
# Visualization Process

PRESENTED BY

Brian Lehman
brian@gnip.com
@WordCrank

# May 2013:

# “Could you create a cool data visualization for us?”

Sincerely,

-marketing

# June 2013:
# “yes.”

Sincerely,
-data science

# TOOLBOX
# for
# QUESTIONING

# Tool box

- Idea developer
  - Create connections among levels of questions
  - Make options accessible (education)
- Project definer
- Industry explorer
  - Invent or be aware of technology
  - Live here

# BUILDING A PROJECT

# Beginning-to-End vs End-to-Beginning

- Which dataset?
  - There is no ‘which’
  - Aggregation station
  - Weigh what you want against what you have
  - Repeatable collection process
  - Skeptics welcome

- Which visualization?
  - Factors:
    - Familiarity
    - Timeline
    - Deliverable
    - Iterative
  - Drives data collection & manipulation

# Beginning with the visualization

- What does the visualization look like? (no restrictions)
  - What story could this visualization tell? (no restrictions)
    - What do we need to build this visualization? (restrictions)

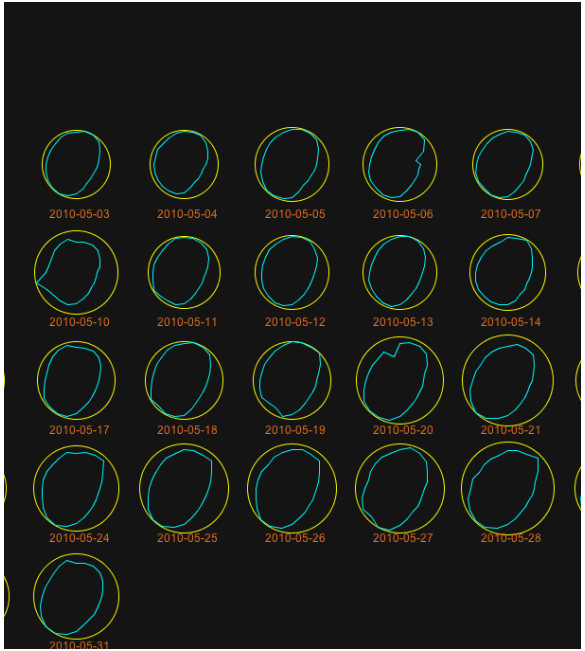
# Beginning with the data

- What data do we have? (restrictions)
  - What story does it tell? (restrictions)
    - What story do we want to tell? (no restrictions)

