Predicting across all US households

http://faraday.io

Seamus Abshere, CTO

My talk

- 2013: defining moment
- 2017: what have we learned?
- Future doubts

Let me take you back to 2013…

Our new AI startup in 2013

- Started at grandma's beach house
- Predicting future customers for biggest contractor in Massachusetts
- Demanding buyers:
  - MIT signals engineer
  - University of Chicago econometrics PhD

Our new AI startup in 2013

- Deadline today
- Support vector machines  exploding
- Servers melting
- Millions of people not predicted yet…

Defining moment

Were we…

- Tool experts? (fix the SVM, unmelt servers)
  or
- Domain experts? (focus on data and signal)

We chose to be domain experts.

Bring us back to 2017

Bring us back to 2017

What have we learned?

- Geocode everything
- Postgres database
- Outsource AI tools
- Labels are secret sauce
- Phi not accuracy
- Watch for cheating
- WhizzML is awesome

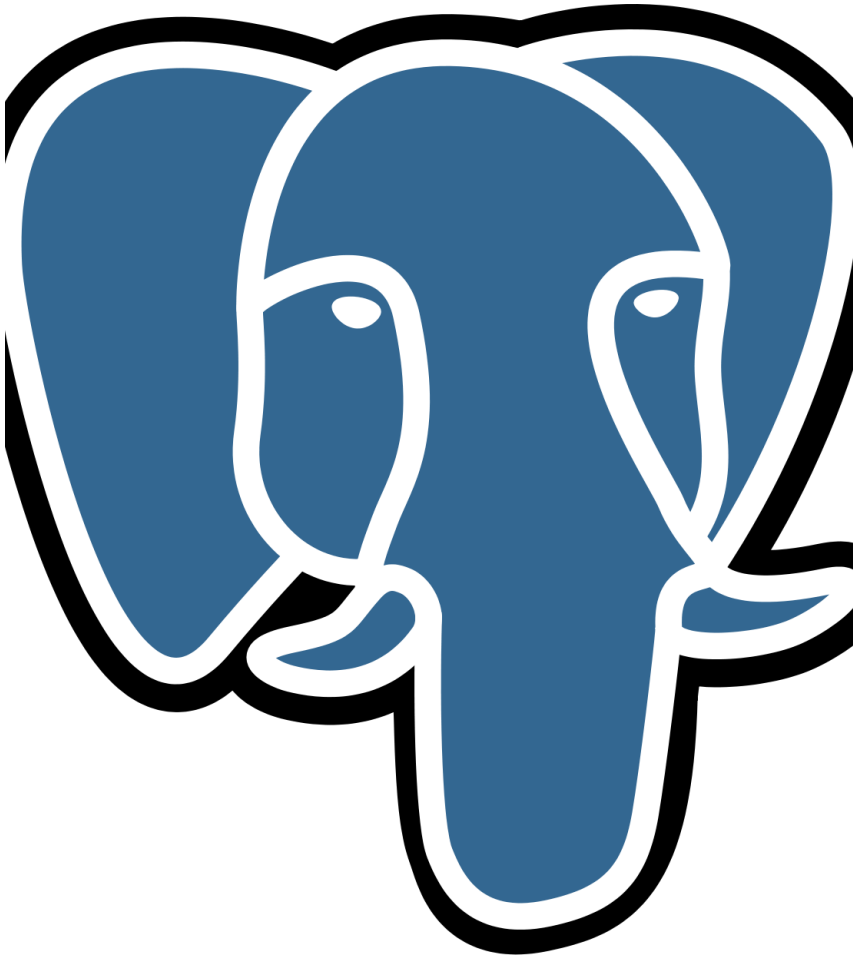
Geocode everything

- Stay sane when you have multiple datasources
- Before
  - Fuzzy match street
  - Pray lat/lons good 🙏
- After
  - Exact match street
  - Fuzzy match person
  - Don't need lat/lon

```
-- loosen the pg_trgm match threshold based on empirical results
-- https://trello.com/c/39fD7BhU/2266-increase-match-rate-by-loosening-name-match
PERFORM set_limit(0.19);

