Comparing Coding of Interviewer Question-Asking Behaviors Using Recurrent Neural Networks to Human Coders

Jerry Timbrook
University of Nebraska-Lincoln

Adam Eck
Oberlin College

BigSurv18, October 2018

Acknowledgements

Acknowledgements:

Kristen Olson and Jolene Smyth

Antje Kirchner
Beth Cochran
Amanda Ganshert
Alexis Swendener
Angelica Phillips
Team of Behavior Coders and Transcriptionists

This work was supported by the National Science Foundation [SES-1132015]. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Behavior Coding

- **Objective, reliable method for coding interaction between interviewers and respondents** (Fowler & Cannell 1996)
- **Description of *actions* during each conversational turn in an interview**

	Behavior Codes		
Conversational turns	Actor	Initial action	Assessment
I: What is your sex?	Interviewer	Question asking	Read exactly as worded
R: Uh, could you repeat the question please?	Respondent	Clarification	Ask for repeat of question
I: What is your sex?	Interviewer	Probing	Repeat question exactly as worded
R: Oh, female.	Respondent	Answer provided	Adequate

Why Do We Need Behavior Coding?

- **Uses:**
  - Pre-test survey questions (Fowler & Cannell 1996)
  - Gain insight into the interviewers' and respondents' cognitive processing (Fowler & Cannell 1996; Schaeffer & Maynard 1996)
  - Evaluate interviewers' performance during the field period (Fowler & Mangione 1990)
  - Evaluate interviewers' effect on measurement (e.g., Dykema et al. 1997; Garbarski et al. 2016; Sarwar et al. 2017; Olson et al. 2018)
- **In short: Behavior Coding helps us understand how respondents and interviewers affect data quality.**

The Present Study

- **The problem:**
  - Behavior coding is typically a manual process
    - Costly and time-consuming
- **The question:**
  - Can we partially automate this coding process using machine learning?
- **The scope:**
  - Focusing on:
    - Interviewer's deviations from exact question readings (question-asking behavior)

Conversational turns	Actor	Initial action	Assessment
I: What is your sex?	Interviewer	Question asking	Read exactly as worded

Why Focus on Question-Asking Behaviors?

- **Standardized Question-Asking** (Fowler & Mangione 1990)
  - Each question is read exactly as worded
  - Why?
    - Ensures all respondents receive the same question wording
      - PIs and data users assume that questions are read exactly as worded in the questionnaire
    - Allows replicability
    - Reduces interviewer variance
      - ↑ interviewer variance = ↓ precision of estimates (wider CI's)
  - Deviations = indicator of problems
    - Lack of interviewer training and monitoring (Fowler & Mangione 1990)
    - Lack of interviewer motivation (Japec 2008)
    - Poor question construction (Oksenberg, Cannell, & Kalton 1991)

Question-Asking Behaviors

- **What are the common classifications for question-asking behaviors?**

We would like to end with a few questions about you and your household. I have to read every question in this survey, even if it seems obvious. What is your sex?

1 MALE
2 FEMALE
3 OTHER, SPECIFY [TEXT BOX]
9 REFUSED

Question-Asking Behaviors

- **Exact Reading**
  - Interviewer reads the question exactly as worded in questionnaire

> We would like to end with a few questions about you and your household. I have to read every question in this survey, even if it seems obvious. What is your sex?

- **Minor Change**
  - Read with one or more words omitted or added
  - Meaning of the question does not change

> *Ok so,* we would like to end with a few questions about you and your household. I *do* have to read every question in *the* survey, even if it seems obvious. What is your sex?

Question-Asking Behaviors

- **Exact Reading**
  - Interviewer reads the question exactly as worded in questionnaire

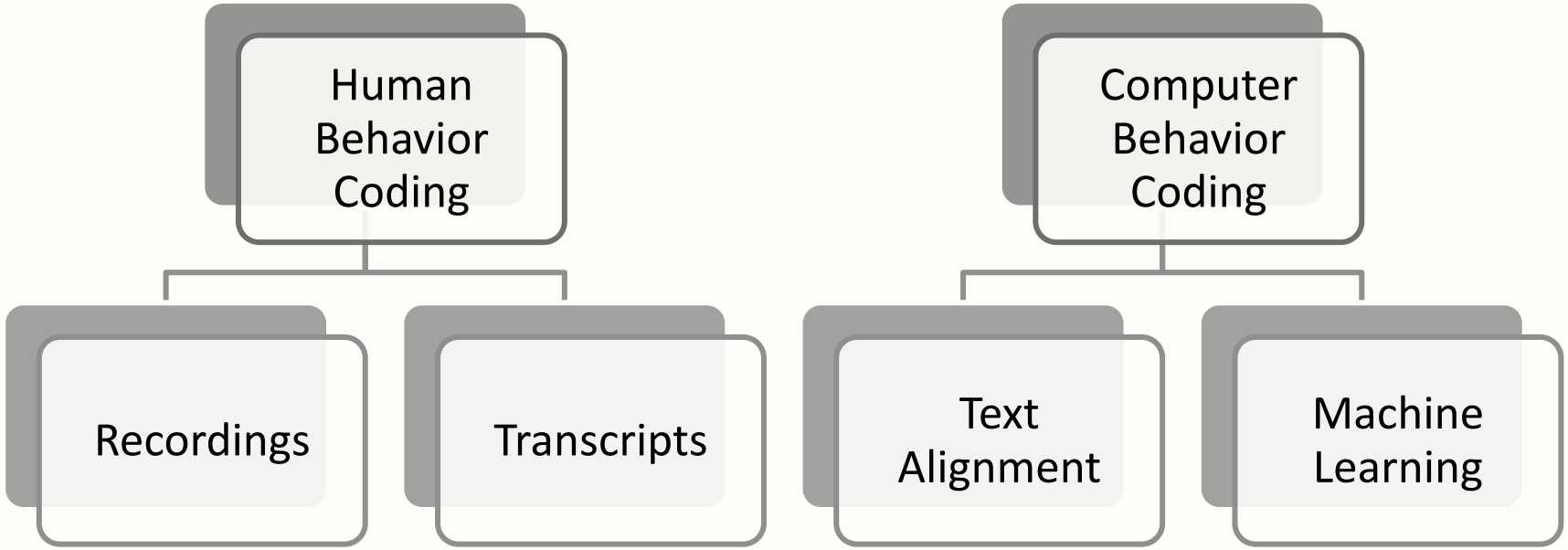
> We would like to end with a few questions about you and your household. I have to read every question in this survey, even if it seems obvious. What is your sex?

- **Major Change**
  - Omission or addition of words that alter question meaning

> We would like to end with a few questions about you and your household.
> I have to read every question in *the* survey, even if it seems obvious.
> What is your sex*, male or female*?

