From data to decisions: Domain knowledge and the machine learning workforce

NBER Economics of AI Conference, U. Toronto

Prasanna (Sonny) Tambe, Wharton School, U. Pennsylvania

Using job listings to track skill trends in the ML workforce

- IT investment has historically been difficult to measure, especially machine learning (ML) or artificial intelligence (AI) investments
- **Job listings** provide a way to track granular changes to the technical skills required from workers
- I analyze an archive of online job listings which includes 1mm+ listings per month between 2010 to 2016

Strengths and limitations of using job listings as a data source

Limitations:

- Challenging to interpret as a vacancy due to sampling
- Within an ad, skills are listed heuristically
- Job listings can be viewed as being aspirational

Strengths:

- Granular skills information (e.g. TensorFlow, Random Forest)
- Within job title variation in skills, i.e. how skills are bundled
- A great deal of intent, and it represents where the firm is headed (the gradient)

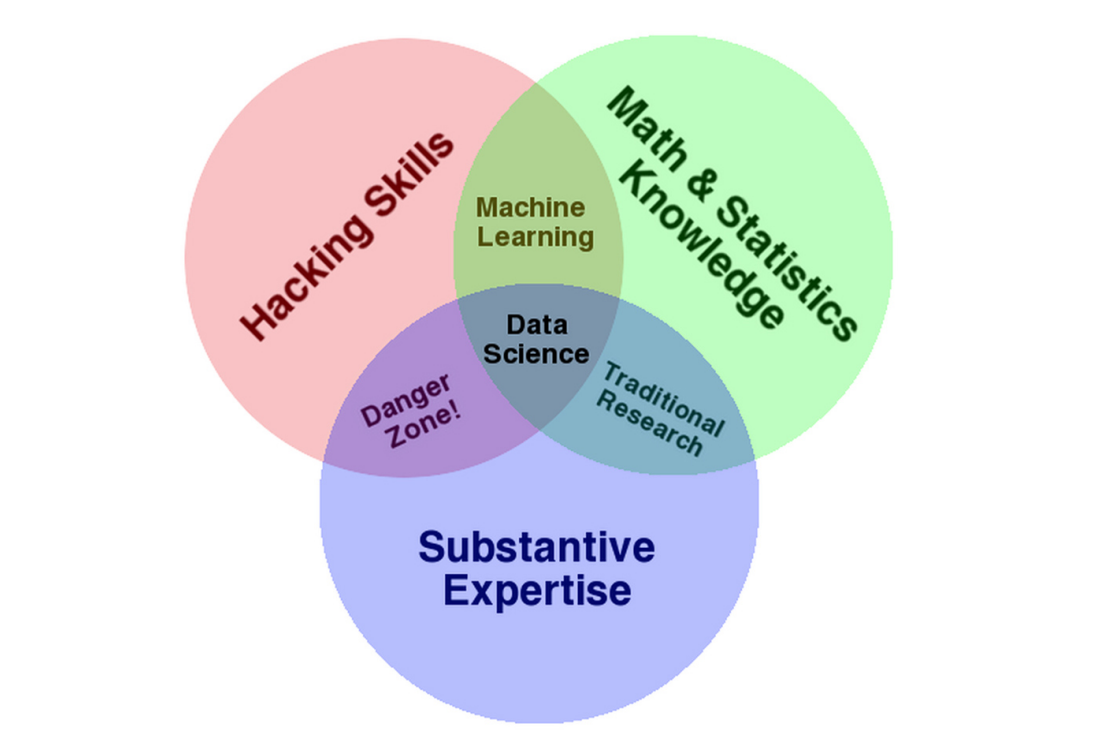
Interpreting the job listing data:

Job listings can be viewed as a large-scale employer survey asking:

> ***What skills would you want from a new hire into a specific job title?***

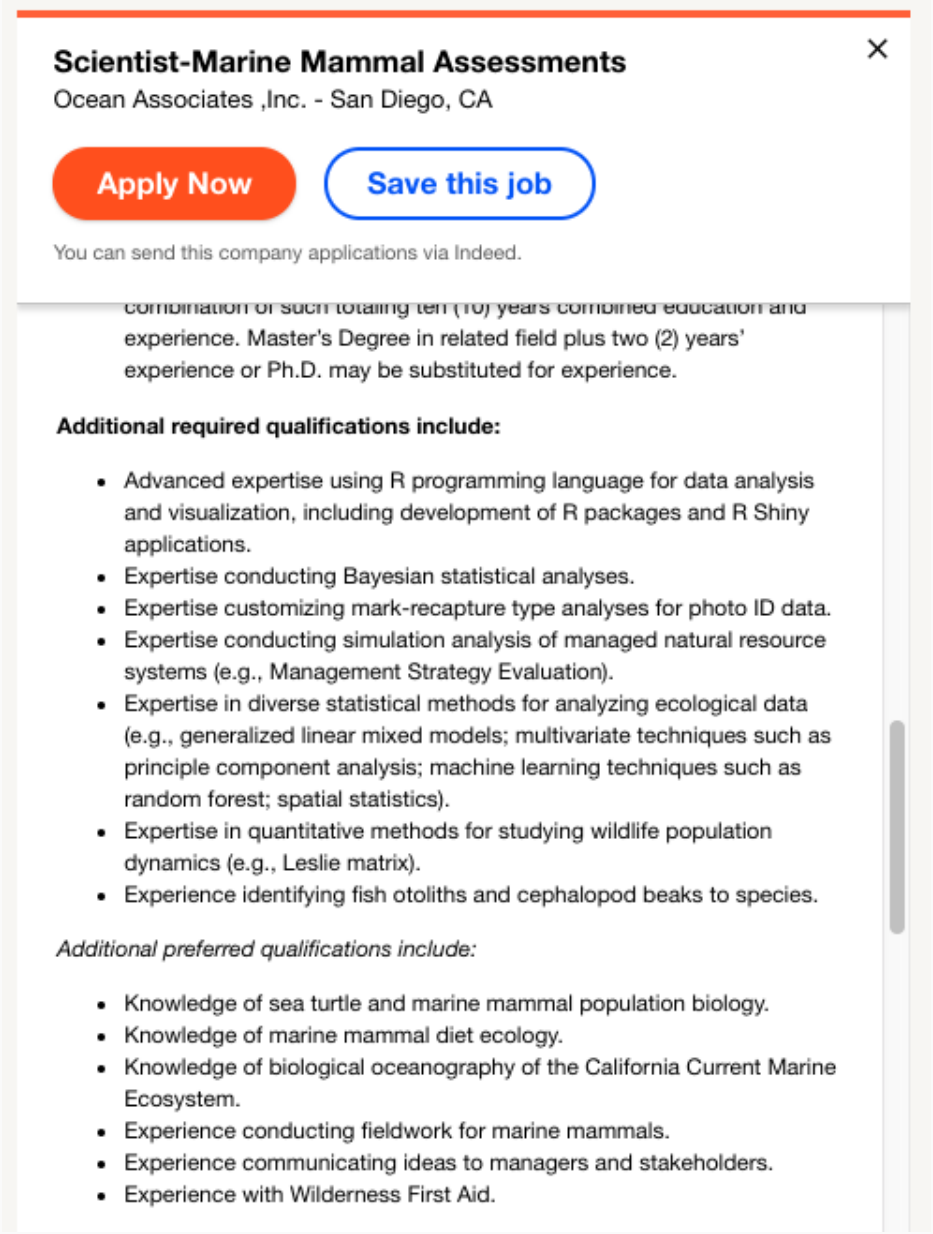
Produces a very large monthly cross-section of responses **where employers bear significant costs for providing inaccurate information**

