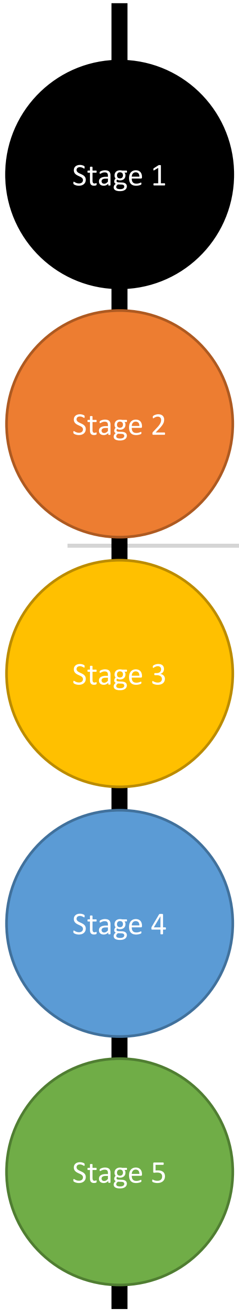
SIOP Machine Learning Competition

Presenter: **Mustafa Akben, Ph.D.**

ELON
UNIVERSITY

SIOP ANNUAL CONFERENCE

BOSTON and ONLINE • April 19–22, 2023



Stage 1

Stage 2

Stage 3

Stage 4

Stage 5

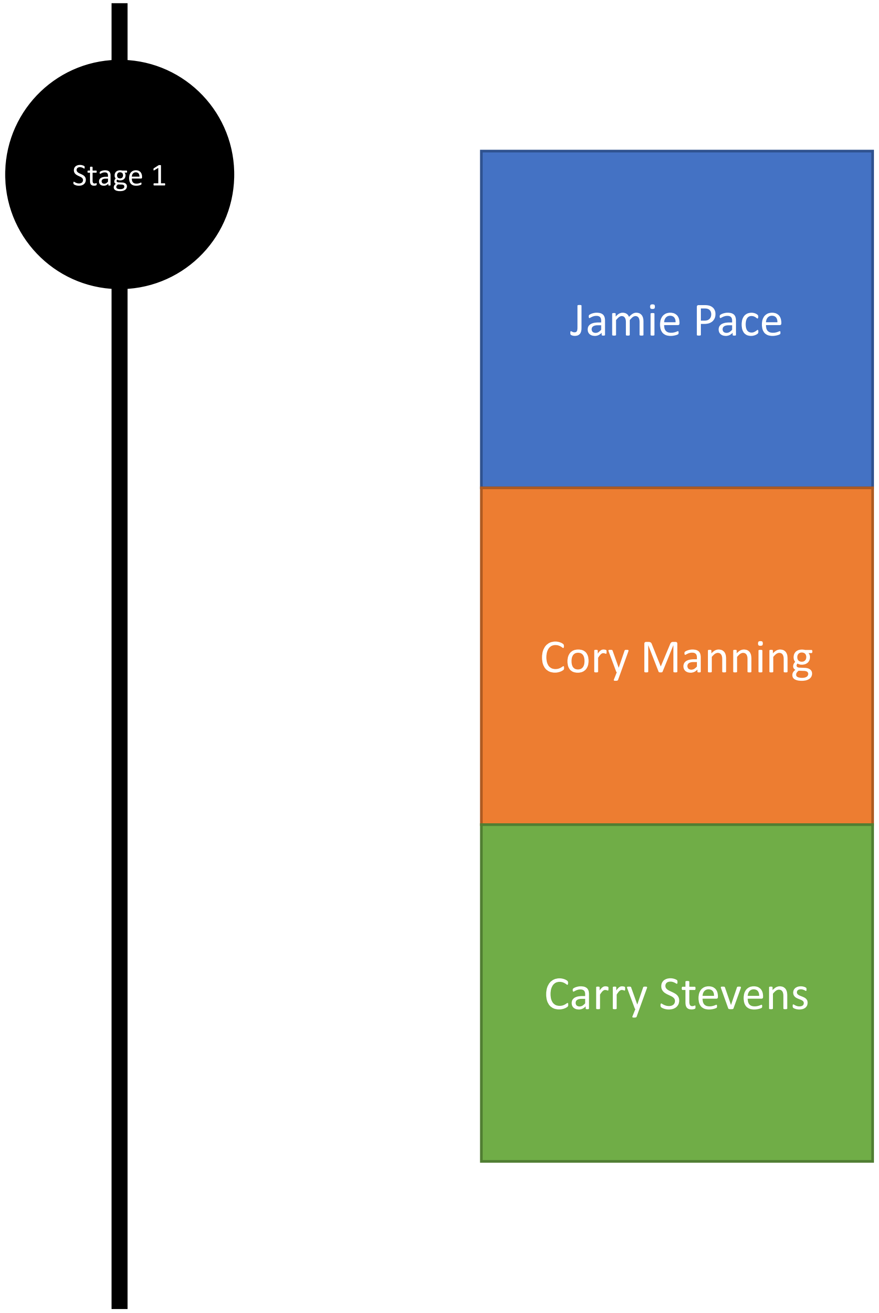
Overall Producer

- Stage 1 — Data Cleaning
- Stage 2 — Model Exploration
- Stage 3 — Feature Engineering
- Stage 4 — Model Building
- Stage 5 — Final Submission

Stage 1

Stage 1 – Data Cleaning: Dirty Data

- Understanding Data
- A lot of NA’s
- Use them for advantages – Missingness based cluster analysis to identify the differences in data
- Two different assessment center simulations with three different player names: Jamie Pace, Cary Stevens, and Cory Manning

