Trowel

A programming language for journalists

David Tagatac	- Language Guru
Hareesh Radhakrishnan	- System Architect
Pucong Han	- System Integrator
Robert Walport	- System Tester
Victoria Mo	- Project Manager

The Problem

- Journalists need to find relevant information online.

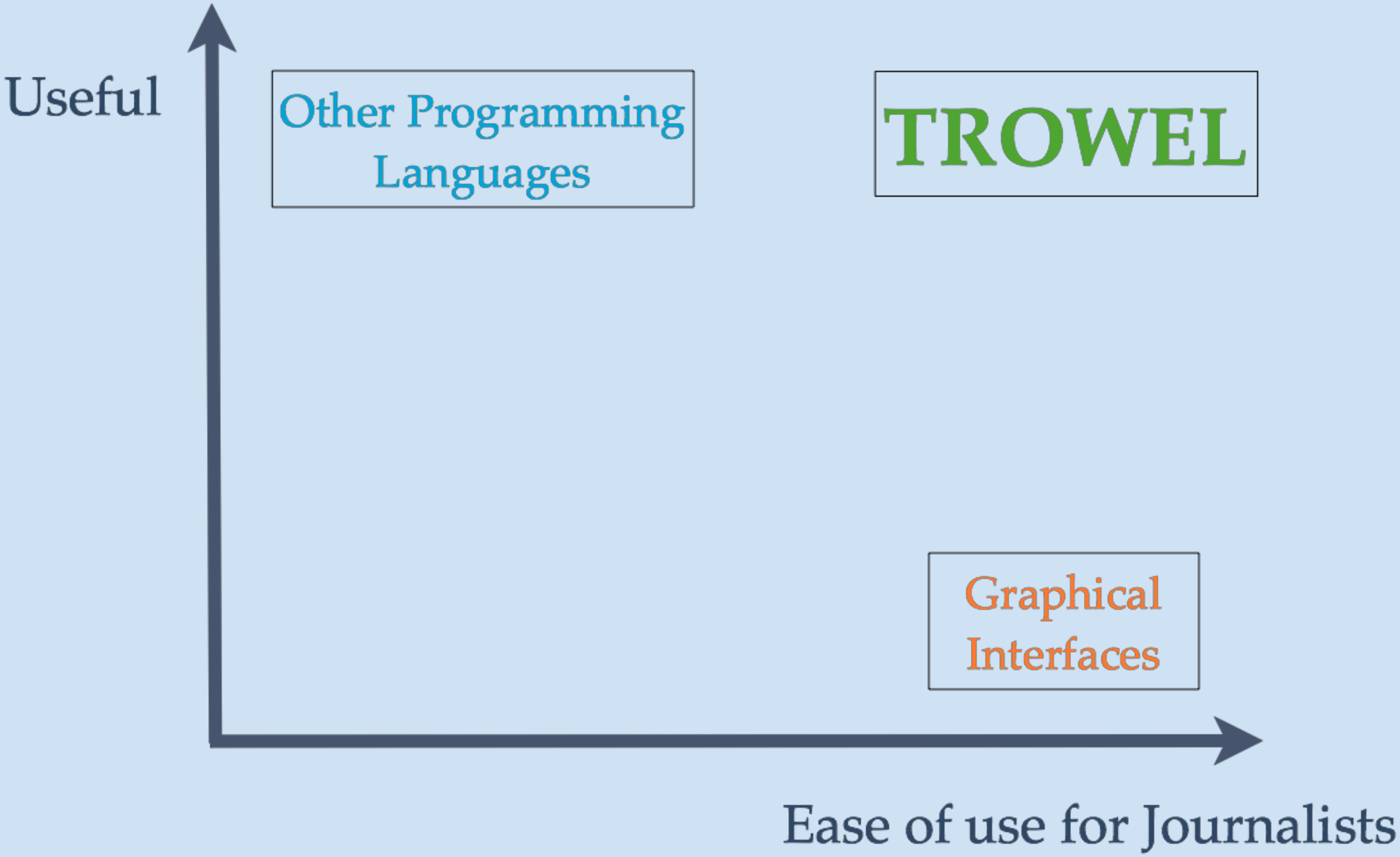
- Web scraping quickens this process: requires knowledge of advanced programming languages

- Typical journalist not technically proficient: does **not** code.

Trowel is...

A web-scraping programming language developed **specifically for journalists.**

Trowel is...



Useful
Other Programming Languages
TROWEL
Graphical Interfaces
Ease of use for Journalists

Trowel is...

Easy-to-Learn

Readable

Accessible

Intuitive

Concise

Domain-Specific

Design Goals

Readability - looks like English!

No Semi-colons

indentation handles scoping

Assignment uses “is” instead of “=”