-- determine if we should use full name
SELECT left_record.person IS NOT NULL AND LENGTH(left_record.person) > 1 INTO use_full_nai

-- determine if we should bother with last name checks
SELECT left_record.person_last_name IS NOT NULL AND LENGTH(left_record.person_last_name)

-- best: match on last name and exact hnst
IF use_last_name THEN
  EXECUTE format('SELECT id FROM %I WHERE
          state = $1
      AND city = $2
      AND house_number_and_street = $3
      AND person_last_name %% $4
    ORDER BY
      person_last_name <-> $4 ASC
    LIMIT 1',
    right_table
  ) USING
    left_record.state,
    left_record.city,
    left_record.house_number_and_street,
    left_record.person_last_name
  INTO best_right_record;
  IF best_right_record IS NOT NULL THEN RETURN best_right_record.id; END IF;
END IF;

-- second best: match on full name and exact hnst
IF use_full_name THEN
  EXECUTE format('SELECT id FROM %I WHERE
```

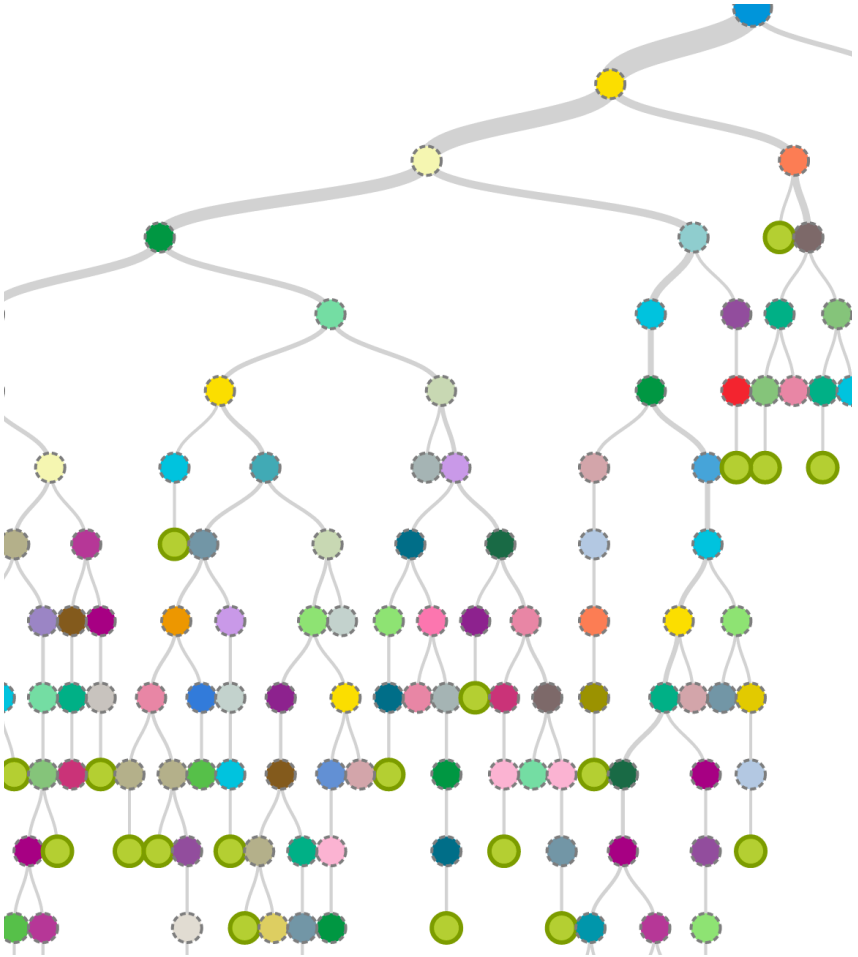
Postgres database

- Don’t mess with inferior databases
- Before
  - MySQL, XLS, etc.
- After
  - Best fuzzy matching
  - Best unstructured fields
  - Best CSV import/export
  - Best user-defined functions
  - 🎉 SQL 🎉



Outsource AI tools

- We chose to be domain experts
- Before
  - Scikit-learn in-house
  - Support Vector Machines (SVM)
- After
  - BigML
  - Decision trees, whatever they recommend



