Vocal Processing with Spectral Analysis

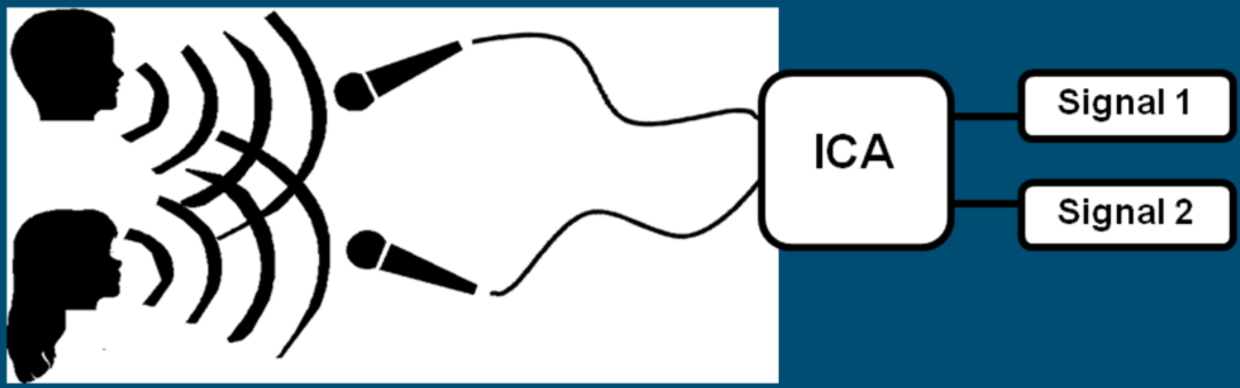
Brad Fitzgerald
Advisor: Professor Joe Makarewicz
Scholar Week – 18 April, 2018

The Cocktail Party Problem

Illustration of Speech in Crowded Room Scenario

History of Analysis

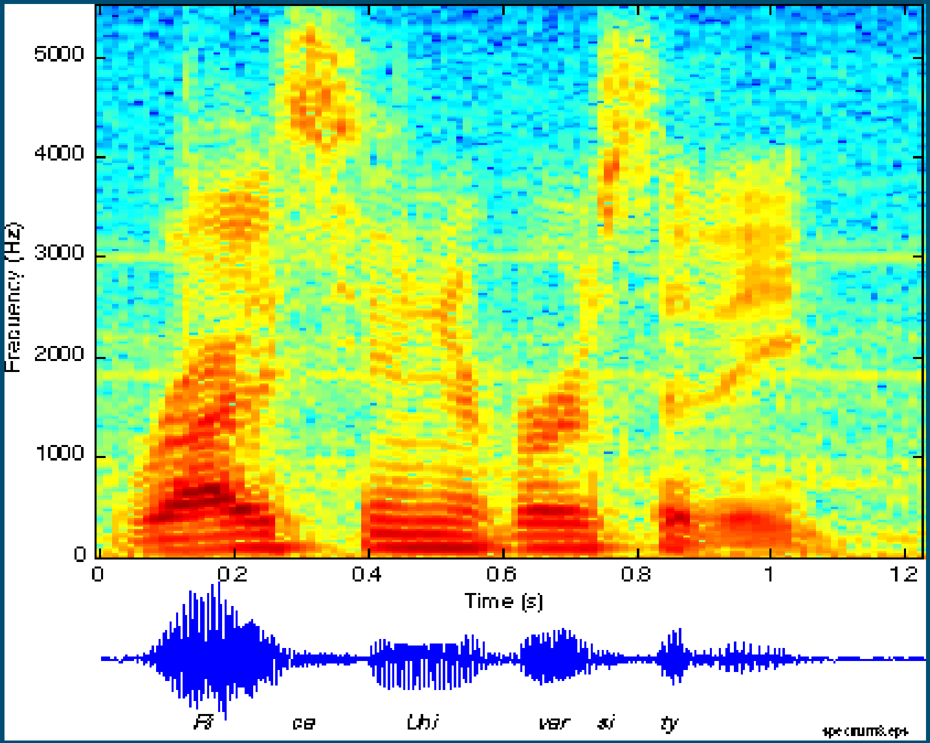
- Little standardization [1]
- Blind Source Separation
  - Optimally requires no prior signal data [2]
  - ICA = Independent Component Analysis
  - PCA = Principal Component Analysis