Focus on domain knowledge and its implications for jobs



- Moving from **Data Management** to **Analysis** to **Prediction** raises the costs of coordinating these activities

Beyond data science: bundling data + domain



Scientist-Marine Mammal Assessments

Ocean Associates ,Inc. - San Diego, CA

Apply Now | Save this job

You can send this company applications via Indeed.

combination of such totaling ten (10) years combined education and experience. Master's Degree in related field plus two (2) years' experience or Ph.D. may be substituted for experience.

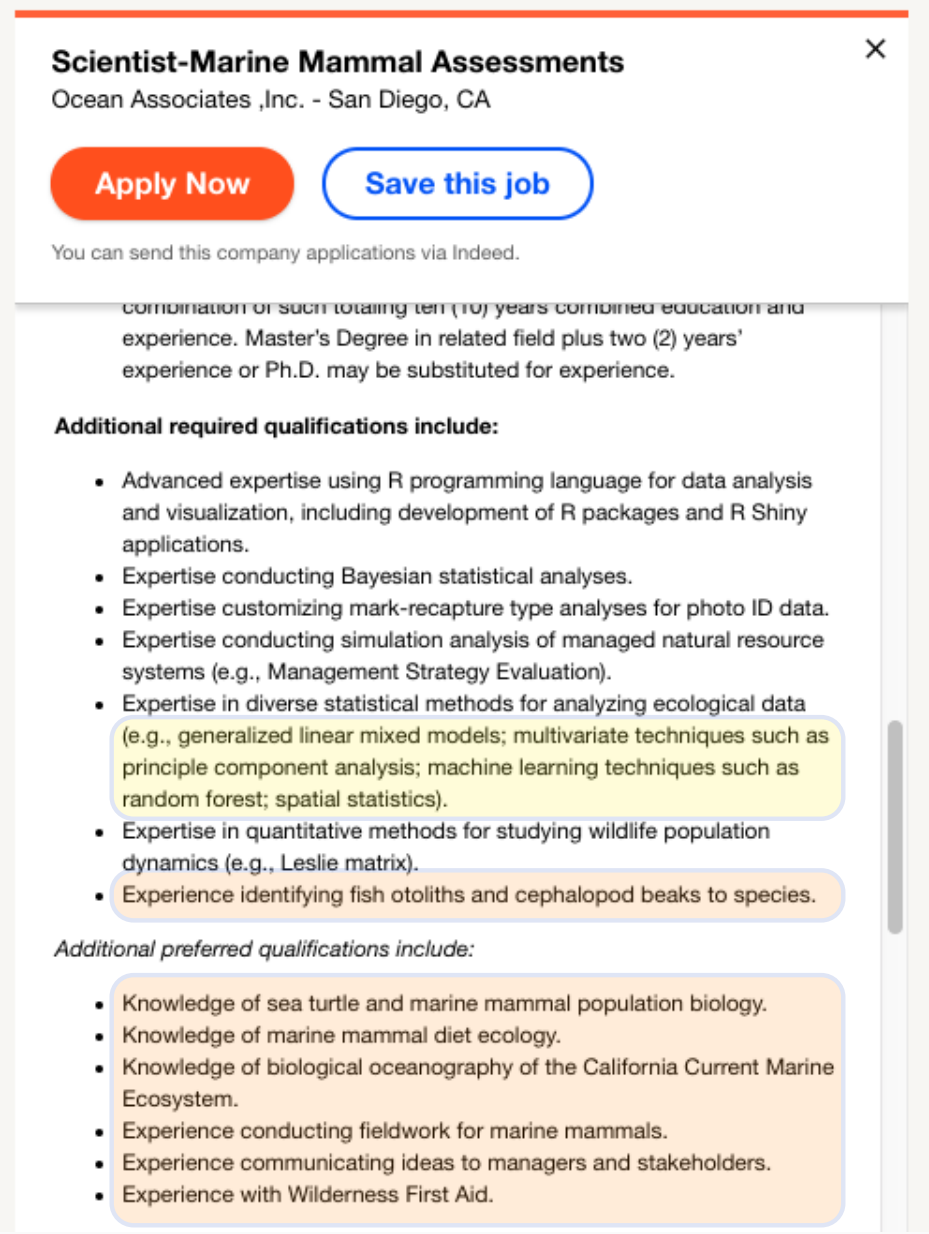
Additional required qualifications include:

- Advanced expertise using R programming language for data analysis and visualization, including development of R packages and R Shiny applications.
- Expertise conducting Bayesian statistical analyses.
- Expertise customizing mark-recapture type analyses for photo ID data.
- Expertise conducting simulation analysis of managed natural resource systems (e.g., Management Strategy Evaluation).
- Expertise in diverse statistical methods for analyzing ecological data (e.g., generalized linear mixed models; multivariate techniques such as principle component analysis; machine learning techniques such as random forest; spatial statistics).
- Expertise in quantitative methods for studying wildlife population dynamics (e.g., Leslie matrix).
- Experience identifying fish otoliths and cephalopod beaks to species.

