Predicting across all US households

http://faraday.io

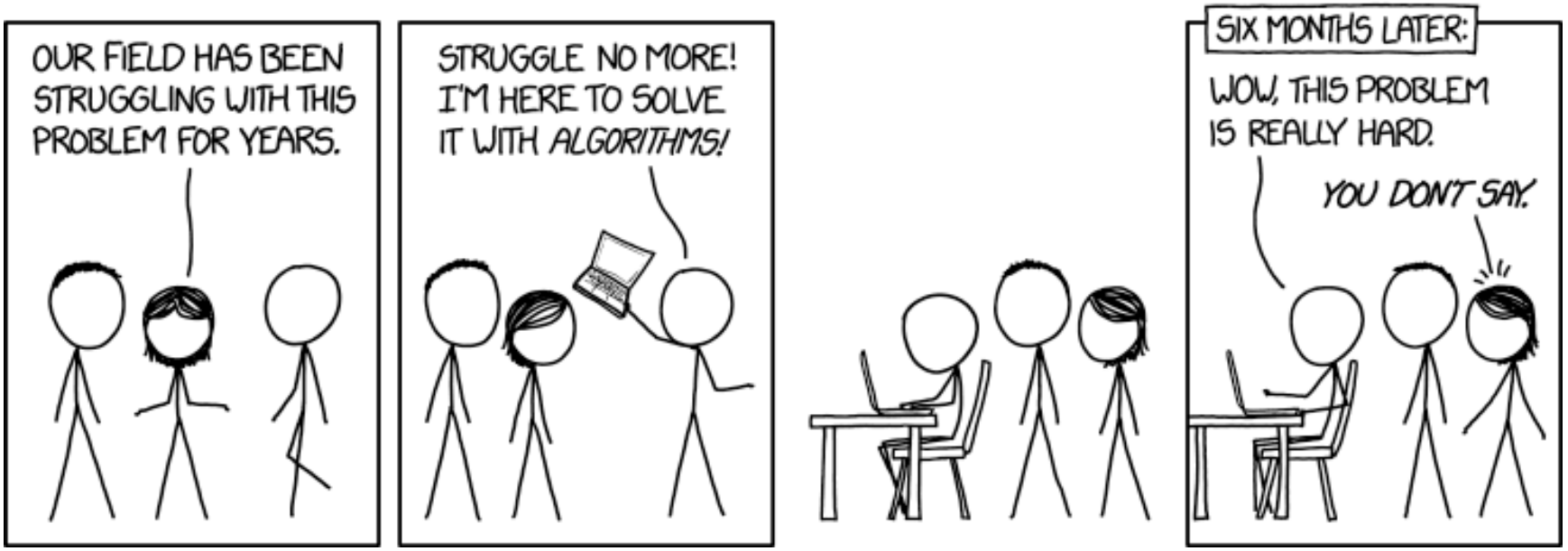
Seamus Abshere, CTO

My talk

- What does it take to create an AI startup?
- Tips & tricks
- Concerns for the future

Faraday

The beginning of Faraday



Faraday

Our new AI startup in 2013

- Problem: who are good customers for home remodeling?

- First customer: biggest contractor in Massachusetts

Faraday

Our new AI startup in 2013

- Manually buying and cleaning data via fax (!)
- Support vector machines  exploding
- Servers melting
- Deadline today…

Faraday

What does it take to be an AI startup?

- Tool experts? (fix the SVM, unmelt servers)

  or

- Domain experts? (focus on data and signal)

Here we are in 2017

Faraday

Here we are in 2017

Faraday

What we did

- Bought databases of all US customers
  - 500+ million records
  - 250+ columns
- Outsourced AI tools

Faraday

What have we learned?