Independent Component Analysis Scheme

History of Analysis

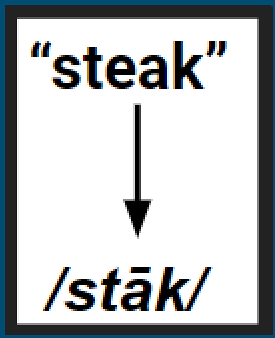
- Fourier Transform
  - Converts signal to frequency domain
  - Allows for spectral analysis [3]



Frequency Spectra Example

Linguistic Theory

- Phoneme – basic unit of speech [4]
  - Examples: */a/* , */t/* , */ch/* , */ng/*
- Phone – further breakdown of speech [4]
  - Example: */t/* pronunciation varies in *steak* vs. *top*



Sample Phonetic Breakdown

Key Prediction: Individuals have unique characteristics in their pronunciation of phonemes/phones

The Question(s):

Can principal component analysis of spectral voice data be used to identify differences between speakers?

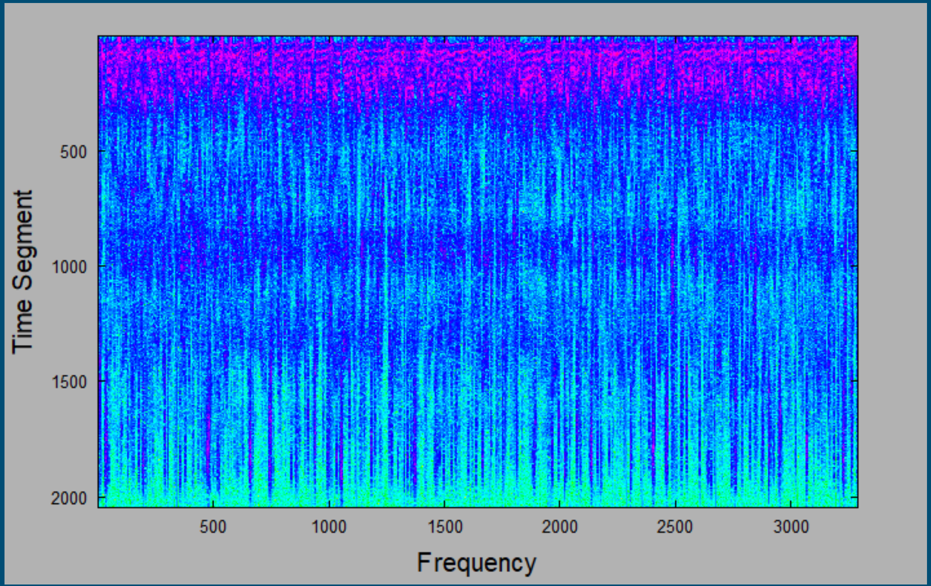
Can such differences be used to develop an algorithm which separates a mixture of vocal signals?

Methodology – Data Collection

- Recorded speech samples from 30 participants
  - 16 Male, 14 Female
- Participants read short story titled “Arthur the Rat”
  - Used by Dictionary of American Regional English[5]
  - Offers full phonetic representation of American English

Methodology – Data Processing

- Speech signal broken up into 2500-3500 time segments
- Fast Fourier Transform performed on each segment
  - Transforms signal to frequency domain for singular value decomposition

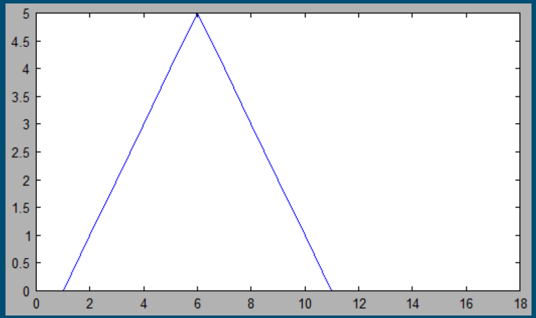


Person 1 Frequency Spectra

Methodology – Data Processing

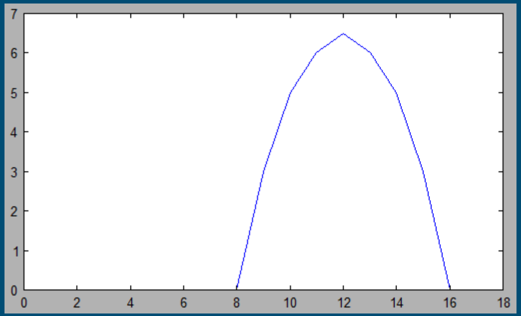
- Principal Component Analysis – using singular value decomposition (SVD) to break up a signal into:
  - Principal Vectors – “building blocks” of a signal
  - Principal Value – corresponding magnitude of a value

