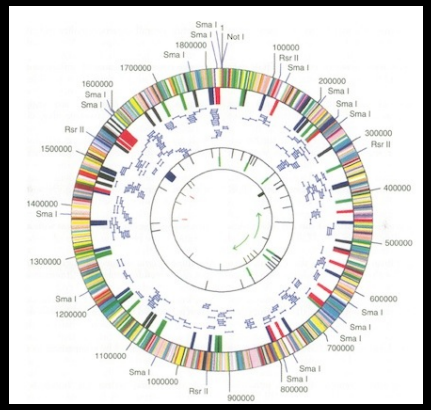
# what is a computational biologist doing at the New York Times?

## (and what can academia do for a 163-year old company?)



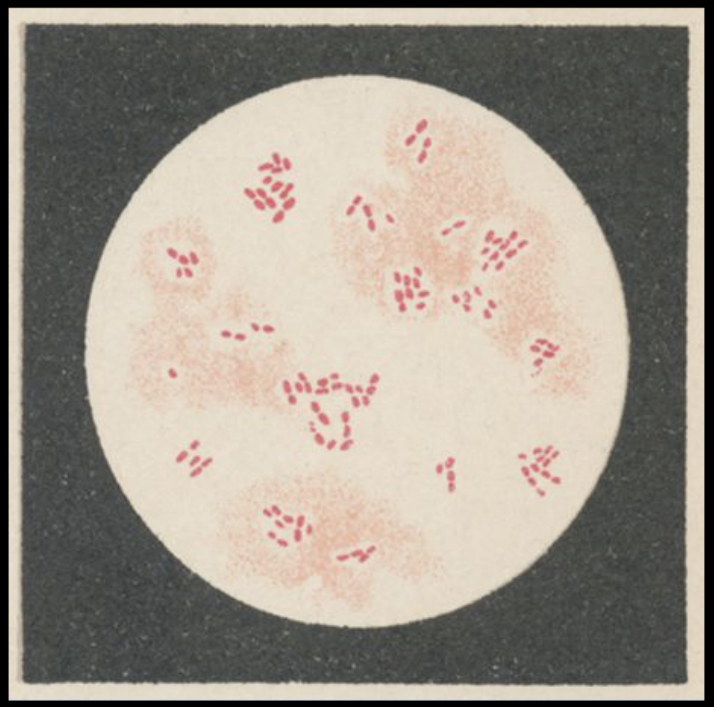

chris.wiggins@columbia.edu
chris.wiggins@nytimes.com
chris.wiggins@hackNY.org
@chrishwiggins

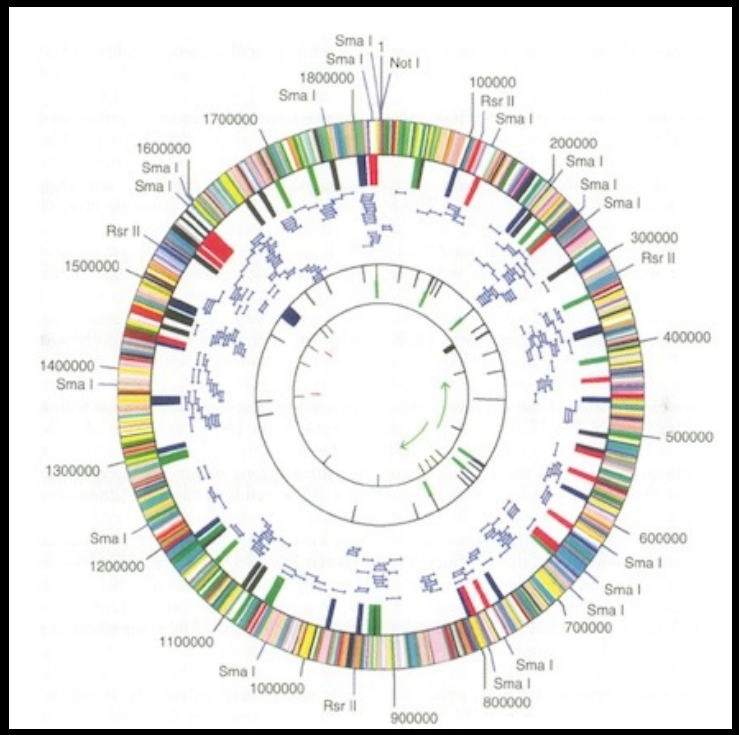
# context/background

# context/background

(before ‘the talk’)