1. Clean data first
2. Geocode everything
3. Use a good database
4. Join data
5. Choose algorithm
6. Scale algorithm
7. Labels are secret sauce
8. Report correctly
9. Watch for cheating
10. Automate workflow

Clean data first

- How do I remove garbage?
- Do it before anything else
- “More expensive data is, the s***** it is”

Faraday

Clean data first

	A	B	C	D	E	F
	Name	**Address**	**City**	**State**	**Gardener**	**Dog breed**
	Shaun Abshere	1038 E Dayton St	Madison	WI	TRUE	beagle
	Joe Rossmeissl	1842 Rutledge St	Madison	WI	~~UNONWN~~	~~UNK~~
	Devin David	123 N Blount St Unit 403	Madison	WI	FALSE	~~NULL~~
	Caity Rose	100 E Main St	Madison	WI	TRUE	pit bull

	A	B	C	D	E	F
	Name	**Address**	**City**	**State**	**Gardener**	**Dog breed**
	Shaun Abshere	1038 E Dayton St	Madison	WI	TRUE	beagle
	Joe Rossmeissl	1842 Rutledge St	Madison	WI		
	Devin David	123 N Blount St Unit 403	Madison	WI	FALSE	
	Caity Rose	100 E Main St	Madison	Wi	TRUE	pit bull

https://github.com/faradayio/scrubcsv
Fixes formatting and removes nulls at 100mb/s

Faraday

Geocode everything

- How do I compare data from multiple sources?

- Need a single source of truth

Faraday

Geocode everything

	A	B	C	D	E	F
	Name	**Address**	**City**	**State**	**Gardener**	**Dog breed**
2	Shaun Absher e	1038 e deyton st.	madison	wi	TRUE	beagle
3	Joe Rossmeis l	1842 RUTLEDGE STREET	MADISON	WI	UNONWN	UNK
4	Devin David	123 North BLOUNT st # 403	MADISON	Wisconsin	FALSE	NULL
5	Caity Rose	100 MainSt	Madison	Wi	TRUE	pit bull

	A	B	C	D	E	F
	Name	**Address**	**City**	**State**	**Gardener**	**Dog breed**
	Shaun Absher e	1038 E Dayton St	Madison	WI	TRUE	beagle
	Joe Rossmeis l	1842 Rutledge St	Madison	WI	UNONWN	UNK
	Devin David	123 N Blount St Unit 403	Madison	WI	FALSE	NULL
	Caity Rose	100 E Main St	Madison	WI	TRUE	pit bull

https://www.npmjs.com/package/smartystreets
3000 records/second - $1000/mo unlimited

