100% Big Data
0% Hadoop
0% Java

Pavlo Baron, codecentric

entwickler.press

Pragmatische IT-Architektur

Schlanke Technologien und Konzepte für dynamische Unternehmen

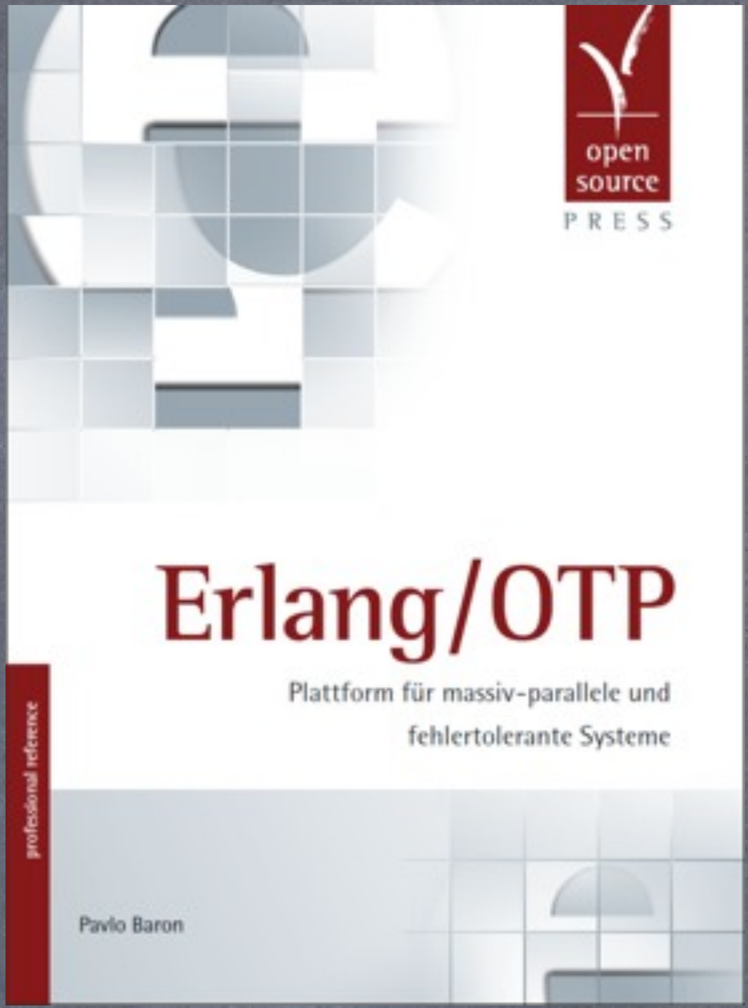
Pavlo Baron

pavlo BARON
michael HÜTTERMANN

FRAGILE AGILE

AGILE SOFTWAREENTWICKLUNG RICHTIG VERSTEHEN UND LEBEN

HANSER

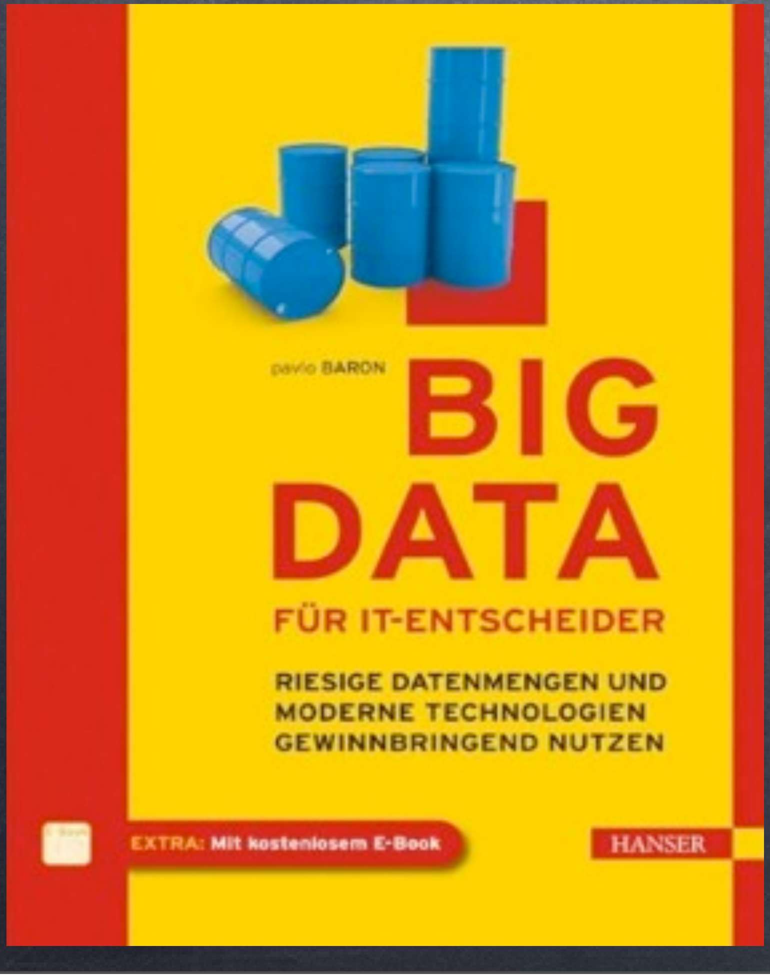

open source PRESS

Erlang/OTP

Plattform für massiv-parallele und fehlertolerante Systeme

professional reference

Pavlo Baron


pavlo BARON

BIG DATA

FÜR IT-ENTSCHEIDER

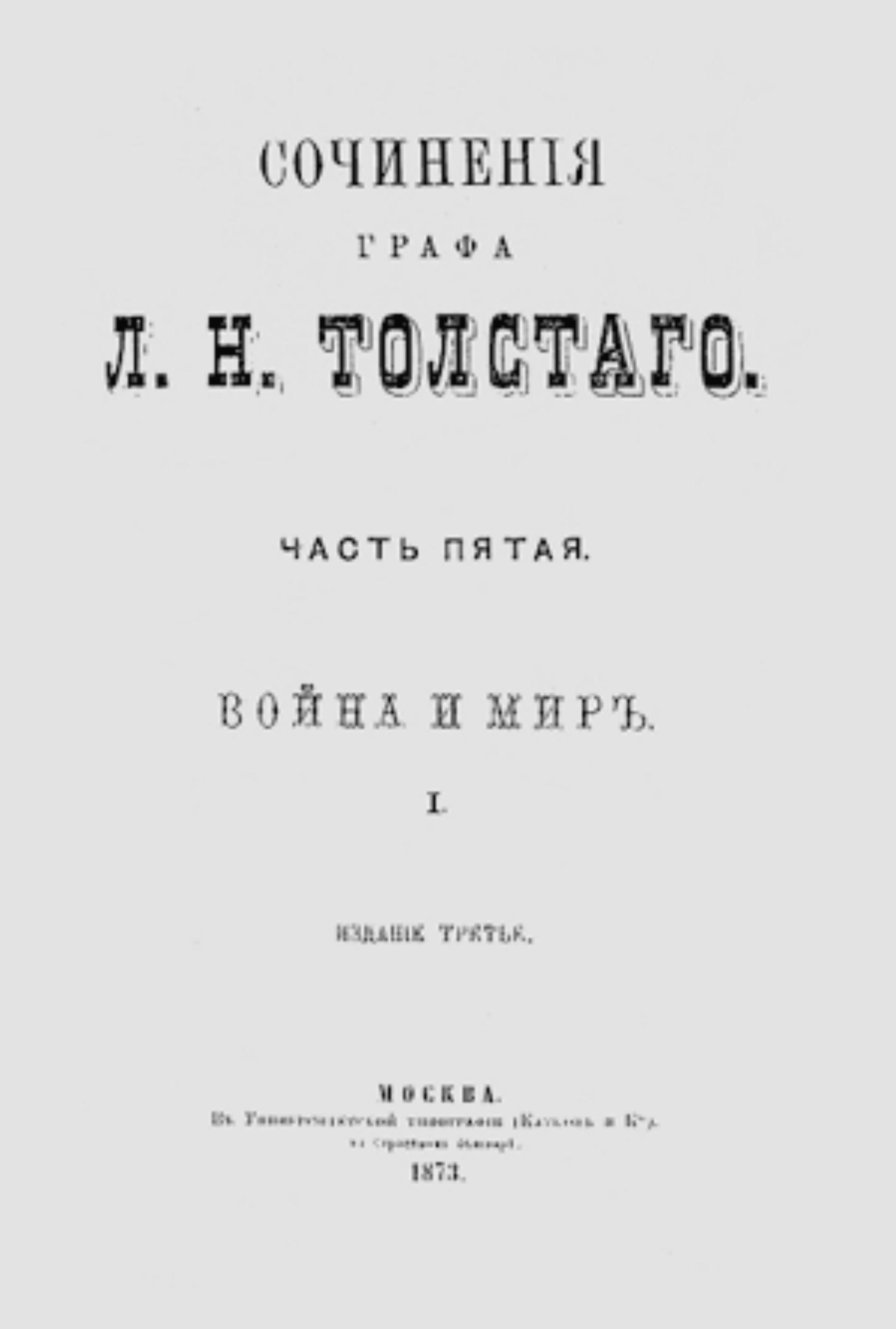
RIESIGE DATENMENGEN UND MODERNE TECHNOLOGIEN GEWINNBRINGEND NUTZEN

EXTRA: Mit kostenlosem E-Book

HANSER

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

So here is the short story…



СОЧИНЕНІЯ
ГРАФА
Л. Н. ТОЛСТАГО.

ЧАСТЬ ПЯТАЯ.

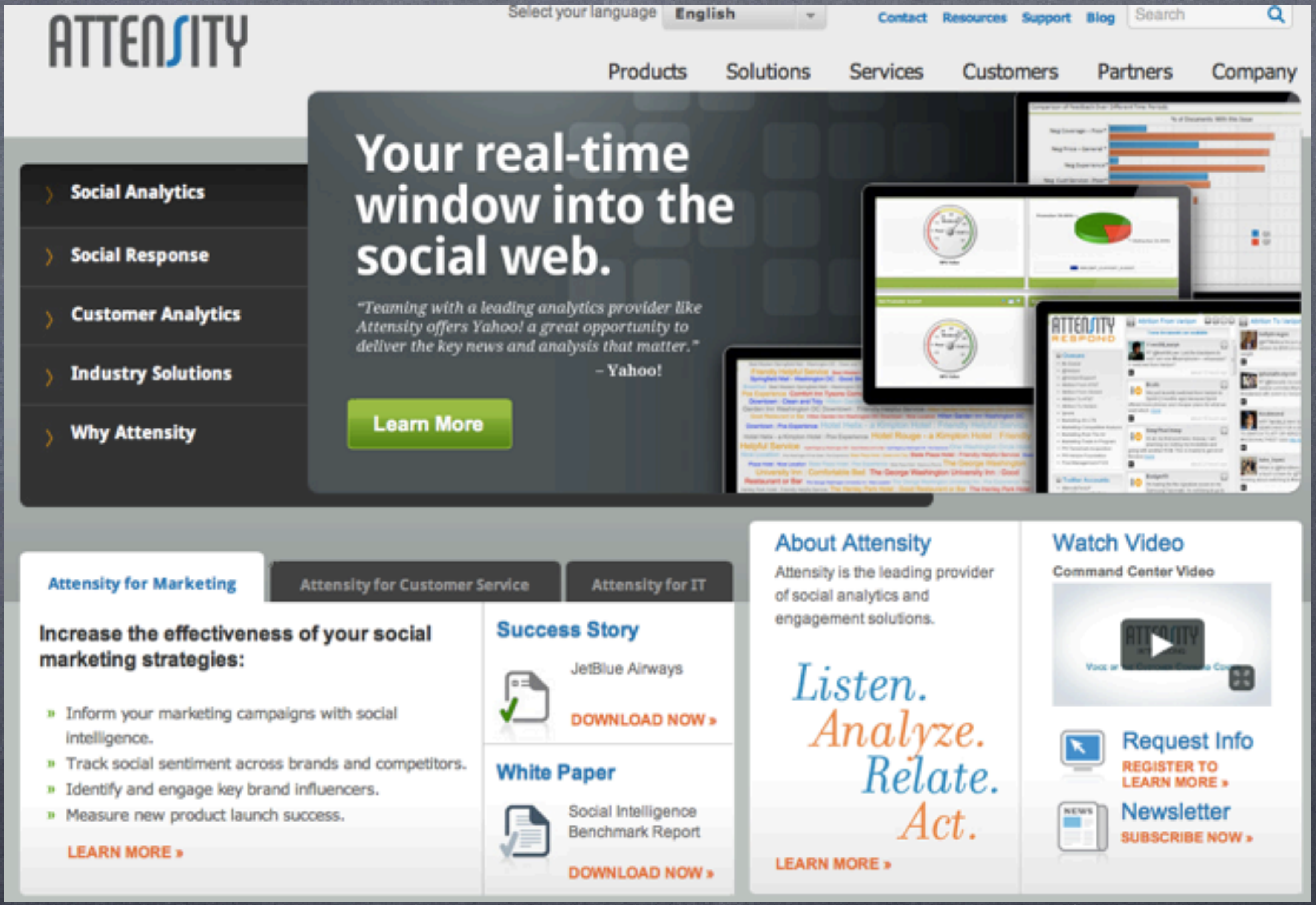
ВОЙНА И МИРЪ.

I.

ИЗДАНІЕ ТРЕТЬЕ.

МОСКВА.
Въ Университетской типографіи (Катковъ и Ко),
на Страстномъ бульварѣ.
1873.

sitting there, listening...



ATTENSITY

Select your language English

Contact Resources Support Blog Search

Products Solutions Services Customers Partners Company

- Social Analytics
- Social Response
- Customer Analytics
- Industry Solutions
- Why Attensity

Your real-time window into the social web.

"Teaming with a leading analytics provider like Attensity offers Yahoo! a great opportunity to deliver the key news and analysis that matter."
– Yahoo!

Learn More

Attensity for Marketing | Attensity for Customer Service | Attensity for IT

Increase the effectiveness of your social marketing strategies:

- Inform your marketing campaigns with social intelligence.
- Track social sentiment across brands and competitors.
- Identify and engage key brand influencers.
- Measure new product launch success.

LEARN MORE »

Success Story

JetBlue Airways

DOWNLOAD NOW »

White Paper

Social Intelligence Benchmark Report

DOWNLOAD NOW »

About Attensity

Attensity is the leading provider of social analytics and engagement solutions.

Listen. Analyze. Relate. Act.

LEARN MORE »

Watch Video

Command Center Video

Request Info

REGISTER TO LEARN MORE »

Newsletter

SUBSCRIBE NOW »

presented as Houdini magic...