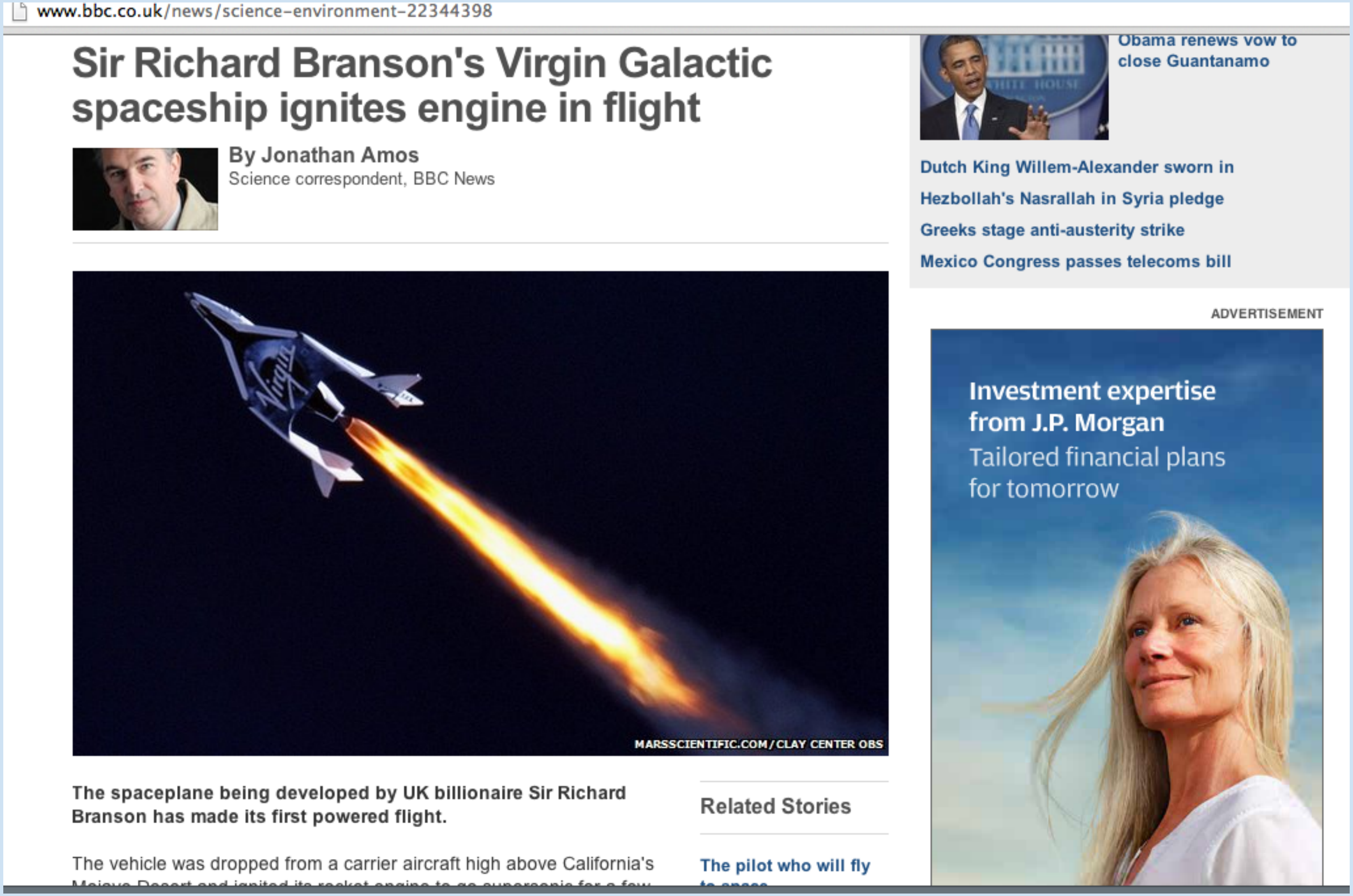
Design Goals

Built in Functions Make Sense

“insert Url into UrlList”

inserts the Url variable into the UrlList variable!

Live Demo: Liftoff!



www.bbc.co.uk/news/science-environment-22344398

Sir Richard Branson's Virgin Galactic spaceship ignites engine in flight

By Jonathan Amos
Science correspondent, BBC News

MARSSCIENTIFIC.COM/CLAY CENTER OBS

The spaceplane being developed by UK billionaire Sir Richard Branson has made its first powered flight.

The vehicle was dropped from a carrier aircraft high above California's Mojave Desert and ignited its rocket engine to go supersonic for a few

Related Stories

The pilot who will fly to space

Obama renews vow to close Guantanamo

Dutch King Willem-Alexander sworn in

Hezbollah's Nasrallah in Syria pledge

Greeks stage anti-austerity strike

Mexico Congress passes telecoms bill

Live Demo: Liftoff!

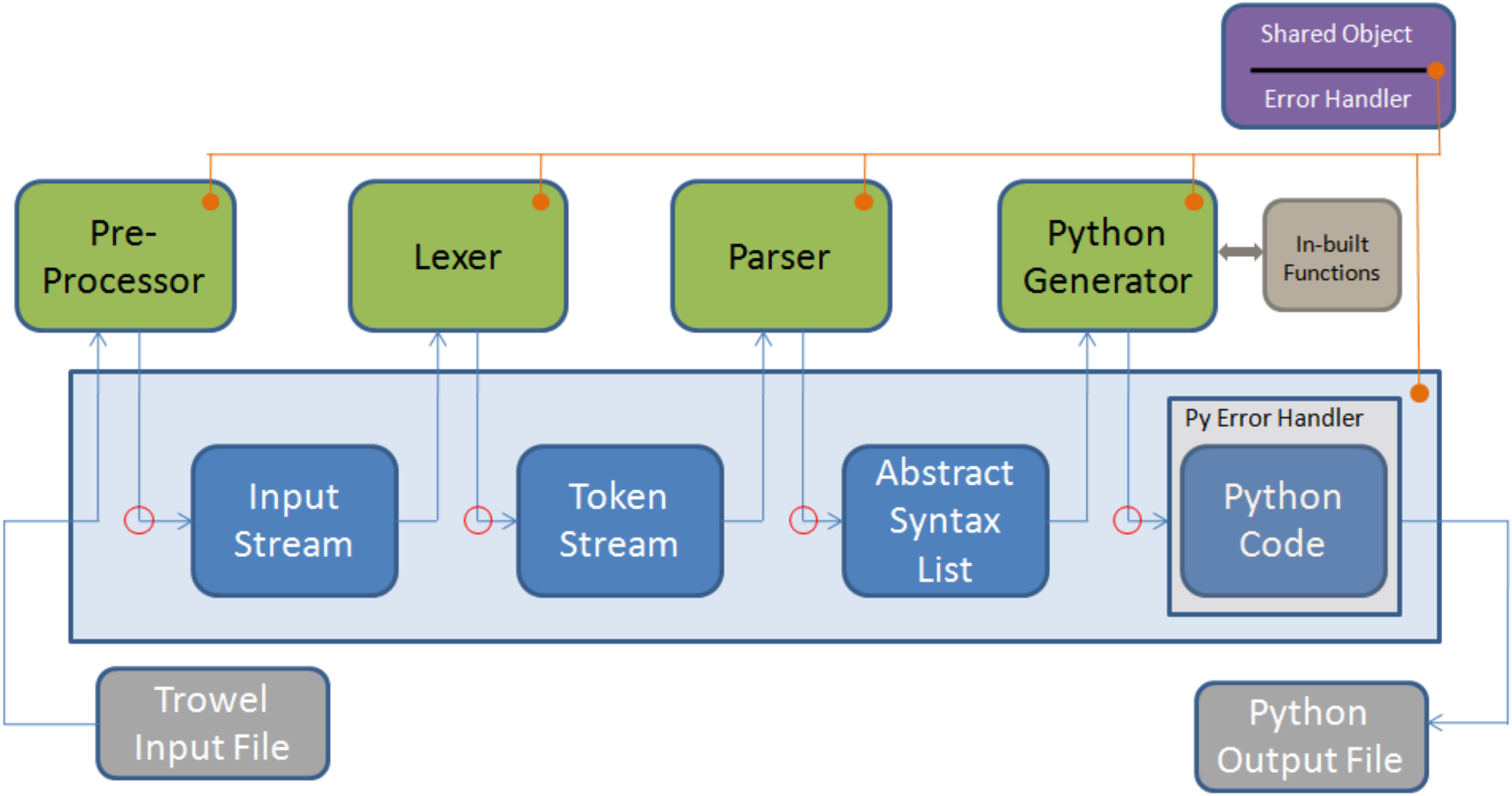
```
textlist paragraphs

url articleurl is ‘http://www.bbc.co.uk/news/science-environment-22344398’

paragraphs is findtext in articleurl with "time" and "flight"

print paragraphs
```