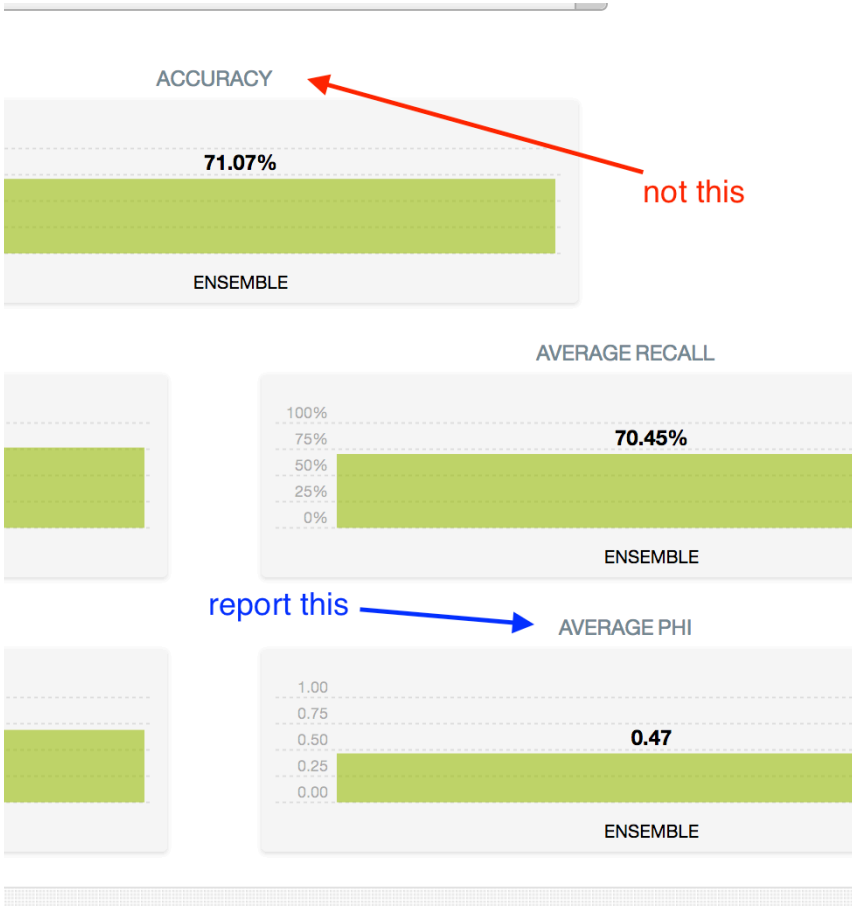
Labels are secret sauce

- Opportunity for creativity and uniqueness
- Before
  - Made purchase = good
  - Don't call me back = bad
- After
  - Per-client goals
  - Time- and sequence-based

```
"goal": "Prevent churn",
"type": "Retention",
"activity": "preventing churn",
"eligible": [{
  "investment": true
}],
"negative_comment": "INVERTED - (old [note 'first' in fil
"negative": [
  [{
    "days_since_first_investment": [730, "Infinity"],
    "days_since_last_rejection": {"not": [0, 730]}
  }]
],
"positive_comment": "INVERTED - rejection AND (no investm
"positive": [{
  "rejection": true,
  "investment_since_rejection": false
}],
```