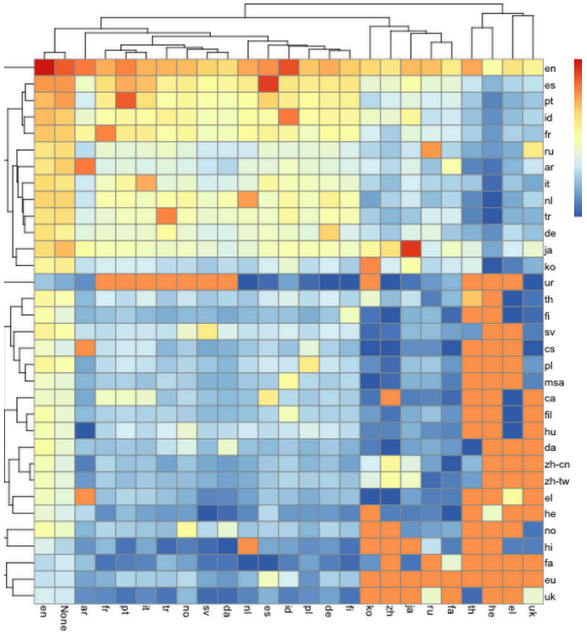
# TWITTER LANGUAGES STORY

# Initial visualizations

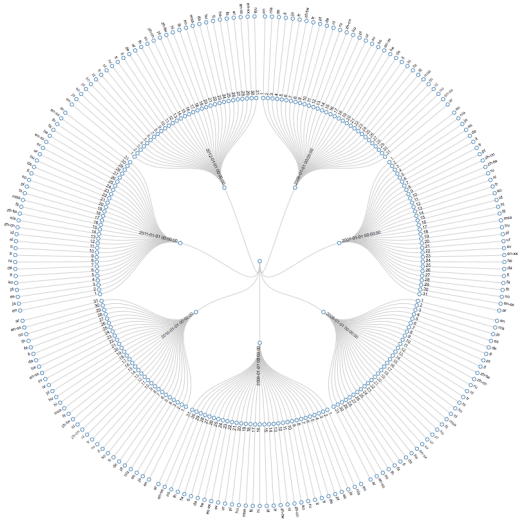
## Tweet volume




## Language comparisons

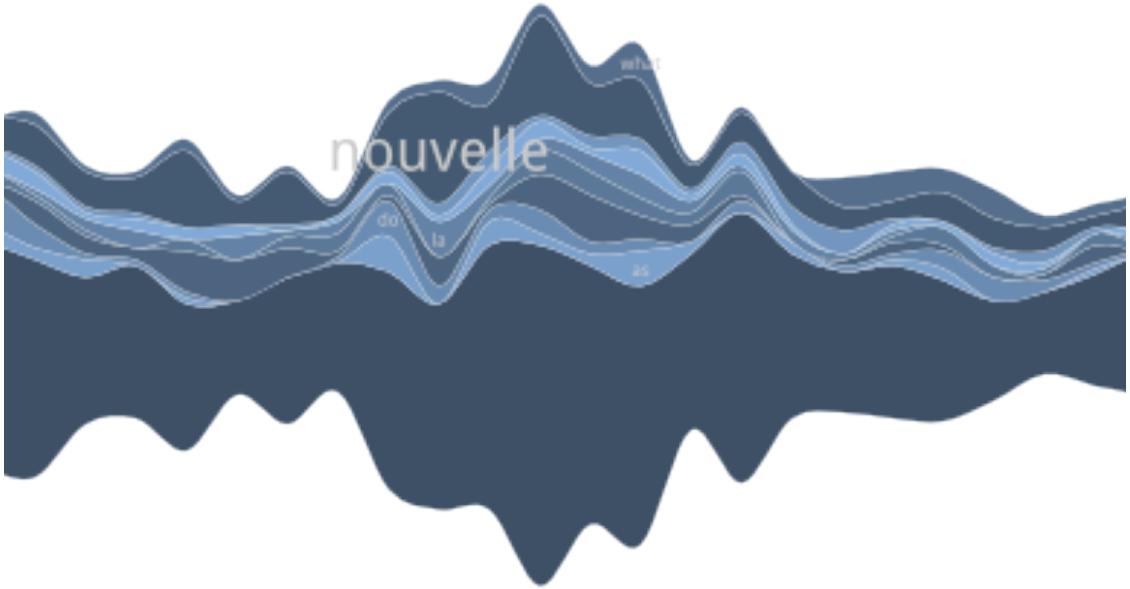



## Say what?



# Anchor on a goal

- Post napkin graffiti and hallway conversations, we found an anchor:
  - **Visualize Twitter languages over time**.
- End to Beginning:
  - Graphing ideas that could accomplish the goal?



