What Do I Do With
This Giant Pile of Data?

Aja Hammerly

http://github.com/thagomizer

@thagomizer_rb

http://www.thagomizer.com

SUBSTANTIAL

COMPELLING INTRO

https://www.flickr.com/photos/sindesign/238419364

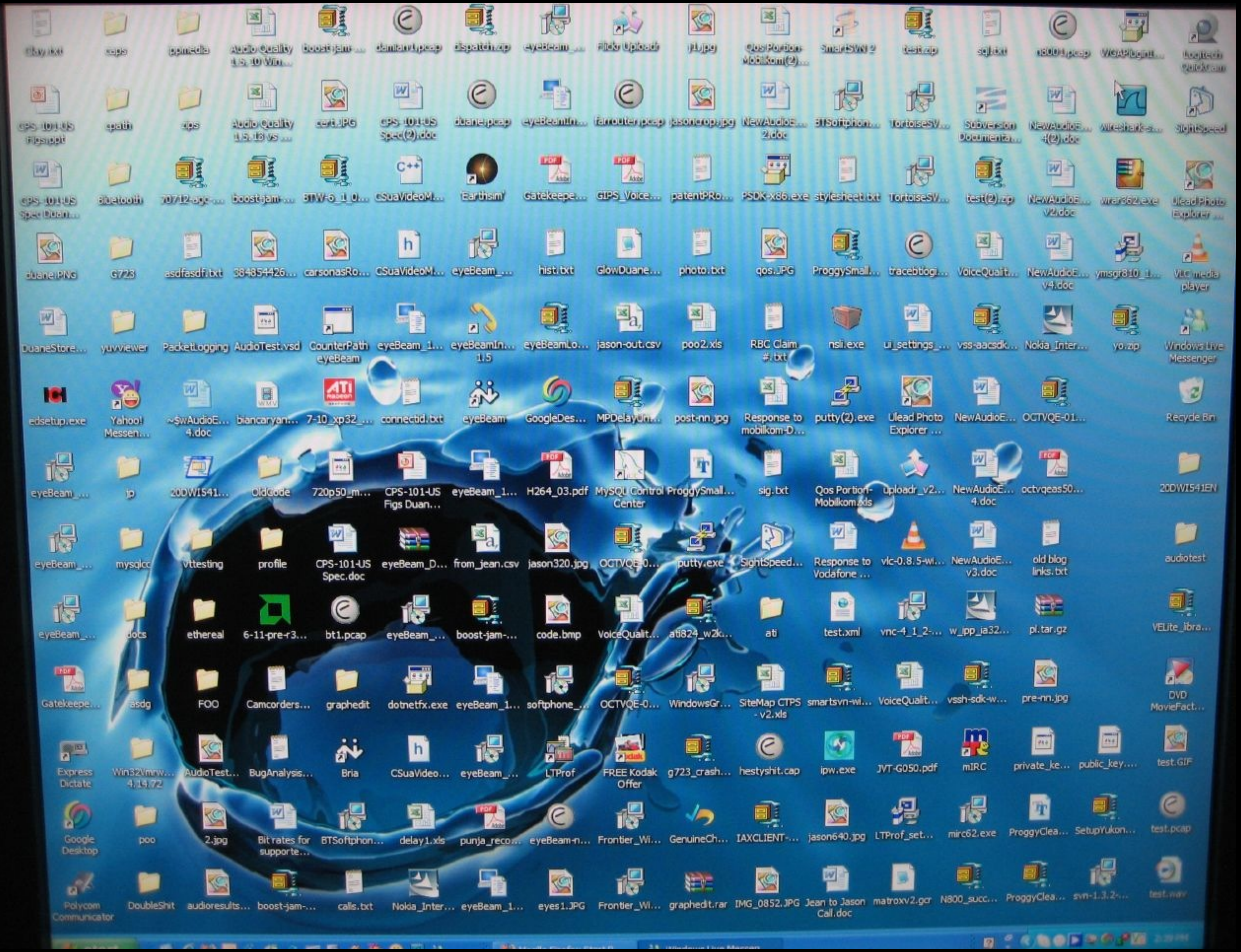
the ultimate guide to clearing your clutter mary lambert
Home Crafts Eagle Editions
THE MANLY MAN MANUAL
TREES OF DELHI
1000 GARDEN IDEAS
courrèges
CATS & KITTENS Joan Moore REBO
Housekeeping Operations, Design and Management Malini Singh Jaya B. George A JAICO BOOK
very important pets Lustre Press Roli Books
THE DOG OWNER'S HANDBOOK Annette Conn
PUPPIES
PRINCIPLES OF DESIGN THROUGH PHOTOGRAPHY
GEORGE LEPP
GEORGE LEPP KATHRYN VINCENT LEPP WILDLIFE PHOTOGRAPHY

https://www.flickr.com/photos/spykster/6221799231

http://commons.wikimedia.org/wiki/File:HP_EVA4400-1.jpg

https://www.flickr.com/photos/creativecommons/3027812285/

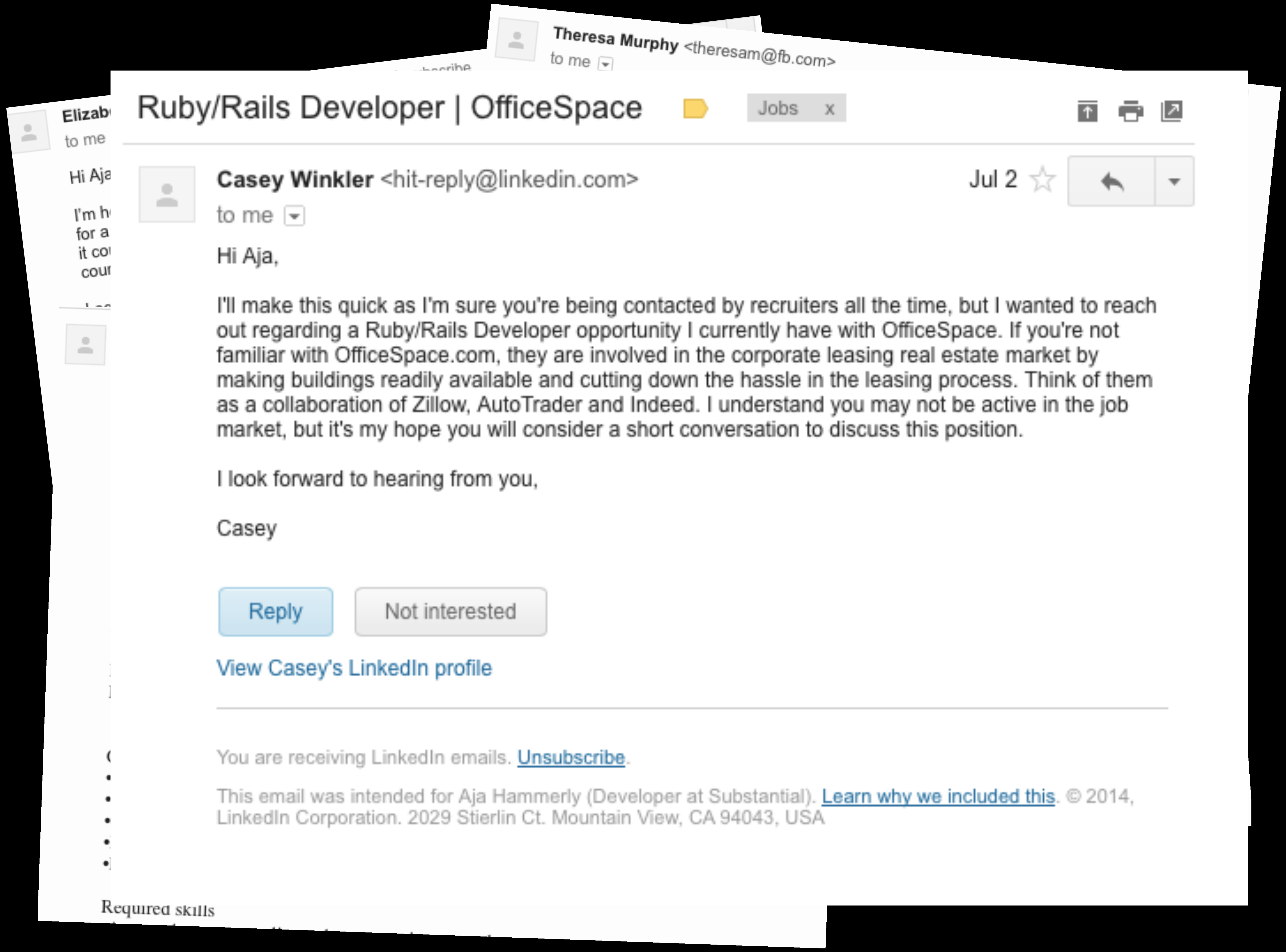
Storage
Is
Cheap



http://www.flickr.com/photos/duanestorey/2053844586/

RANT

CASE STUDY



Ruby/Rails Developer | OfficeSpace

Casey Winkler <hit-reply@linkedin.com> Jul 2

to me

Hi Aja,

I'll make this quick as I'm sure you're being contacted by recruiters all the time, but I wanted to reach out regarding a Ruby/Rails Developer opportunity I currently have with OfficeSpace. If you're not familiar with OfficeSpace.com, they are involved in the corporate leasing real estate market by making buildings readily available and cutting down the hassle in the leasing process. Think of them as a collaboration of Zillow, AutoTrader and Indeed. I understand you may not be active in the job market, but it's my hope you will consider a short conversation to discuss this position.

I look forward to hearing from you,

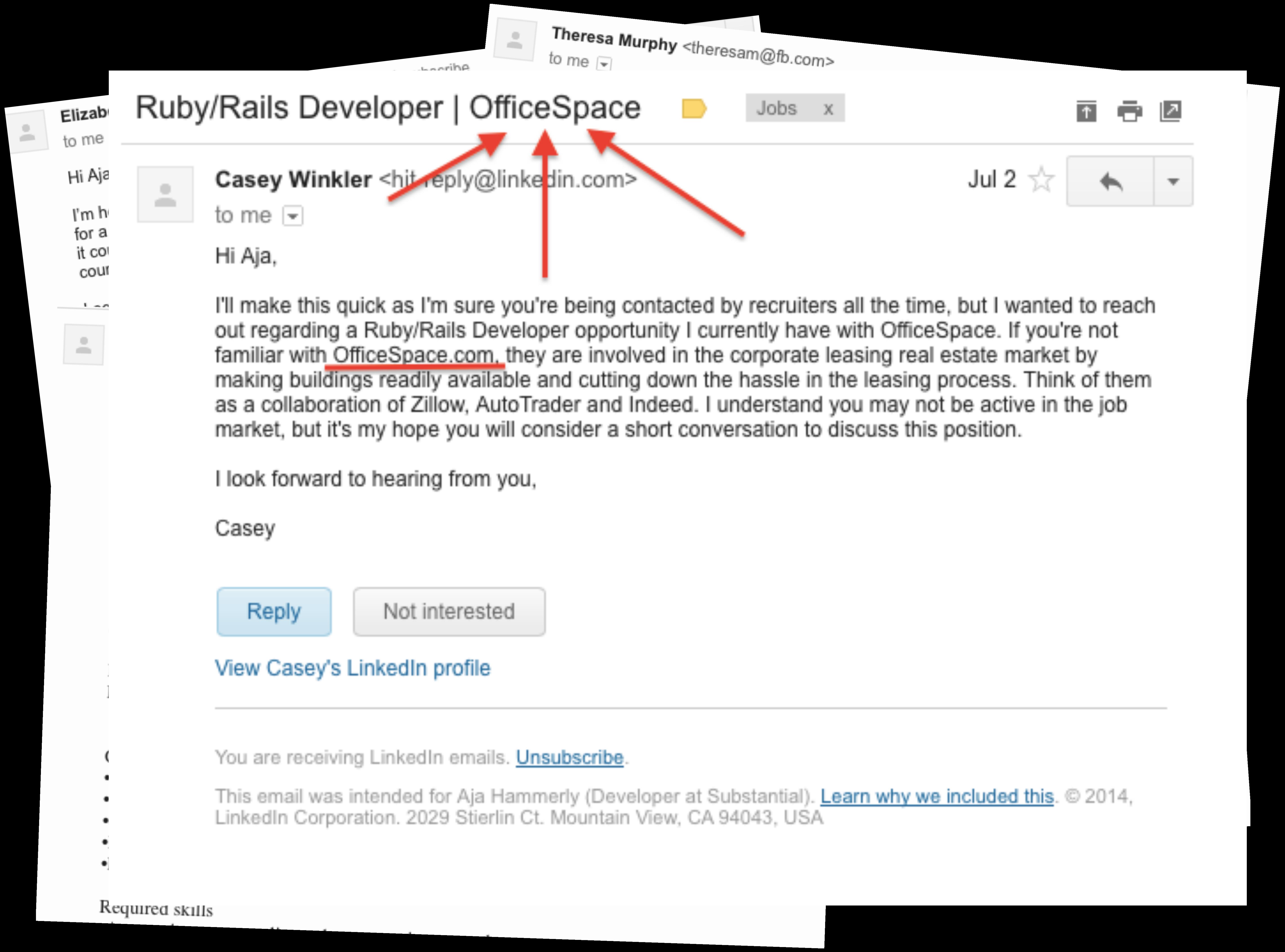
Casey

View Casey's LinkedIn profile

You are receiving LinkedIn emails. Unsubscribe.

This email was intended for Aja Hammerly (Developer at Substantial). Learn why we included this. © 2014, LinkedIn Corporation. 2029 Stierlin Ct. Mountain View, CA 94043, USA

Required skills



Theresa Murphy <theresam@fb.com>
to me

Elizab
to me

Hi Aja

I'm h
for a
it co
cour

Required skills

Ruby/Rails Developer | OfficeSpace

Casey Winkler <hit-reply@linkedin.com>
to me

Jul 2

Hi Aja,

I'll make this quick as I'm sure you're being contacted by recruiters all the time, but I wanted to reach out regarding a Ruby/Rails Developer opportunity I currently have with OfficeSpace. If you're not familiar with OfficeSpace.com, they are involved in the corporate leasing real estate market by making buildings readily available and cutting down the hassle in the leasing process. Think of them as a collaboration of Zillow, AutoTrader and Indeed. I understand you may not be active in the job market, but it's my hope you will consider a short conversation to discuss this position.

