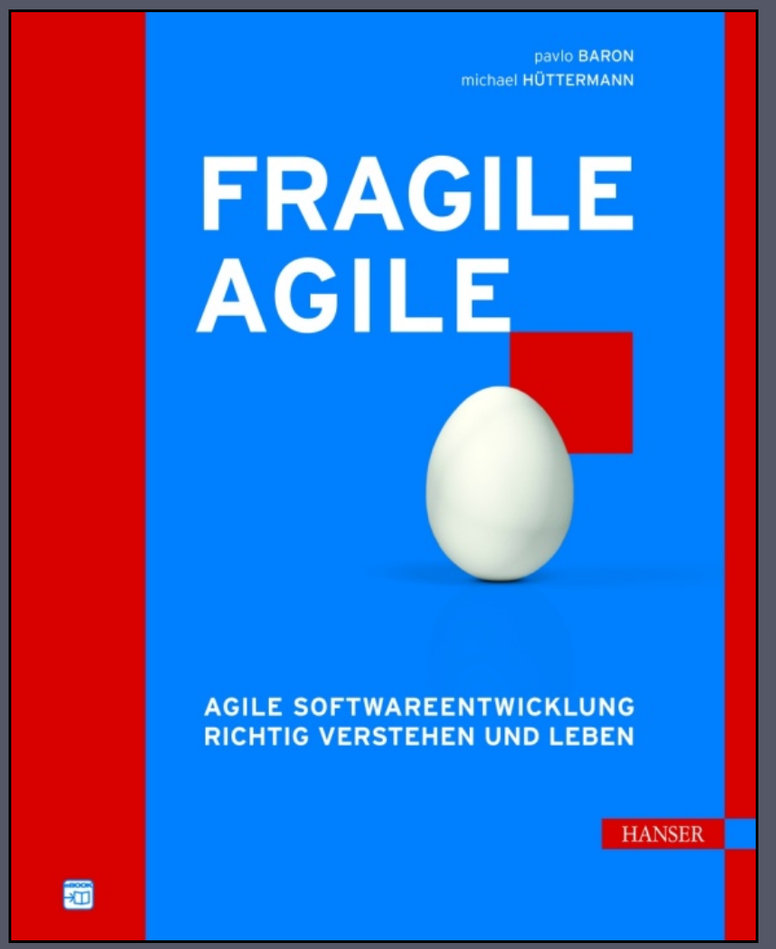
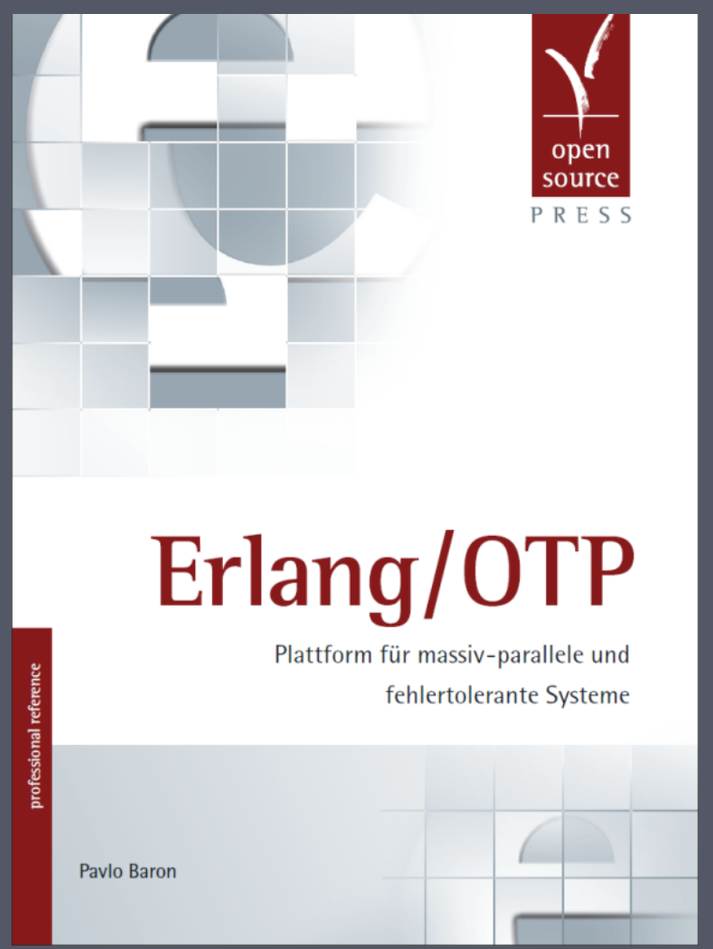
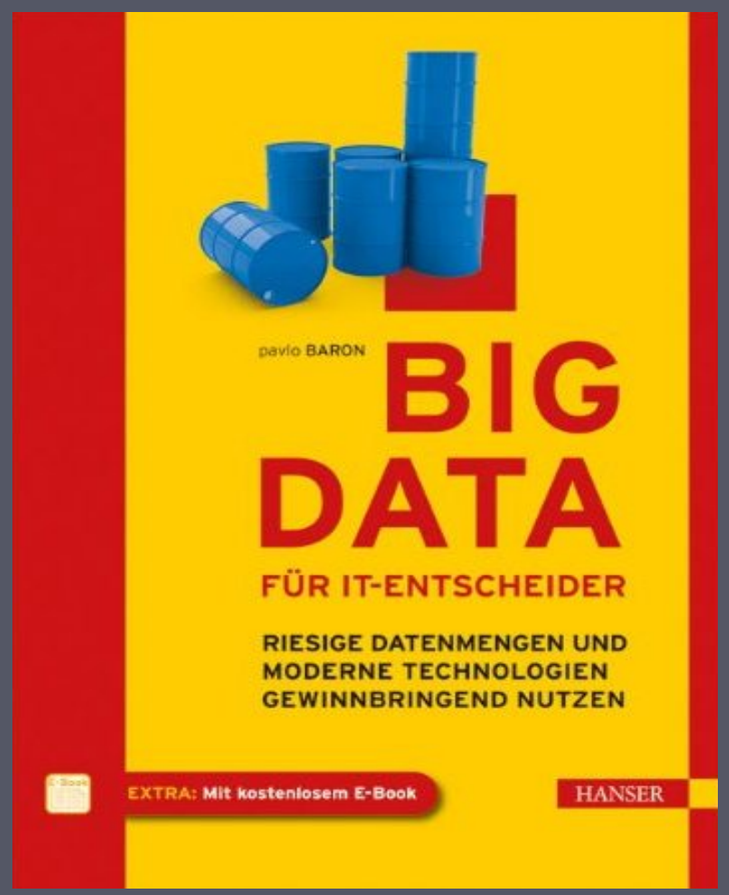
100% Big Data
0% Hadoop
0% Java