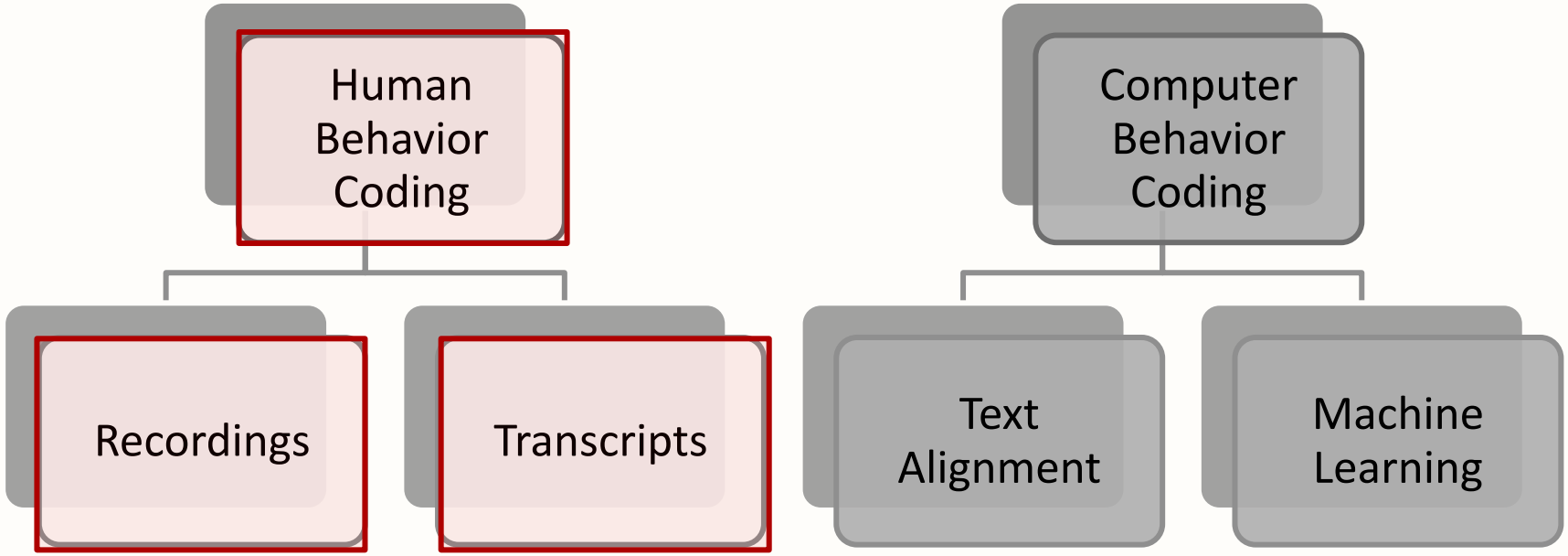
Question-Asking Behaviors

- **What methods are used to code question-asking behaviors?**



Question-Asking Behaviors

- **What methods are used to code question-asking behaviors?**



Human Behavior Coding

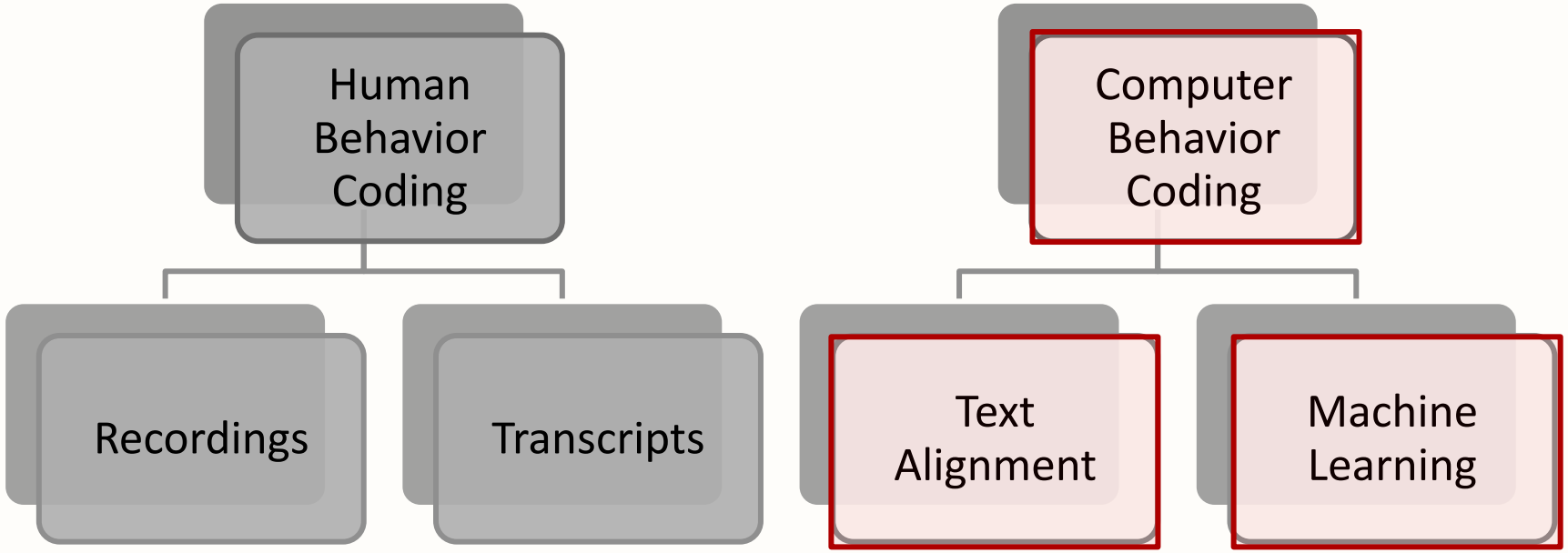
- **Interview Recordings & Transcripts**
  - Transcripts of each case are created using audio recordings of each survey interview
  - A human reads the transcript of each question administration
    - Can also listen if audio recordings are available
  - Assigns a behavior code based on what they read

Human Behavior Coding

- **Interview Recordings & Transcripts**
  - Pros
    - Easy visual review of the question (transcript)
    - Hear the question as it was administered (recording)
      - Timestamps of conversational turns in audio allows for calculating speaking time
  - Cons
    - Potential for human error in coding
    - Error in transcription can lead to error in coding
    - Time consuming and expensive
      - 902 cases
        - (1,680 hours)($11/hour) = $18,480 for transcription
        - (1,163 hours)($11/hour) = $12,793 for coding
        - ($18,480 transcription) + ($12,793 coding) = $31,273

Question-Asking Behaviors

- **How do we classify question-asking behaviors?**



Computer Behavior Coding

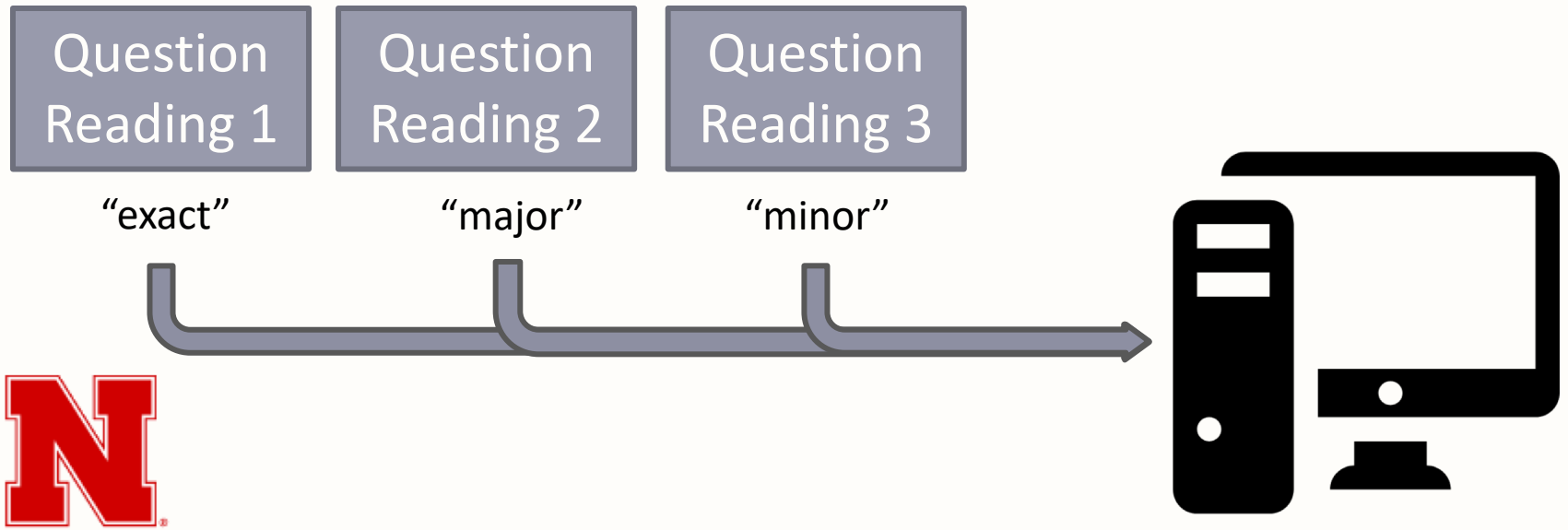
- **Text Alignment**
  - Transcripts required
  - Write computer code that compares the text of each question-reading with the questionnaire
  - Exact Match Method
    - Coded as *exact reading* if the two texts are an identical match
    - Coded as *not exact* if the two texts deviate at all
  - Percent Similarity Method (Distance)
    - Use algorithms to determine percentage of characters in the question-reading text that match the questionnaire text
      - e.g., Levenshtein Algorithm, Trigram Comparison, Jaro-Winkler, Ratcliff/Obershelp
    - Coded as *exact reading* if the percentage is >= some cutoff (e.g., 90%)
    - Coded as *not exact* if the percentage is < the cutoff

Computer Behavior Coding

- **Text Alignment**
  - Pros
    - No human behavior coding required
    - No extra “per case” cost once code is written for a question
    - % Similarity:
      - Allows for some breaks in fluent speech (e.g., uh, um)
      - Allows for some transcription errors (e.g., spelling)
  - Cons
    - Cannot differentiate between major and minor changes
    - % Similarity:
      - Cutoff percentage is arbitrary, may change by question

Methods for Identifying Deviations from Exact Reading

- **Machine Learning**
  - Computer learns to accomplish a task on its own using examples
  - Common Process:
    - Training: Humans provide examples (e.g., transcripts) and their classifications (e.g., behavior codes) as input, computer discovers patterns

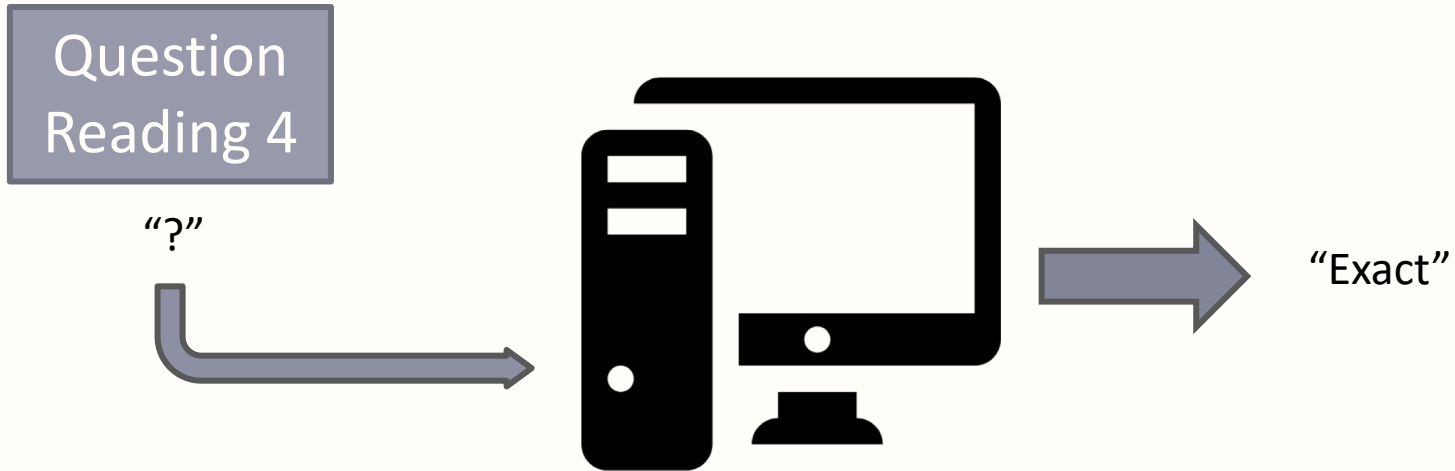


Methods for Identifying Deviations from Exact Reading

- **Machine Learning**
  - Computer learns to accomplish a task on its own using examples
  - Common Process:
    - Training: Humans provide examples (e.g., transcripts) and their classifications (e.g., behavior codes) as input, computer discovers patterns
    - Validation: Given additional examples (e.g., transcripts), check overfitting