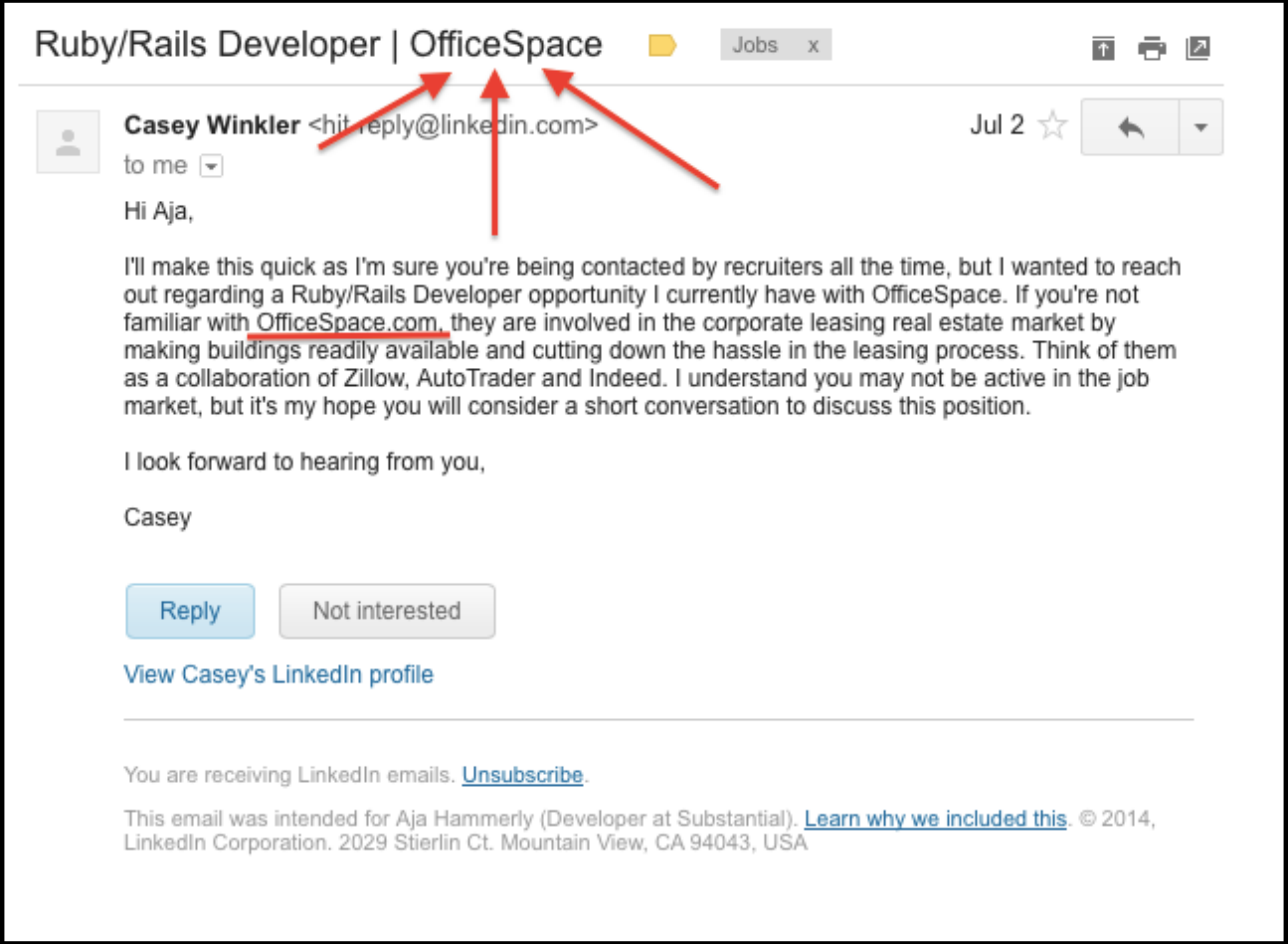
I look forward to hearing from you,

Casey

View Casey's LinkedIn profile

You are receiving LinkedIn emails. Unsubscribe.

This email was intended for Aja Hammerly (Developer at Substantial). Learn why we included this. © 2014, LinkedIn Corporation. 2029 Stierlin Ct. Mountain View, CA 94043, USA



Ruby/Rails Developer | OfficeSpace

Casey Winkler <hit-reply@linkedin.com>

to me

Jul 2

Hi Aja,

I'll make this quick as I'm sure you're being contacted by recruiters all the time, but I wanted to reach out regarding a Ruby/Rails Developer opportunity I currently have with OfficeSpace. If you're not familiar with OfficeSpace.com, they are involved in the corporate leasing real estate market by making buildings readily available and cutting down the hassle in the leasing process. Think of them as a collaboration of Zillow, AutoTrader and Indeed. I understand you may not be active in the job market, but it's my hope you will consider a short conversation to discuss this position.

I look forward to hearing from you,

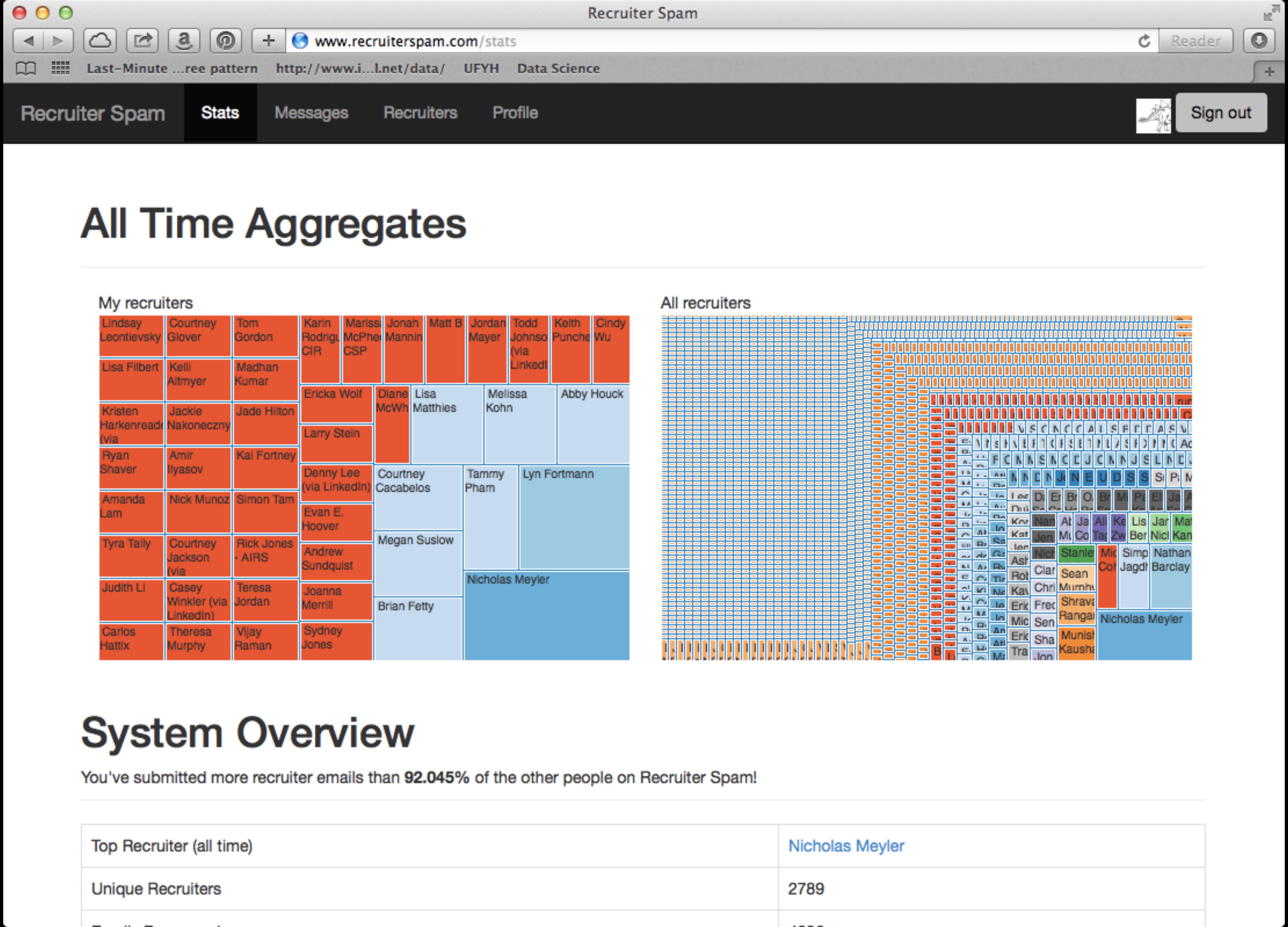
Casey

View Casey's LinkedIn profile

You are receiving LinkedIn emails. Unsubscribe.

This email was intended for Aja Hammerly (Developer at Substantial). Learn why we included this. © 2014, LinkedIn Corporation. 2029 Stierlin Ct. Mountain View, CA 94043, USA

https://www.flickr.com/photos/jdhancock/7823214508



All Time Aggregates

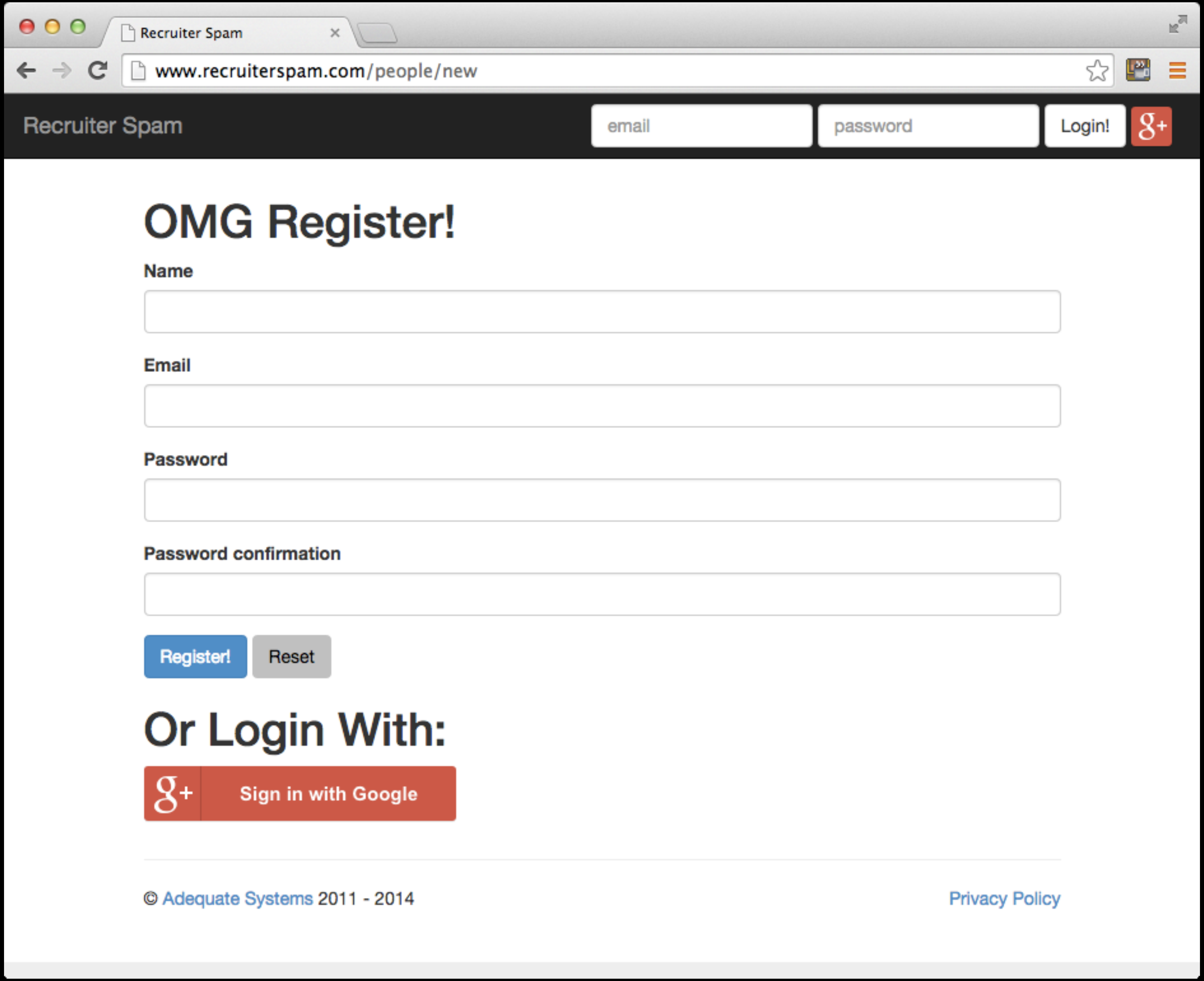
My recruiters

All recruiters

System Overview

You've submitted more recruiter emails than **92.045%** of the other people on Recruiter Spam!

Top Recruiter (all time)	Nicholas Meyler
Unique Recruiters	2789



OMG Register!