Additional preferred qualifications include:

- Knowledge of sea turtle and marine mammal population biology.
- Knowledge of marine mammal diet ecology.
- Knowledge of biological oceanography of the California Current Marine Ecosystem.
- Experience conducting fieldwork for marine mammals.
- Experience communicating ideas to managers and stakeholders.
- Experience with Wilderness First Aid.

Beyond data science: bundling data + domain



Scientist-Marine Mammal Assessments

Ocean Associates ,Inc. - San Diego, CA

Apply Now | Save this job

You can send this company applications via Indeed.

combination of such totaling ten (10) years combined education and experience. Master's Degree in related field plus two (2) years' experience or Ph.D. may be substituted for experience.

Additional required qualifications include:

- Advanced expertise using R programming language for data analysis and visualization, including development of R packages and R Shiny applications.
- Expertise conducting Bayesian statistical analyses.
- Expertise customizing mark-recapture type analyses for photo ID data.
- Expertise conducting simulation analysis of managed natural resource systems (e.g., Management Strategy Evaluation).
- Expertise in diverse statistical methods for analyzing ecological data (e.g., generalized linear mixed models; multivariate techniques such as principle component analysis; machine learning techniques such as random forest; spatial statistics).
- Expertise in quantitative methods for studying wildlife population dynamics (e.g., Leslie matrix).
- Experience identifying fish otoliths and cephalopod beaks to species.

Additional preferred qualifications include:

- Knowledge of sea turtle and marine mammal population biology.
- Knowledge of marine mammal diet ecology.
- Knowledge of biological oceanography of the California Current Marine Ecosystem.
- Experience conducting fieldwork for marine mammals.
- Experience communicating ideas to managers and stakeholders.
- Experience with Wilderness First Aid.

ML and job design

- Tests the notion that as we move from data collection $\rightarrow$ analysis $\rightarrow$ prediction: technical skills are increasingly likely to be bundled along with domain knowledge in jobs

- Implications for education and for most aspects of labor market activity (job search, etc.)
- **Subtext**: Why do so many MBAs want to learn Python?

