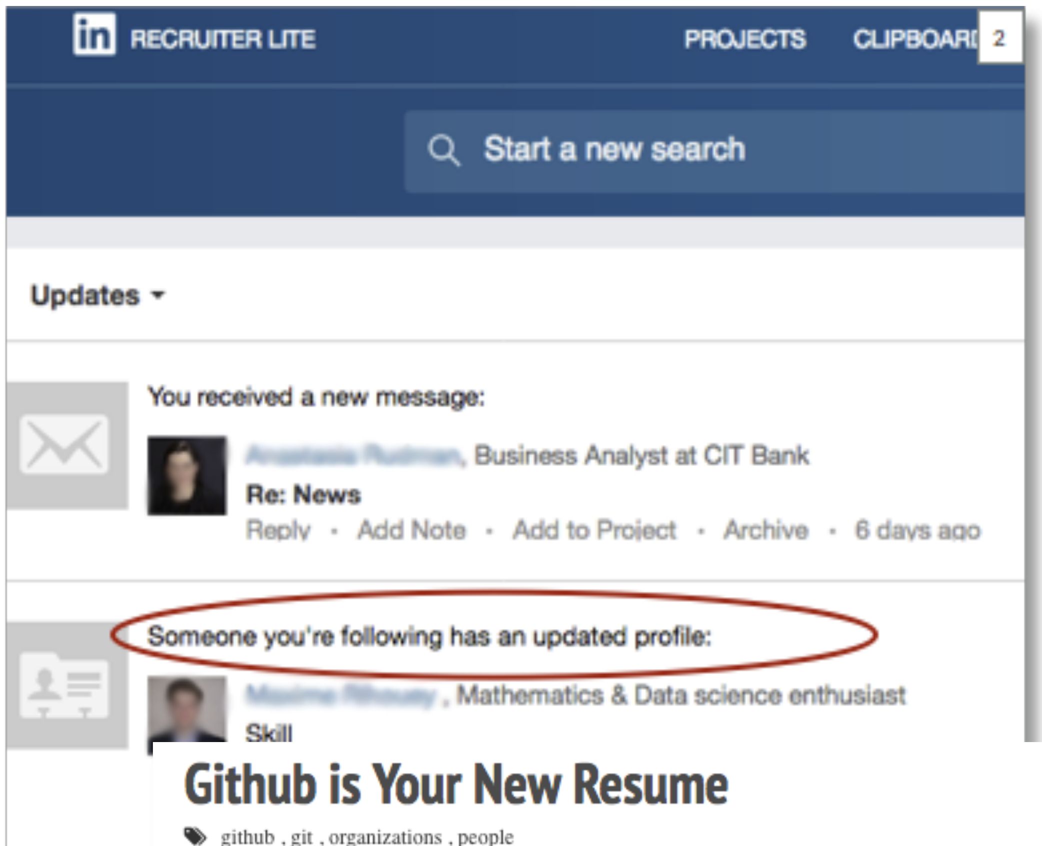
Opera Signal

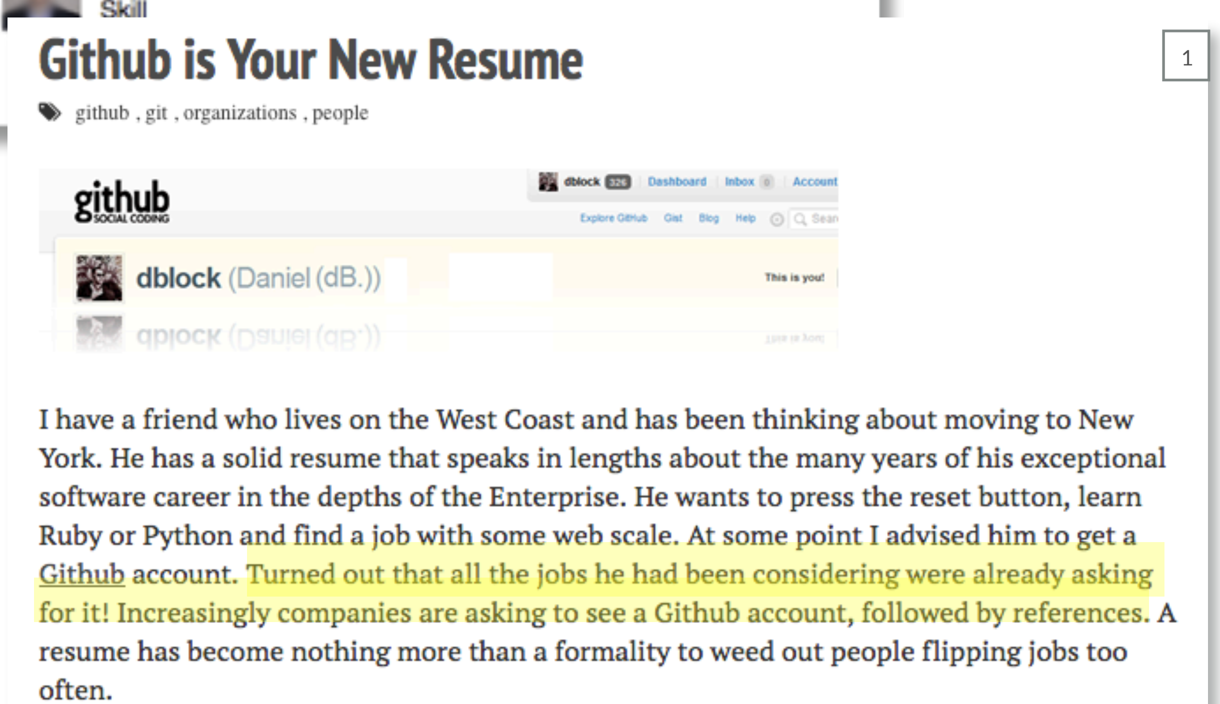
Monitoring tech workers' online activity to predict likely job switchers

Toavina Andriamanerasoa

TECHNOLOGY HAS ALTERED WORKERS' BEHAVIOUR

Workers leave activity traces online which act as useful signals for recruiters




1

Github is Your New Resume

github , git , organizations , people

I have a friend who lives on the West Coast and has been thinking about moving to New York. He has a solid resume that speaks in lengths about the many years of his exceptional software career in the depths of the Enterprise. He wants to press the reset button, learn Ruby or Python and find a job with some web scale. At some point I advised him to get a Github account. Turned out that all the jobs he had been considering were already asking for it! Increasingly companies are asking to see a Github account, followed by references. A resume has become nothing more than a formality to weed out people flipping jobs too often.

- Companies routinely use Github to vet hires
- Coding bootcamps encourage participants to use Github as their portfolio
- Job seekers use Stack Overflow, LN or Hacker News to advertise they are looking for jobs

Sources: (1) http://code.dblock.org/2011/07/14/github-is-your-new-resume.html, (2) LN Recruiter Tool