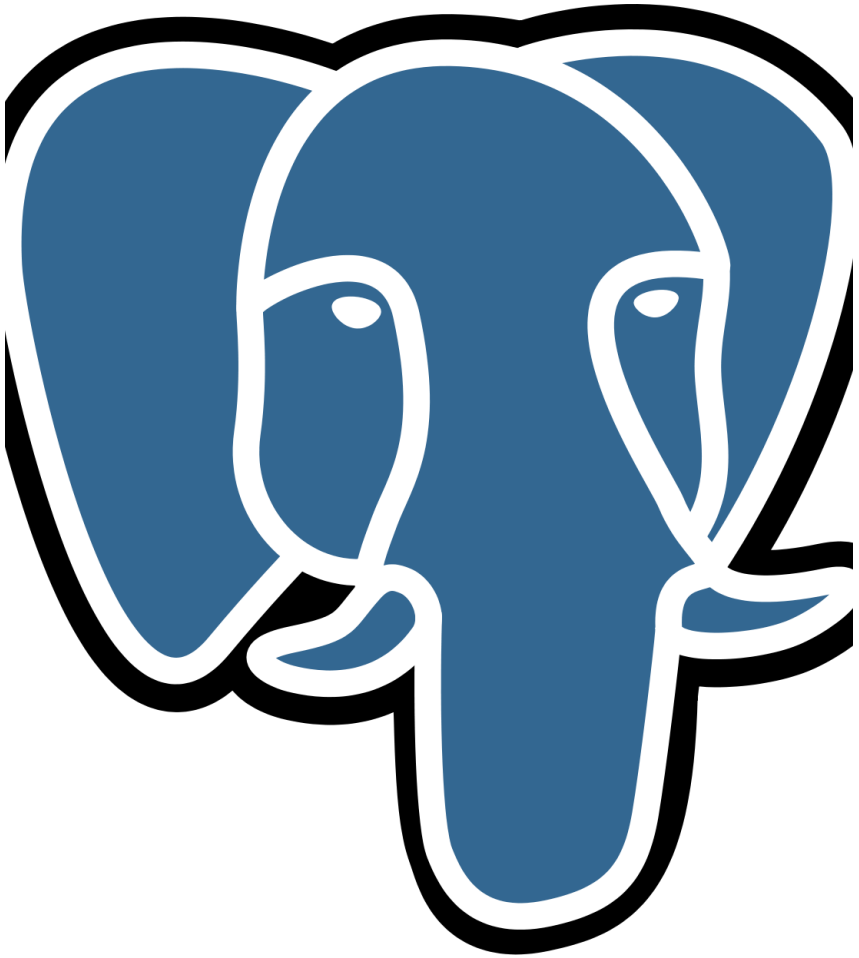
Faraday

Use a good database

- How do I query my data?
- Start simple but powerful

Use a good database

- Try Postgres
- Best fuzzy matching
- Best CSV import/export
- Best user-defined functions
- Best NoSQL support
  - ironic
- Best of all: 🎉 SQL 🎉



Faraday

Join data

- How do I make a single database out of many?
- Fuzzy and exact matching with fallbacks

Join data

- Start with exact matches – these are best
- Fall back to fuzzy
  - N-gram similarity is better than alternatives
- Spatial matching (lat/lon) is slow

```
-- 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 I
END IF;