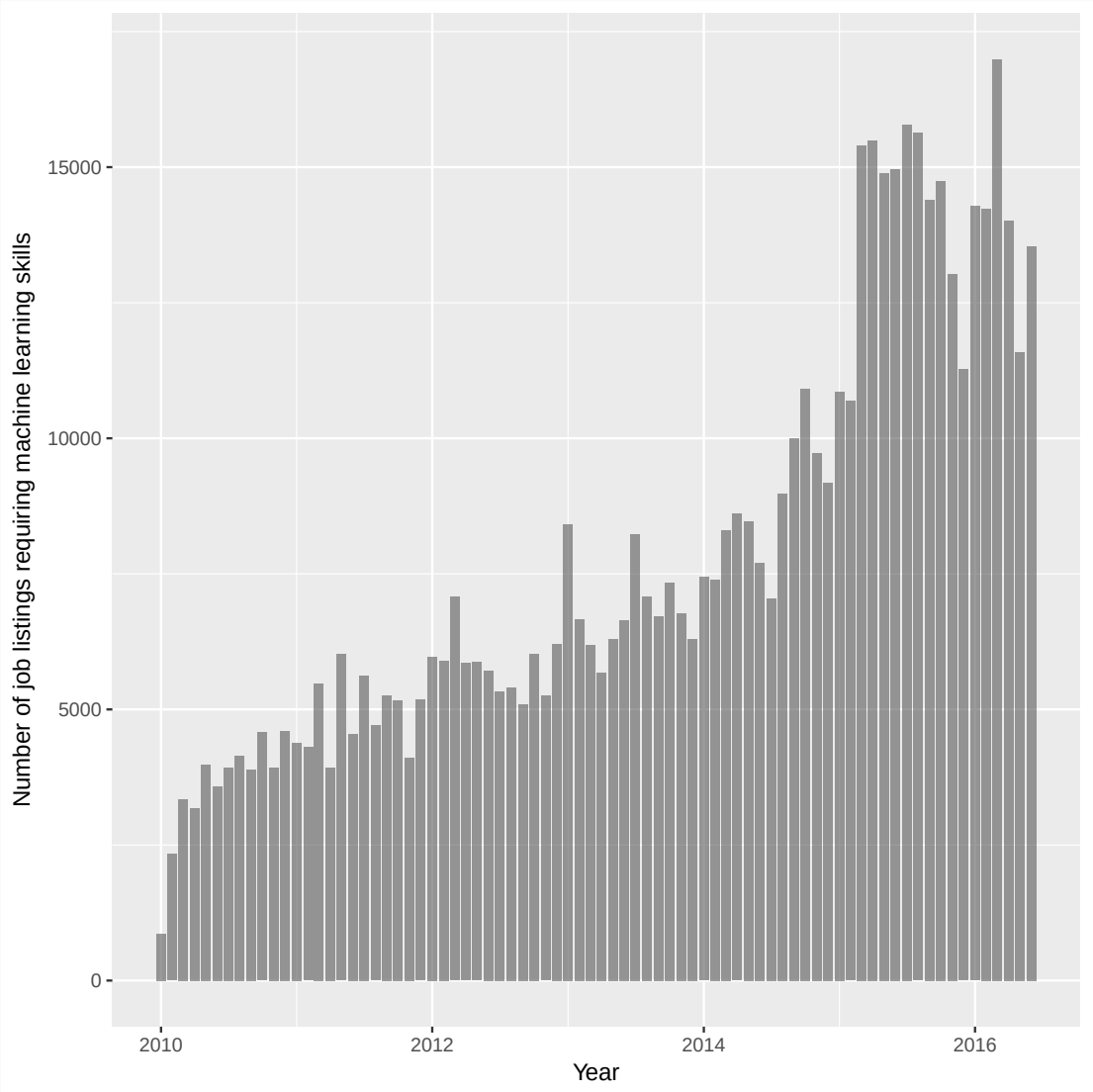
Relevant job skills in the database

- Job skills in database related to Machine Learning

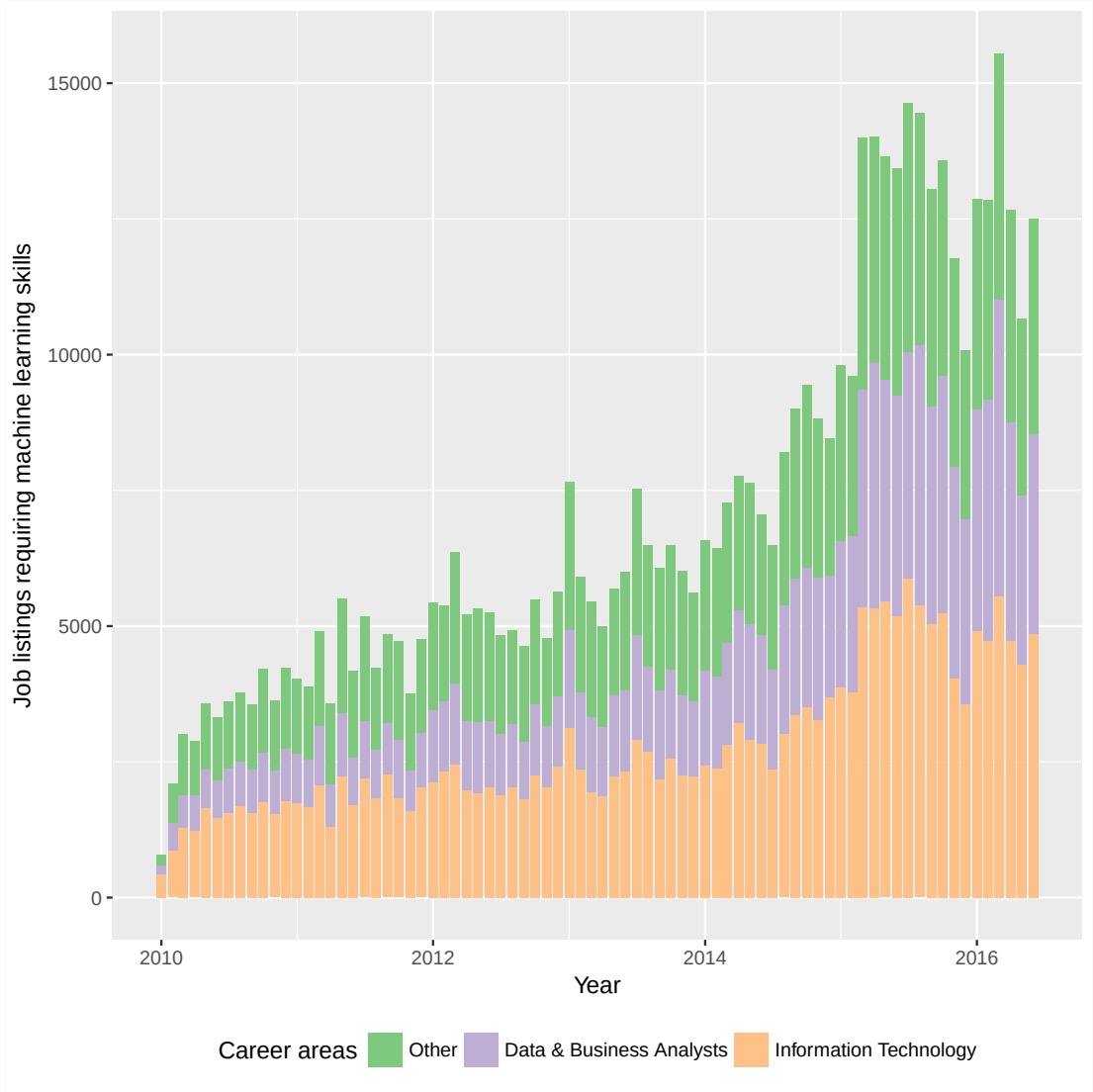
Machine Learning	Artificial Intelligence
Decision Trees	Predictive Analytics
Predictive Models	Data Mining
Deep Learning	Neural Networks
K-Means	Cluster Analysis
Mahout	Random Forests
Language Processing	Support Vector Machines

- Domain knowledge mappings taken from the ONET "Work Knowledge" areas (e.g. biology, political science)

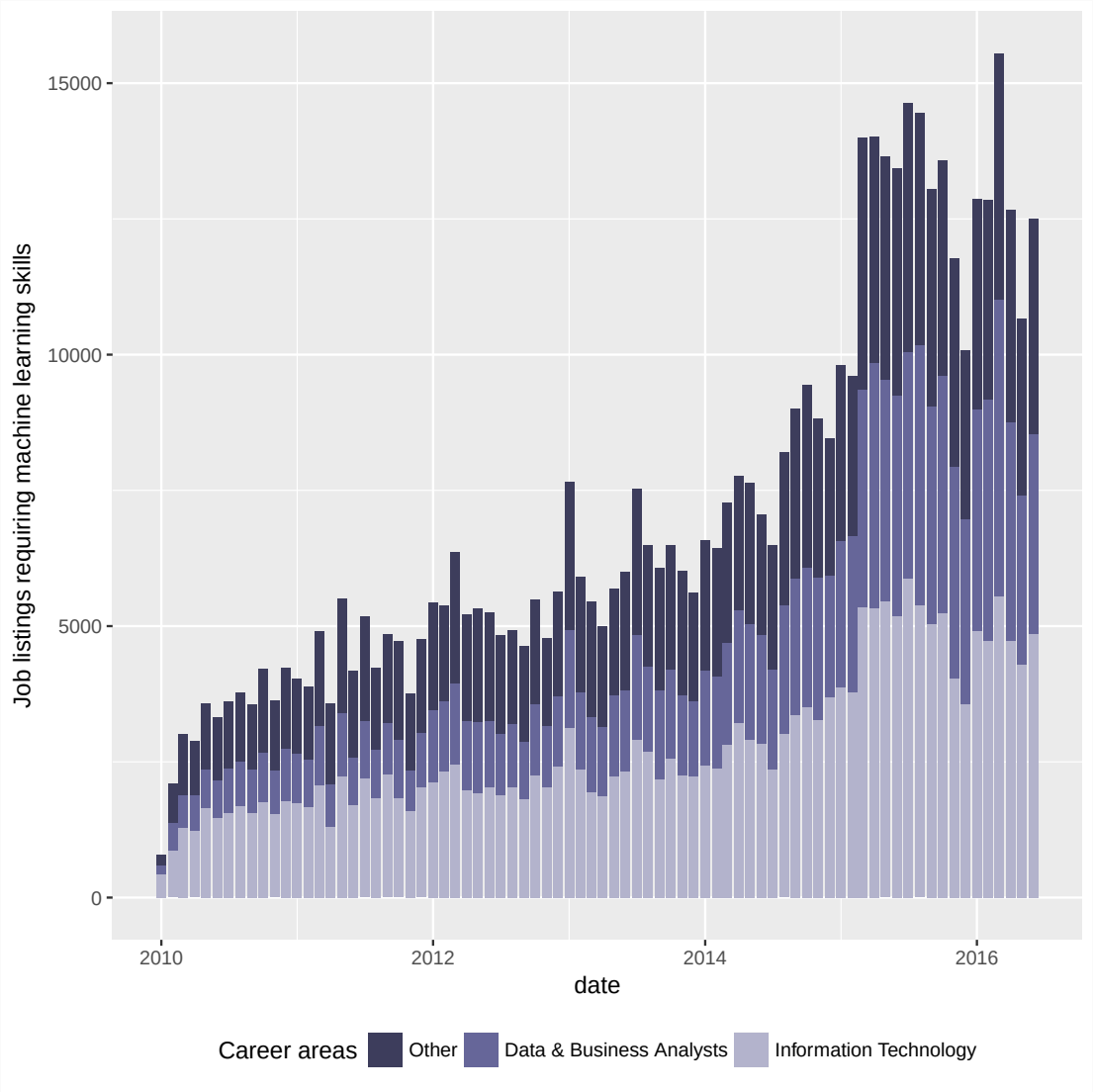
Growth in number of listings in database that include ML skills, 2010-2016 (monthly)



