Apache Mahout

An Extendable Machine Learning Framework
for Flink and Spark (and others)

MAHOUT

About Me

Trevor Grant

PMC Apache Mahout
& Apache Streams

Open Source Evangelist, IBM

MS Applied Math / MBA

@rawkintrevo

rawkintrevo@apache.org

http://rawkintrevo.org

MAHOUT

About Mahout

History of Mahout 2008 - 2014

Lucene Subproject

TLP - May 2010

Feb 2014- v 0.9 (Last Map-Reduce)
  - Lots of Hadoop Vendors froze here

Popular as ML on MR.

The rise of SparkML / MLLib

MAHOUT

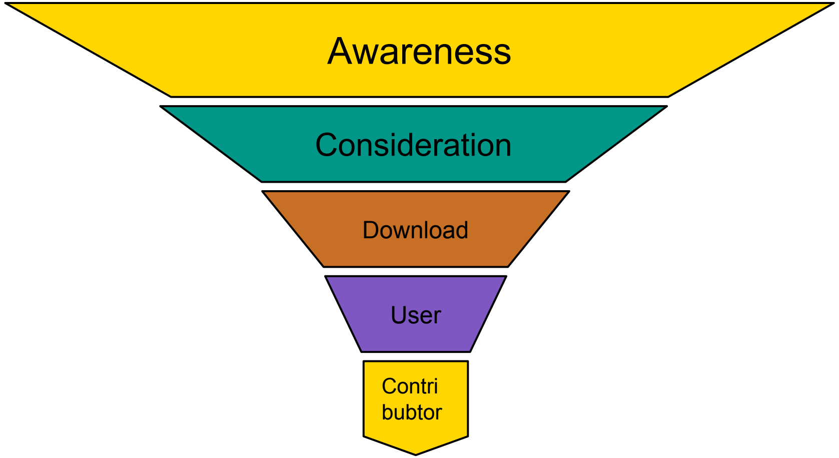
Rebranding

Marketing For Everyone

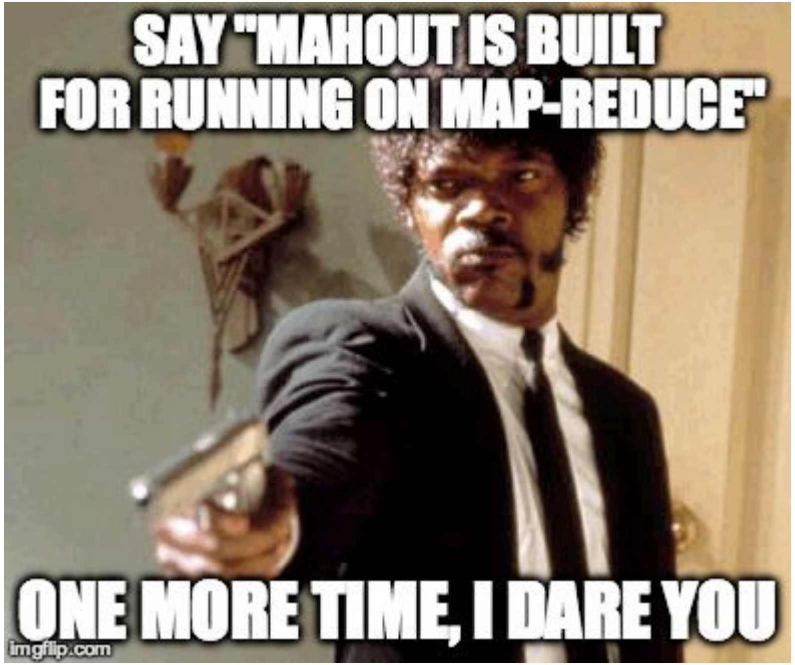
- You started doing open source so you could code whatever you want
- But you want people to use your open source product
- But your open source product doesn’t have a marketing team
- And marketing teams aren’t doing open source.
- …
- So you’re the marketing team
- …
- Or you’re writing a product no one is going to use. /shrug.
- **Exception:** Your OSS product has a company, and the company pays the marketing team (and you probably have too many contributors).

MAHOUT

“Marketing Funnel”



Mahout Evangelism: Rebranding


SAY "MAHOUT IS BUILT FOR RUNNING ON MAP-REDUCE"
ONE MORE TIME, I DARE YOU
imgflip.com

MAHOUT

Evangelism: Rebranding

Problem: High brand-recognition… for something we don’t do anymore

Opportunity: Now relevant to many more areas- GPUs, Spark, Flink, etc.

Solution: Talks and Blog Posts

Corollary: New awareness, and changing perceptions

Mahout Evangelism: Talks (Spring ‘17)

APACHE MAHOUT'S NEW RECOMMENDER ALGORITHM AND USING GPUS TO SPEED MODEL CREATION *Pat Ferrel, Andy Palumbo*. GPU Technology Conference. Silicon Valley, CA- May 11, 2017

EXTENDING MAHOUT-SAMSARA LINEAR ALGEBRA DSL TO SUPPORT GPU CLUSTERS *Suneel Marthi, Trevor Grant*. GPU Technology Conference. Silicon Valley, CA- May 11, 2017

Apache Mahout: An Extendable Machine Learning Framework for Spark and Flink *Trevor Grant*. Apache Big Data. Miami, FL- May 16, 2017

An Apache Based Intelligent IoT Stack for Transportation *Trevor Grant, Joe Olsen*. ApacheCon IoT. Miami, FL- May 18, 2017

(+2 at ApacheCon/Apache Big Data but last minute speaker had conflict)

Apache Mahout: Distributed Matrix Math for Machine Learning *Andrew Musselman*. MLConf. Seattle, WA- May 19, 2017

Weekend Project: Real World AirBnB Data Science and Pricing Bot *Trevor Grant, Andrew Weiner*. Berlin Buzzwords. Berlin, DE- June 13, 2017

Introduction to Online Machine Learning Algorithms *Trevor Grant*. Dataworks Summit. San Jose, CA- June 15, 2017

MAHOUT

Evangelism: Talks (Fall ‘17/ Spring ‘18)

Matrix Math at Scale With Apache Mahout and Apache Spark. *Andrew Musselman*. ODSC. Boston, MA. May 1st, 2018.

The Magnificent Modular Mahout- An extensible library for distributed math and HPC *Trevor Grant*. FOSDEM- High Performance Computing. Brussels, BE- February 4th, 2018.