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

https://www.postgresql.org/docs/9.6/static/pgtrgm.html

Faraday

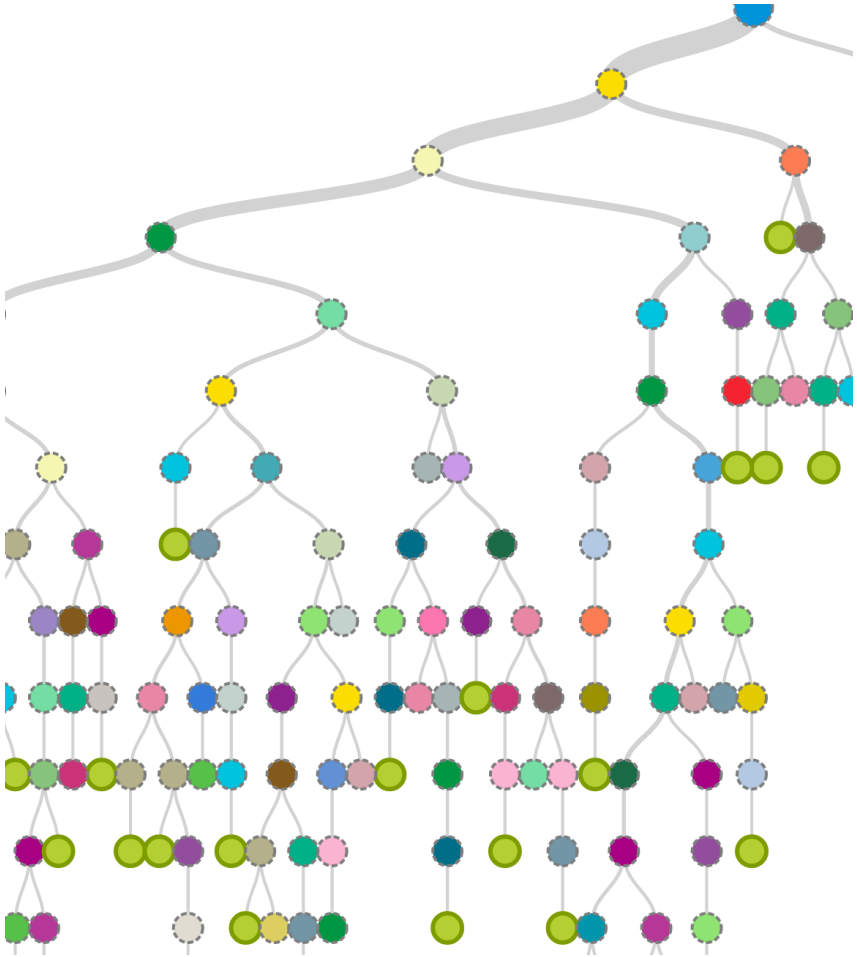
Choose algorithm

- How do I find a signal?

- Lots of confusing options
  - SVM (slow, lots of memory)
  - Decision trees (fast, simple)
  - Neural nets (powerful, complex)

Choose algorithm

- Decision trees are a good default
- Resistant to overfitting
- Fast



Faraday

Scale algorithm

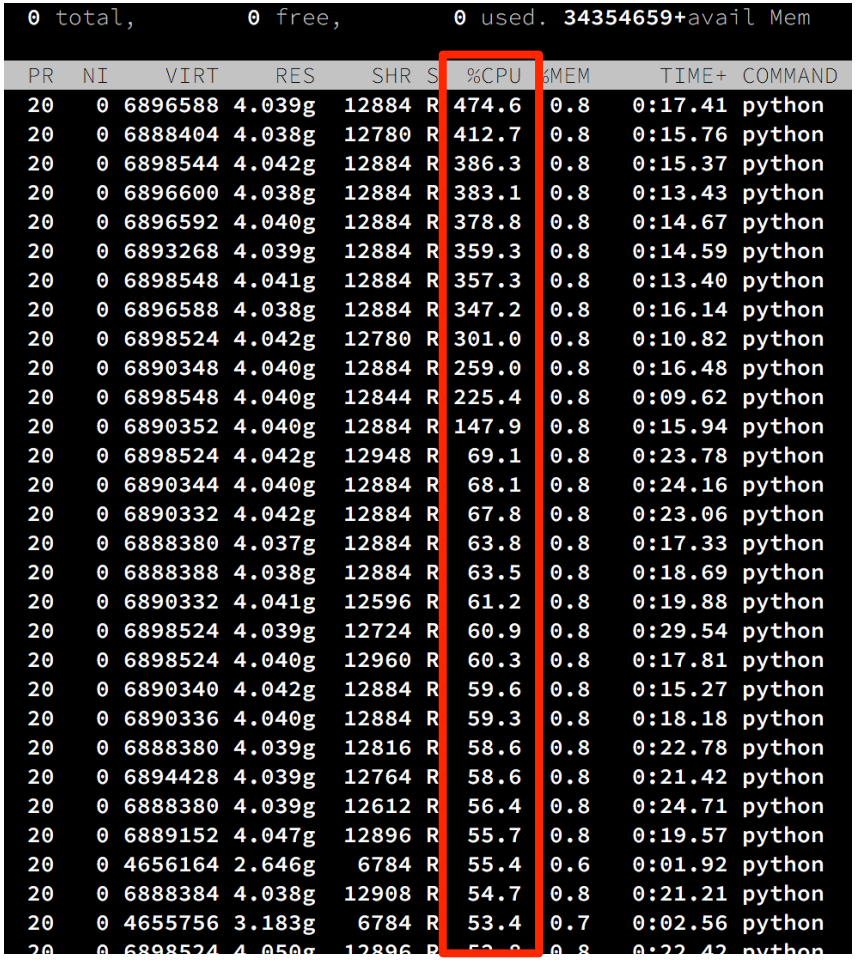
- How do I make predictions across millions of rows?

Scale algorithm

- Buy big servers and parallelize everything

  Or…

- Algorithm-as-a-service
- https://bigml.com/