Name

Email

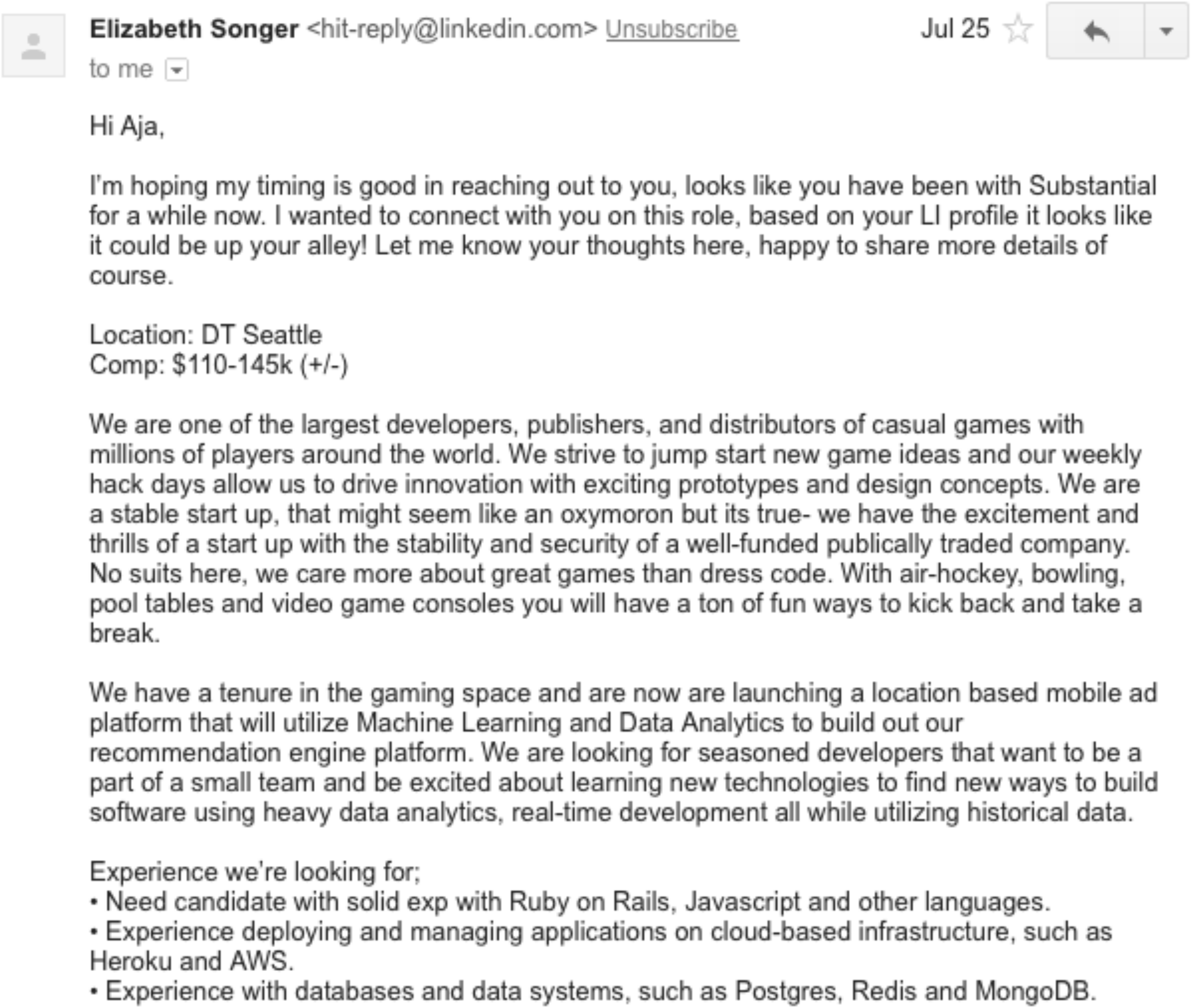
Password

Password confirmation

Or Login With:

© Adequate Systems 2011 - 2014

Privacy Policy



Elizabeth Songer <hit-reply@linkedin.com> Unsubscribe Jul 25

to me

Hi Aja,

I'm hoping my timing is good in reaching out to you, looks like you have been with Substantial for a while now. I wanted to connect with you on this role, based on your LI profile it looks like it could be up your alley! Let me know your thoughts here, happy to share more details of course.

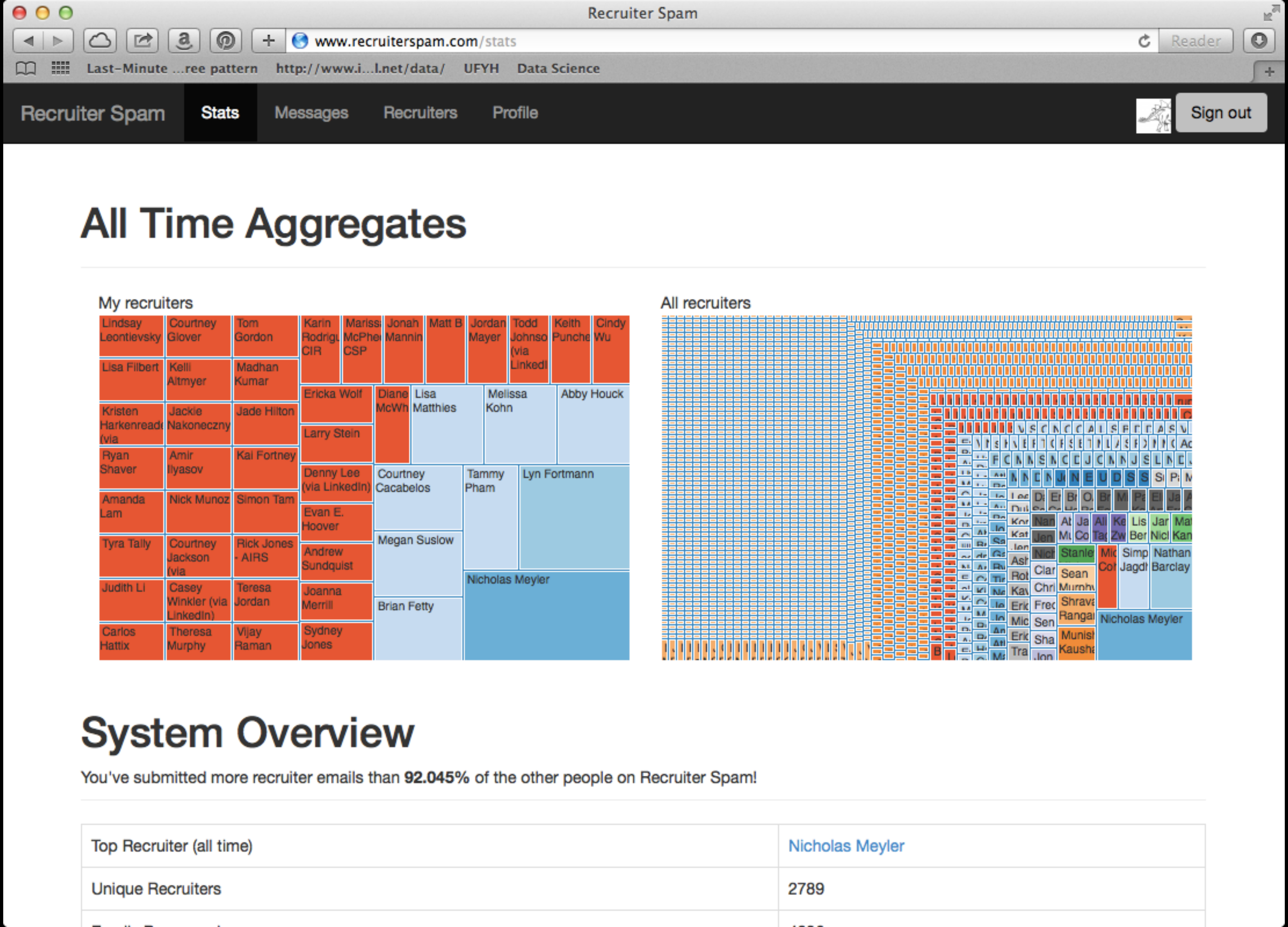
Location: DT Seattle
Comp: $110-145k (+/-)

We are one of the largest developers, publishers, and distributors of casual games with millions of players around the world. We strive to jump start new game ideas and our weekly hack days allow us to drive innovation with exciting prototypes and design concepts. We are a stable start up, that might seem like an oxymoron but its true- we have the excitement and thrills of a start up with the stability and security of a well-funded publically traded company. No suits here, we care more about great games than dress code. With air-hockey, bowling, pool tables and video game consoles you will have a ton of fun ways to kick back and take a break.

We have a tenure in the gaming space and are now are launching a location based mobile ad platform that will utilize Machine Learning and Data Analytics to build out our recommendation engine platform. We are looking for seasoned developers that want to be a part of a small team and be excited about learning new technologies to find new ways to build software using heavy data analytics, real-time development all while utilizing historical data.

Experience we're looking for;

- Need candidate with solid exp with Ruby on Rails, Javascript and other languages.
- Experience deploying and managing applications on cloud-based infrastructure, such as Heroku and AWS.
- Experience with databases and data systems, such as Postgres, Redis and MongoDB.



All Time Aggregates

My recruiters

All recruiters

System Overview

You've submitted more recruiter emails than **92.045%** of the other people on Recruiter Spam!

Top Recruiter (all time)	Nicholas Meyler
Unique Recruiters	2789

Overview

1. Keep Useful Data
2. Make Your Data Usable
3. Using Your Data

Guideline 1:
Keep Useful Data

What Counts As Data?

Database

Logs

App Logs

App

DB

Logs

App
DB

Email Logs

App
DB
Email

Error   Logs

App
DB
Email
Error
Server Logs

Emails

Customer Feedback

Click Stream Data

Backups

Outside Service Data

Everything & Anything

What Data Is Useful?

Data You Use

Data That Is Relevant

Data Has An Expiration

Aggregate

Data That You Will Use Soon

Data For CYA

SOX

HIPPA

COPA

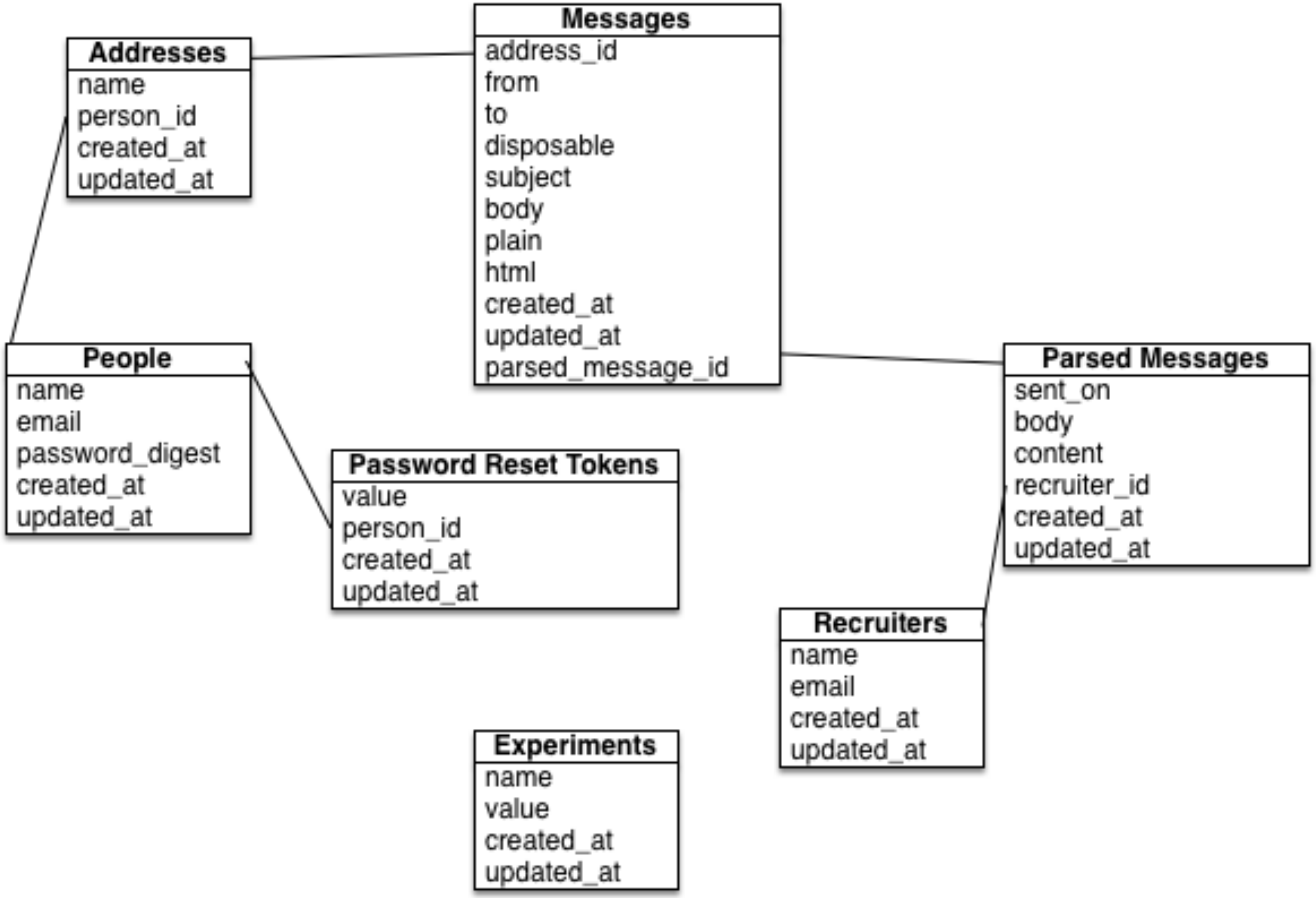
Financial Data

What If…

http://www.webweaver.nu/clipart/img/fantasy/fairies/fairy-godmother-animation.gif

Recruiter Spam

Database



Addresses
name
person_id
created_at
updated_at

Messages
address_id
from
to
disposable
subject
body
plain
html
created_at
updated_at
parsed_message_id