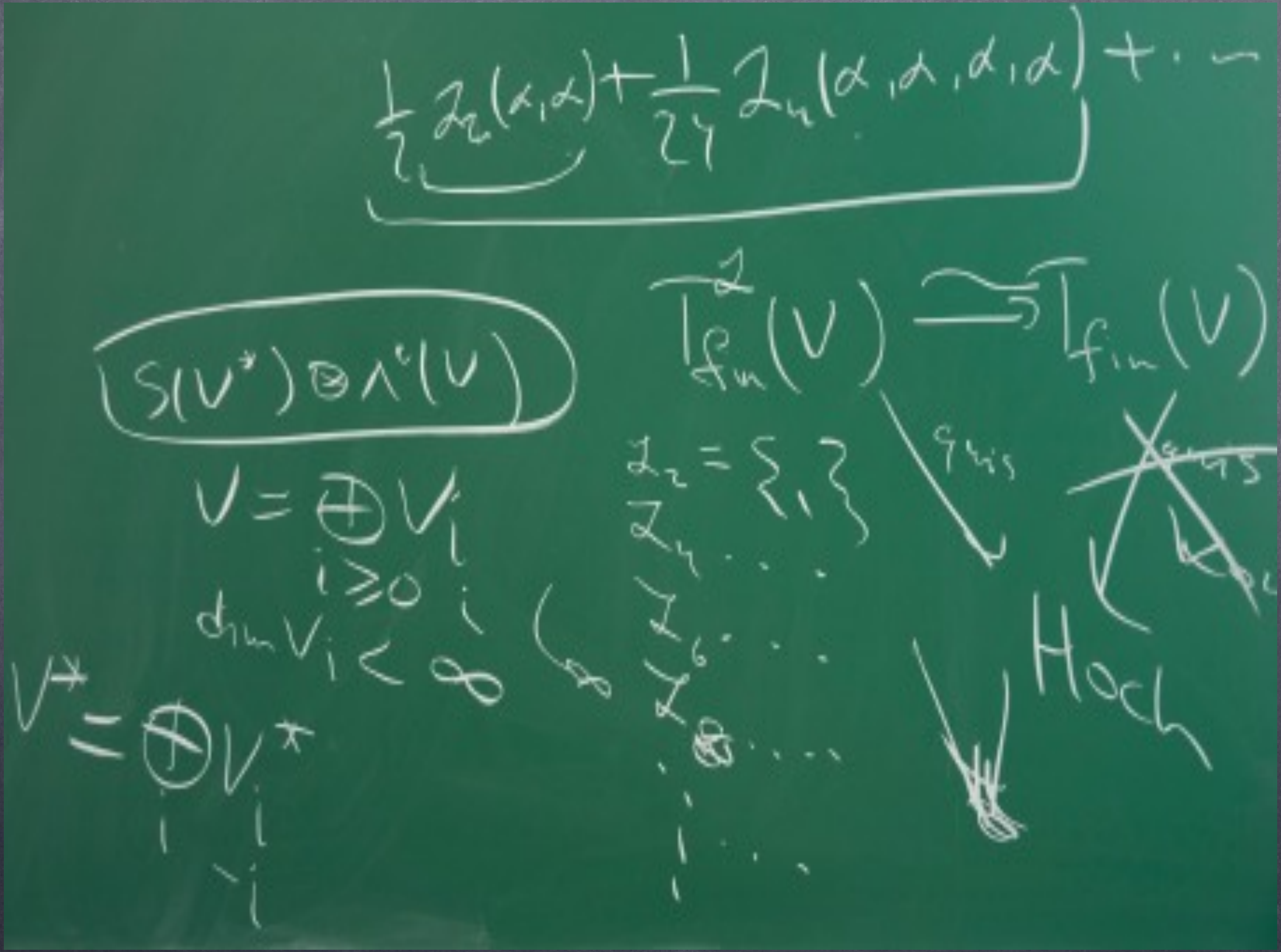
so you telling me...

it's smoke and mirrors?

Smells like a bunch of queues, pipes and filters…

Looks like some NLP…

Sounds like some math...

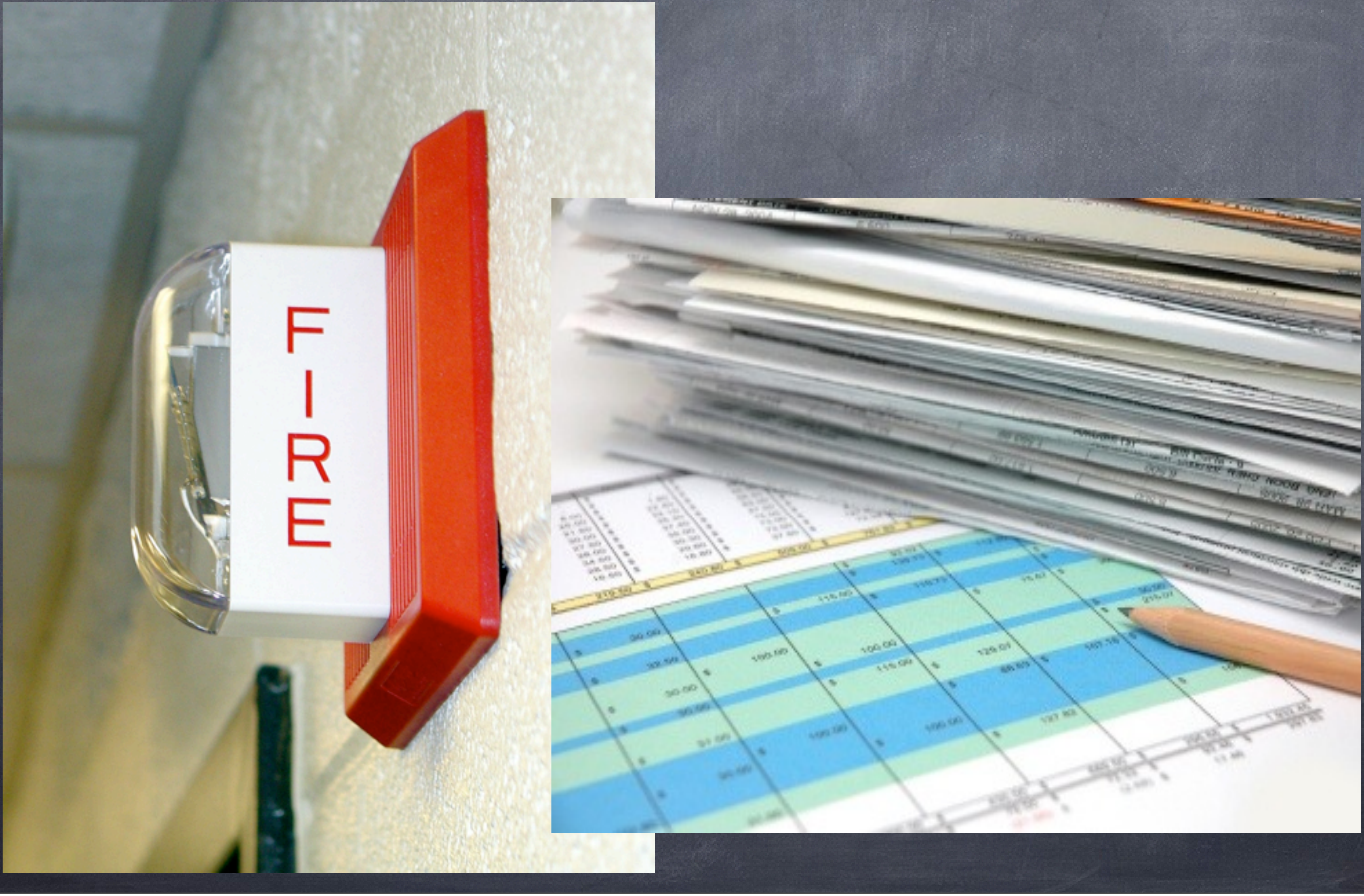

$\frac{1}{2}\mathcal{L}_2(\alpha,\alpha)+\frac{1}{24}\mathcal{L}_4(\alpha,\alpha,\alpha,\alpha)+\cdots$
$T_{fin}(V)$
$T_{fin}(V)$
$S(V^*)\otimes\Lambda^*(V)$
$V=\bigoplus_{i\geq 0} V_i$
$\dim V_i<\infty$
$V^*=\bigoplus_i V_i^*$
$\mathcal{L}_2=\{,\}$
$\mathcal{L}_4\ldots$
$\mathcal{L}_6\ldots$
$\mathcal{L}_8\ldots$
Hoch

Seems like basic ML…

methinks: I can tinker that. I have 2 nights in the hotel...

Fire!

Know the use cases...

