Modeling Worker Career Trajectories with Neural Sequence Models

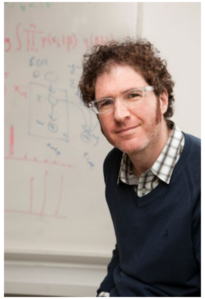
Workshop on Causal Inference & NLP at EMNLP 2021

Keyon Vafa
Columbia University

Emil Palikot
Stanford University

Tianyu Du
Stanford University

Ayush Kanodia
Stanford University



David Blei
Columbia University

Susan Athey
Stanford University

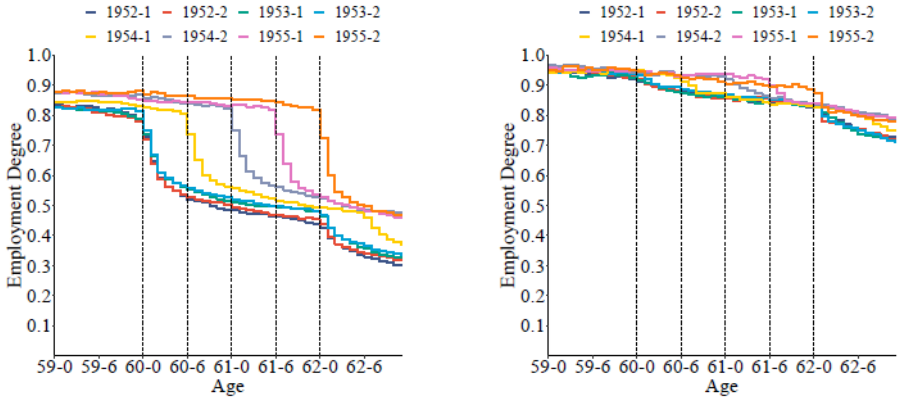
Background: Machine Learning & Natural Experiments for Labor Transitions

Danish Retirement Reform

Retirement reform pushed back age of eligibility for early retirement benefits by six months at a time over several years

When early retirement is available, large chunk of people take it shortly after eligibility

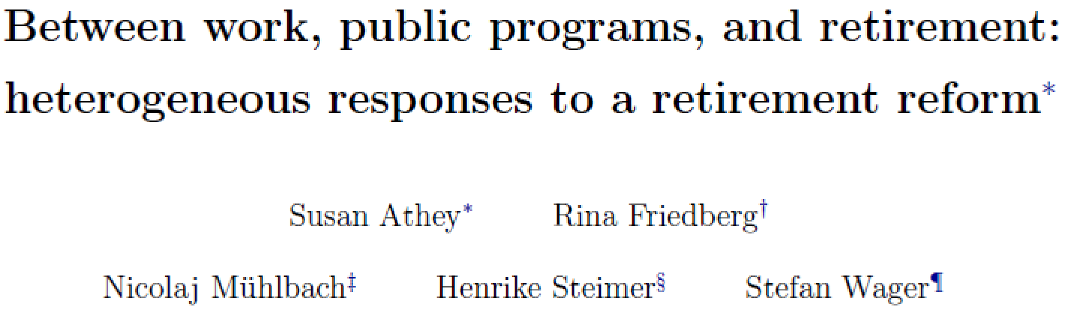
Estimating treatment effect of early retirement benefits on working boils down to predicting who takes the benefit, since without the benefit, very slow decline in employment with age



(a) Lower secondary education (b) Master's degree or equivalent

Figure 2: Employment Trends by Cohort and Education

Notes: This figure shows the trends in employment degree by cohort and education across age. Each cohort is marked by a different color. The figure plots individuals with a lower secondary education as the highest completed education (LHS) as well as individuals with a master's



Between work, public programs, and retirement: heterogeneous responses to a retirement reform[*]

Susan Athey[*] Rina Friedberg[†]

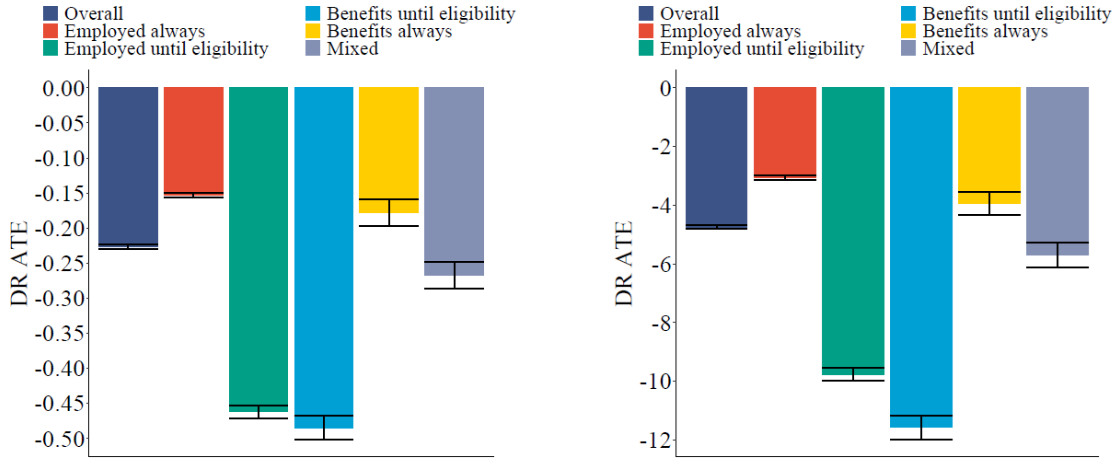
Nicolaj Mühlbach[‡] Henrike Steimer[§] Stefan Wager[¶]

Danish Retirement Reform

We use ML to characterize workers by their predicted “paths” based on characteristics

Bars correspond to different buckets of individuals, classified by their predicted path based on their histories

Our path prediction model successfully predicts who will stop working



(a) Binary IEM

(b) Continuous IEM

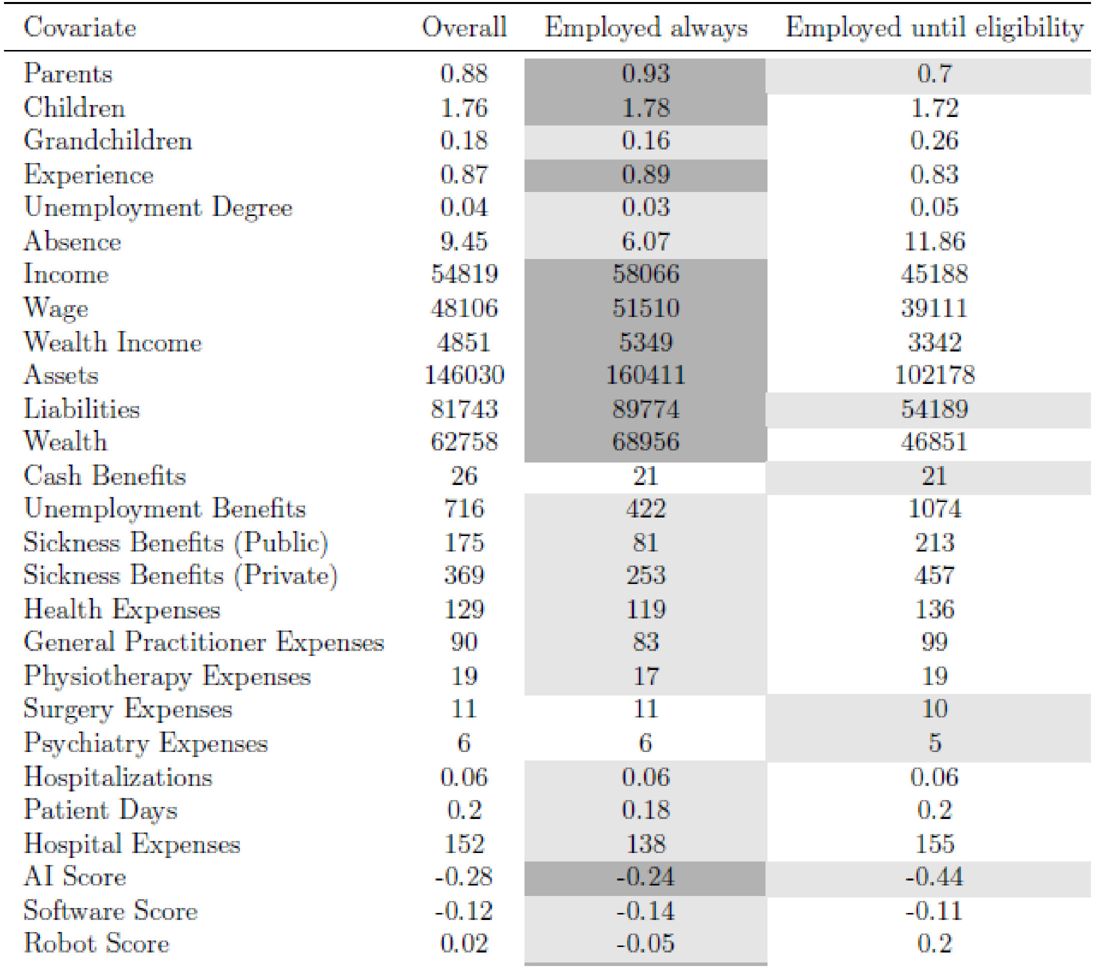
Figure 4: DR ATE on Employment

Notes: This figure shows the estimated DR ATE with a 95% CI on employment overall and by predicted retirement path for the main cohorts in the binary IEM (LHS) and the continuous IEM (RHS). For the binary model, the treatment effect is the probability of being mainly employed six months after becoming eligible to retire early, whereas for the continuous model, the treatment effect is the change in number of weeks employed six months after eligibility.

