@maebert
#pycon2016

@maebert
#pycon2016

+ wordnik

@maebert
#pycon2016

NERD TRIVIA

ROUND 1


GREXIT
PROPOSED WITHDRAWAL OF
GREECE FROM THE EUROZONE

GRINDERS

HACKERS WHO IMPLANT ELECTRONICS INTO THEIR BODIES

LATINX

GENDER-NEUTRAL FORM OF LATINA & LATINO

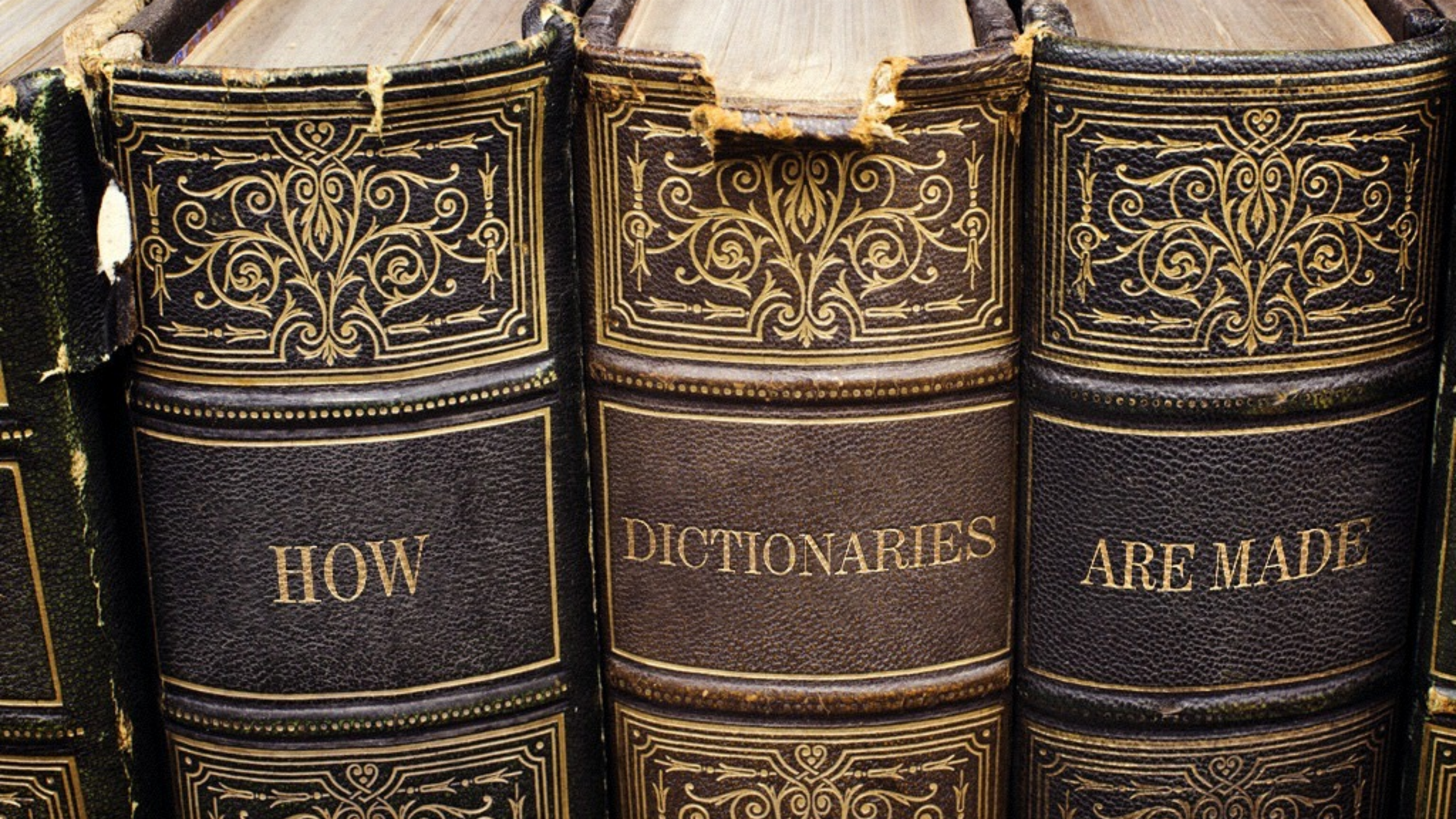

DAD BODS
SMALL FAT DEPOSITS ON OTHERWISE ATHLETICALLY BUILT MALES

BLARF

COMBINATION OF BLANKET AND SCARF

@maebert
#pycon2016


@maebert
#pycon2016
PERFECTLY
CROMULENT


HOW
DICTIONARIES
ARE MADE

@maebert
#pycon2016

SAYS URBANDICTIONARY.COM:

TOP DEFINITION

mansplain

Stating verifiable facts that are inconvenient to the feminist worldview.

by Mansplainer April 17, 2014

4876 2849

BUY THE MUG

2

Mansplain

Telling a woman that she's wrong, even when she actually is.

@maebert
#pycon2016

FREE RANGE DEFINITIONS

"Her gown was of white satin worked with gold, and had long open pendent sleeves, while from her slender and marble neck hung a cordeliere — **a species of necklace imitated from the cord worn by Franciscan friars, and formed of crimson silk twisted with threads of Venetian gold.**"

— WH Ainsworth, Windsor castle

HOW MANY WORDS ARE MISSING?

RESEARCH ARTICLE

Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel,[1,2,3,4,5]*† Yuan Kui Shen,[2,6,7] Aviva Presser Aiden,[2,6,8] Adrian Veres,[2,6,9] Matthew K. Gray,[10] The Google Books Team,[10] Joseph P. Pickett,[11] Dale Hoiberg,[12] Dan Clancy,[10] Peter Norvig,[10] Jon Orwant,[10] Steven Pinker,[5] Martin A. Nowak,[1,13,14] Erez Lieberman Aiden[1,2,6,14,15,16,17]*†

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of 'culturomics,' focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. Culturomics extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities.

pages of 1208 books. The corpus contains 386,434,758 words from 1861; thus, the frequency is 5.5×10^{-5}. The use of "slavery" peaked during the Civil War (early 1860s) and then again during the civil rights movement (1955–1968) (Fig. 1B)

In contrast, we compare the frequency of "the Great War" to the frequencies of "World War I" and "World War II". References to "the Great War" peak between 1915 and 1941. But although its frequency drops thereafter, interest in the underlying events had not disappeared; instead, they are referred to as "World War I" (Fig. 1C).

These examples highlight two central factors that contribute to culturomic trends. Cultural change guides the concepts we discuss (such as "slavery"). Linguistic change, which, of course, has cultural roots, affects the words we use for those concepts ("the Great War" versus "World War I"). In this paper, we examine both linguistic changes, such as changes in the lexicon and grammar, and cul-

@maebert
#pycon[illegible]

HOW MANY WORDS ARE MISSING?

RESEARCH ARTICLE

Quantitative Analysis of [illegible] Using Millions of Digitiz[illegible]

Jean-Baptiste Michel,[1,2,3,4,5]*† Yuan Kui Shen,[2,6,7] Aviva Presser Aiden,[2,6,8] Adrian Veres,[2,6,9] Matthew K. Gray,[10] The Google Books Team,[10] Joseph P. Pickett,[11] Dale Hoiberg,[12] Dan Clancy,[10] [illegible],[10] Steve[illegible]nker,[5] Martin A. Now[illegible]den[1,2,6,[illegible],16,17]*†

We constructed [illegible] ining abo[illegible]% of all [illegible]ks [illegible] prin[illegible] A[illegible] corpus enables u[illegible] ga[illegible] te[illegible] quantitat[illegible] We surv[illegible] he[illegible] ter[illegible] of [illegible] focusing on ling[illegible] a[illegible] cu[illegible] p[illegible] om[illegible] er[illegible] fir[illegible] Eng[illegible] dag[illegible] twe[illegible] 1800 and 2000. We show how this approach can provide insights about fields as diverse as [illegible] the evolution of grammar, collective memory, the adoption of technology, the [illegible] censorship, and historical [illegible]emi[illegible]gy. [illegible]ro[illegible]ds [illegible] inquiry to a wide array of [illegible] pl[illegible]ne[illegible]pa[illegible]g t[illegible]oci[illegible]em[illegible]

pages of [illegible] corpus contains 386,434 [illegible] the frequency is 5.5[illegible] [illegible]ked during the [illegible] [illegible]in during th[illegible] [illegible]ig. 1B) [illegible] of "the G[illegible] [illegible] War I" ar[illegible] [illegible]e Great [illegible] [illegible]ough i[illegible] [illegible]e un- d[illegible] [illegible] they a[illegible] [illegible]actors tha[illegible] [illegible]hange

ONE MILLION WORDS

@maebert
#pycon2016

THE PLAN

WHAT COULD POSSIBLY GO WRONG?