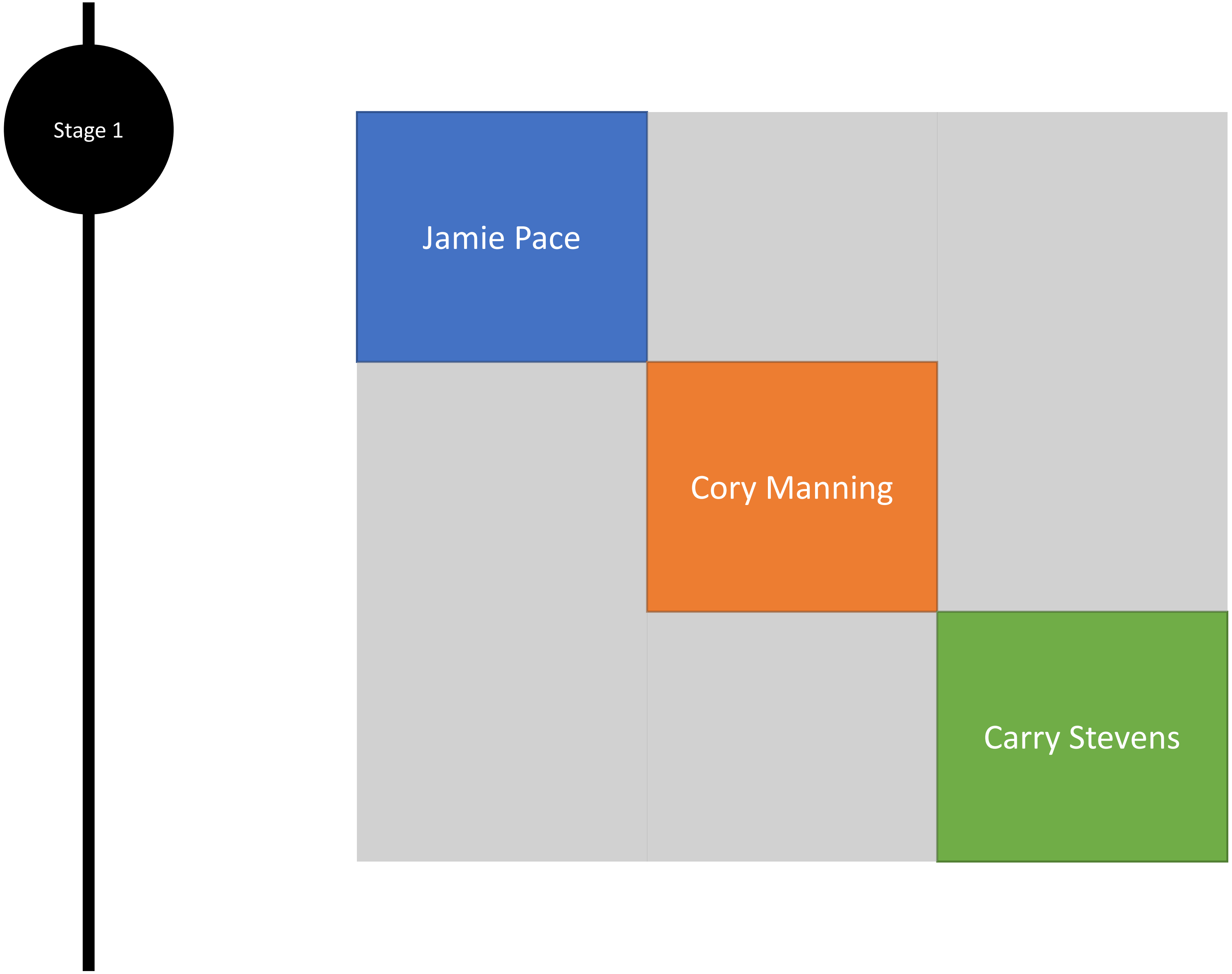


Stage 1

Jamie Pace

Cory Manning

Carry Stevens



Stage 1

Jamie Pace

Cory Manning

Carry Stevens

Stage 1

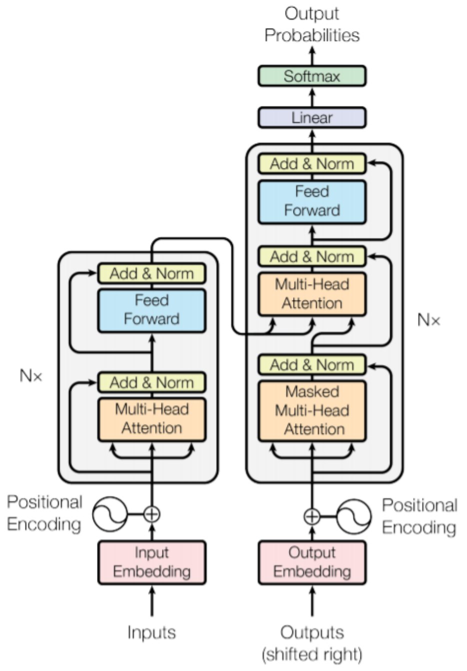
Stage 1 – Data Cleaning: Dirty Data

- Correct the whitespace issues (Good or bad?)
- Extract and create six new columns from the final exercise (meta-reflective parts of the work sample)
- Regex based data extraction
- Missing ratings imputed with K-nn

Stage 2

Stage II: Model Architecture Exploration

- **Deep learning with pre-trained (Transfer learning)**
  - SBERT – T5 – Universal Encoder
  - Initial score: .38 - .40
- Various Deep Learning Model Architectures
  - Self-variables selection network
  - Gated linear units
  - CONV1D Networks
  - LSTM
  - Transformers
- Drawbacks
  - Data Size
  - Very Complex for this small data
- Solutions
  - Add regularizations (Drop out layers, Alpha Drop-outs, or L1-L2)
  - Schedule learning rates
  - Generate Syntactic data with back-translation, added nose, swap noise, T5-based paraphrase model