System Architecture

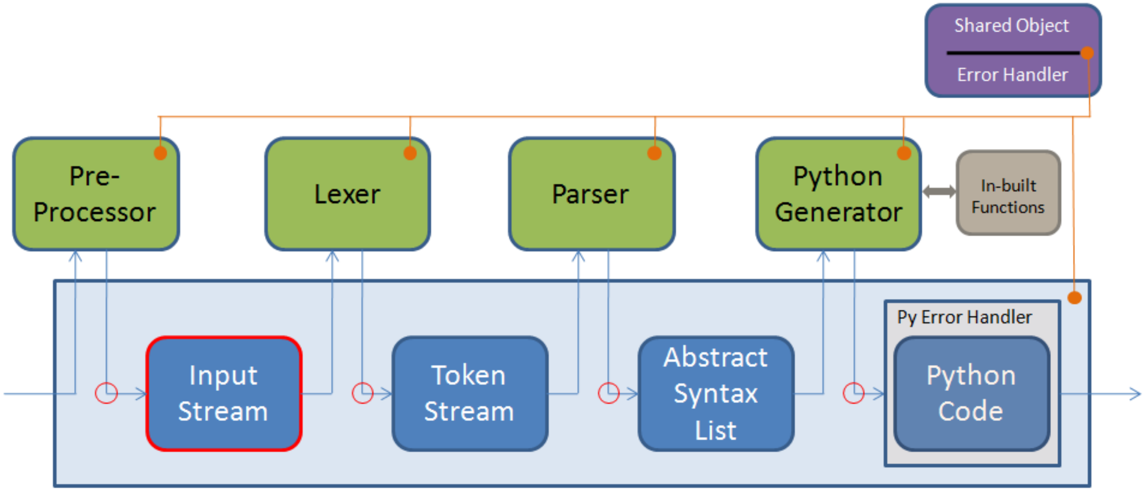
We wanted to make Trowel as close to English as possible

```
# Flexibility vs Robust Design

> read "file.txt" into ul1

> ul2 is findurl in ul1 with "obama" and "romney"
  ul4 is combine ul3 with (findurl in ul2 with "taxes")

> ul5 is mydelete ul3 from ul2 except indexlist1
```

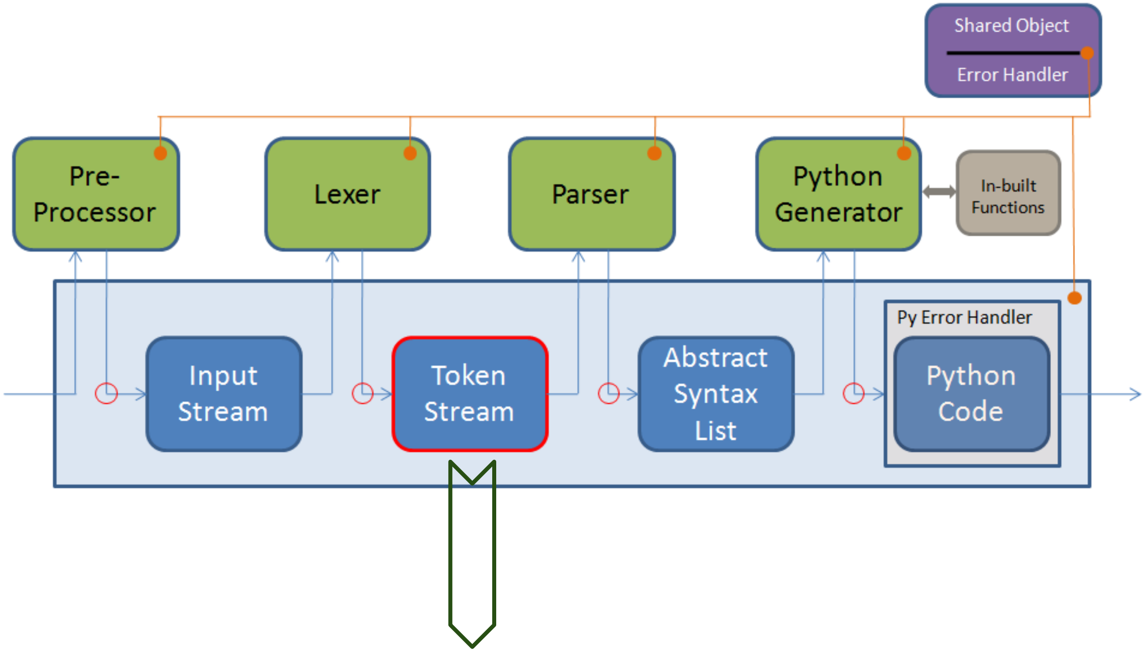
System Architecture



```
textlist para, words
url article is 'bit.ly/trowel'

para is findtext in article
   with "time"
print para
```