Methods for Identifying Deviations from Exact Reading

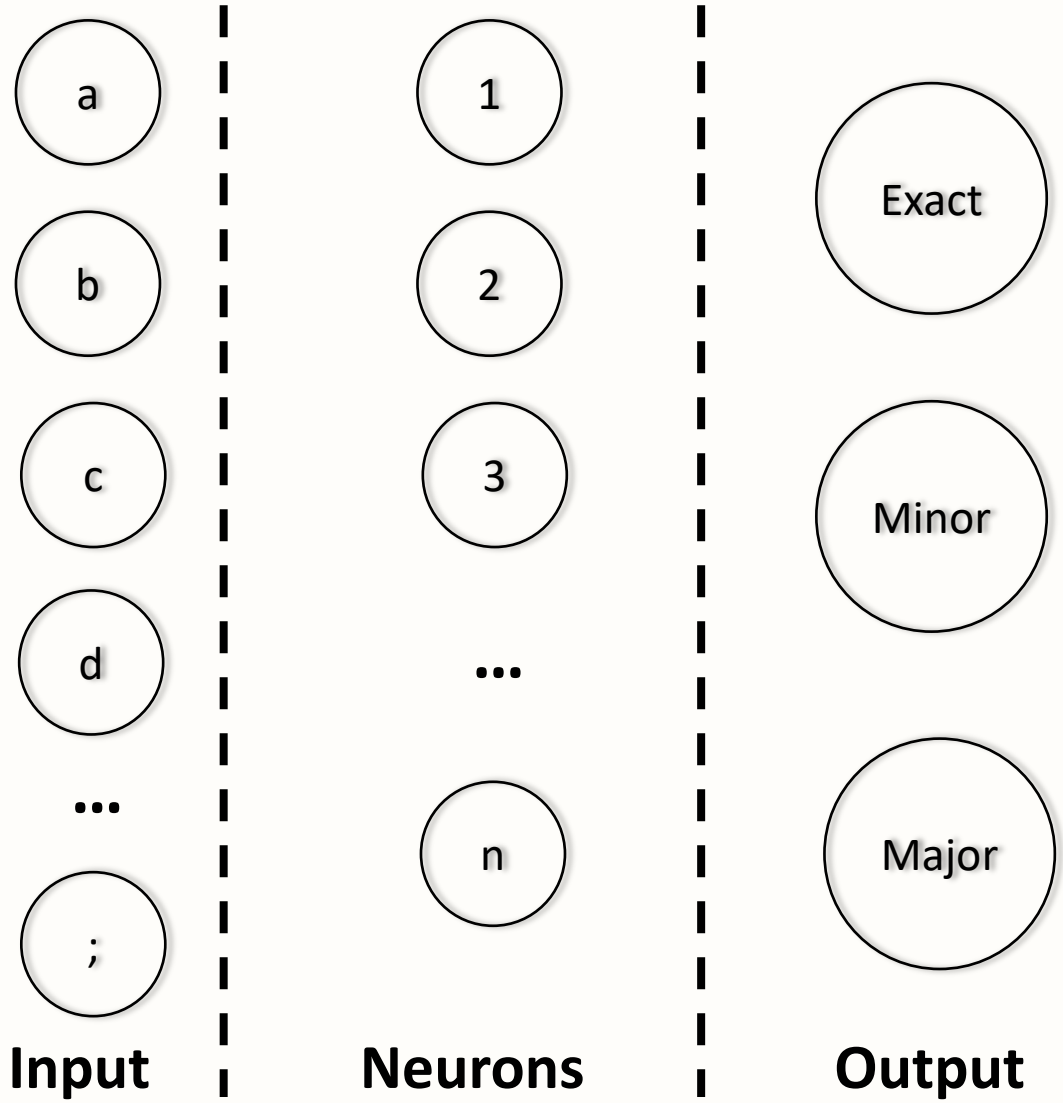
- **Machine Learning**
  - Computer learns to accomplish a task on its own using examples
  - Common Process:
    - Training: Humans provide examples (e.g., transcripts) and their classifications (e.g., behavior codes) as input, computer discovers patterns
    - Validation: Given additional examples (e.g., transcripts), check overfitting
    - Test: Computer classifies new instances (e.g., new sets of transcripts) based on previous examples, compare accuracy to human classification



Methods for Identifying Deviations from Exact Reading

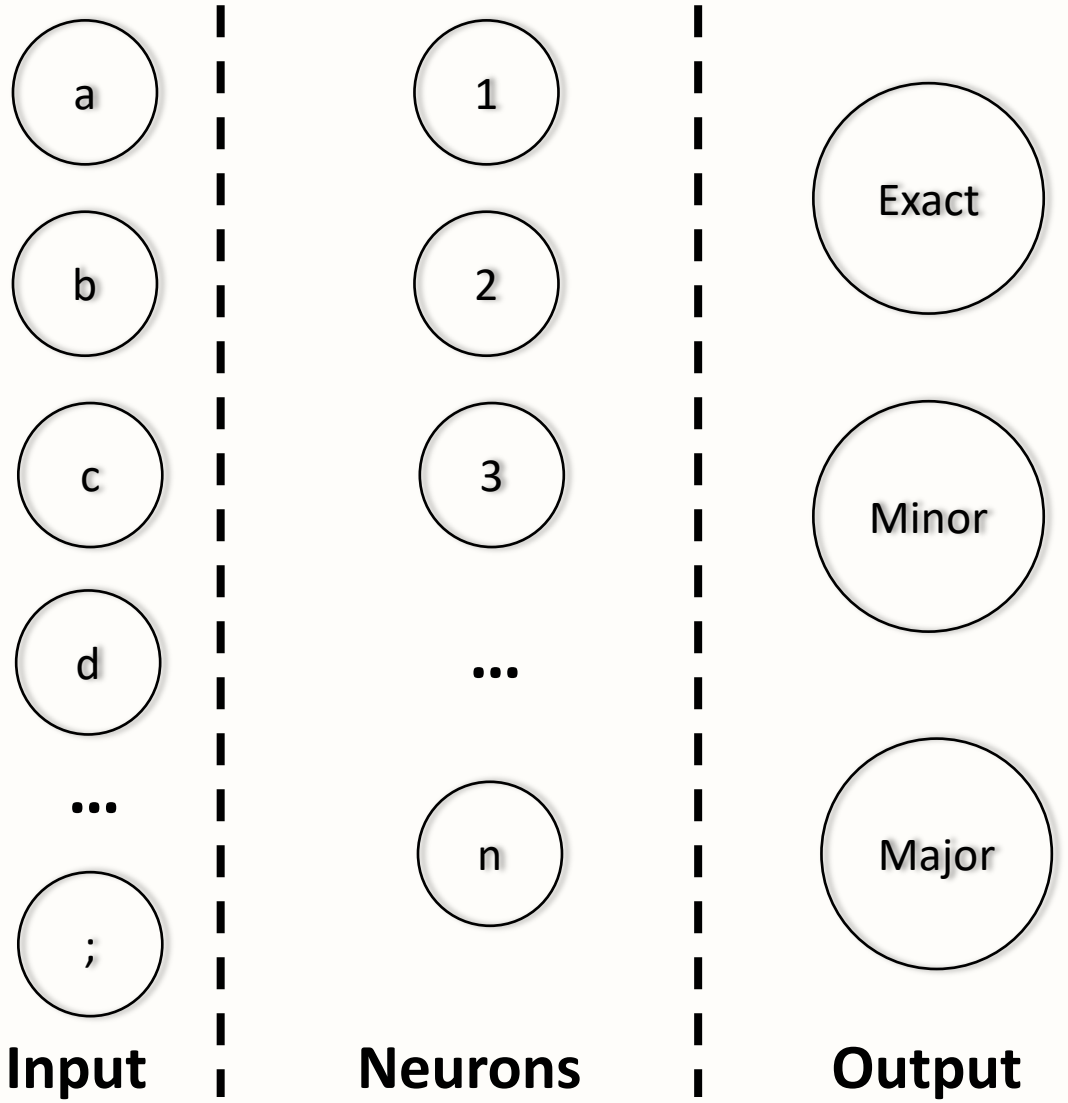
- **Machine Learning**
  - Recent Research Applications:
    - Coding how well therapists adhere to proper treatment methods using transcripts of clinical sessions (Xiao et al. 2016)
    - Coding news articles as relevant/irrelevant for content analysis (Nelson et al. 2018)
  - This project: Recurrent Neural Networks (RNNs)