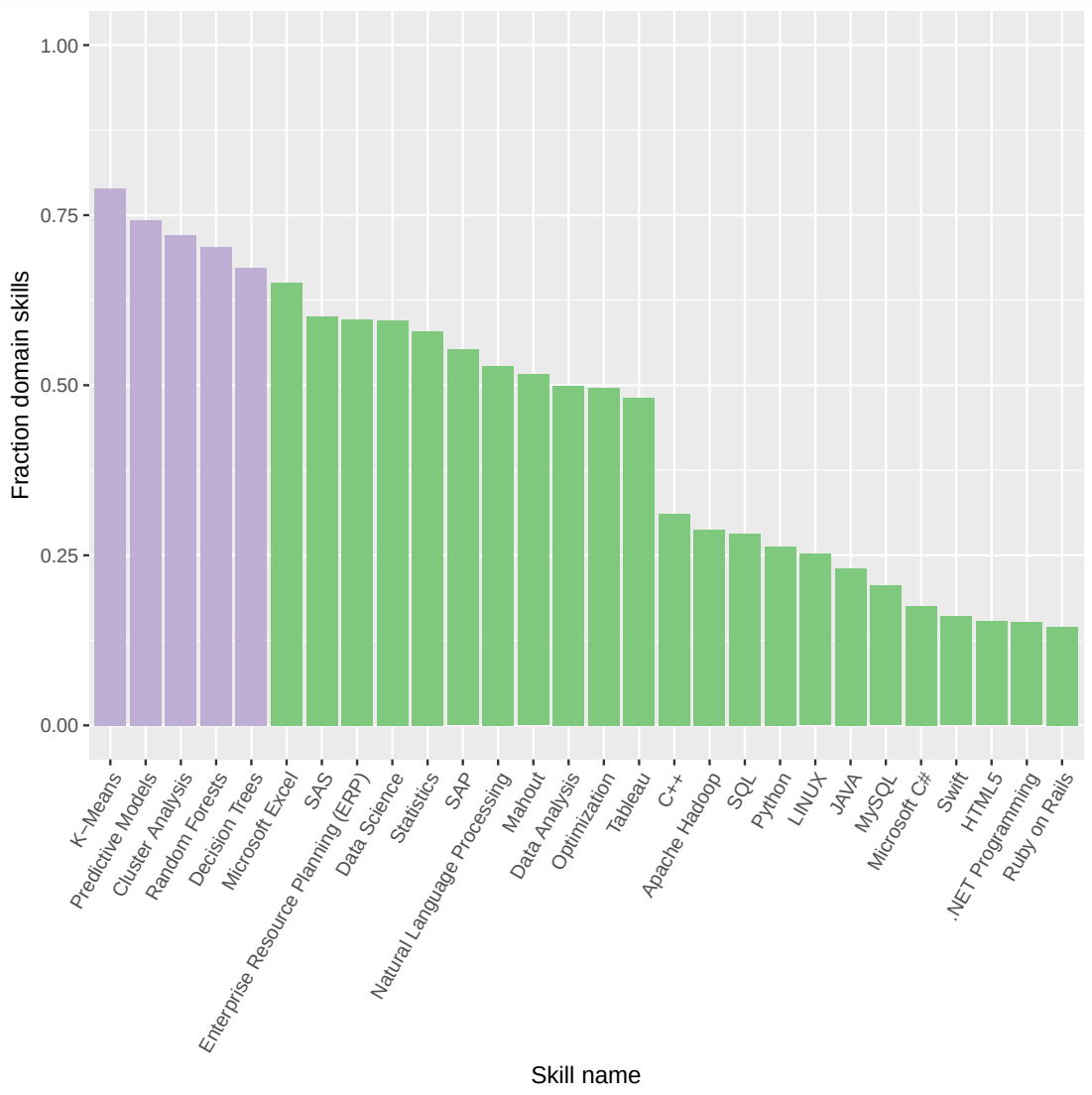
Reflects a combination of listings in IT, data science, and functional areas



Among job listings containing ML skills, greater domain knowledge is required from listings coming from functional areas