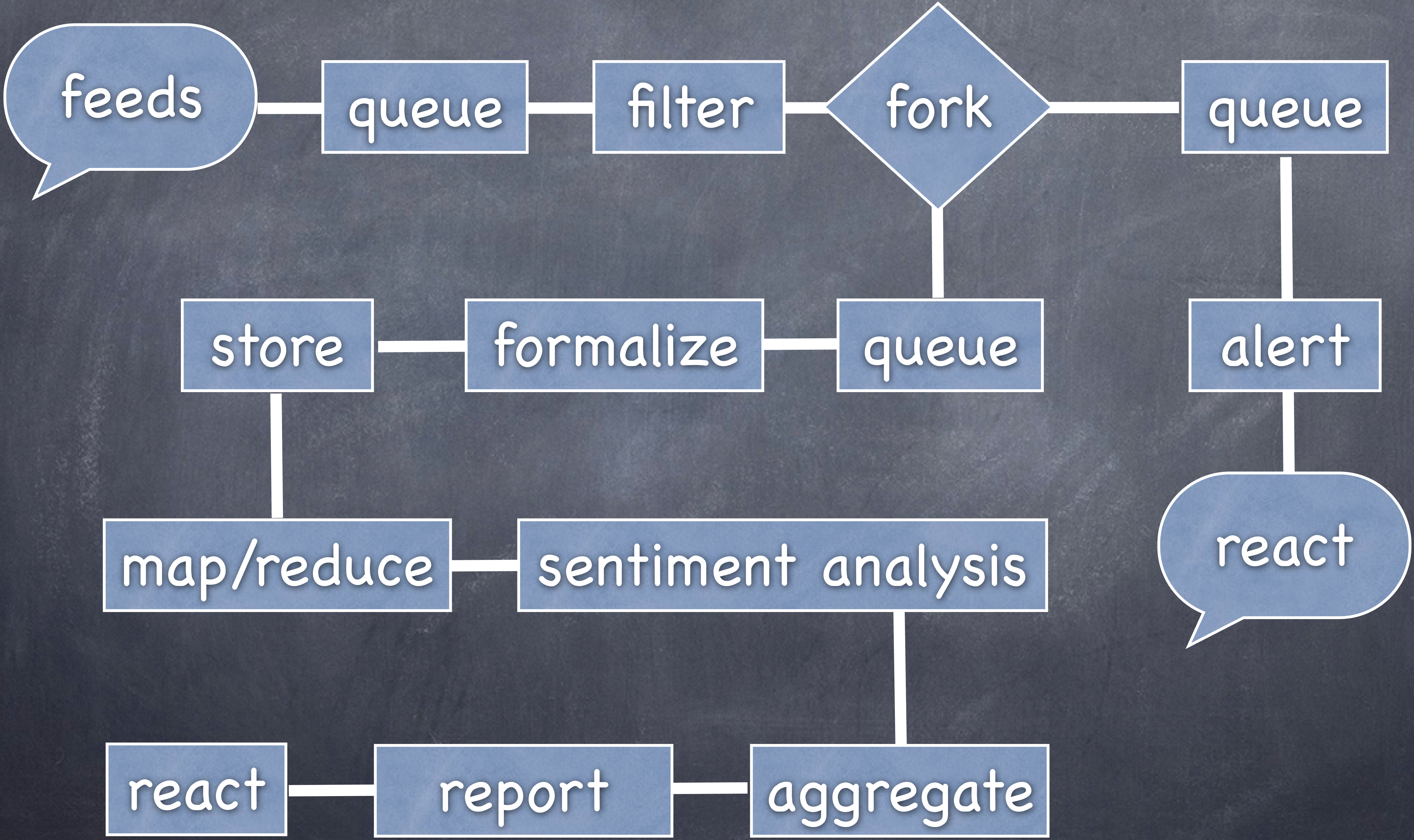


Consume a feed where people say what they think before they think what they say…

Drink Big Data warm, straight from the fire hose...

Then fork for immediate notification and batch analytics...

Some bubbles



Some tech

- Languages: Python, Erlang
- Feeds: Tweepy, crawlers, feed readers
- Queueing: RabbitMQ through Pika
- Store: Riak through protobufs
- Map/reduce: modified Disco to run workers on Riak-nodes data-locally

Some math

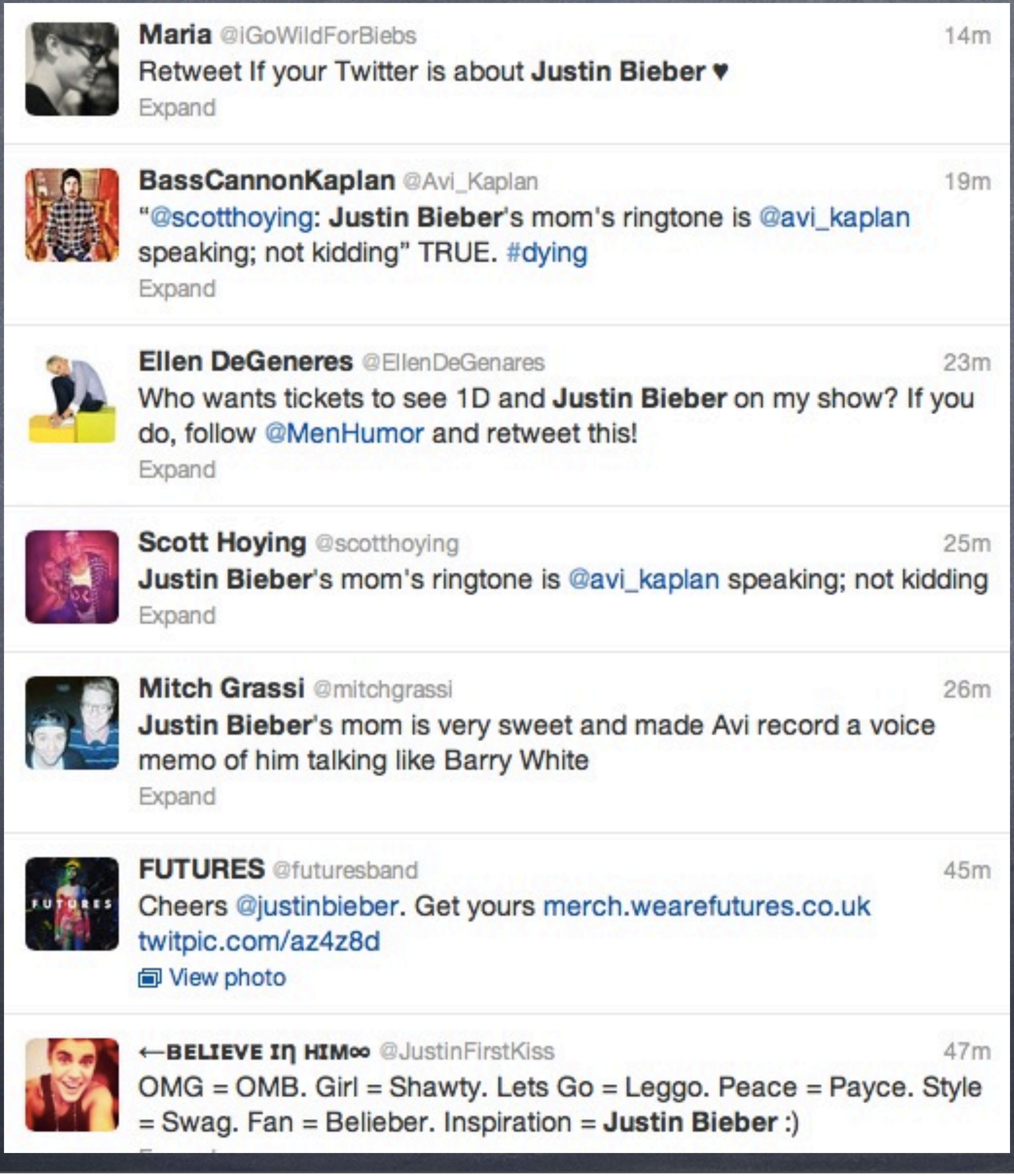
- Analytics: NLP with NLTK
- Algo training: nltk-trainer with pickle=true
- Algos: naive Bayes, decision tree, binary classification based on trigram frequencies
- simple name and antiword filtering based on public and own corpora

Some numbers (on MBA)

- Feed: ~10000 chaotic msg/min
- Store: ~8000 formalized msg/min, N=3, quorum, 3 nodes
- Analytics: ~2000 msg/min (filtered, pos/neg aggregation, location based aggregation)
- Demo: ~1500000 tweets, map/reduce on a handful of tweets for simplicity, pos/neg aggregation