MISSING WORDS

x3M

PREPROCESSING

1.8M

BING SEARCH

x50

HTML PARSING

“...”

DETECT LANGUAGE

“...”

DETECT FRD

“...” → FRD

SAVE TO DB

FRD →

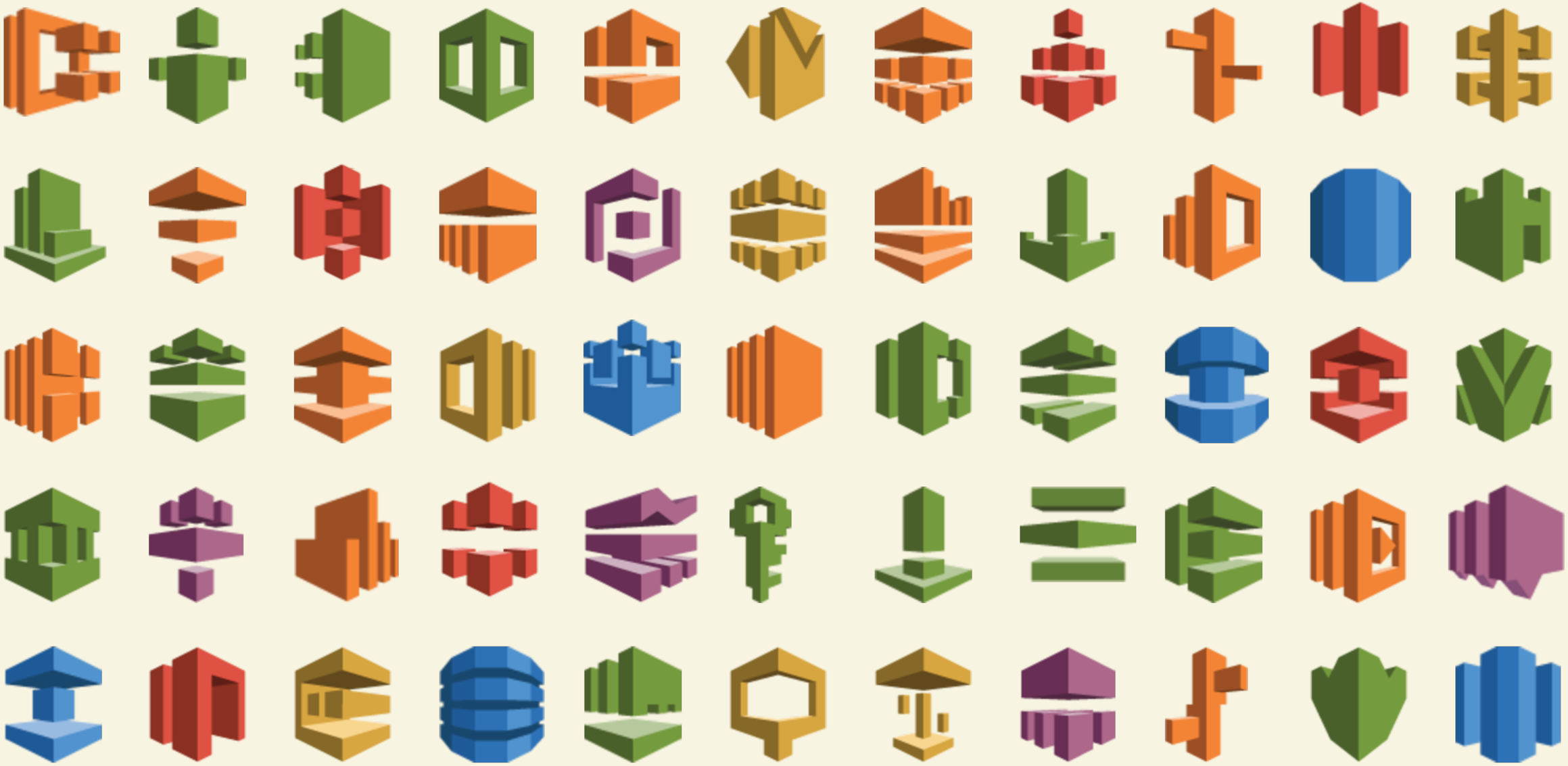
12 YEARS
8 MONTHS
6 DAYS
15 HOURS

@maebert
#pycon2016

NERD TRIVIA

ROUND 2

@maebert #pycon2016