Recurrent Neural Networks



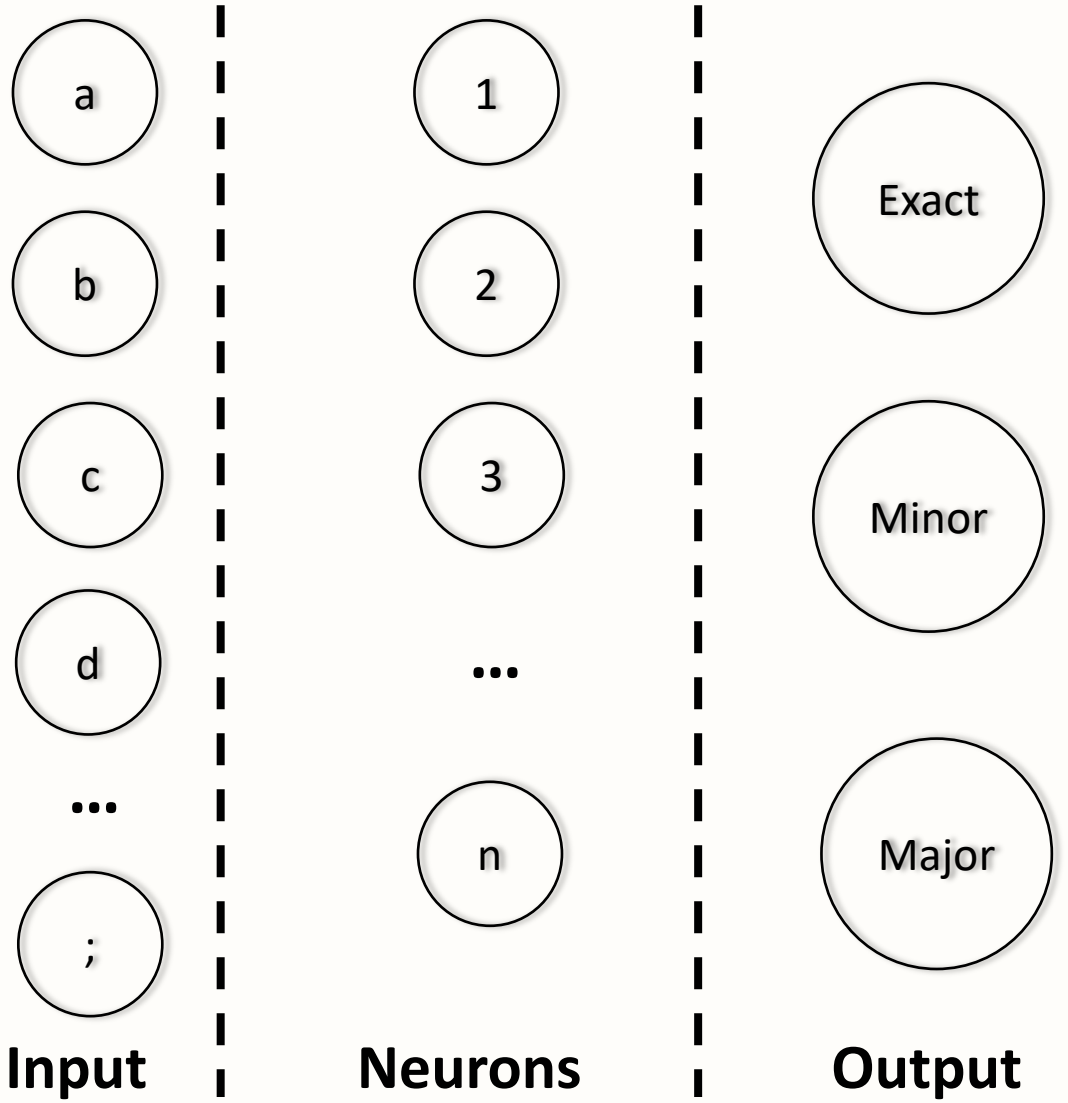
Recurrent Neural Networks

and so, what is your sex?



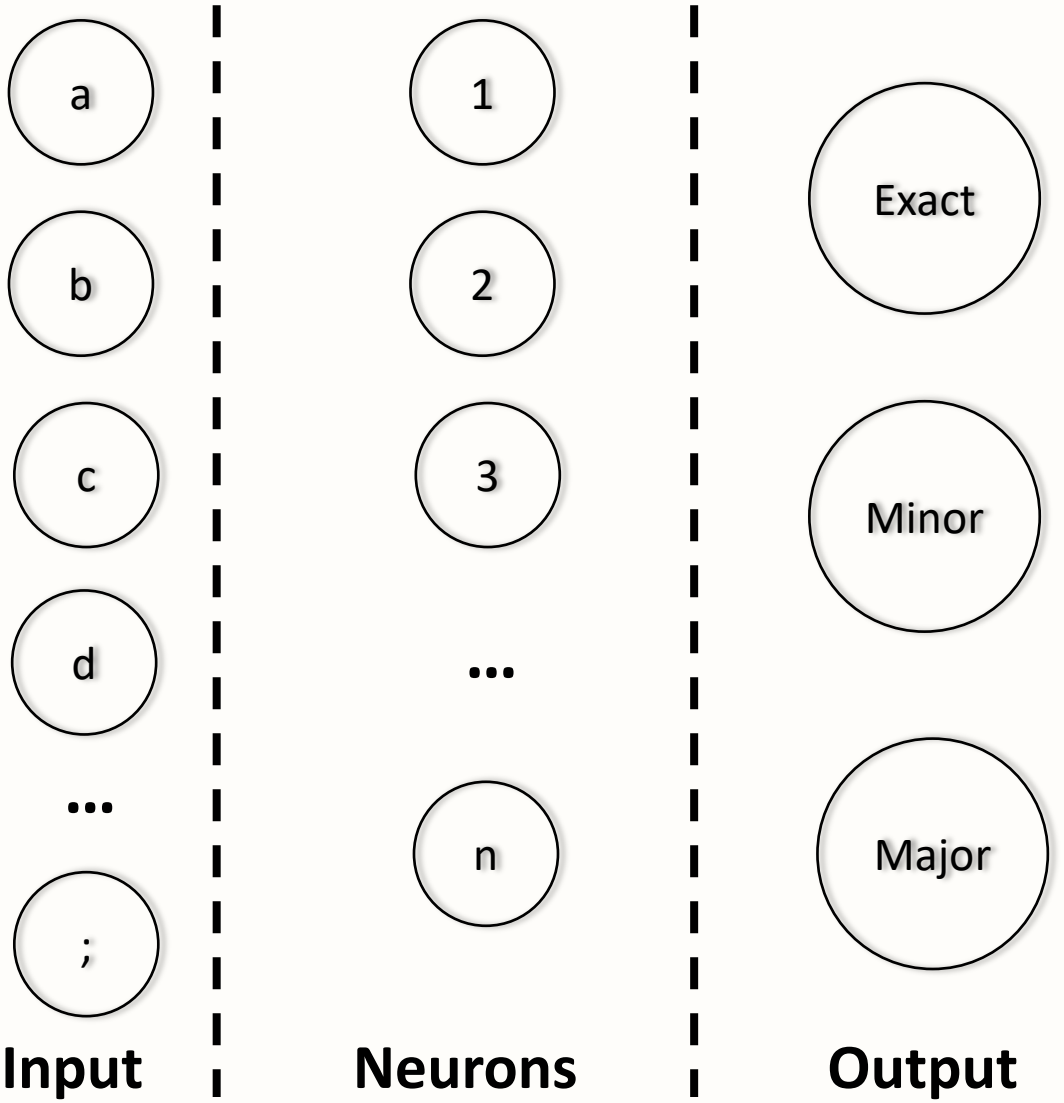
Recurrent Neural Networks

and so, what is your sex?



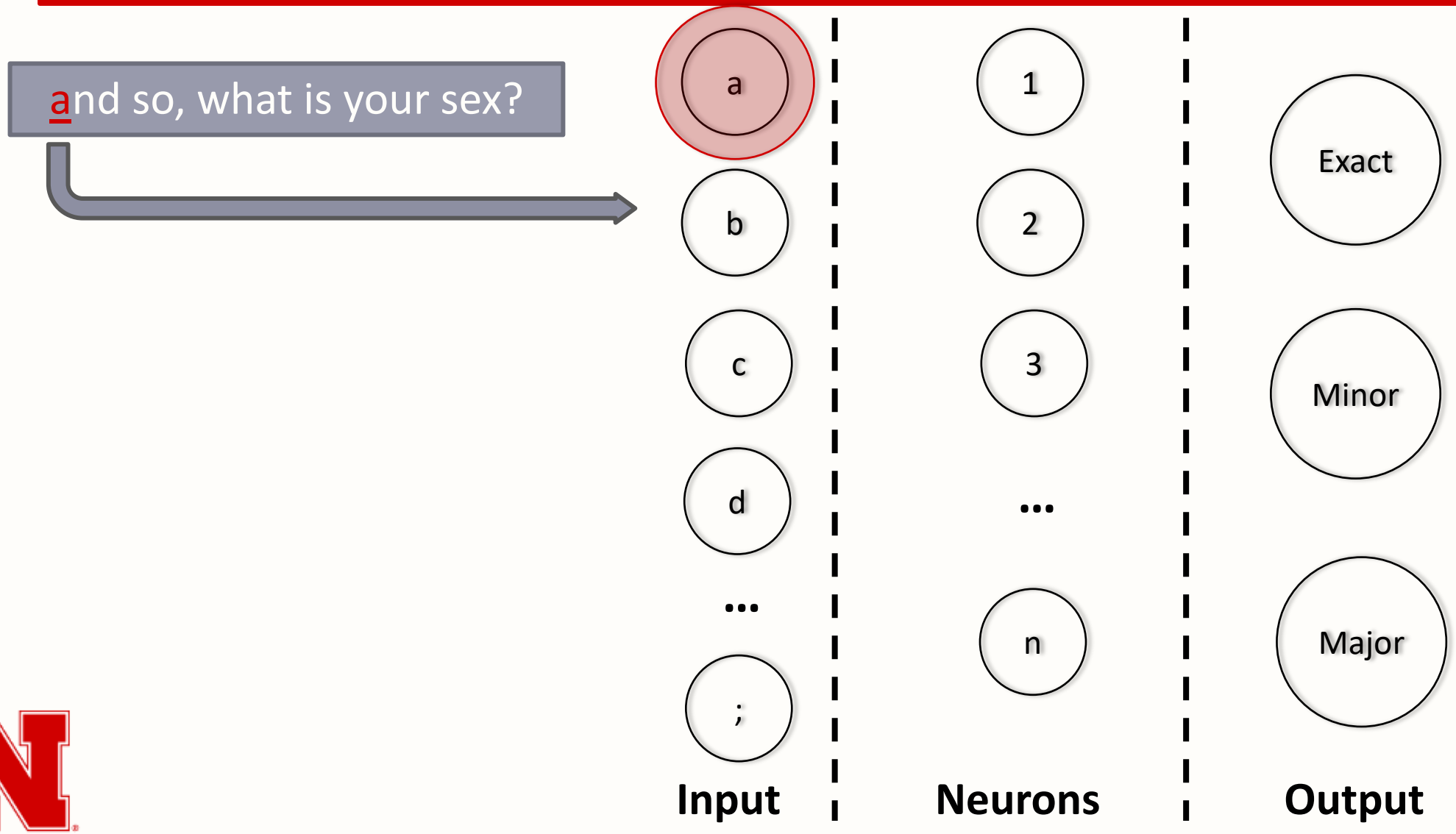
Recurrent Neural Networks

and so, what is your sex?