CREATING A SIGNIFICANT OPPORTUNITY FOR RECRUITERS

By mining employee information, recruiters could increase their success rate

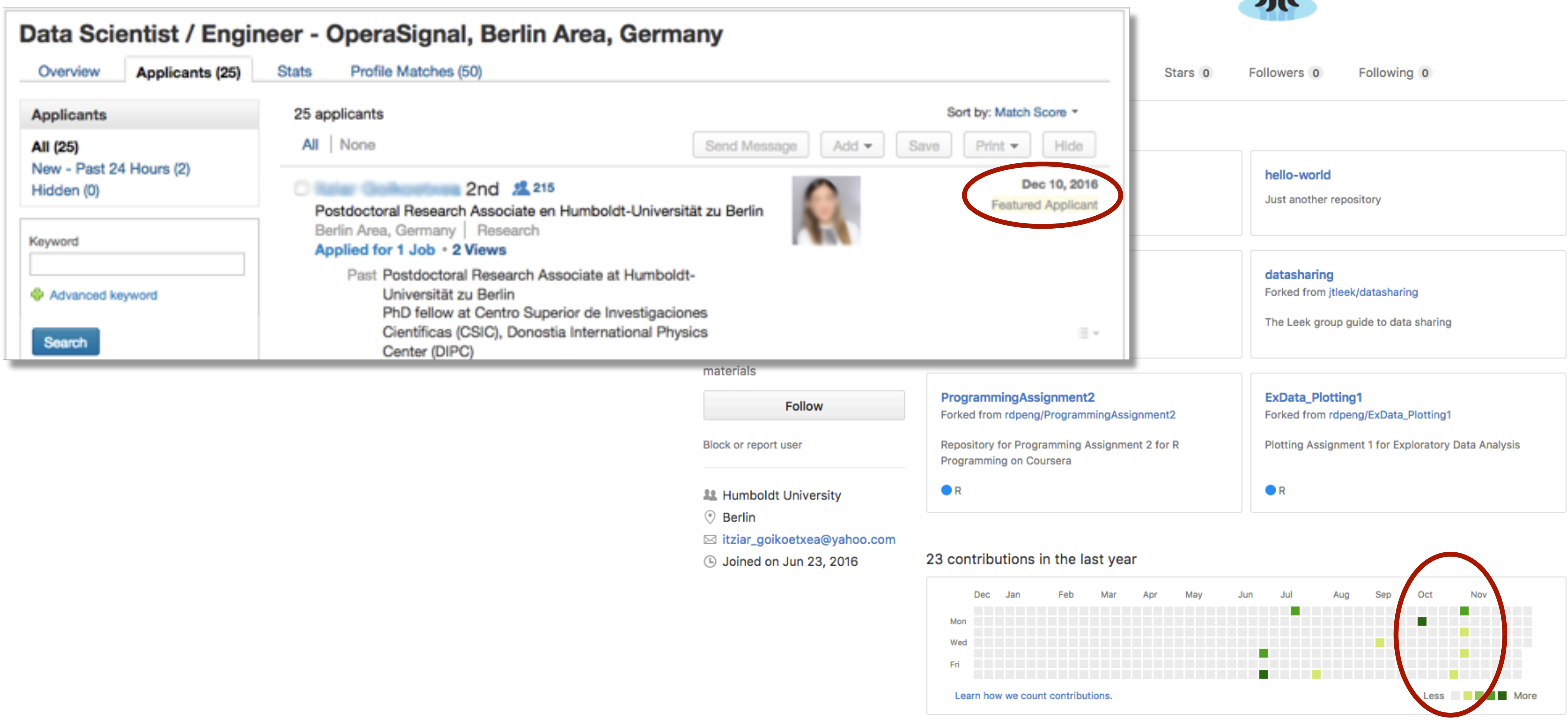
- **Identifying candidates in the market early** is highly valuable edge
- This is especially true for SMEs trying to lure top talent
  - **For these roles recruiters may find only 20% of people they call without preselection are interested** [2]
- By identifying candidates ready to switch jobs, recruiters **save time, improve their call success rate and time to closing**
- **Companies can also use that same information for recruitment and retention**

Sources: (1) ,officevibe.com/blog/12-recruiting-stats, (2) Personal conversation with recruiters

BUT COULD I FIND ANY USEFUL SIGNALS?

By creating a job ad for a fictional startup and browsing Github profiles, I was able to find anecdotal evidence that Github activity could be linked to job-seeking

GitHub



MANY CHALLENGES AROSE IN THE INITIAL STAGE

I found workable solutions to initial issues by narrowing the scope and doing some user research

	Challenges	Solutions
1	Which sources?	• Spoke to very active OSS developer to uncover job-seeking behaviour • Picked LinkedIn and Github given popularity initial validation
2	How to deal with data size and numbers?	• Narrowed GitHub users by limiting scope to named active users in the West (c.2m users) • Narrowed scope further: Users posting on Hacker News (HN): Who Wants to be Hired? (3000+ users) **then filtering for LinkedIn and Github accounts (800 users)**
3	How to get the ground truth? • People don't advertise they are looking to switch jobs	• **Used HN posts and LN job changes as proxies for signalling interest in new jobs**