Some lessons learned

The Beliebers…



Maria @iGoWildForBiebs 14m
Retweet If your Twitter is about **Justin Bieber** ♥
Expand

BassCannonKaplan @Avi_Kaplan 19m
"@scotthoying: **Justin Bieber**'s mom's ringtone is @avi_kaplan speaking; not kidding" TRUE. #dying
Expand

Ellen DeGeneres @EllenDeGenares 23m
Who wants tickets to see 1D and **Justin Bieber** on my show? If you do, follow @MenHumor and retweet this!
Expand

Scott Hoying @scotthoying 25m
Justin Bieber's mom's ringtone is @avi_kaplan speaking; not kidding
Expand

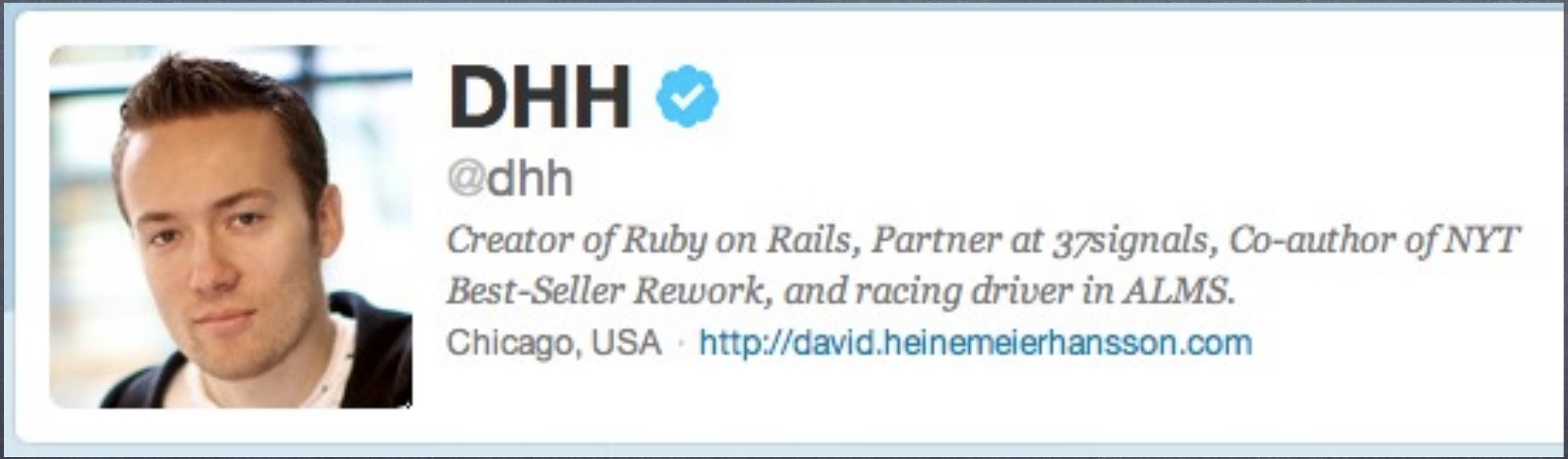
Mitch Grassi @mitchgrassi 26m
Justin Bieber's mom is very sweet and made Avi record a voice memo of him talking like Barry White
Expand

FUTURES @futuresband 45m
Cheers @justinbieber. Get yours merch.wearefutures.co.uk twitpic.com/az4z8d
View photo

←BELIEVE Iŋ HIM∞ @JustinFirstKiss 47m
OMG = OMB. Girl = Shawty. Lets Go = Leggo. Peace = Payce. Style = Swag. Fan = Belieber. Inspiration = **Justin Bieber** :)

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

Real names…



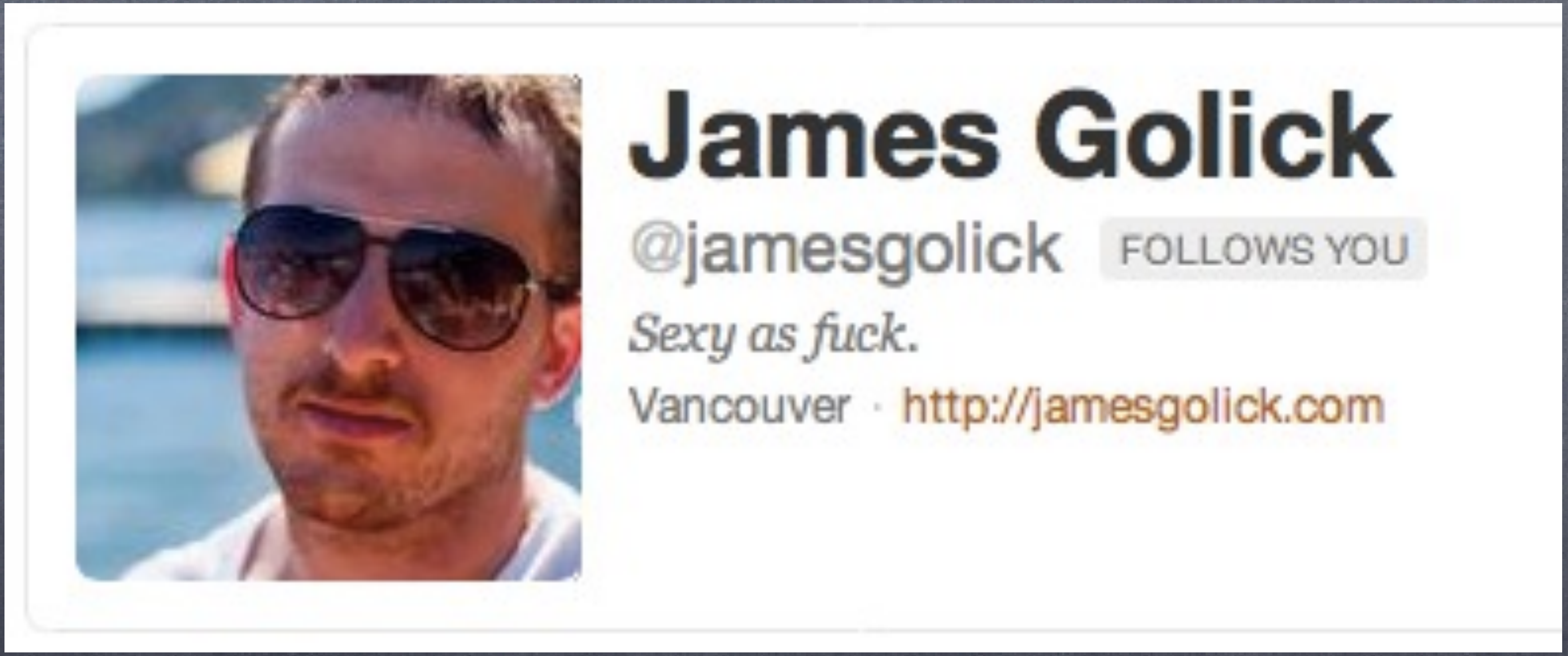
DHH

@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

Absurd profile bios...