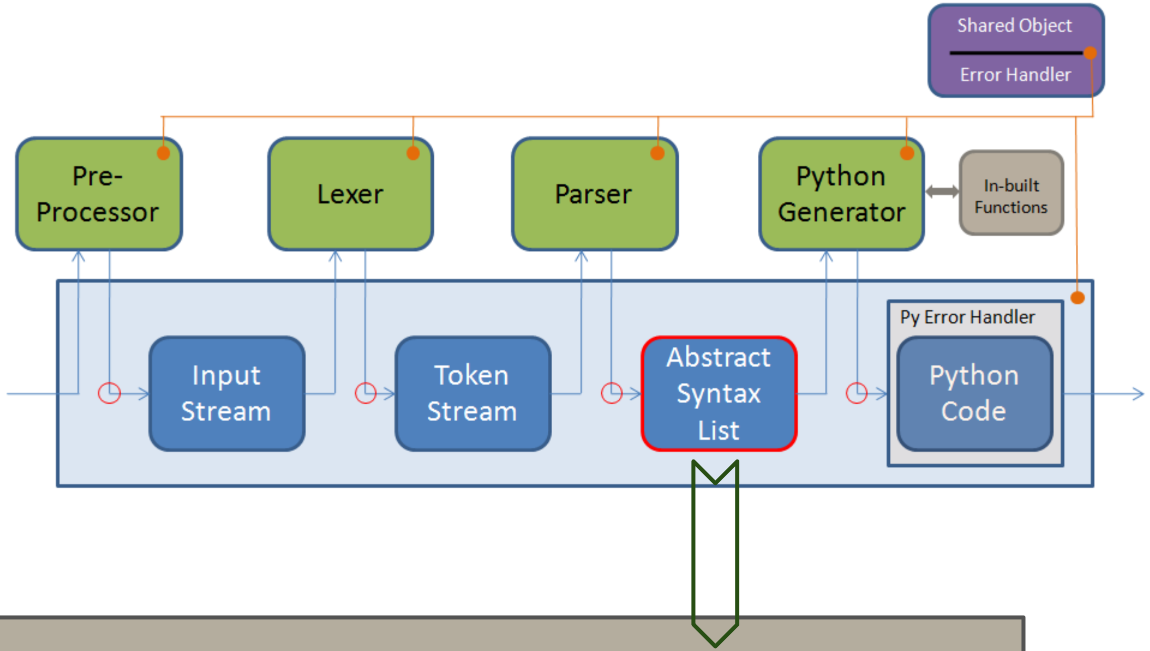


```
textlist para, words
url article is 'bit.ly/trowel'

para is findtext in article
   with "time"
print para
```



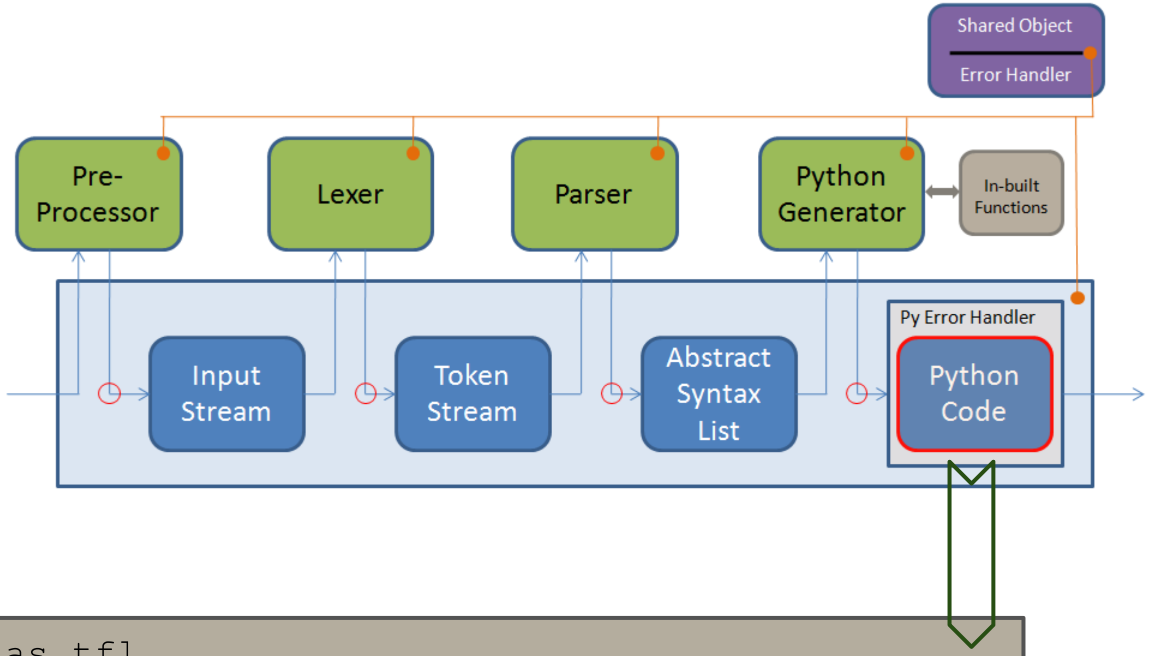
```
[['TEXTLIST', 'textlist'], ['UNKNOWN', 'paragraphs'], ['COMMA',
',''], ['UNKNOWN', 'words']]

[['URL', 'url'], ['UNKNOWN', 'articleurl'], ['IS', 'is'],
['URLVAL', "'bit.ly/trowel'"]]

[['UNKNOWN', 'paragraphs'], ['IS', 'is'], ['UNKNOWN',
'findtext'], ['UNKNOWN', 'in'], ['UNKNOWN', 'articleurl'],
['UNKNOWN', 'with'], ['TEXTVAL', '"time"']]

[['UNKNOWN', 'print'], ['UNKNOWN', 'paragraphs']]
```

```
textlist para, words
url article is 'bit.ly/trowel'

para is findtext in article
   with "time"
print para
```



```
['declaration', ['datatype', 'textlist'], [['paragraphs'],
['words']]]

['declaration', ['datatype', 'url'], [['articleurl',
['expression', ['value', ['url', 'bit.ly/trowel']]]]]]

['assignment', ['variable', 'paragraphs'], ['expression',
['functioncall', ['functionname', 'findtext'], 'arguments',
[['expression', ['insertword', 'in']], ['expression',
['variable', 'articleurl']], ['expression', ['insertword',
'with']], ['expression', ['value', ['text', 'time']]]]]]]

['functioncall', ['functionname', 'print'], 'arguments',
[['expression', ['variable', 'paragraphs']]]]
```

```
▶ textlist para, words
▶ url article is 'bit.ly/trowel'

▶ para is findtext in article
     with "time"
▶ print para
```



```
  import trowelfunctions as tfl

▶ paragraphs = ""
  words = ""

▶ articleurl = 'bit.ly/trowel'

▶ tmp0 = 'in'
  tmp1 = articleurl
  tmp2 = 'with'
  tmp3 = 'time'
  paragraphs = tfl.r_findtext([tmp0,tmp1,tmp2,tmp3])


▶ tmp0 = paragraphs
  tfl.r_print([tmp0])
```