Do I Know You? Realtime Facial Recognition with an Apache Stack *Trevor Grant*. Flink Forward, Berlin. Berlin, DE- September 13, 2017

MAHOUT

Mahout Evangelism: Blog Posts

Multi-domain predictive AI - June 26, 2018 *Pat Ferrel PMC*
https://developer.ibm.com/dwblog/2017/mahout-spark-correlated-cross-occurences/

Introducing Precanned Algorithms in Apache Mahout - May 2, 2017. *Trevor Grant*
https://rawkintrevo.org/2017/05/02/introducing-pre-canned-algorithms-apache-mahout/

Getting Started with Apache Mahout - April 25, 2017. *Trevor Grant*
https://datascience.ibm.com/blog/getting-started-with-apache-mahout-2/

Correlated Cross-Occurrence (CCO): How to make data behave as we want - December 1, 2016. *Pat Ferrel*
http://actionml.com/blog/cco

Starting to Gain Traction: Non PMC Blog Posts

Top Skills Data Scientists Need to Learn in 2018 - Feb 3, 2018.
https://insidebigdata.com/2018/02/03/top-skills-data-scientists-need-learn-2018/

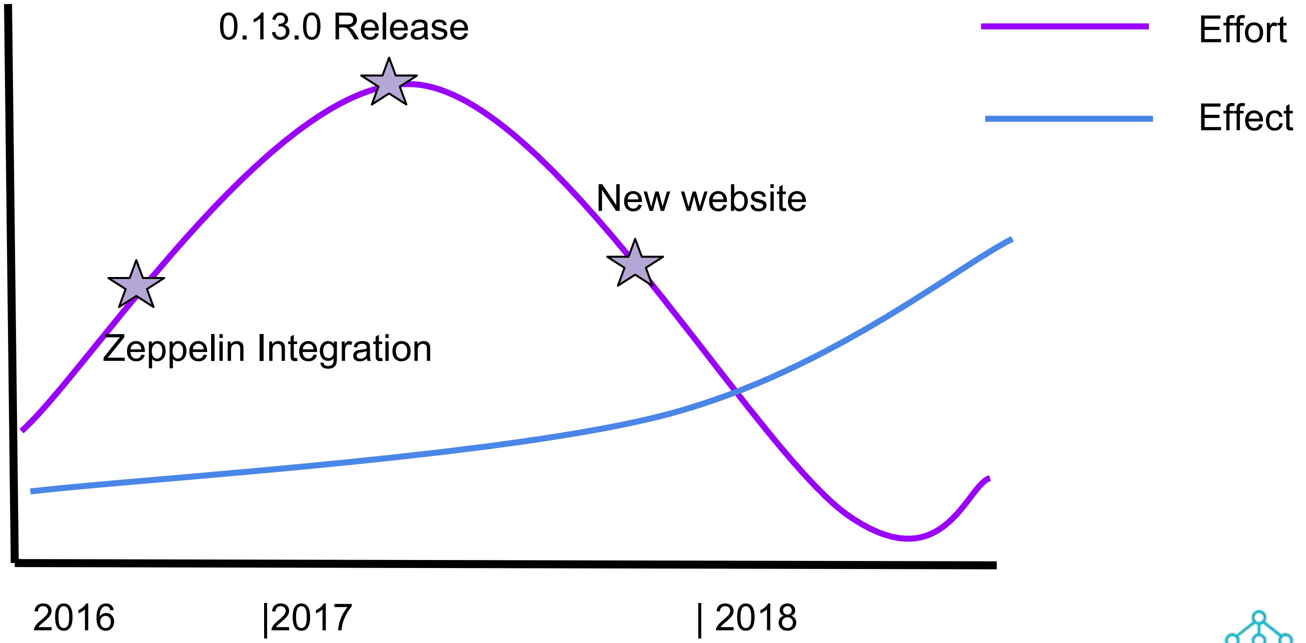
10 Best Machine Learning Software 2018 - Feb 24, 2018. (Map-Reduce)
https://data-flair.training/blogs/machine-learning-software/

Encyclopedia of Big Data Technologies - Jan 29, 2018
https://link.springer.com/referenceworkentry/10.1007/978-3-319-63962-8_144-1

Machine Learning, predictive AI and IoT… Oh My! - May 23, 2017 *Lisa Seacat*
https://developer.ibm.com/dwblog/2017/mahout-spark-correlated-cross-occurences/

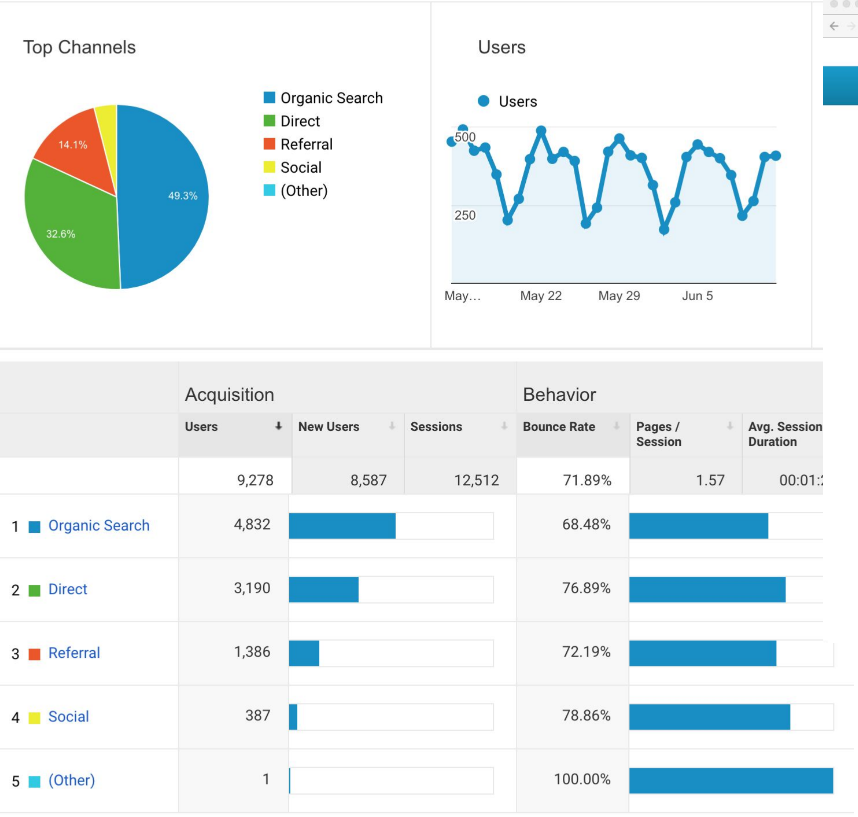
Forgot About Mahout- It's Back and Worth Your Attention. May 18, 2017. *Andrew C. Oliver*
https://www.infoworld.com/article/3197429/machine-learning/forgot-about-mahout-its-back-and-worth-your-attention.html

Non-Science/Data Based Graph

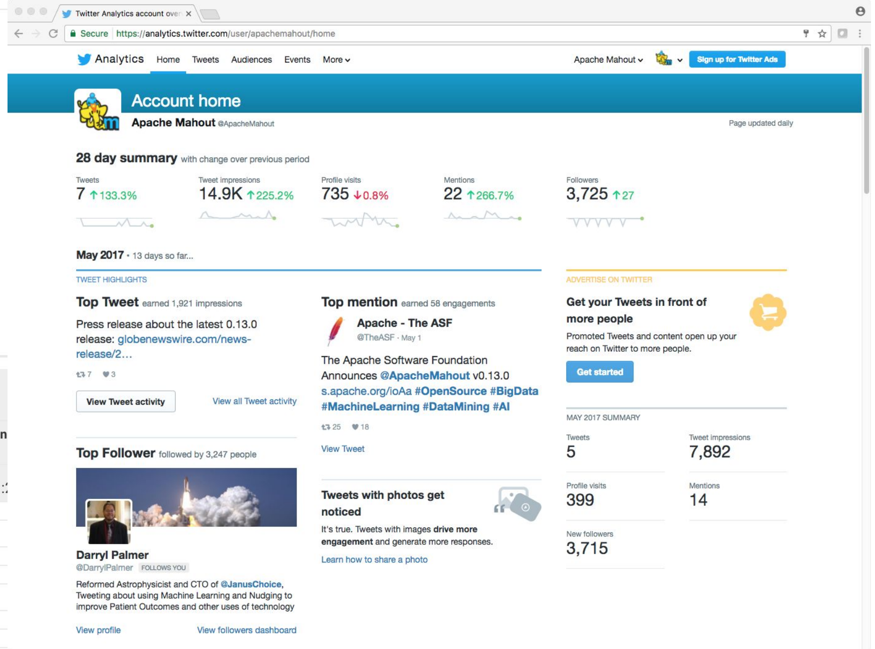


MAHOUT

Monitoring Progress: Analytics



	Acquisition			Behavior		
	Users	New Users	Sessions	Bounce Rate	Pages / Session	Avg. Session Duration
	9,278	8,587	12,512	71.89%	1.57	00:01:
1 🟦 Organic Search	4,832			68.48%		
2 🟩 Direct	3,190			76.89%		
3 🟥 Referral	1,386			72.19%		
4 🟨 Social	387			78.86%		
5 🟦 (Other)	1			100.00%		