```
0 total,        0 free,        0 used. 34354659+avail Mem

  PR  NI    VIRT    RES    SHR S  %CPU  %MEM     TIME+ COMMAND
  20   0 6896588 4.039g  12884 R 474.6   0.8   0:17.41 python
  20   0 6888404 4.038g  12780 R 412.7   0.8   0:15.76 python
  20   0 6898544 4.042g  12884 R 386.3   0.8   0:15.37 python
  20   0 6896600 4.038g  12884 R 383.1   0.8   0:13.43 python
  20   0 6896592 4.040g  12884 R 378.8   0.8   0:14.67 python
  20   0 6893268 4.039g  12884 R 359.3   0.8   0:14.59 python
  20   0 6898548 4.041g  12884 R 357.3   0.8   0:13.40 python
  20   0 6896588 4.038g  12884 R 347.2   0.8   0:16.14 python
  20   0 6898524 4.042g  12780 R 301.0   0.8   0:10.82 python
  20   0 6890348 4.040g  12884 R 259.0   0.8   0:16.48 python
  20   0 6898548 4.040g  12844 R 225.4   0.8   0:09.62 python
  20   0 6890352 4.040g  12884 R 147.9   0.8   0:15.94 python
  20   0 6898524 4.042g  12948 R  69.1   0.8   0:23.78 python
  20   0 6890344 4.040g  12884 R  68.1   0.8   0:24.16 python
  20   0 6890332 4.042g  12884 R  67.8   0.8   0:23.06 python
  20   0 6888380 4.037g  12884 R  63.8   0.8   0:17.33 python
  20   0 6888388 4.038g  12884 R  63.5   0.8   0:18.69 python
  20   0 6890332 4.041g  12596 R  61.2   0.8   0:19.88 python
  20   0 6898524 4.039g  12724 R  60.9   0.8   0:29.54 python
  20   0 6898524 4.040g  12960 R  60.3   0.8   0:17.81 python
  20   0 6890340 4.042g  12884 R  59.6   0.8   0:15.27 python
  20   0 6890336 4.040g  12884 R  59.3   0.8   0:18.18 python
  20   0 6888380 4.039g  12816 R  58.6   0.8   0:22.78 python
  20   0 6894428 4.039g  12764 R  58.6   0.8   0:21.42 python
  20   0 6888380 4.039g  12612 R  56.4   0.8   0:24.71 python
  20   0 6889152 4.047g  12896 R  55.7   0.8   0:19.57 python
  20   0 4656164 2.646g   6784 R  55.4   0.6   0:01.92 python
  20   0 6888384 4.038g  12908 R  54.7   0.8   0:21.21 python
  20   0 4655756 3.183g   6784 R  53.4   0.7   0:02.56 python
  20   0 6898524 4.050g  12896 R  53.0   0.8   0:22.42 python
```

Faraday

Labels are secret sauce

- What is *interesting* about your data?

- Try classifying it in many ways

Labels are secret sauce

- Simple
  - Made purchase = good
  - Don't call me back = bad
- More advanced
  - 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
}],
```