Stage 2

Stage II: Model Architecture Exploration

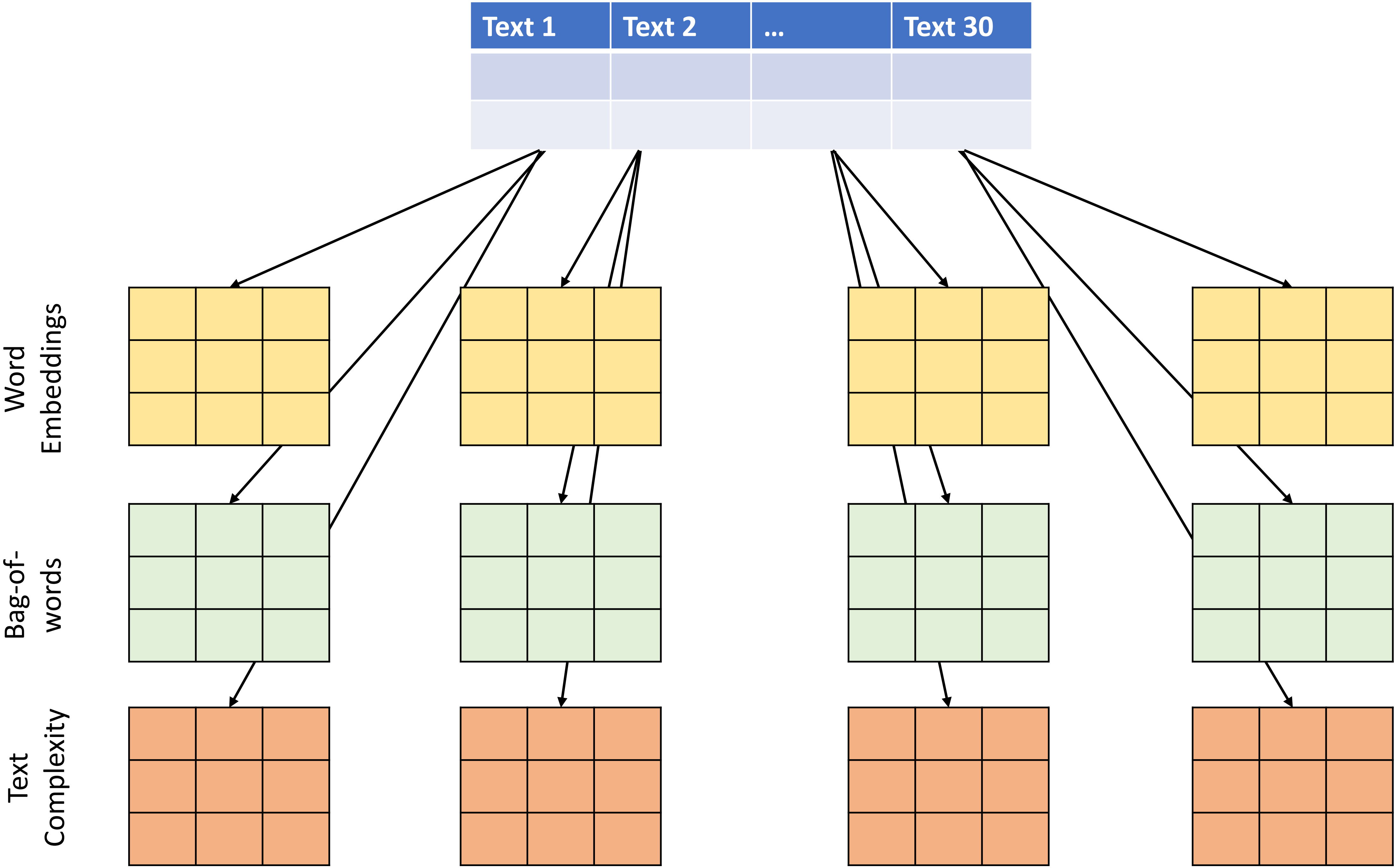
- **Bidirectional Encoder Representations from Transformers** (**BERT**)
  - Train a BERT model with a large corpus including the corpus we have
  - HuffPost + CNBC Data + the competition data
- Masked-language model with TPU
- Use the fined-tuned model's embeddings for the down streamed task
- **Results : ~.41**

- Drawbacks
  - Computational Heavy
  - Complicated and Complex
- Solution
  - Increase the dataset size

Stage III: Feature Engineering

Stage 3

- INSTRUCTION embeddings for semantic meaning
- Bag-of-words with N-grams for each columns for key words
- Text complexity and text statistics such as perplexity, rare word counts, SMOG scores or other text readability scores for each columns