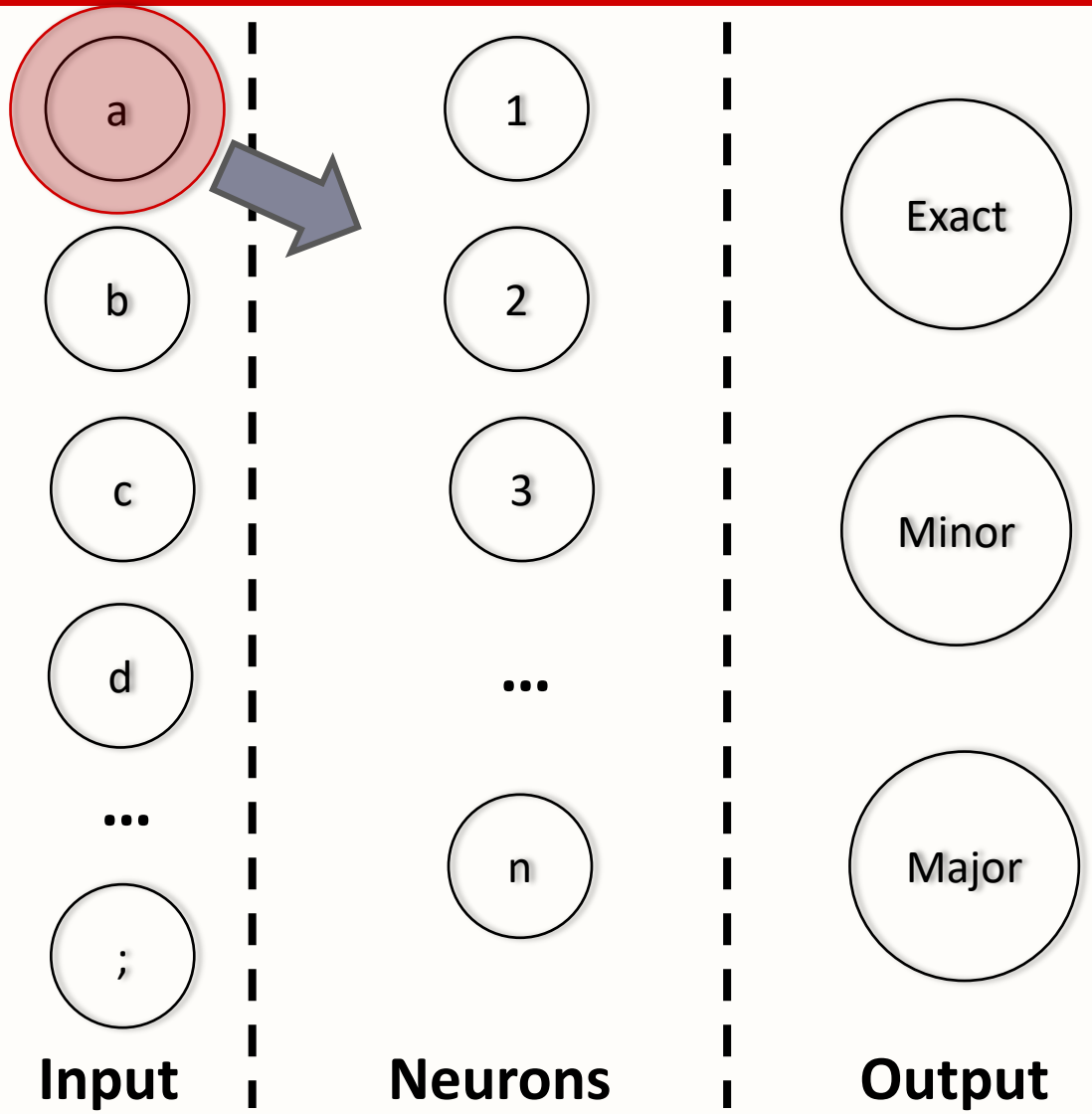
Recurrent Neural Networks

and so, what is your sex?



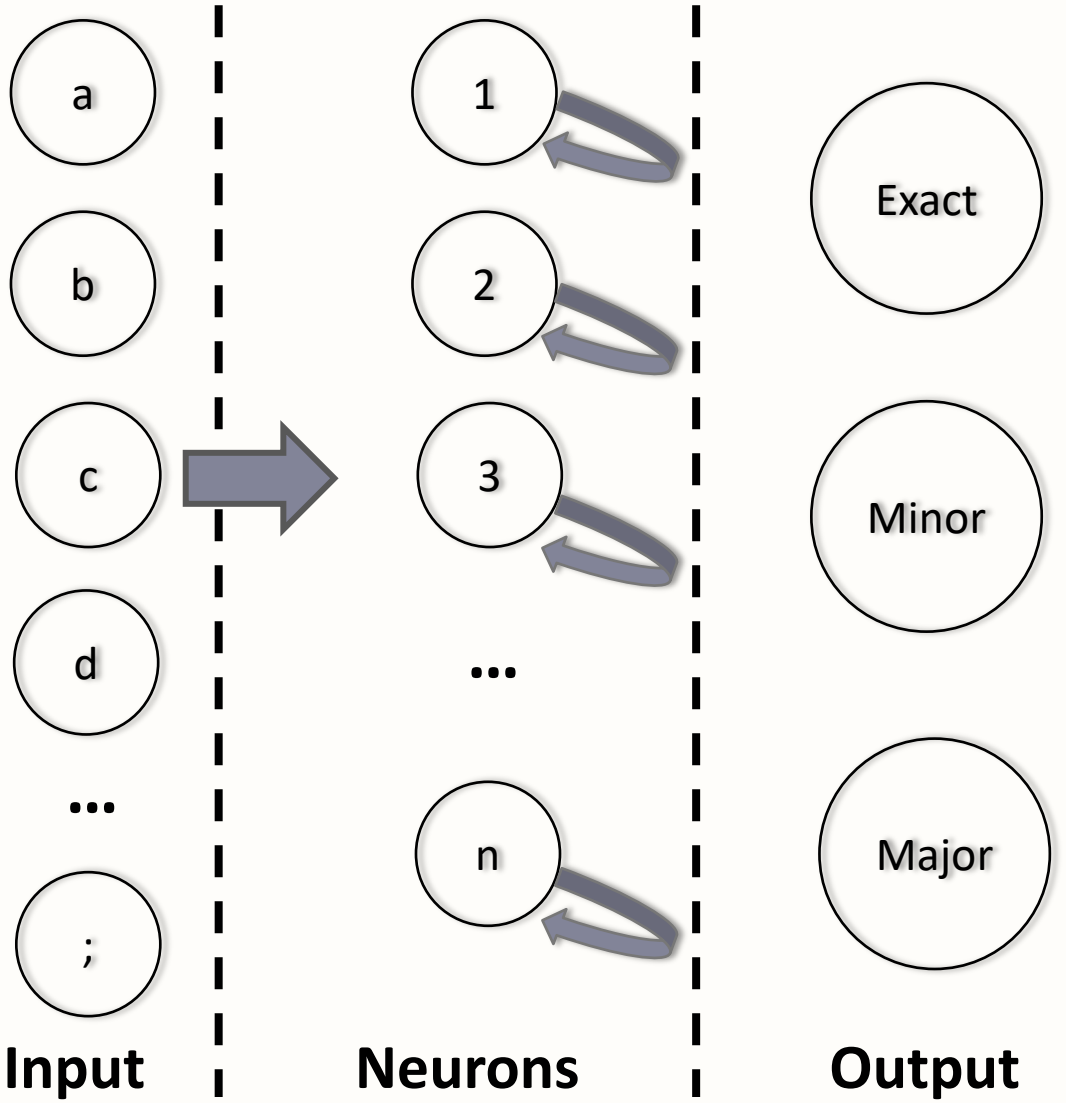
Recurrent Neural Networks

and so, what is your sex?