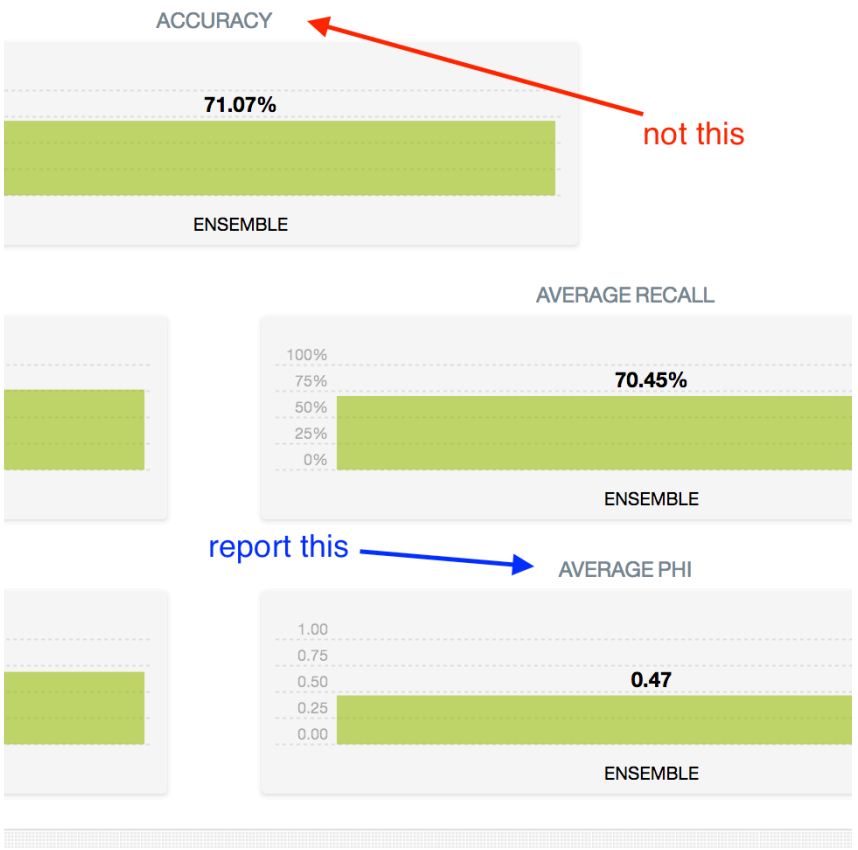
Faraday

Report correctly

- How can I avoid embarrassing myself?
- Don’t be fooled by 99% accuracy

Report correctly

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



Faraday

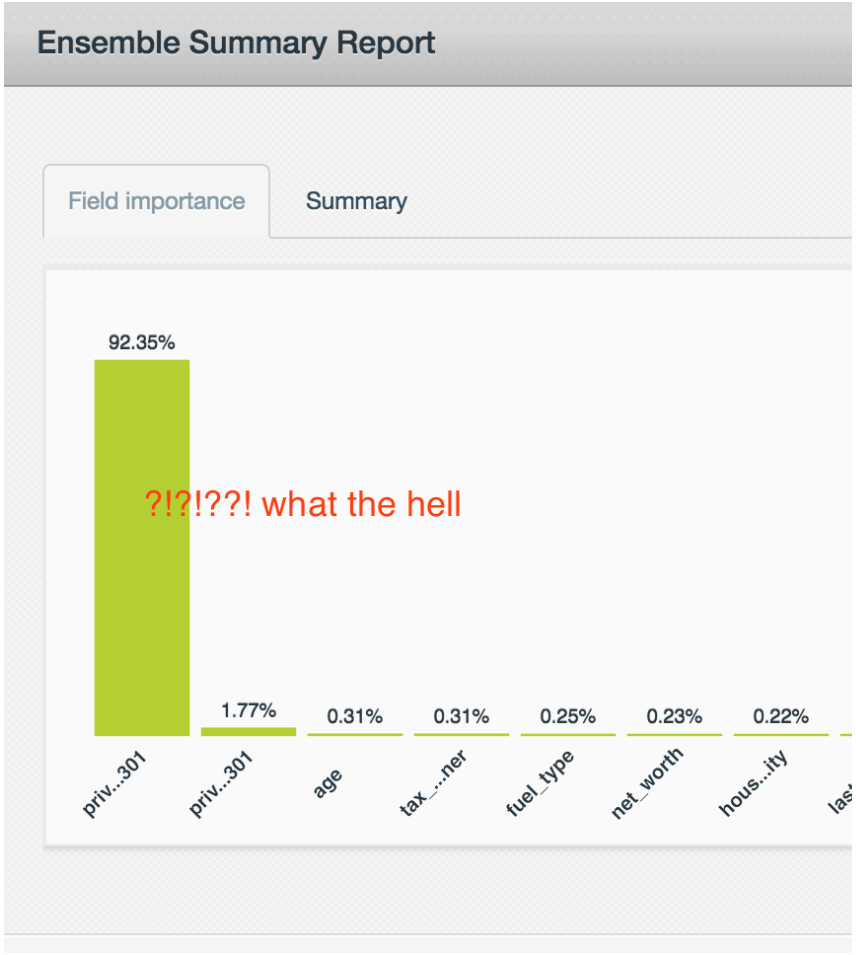
Watch for cheating

- How do I know if I found a signal?

- Be paranoid

Watch for cheating

- Domain expertise
  - If predicting solar customer, cheating based on tax incentives in property data?
- Map it out
  - All goods in one area, all bads in another?



Faraday

Automate workflow

- How do I make predictions in production?