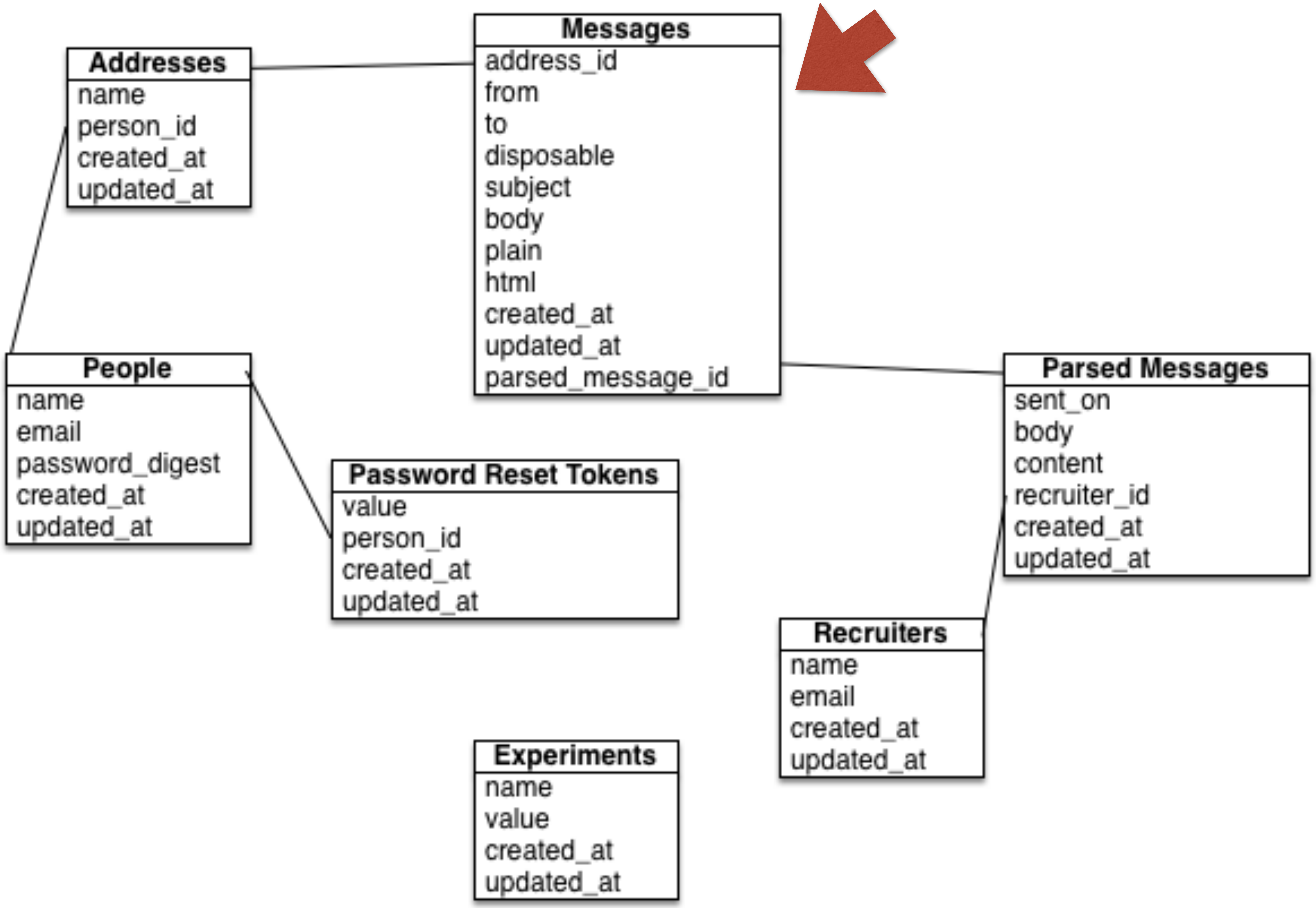
People
name
email
password_digest
created_at
updated_at

Password Reset Tokens
value
person_id
created_at
updated_at

Parsed Messages
sent_on
body
content
recruiter_id
created_at
updated_at

Recruiters
name
email
created_at
updated_at

Experiments
name
value
created_at
updated_at



Addresses
name
person_id
created_at
updated_at

Messages
address_id
from
to
disposable
subject
body
plain
html
created_at
updated_at
parsed_message_id

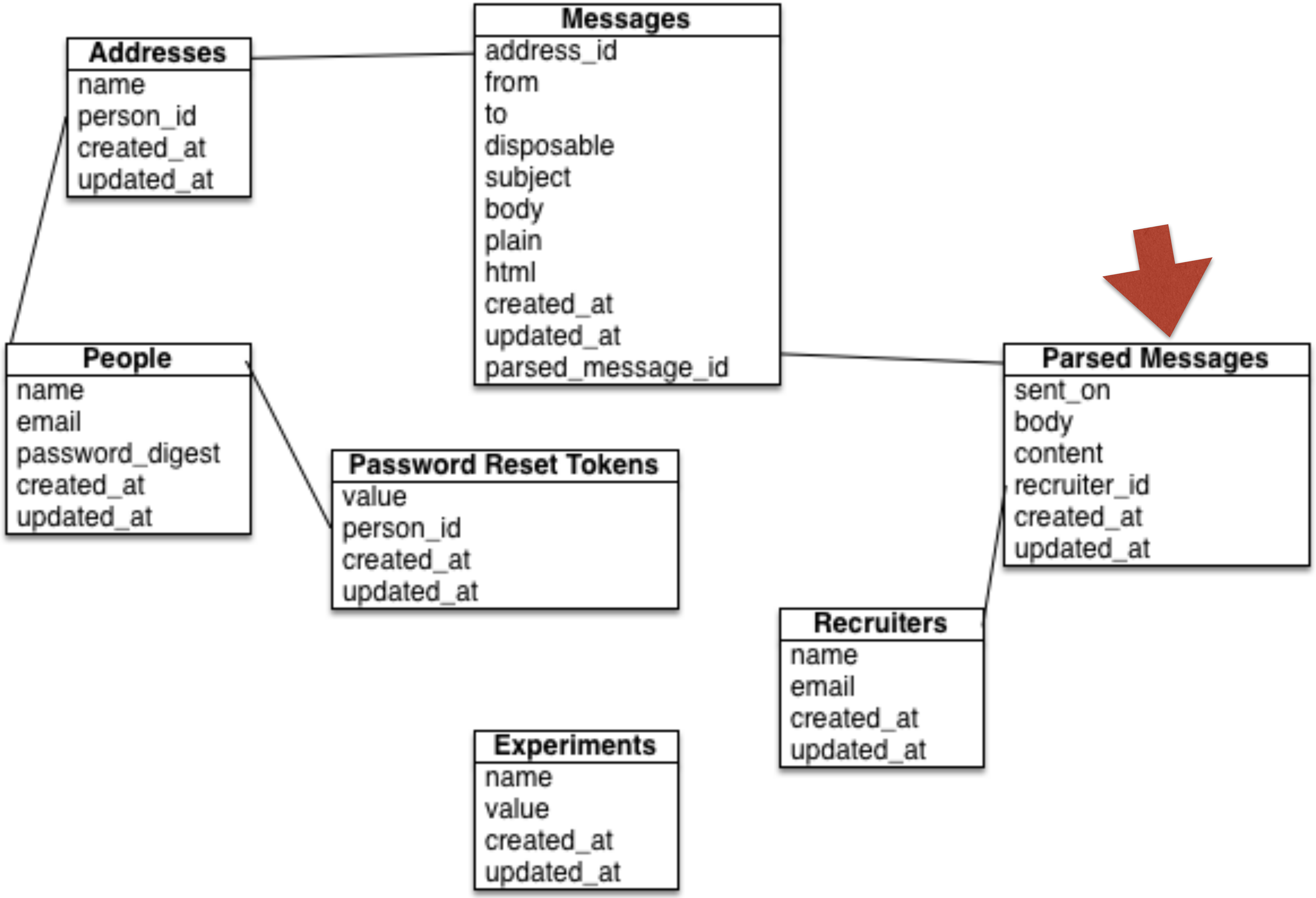
People
name
email
password_digest
created_at
updated_at

Password Reset Tokens
value
person_id
created_at
updated_at

Parsed Messages
sent_on
body
content
recruiter_id
created_at
updated_at

Recruiters
name
email
created_at
updated_at

Experiments
name
value
created_at
updated_at



Addresses
name
person_id
created_at
updated_at

Messages
address_id
from
to
disposable
subject
body
plain
html
created_at
updated_at
parsed_message_id

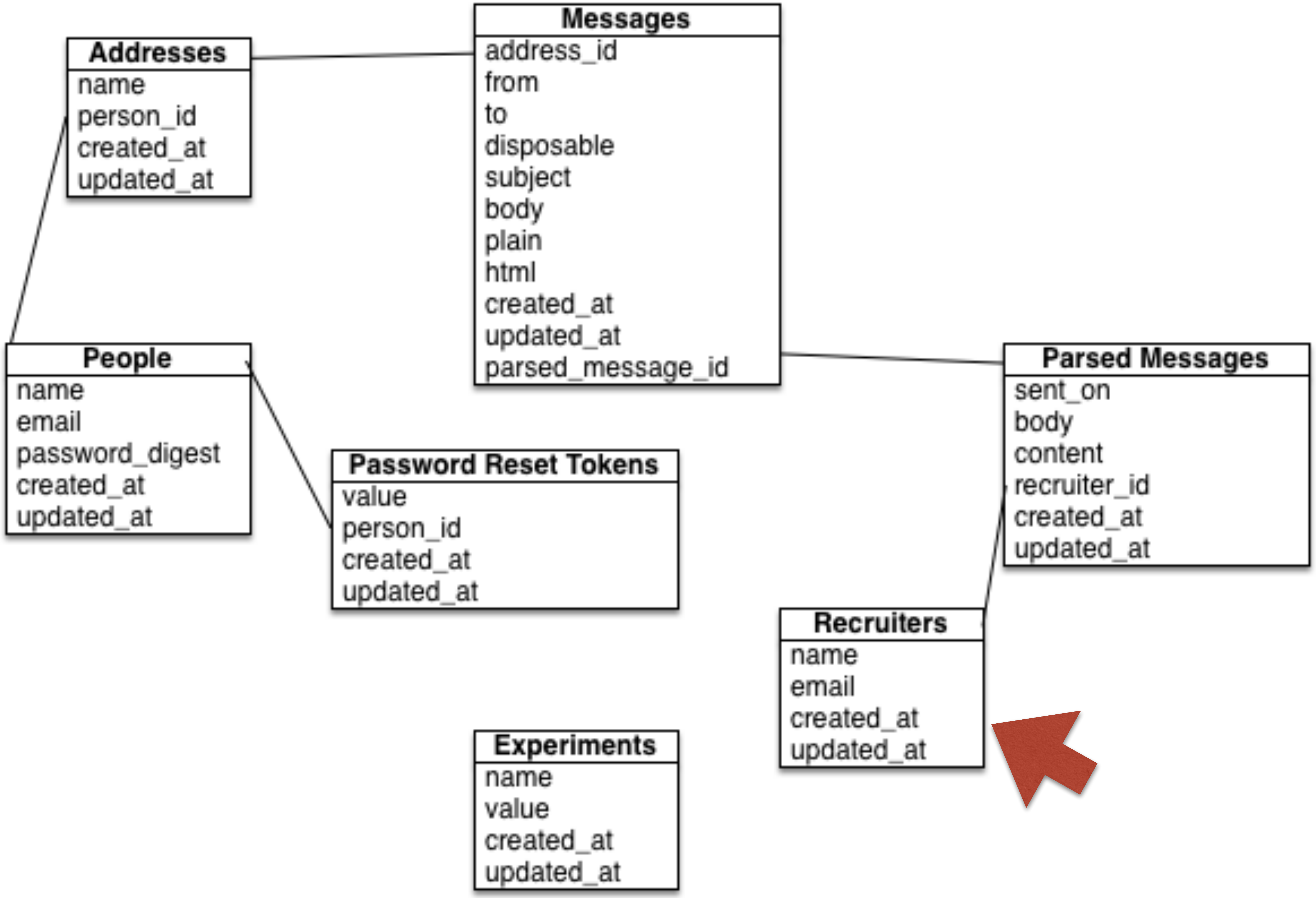
People
name
email
password_digest
created_at
updated_at

Password Reset Tokens
value
person_id
created_at
updated_at

Parsed Messages
sent_on
body
content
recruiter_id
created_at
updated_at

Recruiters
name
email
created_at
updated_at

Experiments
name
value
created_at
updated_at



Addresses
name
person_id
created_at
updated_at

Messages
address_id
from
to
disposable
subject
body
plain
html
created_at
updated_at
parsed_message_id

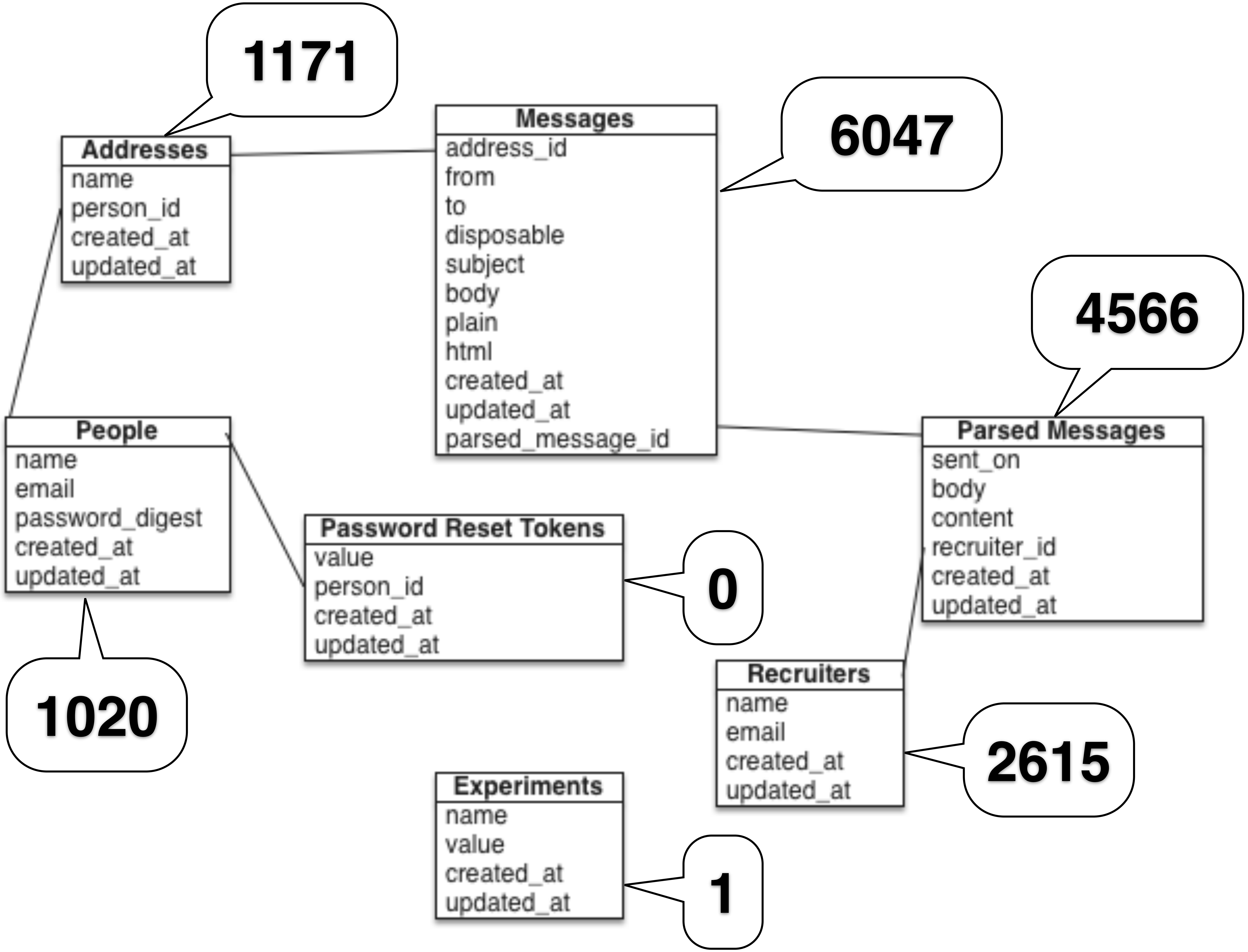
People
name
email
password_digest
created_at
updated_at