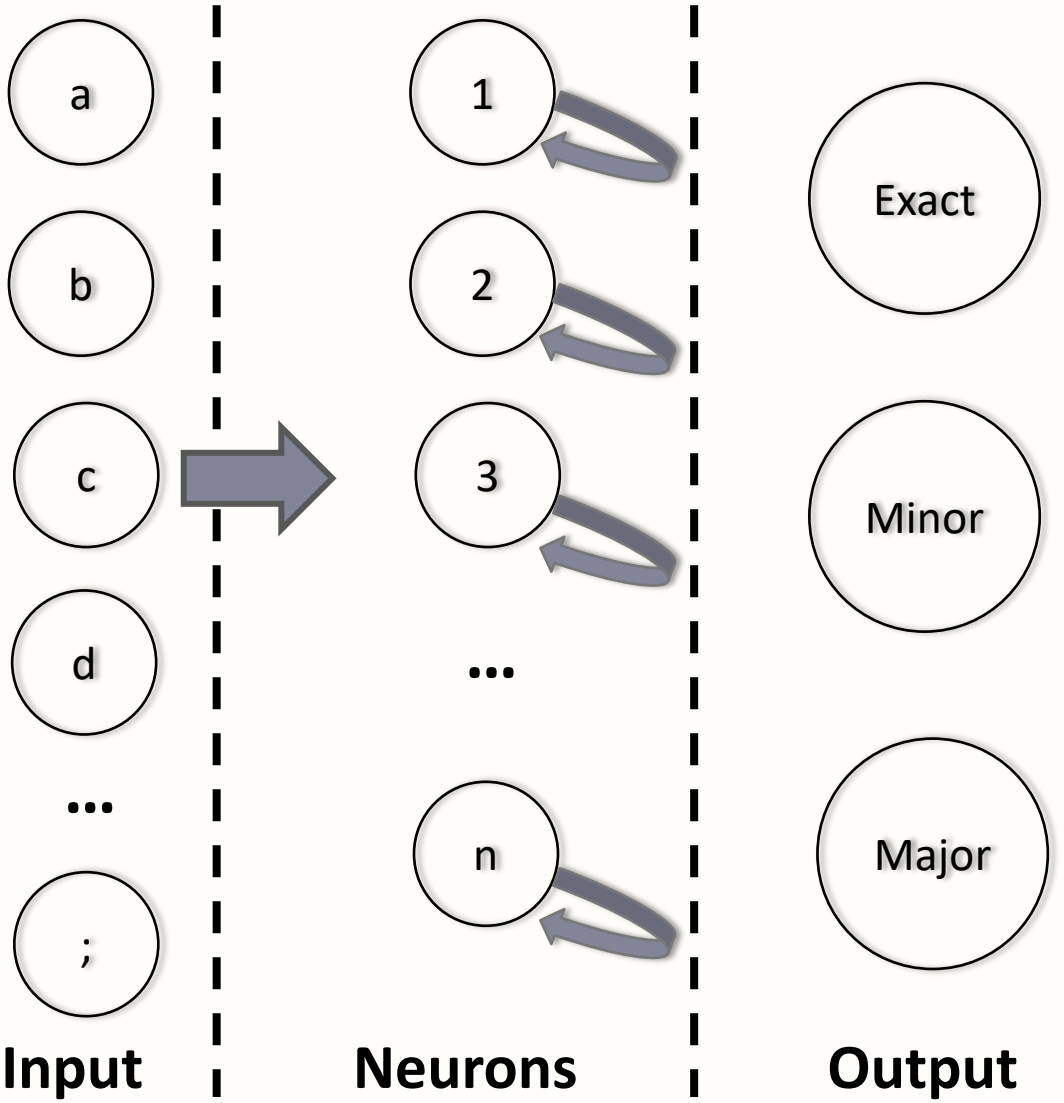
Recurrent Neural Networks

and so, what is your sex?

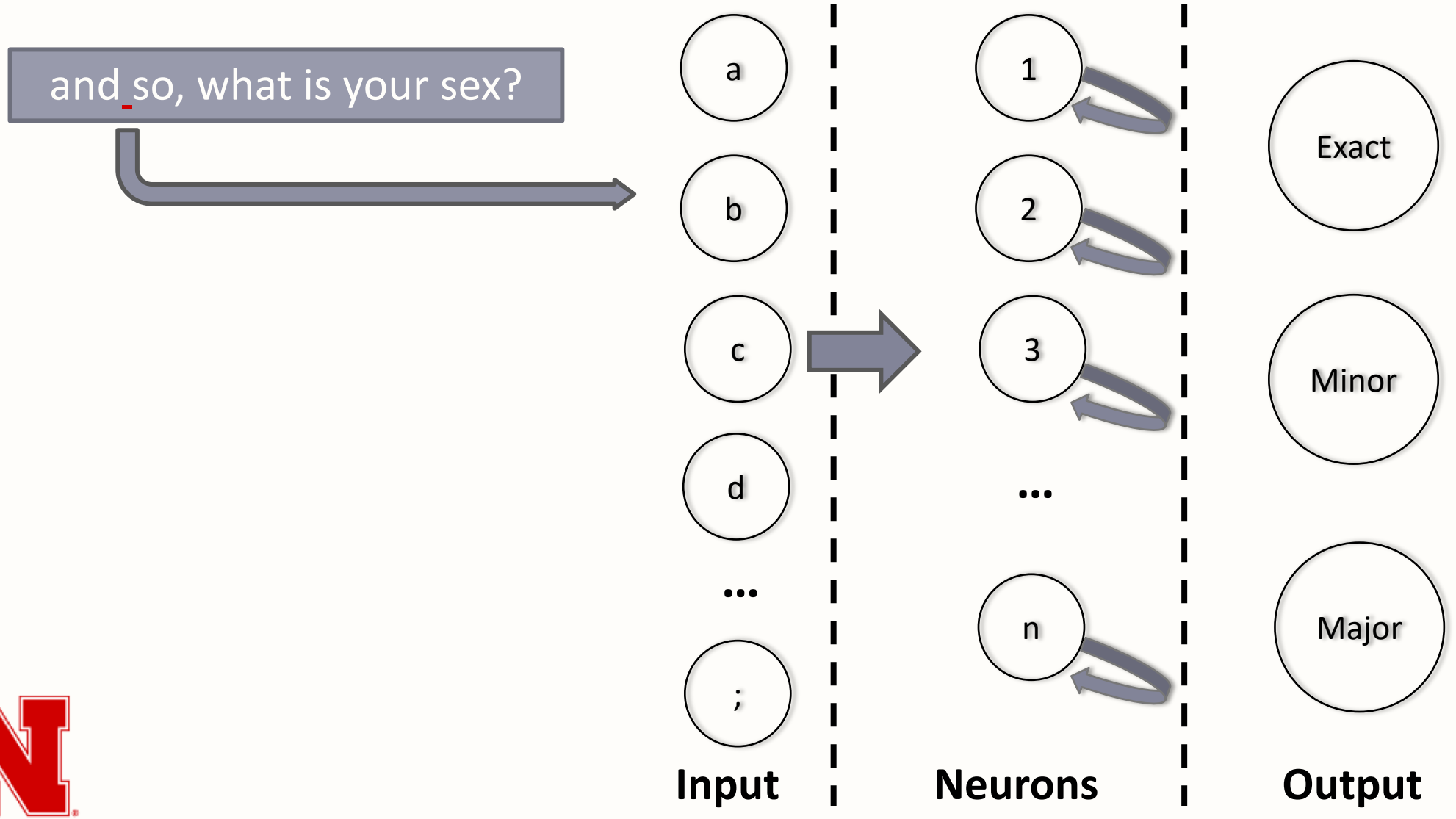


Recurrent Neural Networks

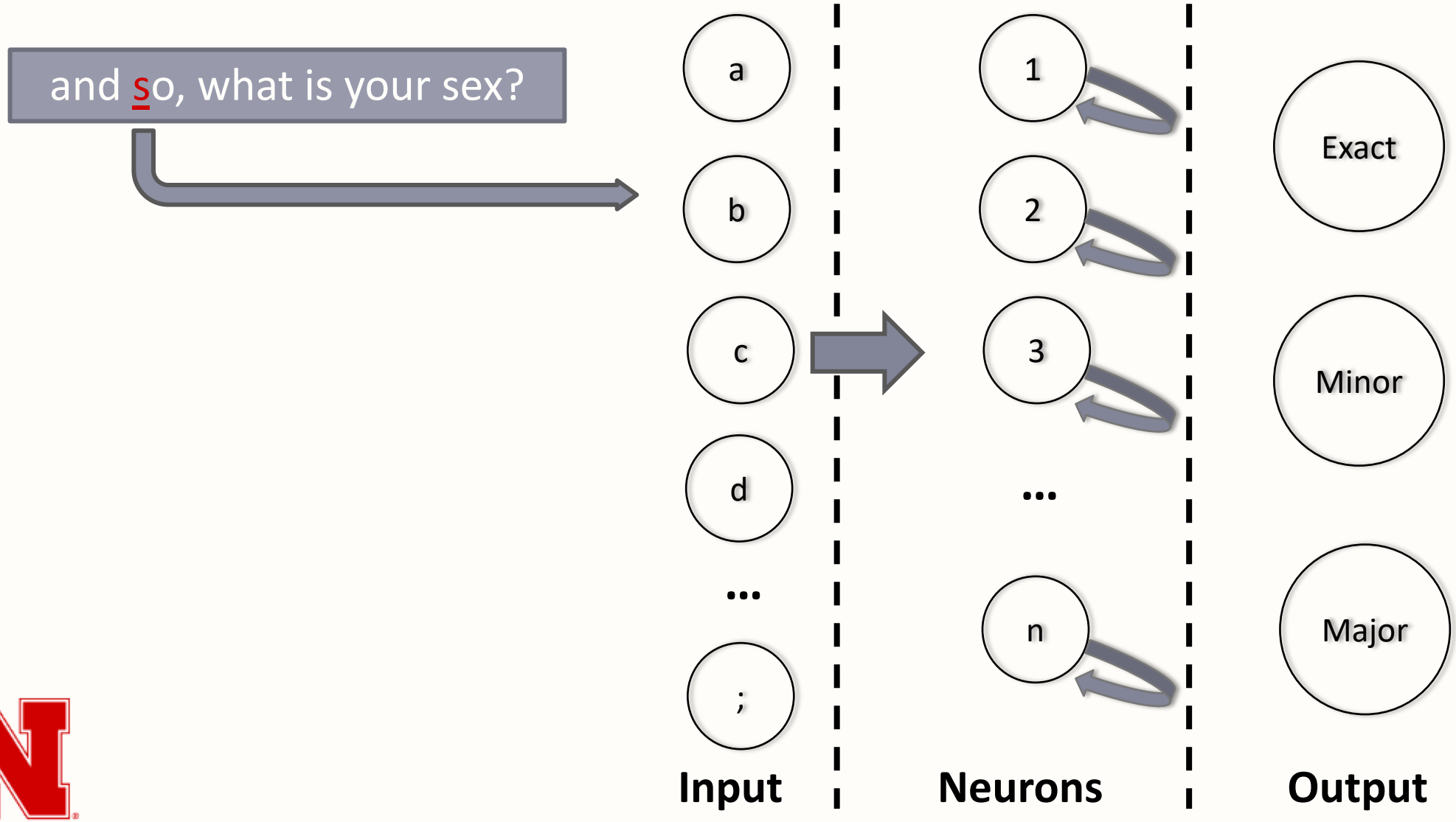
and so, what is your sex?