Twitter Analytics account over ×
Secure https://analytics.twitter.com/user/apachemahout/home
Analytics
Home
Tweets
Audiences
Events
More
Apache Mahout
Sign up for Twitter Ads

Account home

Apache Mahout @ApacheMahout

Page updated daily

28 day summary with change over previous period

Tweets	Tweet impressions	Profile visits	Mentions	Followers
7 ↑133.3%	14.9K ↑225.2%	735 ↓0.8%	22 ↑266.7%	3,725 ↑27

May 2017 · 13 days so far...

TWEET HIGHLIGHTS

Top Tweet earned 1,921 impressions

Press release about the latest 0.13.0 release: globenewswire.com/news-release/2…

7 ♥ 3

View Tweet activity

View all Tweet activity

Top mention earned 58 engagements

Apache - The ASF
@TheASF · May 1

The Apache Software Foundation Announces @ApacheMahout v0.13.0 s.apache.org/ioAa #OpenSource #BigData #MachineLearning #DataMining #AI

25 ♥ 18

View Tweet

Top Follower followed by 3,247 people

Darryl Palmer
@DarrylPalmer FOLLOWS YOU

Reformed Astrophysicist and CTO of @JanusChoice, Tweeting about using Machine Learning and Nudging to improve Patient Outcomes and other uses of technology

View profile

View followers dashboard

Tweets with photos get noticed

It's true. Tweets with images **drive more engagement** and generate more responses.

Learn how to share a photo

ADVERTISE ON TWITTER

Get your Tweets in front of more people

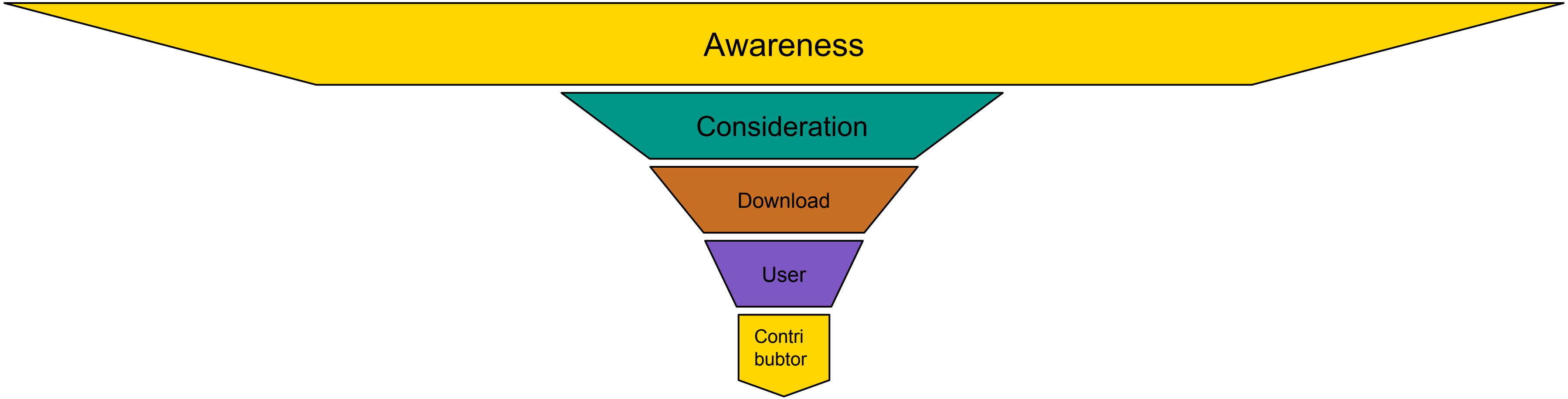
Promoted Tweets and content open up your reach on Twitter to more people.

Get started

MAY 2017 SUMMARY

Tweets	Tweet impressions
5	7,892
Profile visits	**Mentions**
399	14
New followers	
3,715	