Example Function: FindText

- receive a URL to search and a logical expression
    - e.g. "Obama and Romney"
- Use Beautiful Soup to grab website and parse it
- convert the logical expression of terms into a logical expression of booleans for each paragraph
    - e.g. "Obama and Romney" --> "True and True"
- if expression evaluates to true: return the paragraph

Example Function: FindText

Para is findtext in Article

```
def r_findtext(arglist):
    link = arglist[1]
    parts = urlparse.urlsplit(link)
    if not parts.scheme or not parts.netloc:
        link = "http://" + link
    html = urlopen(link)
    soup = BeautifulSoup(html)
    texts = soup.find_all('p')
    keyparas = []

    for para in texts:
        para = para.get_text()
        truthiList = ""
        for entry in arglist[2:]:
            if str(type(entry)) != "<type 'list'>":
                if entry in LOGICALS:
                    truthiList = truthiList + " " + entry
                elif entry in IGNORE:
                    pass
                elif entry in para:
                    truthiList = truthiList + " True"
                else:
                    truthiList = truthiList + " False"
        if eval(truthiList): keyparas.append(para)
    return keyparas
```

Development Environment

Beautiful Soup

Python

Python Lex-Yacc

Testing

Regression testing: git hook disallows “git commit” if any test fails!

- Standard unit testing framework - **unittest**
- Tested the output of the lexer, the parser, the function modules and the whole compiler

Testing: Liftoff!

Lexer test

```
self.assertEqual(parsewrapper().gettokens("url spaceArticle is \
‘http://www.bbc.co.uk/news/science-environment-22344398\’"),
[['indentlevel', 0], ['declaration', ['datatype', 'url'],
[['spacearticle', ['expression', ['value', ['text', \‘http://www.bbc.
co.uk/news/science-environment-22344398\’]]]]]]]
```

Function test

```
self.assertEqual(r_findtext(['in', 'http://www.bbc.co.
uk/news/science-environment-22344398']),
[u"The vehicle was dropped from a carrier aircraft high above
California's Mojave Desert and ignited its rocket engine to go
supersonic for a few seconds...."])
```

Testing Statistics

24% Lexer tests

10% Parser tests

44% Function tests

22% Type Checking

Example: Amy

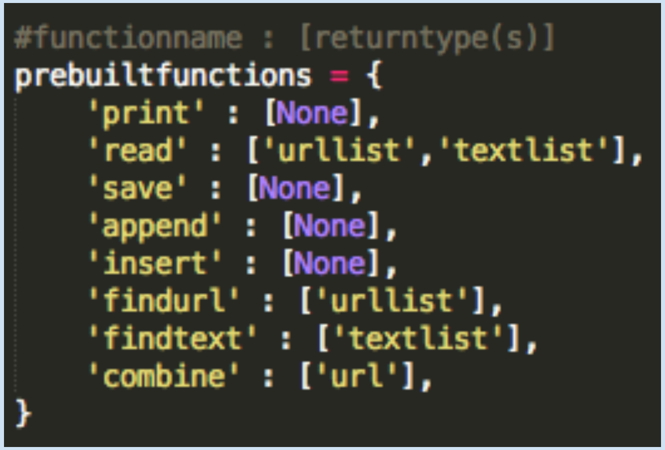
- Amy wants the latest tweets of a list of politicians.
- Amy has been given a similar assignment before.
- She thinks to herself…
- Trowel!

Example: Amy

```
define getTweets of (text person) from twitter:
    text prefix is ‘http://www.twitter.com/'
    url twitterurl is combine prefix and person
    results is findText in twitterurl
    return results

textlist alltweets

for name in [‘BarackObama’, ‘MittRomney’, ‘JoeBiden’]:
    textList tweets is getTweets of name from twitter
    insert tweets into alltweets

save alltweets into "tweetsfile.txt"
```

Example: Amy - User Defined Func

```
define getTweets of (text person) from twitter:
    text prefix is ‘http://www.twitter.com/'
    url twitterurl is combine prefix and person
    results is findText in twitterurl
    return results
```