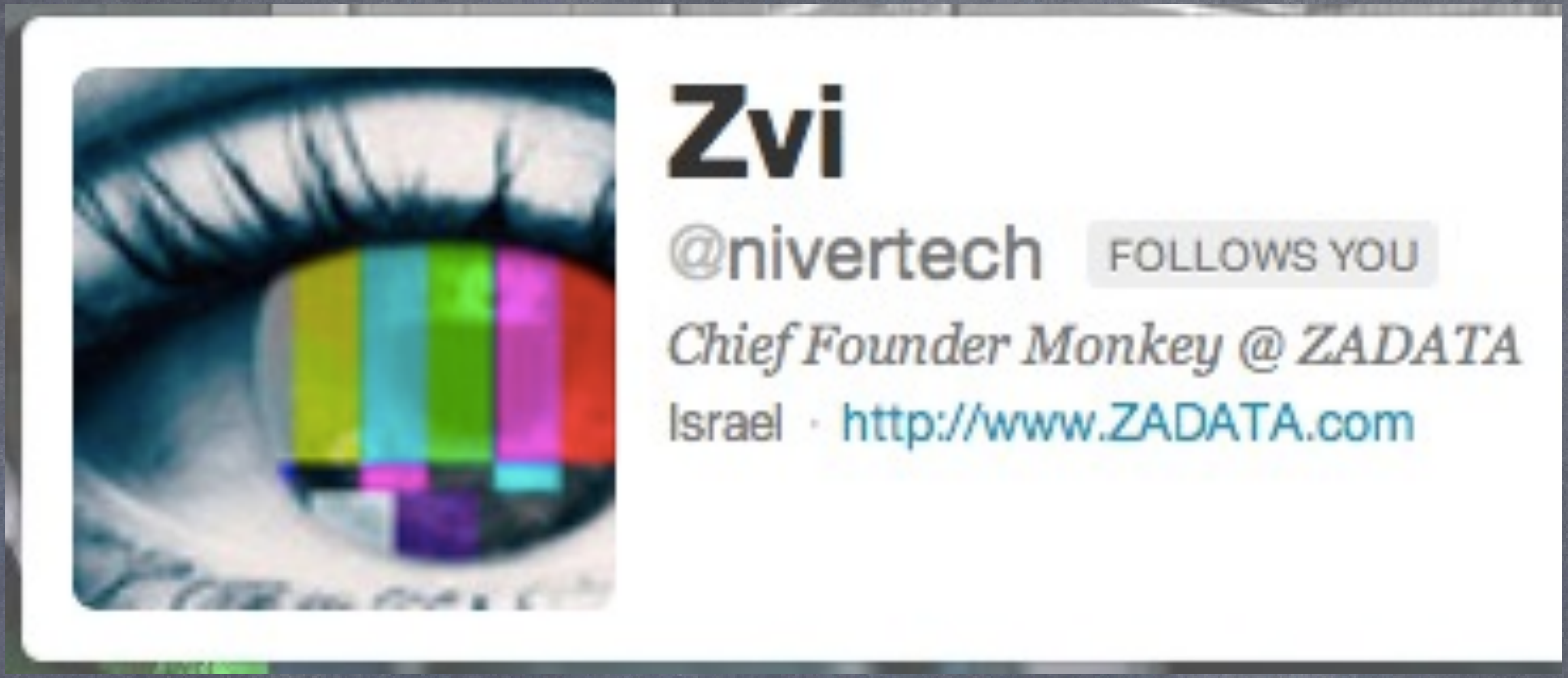
James Golick

@jamesgolick FOLLOWS YOU

Sexy as fuck.

Vancouver · http://jamesgolick.com

Location...

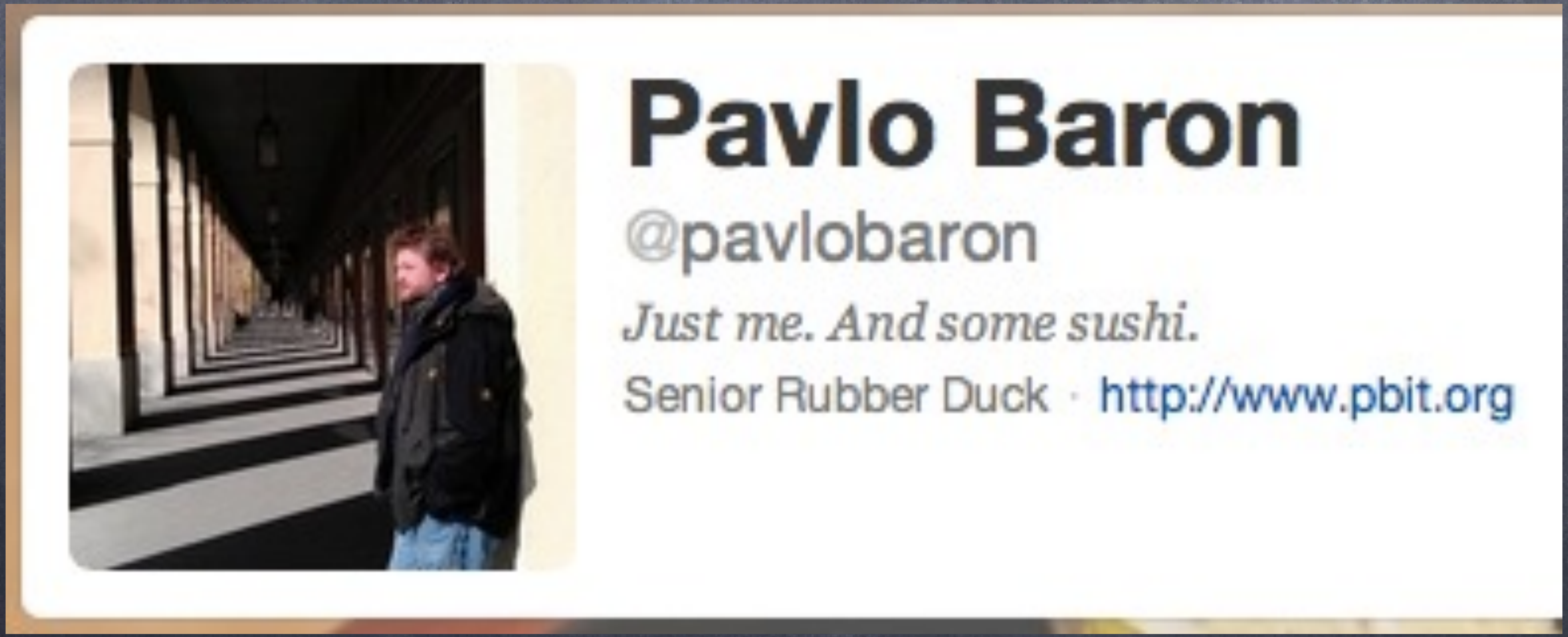


Zvi

@nivertech FOLLOWS YOU

Chief Founder Monkey @ ZADATA

Israel · http://www.ZADATA.com



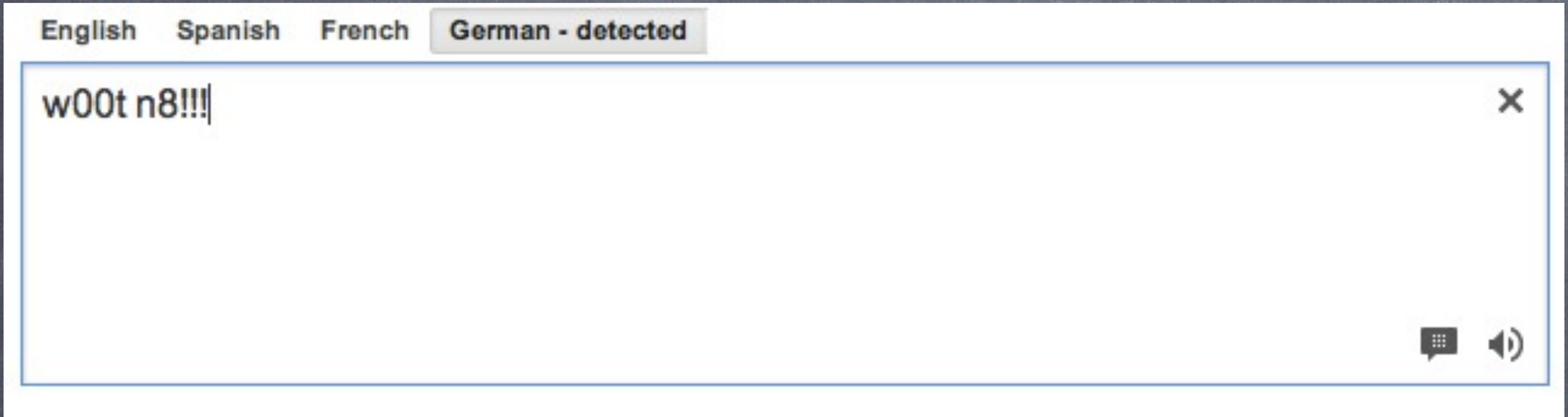
Pavlo Baron

@pavlobaron

Just me. And some sushi.