Evangelism



Awareness

Consideration

Download

User

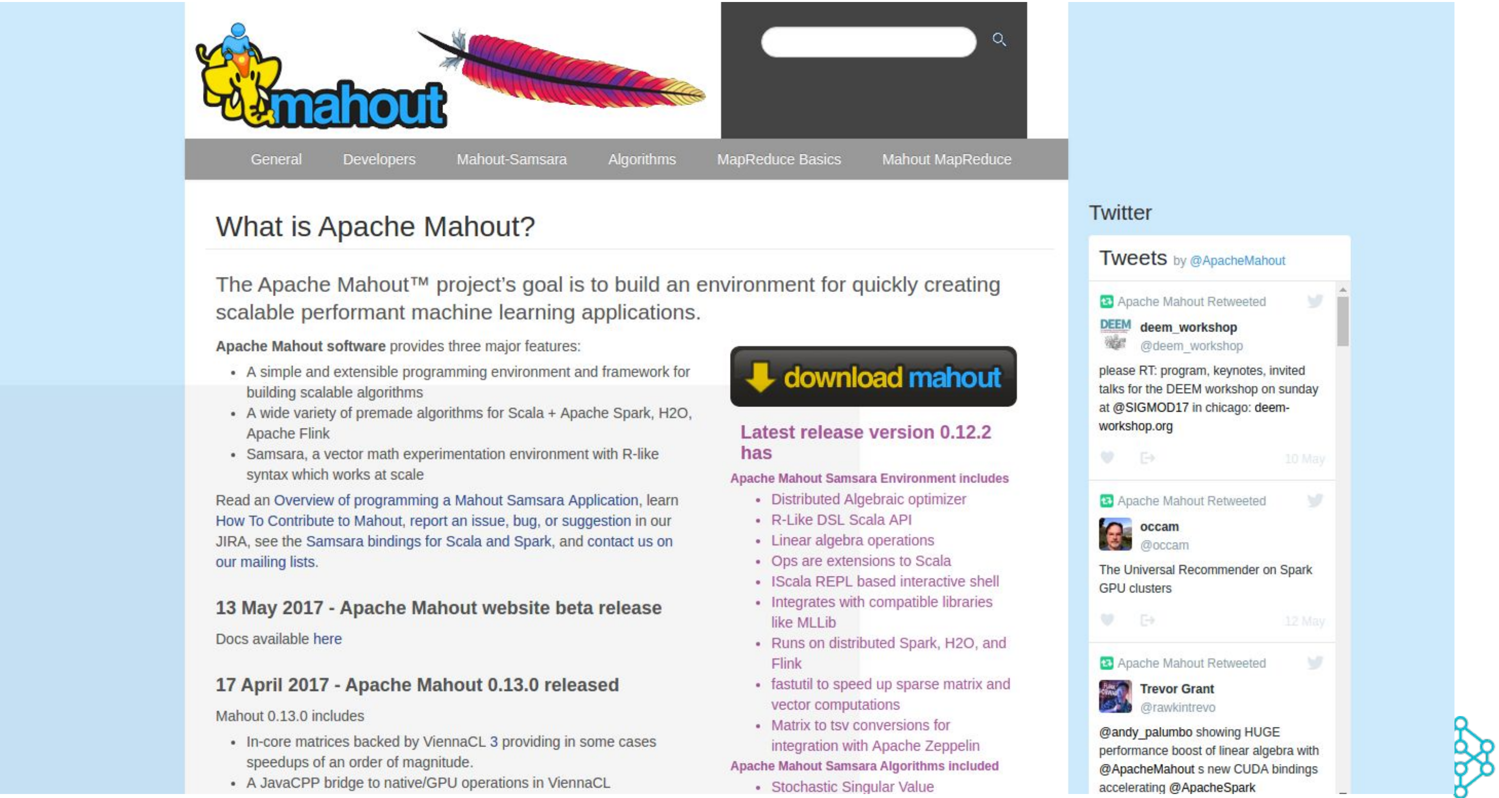
Contri
bubtor

MAHOUT

New Image

MAHOUT

(Old) Website



What is Apache Mahout?

The Apache Mahout™ project's goal is to build an environment for quickly creating scalable performant machine learning applications.

Apache Mahout software provides three major features:

- A simple and extensible programming environment and framework for building scalable algorithms
- A wide variety of premade algorithms for Scala + Apache Spark, H2O, Apache Flink
- Samsara, a vector math experimentation environment with R-like syntax which works at scale

Read an Overview of programming a Mahout Samsara Application, learn How To Contribute to Mahout, report an issue, bug, or suggestion in our JIRA, see the Samsara bindings for Scala and Spark, and contact us on our mailing lists.

13 May 2017 - Apache Mahout website beta release

Docs available here

17 April 2017 - Apache Mahout 0.13.0 released

Mahout 0.13.0 includes

- In-core matrices backed by ViennaCL 3 providing in some cases speedups of an order of magnitude.
- A JavaCPP bridge to native/GPU operations in ViennaCL

Latest release version 0.12.2 has

Apache Mahout Samsara Environment includes

- Distributed Algebraic optimizer
- R-Like DSL Scala API
- Linear algebra operations
- Ops are extensions to Scala
- IScala REPL based interactive shell
- Integrates with compatible libraries like MLLib
- Runs on distributed Spark, H2O, and Flink
- fastutil to speed up sparse matrix and vector computations
- Matrix to tsv conversions for integration with Apache Zeppelin

Apache Mahout Samsara Algorithms included

- Stochastic Singular Value

Twitter

Tweets by @ApacheMahout

Apache Mahout Retweeted

deem_workshop @deem_workshop

please RT: program, keynotes, invited talks for the DEEM workshop on sunday at @SIGMOD17 in chicago: deem-workshop.org

10 May

Apache Mahout Retweeted

occam @occam

The Universal Recommender on Spark GPU clusters

12 May

Apache Mahout Retweeted

Trevor Grant @rawkintrevo

@andy_palumbo showing HUGE performance boost of linear algebra with @ApacheMahout s new CUDA bindings accelerating @ApacheSpark

New Image

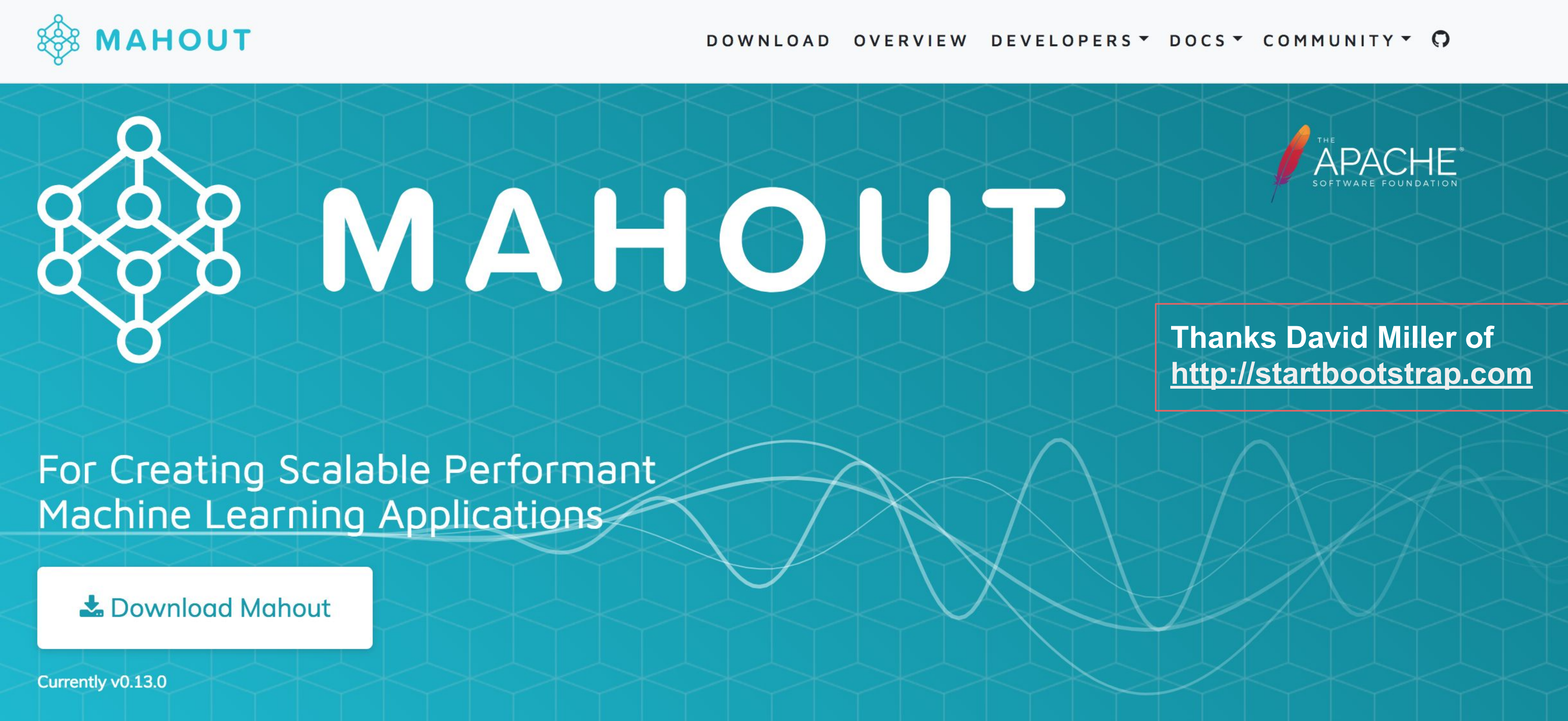
Problem: Old Website

Opportunity: Websites are easy (technically- design on the other hand...)