Danish Retirement Reform

We use ML to characterize workers by their predicted "paths" based on characteristics

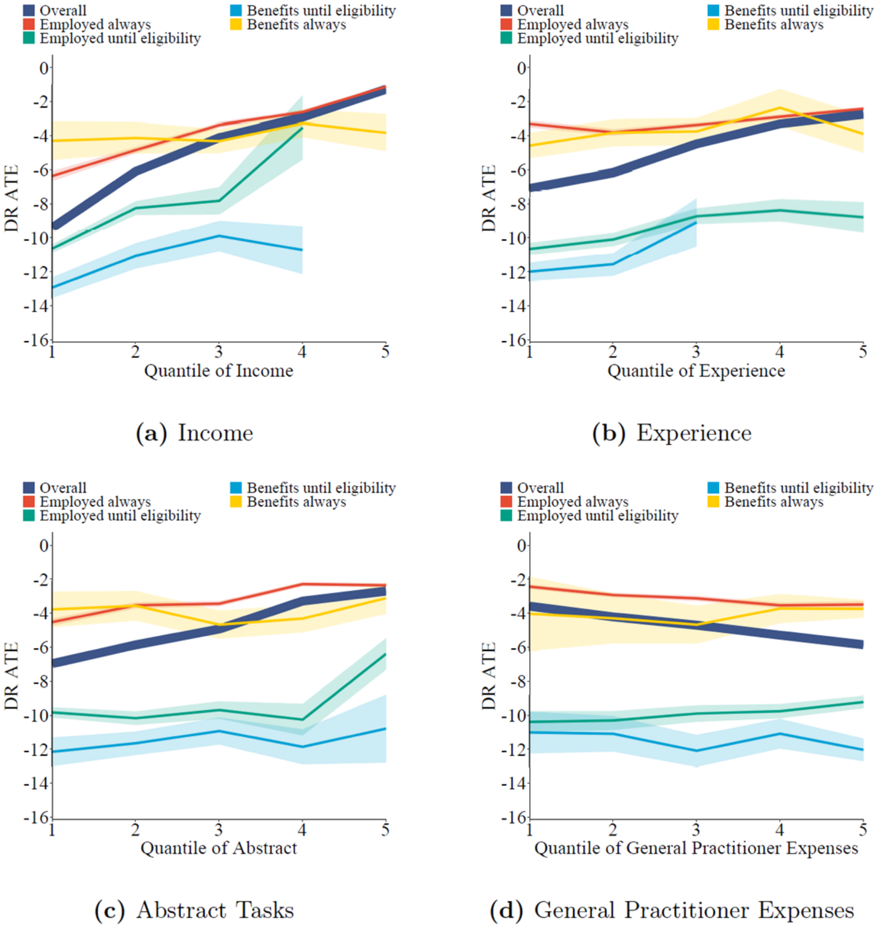
Table shows differences in characteristics across groups



Covariate	Overall	Employed always	Employed until eligibility
Parents	0.88	0.93	0.7
Children	1.76	1.78	1.72
Grandchildren	0.18	0.16	0.26
Experience	0.87	0.89	0.83
Unemployment Degree	0.04	0.03	0.05
Absence	9.45	6.07	11.86
Income	54819	58066	45188
Wage	48106	51510	39111
Wealth Income	4851	5349	3342
Assets	146030	160411	102178
Liabilities	81743	89774	54189
Wealth	62758	68956	46851
Cash Benefits	26	21	21
Unemployment Benefits	716	422	1074
Sickness Benefits (Public)	175	81	213
Sickness Benefits (Private)	369	253	457
Health Expenses	129	119	136
General Practitioner Expenses	90	83	99
Physiotherapy Expenses	19	17	19
Surgery Expenses	11	11	10
Psychiatry Expenses	6	6	5
Hospitalizations	0.06	0.06	0.06
Patient Days	0.2	0.18	0.2
Hospital Expenses	152	138	155
AI Score	-0.28	-0.24	-0.44
Software Score	-0.12	-0.14	-0.11
Robot Score	0.02	-0.05	0.2

Danish Retirement Reform

TREATMENT EFFECTS BY PREDICTED PATH AND QUARTILES OF COVARIATES



(a) Income

(b) Experience

(c) Abstract Tasks

(d) General Practitioner Expenses

Figure 5: DR ATE on Employment by Continuous Covariates

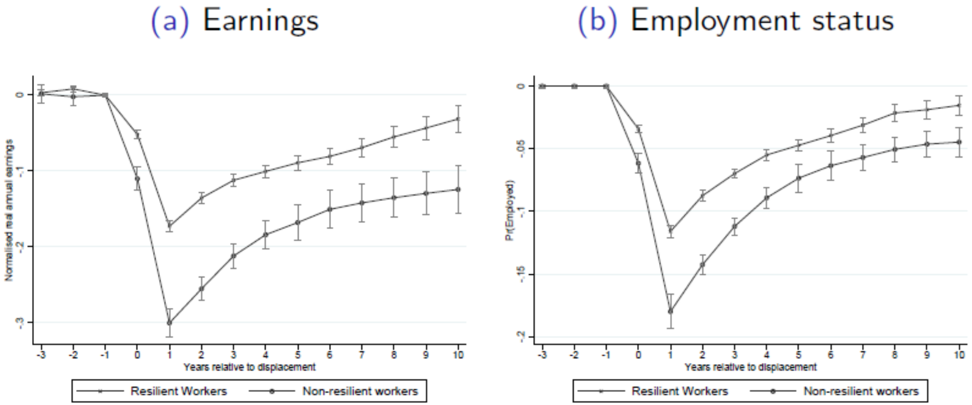
Notes: This figure shows the DR ATE with a 95% CI overall and by predicted retirement path across quintiles of four selected covariates (*income*, *experience*, *abstract tasks*, and *general practitioner expenses*) as estimated by the continuous IEM using a six-month horizon. The sample is the training sample of the main cohorts and all predictions are out-of-bag.

Swedish Plant Closures

Use machine learning to identify “resilient” and “non-resilient” groups.

The separation in the figures shows that resilience is highly predictable.

Gives evaluation of earnings for observations with large earnings loss in the first year (non-resilient) and smaller earnings loss in the first year (resilient)


(a) Earnings

(b) Employment status

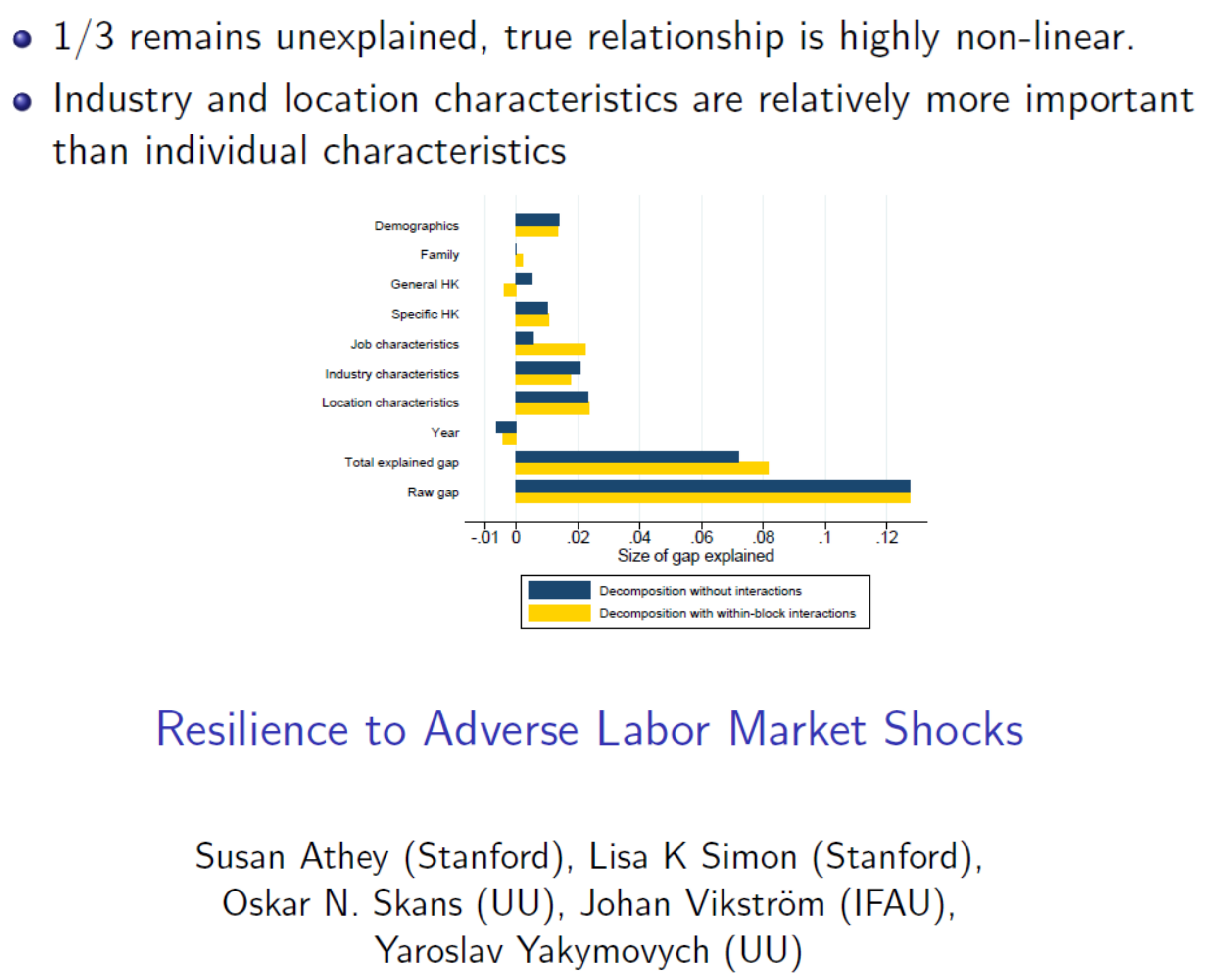
Resilience to Adverse Labor Market Shocks

Susan Athey (Stanford), Lisa K Simon (Stanford),
Oskar N. Skans (UU), Johan Vikström (IFAU),
Yaroslav Yakymovych (UU)

Swedish Plant Closures

What explains resilience?