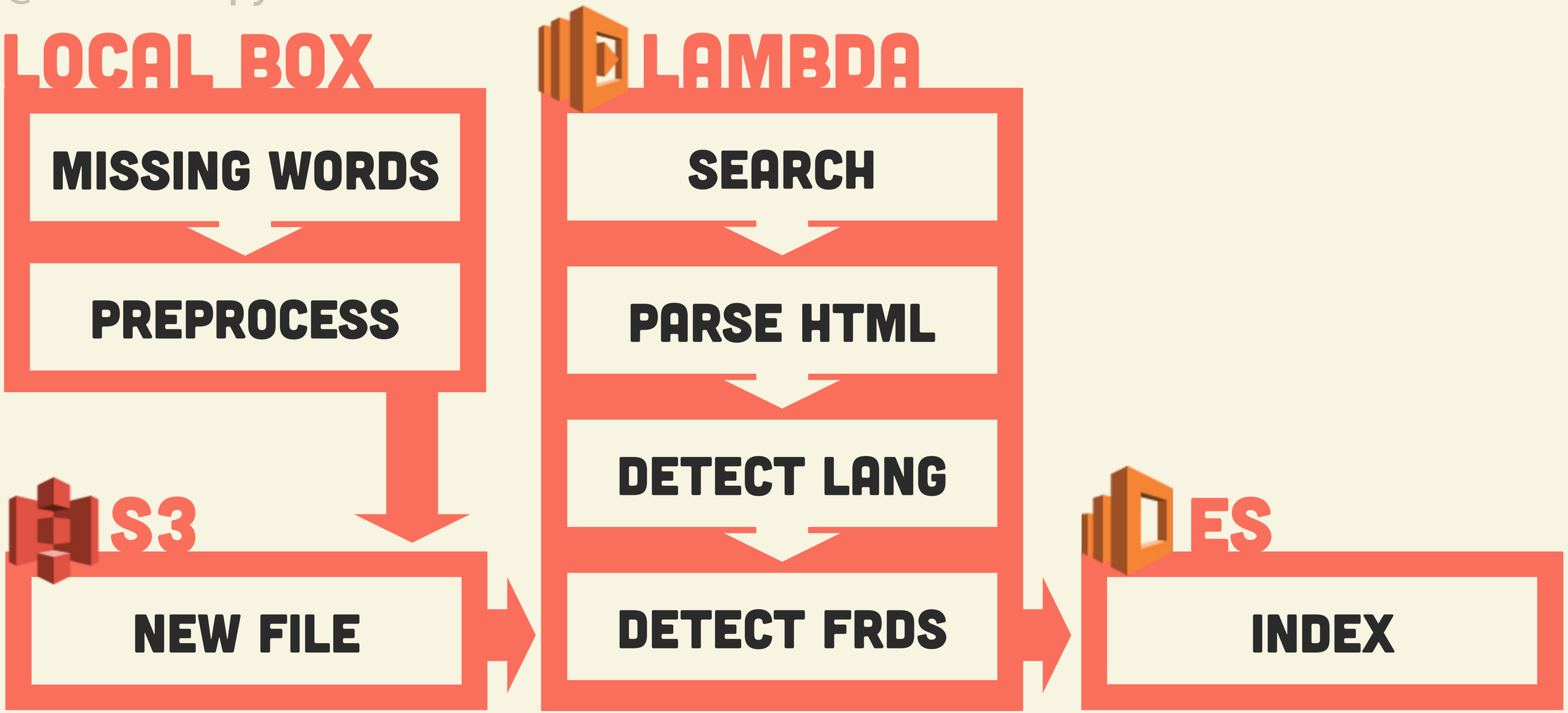


S3

EC2

ElasticSearch

Lambda


LOCAL BOX
MISSING WORDS
PREPROCESS
S3
NEW FILE
LAMBDA
SEARCH
PARSE HTML
DETECT LANG
DETECT FRDS
ES
INDEX

MISSING WORDS

x3M

PREPROCESSING

1.8M

BING SEARCH

HTML x50

HTML PARSING

HTML "..."