Solution: Reboot website with Jekyll Bootstrap (and less emphasis on Map-Reduce)

Corollary: Much easier for committers and contributors to update website, add tutorials, etc. Encourages (requires) new features to be submitted with good docs.

New Website



For Creating Scalable Performant Machine Learning Applications

Currently v0.13.0

Thanks David Miller of http://startbootstrap.com

New Logo

Apache
mahout

Relevant Features

MAHOUT

Killer New Features

GPU Integration

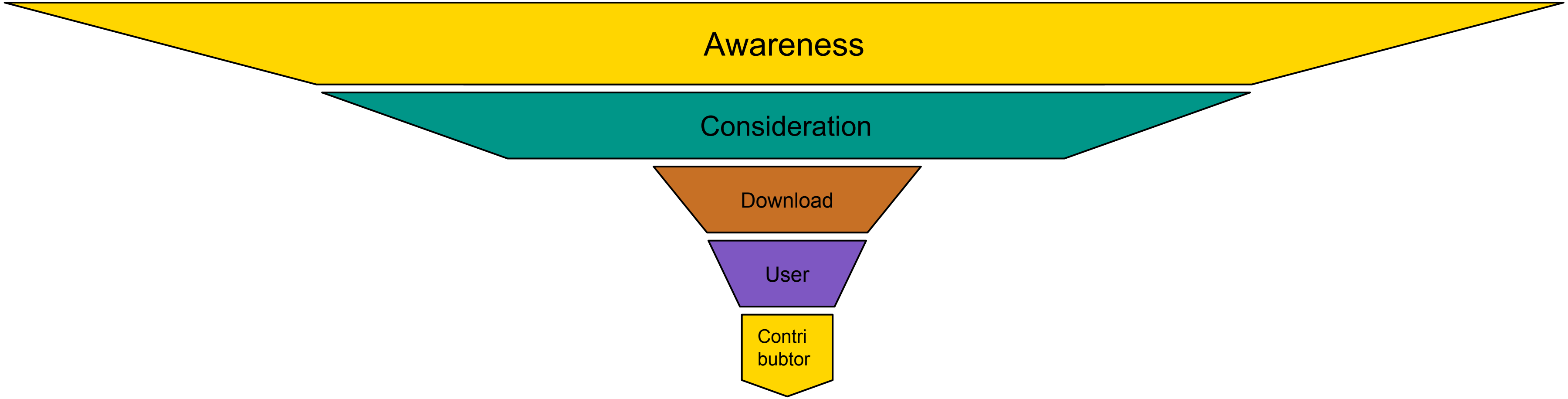
Algo Framework / Precanned Algos

Zeppelin Integration

Mahout-Samsara (Mathematically Expressive Scala DSL)

Etc.

New Features



Awareness

Consideration

Download

User

Contri
bubtor

Apache Zeppelin integration

MAHOUT

Huge win

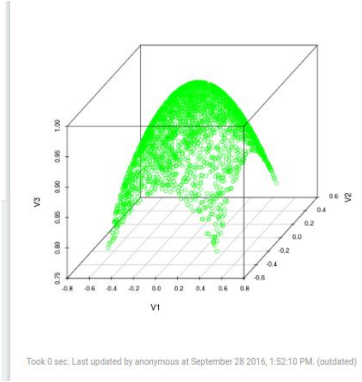
Scala has weak visualization.

R and Python have great visualization!

Zeppelin allows user to hand off variables between interpreters in notebook.

Do work in Mahout (Scala) - Plot in R/Python!

Mahout Gaussian DRM plotted in R



Took 0 sec. Last updated by anonymous at September 28 2016, 1:52:10 PM. (outdated)

Huge win- kinda wonky install.

A script exists for assisting with install… needs update for 0.13.0
https://issues.apache.org/jira/browse/ZEPPELIN-2417

Major Issue- Zeppelin out of the box supports Scala 2.11/Spark 2.x

Mahout 0.13.0 will build, but no binaries exist.

User must either build Mahout (for Spark 2.1) or Build Zeppelin (for Spark 1.6)

Fix coming soon in 0.13.1 (profiles and binaries for Spark 2.1/Scala 2.11)
http://mahout.apache.org/docs/0.13.1-SNAPSHOT/tutorials/misc/mahout-in-zeppelin/

MAHOUT

Or so we thought...

Originally we planned on releasing 0.13.1 weeks after 0.13.0

There were issues with `scopt` being version locked on Scala 2.10 (required for CLI drivers).

Also, this caused issues in `sbt` builds (obvious solution, no one should use sbt?)

Eventually led to...

Which Led to

Disagreements on mailing list / Slack.

People blaming people for things.

Other issues one has when fighting with family around the holidays.

And Finally...

Massive POM ~~cleaning~~ gutting.

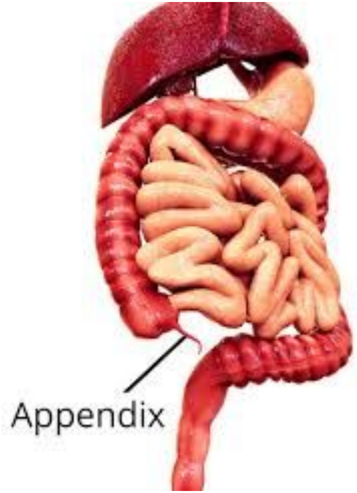
Project structure refactor (MR now in Community, Math components combined).

In progress now.

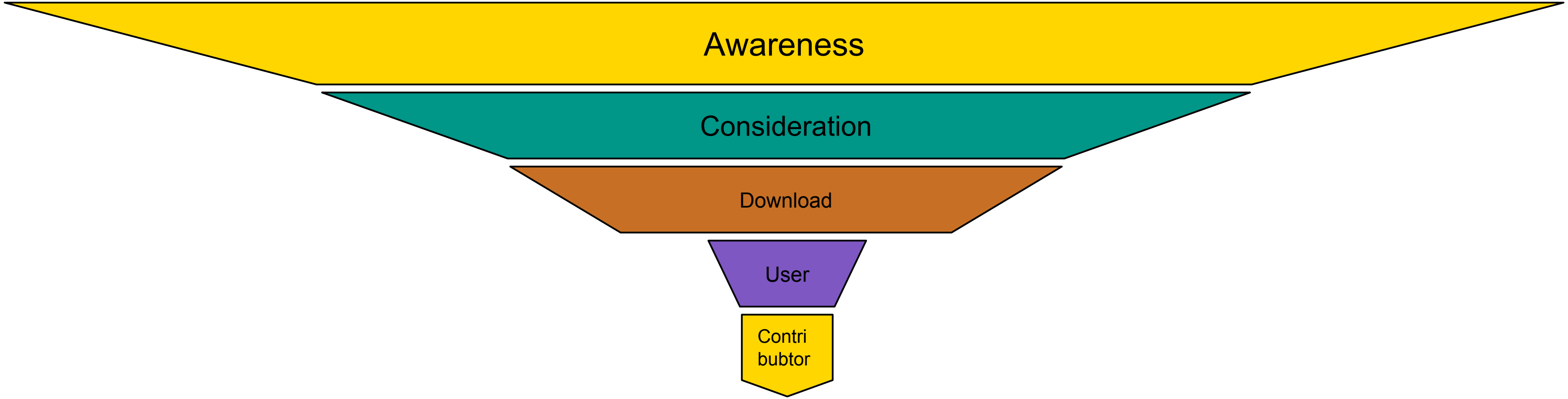
MAHOUT

Sometimes Code is like an Appendix

Once useful, but so long ago- no can remember why.