Use Gelbach decomposition to explain using categories of covariates



- 1/3 remains unexplained, true relationship is highly non-linear.
- Industry and location characteristics are relatively more important than individual characteristics

Resilience to Adverse Labor Market Shocks

Susan Athey (Stanford), Lisa K Simon (Stanford), Oskar N. Skans (UU), Johan Vikström (IFAU), Yaroslav Yakymovych (UU)

Modeling Worker Career Trajectories with Neural Sequence Models

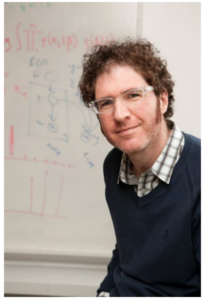
Preliminary!!!!

Keyon Vafa
Columbia University

Emil Palikot
Stanford University

Tianyu Du
Stanford University

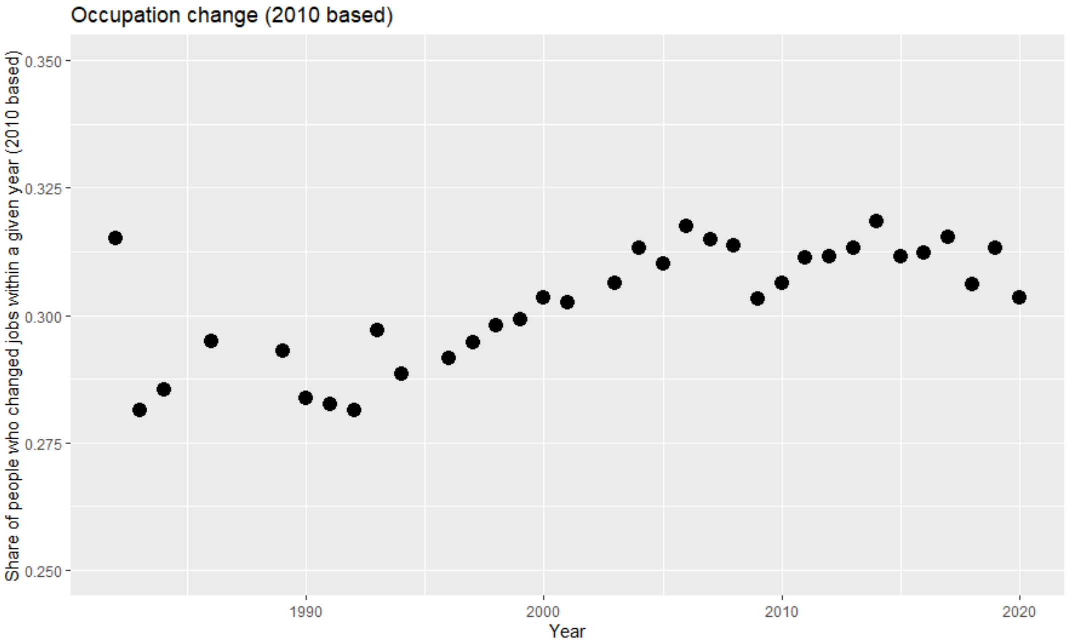
Ayush Kanodia
Stanford University



David Blei
Columbia University

Susan Athey
Stanford University

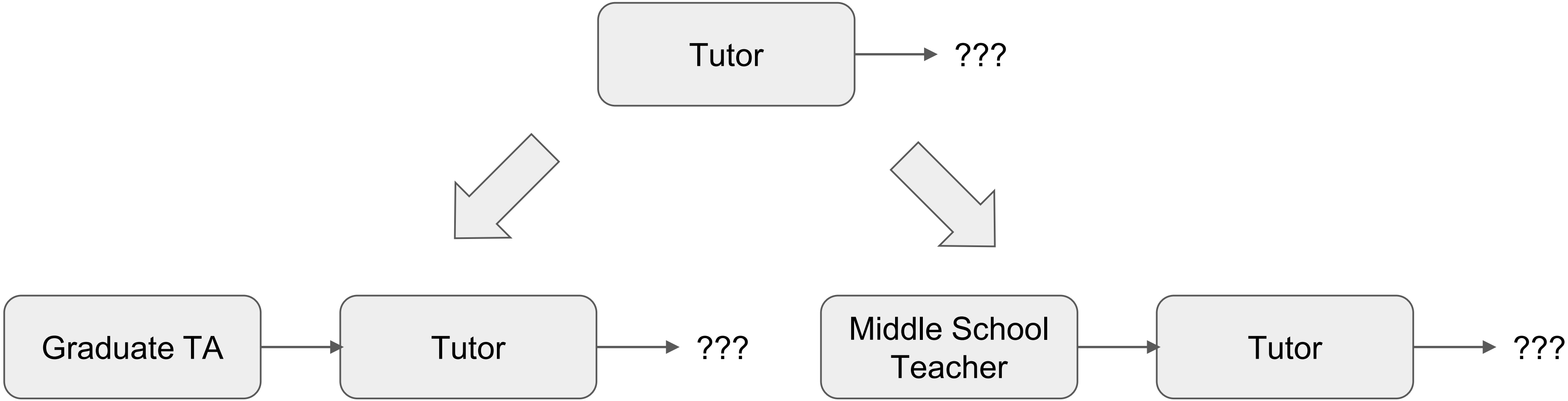
Increasing Rate of Job Transitions



Individuals born between 1957-64 held an average of 12 jobs before turning 55 (Bureau of Labor Statistics)

Motivation

Motivation: Analyzing careers instead of individual jobs.



Transitions depend on entire careers, not just the current job.

Goal

Goal: Modeling labor transitions as a function of career histories and covariates.

Ultimate questions include:

- **Jobs** (which jobs should I take if I want a specific job in five years?)
- **Educational degrees** (which degree will maximize my expected salary?)
- **Individual characteristics** (what is the gender wage gap?)

Challenges

Data: Most economists rely on administrative data, which either follows individuals for a short period of time (CPS) or contains few individuals (PSID).

Model Flexibility: Classic economic analyses (e.g. Toikka, 1976) use 1st order Markov models that are unable to capture complex career trajectories. Subsequent literature typically continues to use models with very small/simple state space (e.g. in/out of labor force), with big focus on unobservables (e.g. worker "ability").

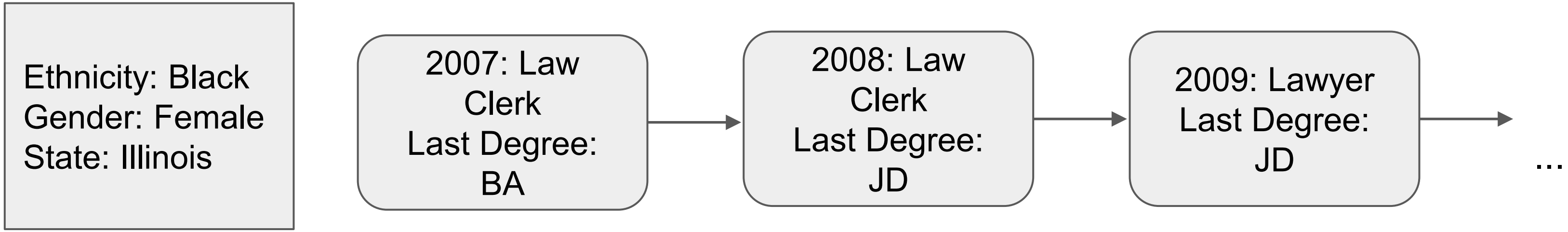
- **Recent work incorporating NLP-inspired dimension reduction for labor transitions:** Karthik Rajkumar, Lisa Simon, and Susan Athey. A Bayesian Approach to Predicting Occupational Transitions. PhD thesis, Stanford University, 2021. Applies matrix factorization to U.S. administrative data (thousands of workers).
    - In turn builds on Donnelly, Ruiz, Athey & Blei (forthcoming), Ruiz, Athey & Blei (2019) which applies NLP-inspired matrix factorization to supermarket shopping. There, dimension reduction helps with confounding; causal inference on price changes. Lots of price variation in data; make use of consumers arriving before & after price change.

Richard S Toikka. A Markovian model of labor market decisions by workers. *The American Economic Review*, pages 821–834, 1976.

Dataset

Zippia dataset: Resumes posted online from 24 million American workers.

Each resume is a sequence of jobs (coded into one of 1080 occupational categories) and covariates:



Ethnicity: Black
Gender: Female
State: Illinois

2007: Law Clerk
Last Degree: BA

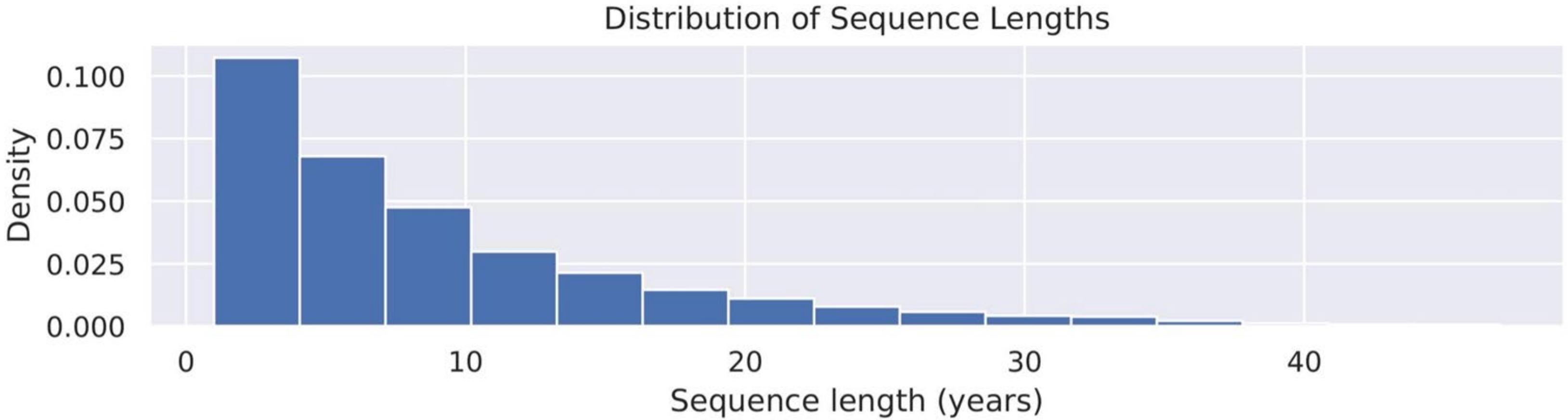
2008: Law Clerk
Last Degree: JD

2009: Lawyer
Last Degree: JD

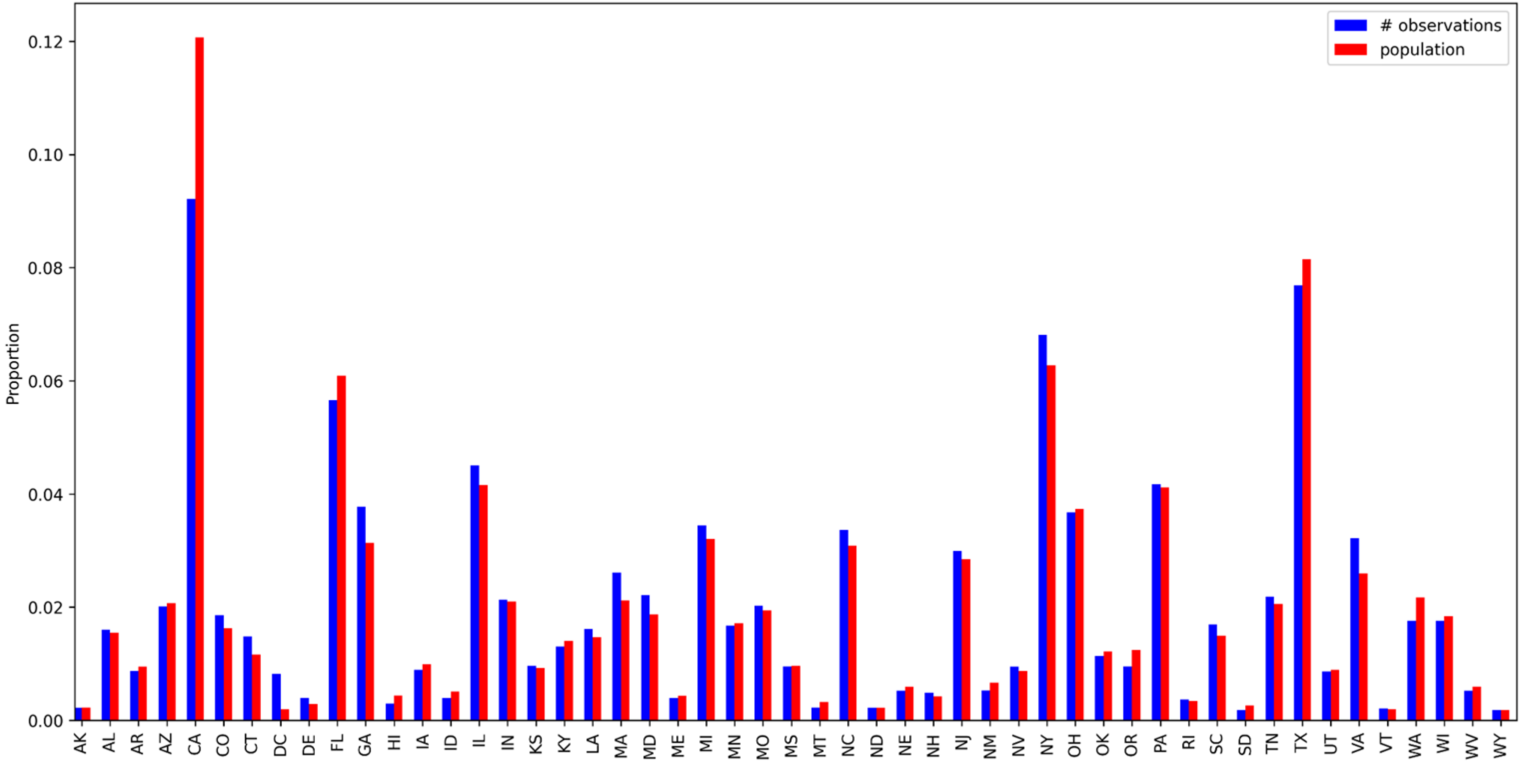
...

x 24 million

Dataset


Distribution of Sequence Lengths
Density
0.100
0.075
0.050
0.025
0.000
0
10
20
30
40
Sequence length (years)

Dataset



Geographical distribution is representative of US population.

Economic Model

Sequence of jobs for individual *u*: $\mathbf{y}_{u,t} = (y_{u,1}, \ldots, y_{u,t})$

Sequence of covariates for individual *u*: $\mathbf{x}_{u,t} = (x_{u,1}, \ldots, x_{u,t})$

Unobservables for individual *u*: z_u

Underlying job-worker matching mechanism:

$$p(y_{u,t} = j | \mathbf{y}_{u,t-1}, \mathbf{x}_{u,t}, z_u)$$

Economic Assumptions

Assumption 1: Existence of low-dimensional job representations.

There exist job representations $h(\mathbf{y}_{u,t-1}, \mathbf{x}_{u,t}) \in \mathbb{R}^D$ that fully specify job transitions: $y_{u,t} \perp\!\!\!\perp (\mathbf{y}_{u,t-1}, \mathbf{x}_{u,t}) \mid h(\mathbf{y}_{u,t-1}, \mathbf{x}_{u,t})$

A1 is WLOG for large enough *D*; useful if holds (approximately) for smaller *D*

Assumption 2: Sequential unconfoundedness:

$$p(y_{u,t} = j | \mathbf{y}_{u,t-1}, \mathbf{x}_{u,t}, z_u) = p(y_{u,t} = j | \mathbf{y}_{u,t-1}, \mathbf{x}_{u,t}).$$

A2 is strong assumption; imposing it enables causal interpretation

A2 is **in general violated** in labor market data, but applying causal techniques can still result in more interpretable descriptive analysis

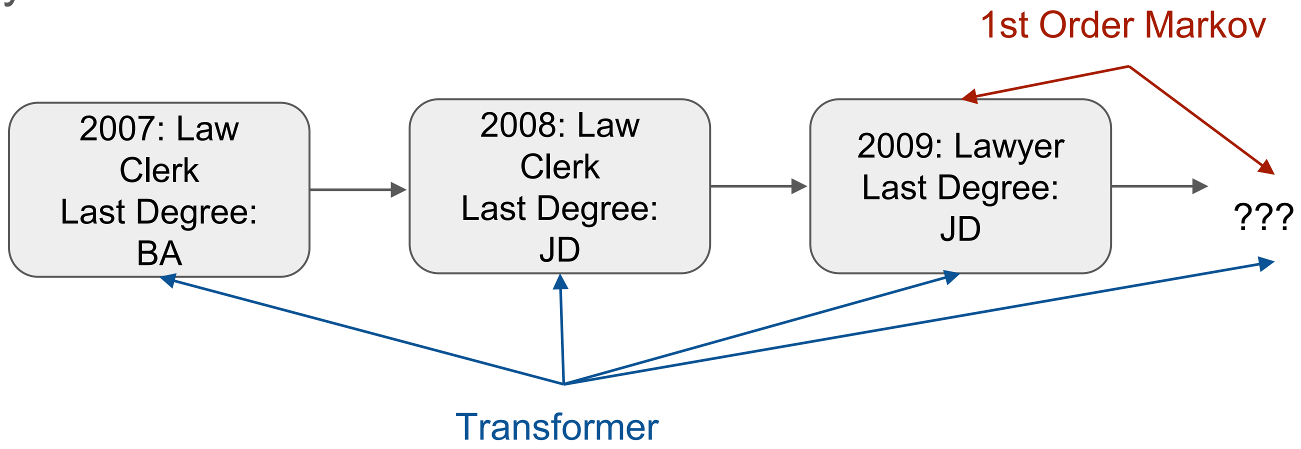
In some cases violations may be mild

Prioritize future investigations (e.g. using natural experiments) & generate hypotheses

Model Parameterization

We use a **transformer neural network** (originally developed for text by Vaswani et al., 2017) to parameterize $p(y_{u,t} = j | \mathbf{y}_{u,t-1}, \mathbf{x}_{u,t})$.

Unlike a first-order Markov model, a transformer can condition on all previous jobs in a history.



1st Order Markov
2007: Law Clerk Last Degree: BA
2008: Law Clerk Last Degree: JD
2009: Lawyer Last Degree: JD
???
Transformer

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

Adapting Transformers for Careers

Sequences of jobs differ from sequences of words:

1. **Repeated jobs**: 76% of individuals remain in the same job from one year to the next.

1. **Covariates**: Job transitions depend on static (e.g. ethnicity) and time-varying (e.g. most recent degree) covariates.

Modeling Repeated Jobs

Transformers estimate representations $h(\mathbf{y}_{t-1}, \mathbf{x}_t) \in \mathbb{R}^D$ for each job and covariate.

We make predictions in two stages:

1. Predict the likelihood an individual stays: $r_t = \sigma\left(\eta \cdot h(\mathbf{y}_{t-1}, \mathbf{x}_t)\right)$

1. Marginalize over the probability a job is repeated:

$$p(y_t = j | \mathbf{y}_{t-1}, \mathbf{x}_t) = \begin{cases} r_t & \text{if } j = y_{t-1} \\ (1 - r_t) * \frac{\exp\{\beta_j \cdot h(\mathbf{y}_{t-1}, \mathbf{x}_t)\}}{\sum_{j' \neq y_{t-1}} \exp\left\{\beta_{j'} \cdot h(\mathbf{y}_{t-1}, \mathbf{x}_t)\right\}} & \text{otherwise.} \end{cases}$$

Incorporating covariates

A progression of representations $h^{(1)}(\mathbf{y}_{t-1}, \mathbf{x}_t), \ldots, h^{(L)}(\mathbf{y}_{t-1}, \mathbf{x}_t)$ are estimated recursively, with the final used to predict the next job.