DETECT LANGUAGE

"..."

DETECT FRD

"..." FRD

SAVE TO DB

FRD

adeiladu

bialya

conquistar

DL 3 Aminoisobutyric

eu oi oa ou

frappul

galen's bondage etymologies

h’ors d’oeuvres

i collect bizarre animal confectionery dioramas

janky

kryogenkryokonitekryoscopy

list of unusual deaths

macÃƒÆ’Ã‚Â£Ãƒâ€

naaaaa

okcupid uranium-thorium dating

paenismus

quaestuary

revoltingly viviparous

shit my pt says

the dogs of slavery, misgovernment, and ostracism

outsnark

vésigniéite

woggle

xenofeminism

yebo

zyxnoid

adeiladu

bialya

conquistar

DL 3 Aminoisobutyric

eu oi oa ou

frappul

galen's bondage etymologies

h’ors d’oeuvres

i collect bizarre animal confectionery dioramas

janky

kryogenkryokonitekryoscopy

list of unusual deaths

macÃƒÆ’Ã‚Â£Ãƒâ€

naaaaa

okcupid uranium-thorium dating

paenismus

quaestuary

revoltingly viviparous

shit my pt says

the dogs of slavery, misgovernment, and ostracism

outsnark

vésigniéite

woggle

xenofeminism

yebo

zyxnoid

adeiladu

bialya

conquistar

DL 3 Aminoisobutyric

eu oi oa ou

frappul

galen's bondage etymologies

h’ors d’oeuvres

i collect bizarre animal confectionery dioramas

janky

kryogenkryokonitekryoscopy

list of unusual deaths

macÃƒÆ’Ã‚Â£Ãƒâ€

naaaaa

okcupid uranium-thorium dating

paenismus

quaestuary

revoltingly viviparous

shit my pt says

the dogs of slavery, misgovernment, and ostracism

outsnark

vésigniéite

woggle

xenofeminism

yebo

zyxnoid

adeiladu

bialya

conquistar

DL 3 Aminoisobutyric

eu oi oa ou

frappul

galen's bondage etymologies

h’ors d’oeuvres

i collect bizarre animal confectionery dioramas

janky

kryogenkryokonitekryoscopy

list of unusual deaths

macÃƒÆ’Ã‚Â£Ãƒâ€

naaaaa

okcupid uranium-thorium dating

paenismus

quaestuary

revoltingly viviparous

shit my pt says

the dogs of slavery, misgovernment, and ostracism

outsnark

vésigniéite

woggle

xenofeminism

yebo

zyxnoid

adeiladu

bialya

conquistar

DL 3 Aminoisobutyric

eu oi oa ou

frappul

galen's bondage etymologies

h’ors d’oeuvres

i collect bizarre animal confectionery dioramas

janky

kryogenkryokonitekryoscopy

list of unusual deaths

macÃƒÆ’Ã‚Â£Ãƒâ€

naaaaa

okcupid uranium-thorium dating

paenismus

quaestuary

revoltingly viviparous

shit my pt says

the dogs of slavery, misgovernment, and ostracism

outsnark

vésigniéite

woggle

xenofeminism

yebo

zyxnoid

```python
def valid_term(term):
    words = term.split()
    exclusion_rules = (
        any(len(word) > 15 for word in words),
        len(words) > 5,
        any(c in term for c in "\u2588,!?:1234567890"),
        sum(ord(c) > 255 for c in term) > 2,
        all(len(word) < 3 for word in words)
    )
    return not any(exclusion_rules)
```

adeiladu
bialya
conquistar
DL 3 Aminoisobutyric
eu oi oa ou
frappul
galen's bondage etymologies
h’ors d’oeuvres
i collect bizarre animal confectionery dioramas
janky
kryogenkryokonitekryoscopy
list of unusual deaths
macÃƒÆ’Ã‚Â£Ãƒâ€
naaaaa
okcupid uranium thorium dating
paenismus
quaestuary
revoltingly viviparous
shit my pt says
the dogs of slavery, misgovernment, and ostracism
outsnark
vésigniéite
woggle
zyxnoid


@maebert
#pycon2016
ExcuseMyFrench
LanguageDetection

MISSING WORDS

x3M

PREPROCESSING

1.8M

BING SEARCH

HTML x50

HTML PARSING

HTML "..."