Pavlo Baron, codecentric



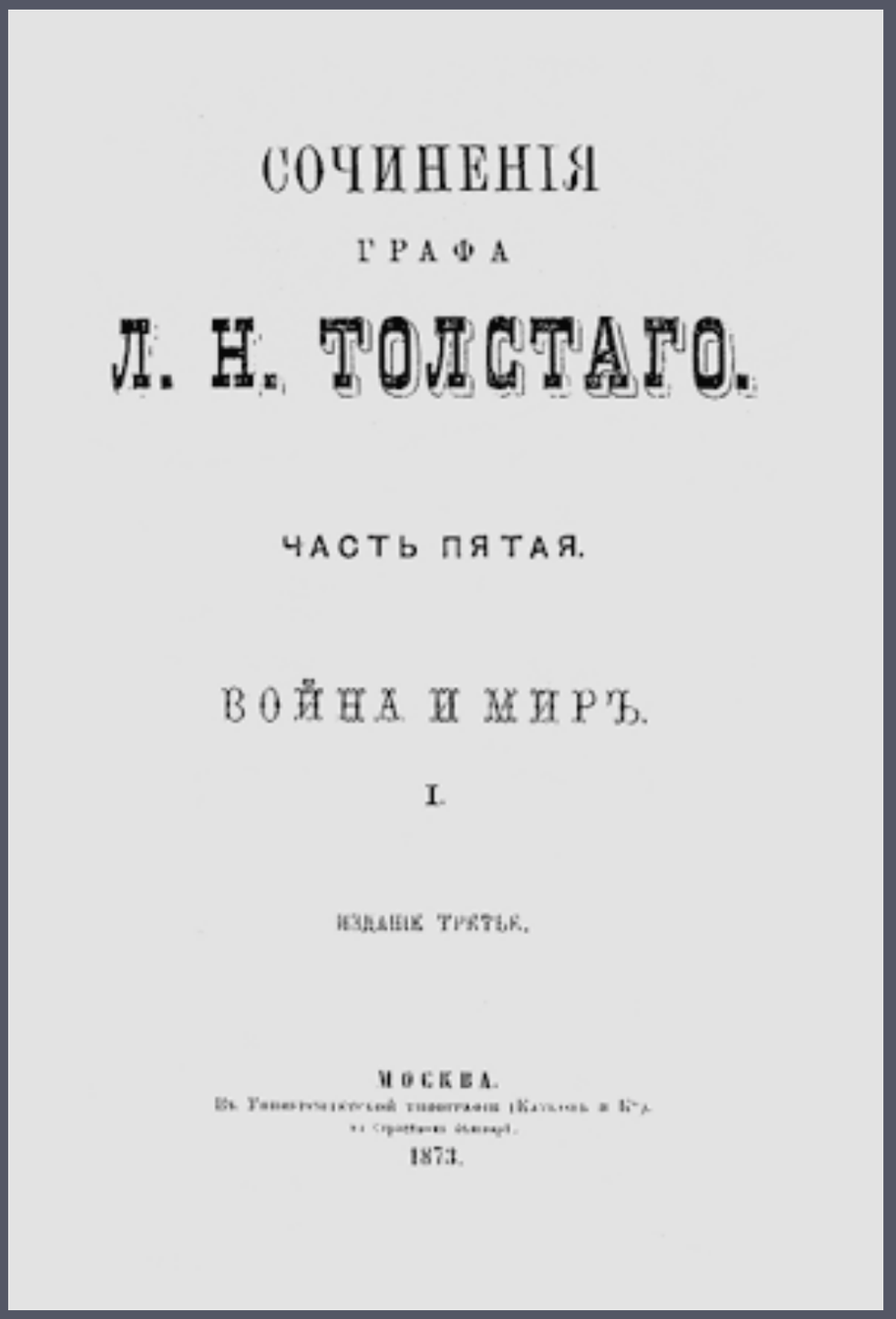




- pavlo.baron@codecentric.de
- @pavlobaron
- github.com/pavlobaron

I don't rant.
I just express
my opinion.

So here is the short story...



СОЧИНЕНІЯ

ГРАФА

Л. Н. ТОЛСТАГО.

ЧАСТЬ ПЯТАЯ.

ВОЙНА И МИРЪ.

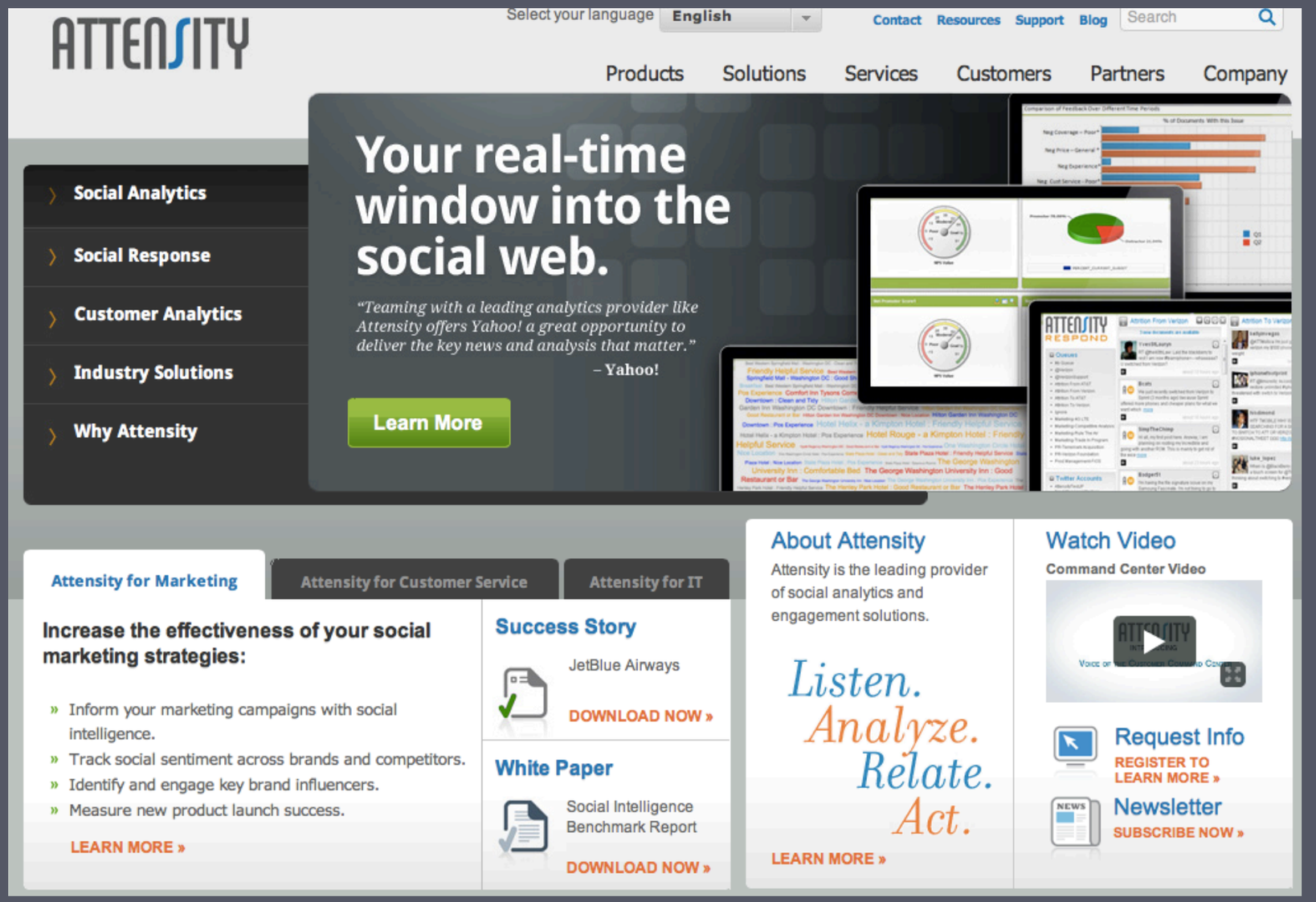
I.

ИЗДАНІЕ ТРЕТЬЕ.

МОСКВА.

1873.

sitting there, listening...