Password Reset Tokens
value
person_id
created_at
updated_at

Parsed Messages
sent_on
body
content
recruiter_id
created_at
updated_at

Recruiters
name
email
created_at
updated_at

Experiments
name
value
created_at
updated_at



Addresses
name
person_id
created_at
updated_at

1171

Messages
address_id
from
to
disposable
subject
body
plain
html
created_at
updated_at
parsed_message_id

6047

Parsed Messages
sent_on
body
content
recruiter_id
created_at
updated_at

4566

People
name
email
password_digest
created_at
updated_at

1020

Password Reset Tokens
value
person_id
created_at
updated_at

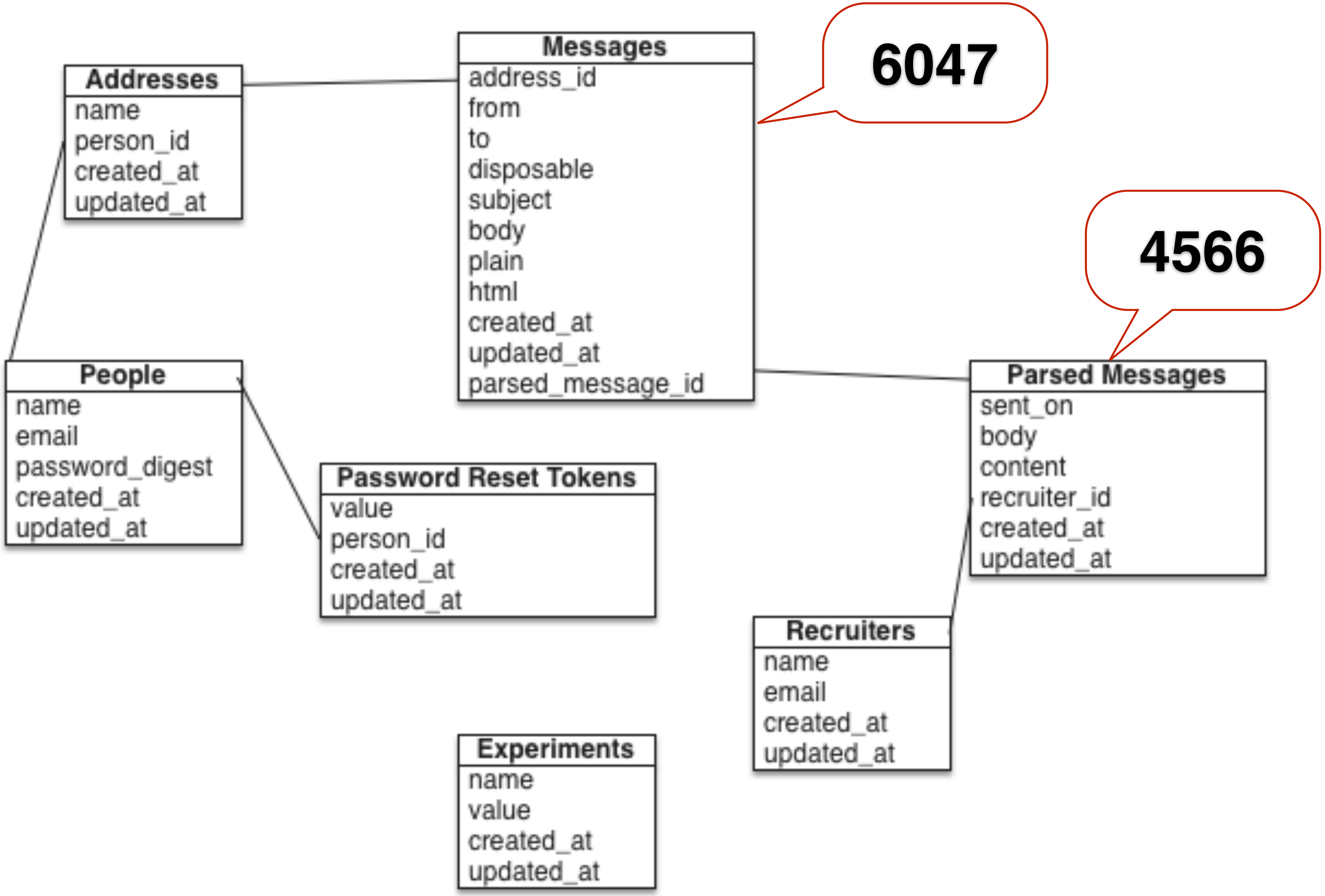
0

Recruiters
name
email
created_at
updated_at

2615

Experiments
name
value
created_at
updated_at

1



Addresses
name
person_id
created_at
updated_at

Messages
address_id
from
to
disposable
subject
body
plain
html
created_at
updated_at
parsed_message_id

6047

4566

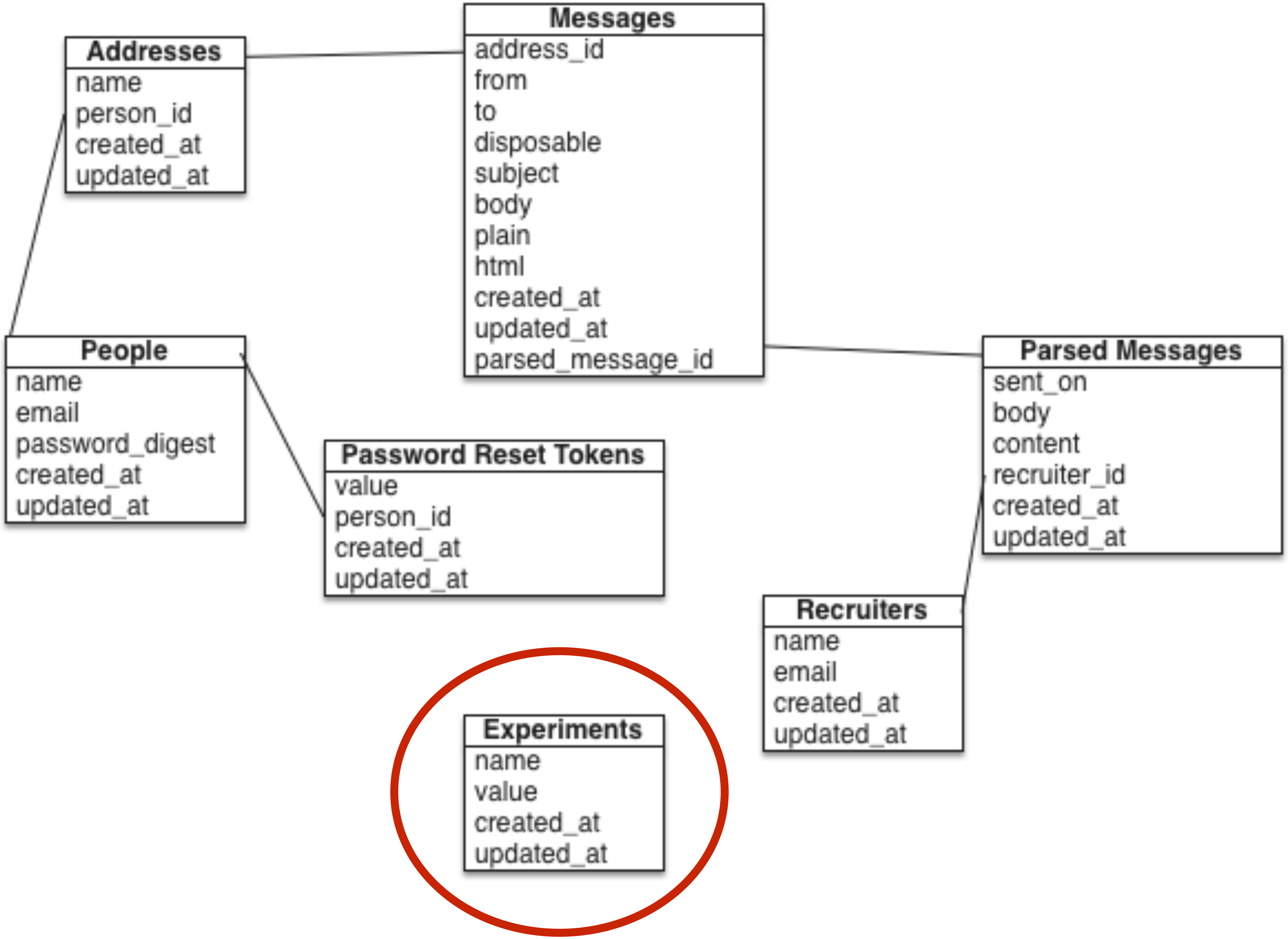
People
name
email
password_digest
created_at
updated_at

Password Reset Tokens
value
person_id
created_at
updated_at

Parsed Messages
sent_on
body
content
recruiter_id
created_at
updated_at

Recruiters
name
email
created_at
updated_at

Experiments
name
value
created_at
updated_at



Addresses
name
person_id
created_at
updated_at

Messages
address_id
from
to
disposable
subject
body
plain
html
created_at
updated_at
parsed_message_id

People
name
email
password_digest
created_at
updated_at

Password Reset Tokens
value
person_id
created_at
updated_at

Parsed Messages
sent_on
body
content
recruiter_id
created_at
updated_at

Recruiters
name
email
created_at
updated_at

Experiments
name
value
created_at
updated_at

Other Data

Guideline 2:
Make Your Data Usable

Accessible

Centralized

Searchable

Idiot Proof

Read Only

Comprehensible

Column Names

Units

Explain Things

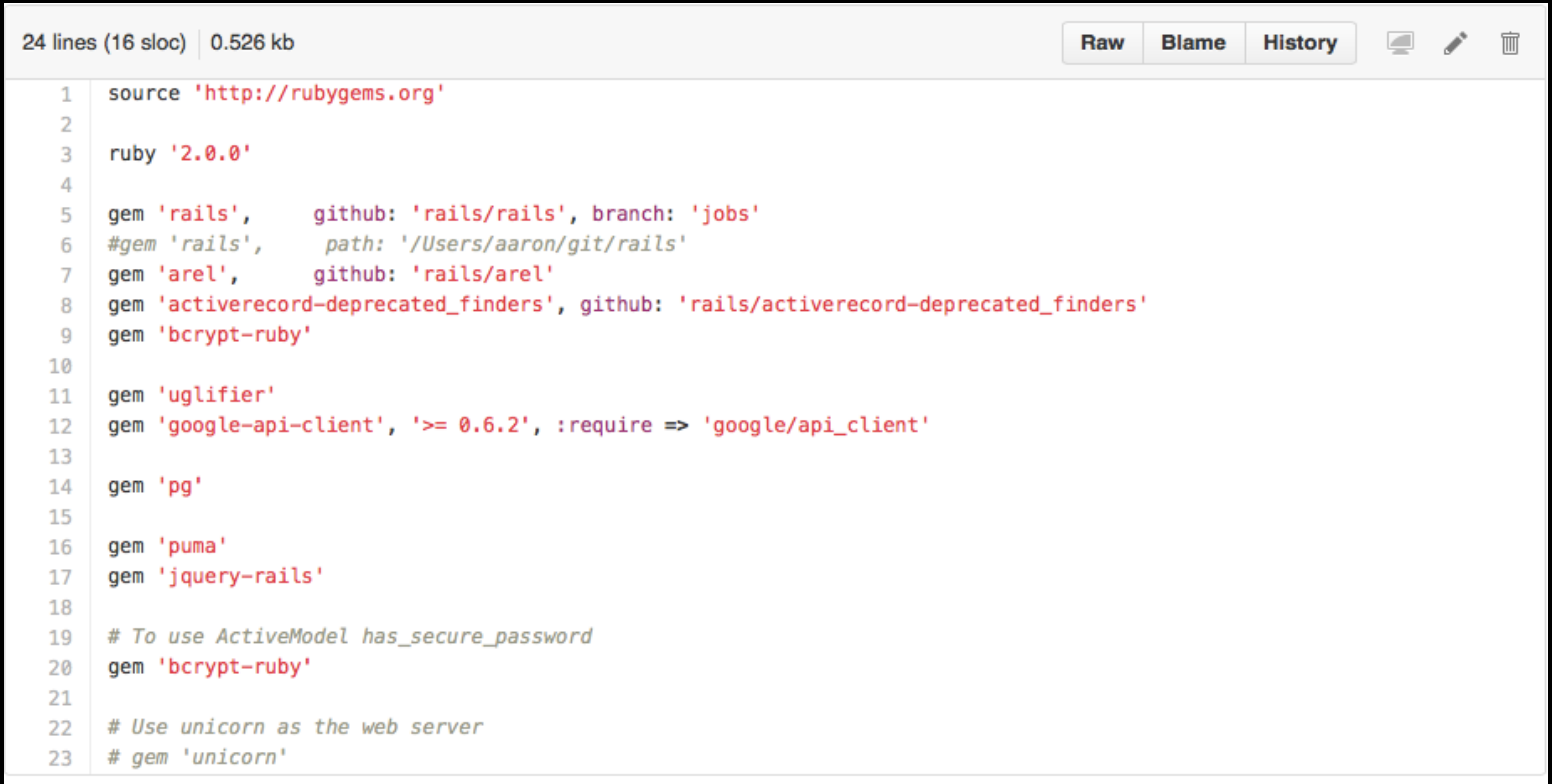
Cleansed

Test Data

Anonymized

Automatic

Recruiter Spam



```ruby
source 'http://rubygems.org'

ruby '2.0.0'

gem 'rails',     github: 'rails/rails', branch: 'jobs'
#gem 'rails',     path: '/Users/aaron/git/rails'
gem 'arel',      github: 'rails/arel'
gem 'activerecord-deprecated_finders', github: 'rails/activerecord-deprecated_finders'
gem 'bcrypt-ruby'

gem 'uglifier'
gem 'google-api-client', '>= 0.6.2', :require => 'google/api_client'

gem 'pg'

gem 'puma'
gem 'jquery-rails'

# To use ActiveModel has_secure_password
gem 'bcrypt-ruby'

# Use unicorn as the web server
# gem 'unicorn'
```