- Automate or die

Automate workflow

- BigML provides WhizzML
- 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
```

Faraday

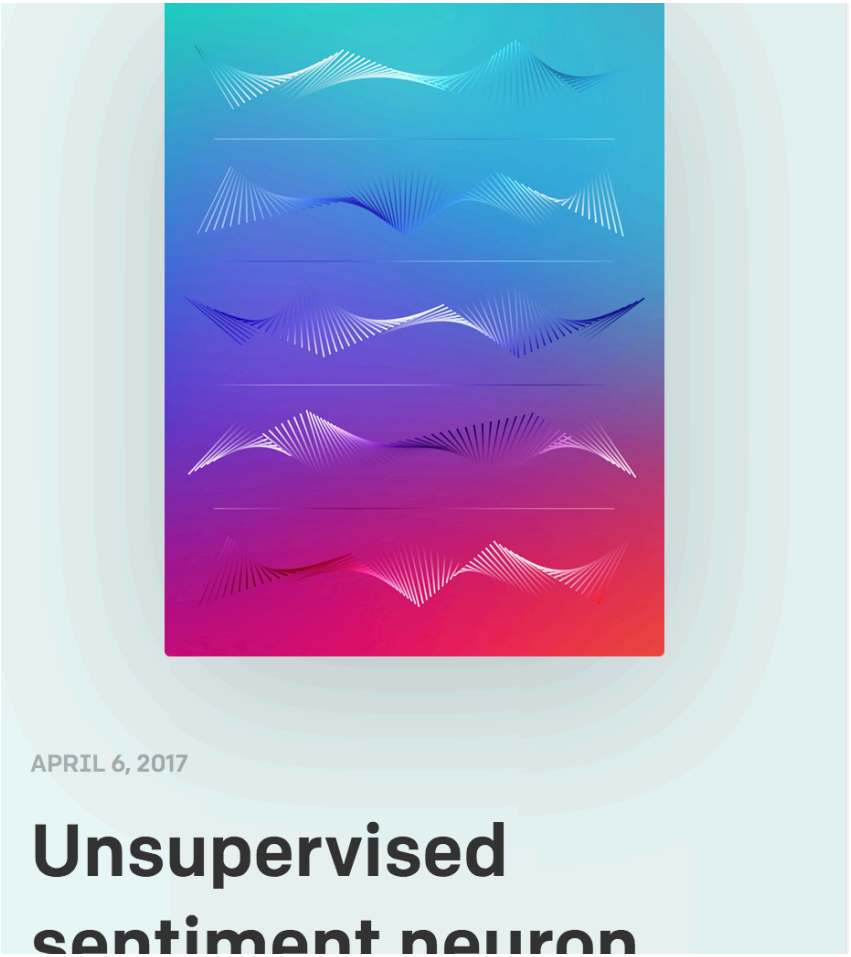
Concerns for the future

Faraday

Convolutional / deep / ? neural nets

- All of the exciting announcements are about neural networks

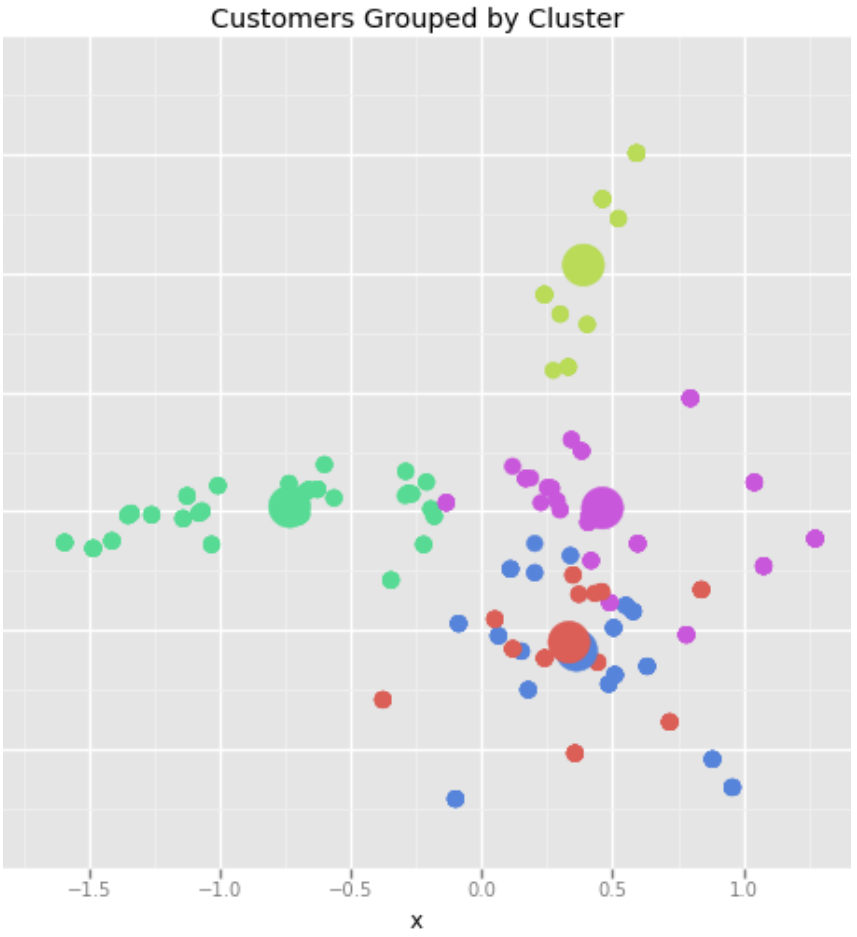
- When are you limited by your choice of algorithm?



Faraday

Beyond classification

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



Faraday

What do I worry about?

- Tool experts?

  or

- Domain experts?