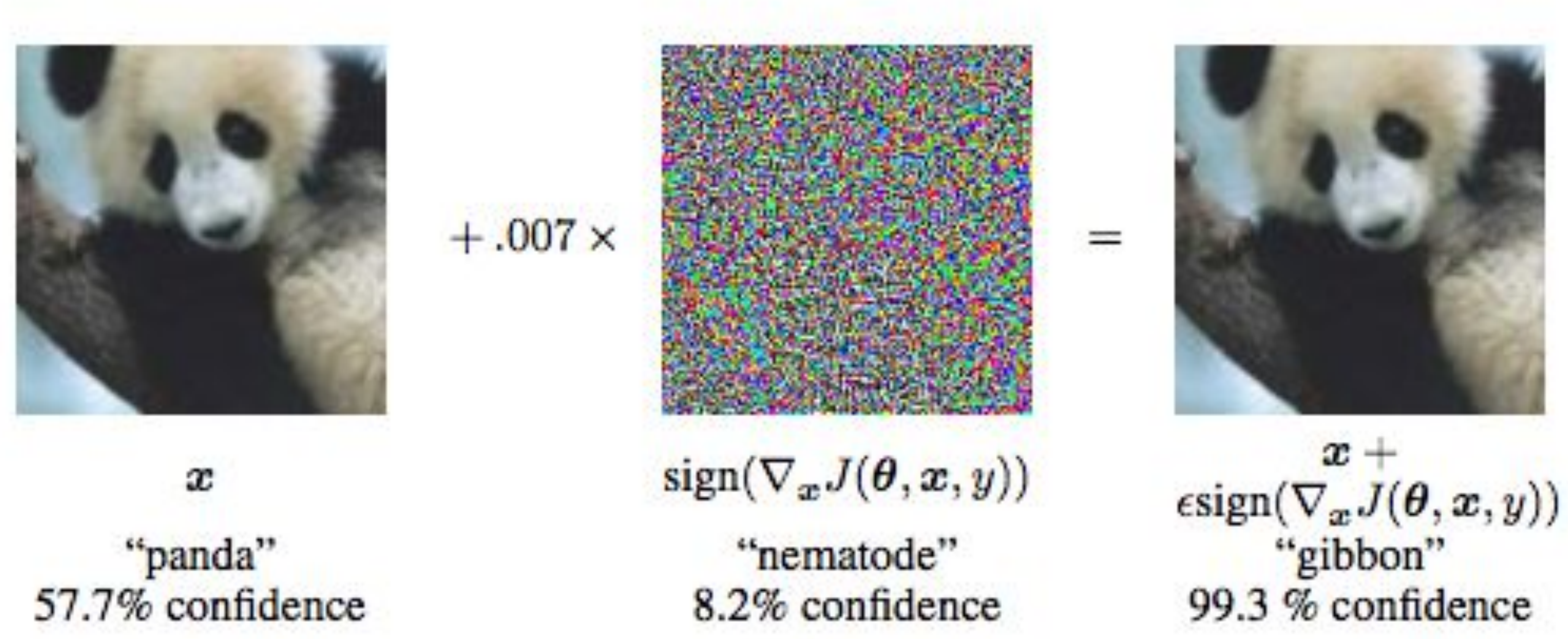
Background Information



Background Information



Background Information



Background Information

- ➔ Niedek (2016)
- ➔ Carlini (2016)
- ➔ Zhang (2017)

Background Information

- → Moosavi-Dezfooli (2016)
- → Papernot (2017)
- → Athalye (2017)

Purpose

- ➔ In order to develop secure speech recognition systems, vulnerabilities must be discovered and mitigated.
- ➔ Developing an exploit based upon adversarial machine learning would highlight the vulnerabilities associated with internal neural networks.

Threat Model

- ➔ The adversary has access to the speech recognition system after training is complete.
- ➔ There is adequate time to implement the adversarial algorithm on any speech recognition system.
- ➔ The adversary can add noise vectors to the input of the system.
- ➔ The scenario is black-box.

Algorithm Design

$$F(x, y) = y$$

$$F((x + v), \hat{y}) = \hat{y}$$

$$\|v_2\| \leq \varsigma$$

$$r(x) = argmin\|x_2\| \leq \varsigma \; subject \; to \; F((x + v), \hat{y}) = \hat{y}$$

Algorithm Design

$$\varsigma_r = \varsigma_F(w(L2(x+v))$$

$$r(x) = argmin||x_2|| \leq \varsigma \; subject \; to \; \frac{1}{k}\sum_{i=1}^{k} F((x+v),\hat{y}) = \hat{y}$$