Improvements

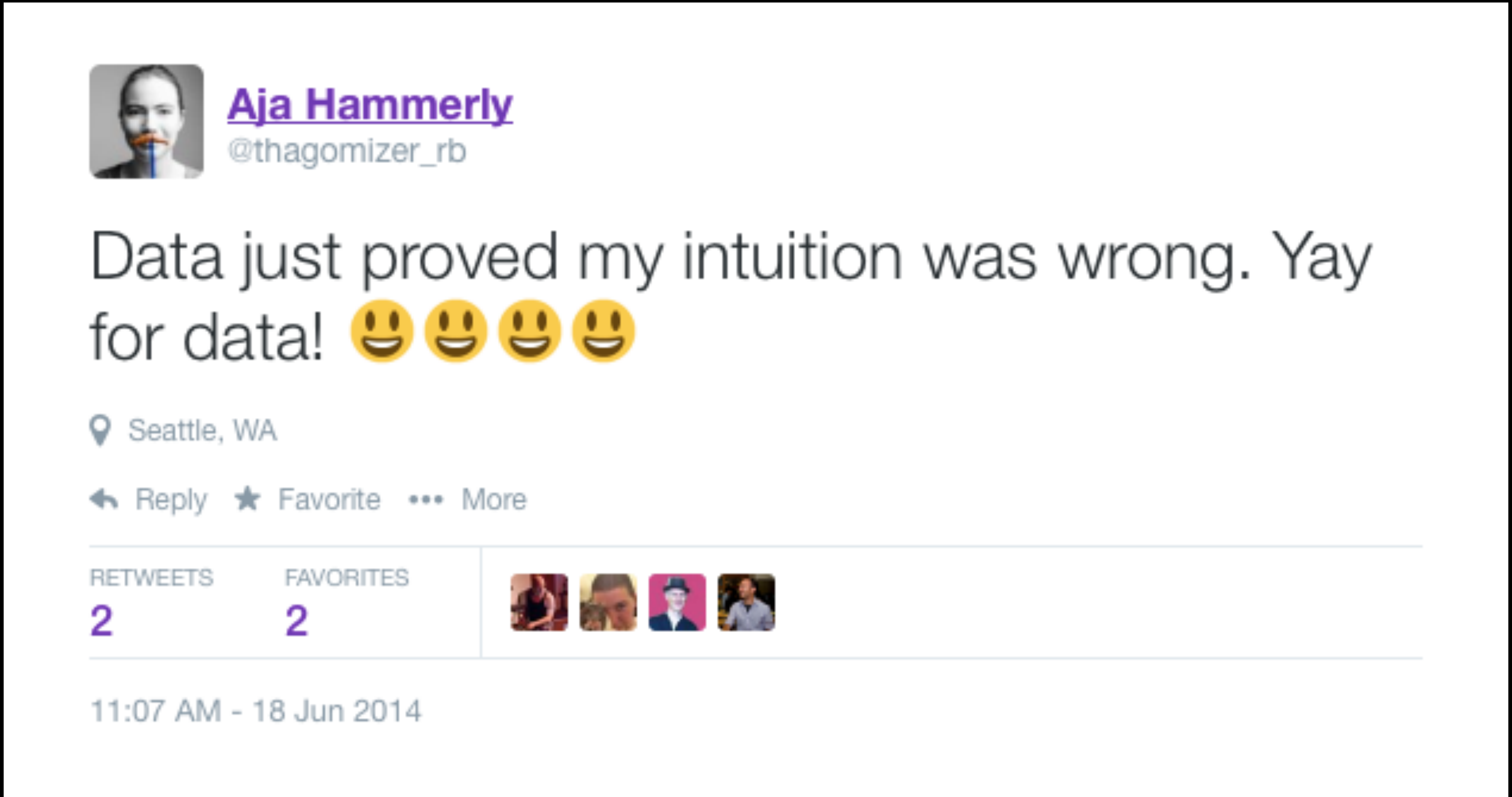
- Automate restoring prod to local
- Fix Gemfile
- Improve Documentation

Guideline 3:
Use Your Data!

Drive Improvement

Justify Decisions

Challenge Assumptions



Aja Hammerly
@thagomizer_rb

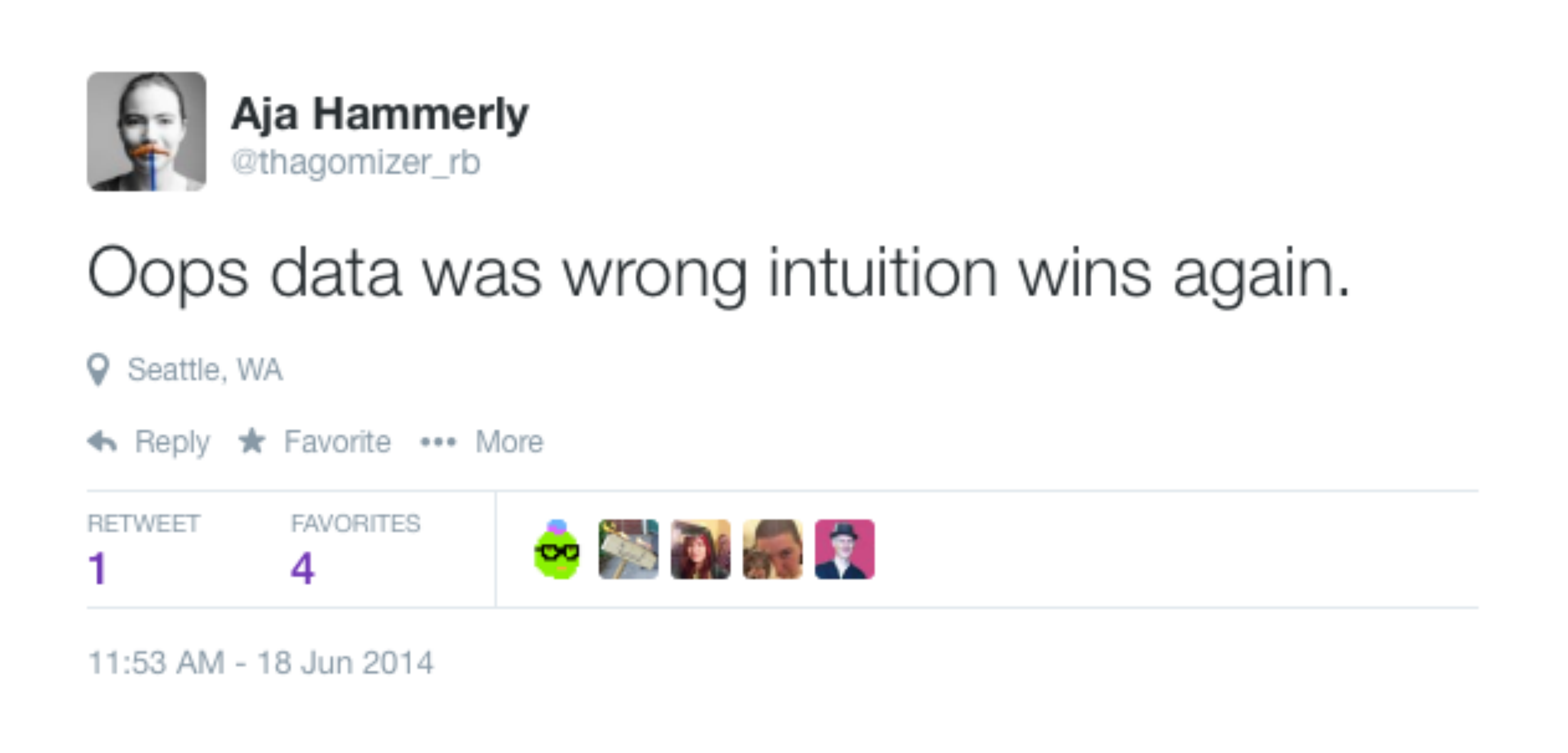
Data just proved my intuition was wrong. Yay for data! 😃😃😃😃

Seattle, WA

RETWEETS 2

FAVORITES 2

11:07 AM - 18 Jun 2014



Aja Hammerly
@thagomizer_rb

Oops data was wrong intuition wins again.