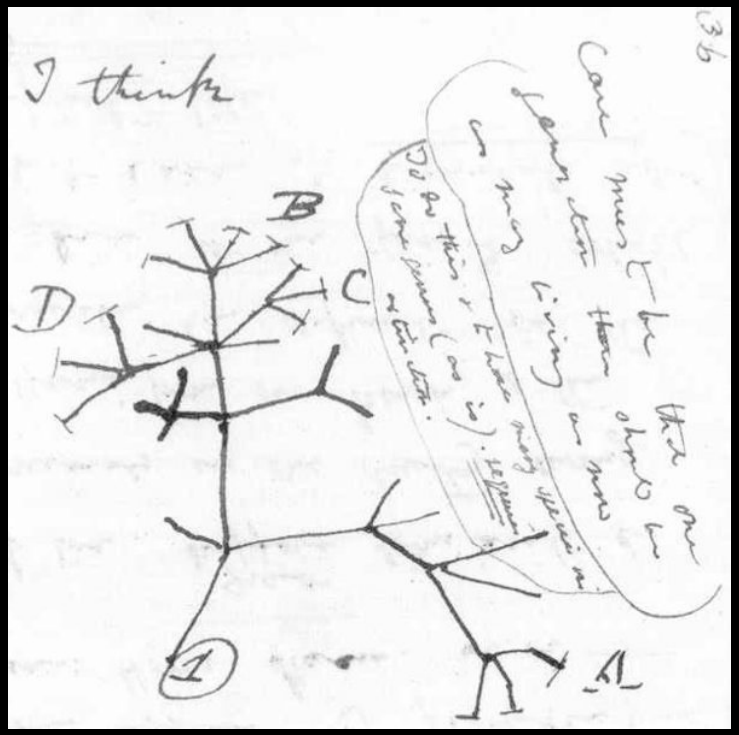
# biology: 1892 vs. 1995

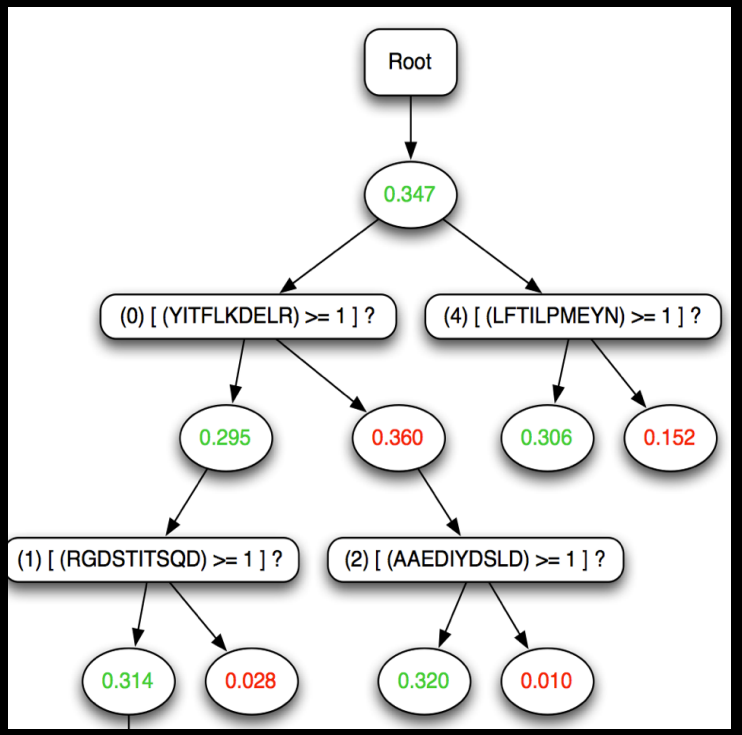





# biology changed for good.

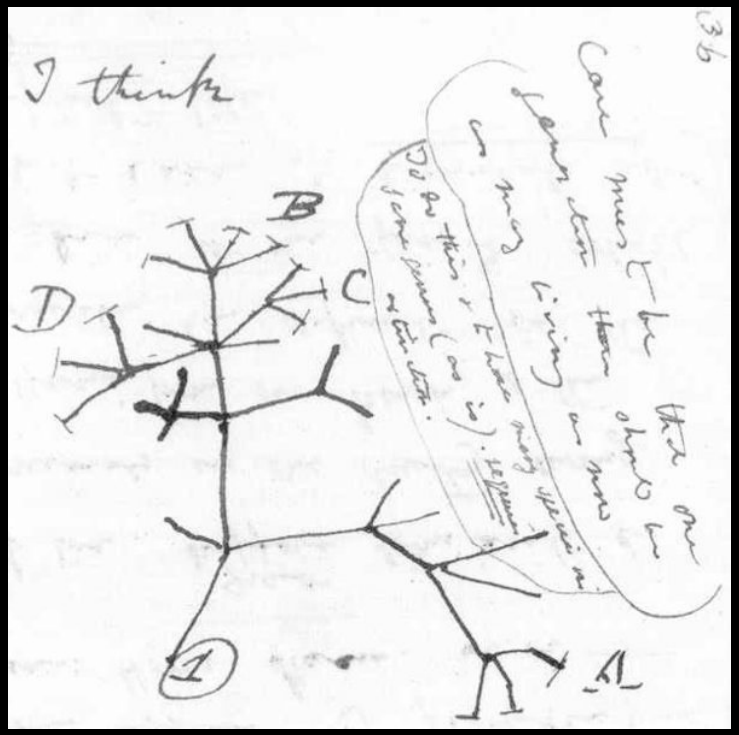
# genetics: 1837 vs. 2012

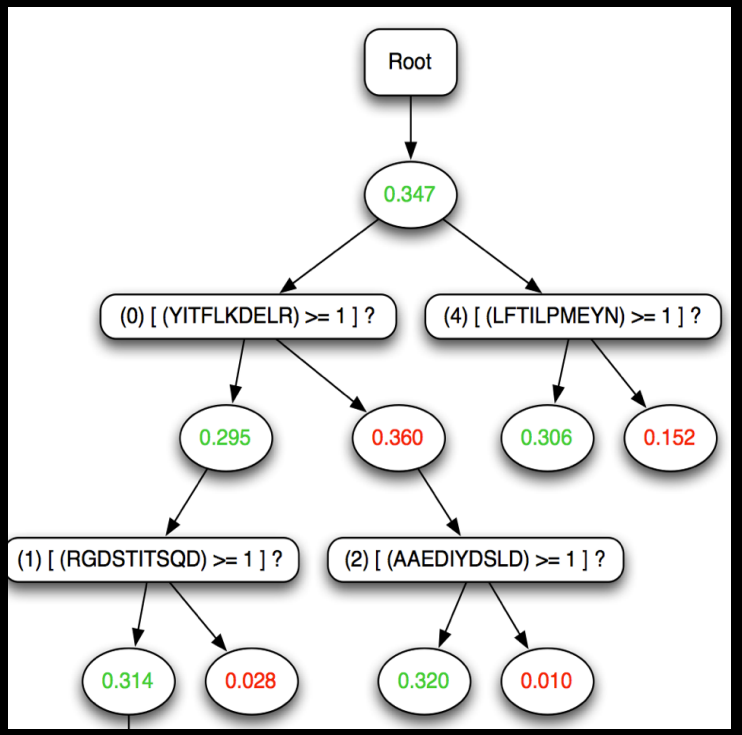



# from “segments” to algorithms

# genetics: 1837 vs. 2012




# from intuition to prediction

data science: web scale

example:

163 yr old

The New York Times

# Reshaping New York

bit.ly/nyt-interactive-2013

From buildings to bike lanes to painting over Broadway, how the city changed in 12 years of Bloomberg

**Begin the tour**

# R+D: nytlabs.com

# developer.nytimes.com: 2008

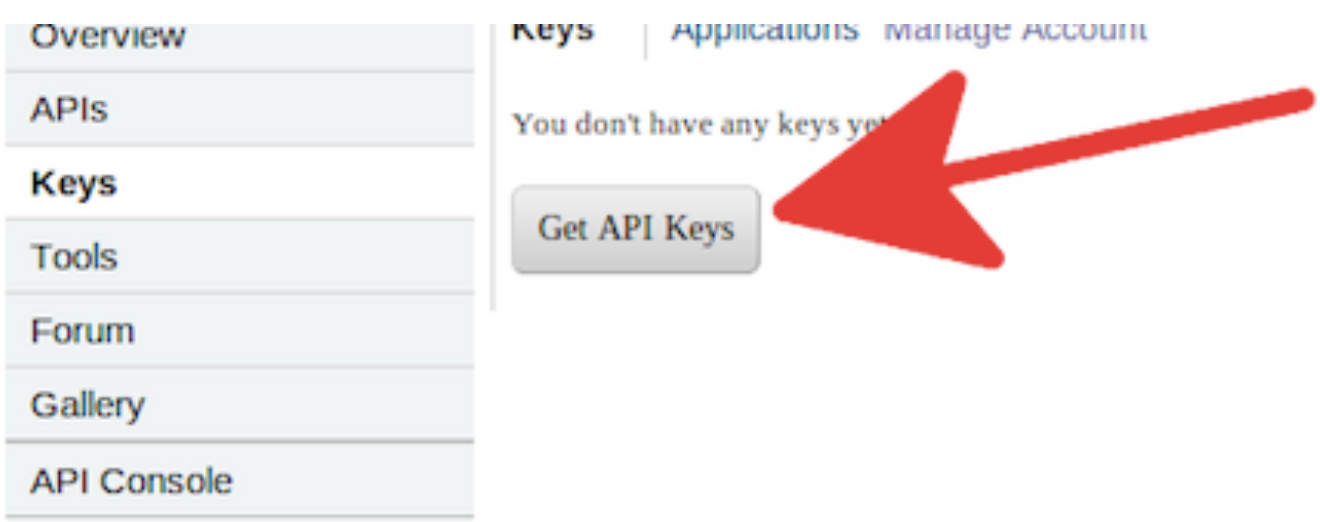
## The New York Times Releases Its First API

**Andres Ferrate**, October 15th, 2008 Commen

After much anticipation, The New York Times has released its first API: a Campaign Finance API that allows developers to retrieve contribution and expenditure data based on United States Federal Election Commission filings (our New York T Campaign Finance API profile). The API is part of the new Times Developer Network, which wi eventually give developers access to several APIs.

As explained in The New York Times Blog:

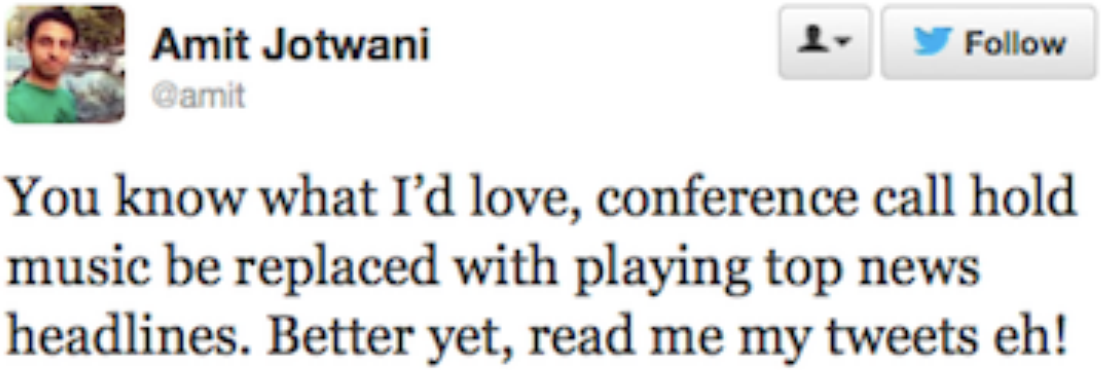
> *The initial version of the Campaign Finance API offers overall figures for presidential candidates, as well as state-by-state and ZIP code totals for specific candidates. In addition, the API supports a contributor name search using any of the following parameters: first name, last name and ZIP code.*




# Replacing Hold Music With New York Times Headlines

Posted by Rob Spectre on July 31, 2013 in Tips, Tricks & Sample Code

So I was straight chillin' on the Internet, fixing to moderate another troll post on my favorite Justin Bieber fan forum, when my buddy Amit posts on Twitter:



www.nytimes.com/newsgraphics/2014/senate-model/

TheUpshot Senate Forecasts

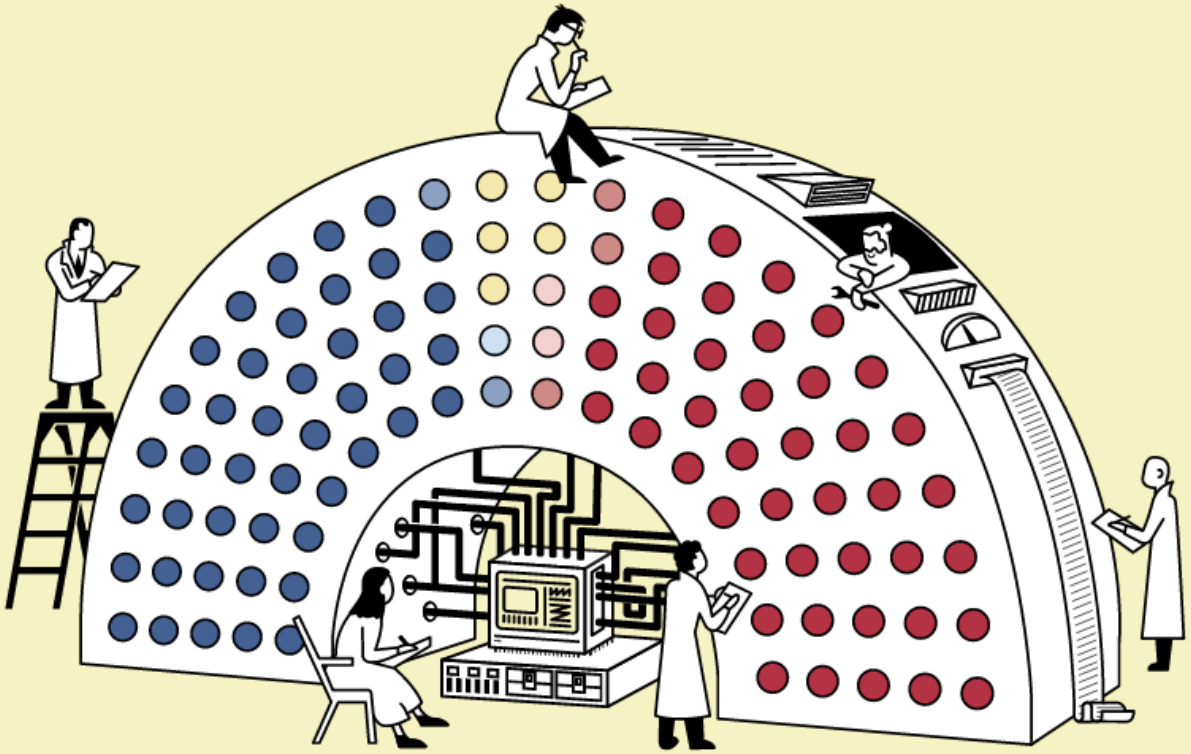
Overview Other Forecasts About

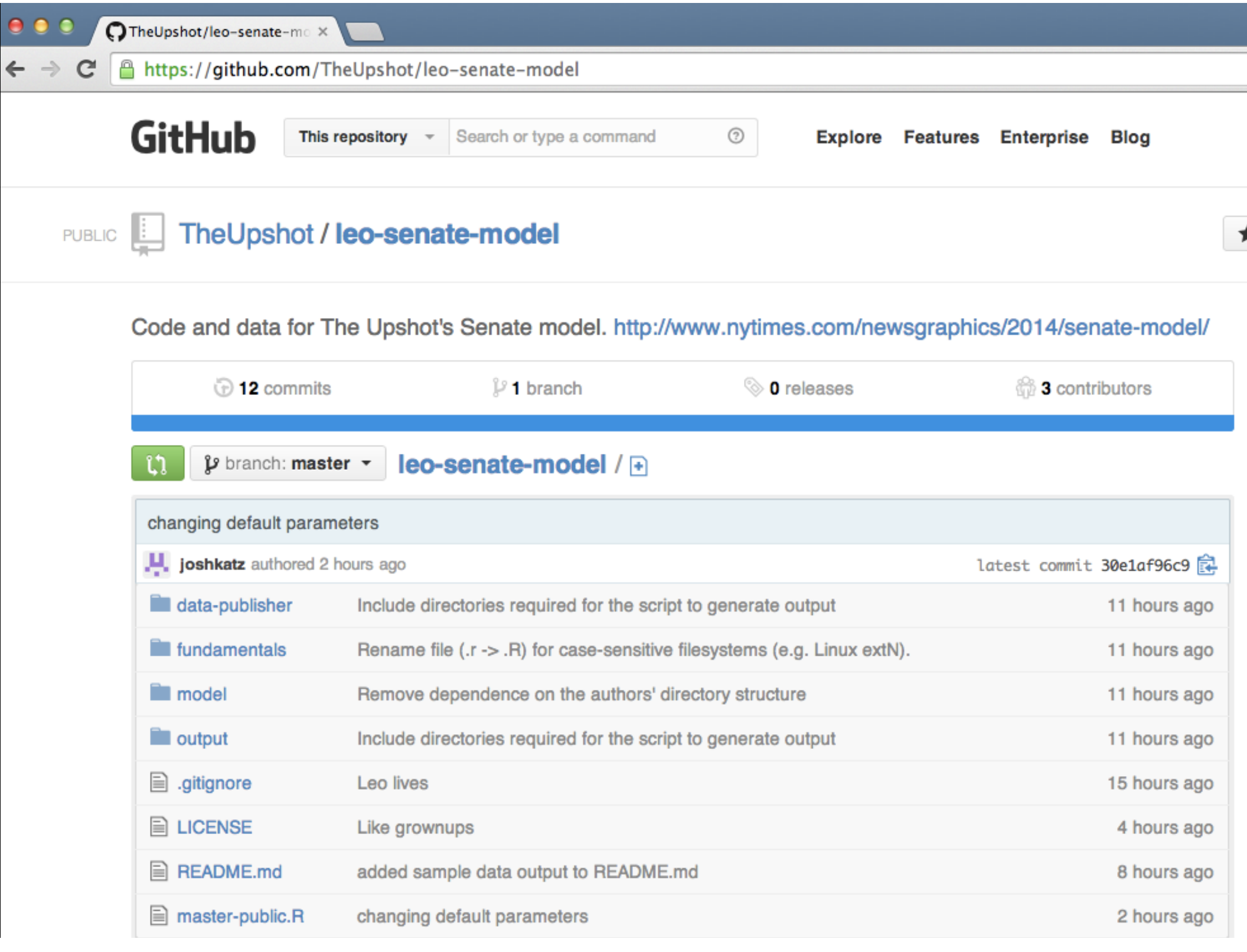
# Who Will Win The Senate?

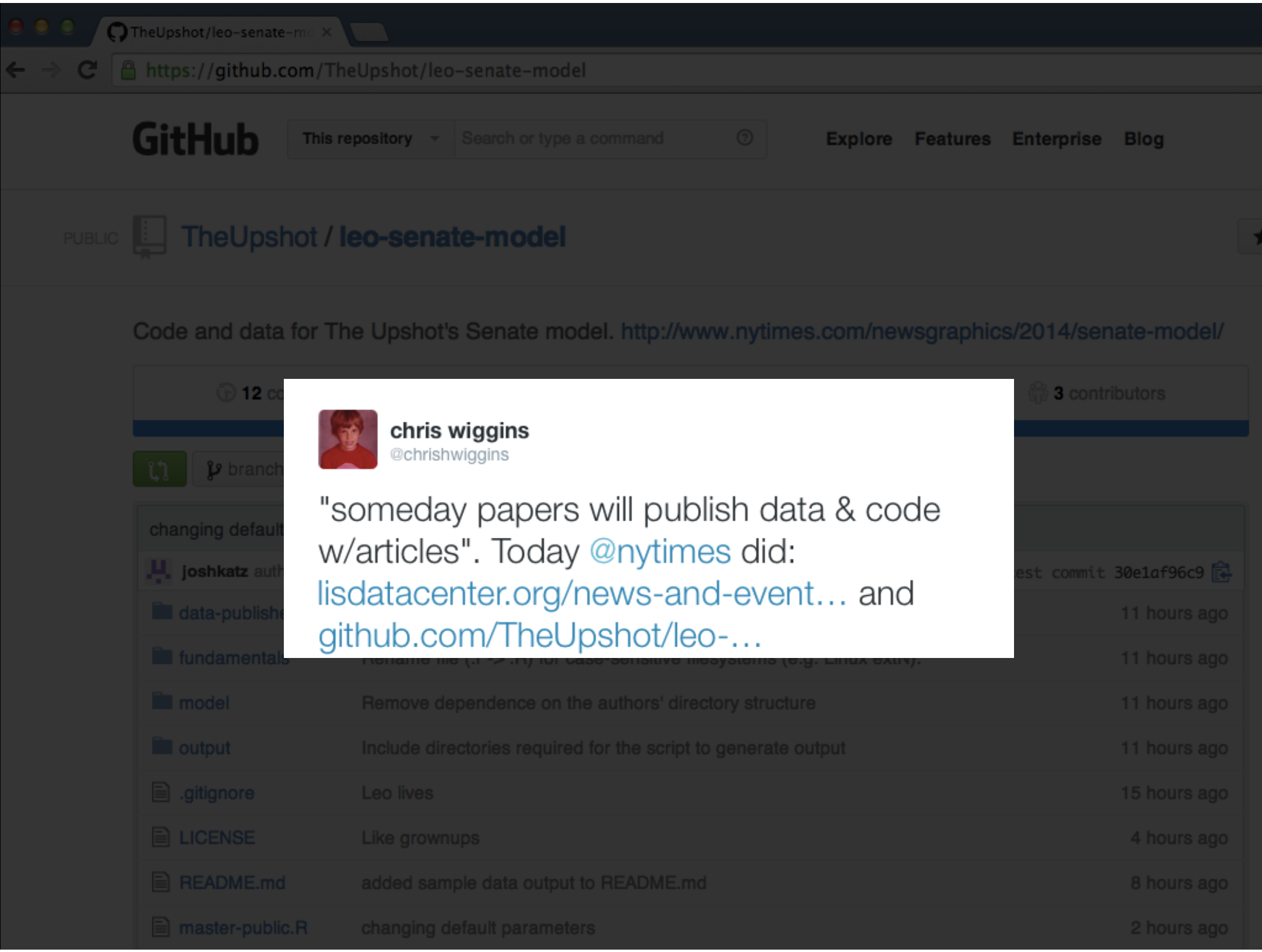
According to our statistical election-forecasting machine, it's a **tossup**. The Democrats have about a 51% chance of retaining a majority.

Tweet or Share on Facebook

Last updated Tuesday, April 22 at 2:30 AM EDT.




TheUpshot/leo-senate-mo
https://github.com/TheUpshot/leo-senate-model
GitHub
This repository
Search or type a command
Explore
Features
Enterprise
Blog
PUBLIC
TheUpshot / leo-senate-model
Code and data for The Upshot's Senate model. http://www.nytimes.com/newsgraphics/2014/senate-model/
12 commits
1 branch
0 releases
3 contributors
branch: master
leo-senate-model /
changing default parameters
joshkatz authored 2 hours ago
latest commit 30e1af96c9
data-publisher Include directories required for the script to generate output 11 hours ago
fundamentals Rename file (.r -> .R) for case-sensitive filesystems (e.g. Linux extN). 11 hours ago
model Remove dependence on the authors' directory structure 11 hours ago
output Include directories required for the script to generate output 11 hours ago
.gitignore Leo lives 15 hours ago
LICENSE Like grownups 4 hours ago
README.md added sample data output to README.md 8 hours ago
master-public.R changing default parameters 2 hours ago