Methodology – SVD Explained



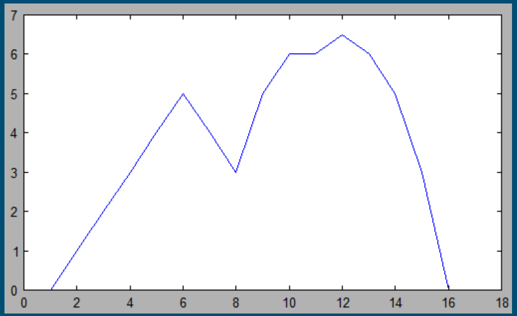
Vector 1

+



Vector 2

=



Mixed Signal

$$Vector_1 * Value_1 \quad + \quad Vector_2 * Value_2 \quad = \quad Mixed\ Signal$$

Methodology – SVD

- SVD on all 30 speakers = principal vector set for each
- Compiled 50 most significant principal vectors from all 30 sets
  - Performed SVD on combined principal vectors, producing finalized set of principal vectors representative of all 30 speakers
- Using final principal vectors, created projection matrix
  - Average principal values for all 30 speakers

Identifying Speakers – Algorithm #1

$$M = \sum\left(\underbrace{\frac{\alpha - \mu}{\sigma}}_{Z\text{ - score}} * W\right)$$

- M = Comparable measurement → $\boxed{\text{Select speaker with lowest } M}$
- α = Measured principal value
- μ = Average speaker principal value
- σ = Speaker's standard deviation
- W = Vector weight

Identifying Speakers – Algorithm #2

$$M = \sum_{i=1}^{10} \underbrace{\left(\frac{\alpha - \mu}{\sigma}\right)}_{Z\text{ - score}}$$

- M = Comparable measurement → $\boxed{\text{Select speaker with lowest } M}$
- α = Measured principal value
- μ = Average speaker principal value
- σ = Speaker's standard deviation
- i = Vector number