Text 1

Text 2

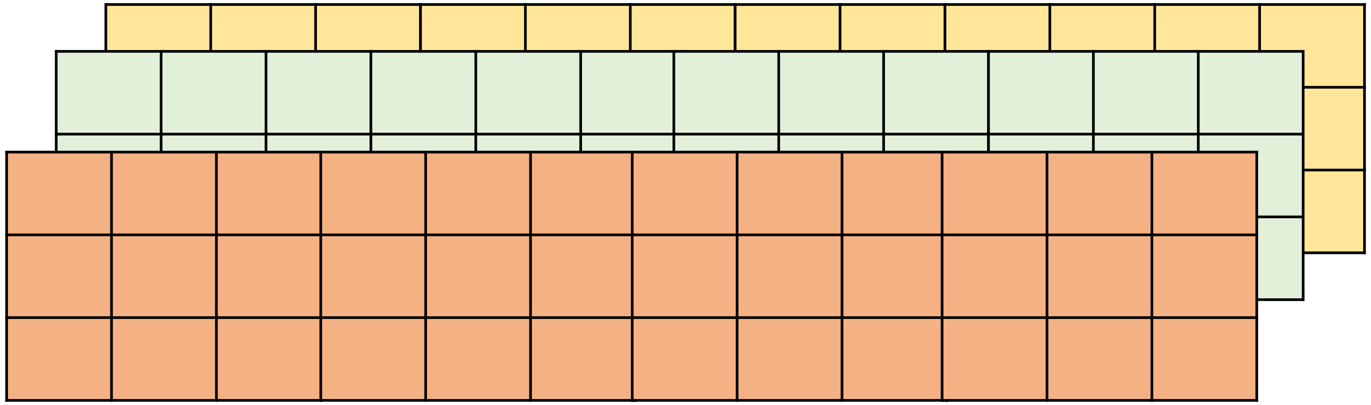
…

Text 30

Word Embeddings

Bag-of-words

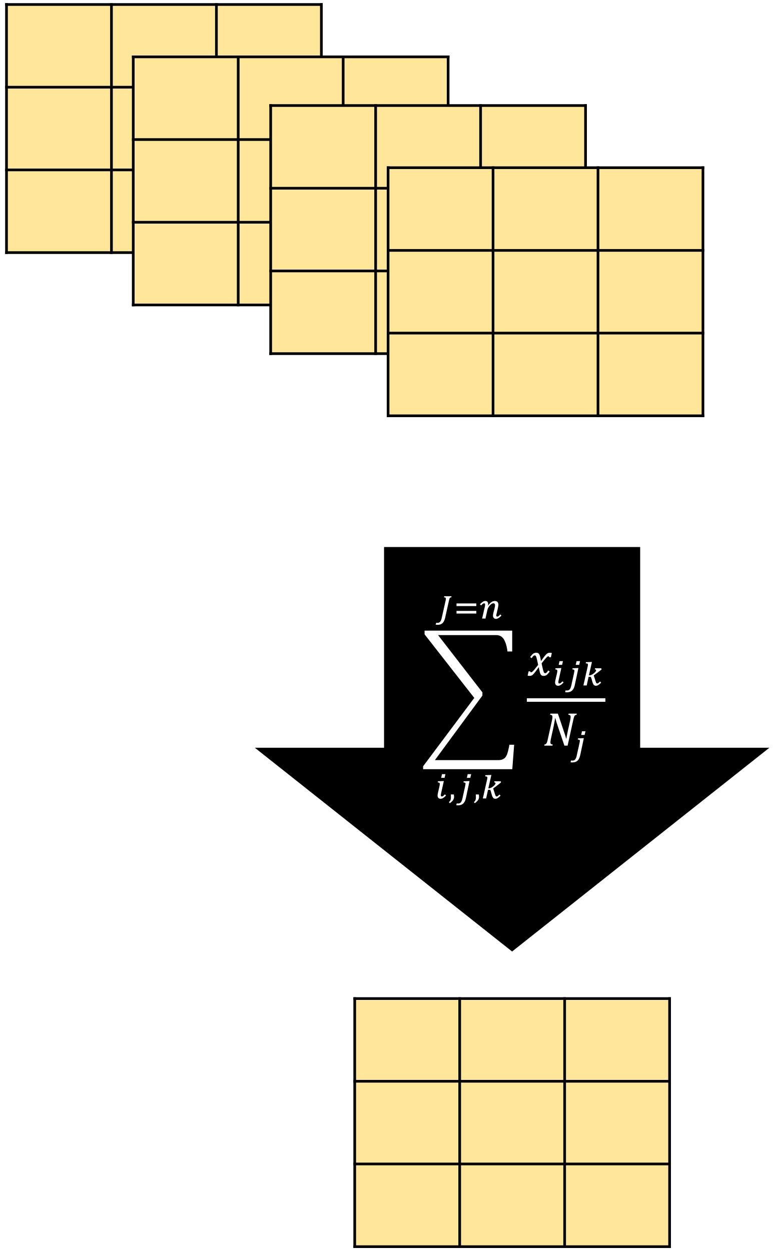