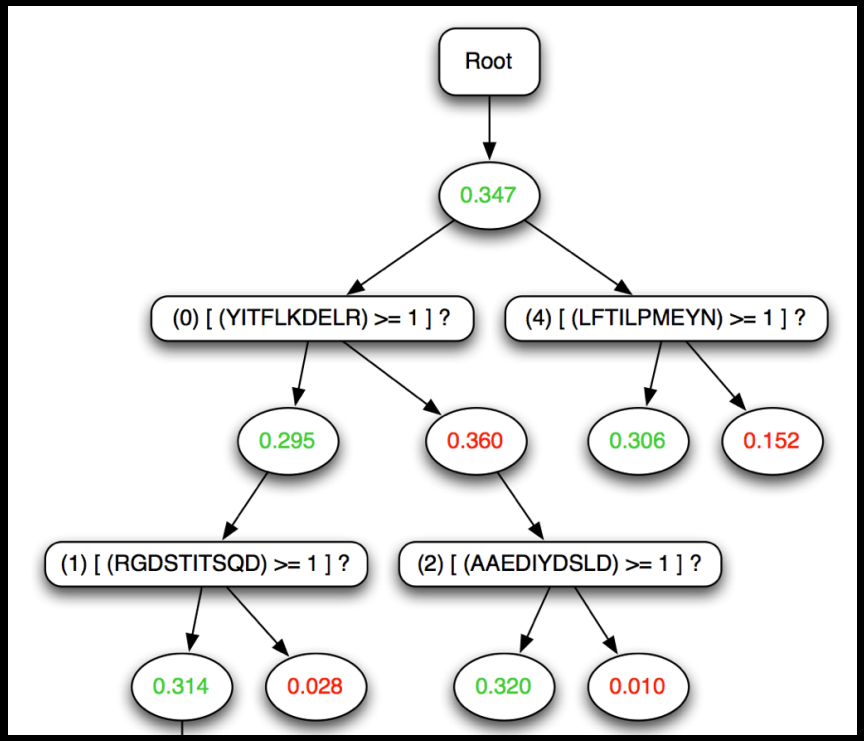

TheUpshot/leo-senate-mo
https://github.com/TheUpshot/leo-senate-model
GitHub
This repository
Search or type a command
Explore
Features
Enterprise
Blog
PUBLIC
TheUpshot / leo-senate-model
Code and data for The Upshot's Senate model. http://www.nytimes.com/newsgraphics/2014/senate-model/
12 co
3 contributors
branch
changing default
joshkatz auth
est commit 30e1af96c9
data-publishe
11 hours ago
fundamentals
11 hours ago
model
Remove dependence on the authors' directory structure
11 hours ago
output
Include directories required for the script to generate output
11 hours ago
.gitignore
Leo lives
15 hours ago
LICENSE
Like grownups
4 hours ago
README.md
added sample data output to README.md
8 hours ago
master-public.R
changing default parameters
2 hours ago
chris wiggins
@chrishwiggins
"someday papers will publish data & code w/articles". Today @nytimes did: lisdatacenter.org/news-and-event… and github.com/TheUpshot/leo-…


The lobby of The New York TImes Building/Nic Lehoux

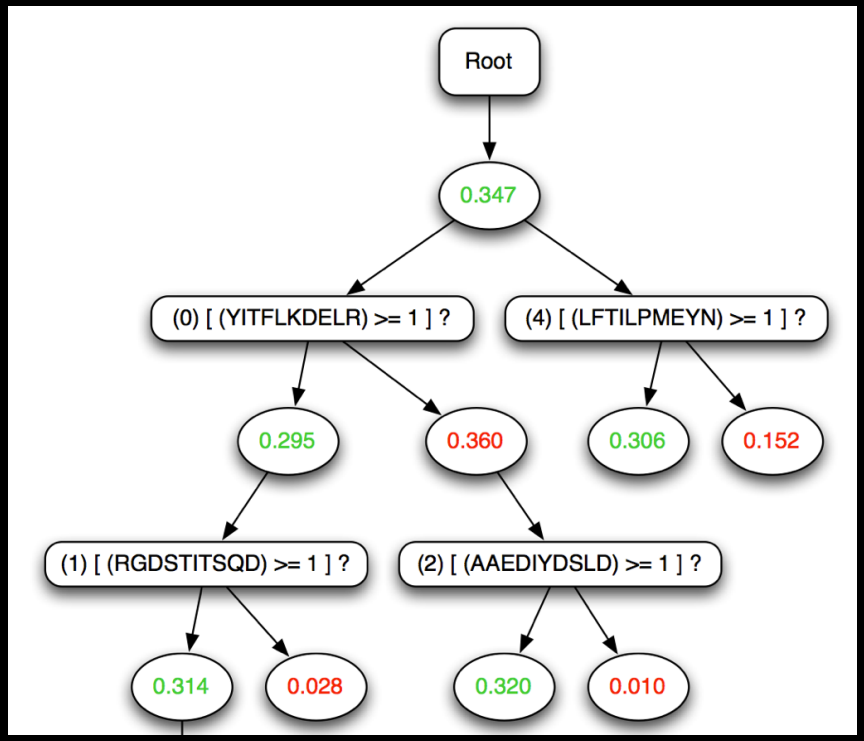
**01/09/2014**
The New York Times Company to Webcast Fourth-Quarter and Full-Year 2013 Earnings Conference Call »

**01/02/2014**
The New York Times to Introduce Redesign of NYTimes.com to All Users Jan. 8 »

**12/12/2013**
The New York Times Company Declares Regular Quarterly Dividend »




# from “segments” to algorithms




# from intuition to prediction

# data science: the web

data science: the web

is your “online presence”

data science: the web

is a microscope

data science: the web

is an experimental tool

data science: the web

is an optimization tool

</header>

</header>

i.e., <body>

common requirements
in data science:

common requirements
in data science:

1. practices

common requirements
in data science:

1. practices
2. skills

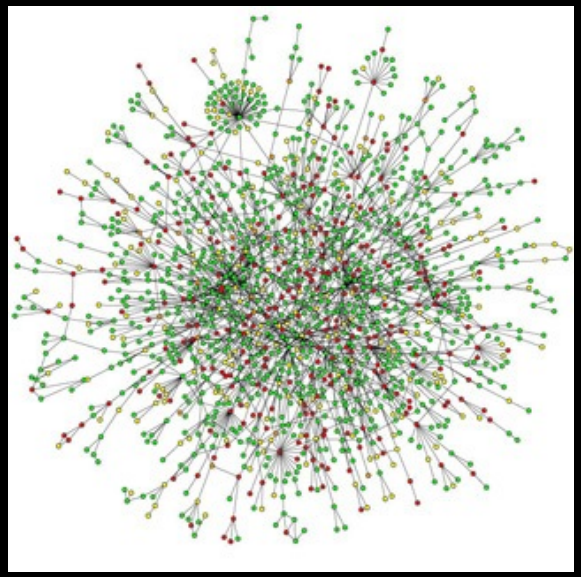
common requirements
in data science:

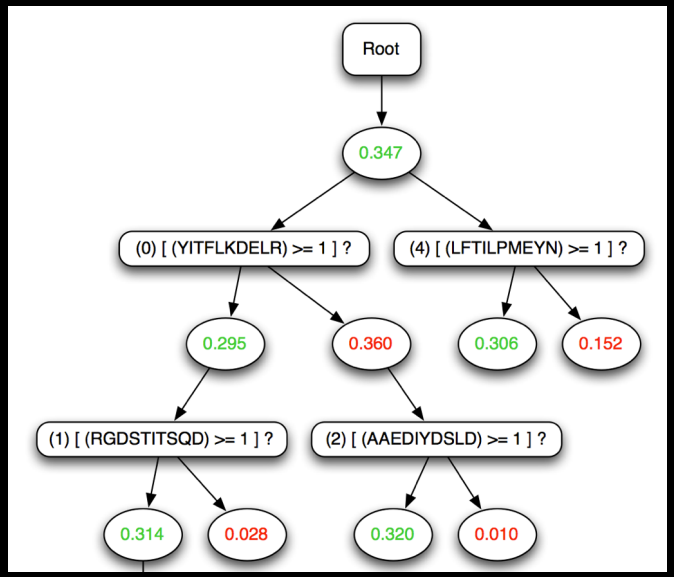
1. practices
2. skills
3. culture

# data science: practice

# data science: practice

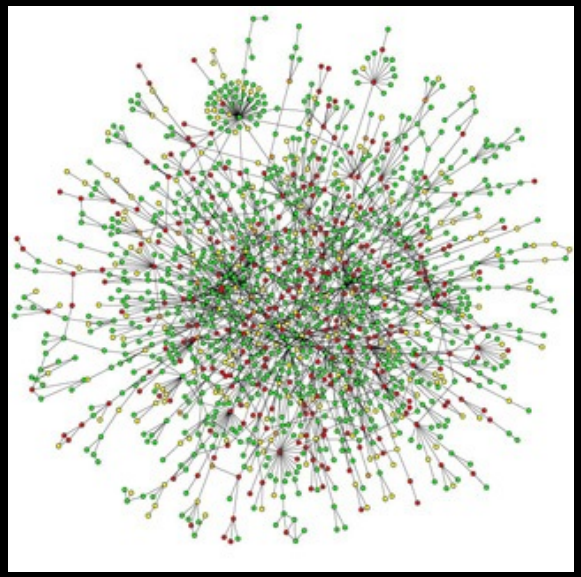
- reframe domain questions
as machine learning tasks

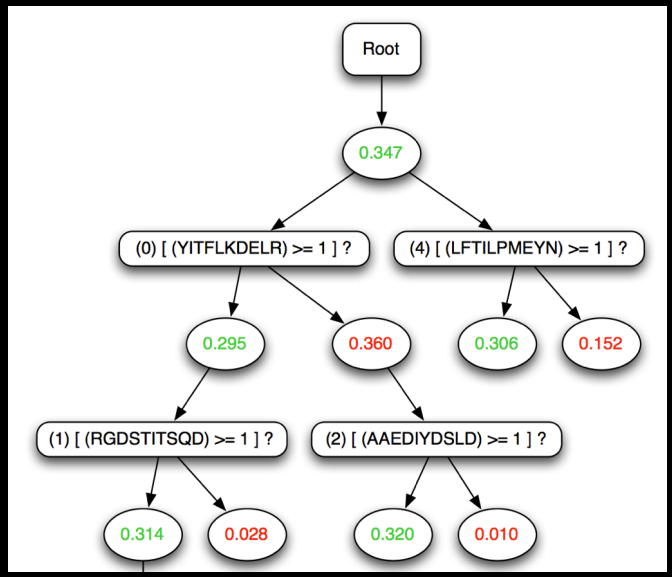




# data science: practice

- better wrong than "nice"