Results – Algorithms 1 & 2 Accuracy

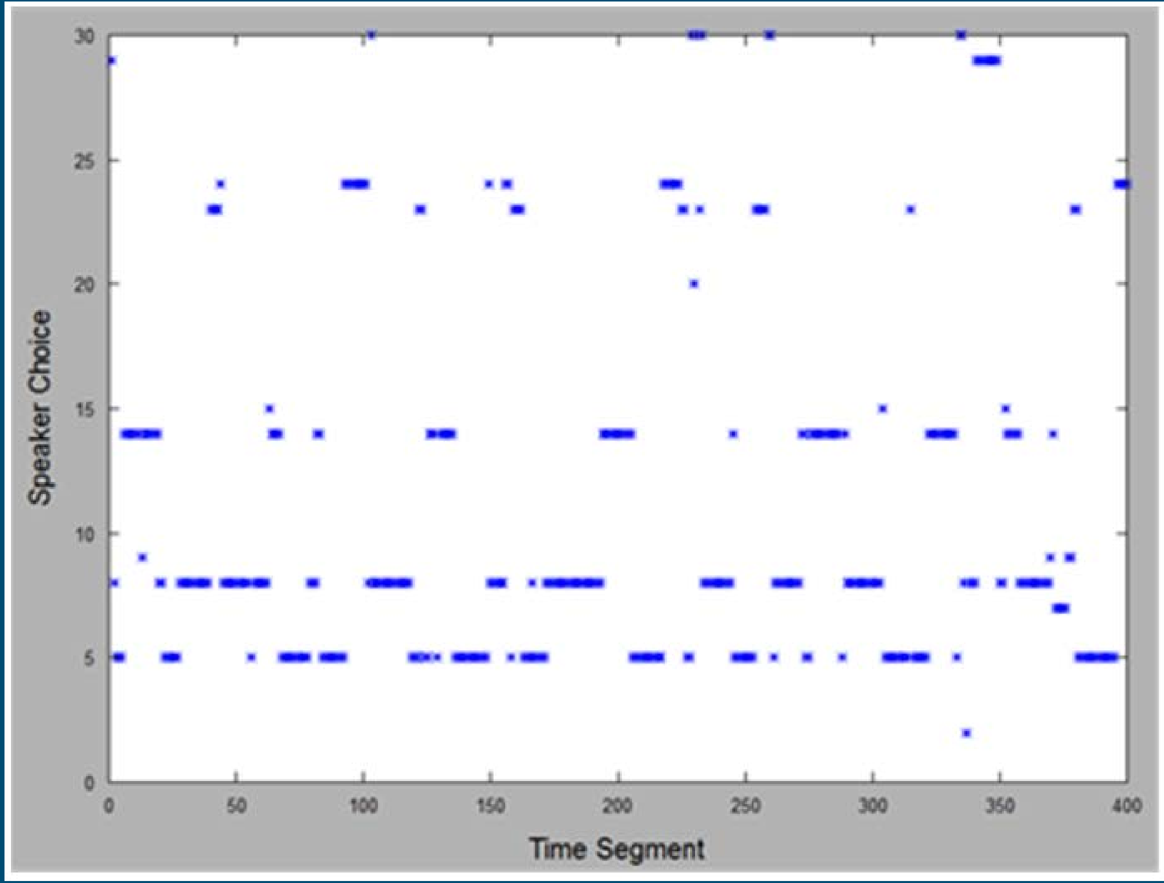
Recording	Correct Prediction Rate	Recording	Correct Prediction Rate
1	0.64%	16	1.04%
2	11.18%	17	3.34%
3	1.67%	18	15.91%
4	1.13%	19	2.04%
5	24.87%	20	1.23%
6	0.36%	21	6.61%
7	1.65%	22	23.20%
8	85.69%	23	43.59%
9	8.49%	24	73.88%
10	0.22%	25	9.69%
11	1.14%	26	2.04%
12	0.00%	27	24.38%
13	1.99%	28	1.68%
14	36.97%	29	36.82%
15	16.11%	30	18.14%

Algorithm 1 Accuracy (Single Speaker)

Recording	Correct Prediction Rate	Recording	Correct Prediction Rate
1	0.00%	16	1.87%
2	3.41%	17	5.35%
3	4.51%	18	10.67%
4	0.97%	19	4.72%
5	30.31%	20	3.88%
6	0.58%	21	3.18%
7	1.46%	22	14.59%
8	65.83%	23	40.79%
9	0.74%	24	38.45%
10	0.58%	25	17.25%
11	0.39%	26	1.37%
12	0.08%	27	14.56%
13	0.97%	28	0.12%
14	24.60%	29	8.43%
15	3.43%	30	18.14%