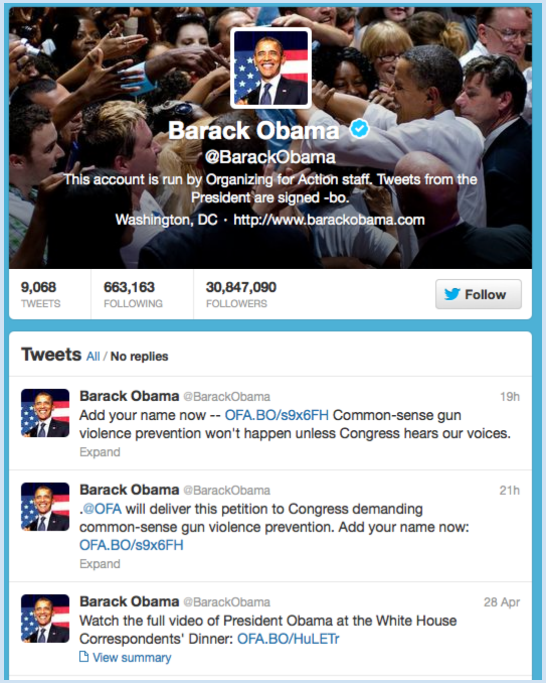


```
#functionname : [returntype(s)]
prebuiltfunctions = {
    'print' : [None],
    'read' : ['urllist','textlist'],
    'save' : [None],
    'append' : [None],
    'insert' : [None],
    'findurl' : ['urllist'],
    'findtext' : ['textlist'],
    'combine' : ['url'],
}
```

```
'gettweets': ['textlist'],
```

Example: Amy

```
textlist alltweets
for person in [‘BarackObama’, ‘MittRomney’, ‘JoeBiden’]:
    textList tweets is getTweets of person from twitter
    insert tweets into alltweets
```



Barack Obama
@BarackObama
This account is run by Organizing for Action staff. Tweets from the President are signed -bo.
Washington, DC · http://www.barackobama.com

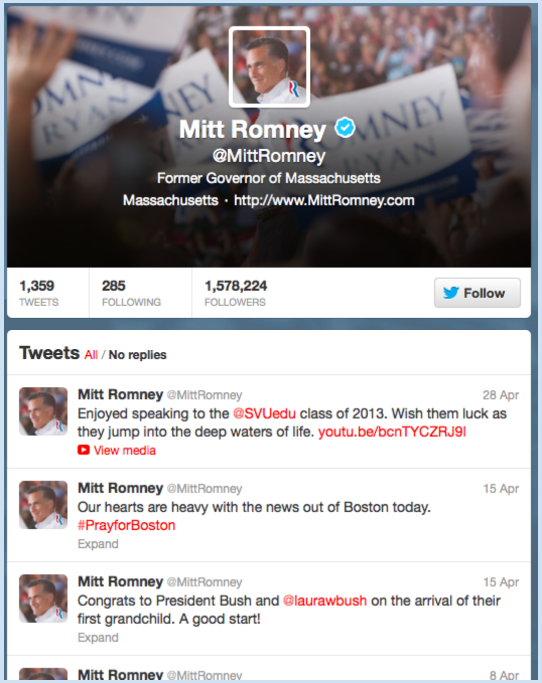
9,068 TWEETS | 663,163 FOLLOWING | 30,847,090 FOLLOWERS | Follow

Tweets All / No replies

Barack Obama @BarackObama 19h
Add your name now -- OFA.BO/s9x6FH Common-sense gun violence prevention won't happen unless Congress hears our voices.
Expand

Barack Obama @BarackObama 21h
.@OFA will deliver this petition to Congress demanding common-sense gun violence prevention. Add your name now: OFA.BO/s9x6FH
Expand

Barack Obama @BarackObama 28 Apr
Watch the full video of President Obama at the White House Correspondents' Dinner: OFA.BO/HuLETr
View summary



Mitt Romney
@MittRomney
Former Governor of Massachusetts
Massachusetts · http://www.MittRomney.com

1,359 TWEETS | 285 FOLLOWING | 1,578,224 FOLLOWERS | Follow

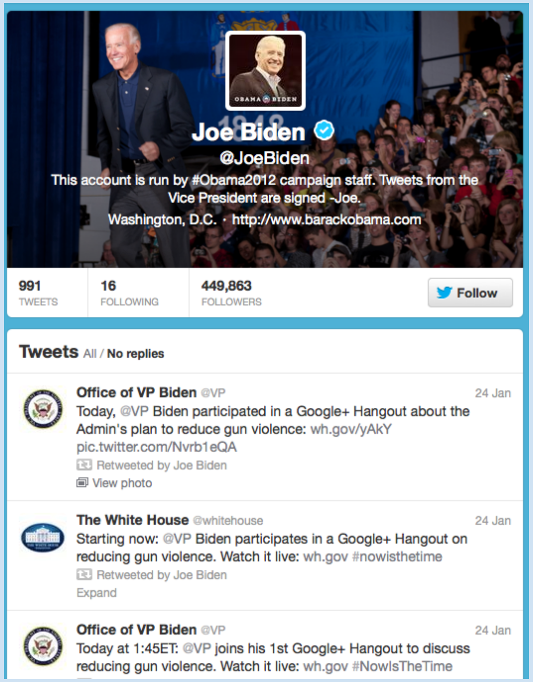
Tweets All / No replies

Mitt Romney @MittRomney 28 Apr
Enjoyed speaking to the @SVUedu class of 2013. Wish them luck as they jump into the deep waters of life. youtu.be/bcnTYCZRJ9I
View media

Mitt Romney @MittRomney 15 Apr
Our hearts are heavy with the news out of Boston today. #PrayforBoston
Expand

Mitt Romney @MittRomney 15 Apr
Congrats to President Bush and @laurawbush on the arrival of their first grandchild. A good start!
Expand

Mitt Romney @MittRomney 8 Apr



Joe Biden
@JoeBiden
This account is run by #Obama2012 campaign staff. Tweets from the Vice President are signed -Joe.
Washington, D.C. · http://www.barackobama.com