$h^{(1)}(\mathbf{y}_{t-1}, \mathbf{x}_t)$ combines the most recent job (y_{t-1}), the current position (t), and **the most recent covariate** (x_t) using three embedding functions:

$$h^{(1)}(\mathbf{y}_{t-1}, \mathbf{x}_t) = e_y(y_{t-1}) + e_t(t) + e_x(x_{t+1})$$

where $e_y : \mathcal{Y} \to \mathbb{R}^D$, $e_t : \mathcal{T} \to \mathbb{R}^D$, and $e_x : \mathcal{X} \to \mathbb{R}^D$ ($\mathcal{Y}$, $\mathcal{T}$, and $\mathcal{X}$ denote the space of all possible jobs, career lengths, and covariates, respectively).

The subsequent representations are calculated using multi-head attention and feedforward neural network layers.

Results: Comparison to 1st Order Markov Model

Evaluate predictive performance using held-out perplexity (lower is better):

$$\exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log p(y_t|\mathbf{y}_{t-1},\mathbf{x}_t)\right)$$

Intuition:

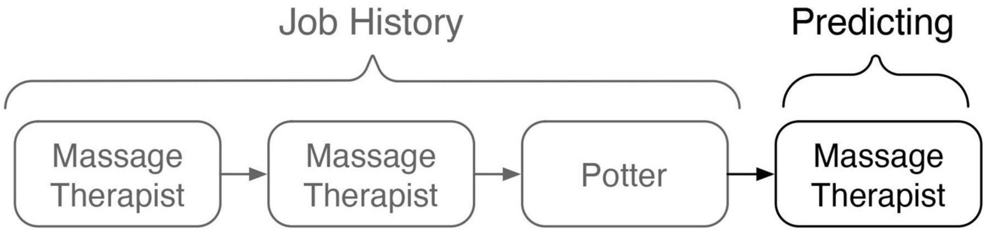
- Monotone transformation of log-likelihood where magnitude is interpretable
- Model with perplexity *k* performs as well as a model that is uniformly uncertain among *k* alternatives

First-Order Markov perplexity: 6.52

Transformer perplexity: 5.00

Results: Comparison to 1st Order Markov Model

Transformer improves upon Markov most when predicting job that is repeated non-consecutively in a sequence.



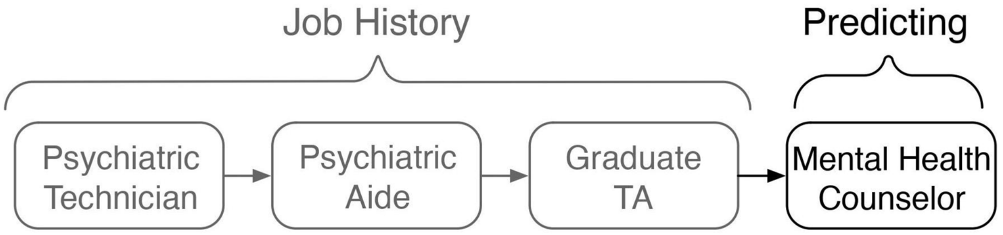
Markov probability of Massage Therapist: 0.002
Transformer probability of Massage Therapist: 0.337

Markov perplexity for non-consecutively repeated jobs: 294.5 (renormalized to condition on changing job: 91.3)

Transformer perplexity for non-consecutively repeated jobs: 19.5 (renormalized: 8.4)

Results: Comparison to 1st Order Markov Model

Transformer also improves when predicting job that hasn't appeared yet in a sequence.



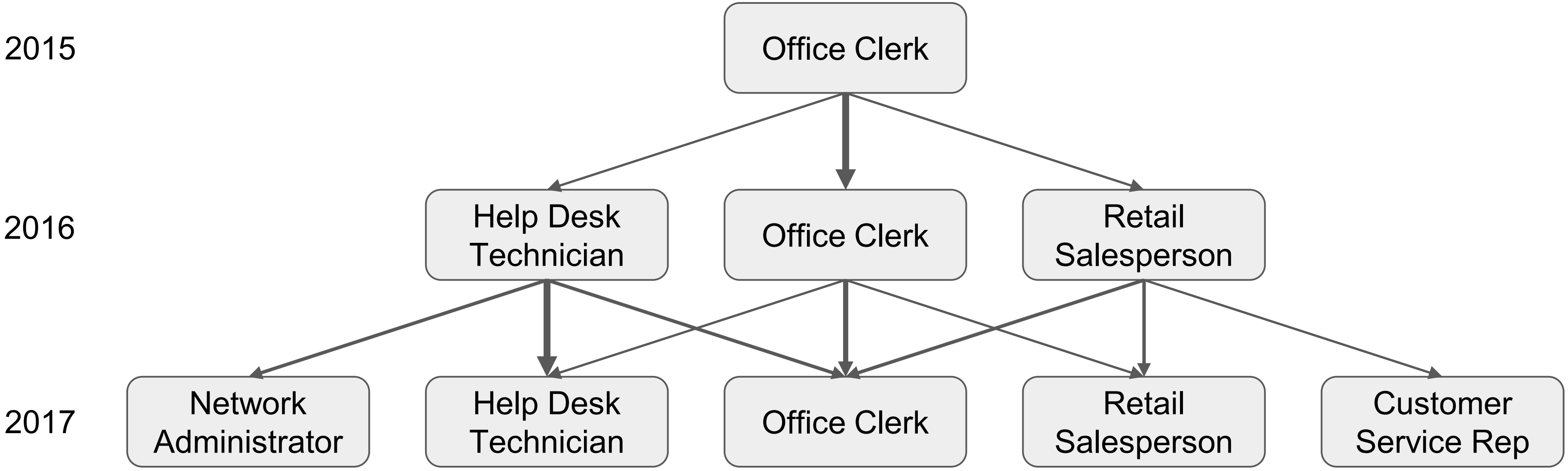
Markov probability of Mental Health Counselor: 0.003
Transformer probability of Mental Health Counselor: 0.060

Markov perplexity for new jobs: 553.9 (renormalized: 168.2)

Transformer perplexity for new jobs: 244.5 (renormalized: 92.3)

Sampling Career Trajectories from the Model

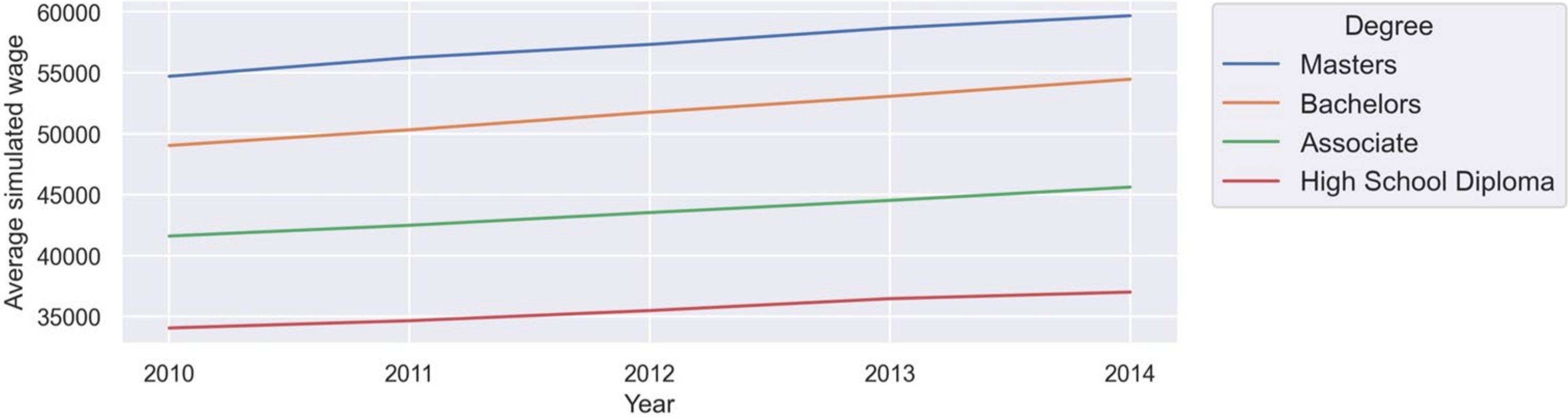
For a Black male in California starting out as an Office Clerk with an associate's degree in 2015:



Model Application:
Simulating future wages for a given worker profile

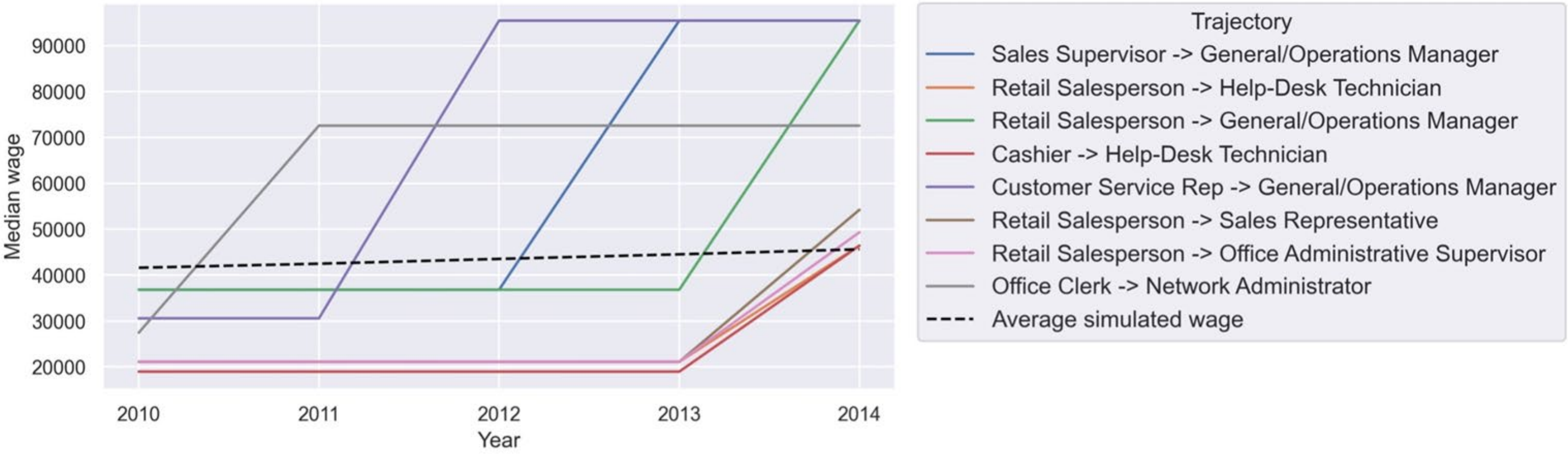
Simulate future wages by linking to median income per simulated job.