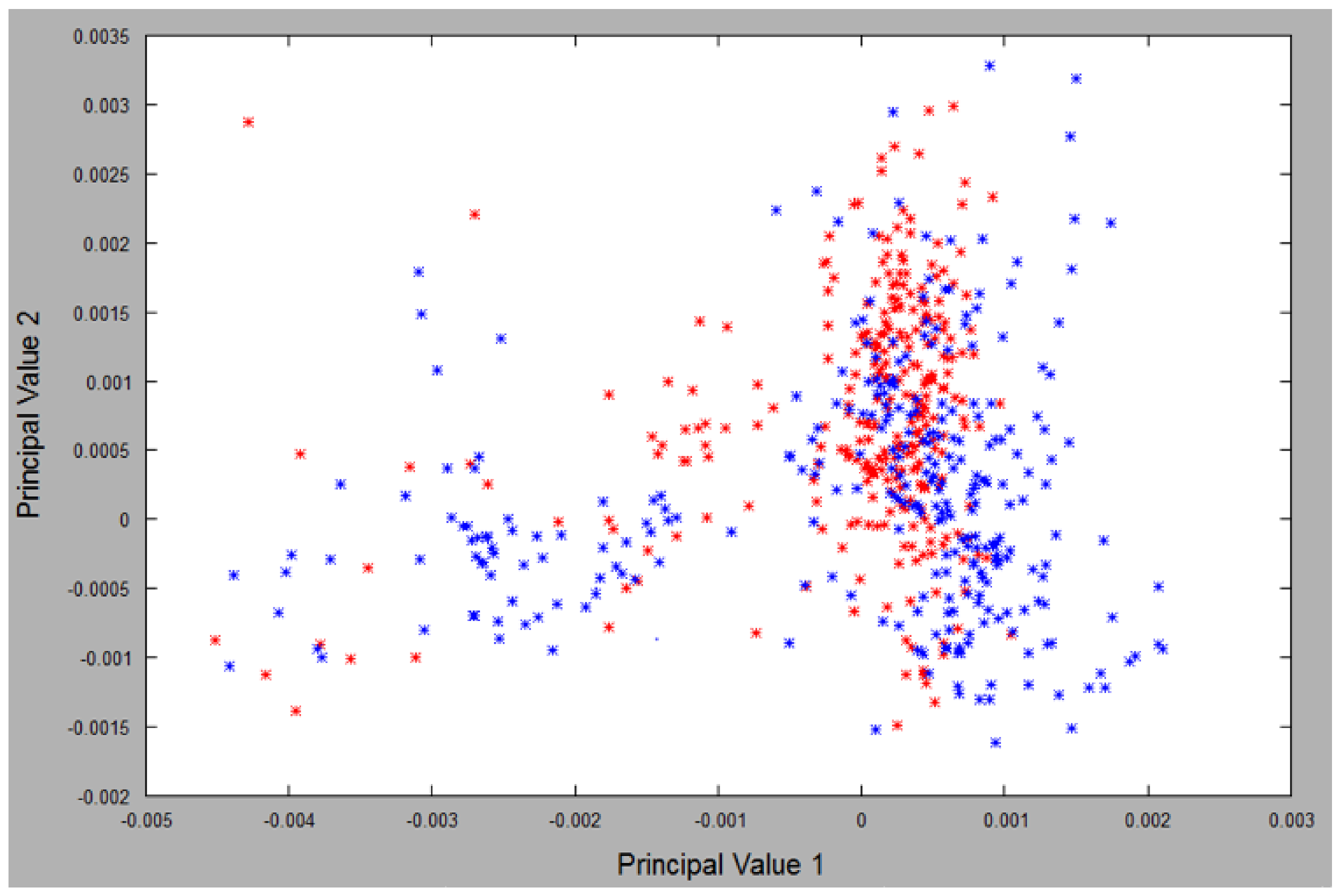
Algorithm 2 Accuracy (Single Speaker)