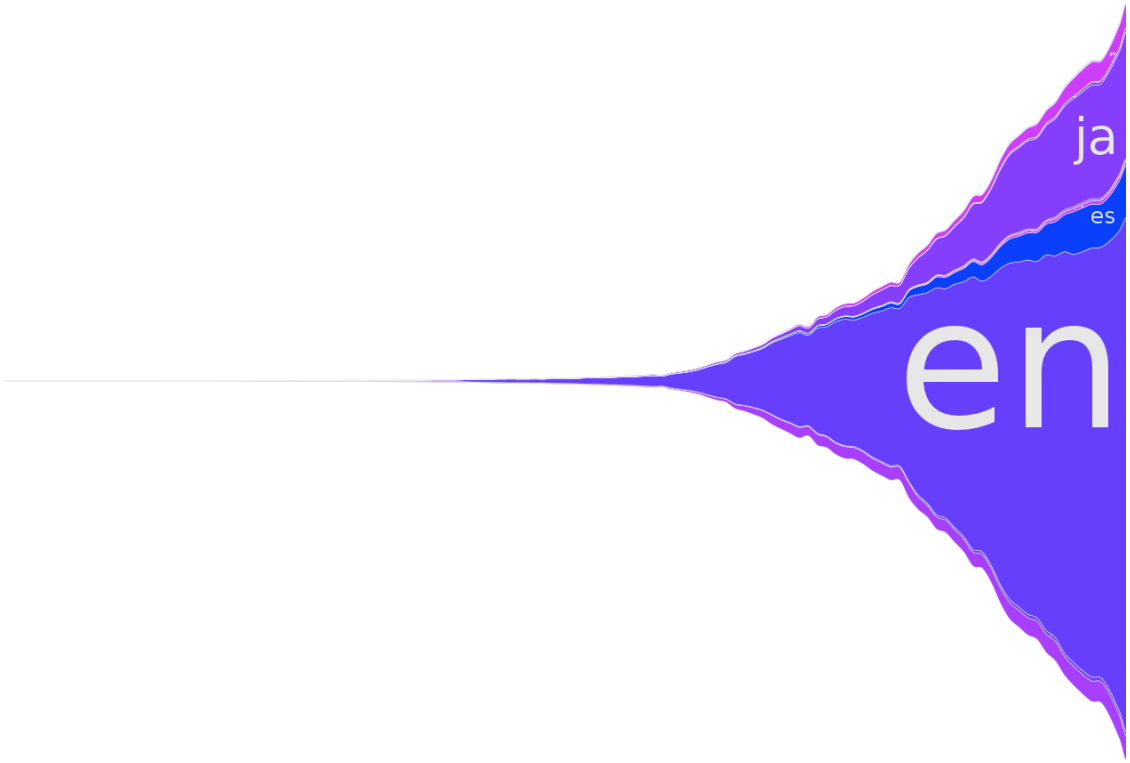
# What we graph: our data

- Needs to create this visualization?
  - Categories
  - Specific fields
  - Format

```
# populate add the counts to inner array with sys.stdin data
#for row in sys.stdin:
bucketcounter = 1 #keeping track of how many buckets we've added to our ti
lastbucket = 0 # to keep track of what our last bucket was, so we know whe
prev_row_date = end
for row_list in csv.reader(sys.stdin):
    row_date = datetime.datetime.strptime(row_list[0],datepattern)
    row_lang = row_list[1]
    row_count = int(row_list[2])
    row_diff = row_date - start # gives timedelta object, difference betwe
    bucket = int(math.ceil(row_diff.total_seconds()/delta.total_seconds()))
#   print >> sys.stderr, "bucket: ", bucket
#   print >> sys.stderr, 'bucket number:', bucket,'last bucket:',lastbucke
    if row_lang == '' or row_lang == 'NULL':
        row_lang = 'n/a'
    print >> sys.stderr, 'bucket number:', bucket,'last bucket:',lastbucke
    if row_lang in langs:
        d[row_lang][bucket][1] += row_count
        if prev_row_date != row_date: # this ensures you only increment bu
            bucketcounter += 1
    else:
        print >> sys.stderr, "invalid row language: ", row_list
    if bucket != lastbucket: # linear interpolation for missing bucket min
        ratio = bucketlimit/float(bucketcounter)
        for key in d:
            d[key][lastbucket][1] *= ratio
```