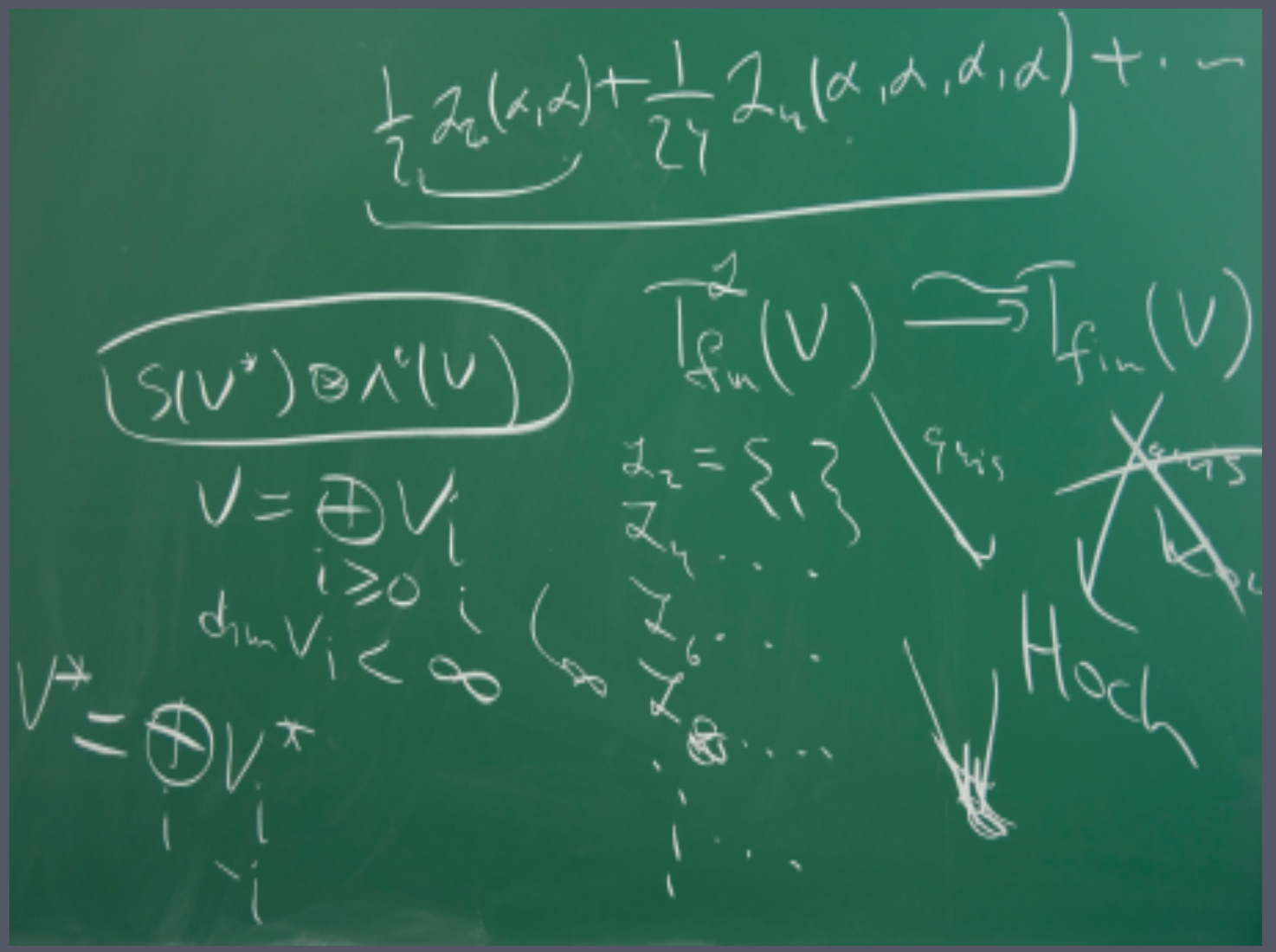
presented as Houdini magic...

so you telling me...

it's smoke and mirrors?

Looks more like NLP to me...

Sounds like a lot of math, too...



And also smells like ML...

methinks: I can tinker that...

So I need some Big Data, where people say what they think before they think what they say...

I need to drink my Big Data warm, straight from the fire hose...

Twitter fire hose, how do I drink you?..

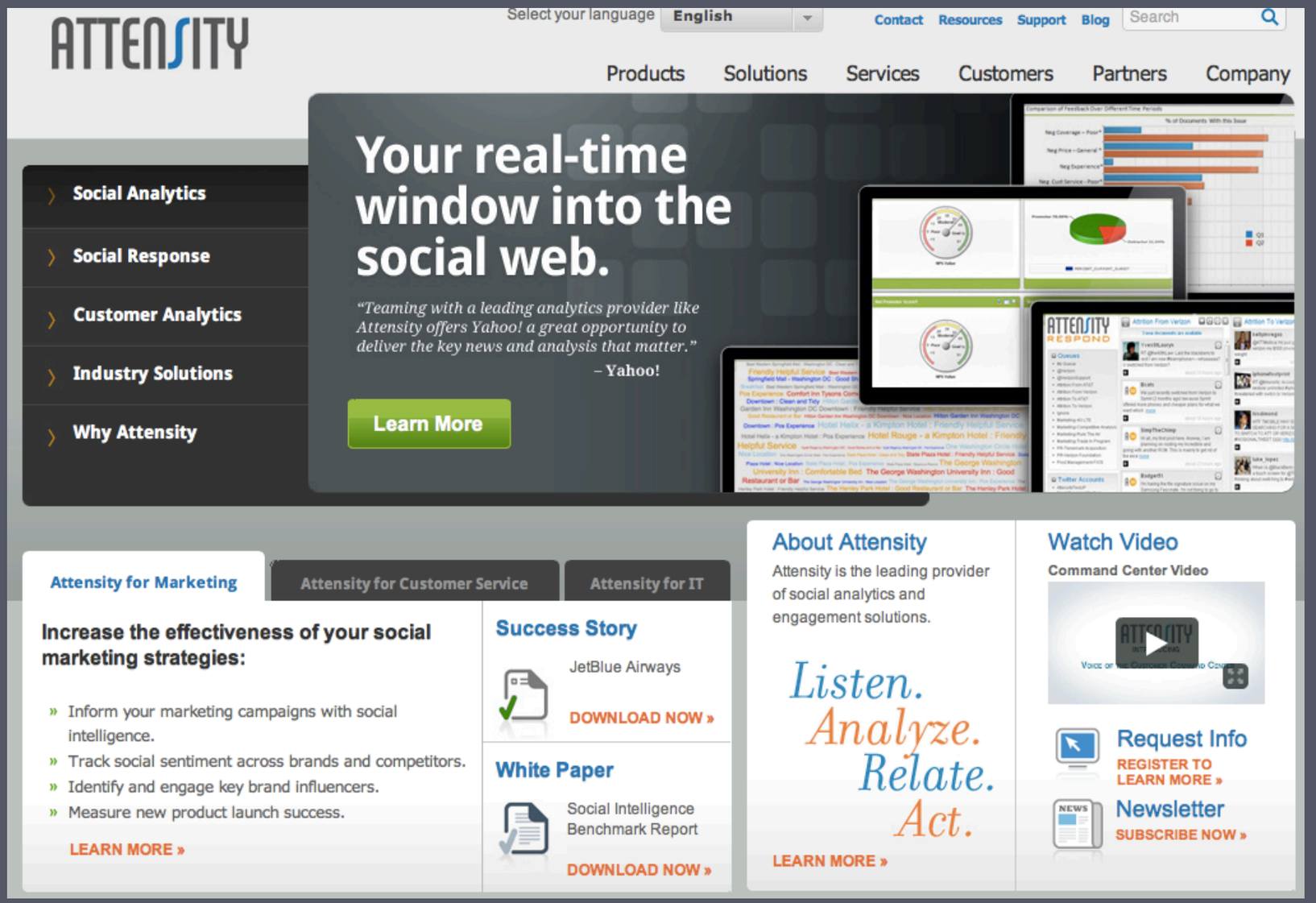
- Firehose can only be accessed by (officially) DataSift and Gnip :(
- Gardenhose access is for research and education only, and seems to be dead :((
- Poor man's alternative is the public stream sampling
  random 1% of the firehose :(((
- But anyway, it's up to 2000 tweets per minute

Wait a minute...

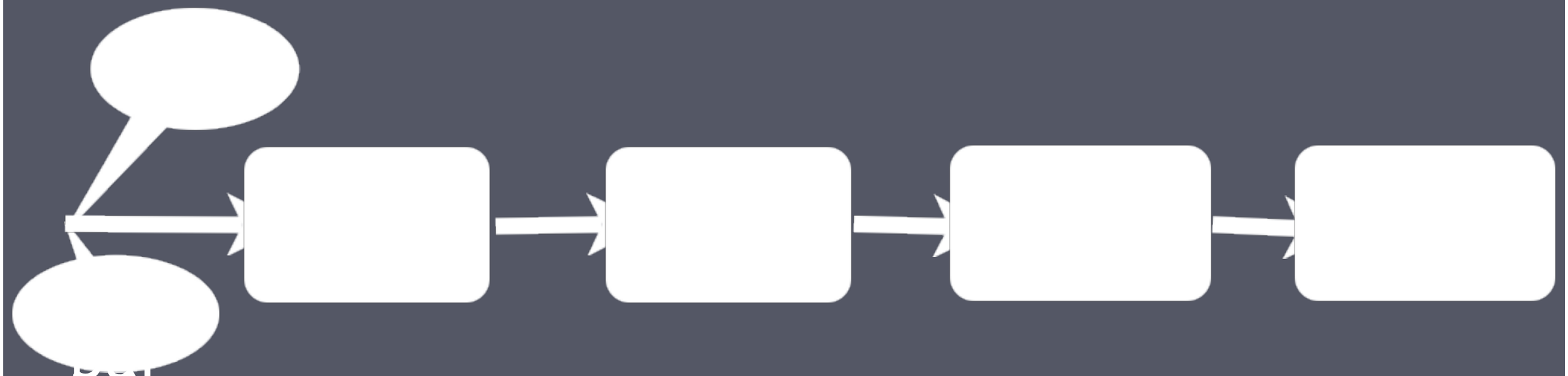
Just 2000 tweets per minute?