Senior Rubber Duck · http://www.pbit.org

Language…



w00t n8!!!

For trigrams in NLTK, use Spanish as “anti-class” to tell English/German from the rest

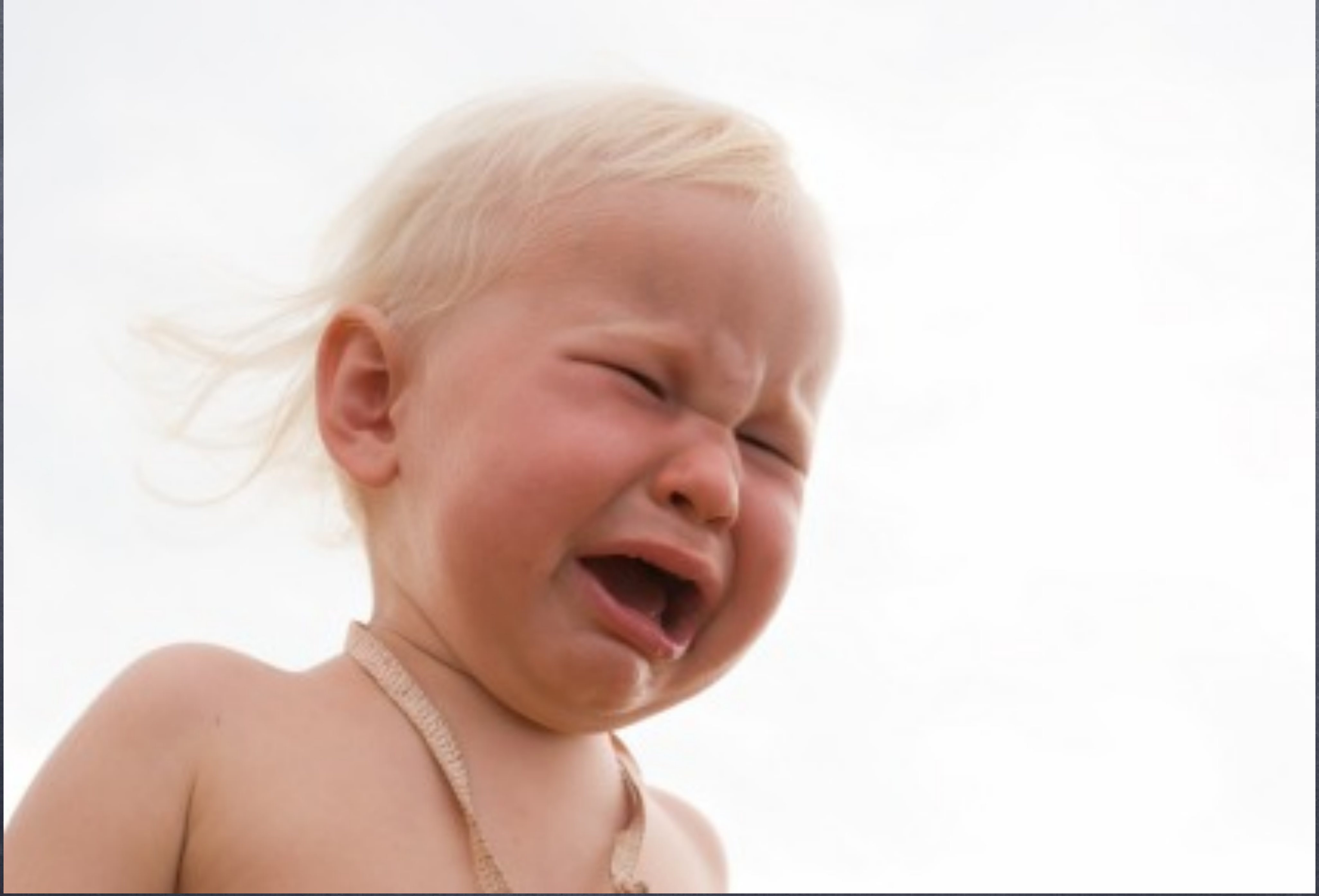
Disco workers on Riak nodes...

- PITA and a lot of tinkering, but necessary for data locality
- Extending Disco is hard...
- Flooding, asynchronous, separate key/value listing in low-level Riak goes very well with Erlang port based Python/Erlang message exchange in Disco. Not
- Evaluating to redo Disco to use RabbitMQ or even ZeroMQ between the worlds (h/t Dan North)

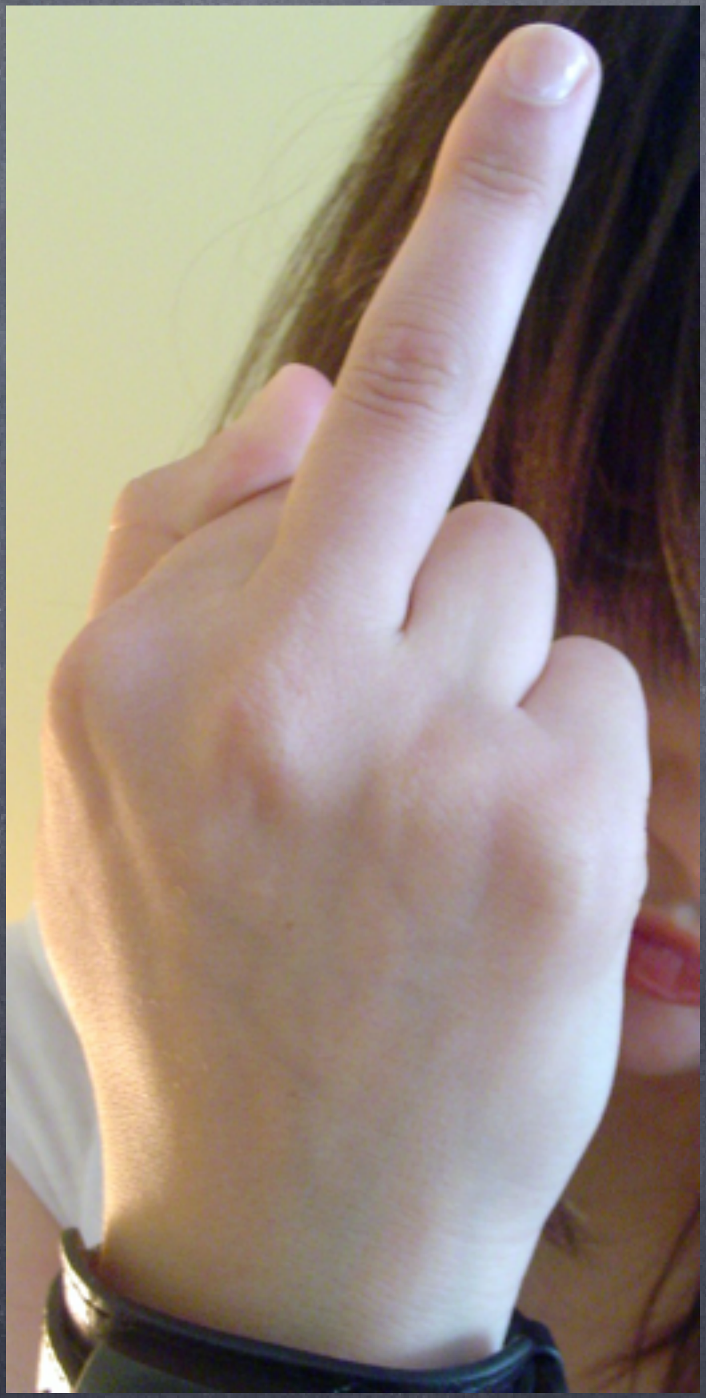
Mixing Python and Erlang in one project…

- Forgetting punctuation in Erlang code all the time when quickly switching from Python
- Terribly missing pattern matching in Python
- Considering to embed Python in Erlang, but it might become a double PITA then

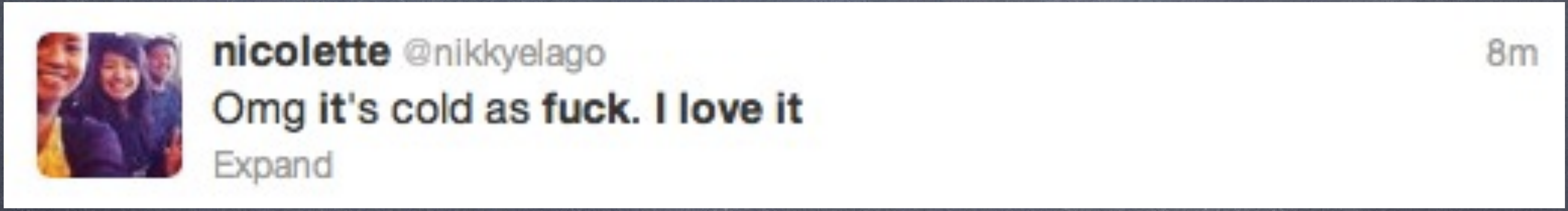
Sentiment analysis...



