Analysis Plan :

Stack Overflow's Developer Survey Data

Team 6 : Sebastian Waz, Sacha Uritis, & Yuxin Fang

Study Details

- Annual online survey last three weeks of January
- 2018:101,592 developers in 183 countries + dependent territories
- 129 questions, ~30 min.

- Stack Overflow :

“the world's largest and most trusted community of professional software developers”

In 2016, 46 million people used Stack Overflow (est. 16 million as professional developers)

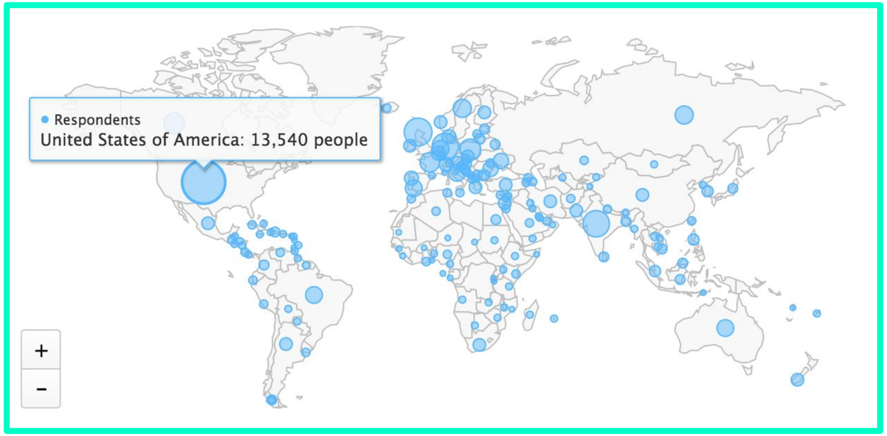
Type of Data

- **Developer Profile** (dev roles, experience, education, demographics, etc.)
- **Technology** (environments and tools)
- **Work** (ethics, salary, culture, etc.)
- **Community** (stack overflow user experience and contribution)

Typical Individual (Developer) Example

- ProgramHobby: "Yes"
- FormalEducation:"Bachelors"
- YearsCodedJob:"3-5 years"
- YearsProgram:"6-8 years"
- Gender: "Male"
- Tabs: "Spaces"

PRIMARY Q1



Who Cares? Job seekers, companies wanting to keep their salaries competitive with national/global rates, students...

What is the predicted **median salary** of software developers for 2019 in each country?

Variables & Methods

Response Var : Median salary among respondents from each country

Potential Covariates : Time, average education level, median age, proportion of tabs or spaces

Methods : LMM using ML estimation will let us handle correlated data. Empirical Bayes estimation will let us estimate individual longitudinal effects and trajectories for each country.

Challenges :

- Salary data has a substantial amount of missingness.
- Salary data collected in different ways over time (e.g. free response vs. pick a range).
- Different labeling or representations (e.g. countries *individually* vs. *cumulatively*)

Secondary Q1:

What countries are correlated in terms of the median salaries over time?

Primary Q2

Who Cares? Organizations that develop such tools (e.g. Microsoft) for the purpose of tailoring their product to their user base.

What individual characteristics are associated with the **usage of a particular software**?

(e.g. Visual Studio, Ruby, Python, C++, Github, etc.)

Variables & Methods

Response Var. : Usage of particular software (Yes/No)

Potential Covariates : Developer Type, years of experience, company size…

Methods : Principal components analysis + clustering to identify subgroups & individual characteristics (covariates). Linear regression to estimate the effect of the covariates.

Challenges :

- Data set contains more than 100 covariates with different degrees of missingness.
- Possible solutions : substantial amount of recoding, imputation, and variable selection will need to take place.

PRIMARY Q3

stack overflow

Who Cares? Stack Overflow Analysts + Potential Advertisers

What are factors are associated with "heavy-users" of Stack Overflow?

Variables & Methods

Response Var. : Whether or not you are a heavy-user of Stack Overflow (Yes/No)

Potential Covariates : Years coding, type of developer...

Methods : Clustering to identify subgroups & attribute patterns of behavior. Logistic regression to estimate the effect of selected characteristics from Clustering.

Challenges :

- Defining a “heavy-user” using questions like how likely are you willing to recommend S.O. or how frequently do you participate in Q&A on Stack Overflow?
- Missingness : these questions are part of the last half of the survey.

Primary Q4

Who Cares? Many people are interested in the progression of non-male presence in tech industry.

What is the rate of change in proportion of males over time across countries?

Secondary Q4:

What is the predicted **proportion of males** for 2019 in each country?

Variables & Methods

Response Var. : Proportion of Males

Potential Covariates : time

Methods : GLMM (Logistic regression) for primary question + Empirical Bayes Estimation for secondary question (to track the trajectory of each country).

Challenges :

- Different categorizations with gender amongst all surveys
- Various Types of Missingness : Nonresponse or Prefer not to Answer

```
<div class="card card-shadow">
    <div class="card-body">
        <h4><?= $language->admin_extra_settings->header; ?></h4>

        <h5><?= $language->admin_extra_settings->display->reset ?></h5>
        <p class="text-muted"><?= $language->admin_extra_settings->display->reset_help ?></p>

        <form action="" method="post" role="form">
            <input type="hidden" name="type" value="reset" />

            <div class="form-check">
                <label class="form-check-label">
                    <input type="checkbox" class="form-check-input" name="users">
                    <?= $language->admin_extra_settings->input->reset_users ?>
                </label>
            </div>

            <?php foreach($sources as $source): ?>
                <div class="form-check">
                    <label class="form-check-label">
                        <input type="checkbox" class="form-check-input" name="<?= $source ?>_users">
                        <?= sprintf($language->admin_extra_settings->input->reset_source, $source) ?>
                    </label>
                </div>

                <div class="form-check">
                    <label class="form-check-label">
                        <input type="checkbox" class="form-check-input" name="<?= $source ?>_logs">
                        <?= sprintf($language->admin_extra_settings->input->reset_source_logs, $source) ?>
                    </label>
                </div>
            <?php endforeach; ?>

            <div class="text-center">
                <button type="submit" name="submit" class="btn btn-primary btn-sm" data-confirm="<?= $language->global->info
            </div>
        </form>
    </div>
</div>
```

Q&A