Text Complexity



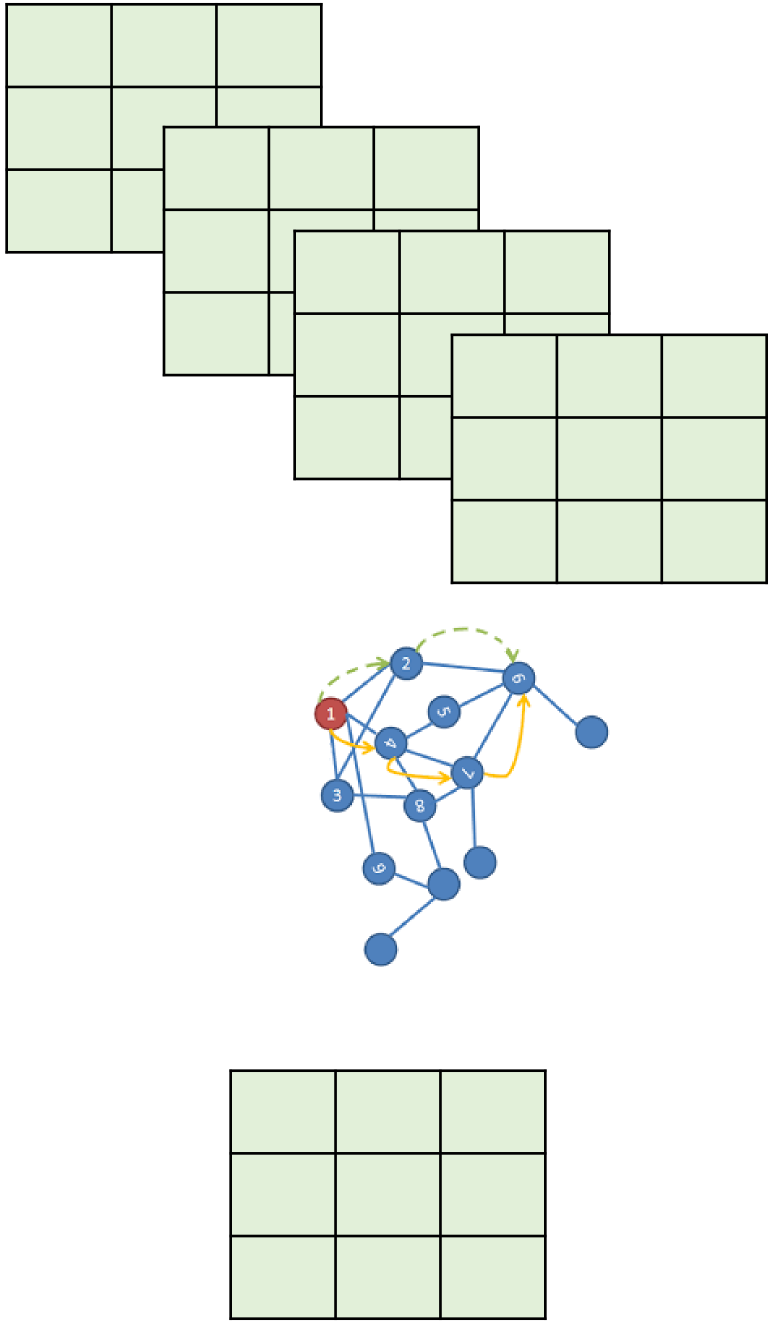
Dimension : (N Observations, 30, 25 000)