Well, actually, strong sentiment analysis...



Very unreliable given the human nature…

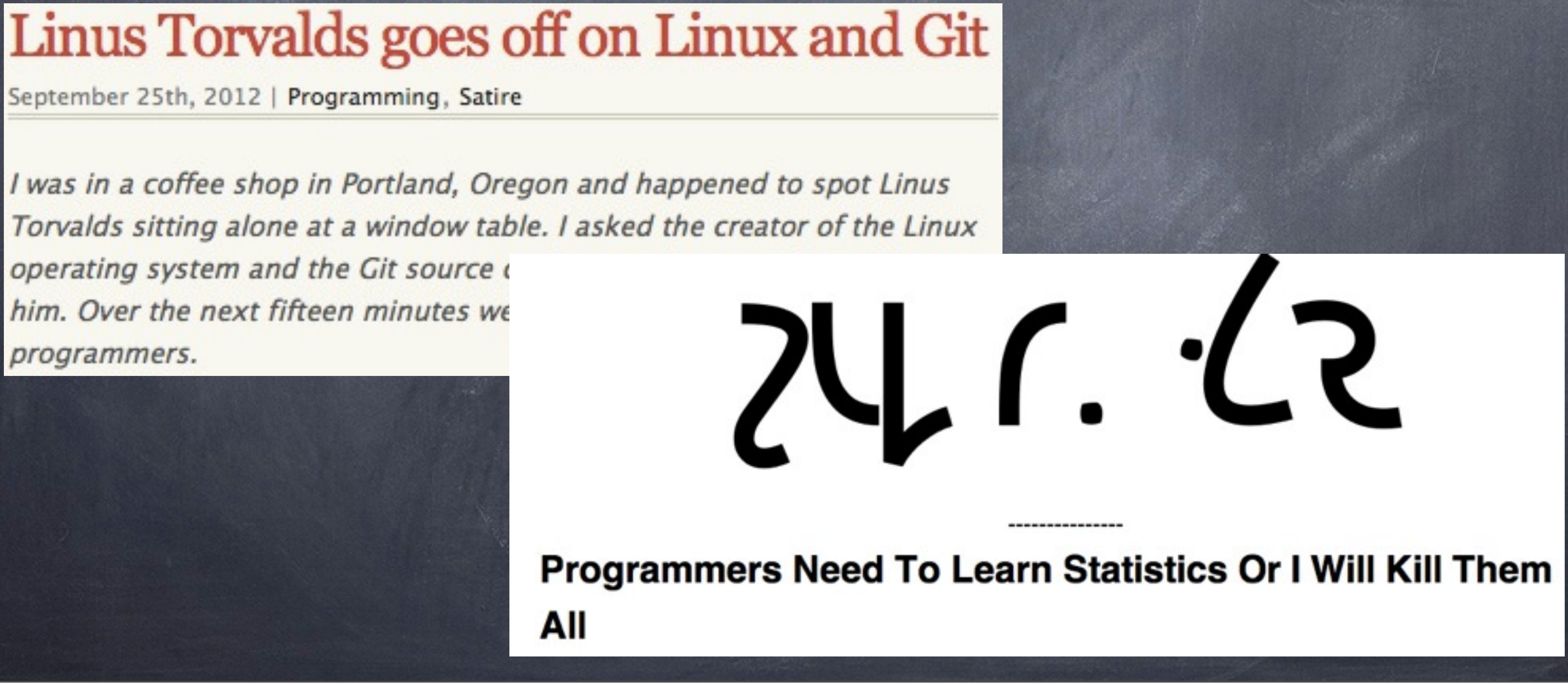


nicolette @nikkyelago

8m

Omg it's cold as fuck. I love it

In addition to the NLTK's movie reviews corpus, use these for "neg" classification



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

him. Over the next fifteen minutes we

programmers.

Programmers Need To Learn Statistics Or I Will Kill Them All

FAQ

Q: Why the heck are you doing this?

A

- Because I can
- Because I want
- Because I want to learn
- Because I want to go deep on low-level
- Because it's very interesting to combine computer science with math

Q: Why not just use Hadoop?

A

- Because I didn't want to run this on the JVM
- Because I have 2 use cases, and only one of them is suitable for batch map/reduce

Q: Why didn't you want to run this on the JVM?

A: well, technically seen, Big Data area is growing on the JVM

- Hadoop
- Pig
- Storm, Kafka, Esper
- Mahout
- OpenNLP

A: but I didn't want this Big Data on my drive

~/.m2

A: and I am evaluating some alternatives to the ecosystem

Q: Why are you queueing at all? Others do gazillions of msg/sec without queues

A

- I could, if instead of filters and batch analytics of chaotic text, it would be just about building trivial sums
- with growable numbers like this, you want to protect any sort of reliable data store from getting flooded by writes, RDBMS or NoSQL store
- Because I need to do some pipes and filters
- Because I'm mixing and crossing borders of data sources and technologies
- Because almost all frameworks that you might consider also do some queueing or buffering

Q: Why did you use Erlang and Python?

A

- Because reliability and distribution are built into the Erlang VM and I don’t need separate coordinators or to reinvent the wheel
- Because both, Python and Erlang, are “functional” enough for what I need day-by-day
- Because Python has been for many years the platform of choice for scientists, thus there are available clever and mature math libraries
- Because Disco is on Python and Erlang, Riak and RabbitMQ are on Erlang

Q: isn't Python slow like hell?

A

- it's not operating at the speed of light
- yes, it is slower at some points
- I'm also testing PyPy to improve performance for the case I should need it, 'cause right now it works just fast enough without explicit bottle-necks in the given architecture, even on one single MBA

Q: MBA is boring. Can you make it real web scale?

A

- well, to be precise, I'm operating on web data
- I can scale queues with RabbitMQ
- I can scale storage with Riak
- I can scale the map/reduce supported analytics with Disco/Riak
- I can scale data sources/feeds, machines, hardware, networks, infrastructure, logins etc. You name it

Q: what's in the future?

A

- I don't have my crystal ball with me
- I've started to implement Pig Latin engine in Python called "Sau" (German for pig), to offer data scientists a comfortable interface and to allow them to run existing Pig scripts on this stack
- I'm going to add more data sources, improve throughput where necessary and work on some low level Disco modifications to change the way it utilizes Erlang in my case

Q: what do we learn about Big Data here?

A

- Big Data is about the “what”, followed by the “how” and enabled by the “what with”

A

- It’s about gathering data, analyzing it, gaining useful information out of it, finding new ways to gather and use information and deriving steps for business improvements, strategy planning, doing soft intelligence aka enterprise level stalking or, even more important, helping make the world a better place - it’s up to you

A

- It's not about building SkyNet - even if this will be built one day, it will be pretty boring. It's about building recommender and decision support systems, thus letting machines do stupid, repeated jobs fast and human beings make high quality decisions

A

- It’s a huge field for geeks with aspiration to learn new things, dig into math and computer science, play with different platforms and tools and pick the right tool chain

Oh, and did the demo run?

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