2000???????

Don't ask me. Remember? Sitting there, listening...



Anyway, I sketched some bubbles...



Now I need some adequate basic tech...

There is a lot of stuff in the Java world I can use for that...

But strange things come to my mind...

I like the JVM

- complex, proved tech
- “mechanical sympathy” possible
- big ecosystem
- large community
- bright guys working on it

But strange things come to my mind...

~/.m2

Big Data on the JVM

- Hadoop
- Pig
- Storm, Esper and whatnot (CEP)
- Mahout
- tons of libs and frameworks and middleware
- big part of the hype

But strange things come to my mind...

And there is also this...

And this...

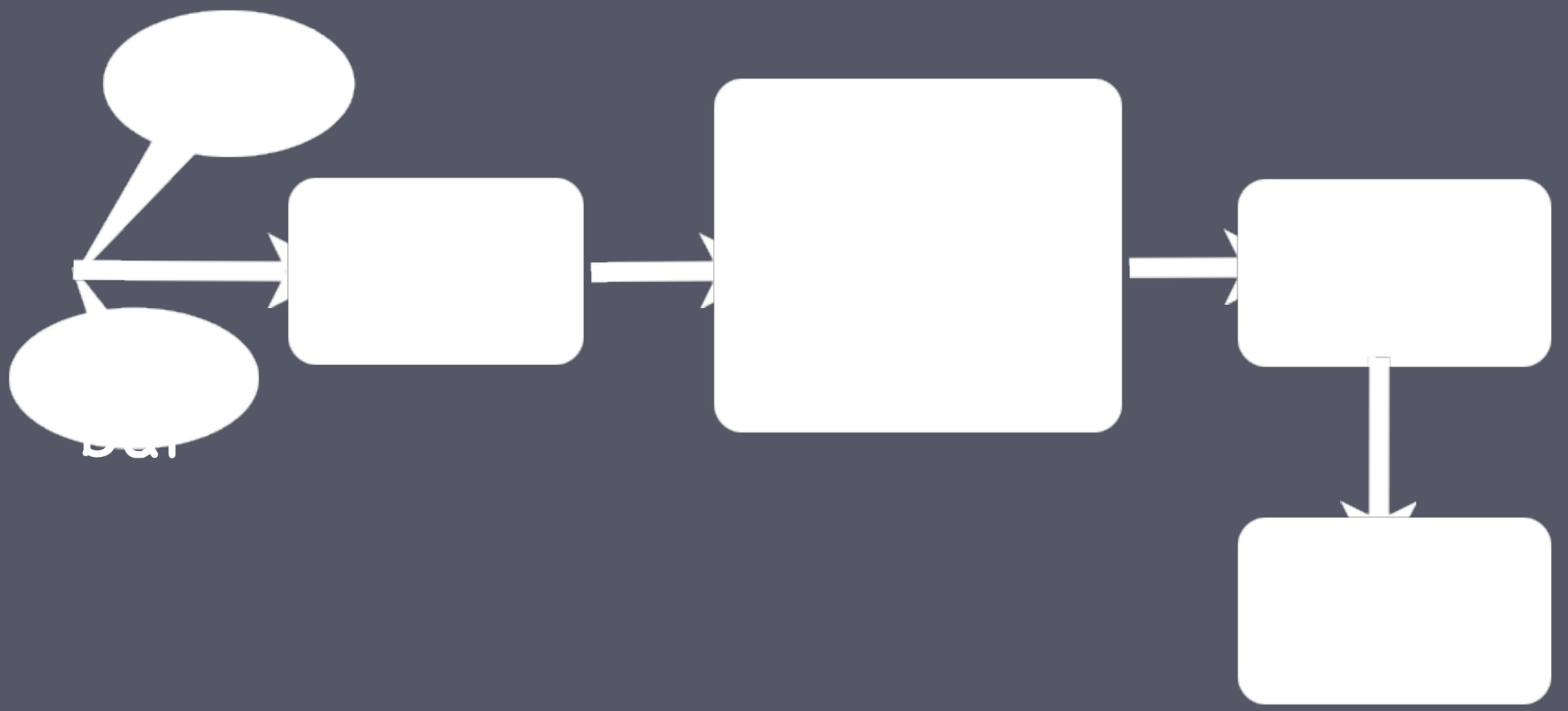
Pavlo Baron @pavlobaron 27 Sep

slowly realising that I will never have something they call success 'cause I simply refuse riding the mainstream wave...

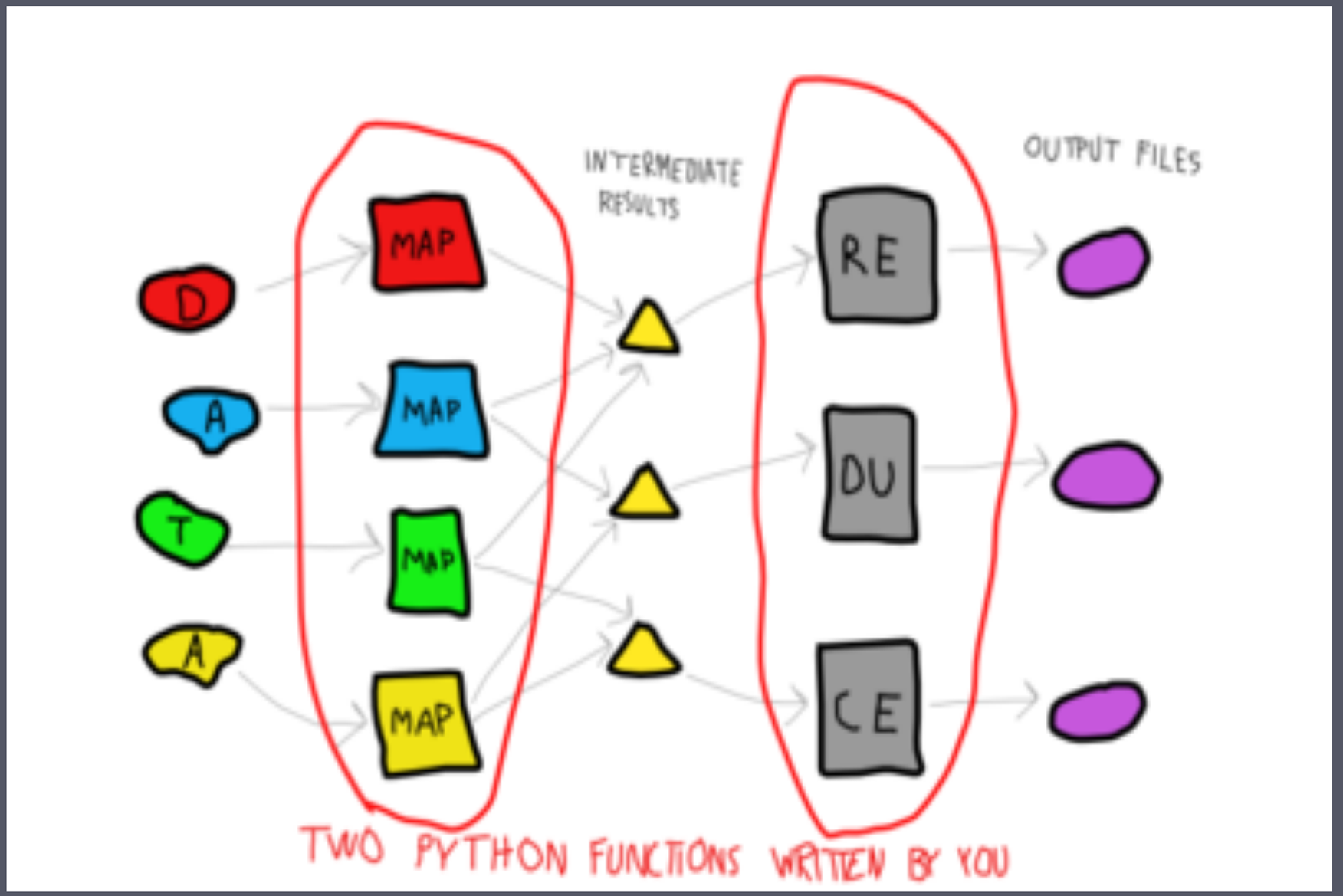
Expand

So I just decided to combine Erlang based software that I delved into with Python hacking that I wanted to do more of...