Recurrent Neural Networks

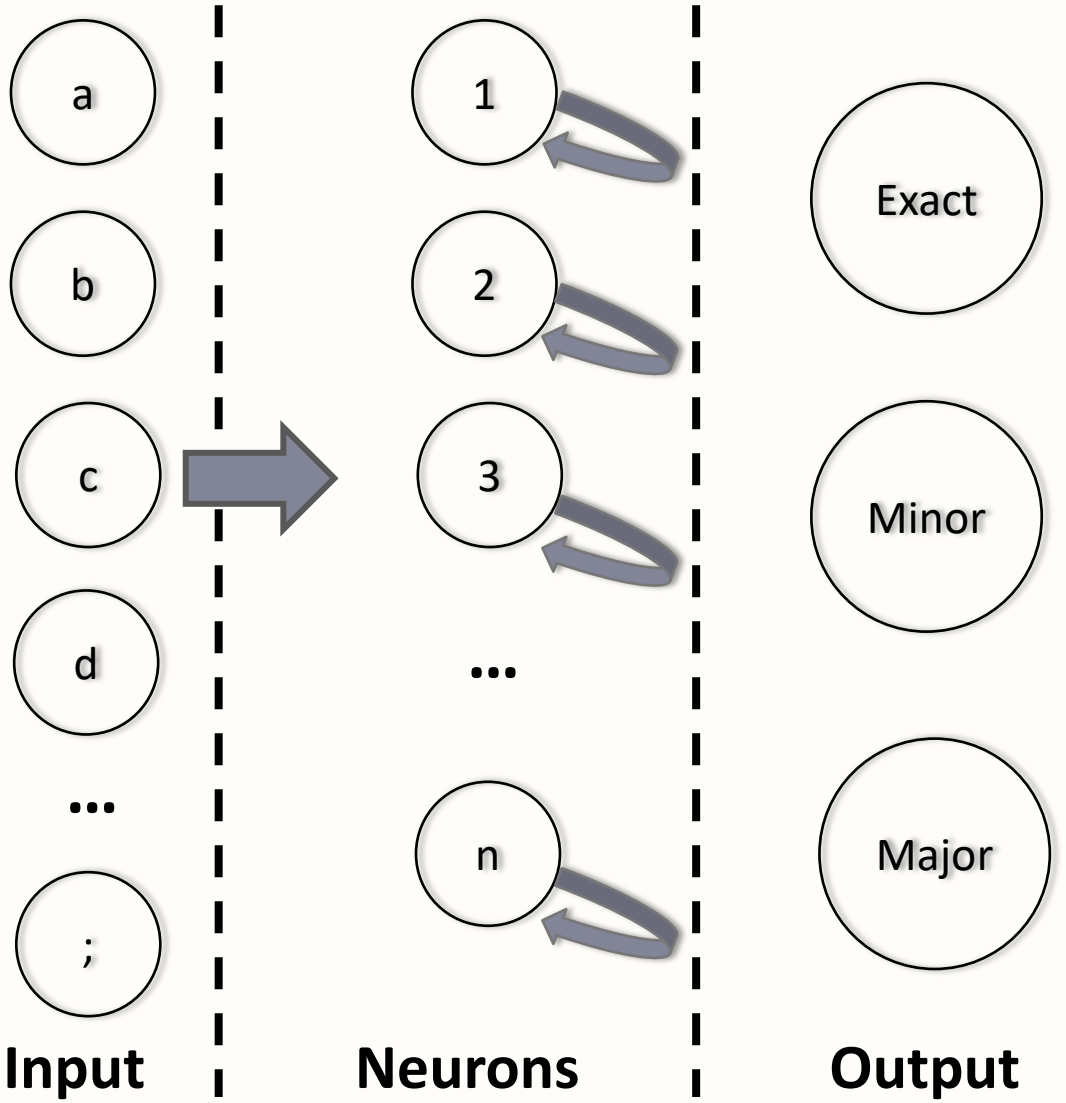


Recurrent Neural Networks



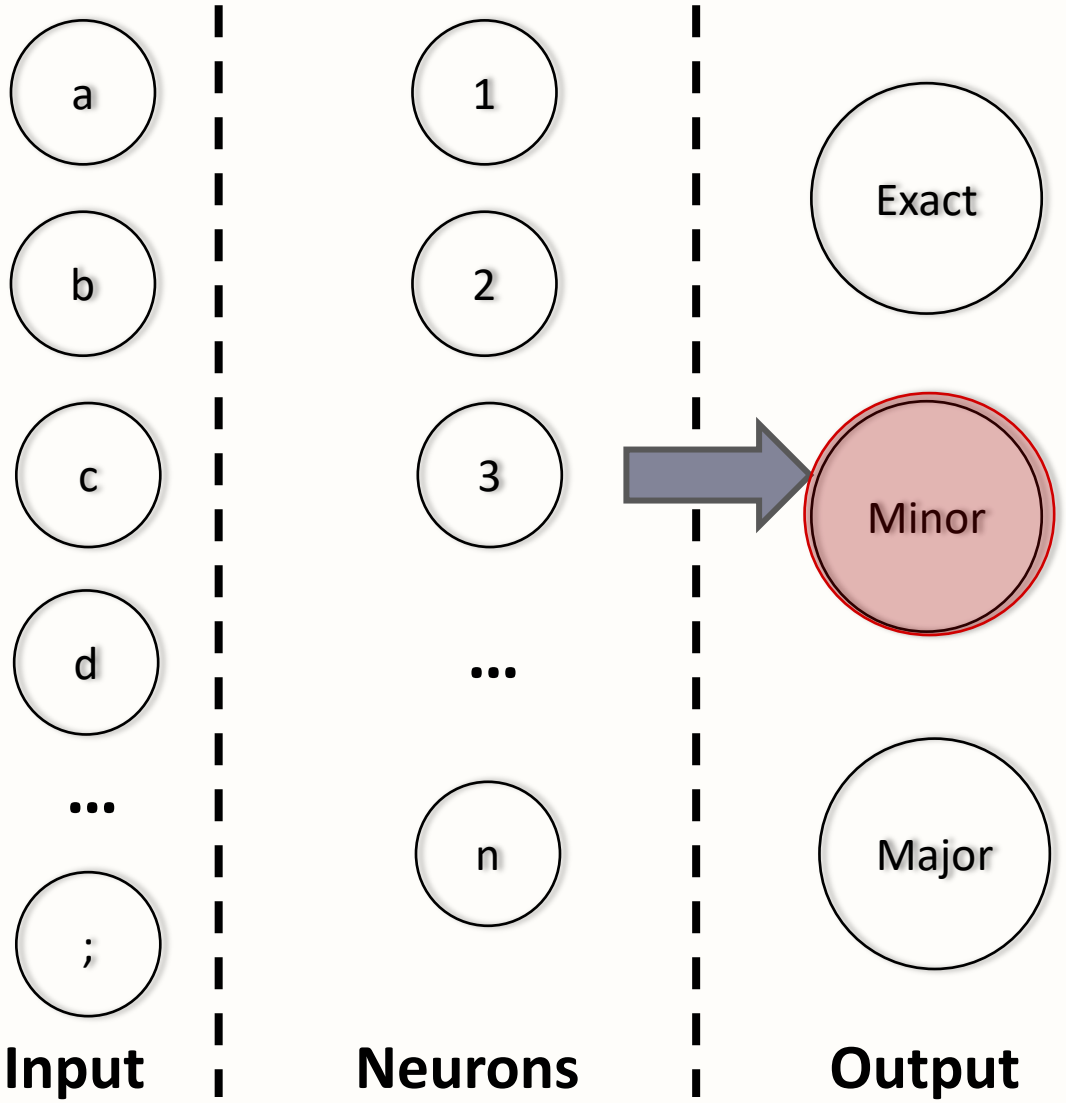
Recurrent Neural Networks

and so, what is your sex?