Seattle, WA

RETWEET 1 FAVORITES 4

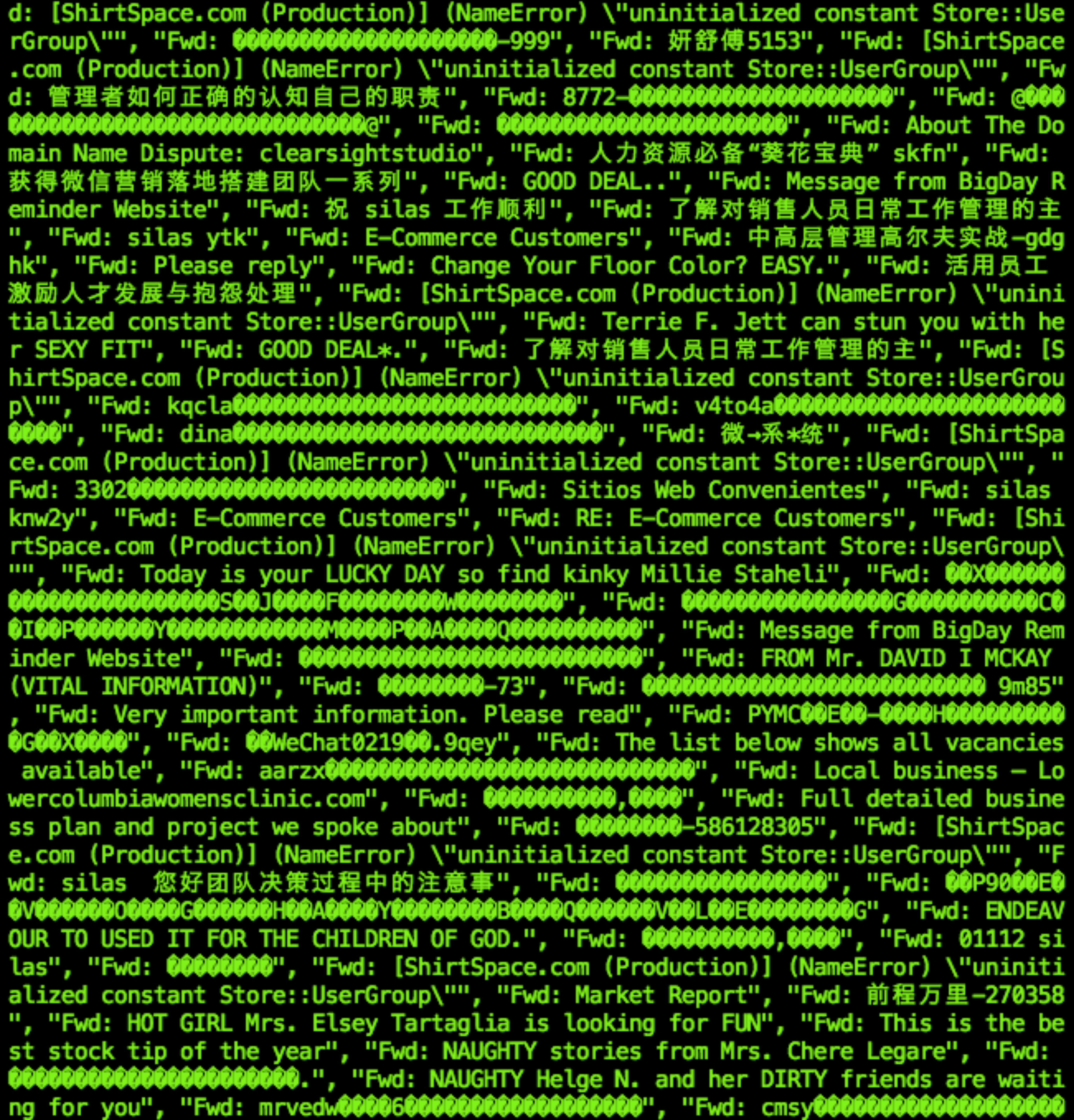
11:53 AM - 18 Jun 2014

Experiment

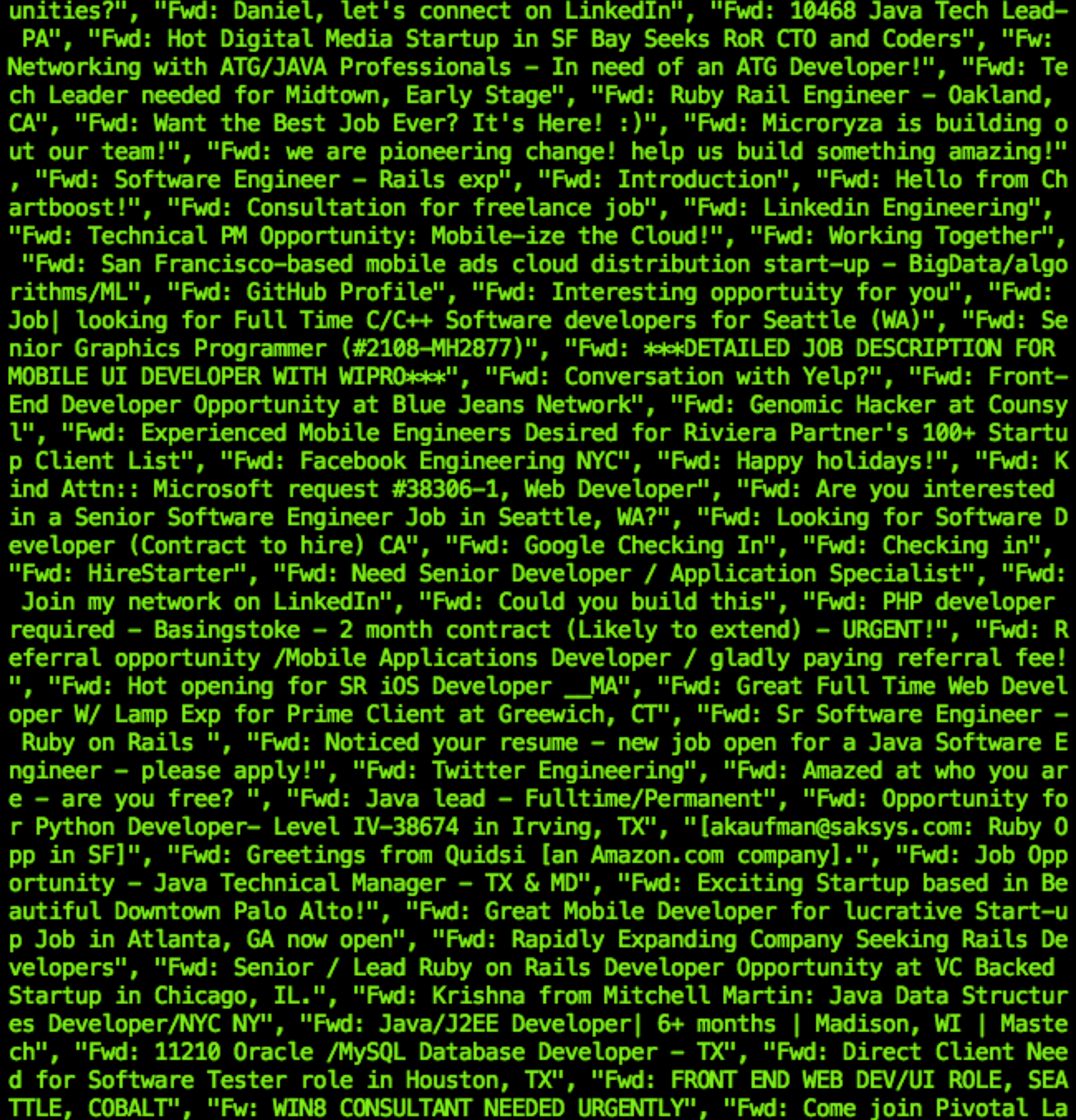
Recruiter Spam

Fewer Unparsed Emails

What can we add to reduce the unparsable emails?



```
d: [ShirtSpace.com (Production)] (NameError) \"uninitialized constant Store::Use
rGroup\"", "Fwd: ��������������������-999", "Fwd: 妍舒傅5153", "Fwd: [ShirtSpace
.com (Production)] (NameError) \"uninitialized constant Store::UserGroup\"", "Fw
d: 管理者如何正确的认知自己的职责", "Fwd: 8772-��������������������", "Fwd: @���
���������������������������@", "Fwd: ����������������������", "Fwd: About The Do
main Name Dispute: clearsightstudio", "Fwd: 人力资源必备“葵花宝典” skfn", "Fwd:
获得微信营销落地搭建团队一系列", "Fwd: GOOD DEAL..", "Fwd: Message from BigDay R
eminder Website", "Fwd: 祝 silas 工作顺利", "Fwd: 了解对销售人员日常工作管理的主
", "Fwd: silas ytk", "Fwd: E-Commerce Customers", "Fwd: 中高层管理高尔夫实战-gdg
hk", "Fwd: Please reply", "Fwd: Change Your Floor Color? EASY.", "Fwd: 活用员工
激励人才发展与抱怨处理", "Fwd: [ShirtSpace.com (Production)] (NameError) \"unini
tialized constant Store::UserGroup\"", "Fwd: Terrie F. Jett can stun you with he
r SEXY FIT", "Fwd: GOOD DEAL*.", "Fwd: 了解对销售人员日常工作管理的主", "Fwd: [S
hirtSpace.com (Production)] (NameError) \"uninitialized constant Store::UserGrou
p\"", "Fwd: kqcla�������������������������", "Fwd: v4to4a����������������������
����", "Fwd: dina����������������������������", "Fwd: 微→系*统", "Fwd: [ShirtSpa
ce.com (Production)] (NameError) \"uninitialized constant Store::UserGroup\"", "
Fwd: 3302������������������������", "Fwd: Sitios Web Convenientes", "Fwd: silas 
knw2y", "Fwd: E-Commerce Customers", "Fwd: RE: E-Commerce Customers", "Fwd: [Shi
rtSpace.com (Production)] (NameError) \"uninitialized constant Store::UserGroup\
"", "Fwd: Today is your LUCKY DAY so find kinky Millie Staheli", "Fwd: ��X������
����������������S��J����F��������W��������", "Fwd: ����������������G����������C�
�I��P������Y������������M����P��A����Q����������", "Fwd: Message from BigDay Rem
inder Website", "Fwd: ��������������������������", "Fwd: FROM Mr. DAVID I MCKAY 
(VITAL INFORMATION)", "Fwd: ��������-73", "Fwd: �������������������������� 9m85"
, "Fwd: Very important information. Please read", "Fwd: PYMC��E��-����H���������
�G��X����", "Fwd: ��WeChat0219��.9qey", "Fwd: The list below shows all vacancies
 available", "Fwd: aarzx����������������������������", "Fwd: Local business – Lo
wercolumbiawomensclinic.com", "Fwd: ����������,����", "Fwd: Full detailed busine
ss plan and project we spoke about", "Fwd: ��������-586128305", "Fwd: [ShirtSpac
e.com (Production)] (NameError) \"uninitialized constant Store::UserGroup\"", "F
wd: silas  您好团队决策过程中的注意事", "Fwd: ����������������", "Fwd: ��P9���E�
�V������O����G������H��A����Y��������B����Q������V��L��E��������G", "Fwd: ENDEAV
OUR TO USED IT FOR THE CHILDREN OF GOD.", "Fwd: ����������,����", "Fwd: 01112 si
las", "Fwd: ��������", "Fwd: [ShirtSpace.com (Production)] (NameError) \"uniniti
alized constant Store::UserGroup\"", "Fwd: Market Report", "Fwd: 前程万里-270358
", "Fwd: HOT GIRL Mrs. Elsey Tartaglia is looking for FUN", "Fwd: This is the be
st stock tip of the year", "Fwd: NAUGHTY stories from Mrs. Chere Legare", "Fwd:
����������������������.", "Fwd: NAUGHTY Helge N. and her DIRTY friends are waiti
ng for you", "Fwd: mrvedw����6������������������", "Fwd: cmsy�������������������
```



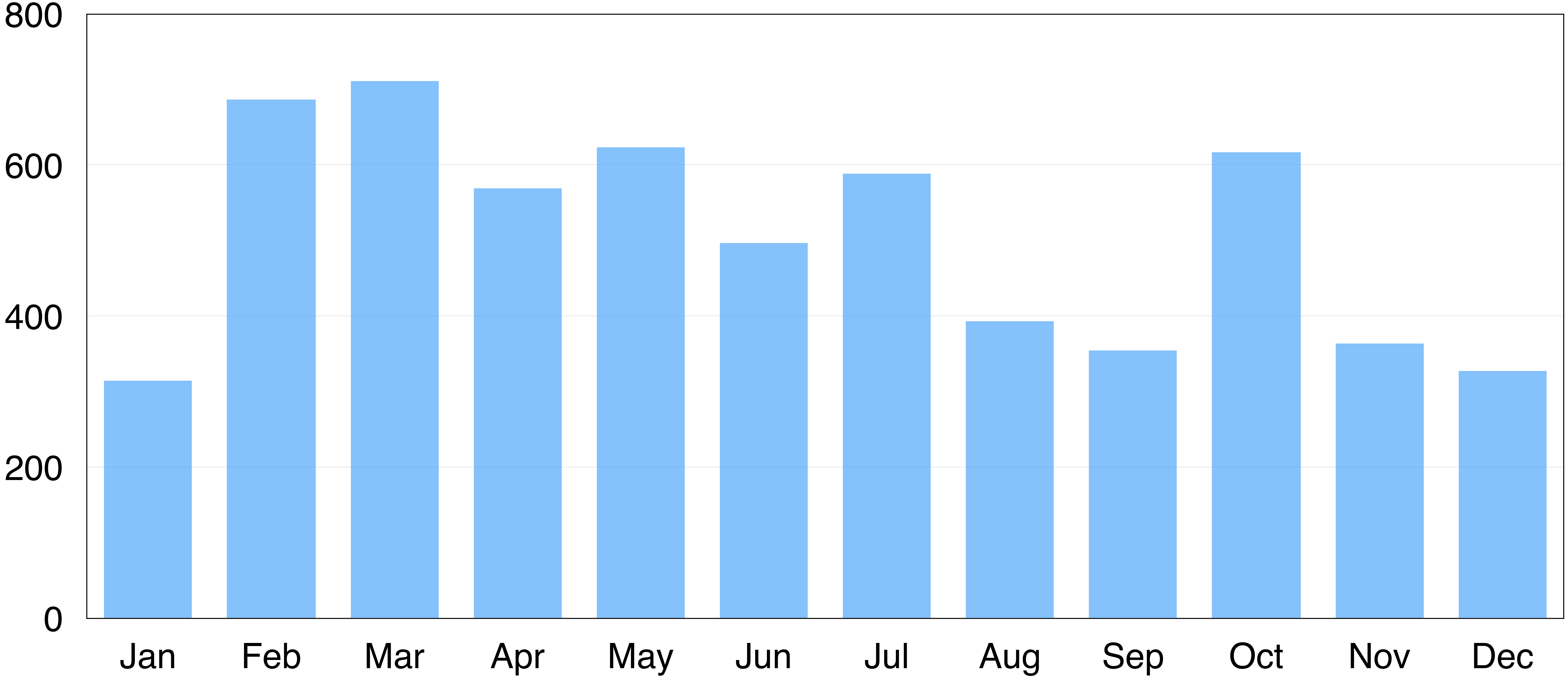
```
unities?", "Fwd: Daniel, let's connect on LinkedIn", "Fwd: 10468 Java Tech Lead-
 PA", "Fwd: Hot Digital Media Startup in SF Bay Seeks RoR CTO and Coders", "Fw:
Networking with ATG/JAVA Professionals - In need of an ATG Developer!", "Fwd: Te
ch Leader needed for Midtown, Early Stage", "Fwd: Ruby Rail Engineer - Oakland,
CA", "Fwd: Want the Best Job Ever? It's Here! :)", "Fwd: Microryza is building o
ut our team!", "Fwd: we are pioneering change! help us build something amazing!"
, "Fwd: Software Engineer - Rails exp", "Fwd: Introduction", "Fwd: Hello from Ch
artboost!", "Fwd: Consultation for freelance job", "Fwd: Linkedin Engineering",
"Fwd: Technical PM Opportunity: Mobile-ize the Cloud!", "Fwd: Working Together",
 "Fwd: San Francisco-based mobile ads cloud distribution start-up - BigData/algo
rithms/ML", "Fwd: GitHub Profile", "Fwd: Interesting opportuity for you", "Fwd:
Job| looking for Full Time C/C++ Software developers for Seattle (WA)", "Fwd: Se
nior Graphics Programmer (#2108-MH2877)", "Fwd: ***DETAILED JOB DESCRIPTION FOR
MOBILE UI DEVELOPER WITH WIPRO***", "Fwd: Conversation with Yelp?", "Fwd: Front-
End Developer Opportunity at Blue Jeans Network", "Fwd: Genomic Hacker at Counsy
l", "Fwd: Experienced Mobile Engineers Desired for Riviera Partner's 100+ Startu
p Client List", "Fwd: Facebook Engineering NYC", "Fwd: Happy holidays!", "Fwd: K
ind Attn:: Microsoft request #38306-1, Web Developer", "Fwd: Are you interested
in a Senior Software Engineer Job in Seattle, WA?", "Fwd: Looking for Software D
eveloper (Contract to hire) CA", "Fwd: Google Checking In", "Fwd: Checking in",
"Fwd: HireStarter", "Fwd: Need Senior Developer / Application Specialist", "Fwd:
 Join my network on LinkedIn", "Fwd: Could you build this", "Fwd: PHP developer
required - Basingstoke - 2 month contract (Likely to extend) - URGENT!", "Fwd: R
eferral opportunity /Mobile Applications Developer / gladly paying referral fee!
", "Fwd: Hot opening for SR iOS Developer __MA", "Fwd: Great Full Time Web Devel
oper W/ Lamp Exp for Prime Client at Greewich, CT", "Fwd: Sr Software Engineer -
 Ruby on Rails ", "Fwd: Noticed your resume - new job open for a Java Software E
ngineer - please apply!", "Fwd: Twitter Engineering", "Fwd: Amazed at who you ar
e - are you free? ", "Fwd: Java lead - Fulltime/Permanent", "Fwd: Opportunity fo
r Python Developer- Level IV-38674 in Irving, TX", "[akaufman@saksys.com: Ruby O
pp in SF]", "Fwd: Greetings from Quidsi [an Amazon.com company].", "Fwd: Job Opp
ortunity - Java Technical Manager - TX & MD", "Fwd: Exciting Startup based in Be
autiful Downtown Palo Alto!", "Fwd: Great Mobile Developer for lucrative Start-u
p Job in Atlanta, GA now open", "Fwd: Rapidly Expanding Company Seeking Rails De
velopers", "Fwd: Senior / Lead Ruby on Rails Developer Opportunity at VC Backed
Startup in Chicago, IL.", "Fwd: Krishna from Mitchell Martin: Java Data Structur
es Developer/NYC NY", "Fwd: Java/J2EE Developer| 6+ months | Madison, WI | Maste
ch", "Fwd: 11210 Oracle /MySQL Database Developer - TX", "Fwd: Direct Client Nee
d for Software Tester role in Houston, TX", "Fwd: FRONT END WEB DEV/UI ROLE, SEA
TTLE, COBALT", "Fw: WIN8 CONSULTANT NEEDED URGENTLY", "Fwd: Come join Pivotal La
```

Start With A Question

What months have the most spam?