So I sketched some concrete bubbles...



But wait, why don't I do multi-phase map/reduce?..



´cause if you want to process (Big) data being streamed and you want it to work, you don't map/reduce it. You simply can't...

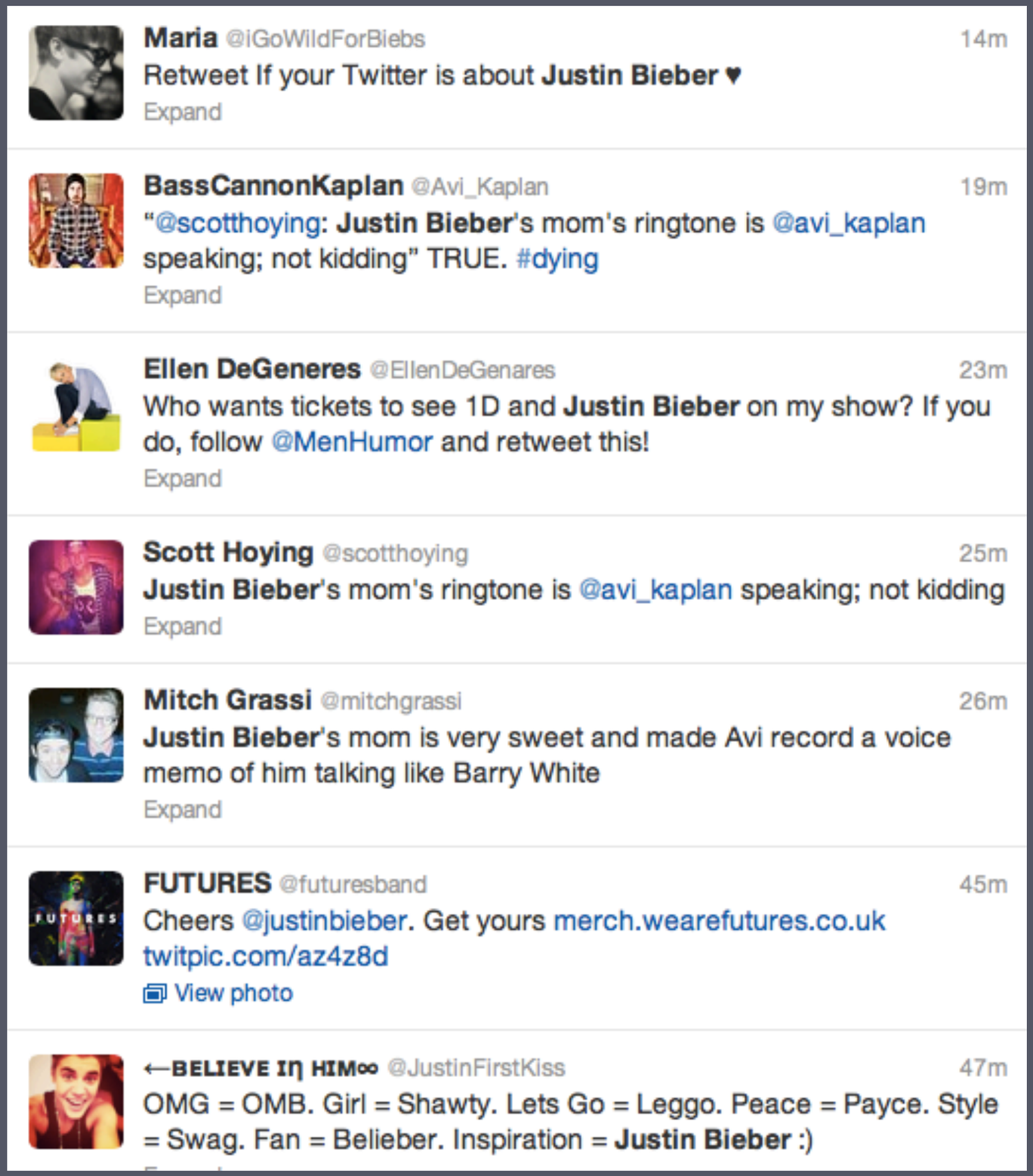
And still it has nothing to do with real-time. You can call it "near real-time", or even "as fast as possible" or "while I order a pizza"

Anyway, everything is "boringly simple" in this picture.

Except NLTK...

What I thought first is that I will use NLTK to analyze if someone rants, but it came different...

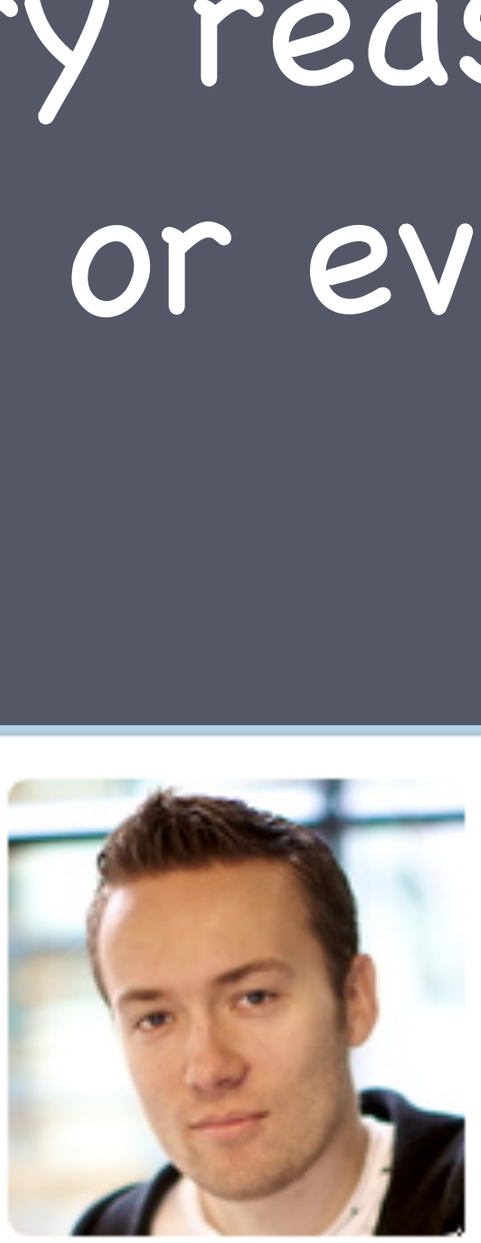
The flood of the Beliebers...



More than 60% of the sample stream is useless garbage...

So I need to filter it.
Beliebers are clear, but
what are the other criteria?..

Try reasonable user names or even real names?..



DHH

@dhh

Creator of Ruby on Rails, Partner at 37signals, Co-author of NYT Best-Seller Rework, and racing driver in ALMS.

Chicago, USA · http://david.heinemeierhansson.com

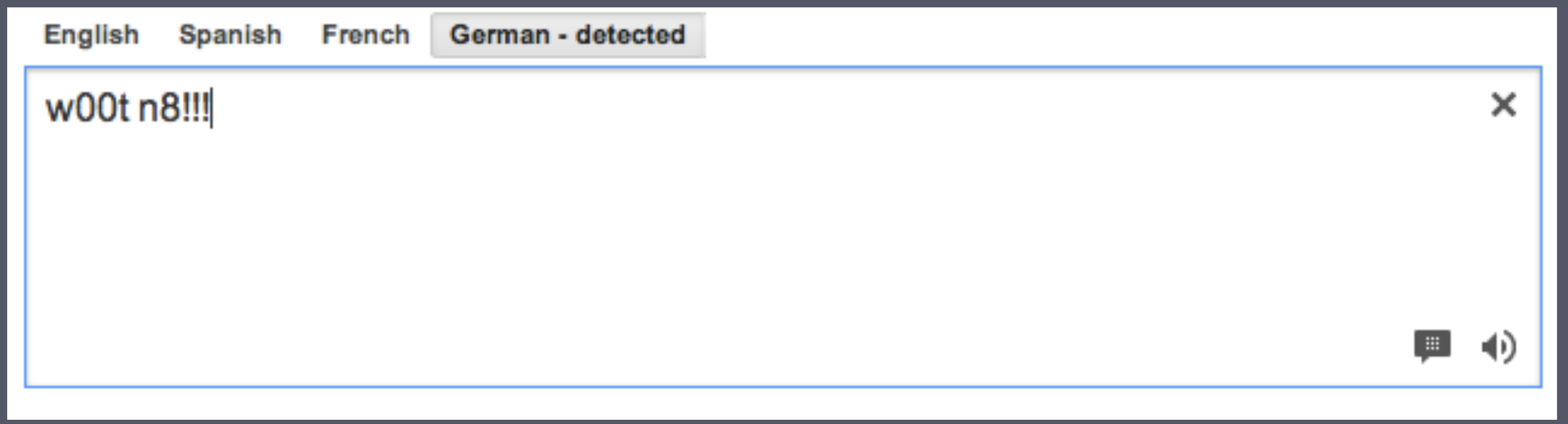
How to tell a bot from a human, well knowing that (user) names can be, well, anything?..