Other Project Integration



Awareness

Consideration

Download

User

Contri
bubtor

MAHOUT

Mathematically Expressive Scala DSL

MAHOUT

Big Math

Problem: May not know Scala/Apache Spark/Apache Flink

Opportunity: Do know R

Solution: Create an abstracted language- mathematically expressive Scala DSL

Big Math

```
implicit val sdc: org.apache.mahout.sparkbindings.SparkDistributedContext = sc2sdc(sc)

val A = drmWrap(rddA)
val B = drmWrap(rddB)

val C = A.t %*% A + A %*% B.t
```

`C` is a `RDD[(Int, org.apache.mahout.Vector)]`

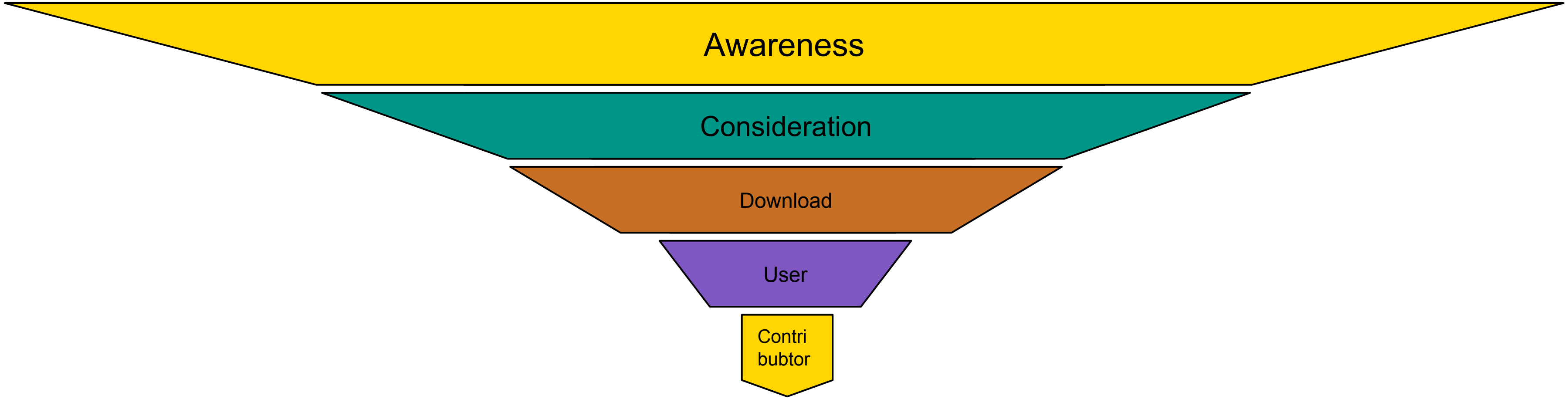
Also have truly distributed matrix decompositions.

“Possibly” Coming soon: Mahout for Tensorflow

Problem- Tensorflow is very very ugly.

We can use Tensorflow 2d Matrices with Mahout Scala DSL.

Big Math



Engine Neutrality

Engine Neutrality

Problem: Distributed engines come and go (we learned this first)

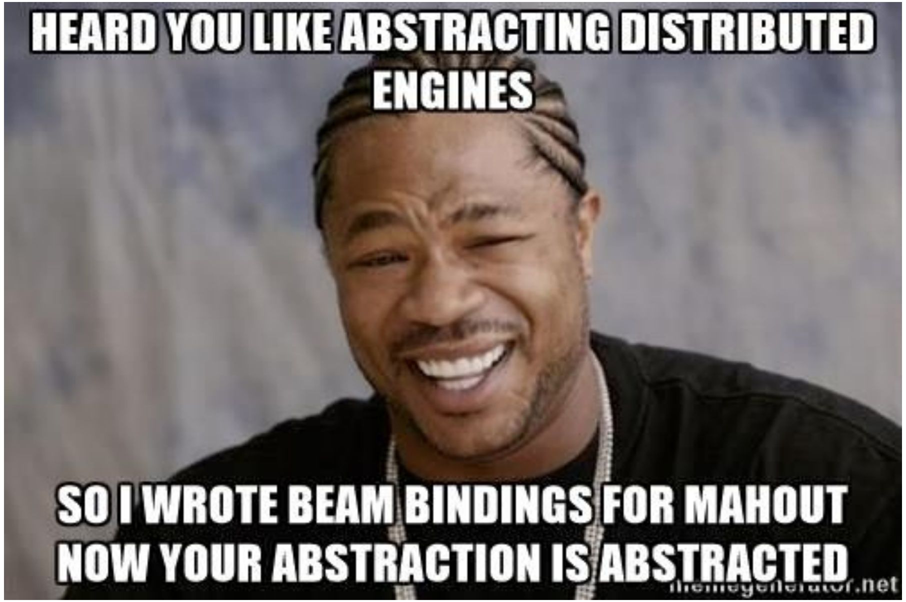
Opportunity: We learned this first!

Solution: Create Engine Neutral Libraries which can bind to new engines

Corollary: Implement algorithm once- run it anywhere*

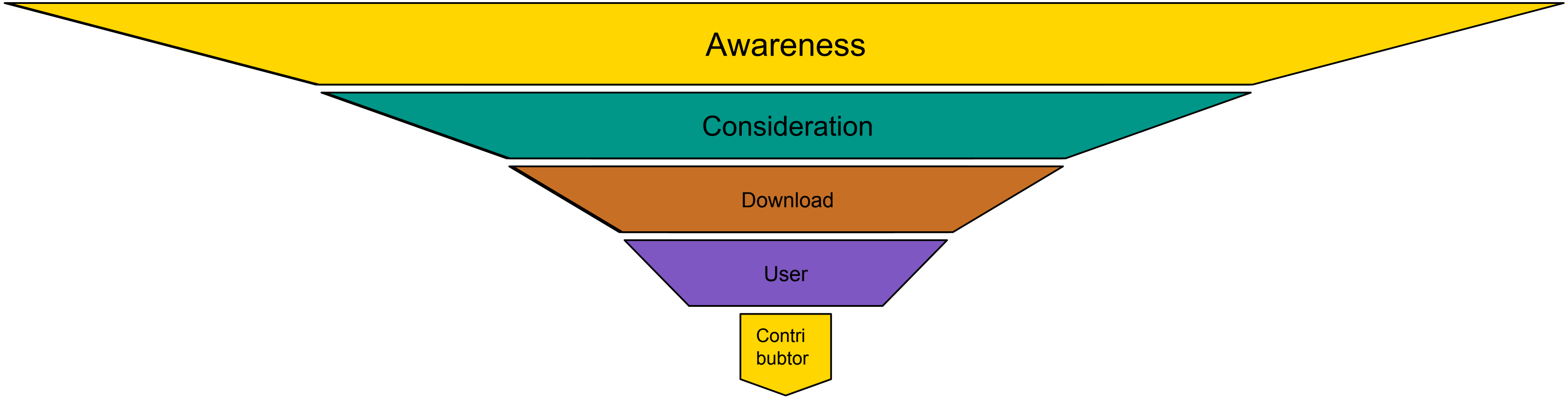
MAHOUT

We don’t actually have Beam bindings (yet), this is just for lulz


HEARD YOU LIKE ABSTRACTING DISTRIBUTED ENGINES
SO I WROTE BEAM BINDINGS FOR MAHOUT
NOW YOUR ABSTRACTION IS ABSTRACTED
memegenerator.net