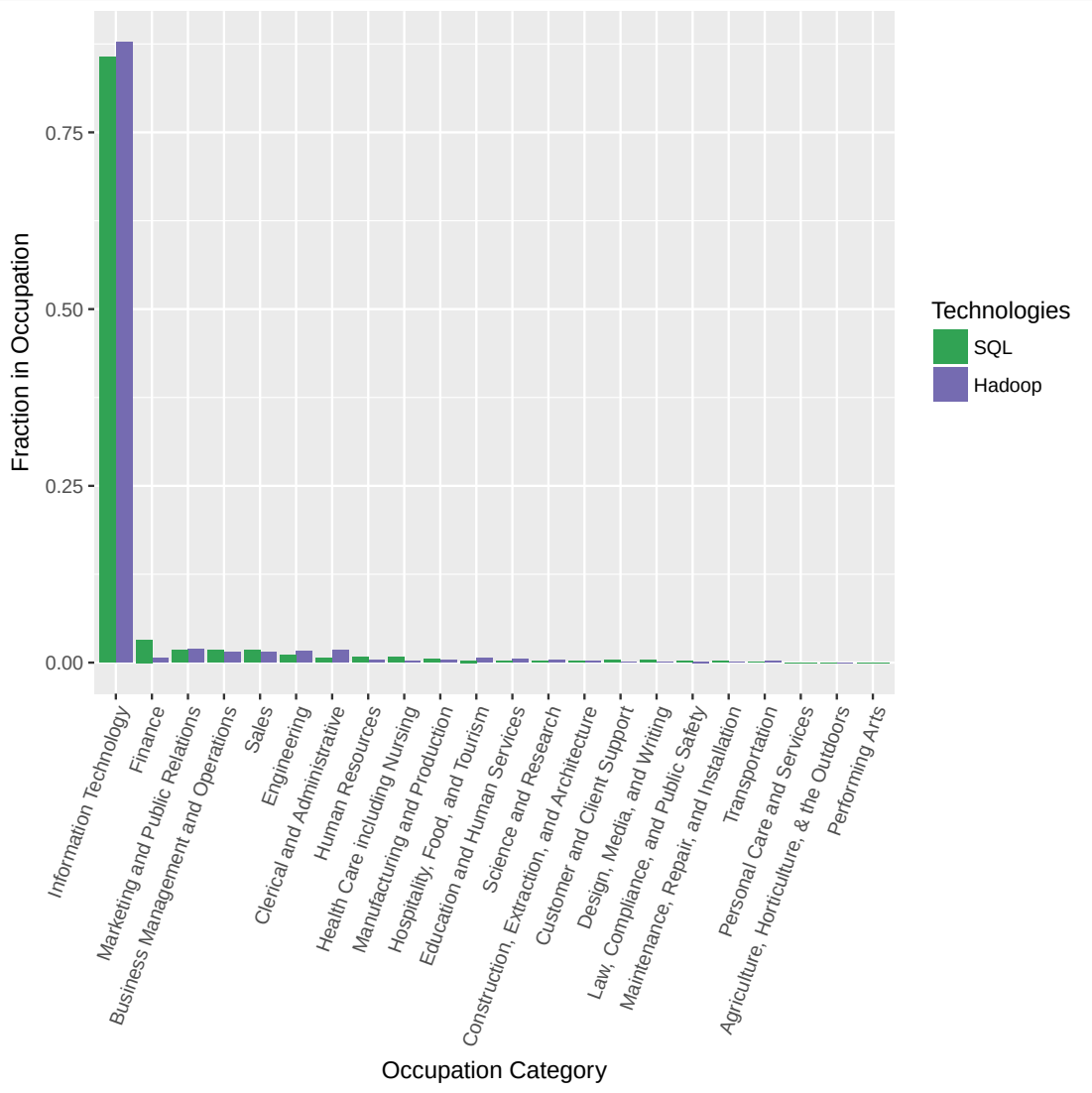
Prediction tools are relatively more frequently bundled with domain knowledge



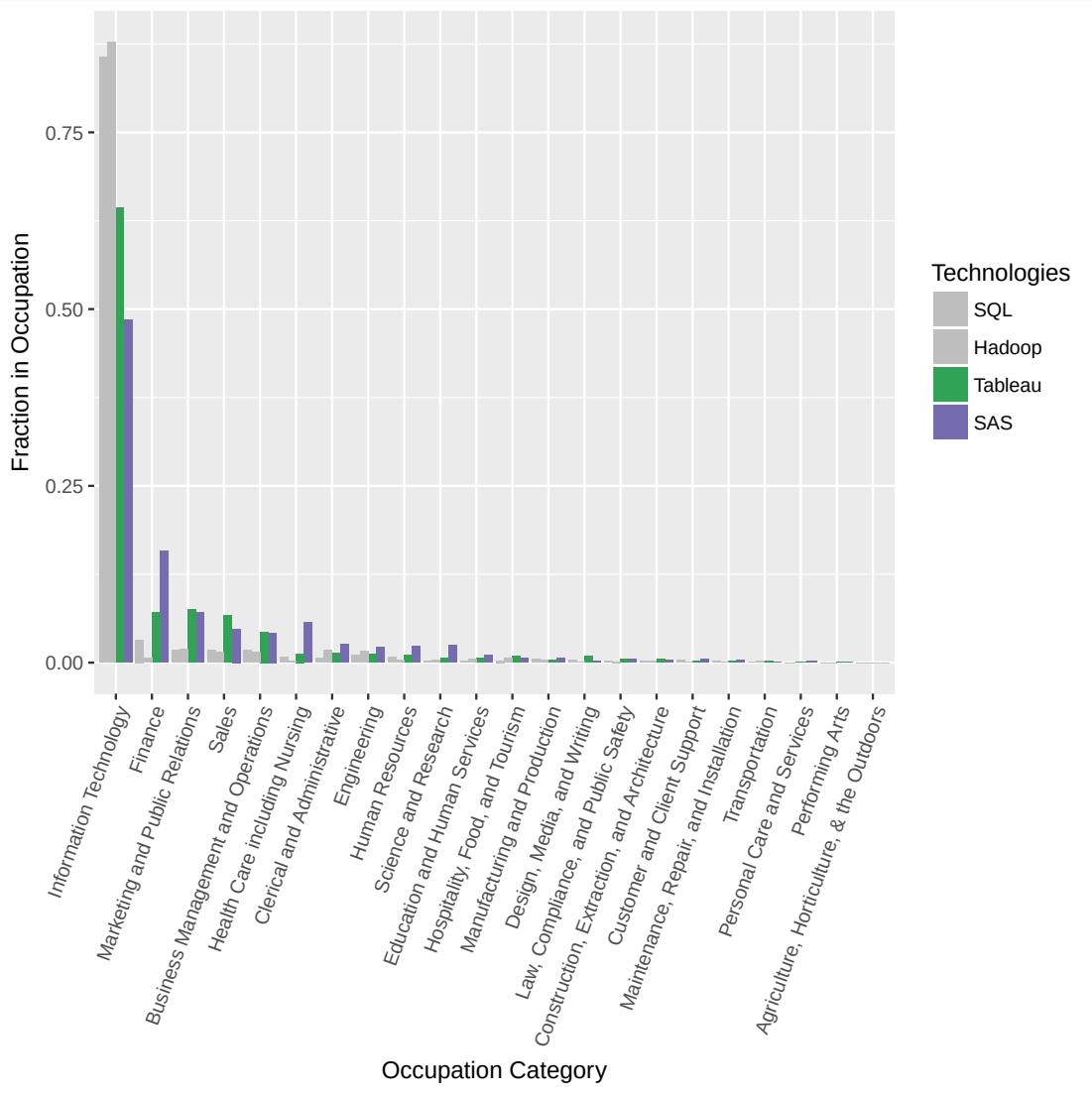
Jobs experiencing the fastest % growth in ML skills