Absurd profile bio? Forget it!..

Correct location? Forget it!..

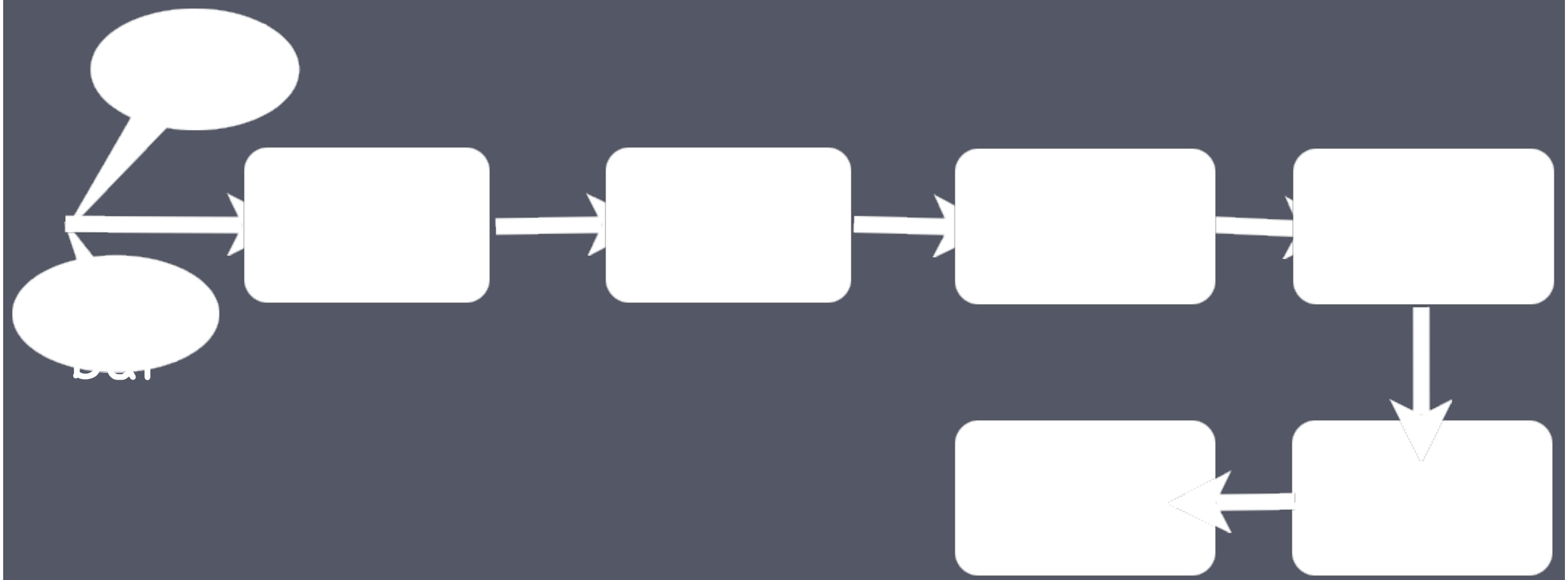
Correct user specified location? Forget it!..

Correct language? Forget it!..

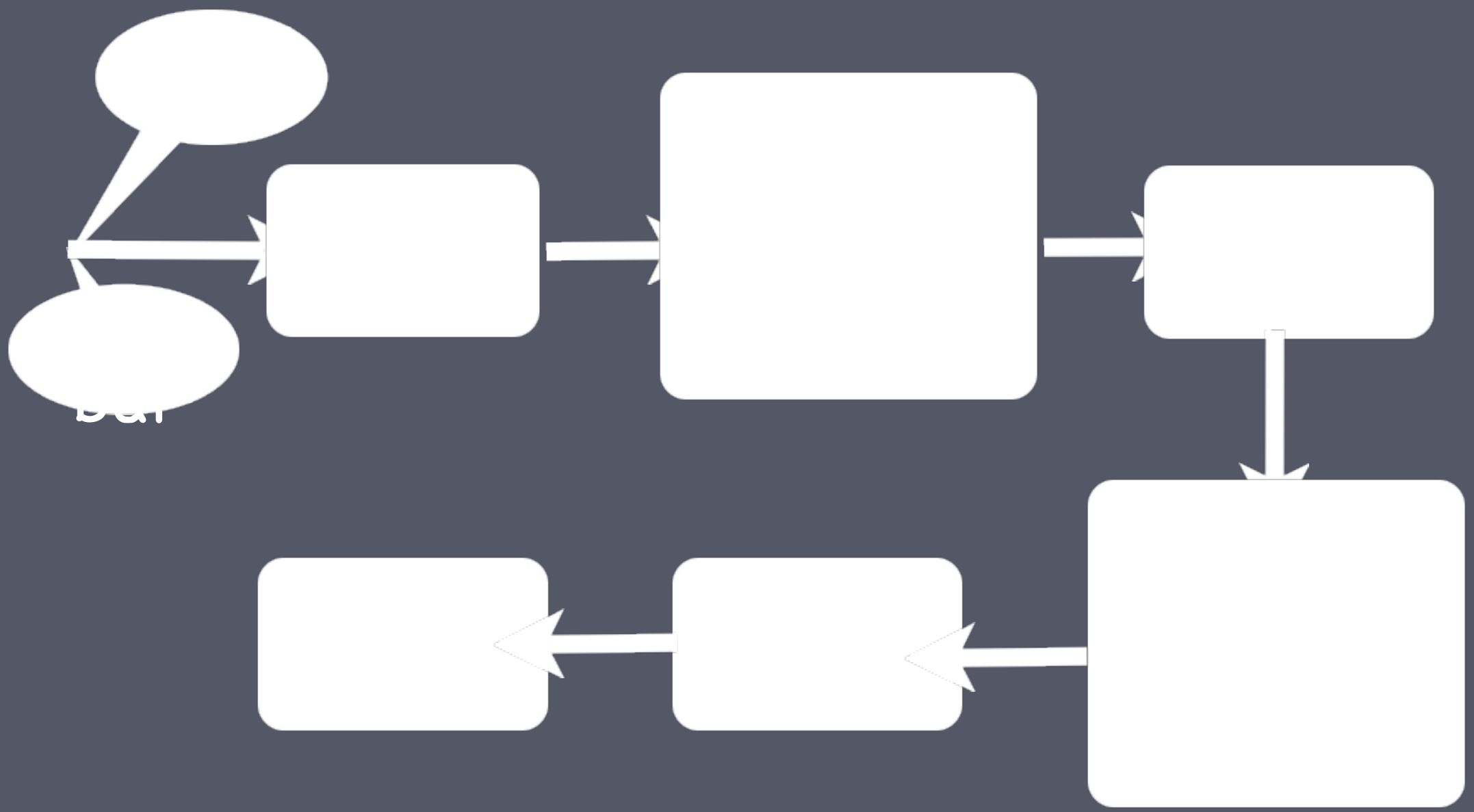


I can only do my best.
That means no filter on location. No filter on profile bio. Using NLTK to classify between English and Spanish (!!!) through tinkering

So I resketch my bubbles...



And my concrete bubbles...