991 TWEETS | 16 FOLLOWING | 449,863 FOLLOWERS | Follow

Tweets All / No replies

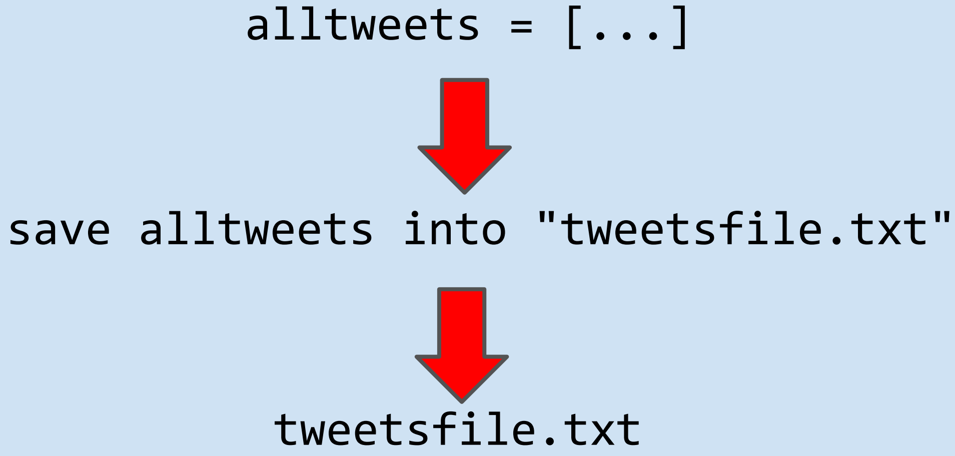
Office of VP Biden @VP 24 Jan
Today, @VP Biden participated in a Google+ Hangout about the Admin's plan to reduce gun violence: wh.gov/yAkY pic.twitter.com/Nvrb1eQA
Retweeted by Joe Biden
View photo

The White House @whitehouse 24 Jan
Starting now: @VP Biden participates in a Google+ Hangout on reducing gun violence. Watch it live: wh.gov #nowisthetime
Retweeted by Joe Biden
Expand

Office of VP Biden @VP 24 Jan
Today at 1:45ET: @VP joins his 1st Google+ Hangout to discuss reducing gun violence. Watch it live: wh.gov #NowIsTheTime

```
alltweets = [...]
```

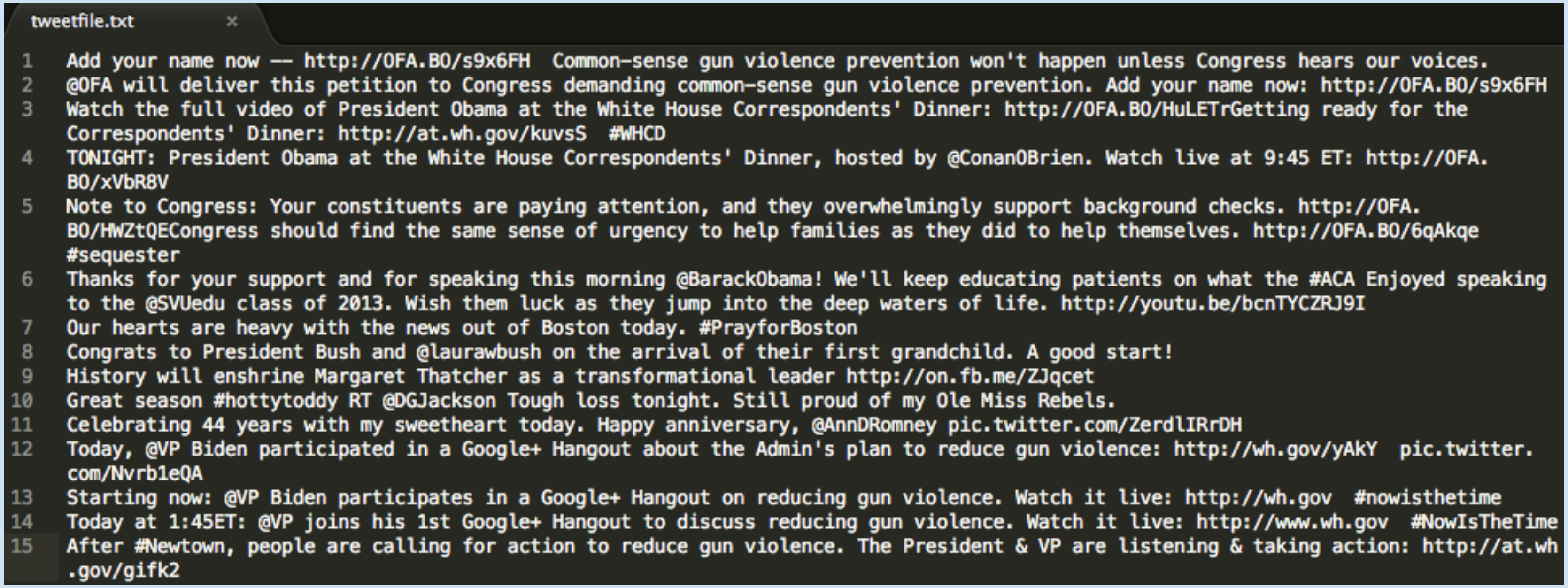
Example: Amy



```
alltweets = [...]
```

```
save alltweets into "tweetsfile.txt"
```

```
tweetsfile.txt
```



```
Add your name now -- http://OFA.BO/s9x6FH  Common-sense gun violence prevention won't happen unless Congress hears our voices.
@OFA will deliver this petition to Congress demanding common-sense gun violence prevention. Add your name now: http://OFA.BO/s9x6FH
Watch the full video of President Obama at the White House Correspondents' Dinner: http://OFA.BO/HuLETrGetting ready for the
Correspondents' Dinner: http://at.wh.gov/kuvsS  #WHCD
TONIGHT: President Obama at the White House Correspondents' Dinner, hosted by @ConanOBrien. Watch live at 9:45 ET: http://OFA.
BO/xVbR8V
Note to Congress: Your constituents are paying attention, and they overwhelmingly support background checks. http://OFA.
BO/HWZtQECongress should find the same sense of urgency to help families as they did to help themselves. http://OFA.BO/6qAkqe
#sequester
Thanks for your support and for speaking this morning @BarackObama! We'll keep educating patients on what the #ACA Enjoyed speaking
to the @SVUedu class of 2013. Wish them luck as they jump into the deep waters of life. http://youtu.be/bcnTYCZRJ9I
Our hearts are heavy with the news out of Boston today. #PrayforBoston
Congrats to President Bush and @laurawbush on the arrival of their first grandchild. A good start!
History will enshrine Margaret Thatcher as a transformational leader http://on.fb.me/ZJqcet
Great season #hottytoddy RT @DGJackson Tough loss tonight. Still proud of my Ole Miss Rebels.
Celebrating 44 years with my sweetheart today. Happy anniversary, @AnnDRomney pic.twitter.com/ZerdlIRrDH
Today, @VP Biden participated in a Google+ Hangout about the Admin's plan to reduce gun violence: http://wh.gov/yAkY  pic.twitter.
com/Nvrb1eQA
Starting now: @VP Biden participates in a Google+ Hangout on reducing gun violence. Watch it live: http://wh.gov  #nowisthetime
Today at 1:45ET: @VP joins his 1st Google+ Hangout to discuss reducing gun violence. Watch it live: http://www.wh.gov  #NowIsTheTime
After #Newtown, people are calling for action to reduce gun violence. The President & VP are listening & taking action: http://at.wh
.gov/gifk2
```