MAHOUT

Engine Neutrality



Awareness

Consideration

Download

User

Contri bubtor

Algorithm Framework

MAHOUT

Attracting Contributors

Problem: Very small pool of qualified contributors

Opportunity: Mathematically expressive scala makes it easy to write and review “math part”

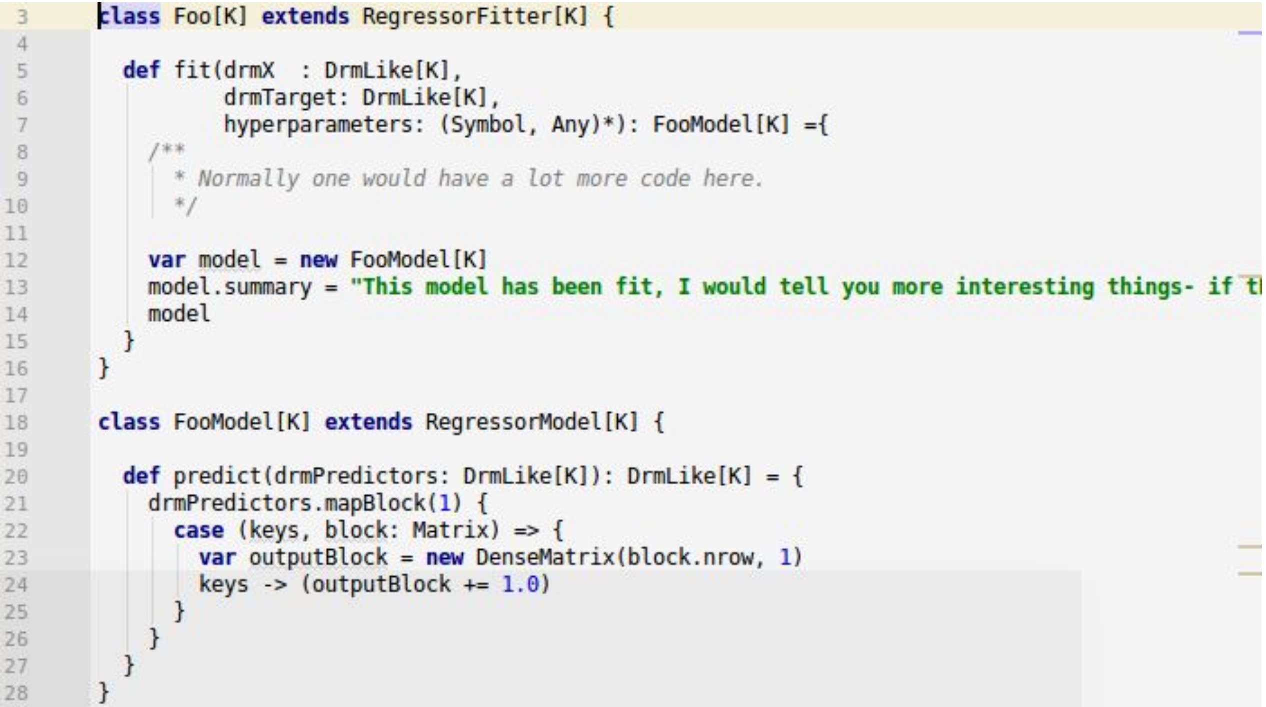
Solution: Create templates and tutorials showing users how to add algorithm

Drop your algorithms in easily

Algorithms are easy to compose in Mahout (as easy* as R, often can use R implementation for guidance)

Scala classes / package layout may still be overwhelming for our target users

Algorithm Template



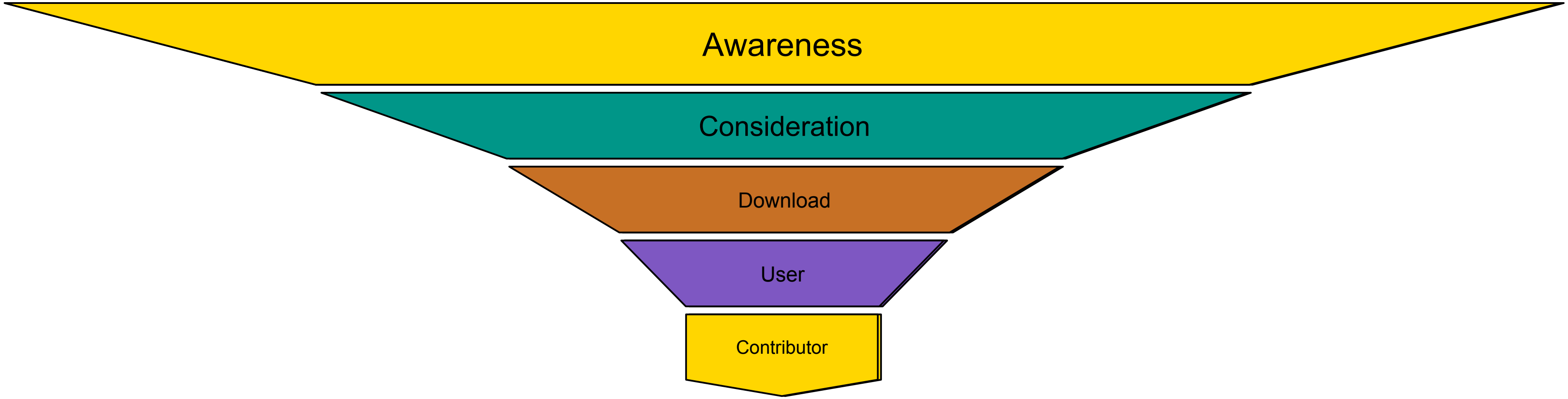
```
class Foo[K] extends RegressorFitter[K] {

  def fit(drmX  : DrmLike[K],
          drmTarget: DrmLike[K],
          hyperparameters: (Symbol, Any)*): FooModel[K] ={
    /**
      * Normally one would have a lot more code here.
      */

    var model = new FooModel[K]
    model.summary = "This model has been fit, I would tell you more interesting things- if t
    model
  }
}

class FooModel[K] extends RegressorModel[K] {

  def predict(drmPredictors: DrmLike[K]): DrmLike[K] = {
    drmPredictors.mapBlock(1) {
      case (keys, block: Matrix) => {
        var outputBlock = new DenseMatrix(block.nrow, 1)
        keys -> (outputBlock += 1.0)
      }
    }
  }
}
```

Encourage using dev

A lot of hand holding with first time contributors- encourage them to pay it forward.

(Mailing lists can be hard. Guy who did the website, DID the website, took him months to figure out how to subscribe to dev).