Curse of Dimensionality : N < p

Semantic Embeddings – Average of Embeds over Dimensions

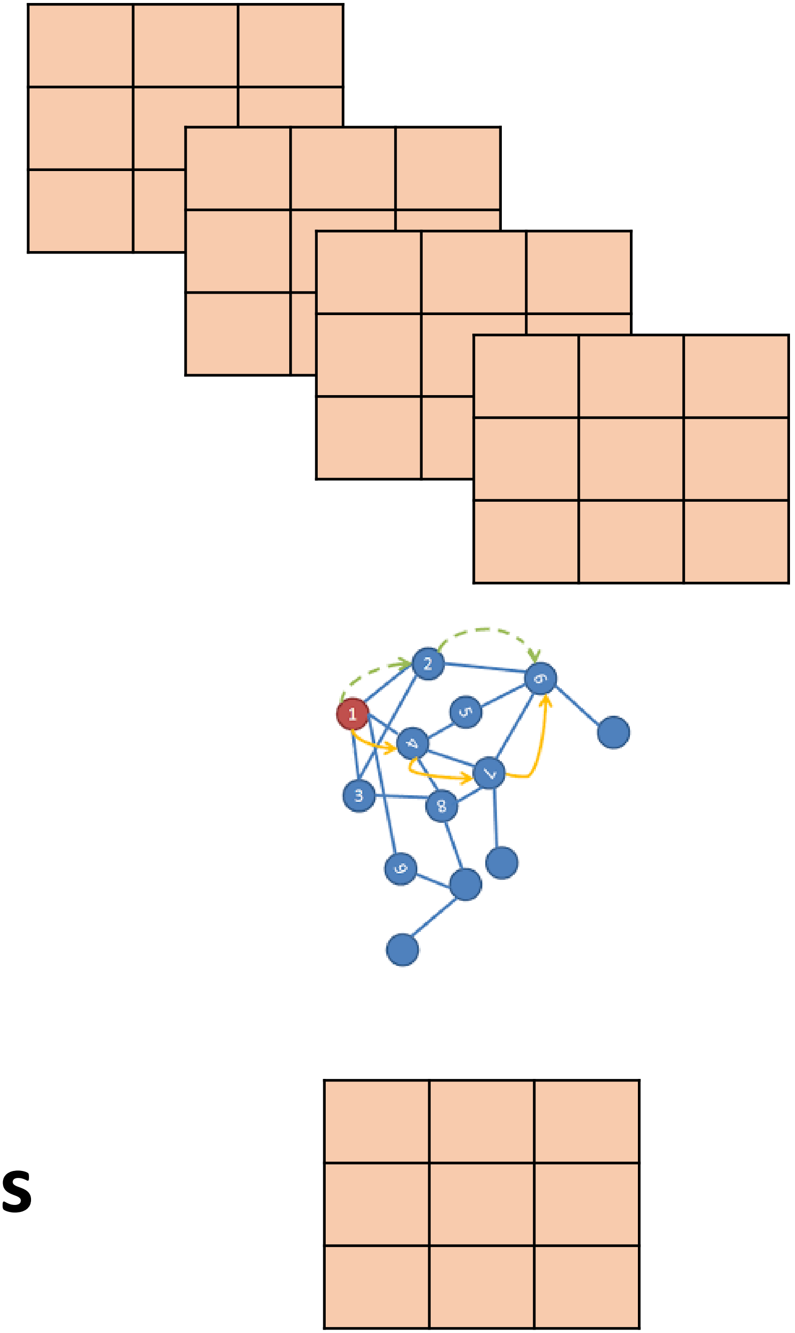


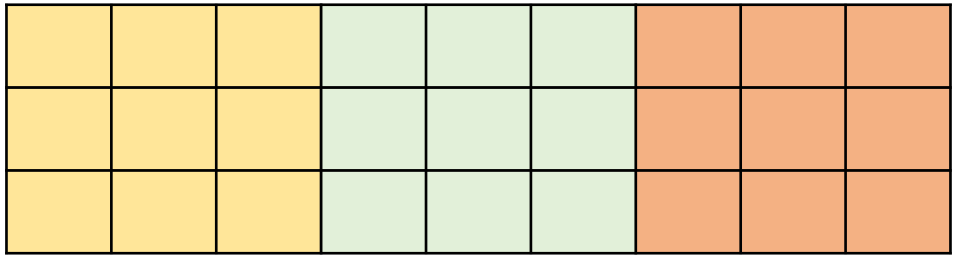
$$\sum_{i,j,k}^{J=n} \frac{x_{ijk}}{N_j}$$



Bag-of-words and Text Complexity

Diffusion Map Embeddings





Dimension : (N Observations, 2000~)

Stage 4

Stage IV: Model Building

- Multiple Ensembled model for each ratings

- K-fold cross validation (k = 15)
  - Validation variation is very large

- Select K-folds based on energy distance

$$D^2(F,G) = 2\,\mathrm{E}\,\|X - Y\| - \mathrm{E}\,\|X - X'\| - \mathrm{E}\,\|Y - Y'\| \geq 0,$$

Stage IV: Model Building

- Trained with various model with H2O in Python
  - XGBoost, GBM, GLM, Random Forest, Deep Learning
    - 100 Different Model – 4 to 5 Hr Training
  - Ensembled scores calculated with penalized linear models such as Elastic Net
  - Combined the top 3 ensembled model as a **simple average voting schema based on the cross-validated r^2value**
  - Development Phase Score : **.51**

Stage 4

Stage V: Submission

- Test data processed in the same way

- Due to the birth of my first child, I can only submit 4 out of 10 test submissions

- Issues
  - Time management

- Solution
  - Find good collaborators to work with

Stage 5

Lesson Learned

- Small Data = Feature Enginnering
- Semantic, Structure, and Complexity of Text
- Dimensionality Reduction with Diffusion Map – Preserves the local manifold geometry
- Time Management and Collaboration

Thank you for your attention!

Mustafa Akben, Ph.D.

makben@elon.edu