Results – Speaker Predictions



Algorithm 1 Identifications for Speaker 5

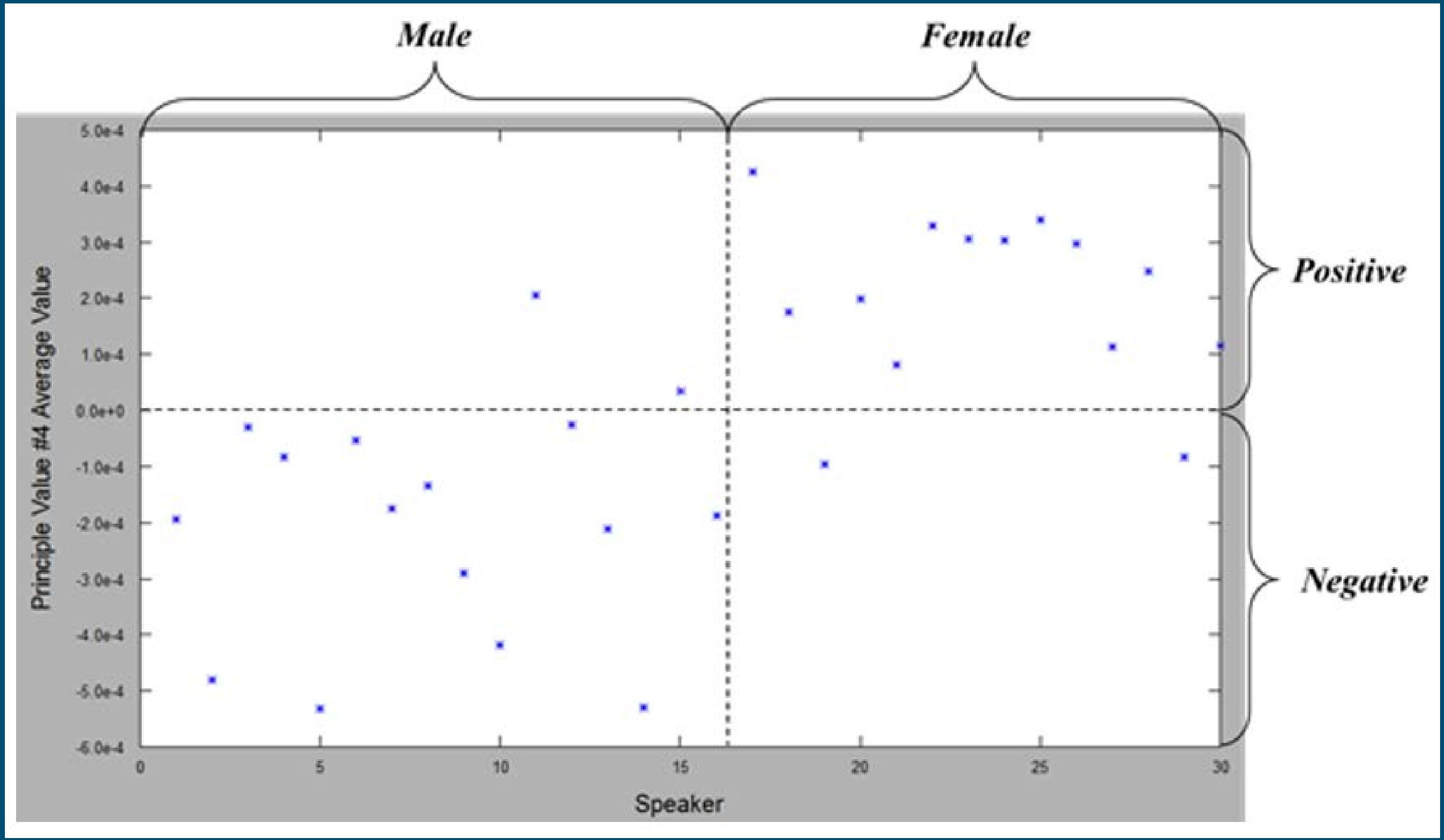
Results – Principal Values



Interaction of two principal values for Speaker 1 (blue) and Speaker 2 (red)

The principal values overlap between the two speakers for most of the region, making it difficult to use the interaction of the principal values to separate the speakers.

Results – Speaker Predictions



Principal Values (from PV#4) for males and females

Conclusions

- Algorithms 1and 2 were not successful in correctly identifying speakers
  - Algorithms tended towards guessing one specific speaker to often
  - Could not move forward to separation of mixed signals
- Principal Vector #4 = good predictor of gender
- Moving Forward
  - Revise principal component analysis process
  - Account for empty space, or pauses in speech

References

[1] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," Speech Commun., vol. 49, no. 7/8, pp. 588–601, Jul. 2007.

[2] Y. Mori et al., "Blind Separation of Acoustic Signals Combining SIMO-Model-Based Independent Component Analysis and Binary Masking," EURASIP J. Adv. Signal Process., vol. 2006, no. 1, p. 034970, Dec. 2006.

[3] H. Nakatsuji and S. Omatu, "Real Time Spectral Analysis," IEEJ Trans. Electron. Inf. Syst., vol. 126, no. 3, pp. 383–388, 2006.

[4] J. S. Bowers, N. Kazanina, and N. Andermane, "Spoken word identification involves accessing position invariant phoneme representations," J. Mem. Lang., vol. 87, pp. 71–83, Apr. 2016.

[5] Dictionary of American Regional English, (2013). "Arthur the Rat." [Online]. Available: http://dare.wisc.edu/audio/arthur-the-rat. [Accessed: 6-Sept-2016].

Acknowledgements

Thank you so much to all those who impacted and guided me through the Honors Program during my time at Olivet:

Prof. Joe Makarewicz
Dr. Steven Case
Dr. Beth Schurman
Prof. Erik Young
Dr. Brian Stipp
Dr. Dan Sharda

Thank You!

Questions?