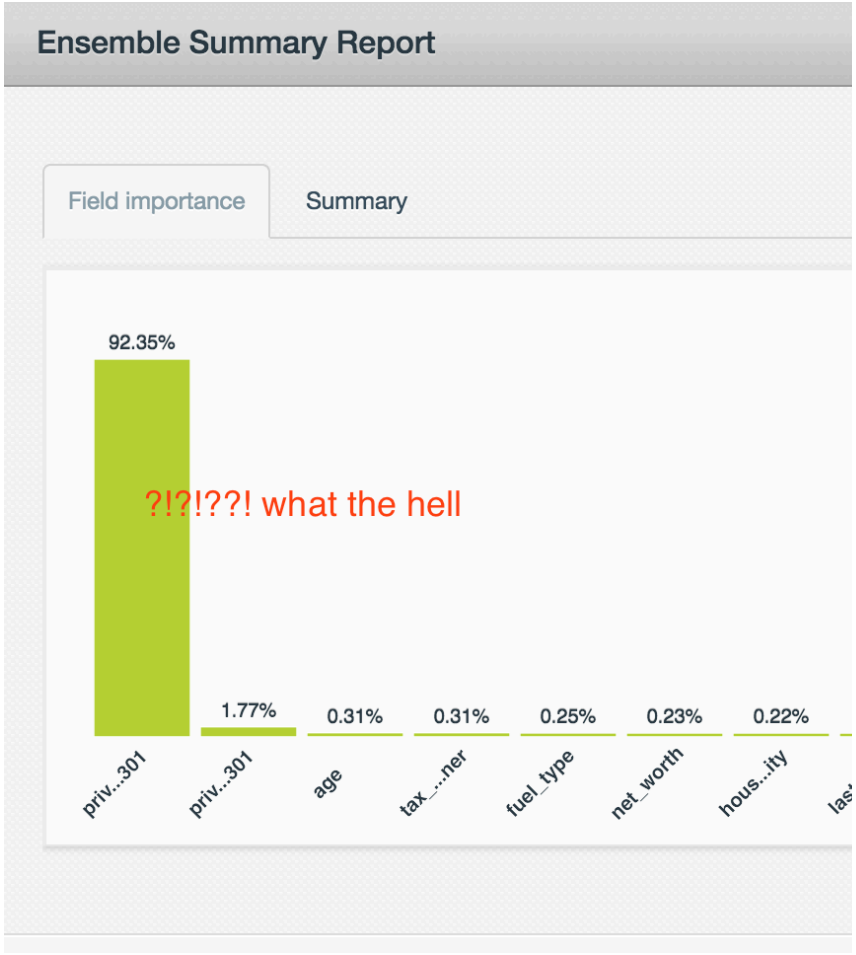
Phi not accuracy

- Use meaningful metrics
- Before
  - Accidentally reported training as test 😳
  - Accuracy
- After
  - Automated train/test
  - Phi
  - Class balancing



Watch for cheating

- If a model is too good to be true, it probably is
- Before
  - so distracted by tools we couldn't focus on context
- After
  - obsessive focus down to the feature level



WhizzML is awesome

- Make the hard network / scaling go away
- Before
  - Polling HTTP
- After
  - Single script

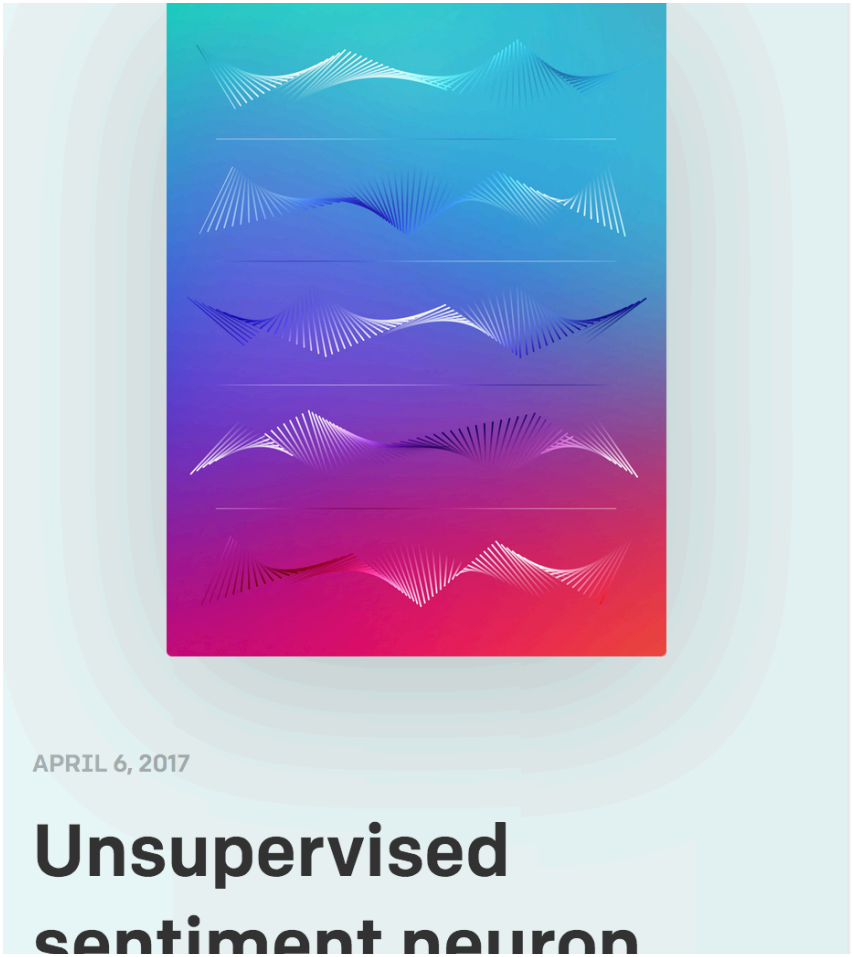
Source code

```
26      (filter (lambda (x) (contains? map2 x)))
27
28  (define (create-split name dataset-id sample
29    (create-dataset
30      {"name" (full-name name)
31       "origin_dataset" dataset-id
32       "sample_rate" sample-rate
33       "out_of_bag" out-of-bag
34       "seed" "1234"}))
35
36  (define goods-unfiltered (create-dataset-fr
37  (define bads-unfiltered (create-dataset-from
38
39  (define fields
40    (list-intersection
41      (fields-with-high-coverage goods-unfilte
42      (fields-with-high-coverage bads-unfilter
43
44  (define goods-all (create-dataset
45    {"name" (full-name "goods-all")
46     "origin_dataset" goods-unfiltered
47     "input_fields" fields }))
48  (define bads-all (create-dataset
49    {"name" (full-name "bads-all")
50     "origin_dataset" bads-unfiltered
```