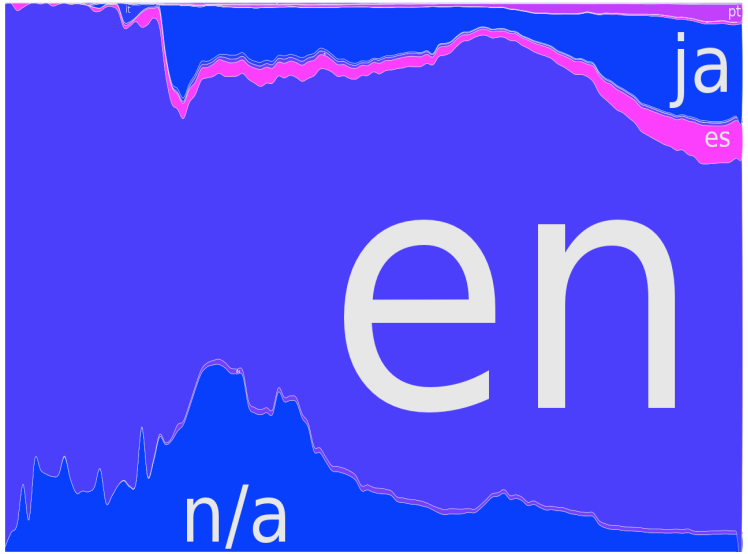
# What we see: our data

- Streamgraph technique
  - What!? That's not what I wanted to see!
- Decisions:
  - Refactor technique.
  - Data is messed up.
  - Story does not exist.



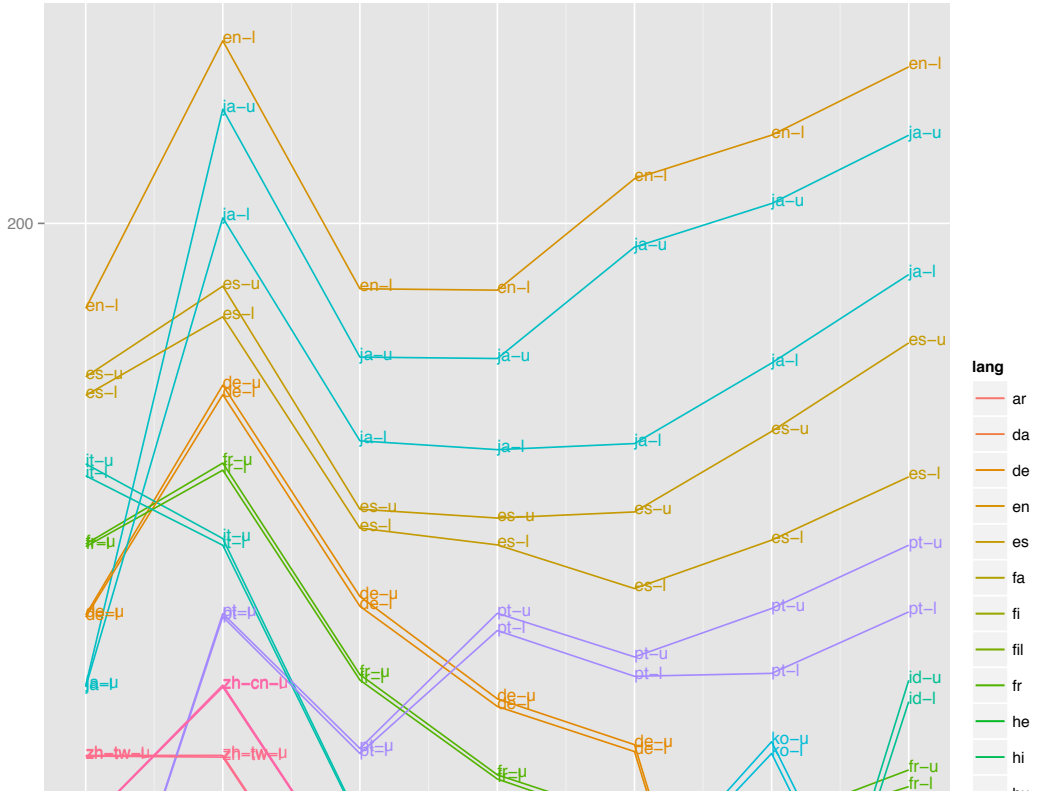
# What we do: iterate

- In this draft: languages as percentage of Tweets.
- Scale is still an issue
- Go back to the goal