Recurrent Neural Networks

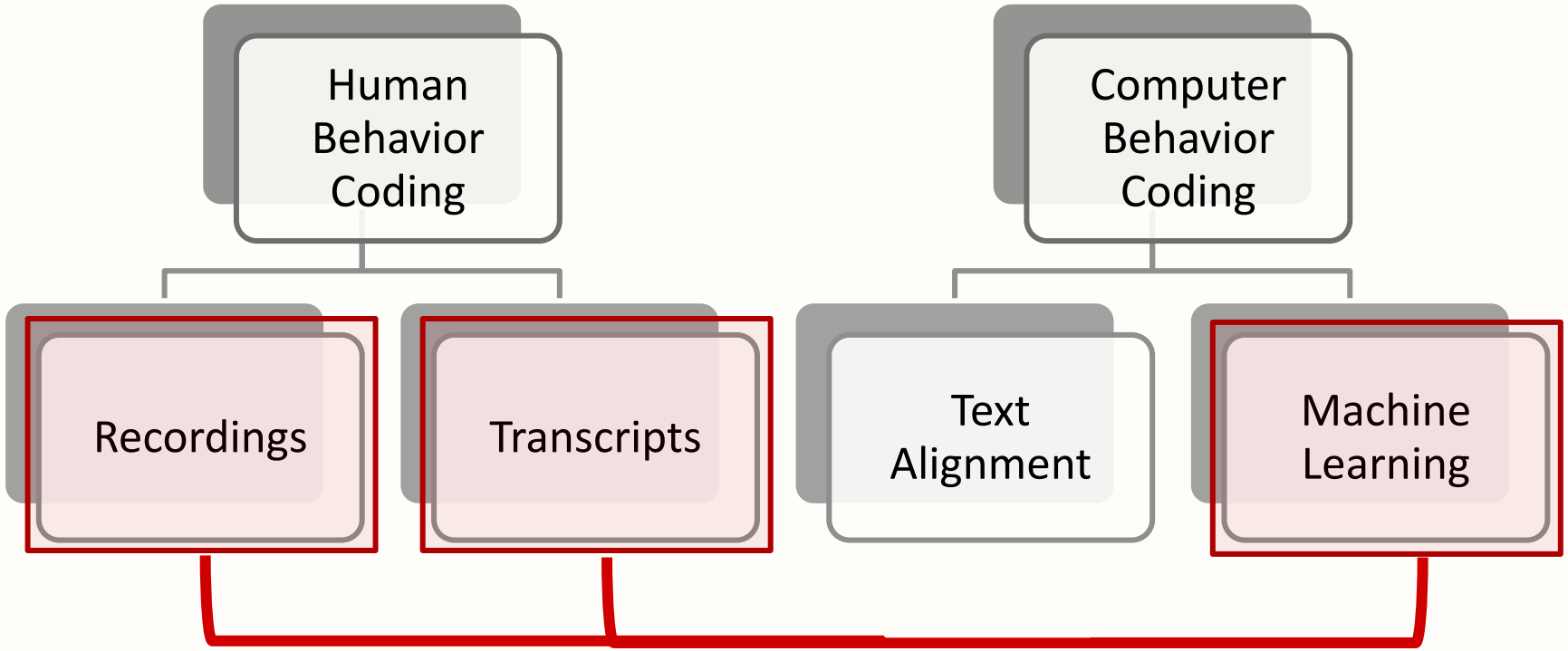
and so, what is your sex?



Recurrent Neural Networks (RNNs)

- **Pros**
  - Once a network is trained for a question, the “per case” cost does not increase
  - Accounts for the sequential nature of text data
    - Discovers the contextual importance of words (e.g., exact readings vs. major changes)
- **Cons**
  - Transcripts required for every case
  - Technical knowledge required
  - Computing power required

The Present Study



Research Questions

- **Can recurrent neural networks partially automate the coding of interviewer question-asking behaviors?**
  - Is the accuracy of RNN coding comparable to human coders?
  - Does the accuracy of RNN coding depend on having a high prevalence of each behavior in the training data?

Data

- **Work and Leisure Today 2 Survey**
  - National telephone survey of U.S. adults conducted by Abt SRBI in Aug and Sept of 2015.
  - Dual-Frame (Landline and Mobile)
  - n=902; AAPOR RR3 = 7.8%
  - 58 questions
  - All interviews were recorded and transcribed

Operationalizing Question-Asking Turns and the Outcome

- **Question-Asking Turns**
  - The interviewer's first conversational turn for each question
    - Conversational turn = uninterrupted speech by one actor
  - Excludes turns where:
    - The interviewer was interrupted by the respondent
    - The interviewer stuttered
- **Outcome variable**
  - Behavior Code assessing interviewer's question-asking
    - Exact Reading
    - Minor Change
    - Major Change

Questions Selected for Analysis

- **6 Questions selected for analysis**
  - 3 questions with a high prevalence of major/minor changes in training data

Question Topic	Question Type	% Exact Readings	% Minor Changes	% Major Changes
Respondent Sex	Demographic	64.7%	20.4%	14.9%
Leisure Activities	Open-Ended	80.5%	9.3%	10.2%
Access to Job Equipment	Attitude/Opinion	69.4%	14.8%	15.8%

Questions Selected for Analysis

- **6 Questions selected for analysis**
  - 3 questions with a low prevalence of major/minor changes in training data

Question Topic	Question Type	% Exact Readings	% Minor Changes	% Major Changes
# Adults in HH	Demographic	85.7%	11.8%	2.5%
Threats to Personal Privacy	Open-Ended	90.4%	4.0%	5.6%
Own Cell or Smartphone	Frame	72.4%	25.7%	1.9%

Human Coding

- **Undergrad-Coded Data**
  - 16 undergraduate students
  - Coded question-asking behaviors for 899 interviews
    - 45,078 question-asking turns
  - Used Sequence Viewer software (Dijkstra 1999)
    - Allows coders to read and hear question-asking
- **Master-Coded Data (Ground Truthing)**
  - A 10% random subsample of the all data (94 cases, 4,688 question-asking turns)

Recurrent Neural Network Coding

- **Network Creation (for each question)**
  - Training Data (80% of undergrad-coded data)
    - Random subsample of the data
  - Validation Data (20% of undergrad-coded data)
    - Ensures model does not overfit data
  - k-fold cross-validation for each question
    - *k*=5 different networks with *k*=5 different subsamples per question
      - Most accurate network on validation method retained
      - Minimize selection of a “bad” training set
- **Testing Data (100% of master-coded data)**
  - Cases not used for training or validation

Analyses

- **Kappa Scores**
  - Calculated by question between:
    - Undergrad coders and master coders
    - RNN coding and master coders
  - Test for differences in kappas across undergrads and RNNs by question:
    - Bootstrapped 95% confidence intervals (*B*=500)
    - Non-overlapping CI's as statistically significantly different

Analyses

- **Percent Agreement / Accuracy by Question**
  - An alternative to kappa for rare events (Viera & Garrett 2005)
  - $\frac{\#\ of\ correct\ predictions}{Total\ \#\ of\ predictions}$

- **Recall**
  - For each behavior, what proportion of master-coder identified behaviors (i.e., exact reading, minor change, or major change) was identified by the RNN or the undergrad coders?
  - $\text{Recall}_{minor} = \frac{\#\ of\ times\ RNN\ AND\ master\ coder\ identified\ a\ minor\ change}{\#\ of\ times\ master\ coder\ identified\ a\ minor\ change}$

Results – Kappas and % Agreement / Accuracy

- **<u>High Prevalence</u> of Major/Minor Changes**

Results – Kappas and % Agreement / Accuracy

- **High Prevalence of Major/Minor Changes**

Question Topic	Question Type	Kappa with Master Coder				% Agree / Accuracy with Master Coders		
		Undergrad	RNN	Diff		Undergrad	RNN	Diff
Respondent Sex	Demographic	0.39	0.47	-0.08		61.73%	65.88%	-4.15%
Leisure Activities	Open-Ended	0.60	0.57	0.03		84.52%	82.76%	1.76%
Access to Job Equipment	Attitude/Opinion	0.70	0.80	-0.10		85.71%	90.48%	-4.77%

*p<.05

Results - Recall

- **High Prevalence of Major/Minor Changes**

Question Topic	Question Type	Exact Reading			Minor Change			Major Change	
		Undergrad	RNN		Undergrad	RNN		Undergrad	RNN
Respondent Sex	Demographic	80.55%	94.44%		29.63%	16.67%		72.22%	89.47%
Leisure Activities	Open-Ended	100.00%	98.36%		21.43%	13.33%		81.82%	90.09%
Access to Job Equipment	Attitude/Opinion	96.55%	100.00%		50.00%	50.00%		80.00%	100.00%

$$\text{Recall}_{minor} = \frac{\#\ of\ times\ RNN\ AND\ master\ coder\ identified\ a\ minor\ change}{\#\ of\ times\ master\ coder\ identified\ a\ minor\ change}$$

Results – Kappas and % Agreement / Accuracy

- **Low Prevalence of Major/Minor Changes**

Question Topic	Question Type	Kappa with Master Coder				% Agree / Accuracy with Master Coders		
		Undergrad	RNN	Diff		Undergrad	RNN	Diff
# Adults in HH	Demographic	0.20	-0.02	0.22*		73.47%	67.31%	6.16%
Threats to Personal Privacy	Open-Ended	0.80	0.55	0.25		95.29%	89.77%	5.52%
Own Cell or Smartphone	Frame	0.38	0.11	0.27*		63.64%	45.68%	17.96%

*p<.05

Results - Recall

- **Low Prevalence of Major/Minor Changes**

Question Topic	Question Type	Exact Reading			Minor Change			Major Change	
		Undergrad	RNN		Undergrad	RNN		Undergrad	RNN
# Adults in HH	Demographic	100.00%	97.22%		14.28%	0.00%		100.00%	0.00%
Threats to Personal Privacy	Open-Ended	98.65%	98.67%		83.33%	12.50%		60.00%	80.00%
Own Cell or Smartphone	Frame	96.30%	89.29%		48.89%	26.09%		20.00%	0.00%

$$\text{Recall}_{minor} = \frac{\#\ of\ times\ RNN\ AND\ master\ coder\ identified\ a\ minor\ change}{\#\ of\ times\ master\ coder\ identified\ a\ minor\ change}$$

Summary of Findings

- **RNN-coding is comparable to undergrad coding when there is a *high prevalence* of deviations from exact reading in the training data**
  - More deviations to train on = RNN coding comparable to undergrads
- **RNN-coding performs worse than undergrad coding when there is a *low prevalence* of deviations from exact reading in the training data**
  - Fewer deviations to train on = worse RNN coding
- **RNN-coding has trouble identifying minor changes**
  - But minor changes do not affect question meaning, and may not be of interest for most researchers.

Limitations and Future Directions

- **Many opportunities to refine RNN model**
  - Collapsing Exact Readings and Minor Changes
  - Change RNN parameters to change performance (e.g., # of hidden neurons, layers, learning rate)
  - Minimum amount of training data required to produce results comparable to undergrads
- **More behaviors!**
- **Refining training/validation/test sources**
  - Mixing undergrad- and master-coded data into training/validation?

Final Takeaway

- **RNNs are a promising method for automating behavior coding of interviewer question-reading behaviors, especially when many major changes are expected.**
  - More results after more optimization!

Thank you!

Contact:
Jerry Timbrook
jerry.timbrook@gmail.com