Application: compare trajectories by education level. The comparisons do NOT have a causal interpretation, because people with different degrees have different unobservables. But model allows you to focus on specific demographic groups

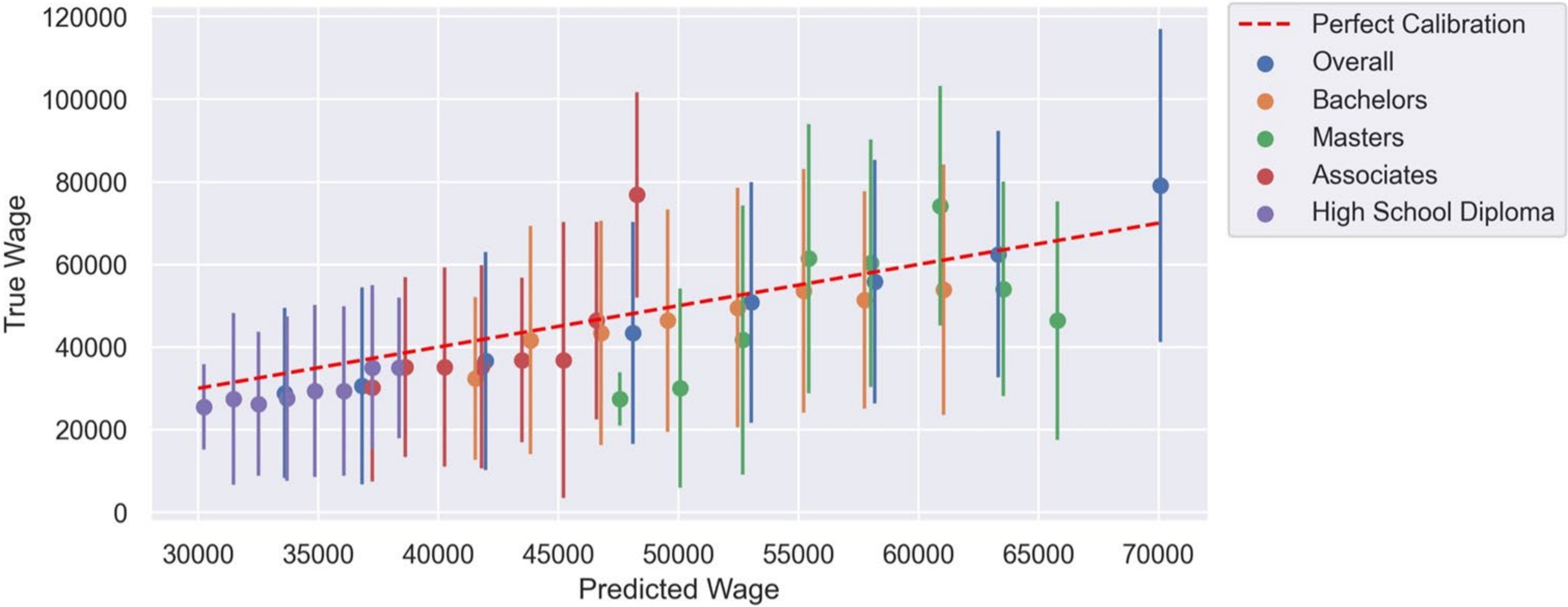


Model Application:
Simulating future wages for a given worker profile

Most common simulated upward trajectories (here, trajectories where median wages for the occupation double in a 5 year period) for Hispanic/Latino employee in CA with first recorded job in 2010 and with **associate's degree**:

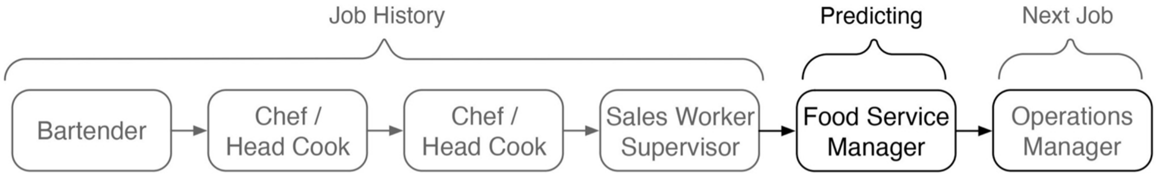


Calibration of 5-Year Wage Predictions



Model Application: Imputing Intermediate Jobs

If a worker is employed at job A in year *t* and job C in year *t* + 2, what is the distribution of jobs at time *t + 1*?



Job History

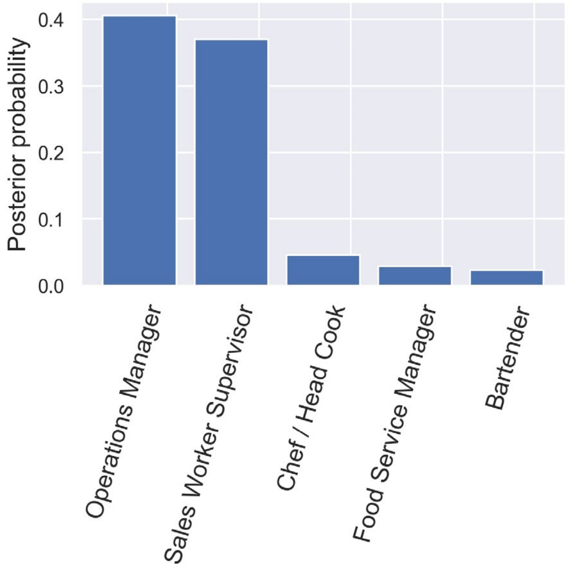
Predicting

Next Job

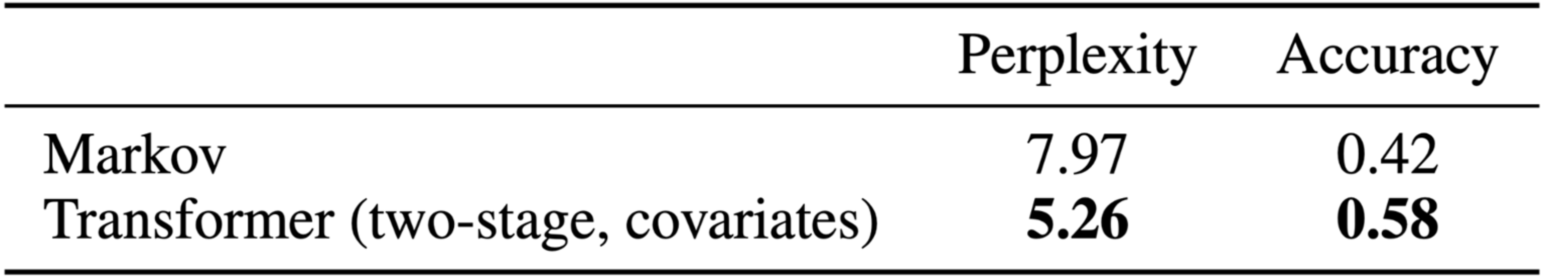
Bartender → Chef / Head Cook → Chef / Head Cook → Sales Worker Supervisor → Food Service Manager → Operations Manager

Markov probability of Food Service Manager: 0.002
Transformer probability of Food Service Manager: 0.029

Transformer posterior distribution:



Model Application: Imputing Intermediate Jobs



	Perplexity	Accuracy
Markov	7.97	0.42
Transformer (two-stage, covariates)	**5.26**	**0.58**

Transformer can correctly predict 58% of intermediate jobs between differing jobs.

Case Study: Gender Gap

Do men and women have different career trajectories?

Specifically: Do men and women with similar histories and covariates transition to occupations with different median incomes?

Differences can arise at different points in career:

- First job (conditioning on covariates: education, state, ethnicity, etc.)
- Mid-career (conditioning on all previous jobs and covariates)

Interpretation:

- Gender affects preferences & opportunities, thus choices and outcomes
- Men & women with similar observables will in general have different unobservables
- Finding large gaps suggests further study to disentangle sources of gap

Case Study: Gender Gap

Conditional on job history and covariates, is there a gender gap in next-job median salary?

For example: Consider all customer service representatives with 5 years of work experience.

Unadjusted estimator doesn't account for different job histories:

$$\mathbb{E}[\text{next-job median salary}|\text{male}] - \mathbb{E}[\text{next-job median salary}|\text{female}]$$

The first 5 years of work experience may be different for males and females.

Curse of Dimensionality prevents from matching on previous job experience.

Adjusted Estimator

Recall **Assumption 1**: There exist low-dimensional job representations that fully specify the distribution of job transitions as a function of previous jobs and covariates:

$$y_{u,t} \perp\!\!\!\perp (\mathbf{y}_{u,t-1}, \mathbf{x}_{u,t}) \mid h(\mathbf{y}_{u,t-1}, \mathbf{x}_{u,t})$$

The **adjusted estimator** uses estimated representations to compare average outcomes for men and women who have similar histories/covariates.

Adjusting for Predicted Wage

Since estimated representations are high-dimensional, adjusted estimator uses **predicted next-occupation wage** to group individuals.

For median wage m_j of occupation j, predicted wage $\hat{w}_t$ is given by

$$\hat{w}_t = \sum_j p(y_t = j | \mathbf{y}_{t-1}, \mathbf{x}_t) * m_j.$$

Problem: If p is trained on all career sequences, $\hat{w}_t$ may encode information about gender, so men and women with similar $\hat{w}_t$ may not have similar backgrounds.

Solution: Train p on only male careers, interpret $\hat{w}_t$ as **predicted male-wage**.

Decomposing Gender Gaps: 2nd Year

Unadjusted gender gap for individuals who are customer service reps in their 1st yr: $1586.