90% OF MY TIME WAS SPENT ACQUIRING, CLEANING MERGING AND TRANSFORMING DATA…

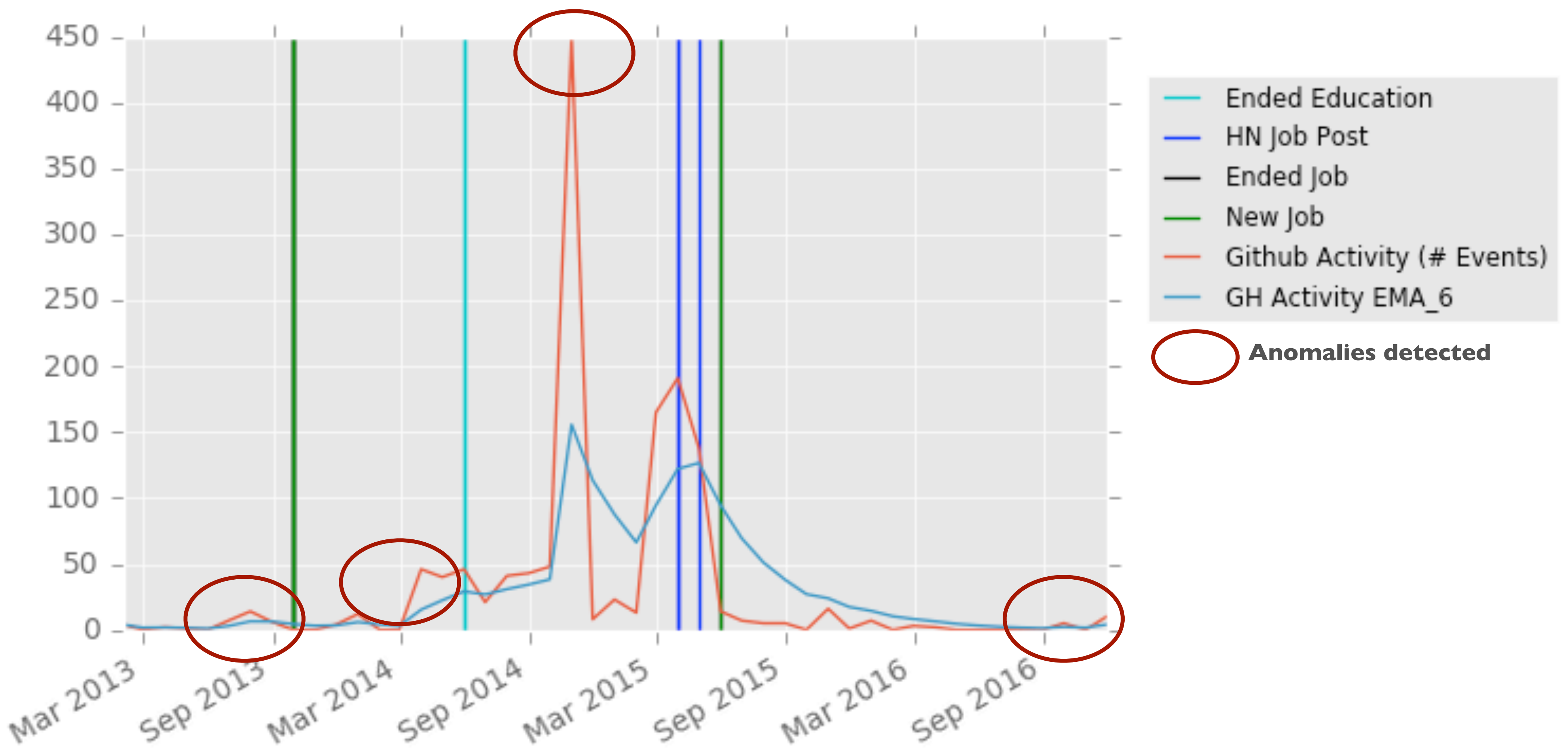
Acquiring, cleaning and merging the data was challenging and very time-consuming due to data size and variety of tools and hacks required

	GitHub events on Google BigQuery	GitHub REST API	Hacker News	LN User profiles
Data	• GitHub events (creating repos, push requests…) by user, date	• Personal user data required to identify users on other resources	• Posts on HTML pages from users looking for jobs	• **Full** user profiles (public profiles insufficient)
Challenges	• 4TB compressed data, impossible to deal with in memory (Dask too slow)	• 5,000 requests per hour rate limit	• Messy HTML, unstructured data	• Very hard to scrape private profiles • Often multiple matches
Tools / Acquisition Process	• Google BigQuery (SQL-like) to get results in seconds • Joblib to parallelize local operations	• 8 remote Google Compute machines to beat rate limit	• BeautifulSoup to parse HTML	• Found a workaround by exporting PDF CVs • Used cross-references from other databases to find right profile
Munging	• Regex to filter countries and narrow user scope	• None	• Extensive Regex to clean email addresses, links that people try to hide from spammers	• Parsed pdf into text with pdfminer • Regex and line parsing to find structure

TO GENERATE TIME-SERIES AND STATIC DATA TO PREDICT WHO WILL BE JOB-HUNTING IN THE NEXT 3 MONTHS

When observing the data it became apparent that some signals preceded job moves or HN Posts by a few months, as expected