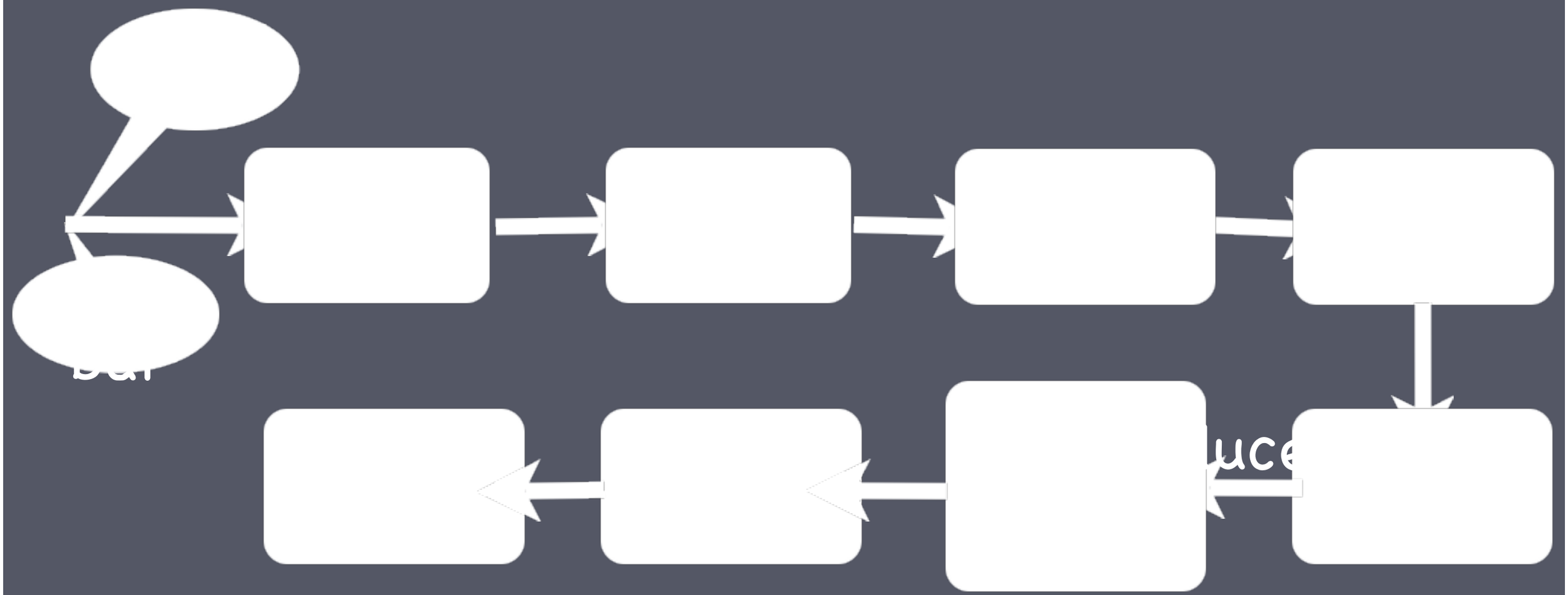
I'm careful now. What else can I filter out?..

Nothing. And I also need to find more users - it's not enough to accept that few mostly useless data coming through the sample stream...

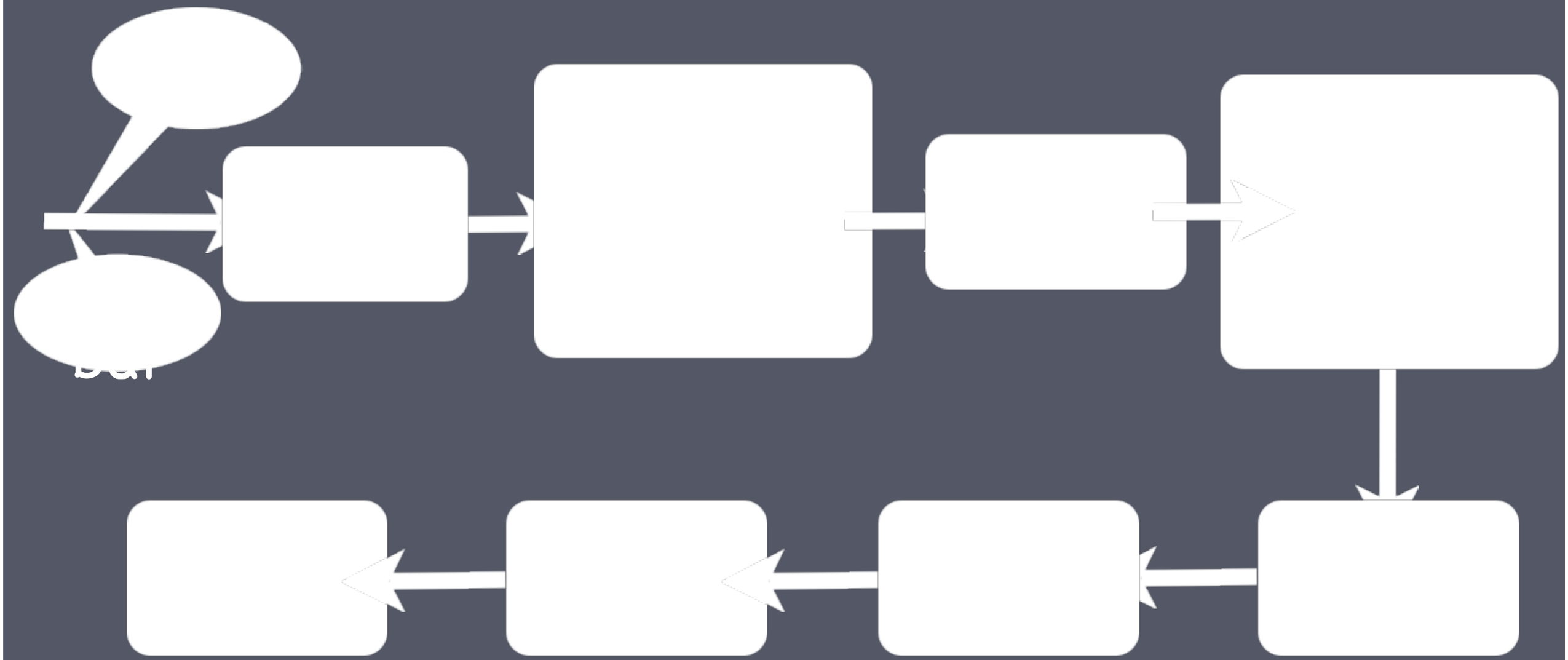
Twitter, how do I stalk?..

- 150 unauthenticated API calls per hour :(
- 350 authenticated API calls per hour :((
- Limits per IP address ((((
- “scalable” through more IP addresses? :/
- “scalable” through more users? :/
- “scalable” through more apps per user? :/
- every step close to hurting yourself through the Terms Of Service :(((((

Anyway, time to resketch my bubbles...



And my concrete bubbles...



wait, Riak, Disco,

Map/Reduce???..

Time to explain the tech,
huh?..

What I didn't explain before:
I picked RabbitMQ. Because
it's fast, reliable, flexible.
And it's written in Erlang.
"Erlang" like in "reliable"

I picked Riak because it stores distributed, redundant and reliable. And it's written in Erlang.

"Erlang" like in "distributed, redundant, reliable"

I picked Disco because it comfortably runs distributed map/reduce jobs written in Python. And its core is written in Erlang.