# data science: practice

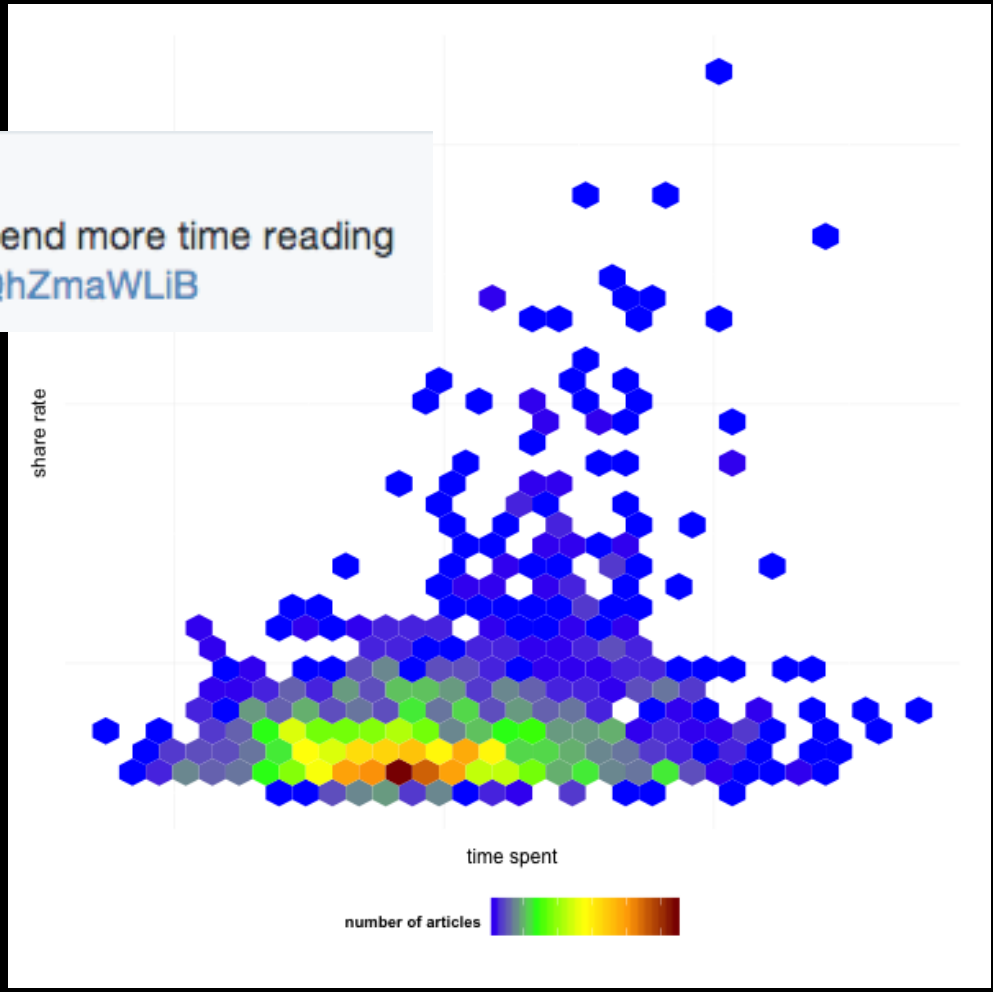
- be relevant

# data science: practice

- be relevant

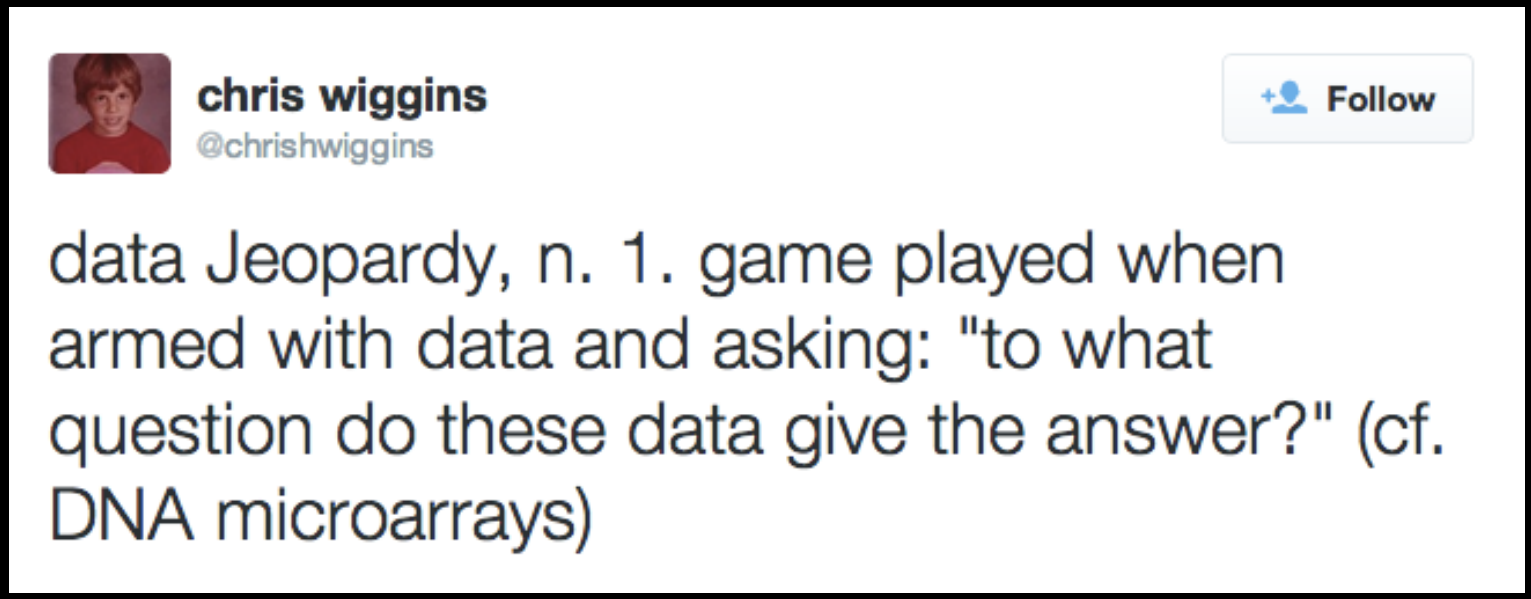
# data science: practice

- be relevant

# data science: practice

- hypotheses are not data jeopardy

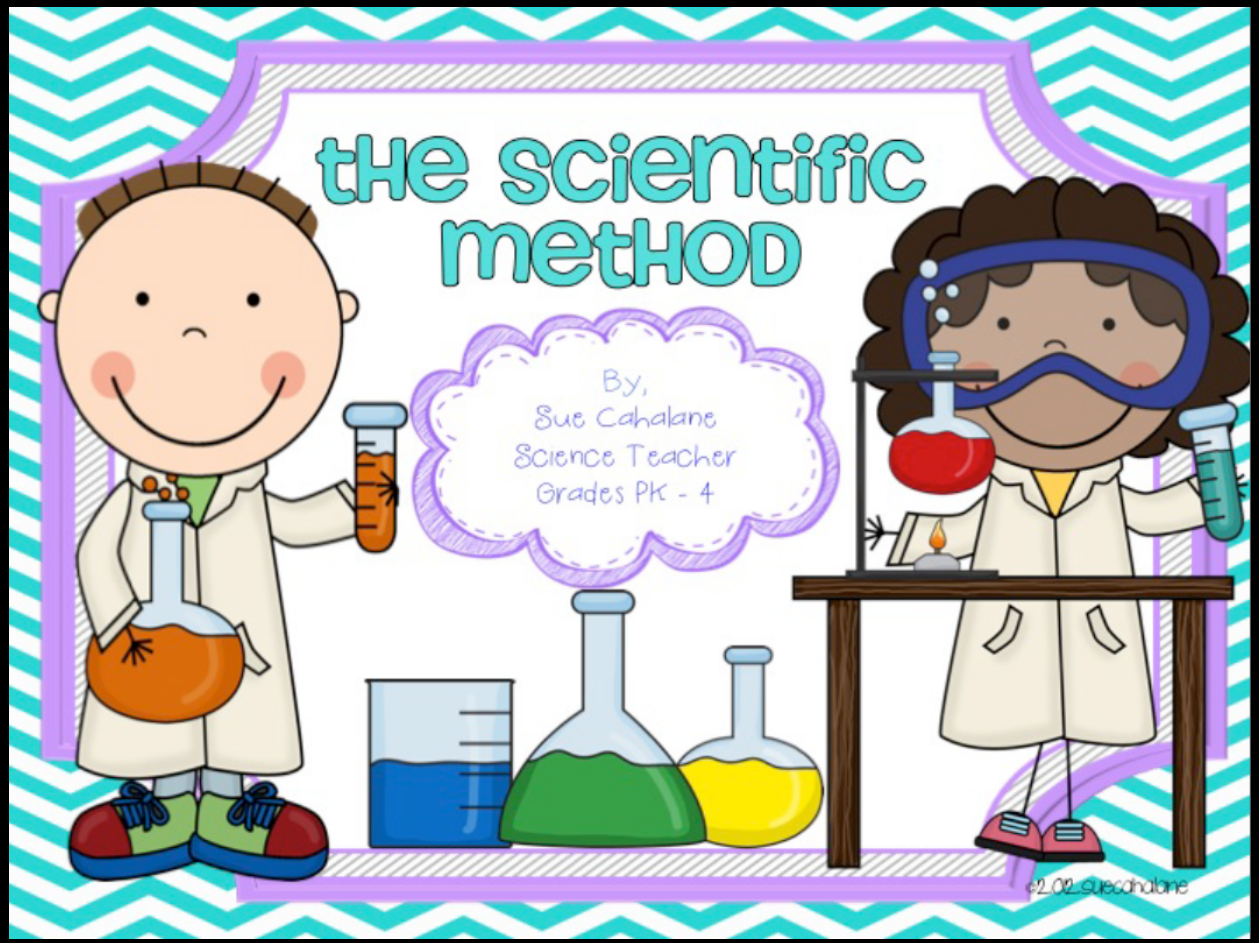


# data science: practice

- befriend experimentalists

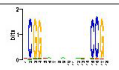
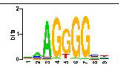
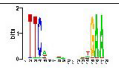
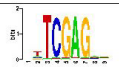
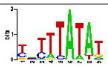
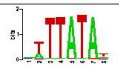
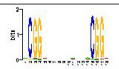
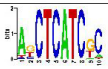
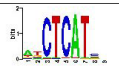
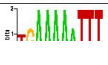
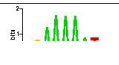
# data science: practice