Sample User Timeline



- Besides the information in the chart above, I extracted a lot of data about users' education, jobs, bootcamp attendance, highest degree achieved…
- I used an exponential weighted moving average to detect unusually high Github activity[1]

Note: (1) No focus on lower than usual activity

THE MODELLING PROCESS HELPED PERFORMANCE

Reducing dimensionality, resampling, selecting the appropriate model led to significant improvements in Area under ROC curve

PCA & Scaling

- From 175 features to 25 components

Improved AUC by c. 40 basis points(1)

Resampling

- Imbalanced dataset (c. 74% in class 0)
- Combined over-sampling minority class and under-sampling other class

Improved AUC by c. 156 basis points(1)

Fitting and Predicting

- Tried many models
- Extremely Randomised Trees best performing model by far
- Through tweaking and GridSearch, selected optimal parameters for the main model

Improved AUC by c. 286 basis points(1)

Opera Signal

Note: (1) vs. Scikitlearn's standard Extremely Randomised Trees model with default parameters

THE IMPORTANT FEATURES MAKE SENSE

Reducing dimensionality, resampling, selecting the appropriate model led to significant improvements in Area under ROC curve

- Among the key features for the tree classification model are:
  - Moving average of GitHub activity and their difference with actual value
  - Number of GitHub events in that particular month
  - Number of public repos, GitHub Followers
  - Number of public Gists

THE RESULTS ARE ENCOURAGING

It is important to note that in this case, minimising false positives is not a life or death situation - false negatives could be caused by "missing ground truth" in some observations