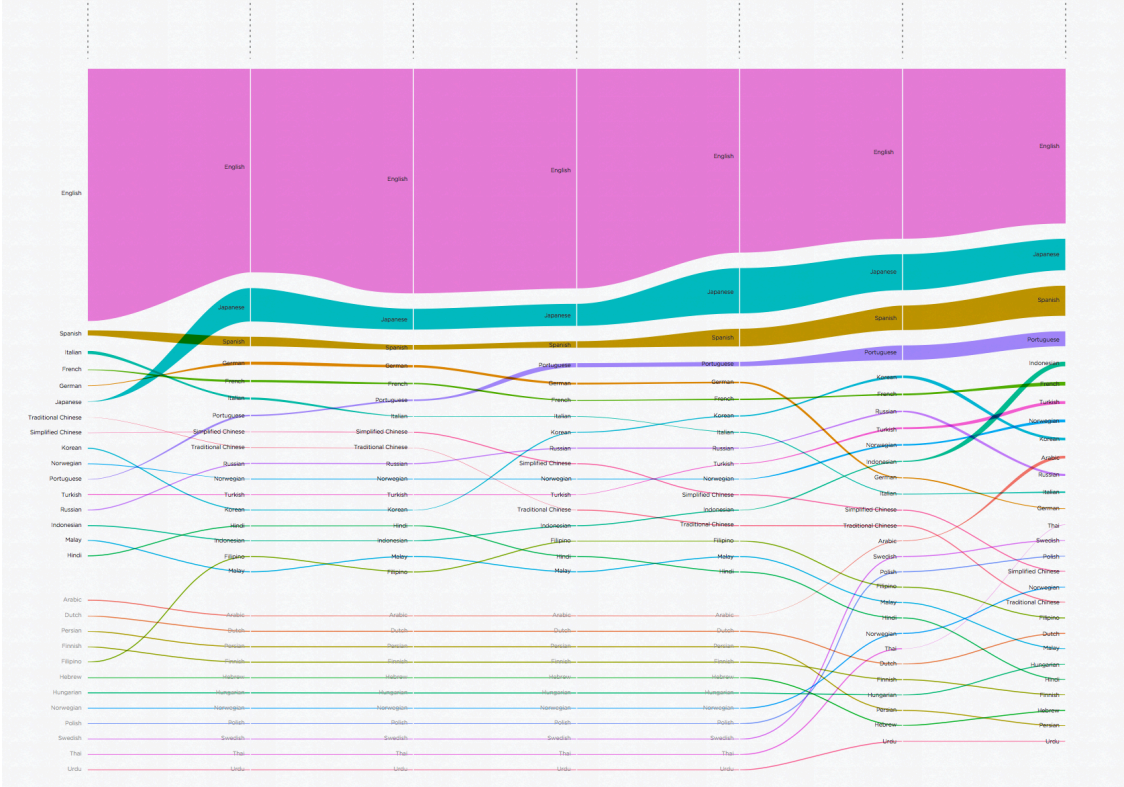
# What we do again: iterate

- The streams became individual stories
- The graph became fugly
- Go back to the goal



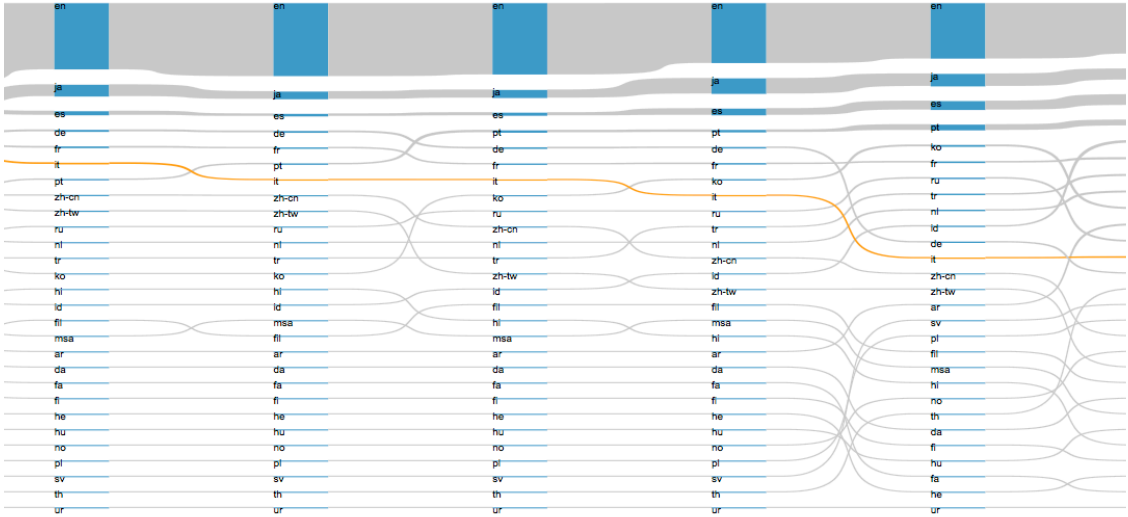
# What is non-iterative?