Computer Scientists
Statisticians
Materials Scientists
Financial Quantitative Analysts
Physicists
Business Intelligence Analysts
Biological Scientists
Bioinformatics Scientists
Social Science Researchers
Database Architects
Biostatisticians
Remote Sensing Scientists
Electric Installers
Robotics Engineers
Economists

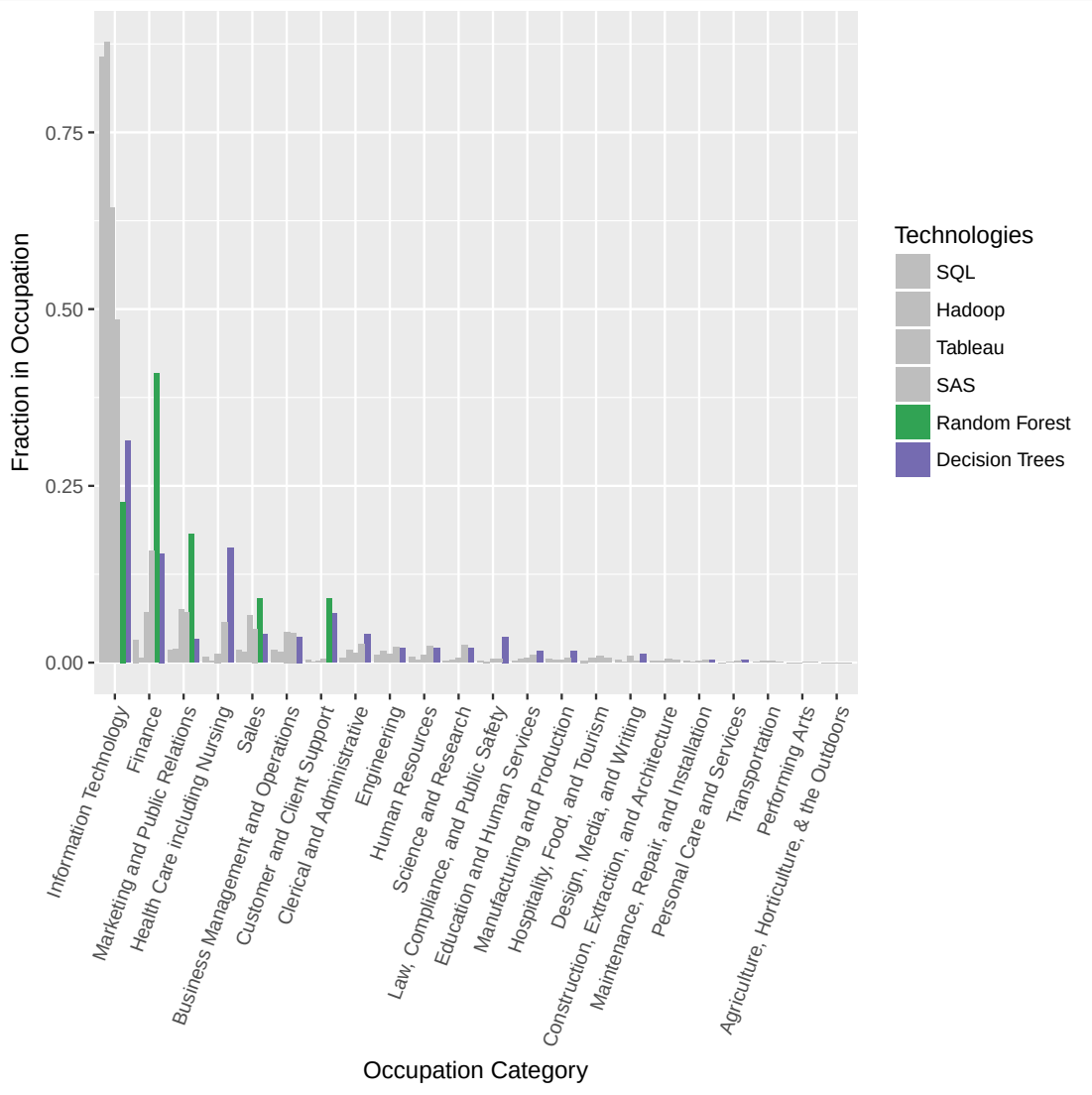
Co-locating data tech with domain skills flattens the distribution across occupations: data collection



Co-locating data tech with domain skills flattens the distribution across occupations: data analysis



Co-locating data tech with domain skills flattens the distribution across occupations: prediction



Regression framework: Within job-title comparisons

Regressions of the form:

$$ML_i = SKILLS_i + title_i + employer_i + industry_i + \epsilon_i$$

- Allows for measurement of correlations between job skills and workers' use of new technologies, conditional on job title
- Mitigates some of the concerns about the sampling frame

Expected correlations with ML and data languages, stats, math, the cloud, and research

	Dependent variable:				
	Data Lang	**Stats**	**Math**	**Cloud/Dist**	**Research**
	(1)	(2)	(3)	(4)	(5)
ML	0.148***	0.069***	0.191***	0.185***	0.105***
	(0.019)	(0.003)	(0.014)	(0.011)	(0.014)
Log(# skills)	0.100***	0.002***	0.074***	0.044***	0.069***
	(0.006)	(0.001)	(0.004)	(0.003)	(0.004)
Observations	12,114	12,114	12,114	12,114	12,114
R^2	0.315	0.060	0.096	0.081	0.082
Adjusted R^2	0.307	0.050	0.087	0.072	0.072