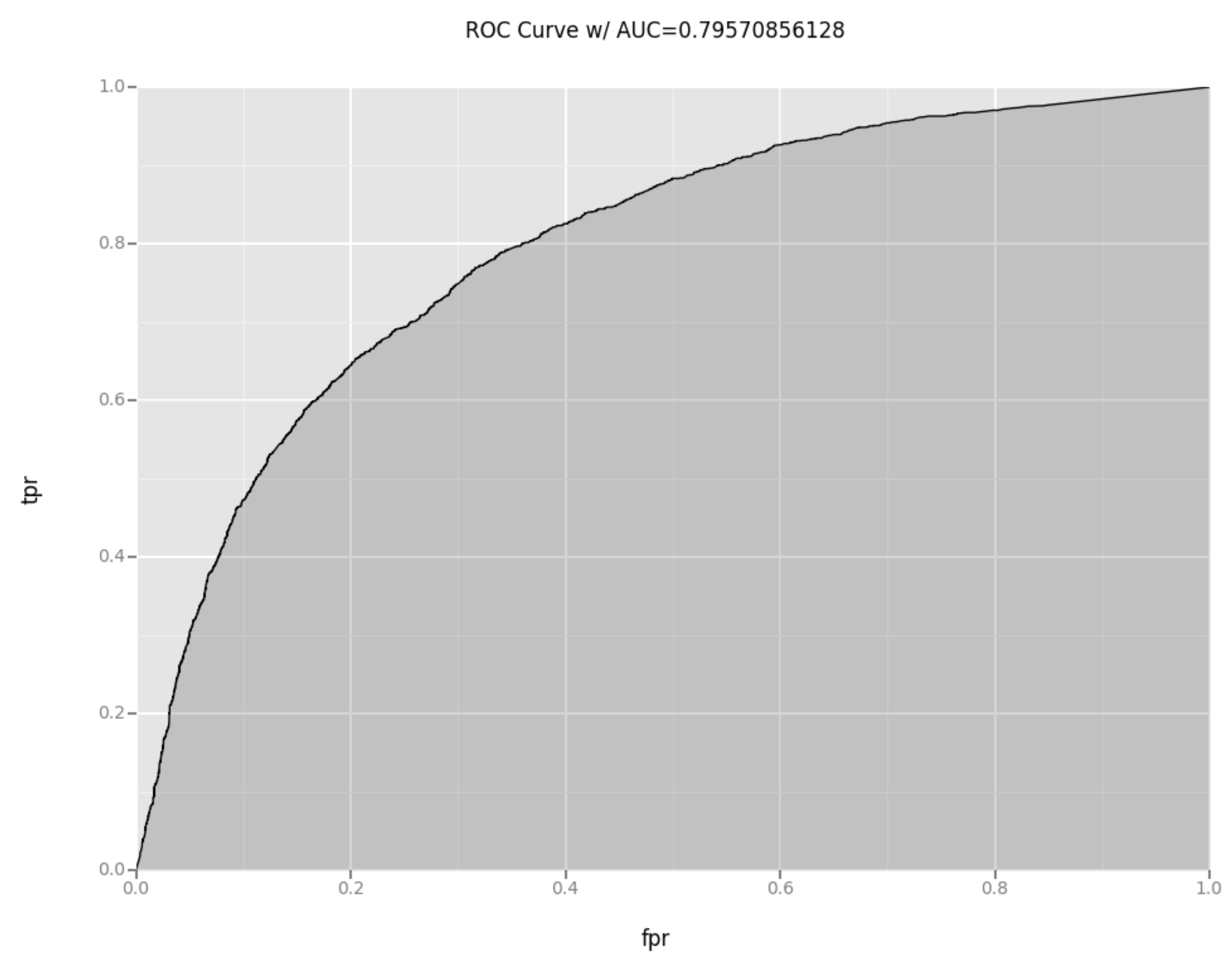
Key Stats

- Test set: 30% of observations
- % of observations in class 0 in data (training and test): 73%
- Accuracy in test set: c.78%
- Area under ROC Curve: c.0.796

Confusion Matrix

	0	1
0	TN 4241	FP 625
1	FN 813	TP 934

ROC Curve



THERE ARE MANY OTHER WAYS THE MODEL COULD BE IMPROVED

Additional Sources

stack overflow
dribbble
glassdoor
gems Ruby
Meetup
HackerRank
kaggle
XING
Speaker Deck

Data Acquisition and Features

- Increase in frequency to weekly or even daily data
- NLP on institutions to identify type (universities, tech company…) - also useful for semi-supervised learning
- Analysing events by repo type (e.g. bootcamp repo…)
- More extensive regex to uncover additional user info and increase size of dataset
- Testing model with potential clients, with recruiters able to insert their notes and label users in the system

Other

- Selling services to companies for their own recruitment effort
- Optimisation of anomaly detection parameters and/or algorithm

Different Models

- Recommendation models for similar potential employees
- Enabling point system for users and using machine learning to infer rules as users select signals and users they like
- Analysing unusual drops in activity
- Clustering users and running separate models on each cluster

A FULLY DEVELOPED AND TESTED PRODUCT COULD MAKE A SIGNIFICANT IMPACT ON RECRUITERS' BOTTOM LINE