DETECT LANGUAGE

"..."

DETECT FRD

"..." FRD

SAVE TO DB

FRD

DETECTING A LANGUAGE

```
from collections import defaultdict

def trigram_freq(text):
    trigrams = [text[k:k+3] for k in range(len(text)-2)]
    freq = defaultdict(float)
    for trigram in trigrams:
        freq[trigram] += 1.0 / len(trigrams)
    return freq
```

DETECTING A LANGUAGE

```
languages = {
    "english": trigram_freq("a quick brown fox jumps…”),
    "italian": trigram_freq("ma la volpe col suo balzo…”),
    "klingon": trigram_freq("SoH 'ej SenwI' rIlwI' je …”)
}

def detect_language(text):
    scores = defaultdict(float)
    for trigram, text_freq in trigram_freq(text).items():
        for lang, lang_freq in languages.items():
            scores[lang] += lang_freq[trigram] * text_freq
    return max(scores, key=scores.get)
```

DETECTING A LANGUAGE (GHETTO REMIX)

```
stopwords = """all just being over both through its
before herself had should to only under ours has do
them his very they not now him nor did""".split()

def is_english(text):
    words = text.split()
    n_stopwords = sum(w in stopwords for w in words)
    return float(n_stopwords) / len(words) >= 0.12
```


@maebert
#pycon2016
MACHINE
LEARNING

MISSING WORDS

x3M

DETECT FRD

“...” → FRD

SAVE TO DB

FRD →

PREPROCESSING

1.8M

DETECT LANGUAGE

“...” →

BING SEARCH

— → HTML x50

HTML PARSING

HTML → “...”

TEXT CLASSIFICATION

Romney is trying to prevent a stampede to Trump of **Vichy Republicans**, collaborationists coming to terms with the occupation of their party.

vs.

There are a lot of elected **Vichy Republicans** who don't know how to do anything but lose, or kowtow to an authority figure.

WHAT PEOPLE THINK MACHINE LEARNING IS ABOUT

ALLURING BRAIN-INSPIRED ALGORITHMS

"BIG DATA"

WHAT MACHINE LEARNING IS ACTUALLY ABOUT

PICKING THE RIGHT FEATURES

HAVING CLEAN, STRONG DATA

CHOOSING THE RIGHT ALGORITHM

@maebert
#pycon2016

TEXT CLASSIFICATION

TEXT CLASSIFICATION IN PYTHON

```
training_data = [
    ("Horror vacui is a latin expression
      that means 'fear of emptiness'", 1),

    ("She put the cordeliere down next to a cap
      of black velvet faced with white satin", 0),

    ("Abusing anyone, I was told, violated Islamic
     tenets against zulm, or cruelty.", 1)
]

sentences, classes = zip(*training_data)
```