- An amazing design: one off is not an option.
- Great work, but when using one off designs, time will become a major hurdle.
- Go back to the goal.



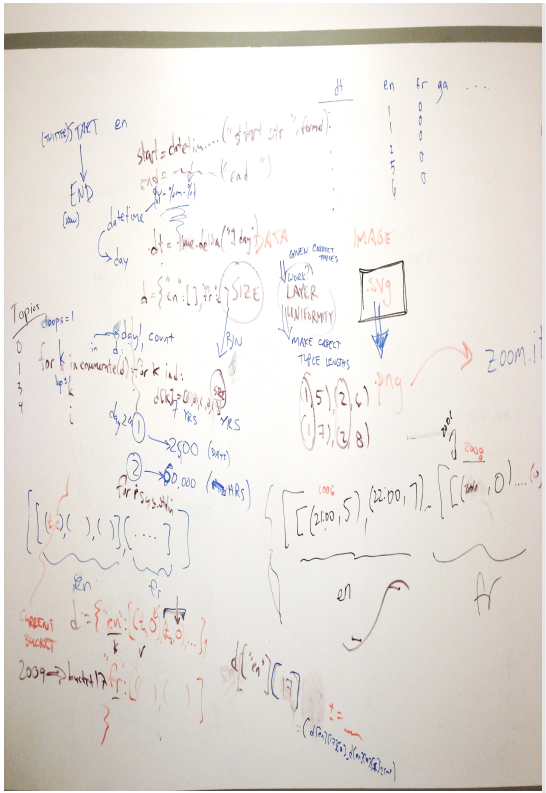
# Why do we what we do?

- Creating iterative designs is essential when change is inevitable
- Finding a code based design, we found the secret sauce.
- We were close to meeting our goal, but still went back to it for refinement.