Example: Amy

```
define getTweets of (text person) searching twitter:
    text prefix is ‘http://www.twitter.com/'
    url twitterurl is combine prefix and person
    results is findText in twitterurl
    return results

for person in [‘BarackObama’, ‘MittRomney’, ‘JoeBiden’]:
    textList tweets is getTweets of person searching twitter
    insert tweets into (textList alltweets)

save alltweets into "tweetsfile.txt"
```

Project Management

1. Touch Base

2. Focus on the Big Picture

3. Respect Your Teammates

Lessons Learned

1. Know Git!

- **Version Control: Git (GitHub)**

- Repository Structure:
    - Use master by default
    - New branch for code that breaks regression tests

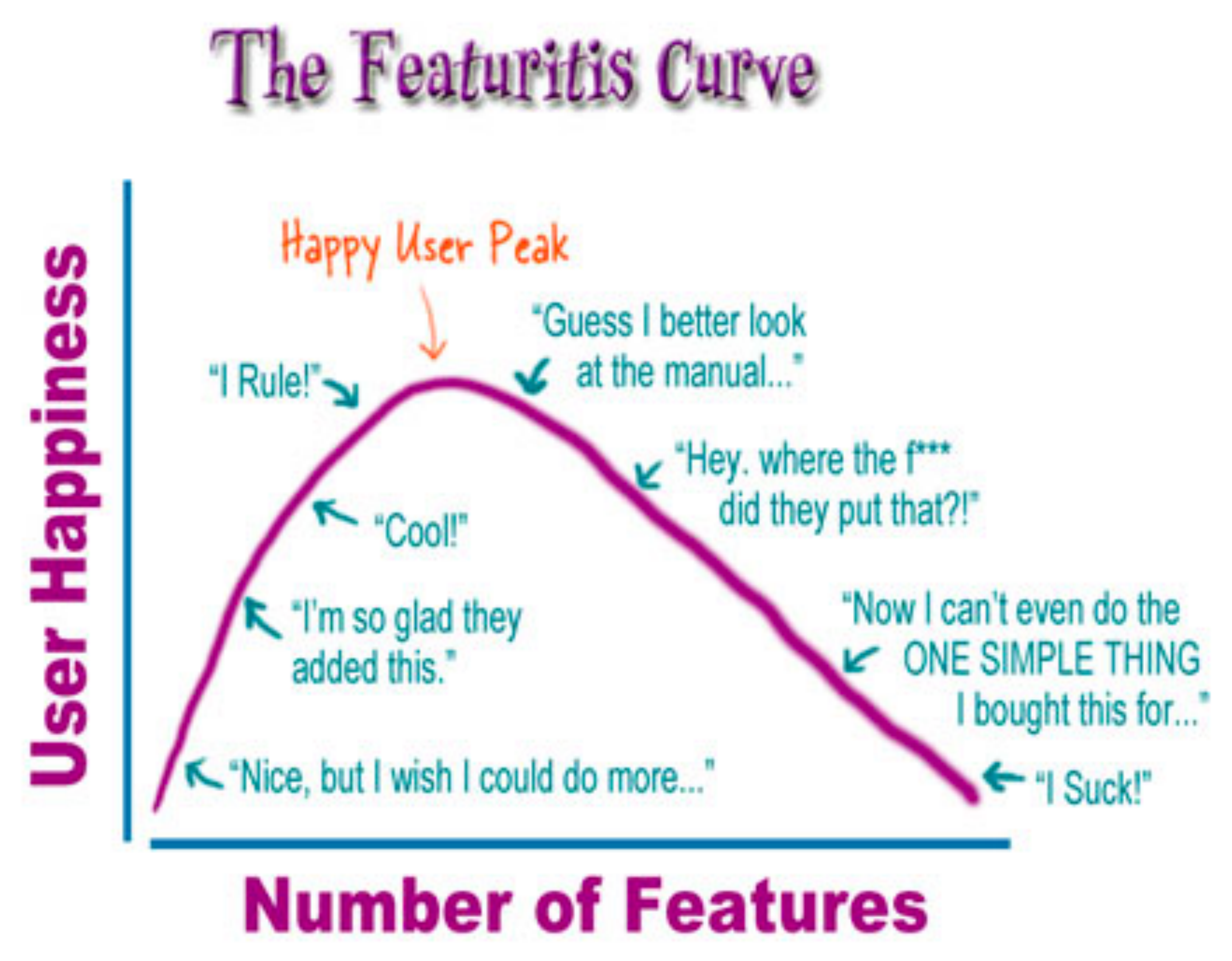
- Use descriptive commit messages!

2. Plan Well and Iterate Quickly

- Being modular up-front saves time in testing and bug-fixing

- Planning before coding is important, but early coding helps the planning process

Possible Expansions to Trowel



1. Help Journalists be Good People

- GET request header with journalist's name and email
- Easy setting of a delay between GET requests

Anderson Cooper likes being a good person.

2. Other Cool, Useful Functionality

- Crawl a news website
- findText using other html tags
- Find the similarity index between two articles

Conclusion

Journalists are **not** Computer Scientists

They have **different** goals, challenges and thought patterns

Trowel works for **them**