- befriend experimentalists



# data science: practice

## - befriend experimentalists

| TFNAME | DB-MOTIF | MOTIF | DBNAME | d(p,q) |
|---|---|---|---|---|
| CBF1 | CACGTG |  | YPD | 0.032635 |
| CGG everted repeat | CGGN*CCG |  | YPD | 0.032821 |
| MSN2 |  |  | TRANSFAC | 0.085626 |
| HSF1 | TTCNNNGAA |  | SCPD | 0.102410 |
| XBP1 |  |  | TRANSFAC | 0.140561 |
| STE12 |  |  | TRANSFAC | 0.256750 |
| GCN4 |  |  | SCPD | 0.292221 |
| TBP |  |  | TRANSFAC | 0.376601 |
| HAP1 | CGGNNNTWNCGG |  | YPD | 0.423004 |
| RAP1 | RMACCCA |  | SCPD | 0.523059 |
| mPAC |  |  | AlignACE | 0.552493 |
| |  |  | | |

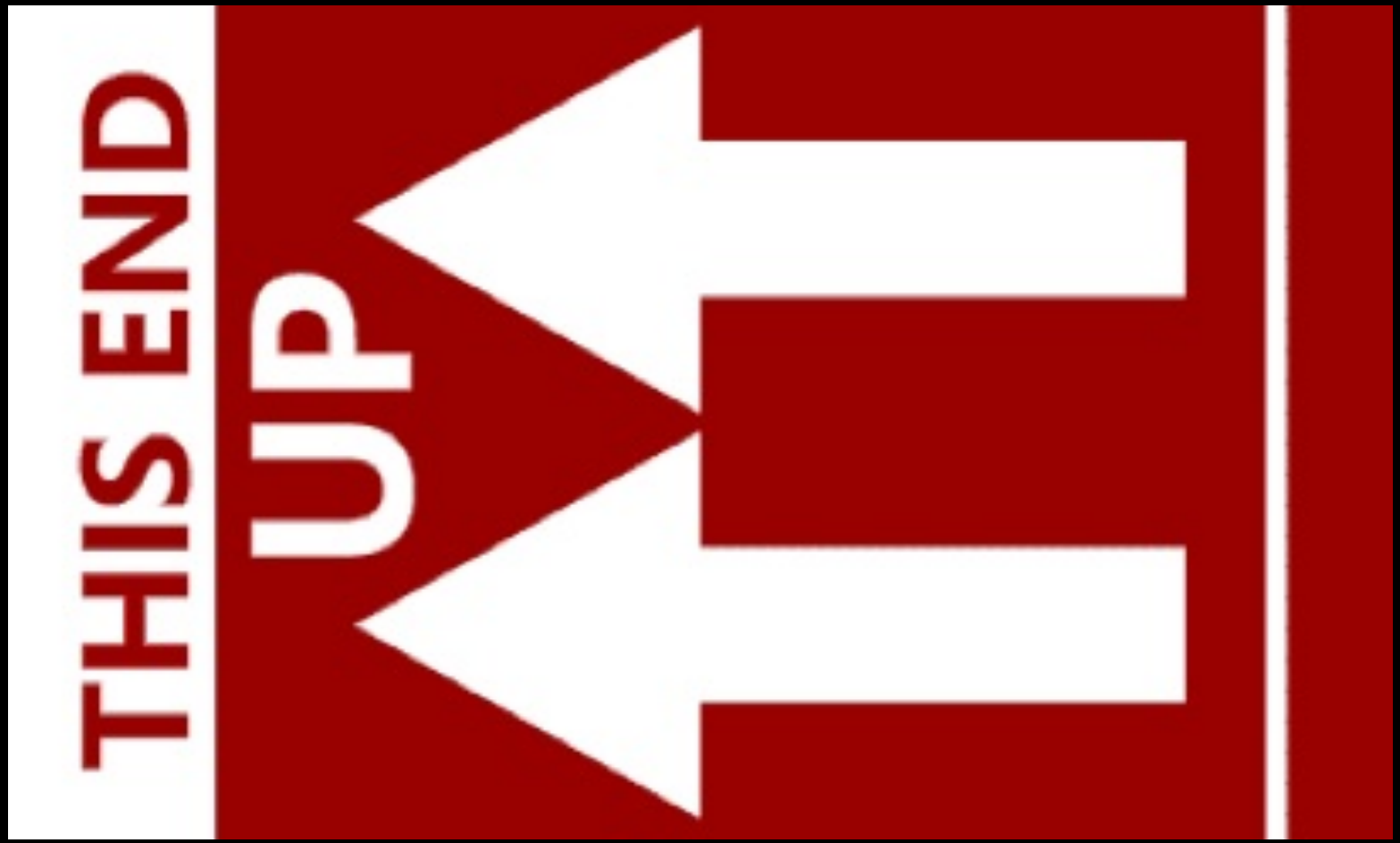
# data science: skills

# data science: skills

- find quantifiables

# data science: skills

- find quantifiables (choose carefully)