# Iterative: applies to related processes

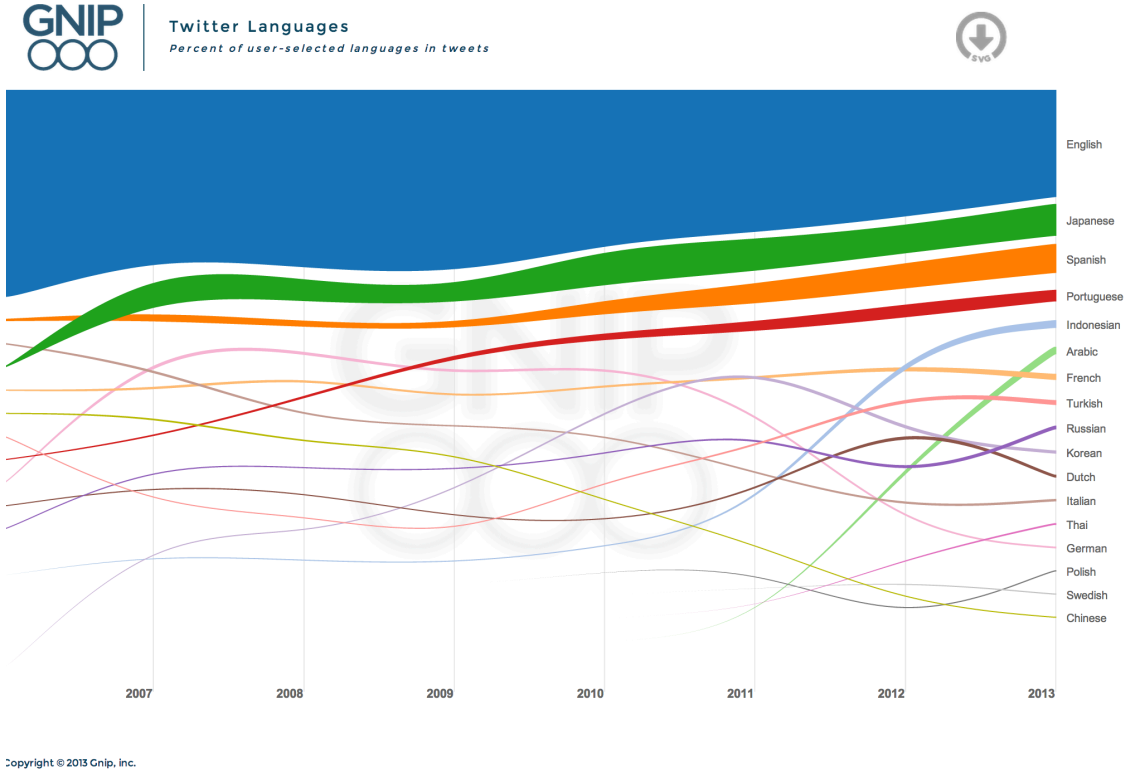
- Easily change the aggregation technique
- Quickly create various versions
- Time was starting to be more of a concern.



```
2013-10-01,bn,9,1.15097815607e-09,60
2013-10-01,bs,25,3.19716154463e-09,55
2013-10-01,ca,3807937,0.000486983589632,18
2013-10-01,cs,959858,0.000122752843436,27
2013-10-01,da,2081181,0.000266154874425,24
2013-10-01,de,29594901,0.00378478717578,14
2013-10-01,el,2743028,0.000350796145498,19
2013-10-01,en,3704922383,0.473809414751,1
2013-10-01,es,1077466635,0.137793395642,3
2013-10-01,et,544,6.95702352112e-08,44
2013-10-01,eu,544603,6.96473507477e-05,30
2013-10-01,fa,456672,5.84021662764e-05,31
2013-10-01,fi,1984380,0.000253775337038,25
2013-10-01,fil,2384990,0.000305007932493,21
2013-10-01,fr,177475363,0.0226966962281,7
2013-10-01,gl,624878,7.99134364675e-05,29
2013-10-01,hd,6,7.67318770712e-10,63
2013-10-01,he,2718261,0.000347628781499,20
2013-10-01,hi,88518,1.13202538243e-05,33
2013-10-01,hr,1086,1.38884697499e-07,42
2013-10-01,hu,1486769,0.000190137626902,26
2013-10-01,id,290564416,0.037159255083,6
2013-10-01,in,29114,3.72328644842e-06,34
2013-10-01,is,46,5.88277724213e-09,52
2013-10-01,it,52370939,0.00669753408909,11
2013-10-01,iw,33,4.22025323892e-09,53
2013-10-01,ja,1251055566,0.159993069833,2
2013-10-01,jv,10,1.27886461785e-09,59
2013-10-01,ka,2,2.55772923571e-10,68
2013-10-01,kh,1,1.27886461785e-10,71
2013-10-01,kk,7,8.95205232498e-10,61
2013-10-01,km,52,6.65009601284e-09,51
2013-10-01,ko,84436551,0.0107982917528,10
2013-10-01,ld,29,3.70870739178e-09,54
2013-10-01,lo,1,1.27886461785e-10,74
```

# Finding a final version

- Design critique:
  - Taking it all personally can damage a final result.
  - Giving it constructively will smooth over the process.
- Trust that everyone is aiming to accomplish this goal.
- Always go back to the goal.
- NYT




# Thank you!

@gnip
www.gnip.com

PRESENTED BY

Brian Lehman
brian@gnip.com
@WordCrank