Future doubts

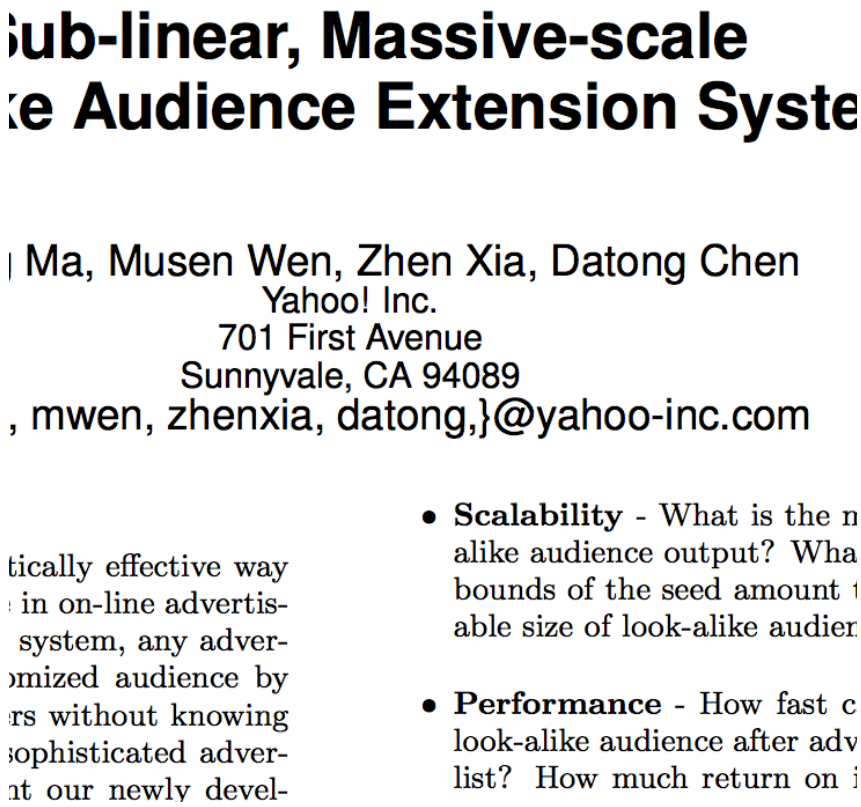
Convolutional / deep / ? neural nets

- All of the exciting announcements are from this space
- I know we'll have to *try* neural nets sometime…
- When will decision trees not be enough?



1 class modeling

- There's more to AI than just classification
- What if we could do more with less?



ub-linear, Massive-scale
e Audience Extension Syste

Ma, Musen Wen, Zhen Xia, Datong Chen
Yahoo! Inc.
701 First Avenue
Sunnyvale, CA 94089
, mwen, zhenxia, datong,}@yahoo-inc.com

tically effective way
in on-line advertis-
system, any adver-
mized audience by
rs without knowing
ophisticated adver-
nt our newly devel-

- **Scalability** - What is the n
alike audience output? Wha
bounds of the seed amount
able size of look-alike audier

- **Performance** - How fast c
look-alike audience after adv
list? How much return on

What do I worry about?

- Are we missing an opportunity to become experts on AI tools?
- Will insight come from our data or the techniques we use?