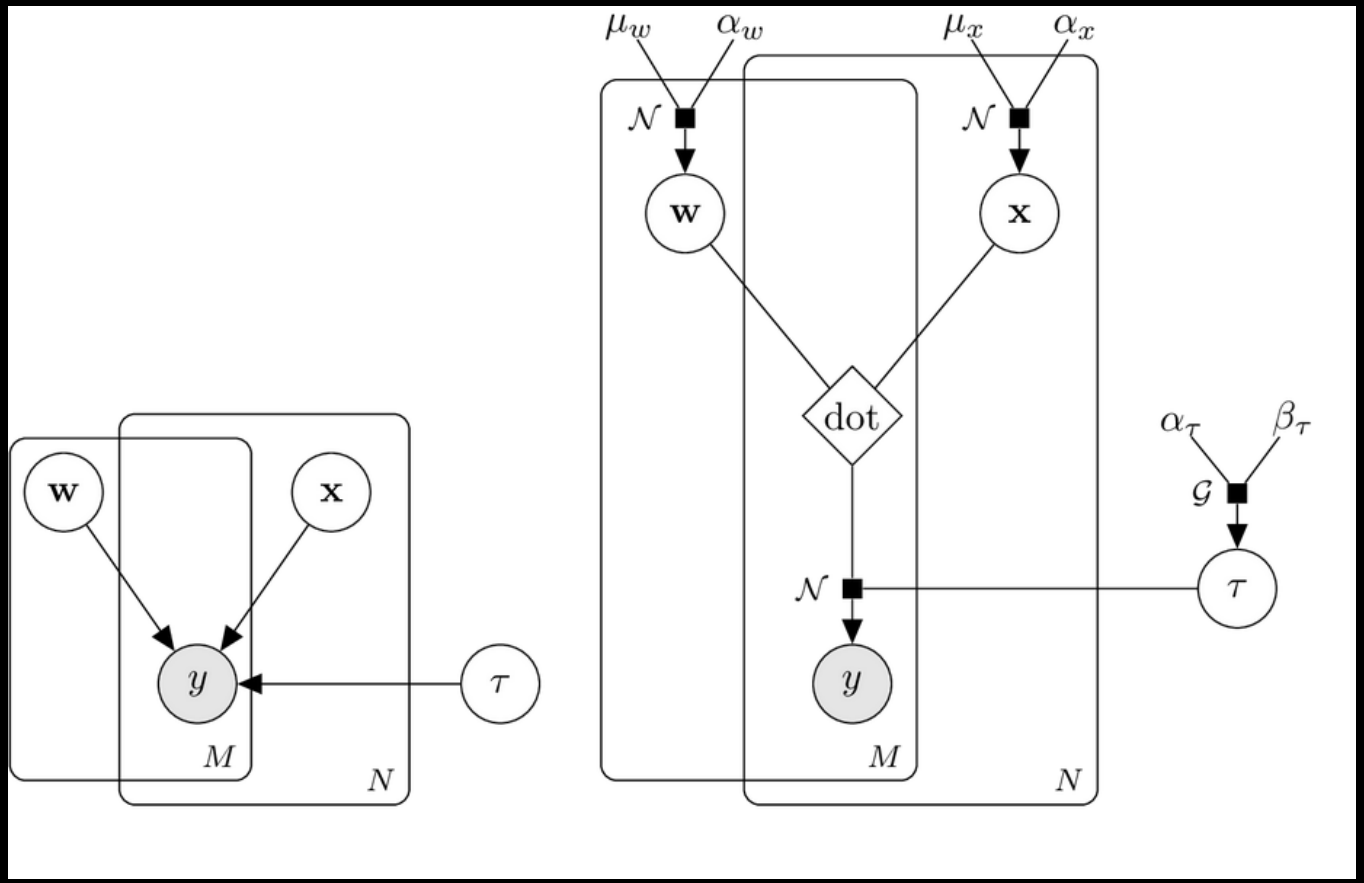


# data science: skills

- straw man first

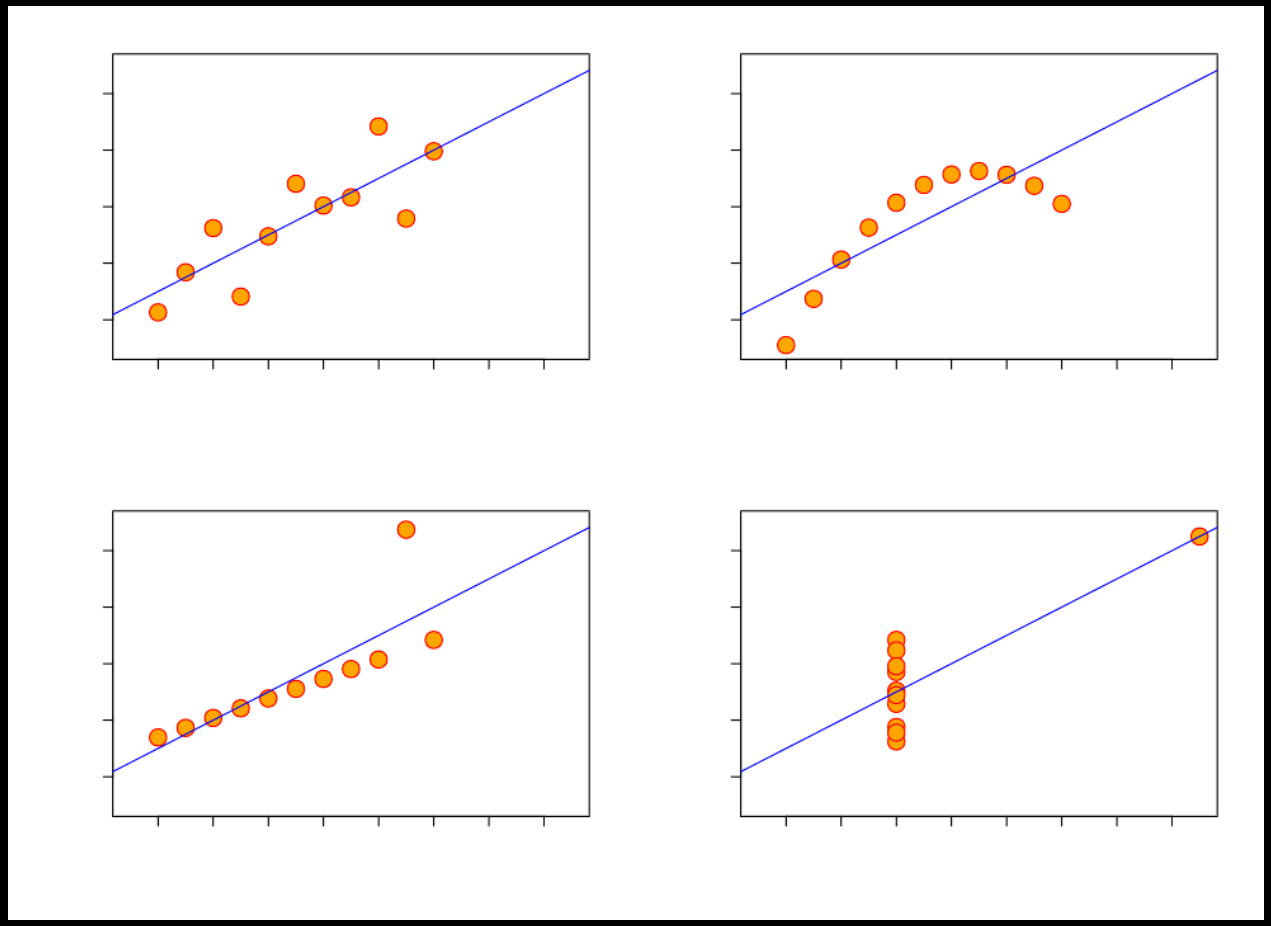
# data science: skills

- straw man first



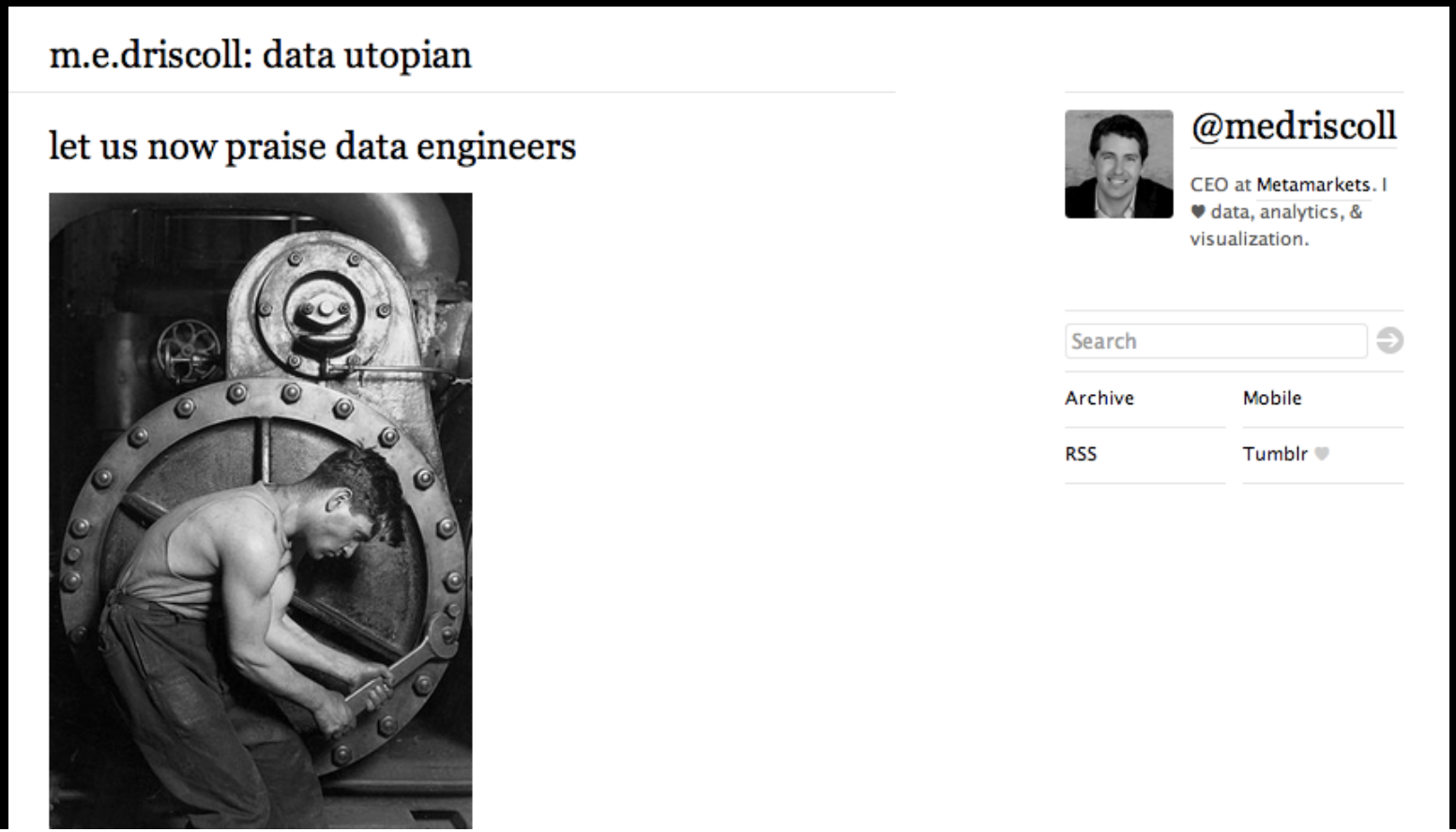
# data science: skills

- small wins before feature engineering



# data science: skills

- data engineering before data science



# data science: culture

# data science: culture

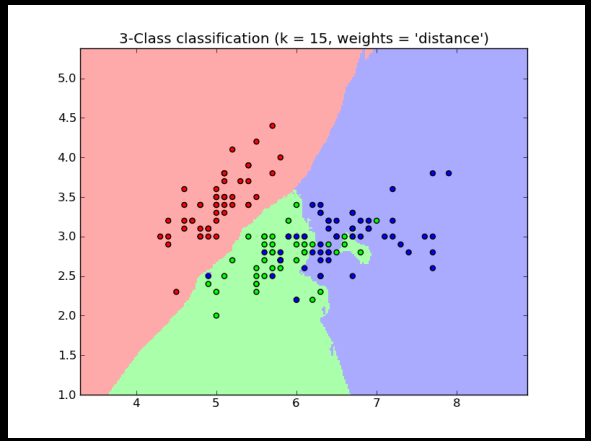
- be communicative

# data science: culture

- be communicative
  (promote rhetorical literacy)

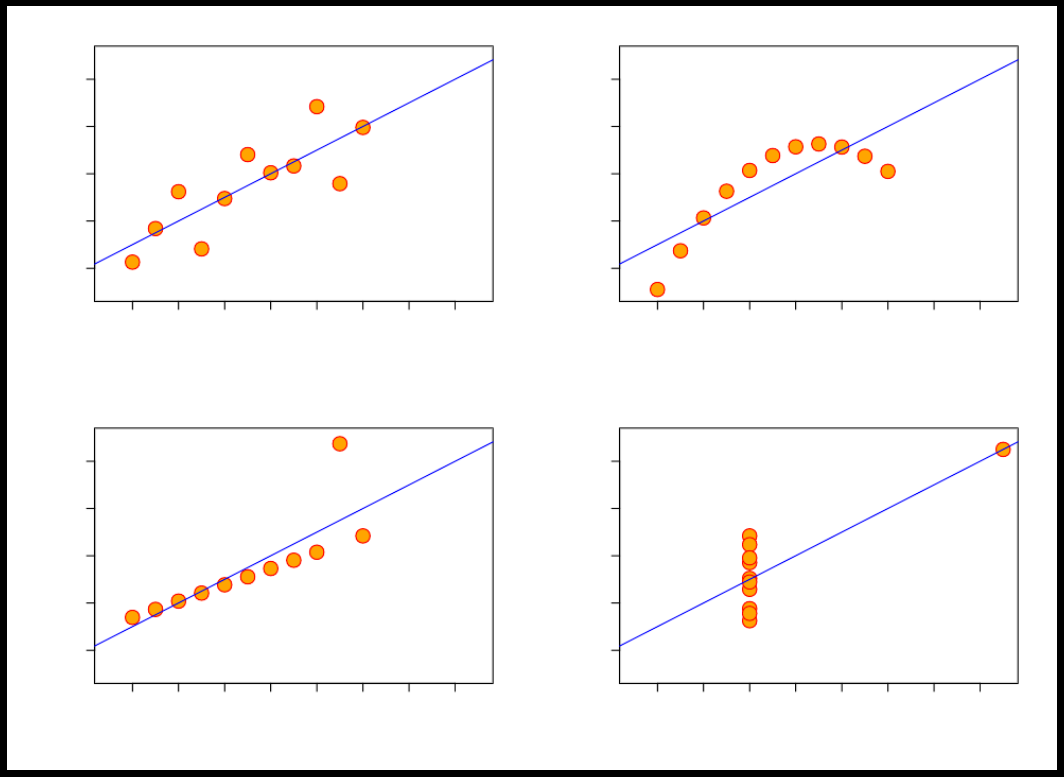
# data science: culture

- be communicative
  (promote rhetorical literacy)
- related: strive to build models which are both predictive and interpretable