"Erlang" like in "distributed"

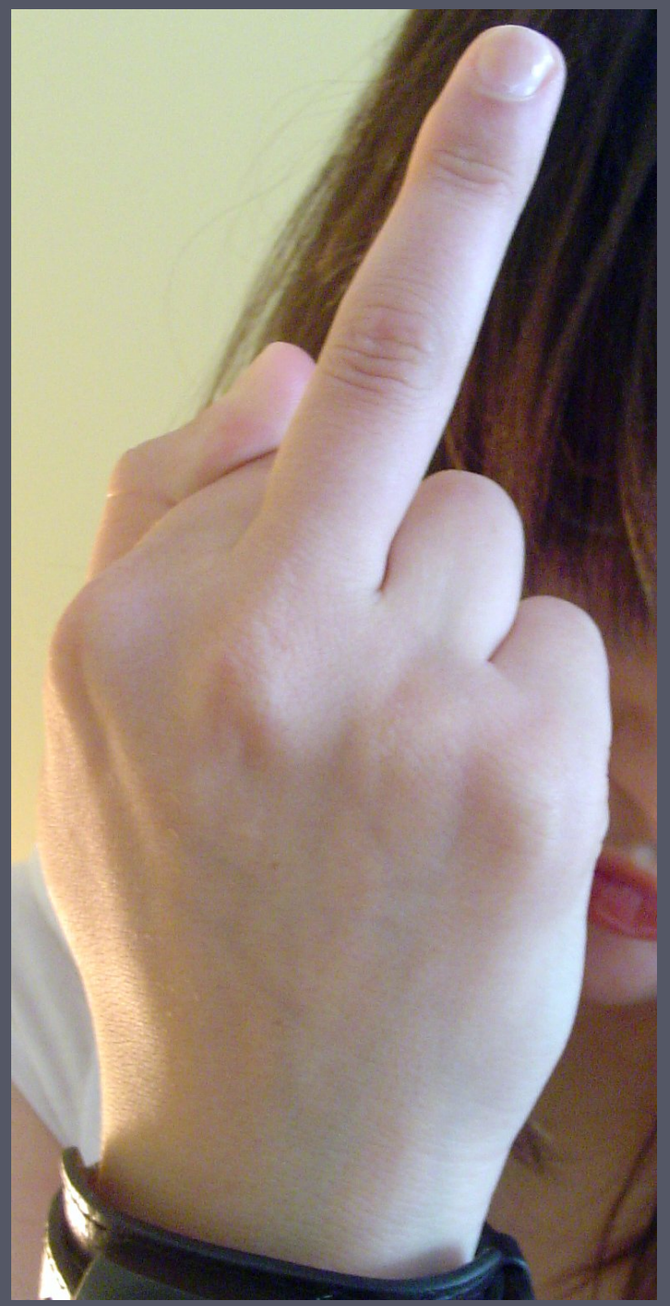
So what I do is to store users in the Riak data store, run data-local map/reduce jobs with Disco on them and ask Twitter for their followers and their recent tweets. Recursively. And very slow through API limits...

And why queueing at all? I want to drink from the sample stream through basic filter only, then store the data without Riak distributed writes eventually slowing down the chain and drink from Riak afterwards...

And the Python stuff? Yes, it is slow(er) at some points. But the whole tool chain balances this out. What I win is a solid platform for analytics...

Sure I could have done this with some other tools, running on the JVM. But remember the strange things coming to my mind?..

So finally I'm at this rant analysis point...



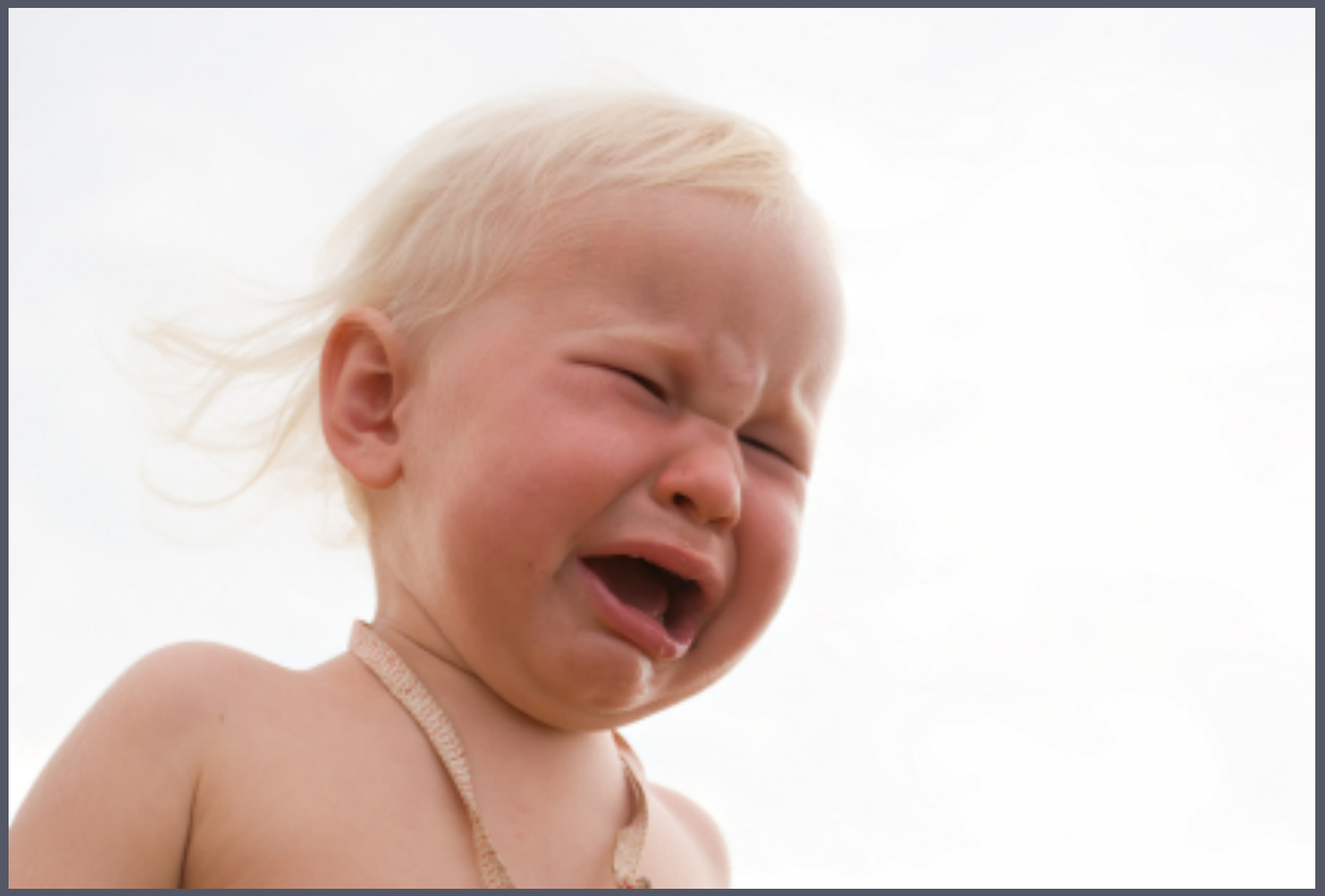
The naive way is to look for swear words etc. But how about this?..

nicolette @nikkyelago 8m

Omg **it**'s cold as **fuck. I love it**

Expand

The right way: sentiment analysis, e.g. through naive Bayes classification



That's home of NLTK being able to tell A from B on text, aka classify. But you need better corpora for rants than what NLTK offers out of the box. Where can I get them?..

Easy - just tag and train using these for

Linus Torvalds goes off on Linux and Git

September 25th, 2012 | **Programming**, **Satire**

I was in a coffee shop in Portland, Oregon and happened to spot Linus Torvalds sitting alone at a window table. I asked the creator of the Linux operating system and the Git source code control system if I could join him. Over the next fifteen minutes we talked about programming and programmers.

Programmers Need To Learn Statistics Or I Will Kill Them All

And in the end, I get my file with rants on some thing or person. And still garbage in there. Like 5 qualified rants per 50'000 users per week. And no colorful charts. Still worth the experiment :)

Learned a lot of useful stuff, became even more allergic against Kool- Aid...

Taught Disco run jobs on prestarted nodes, call Erlang functions and stream back their results to Python, running Disco workers on Riak nodes, asking local vnodes for data locally...

Started implementing Sau - the 100% Python implementation of the Pig Latin processor, so Pig scripts can be ran on Disco workers once I'm done...

Running this whole thing while experimenting on one single W520...

But what do we learn about Big Data here?..

Big Data is...

- Chaos
- Mostly garbage
- Tinkering
- Filtering
- Math, statistics, ML, analytics
- NLP
- Tool selection freedom
- Endless playground for geeks with aspiration

More abstract, Big Data is...

- about what you are trying to find in it
- about finding the best mathematical way to find it
- about filtering out what you don't want to see
- about knowing the limits and hot spots
- about picking the right tool chain

Big Data is 100% data

0-100% Hadoop

0-100% Java

0-100% SQL

100% common sense

100% science

100% analytics

100% experimenting

Thank you!

Most images originate from

istockphoto.com

except few ones taken
from Wikipedia or Flickr (CC)
and product pages
or generated through public
online generators