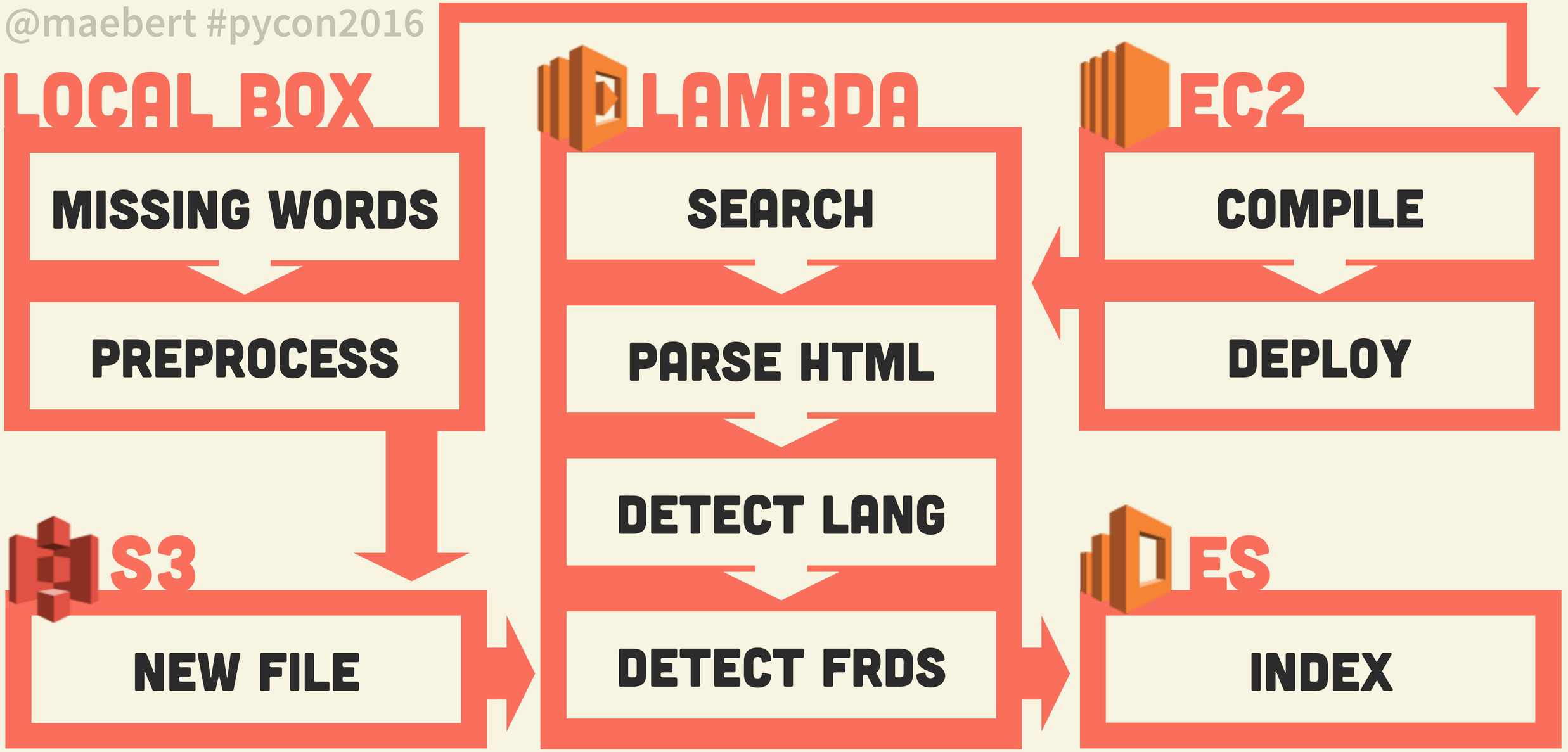
TEXT CLASSIFICATION IN PYTHON

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB

tfidf = TfidfVectorizer()
vectors = tfidf.fit_transform(sentences)
classifier = MultinomialNB()
classifier.fit(vectors, classes)

# Predict something
s = "A rose is a rose"
vectorised_s = tfidf.transform([s])
classifier.predict(vectorised_s)
```

@maebert #pycon2016




SETTING UP AN
AWS LAMBDA ENV
ON EC2

SETTING UP EC2

```
$ yum -y install blas lapack atlas-sse3-devel
$ virtualenv ~/stack; source ~/stack/bin/activate
$ dd if=/dev/zero of=/swapfile bs=1024 count=1500000
$ mkswap /swapfile; chmod 0600 /swapfile; swapon /swapfile
$ pip install numpy scipy pandas sklearn
$ strip `find ~/stack/lib/python2.7/ -name="*.so"`
$ pushd ~/stack/lib/python2.7/site-packages/
$ zip -r9q ~/lambda.zip * ; popd
$ aws s3 cp ~/lambda.zip s3://my_bucket/lambda.zip
$ aws lambda update-function-code --s3-bucket my_bucket \
  --s3-key lambda.zip --function-name lambda_function
```

FULL AWS LAMBDA WALKTHROUGH:

BIT.LY/ML_AWS_LAMBDA

ALL THE CODE:

SERAPIS
SCALABLE
WORD GOBBLER

GITHUB.COM/SUMMER.AI/SERAPIS


MANUEL EBERT @MAEBERT