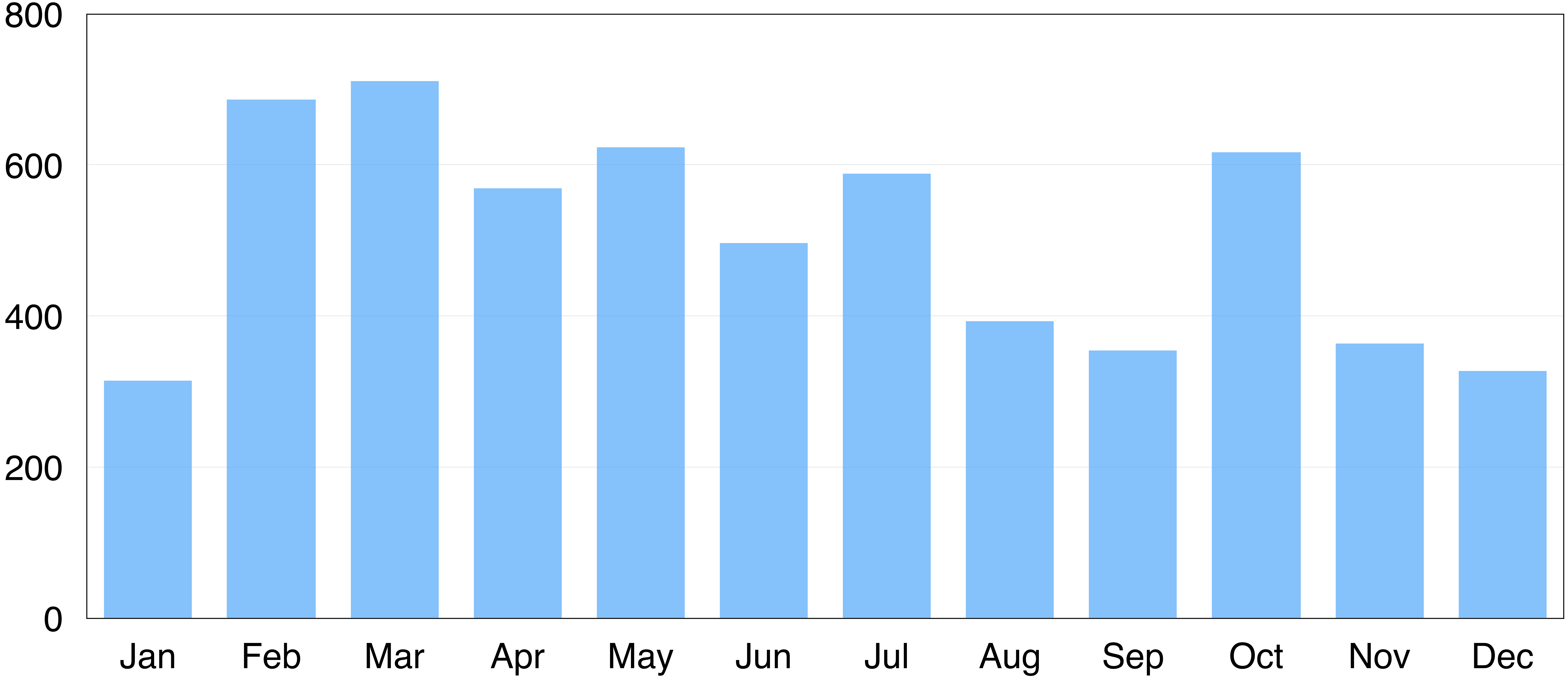
```
> m = Message.all

> months = m.group_by {|x| x.created_at.month}.sort

> counts = months.map {|k, v| v.count}

> puts counts.join("\t")
315  687  711  569  623  497  589  393  355  617  364
327
```

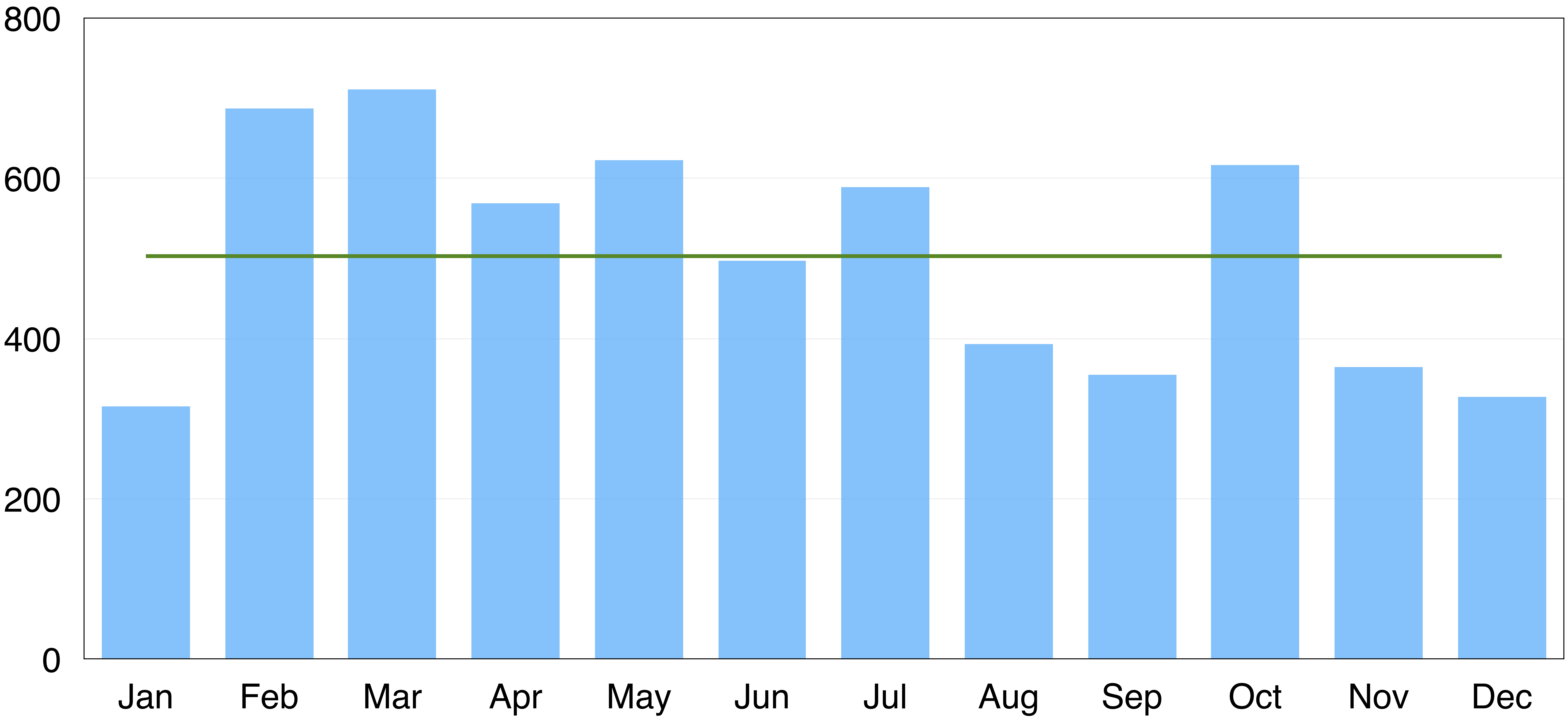
Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct	Nov	Dec
315	687	711	569	623	497	589	393	355	617	364	327


800
600
400
200
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

Start With A Model

Recruiting budgets & targets are set quarterly.


800
600
400
200
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec


800
600
400
200
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

Statistics

Statistics

```
> require 'statsample'

> counts
=> {1=>315, 2=>687, 3=>711, 4=>569,
5=>623, 6=>497, 7=>589, 8=>393, 9=>355,
10=>617, 11=>364, 12=>327}

> v = counts.values.to_scale

> puts v.summary
```

```
= Vector 14
  n :12
  n valid:12
  median: 533.0
  mean: 503.9167
  std.dev.: 146.4547
  std.err.: 42.2778
  skew: -0.0170
  kurtosis: -1.7905
```

Recruiting budgets & targets are set quarterly.

There's more recruiter spam at the beginning of the quarter.

The first month of a quarter has more spam than the other months.

The average spam count in months 1, 4, 7, 10 is greater than the average spam count in the other months.

MSC = Mean Spam Count

$$\text{MSC} \ \forall \ \text{mos.} \in \{1, 4, 7, 10\}$$

$$>$$

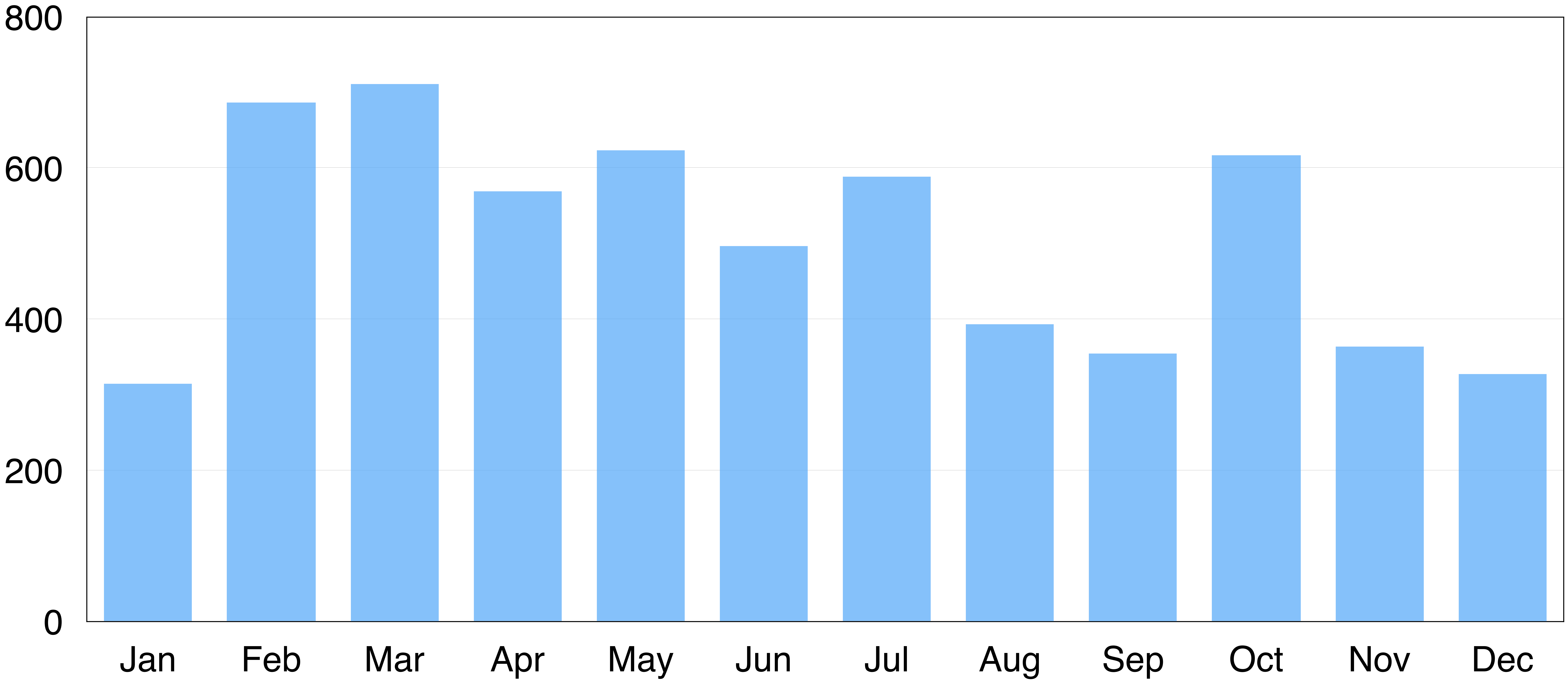
$$\text{MSC} \ \forall \ \text{mos.} \notin \{1, 4, 7, 10\}$$

Comparing Two Means

Mann–Whitney U test

```
> require 'statsample'

> f = counts.select { |k, v| k % 3 == 1 }
> f = f.values.to_vector

> r = counts.reject { |k, v| k % 3 == 1 }
> r = r.values.to_vector

> u = Statsample::Test::UMannWhitney.new(f,r)

> puts u.summary
```

```
+--------------------------------------+-------------------+
| Sum of ranks Vector 1                | 25.000            |
| Sum of ranks Vector 2                | 53.000            |
| U Value                              | 15.000            |
| Z                                    | -0.170 (p: 0.865) |
| Exact p (Dinneen & Blakesley, 1973): | 0.933             |
+--------------------------------------+-------------------+
```


800
600
400
200
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

Other Ideas For Recruiter Spam

Some recruiters are "spammier" than others.

All recruiter emails are basically the same.

What technologies are hot in what regions right now?

Thank you

Aja Hammerly

http://github.com/thagomizer

@thagomizer_rb

http://www.thagomizer.com

SUBSTANTIAL