Adjusted gender gap accounts for differences of education, state, ethnicity, and year of entry:

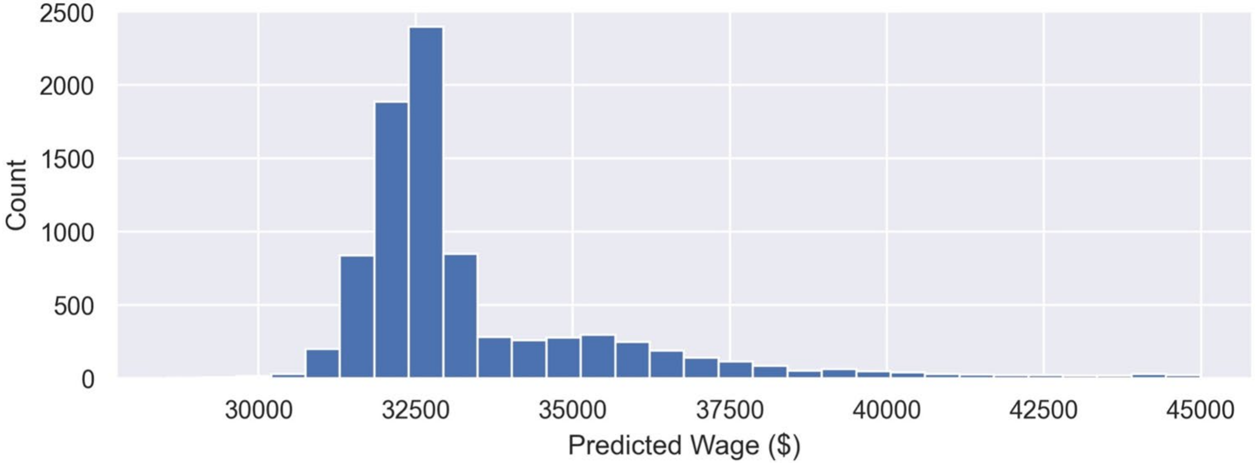
1. Predict 2nd-year wage for all 1st-year customer service reps (based on covariates and 1st-year customer service job).
2. Divide the salary predictions into 10 deciles.
3. The gender gap for each decile is the difference between sample means for males and females.
4. The adjusted gender gap is the average over all deciles.

Adjusted gender gap: 1st-year male customer service representatives transition to jobs with median annual income **$1357 higher** than females.

- Interpretation: 14% of gap accounted for by differences in the gender distribution by education, state, ethnicity, and year of entry.

Predicted 2nd-Year Wages

Predicted "male-wage" for all 1st-year customer service reps:



Quartile 1
Most common degree: High School Diploma

Quartile 2
Most common degree: High School Diploma

Quartile 3
Most common degree: Associate

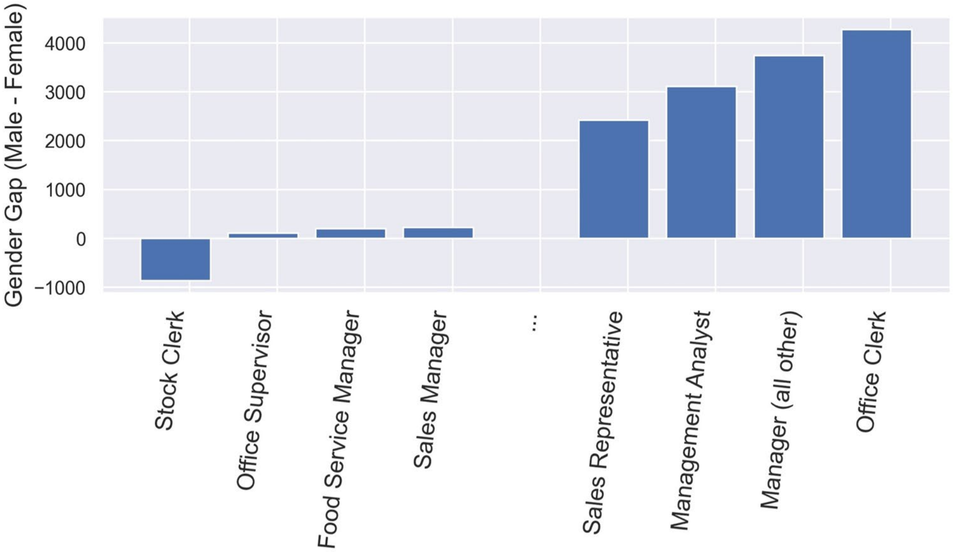
Quartile 4
Most common degree: Bachelor

Adjusted estimator: Calculate gender gap in each decile and average.

Covariate-Adjusted 2nd-Year Gender Gaps

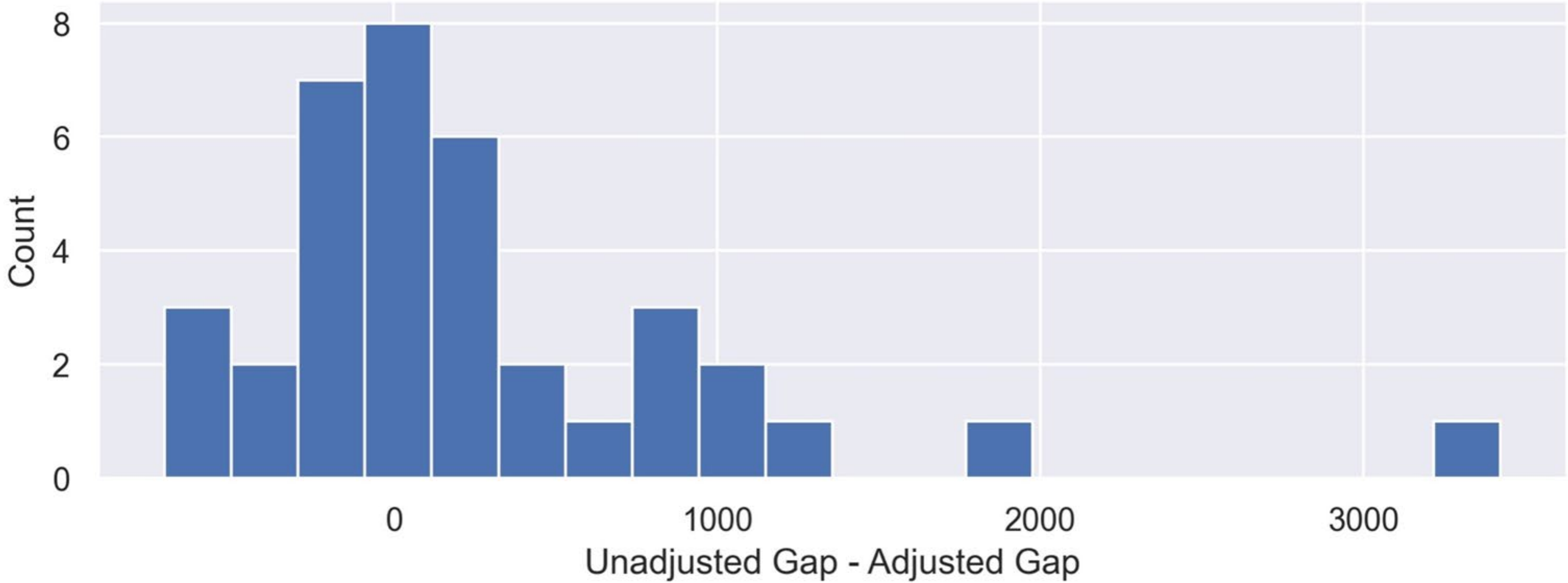
Repeat analysis for various 1st-year jobs.

Interpretation: for men & women with similar observed characteristics and 1st job, trajectories thereafter diverge by gender in a way that matters for wages



Comparing Estimators for 2nd-Year Gender Gap

Unadjusted gap is larger than covariate-adjusted gap 65% of time, avg $277 (relative to gap of $1756):



Decomposing Gender Gaps: 6th Year

Unadjusted gender gap for indiv. who are customer service reps in yr 5: **$1667**.

Adjusted gender gap accounts for differences in covariates and previous jobs:

1. Predict 6th-year wage for all 5th-year customer service reps (based on covariates and previous jobs).
2. Divide the salary predictions into 10 deciles.
3. The gender gap for each decile is the difference between sample means for males and females.
4. The adjusted gender gap is the average over all deciles.

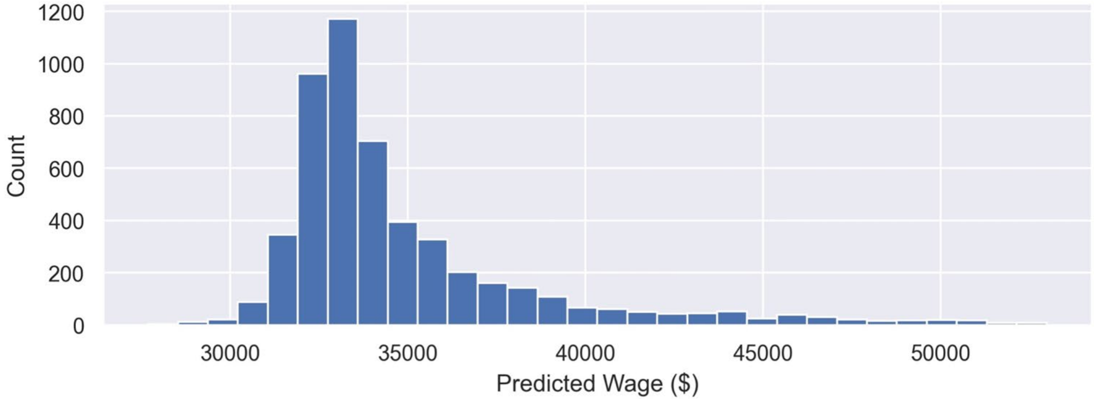
Adjusted gender gap: 5th-year male customer service reps transition to jobs with median annual income **$1258 higher** than females, so 25% of gap accounted for by differences in covariates & previous jobs

Predicted 6th-Year Wages

Predicted "male-wage" for all 5th-year customer service reps:

Quartile 1
Most common degree: High School Diploma
Most common jobs: Cashier, Retail Salesperson, Office Clerk

Quartile 2
Most common degree: Associate
Most common jobs: Cashier, Secretary, Retail Salesperson

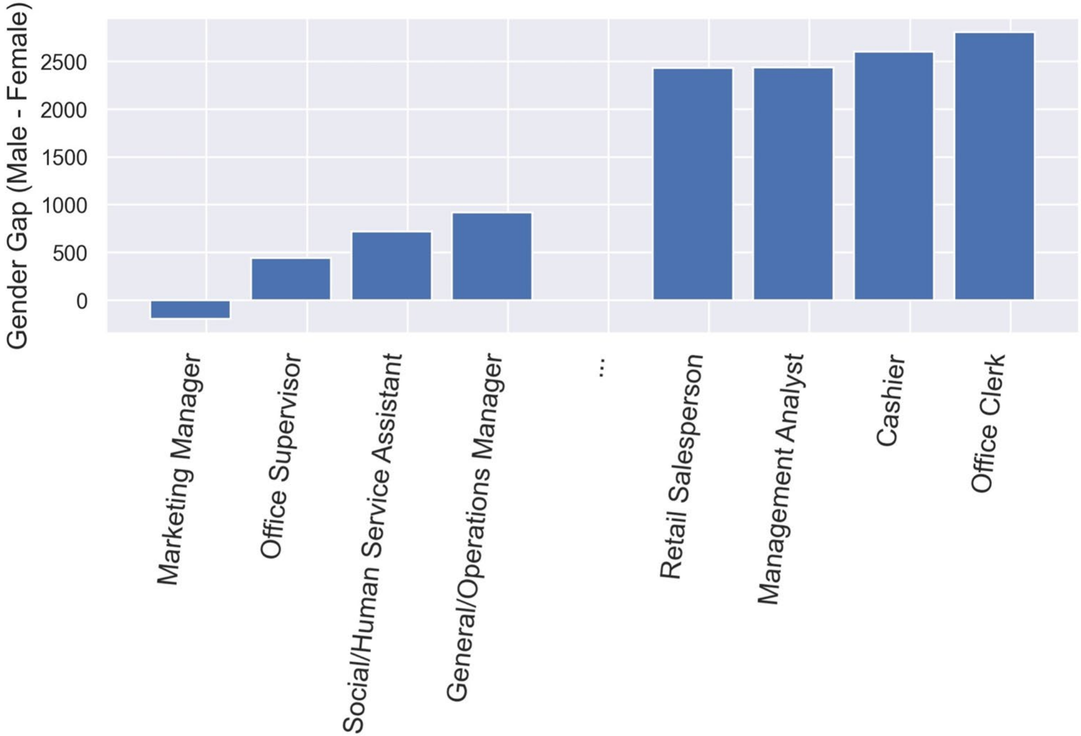


Quartile 3
Most common degree: Bachelor
Most common jobs: Retail Supervisor, Retail Salesperson, Secretary

Quartile 4
Most common degree: Bachelor
Most common jobs: General/Operations Manager, Retail Supervisor, Sales Rep

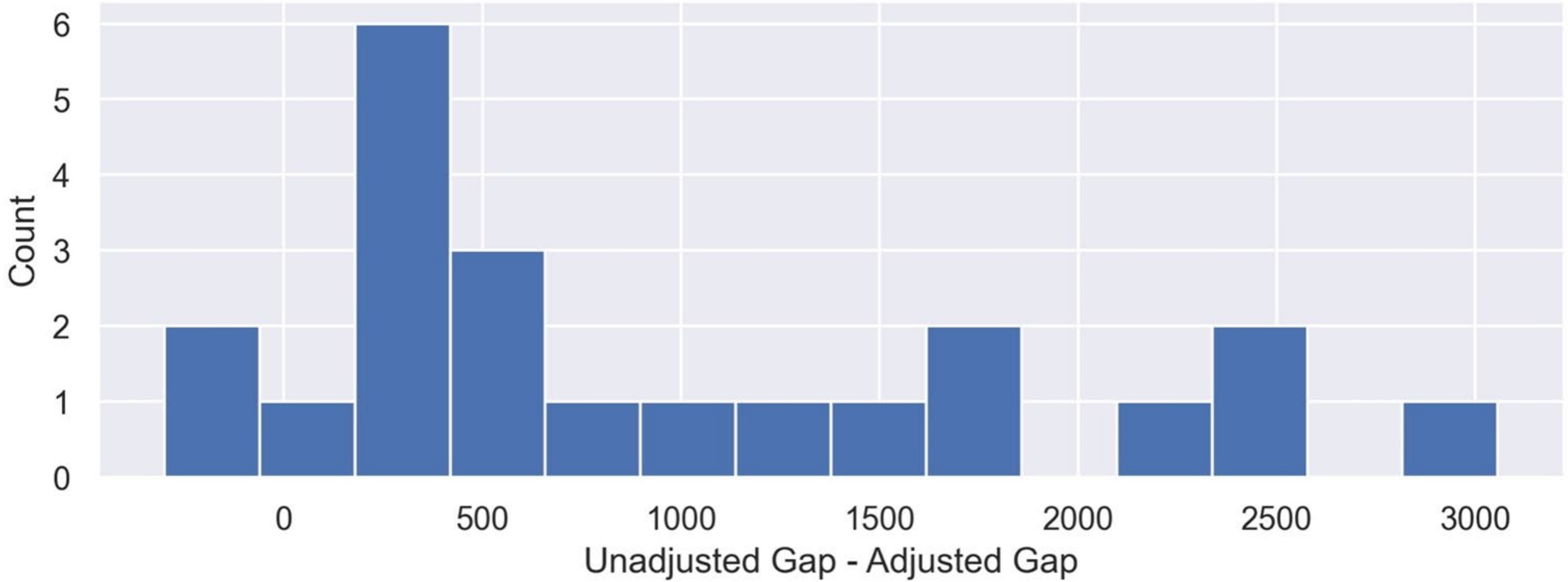
History-Adjusted 6th-Year Gender Gaps

Repeat the analysis for various 5th-year jobs:



Comparing Estimators for 6th-Year Gender Gap

Unadjusted gap is larger than history- and covariate-adjusted gap 91% of time, avg $960 (relative to gap of $2560):



Summary

- Goal: Studying career trajectories (rather than single job transitions) using a dataset of 24 million American worker resumes.
- Transformer adjustments for modeling careers:
  - Two-stage prediction
  - Incorporating covariates
- Model applications:
  - Simulating future wages for a given profile
  - Imputing intermediate jobs
- Gender gap case study:
  - What is the gender wage gap when conditioning on history?
  - Adjust estimator with learned representations to compare similar groups of men and women.
  - Adjusted gender wage gaps are usually smaller than unadjusted gap.

Thank you!

Appendix: Term Definitions

SOC code: The dataset and model use the 7-digit O*NET-SOC Codes from 2010. Some occupation names are shortened due to space constraints (e.g. "First-Line Supervisor of Retail Sales Workers" becomes "Sales Worker Supervisor").

Median wage: Wage information is at SOC-level, from BLS (May 2012). Unavailable median wages were replaced with average wages. SOC codes that did not have any wage information (<2% of SOC codes, none of the 150 most common occupations) were replaced with the national median.

Model architecture: We use the GPT model architecture (Radford et al., 2018): 12 layers, 12 attention heads per-layer, 768-dimensional latent representations (D = 768), and 3072 hidden dimensions for the feed-forward neural networks.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

Appendix: Comparison to Markov Model [15-17, 22-23]

The transformer perplexity results in slide 15-17 and 23 are comparing the two-stage transformer model with all covariates to the first-order Markov model.

In the illustrated examples in slides 16-17 and 22, the transformer probabilities come from a two-stage model trained without covariates -- this is to make clear that the gains in these specific examples are coming strictly from including the job histories.

Appendix: Simulating Wages for a Given Worker Profile [Slides 18-20]

Slide 18: We sample 5,000 trajectories. We keep the three-most frequent second-year jobs, and the three-most frequent third-year jobs conditional on the included two-year trajectories. The width of each arrow is roughly proportional to the frequency of the transition.

Slide 19: We keep the educational degree fixed for each sample for all 5 years. The job sampled for 2010 is always the beginning of the career. We sample 10,000 trajectories for each degree and plot the mean simulated median wage.

Slide 20: We sample 50,000 5-year trajectories. We filter to only include trajectories where the median occupational wage of the 5th year job was at least twice the median occupational wage of the 1st year job. We plot the 8 most frequent of these trajectories (we don't plot trajectories that contain the same jobs as other trajectories that have already been plotted whose only difference is the year in the sequence the transition took place).

Appendix: Calibration Figure [Slide 21]

Population: All individuals in test set who had their first job in 2010, worked for 5 consecutive years, and didn't add any educational degrees over the first 5 years.

Prediction: The model predicts occupational wage by sampling 5-year trajectories for each individual, using only individual covariates (and no occupational information), and pairing with the median occupational wages. The model's prediction is the mean over all 1,000 samples.

Predictions for all 5 years are bucketed into 10 evenly spaced buckets, and the median true wage for each bucket is plotted against the bucket's median predicted wage.The largest bucket for each group is not plotted because it has only a few observations.

The error bars are one standard deviation in each direction. These are large because the model doesn't use observed jobs when simulating trajectories (it only uses covariates), so each bucket contains a variety of job histories.

Appendix: Gender Gap Details

The terms "2nd-year wage" and "6th-year" wage are simplifying: sequences do not always contain one observation per year. For example, the first job for a worker may be in 2010, and they may not work again until 2014. In this case, the 2014 wage is considered the 2nd-year wage. The wage gap we're calculating is precisely the next-job median wage (rather than the "2nd-year median wage").

The most common degrees in slides 29 and 34 do not include examples where the degree is unknown. The extreme right-tails of the histograms are not shown due to space constraints.

Wage gaps are calculated only for jobs that have more than 200 male and 200 female observations in the test set (and does not include examples where the gender is unknown).