$$v \leftarrow P_p(v+v_i)$$

Algorithm Design

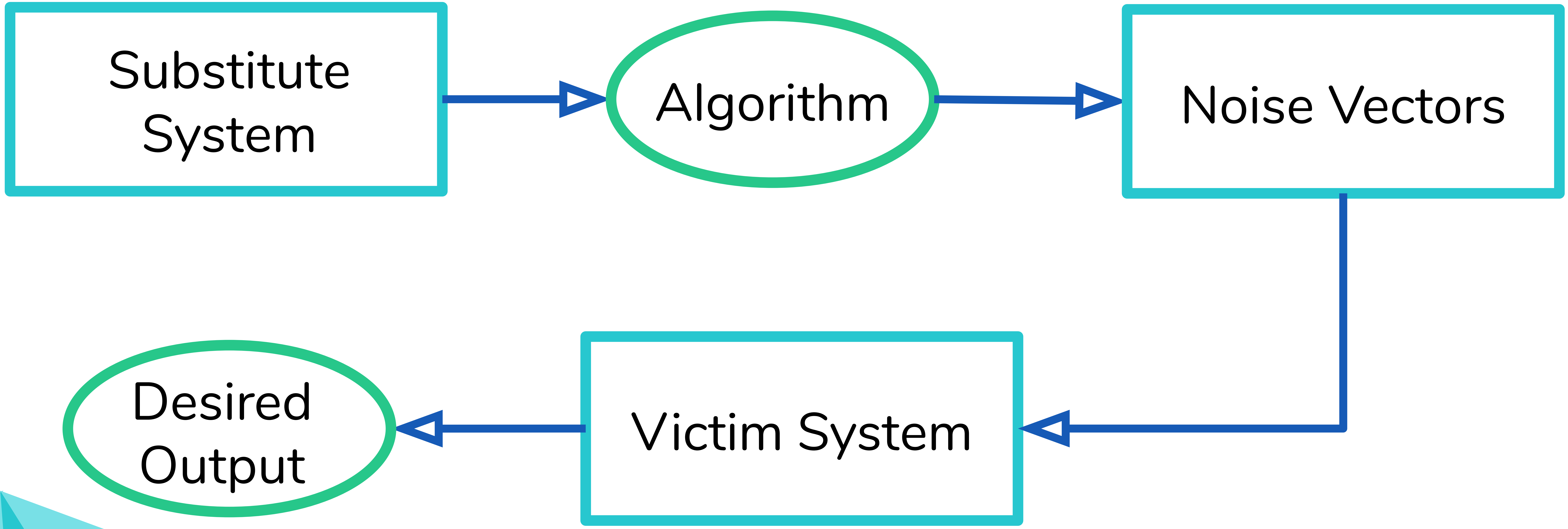
Computing Universal, Transformable Perturbation Vectors for a Specific Target Class:

Input: Data X with data points x_i, neural network F, norm of perturbation ς, desired target class ŷ, maximum iterations I

Output: Universal, transformable noise vector v

1. Initialize $v \leftarrow 0$
2. Initialize $i \leftarrow 0$
3. While $i \leq I$
4. For each datapoint do
5. If $F((x+v),\hat{y}) = \hat{y}$
6. Compute the minimal perturbation that sends input to decision boundary while incorporating gradient sampling:
   $r(x) = argmin\|x_2\| \leq \varsigma \ subject \ to \ \frac{1}{k}\sum_{i=1}^{k} F((x+v),\hat{y}) = \hat{y}$
7. Update the perturbation: $v \leftarrow P_p(v+v_i)$
8. End if
9. End for
10. End while

Attack Overview



Instrumentation

- ➔ Python 2.7
- ➔ TensorFlow
- ➔ NumPy
- ➔ Jupyter Notebook
- ➔ NVIDIA GPU
- ➔ Google Sheets
- ➔ TIMIT dataset

Procedure

- ➔ Conduct preprocessing and normalization procedures on the TIMIT dataset.
- ➔ Program two fully connected neural networks with 5 layers and 600 hidden neurons
- ➔ Designate one as the substitute system and one as the victim system.
- ➔ Train the networks on different subsets of TIMIT for 200000 iterations using the Adagrad optimizer.

Procedure

- ➔ Construct validation sets from TIMIT for five randomly chosen targets for the victim system.
- ➔ Train the proposed adversarial algorithm on the substitute system.
- ➔ Apply noise vectors to validation sets.
- ➔ Record the maximum accuracy for each validation set.

Results

Validation Set	Accuracy
1	64.08%
2	60.25%
3	59.25%
4	63.25%
5	55.25%
Average	60.42%

Limitations

- An environment to simulate a realistic situation was not developed.
- End-to-end speech recognition systems were not tested separately.
- Only the TIMIT dataset was utilized.

Future Work

- ➔ Develop defenses for adversarial exploits.
- ➔ Develop algorithms that enable real-time exploitation with less preparation.
- ➔ Develop algorithms that leverage both hidden Markov models and neural networks.
- ➔ Focus on exploiting and protecting specific speech recognition applications.

Acknowledgements

- ➔ Professor Ramesh Karri of New York University Tandon School of Engineering
- ➔ Zahra Ghodsi of New York University Tandon School of Engineering
- ➔ Jose Mondestin of Queens High School for the Sciences at York College

References

Niedek, T. V. (2016). Phonetic Classification in TensorFlow (Bachelor's thesis). Radboud University.

Carlini, N. et al.(2016). Hidden Voice Commands. 25th USENIX Security Symposium, 513-530. Retrieved from https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/carlini

Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., & Xu, W. (2017). DolphinAttack. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security - CCS 17*. doi:10.1145/3133956.3134052

Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O., & Frossard, P. (2017). Universal Adversarial Perturbations. *2017 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR). doi:10.1109/cvpr.2017.17

Papernot, N., Mcdaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical Black-Box Attacks against Machine Learning. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security - ASIA CCS 17*. doi:10.1145/3052973.3053009

Thank you for your time!

Any questions?