The illustrative model below(1) shows that improvements in and earlier identification of potential candidates would provide substantial business value

	Base	With OperaSignal								
Increase in Perc. Interested	0%	2%	3%	4%	5%	6%	7%	8%	9%	10%
Calls per day per Recruiter	40									
Perc. Candidates Interested	20%	22%	23%	24%	25%	26%	27%	28%	29%	30%
Perc. CV pass screen test	50%									
Perc. Candidates Interviewed	20%									
Perc.Interviews into Jobs	15%									
Avg. Hire per Day	**0,12**	**0,13**	**0,14**	**0,14**	**0,15**	**0,16**	**0,16**	**0,17**	**0,17**	**0,18**
Working Days per Month	22									
Avg. Hires per Month	**2,64**	**2,90**	**3,04**	**3,17**	**3,30**	**3,43**	**3,56**	**3,70**	**3,83**	**3,96**
Avg. Salary per Hire (€)	60.000									
% cut	17%									
Avg. Revenue per Month per Recruiter	26.928	29.621	30.967	32.314	33.660	35.006	36.353	37.699	39.046	40.392
Absolute Revenue Uplift		**2.693**	**4.039**	**5.386**	**6.732**	**8.078**	**9.425**	**10.771**	**12.118**	**13.464**
% Revenue Uplift		**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**
Illustrative Fee per User (€)		250	250	250	250	250	250	250	250	250
Fee as % of Revenue Uplift		**9,3%**	**6,2%**	**4,6%**	**3,7%**	**3,1%**	**2,7%**	**2,3%**	**2,1%**	**1,9%**

Source: (1) Conversation with ex-tech recruiter friend of mine

SO WHAT WILL THE MODEL PREDICT FOR ME?

I'm ready for my next experience!