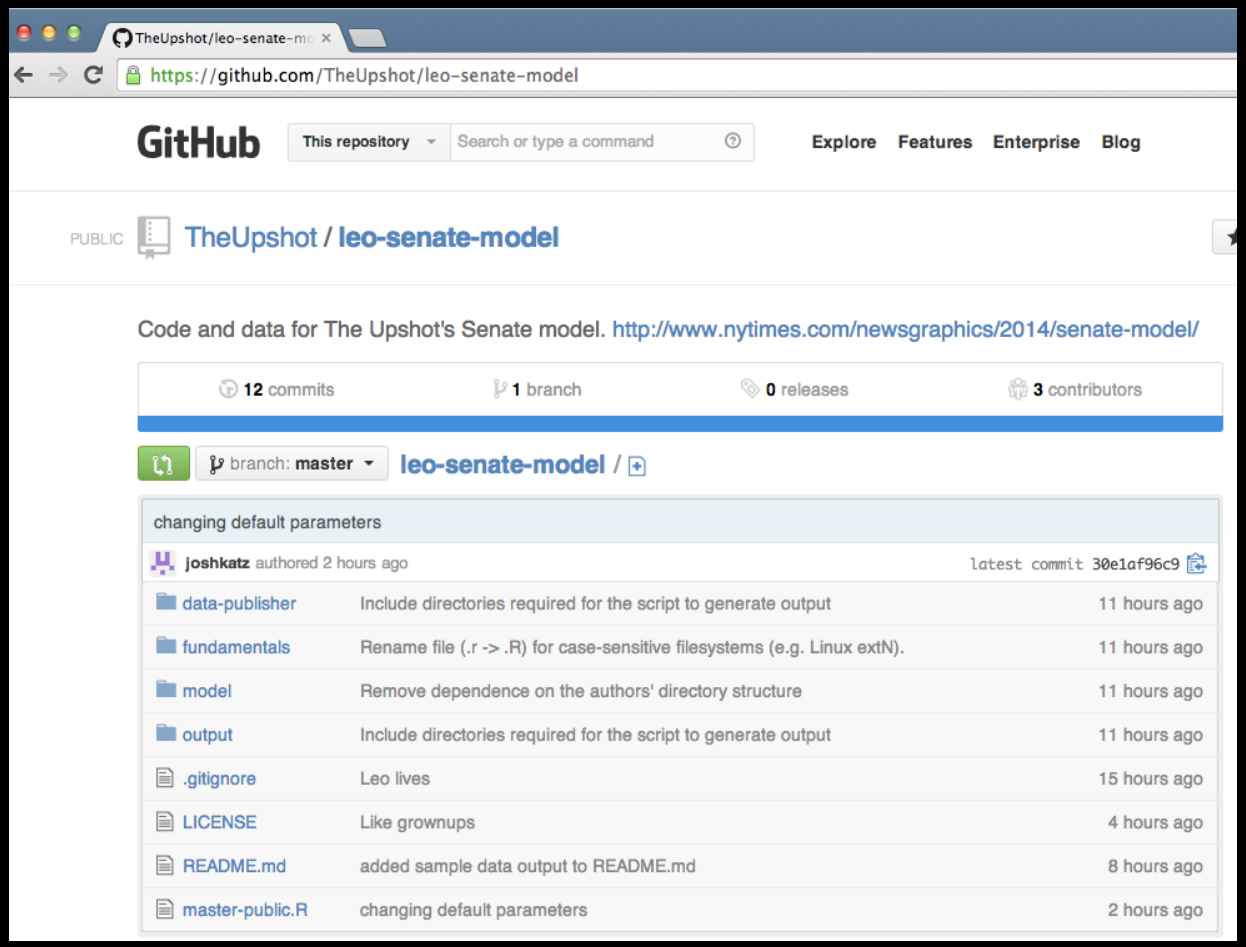
# data science: culture

- be skeptical
  (promote critical literacy)



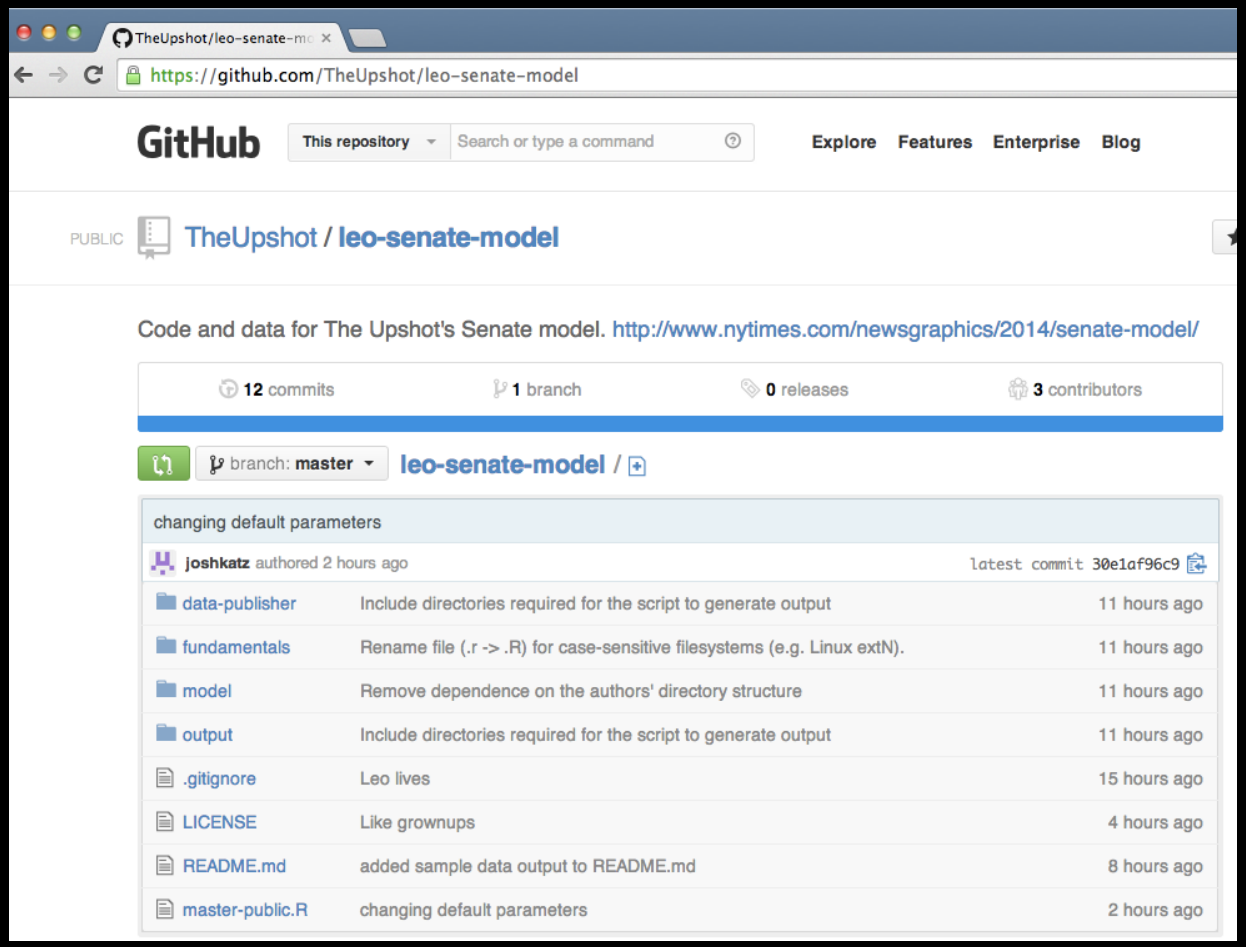
# data science: culture

- be empowering



# data science: culture

- be transparent



# data science: culture

- promote literacy

# data science: culture

- promote literacies:
  1. functional

# data science: culture

- promote literacies:
  1. functional
  2. critical

# data science: culture

- promote literacies:
  1. functional
  2. critical
  3. rhetorical

# data science: culture

- promote literacies:
  1. functional
  2. critical
  3. rhetorical

(cf. Selber, Multiliteracies for a Digital Age. 2004)

</body>

i.e., <footer>

summary:

summary:
pay attention to:

1. practices
2. skills
3. culture

# practices:

1. reframe questions as ML
2. better wrong than "nice"
3. be relevant
4. aim for hypothesis vs data jeapordy
5. befriend experimentalists

skills:

1. find quantifiables
2. straw man first
3. small wins before feature engineering
4. data engineering before data science

# culture:

1. be communicative
2. be skeptical
3. be empowering
4. be transparent
5. promote literacies

# find out more!

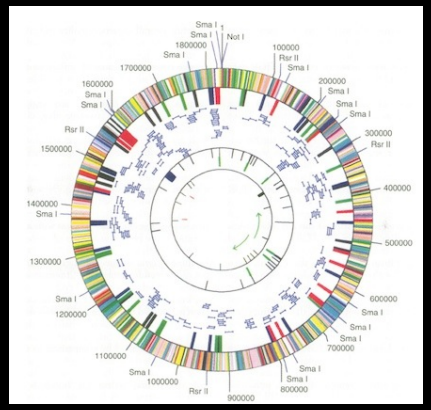
1. postdoc/student opportunities:
chris.wiggins@columbia.edu

2. always hiring:
chris.wiggins@nytimes.com

3. let's talk:
- @chrishwiggins
- gist.github.com/chrishwiggins/

# what is a computational biologist doing at the New York Times?

## (and what can academia do for a 163-year old company?)




chris.wiggins@columbia.edu
chris.wiggins@nytimes.com
chris.wiggins@hackNY.org
@chrishwiggins