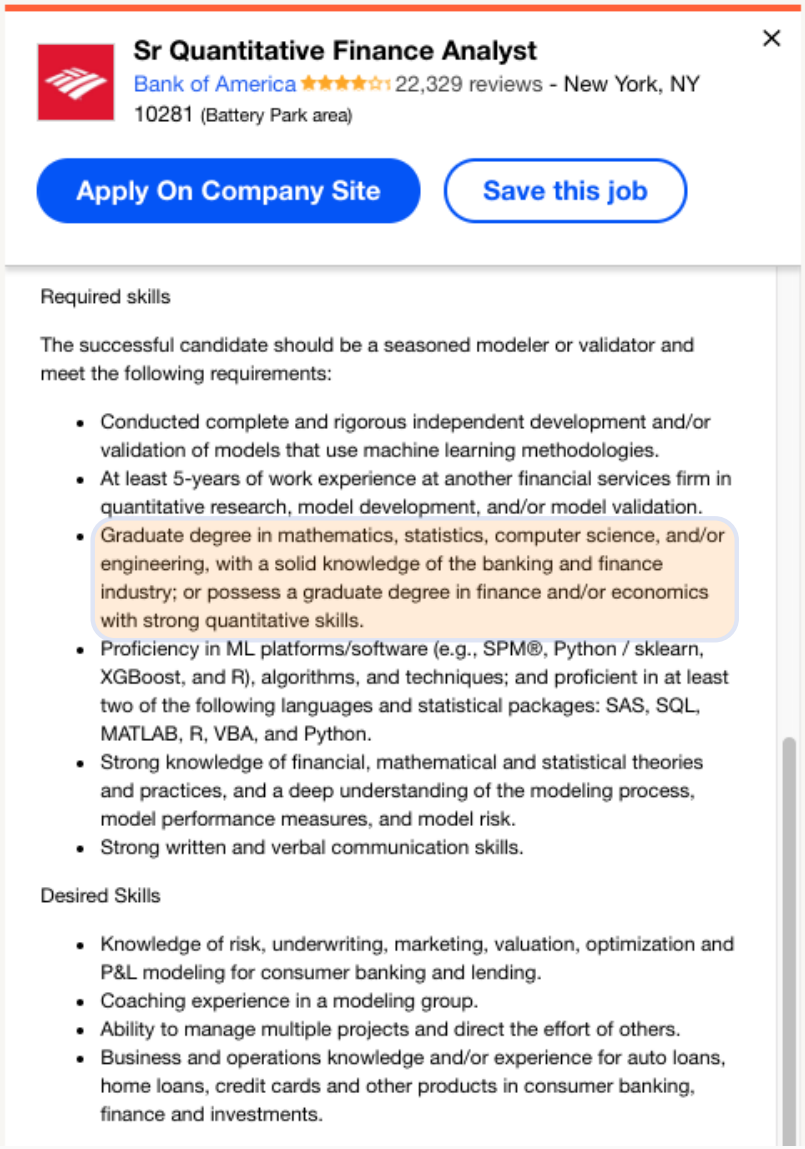
Notes: Conditional correlations of ML skills on other skills listed in job ads. $SKILL_i = ML_i + Log(Num.skills) + jobtitle_i + industry_i + \epsilon_i$. Sample includes listings with ML skills and "matched" listings requiring C++ or Java skills.

Domain knowledge has a stronger relationship with ML than with other data skills

	Dependent variable:				
	SQL	**Hadoop**	**Analytics**	**SAP**	**ML**
	(1)	(2)	(3)	(4)	(5)
Domain	-0.038***	0.011*	0.005	-0.001	0.063***
	(0.011)	(0.006)	(0.004)	(0.003)	(0.004)
Log(# skills)	0.290***	0.045***	0.019***	0.017***	0.00005
	(0.006)	(0.003)	(0.002)	(0.002)	(0.002)
Observations	12,026	12,026	12,026	12,026	12,026
R^2	0.257	0.057	0.123	0.059	0.065

Notes: Conditional correlations of ML skills on other skills listed in job ads. $ML_i = SKILL_i + Log(Num.skills) + jobtitle_i + industry_i + \epsilon_i$. Sample includes listings with ML skills and "matched" listings requiring C++ or Java skills.

A correlate: Listings recommend a CS/IS/tech education or domain-relevant college majors (e.g. biology, economics)



Sr Quantitative Finance Analyst

Bank of America ★★★★☆ 22,329 reviews - New York, NY 10281 (Battery Park area)

Apply On Company Site | Save this job

Required skills

The successful candidate should be a seasoned modeler or validator and meet the following requirements:

- Conducted complete and rigorous independent development and/or validation of models that use machine learning methodologies.
- At least 5-years of work experience at another financial services firm in quantitative research, model development, and/or model validation.
- Graduate degree in mathematics, statistics, computer science, and/or engineering, with a solid knowledge of the banking and finance industry; or possess a graduate degree in finance and/or economics with strong quantitative skills.
- Proficiency in ML platforms/software (e.g., SPM®, Python / sklearn, XGBoost, and R), algorithms, and techniques; and proficient in at least two of the following languages and statistical packages: SAS, SQL, MATLAB, R, VBA, and Python.
- Strong knowledge of financial, mathematical and statistical theories and practices, and a deep understanding of the modeling process, model performance measures, and model risk.
- Strong written and verbal communication skills.

Desired Skills

- Knowledge of risk, underwriting, marketing, valuation, optimization and P&L modeling for consumer banking and lending.
- Coaching experience in a modeling group.
- Ability to manage multiple projects and direct the effort of others.
- Business and operations knowledge and/or experience for auto loans, home loans, credit cards and other products in consumer banking, finance and investments.

These trends are echoed in the larger job listings data

	Dependent variable:				
	CS/IS	**Science**	**Business**	**Economics**	**Either**
	(1)	(2)	(3)	(4)	(5)
ML	-0.109^{***}	0.022^{***}	0.064^{***}	0.018^{***}	0.104^{***}
	(0.012)	(0.004)	(0.009)	(0.003)	(0.016)
Log(# skills)	-0.001	-0.00001	0.004	-0.004^{***}	0.017^{**}
	(0.005)	(0.002)	(0.004)	(0.001)	(0.007)
Observations	4,101	4,101	4,101	4,101	4,101
R^2	0.432	0.134	0.533	0.047	0.095

Notes: Conditional correlations of machine learning skills on degrees listed in job ad. $Degree_i = ML_i + jobtitle_i + industry_i + \epsilon_i$. Sample includes listings with machine learning skills and "matched" workers requiring C++ or Java skills.

Key points

- Suggests that hybrid workers may have a particularly important role to play in the integration of ML tools into business domains
- Flattens the distribution across occupations, continues a trend towards the digitization of new job categories
- Potentially broadens the locus of technical investment within organizations

Comments are very welcome: tambe@wharton.upenn.edu