Engine Neutrality

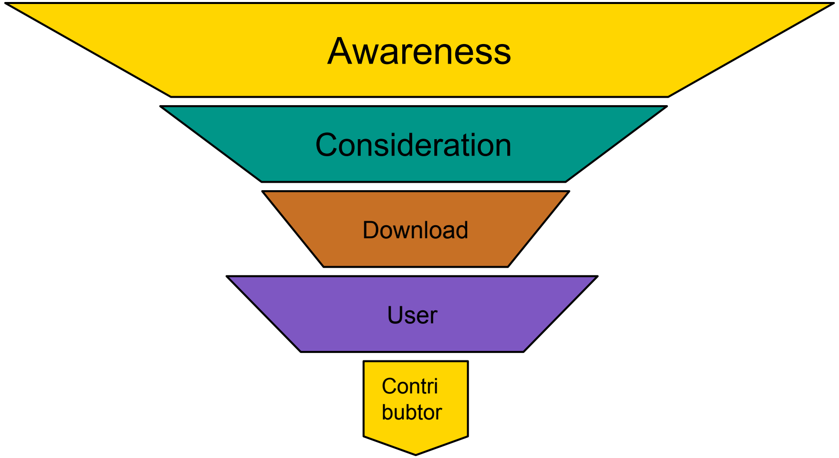


MAHOUT

How this actually progressed...

MAHOUT

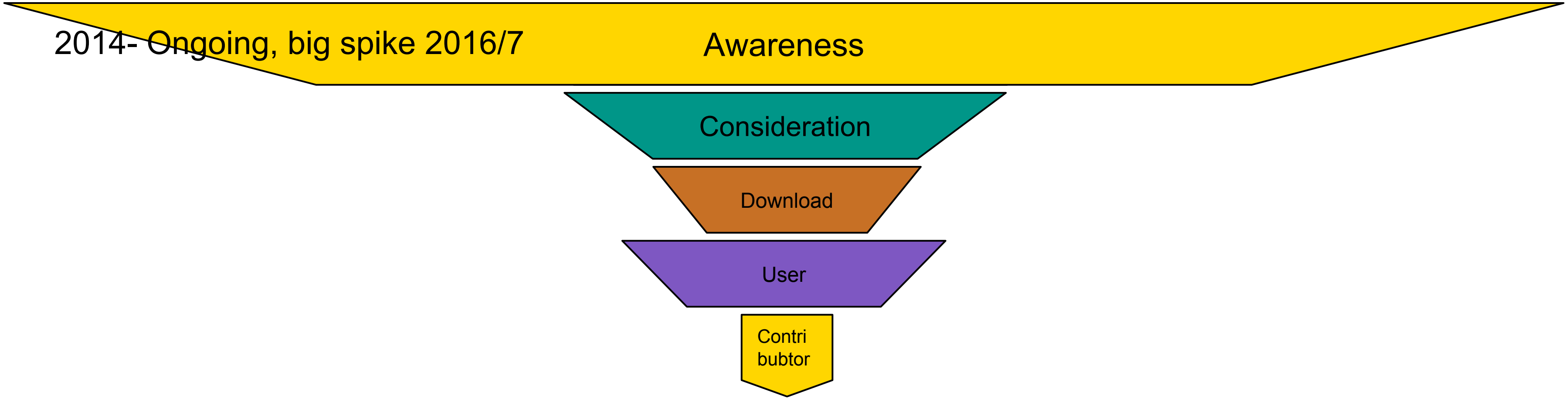
First The Math and Engine Neutrality



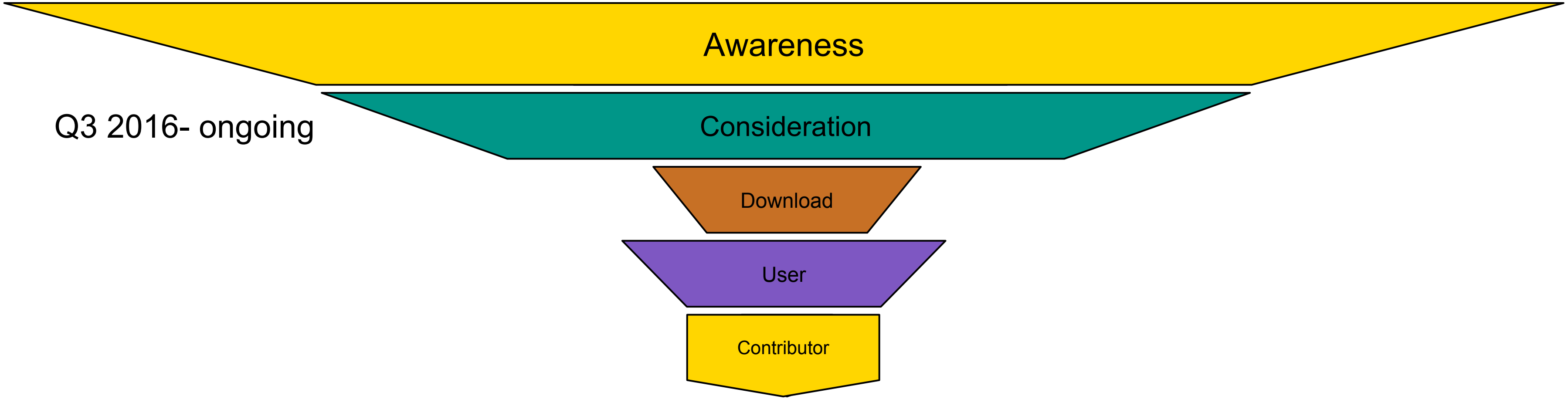
2014-2016

MAHOUT

Then we started our evangelism

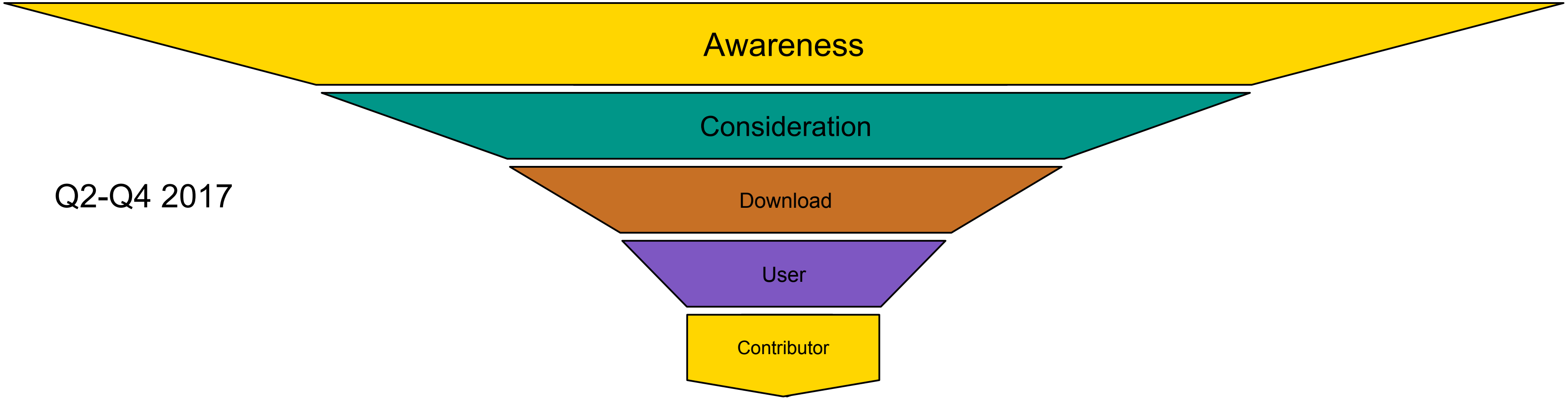


Then the GPU work and algorithm framework



Q3 2016- ongoing

New website...



Awareness

Consideration

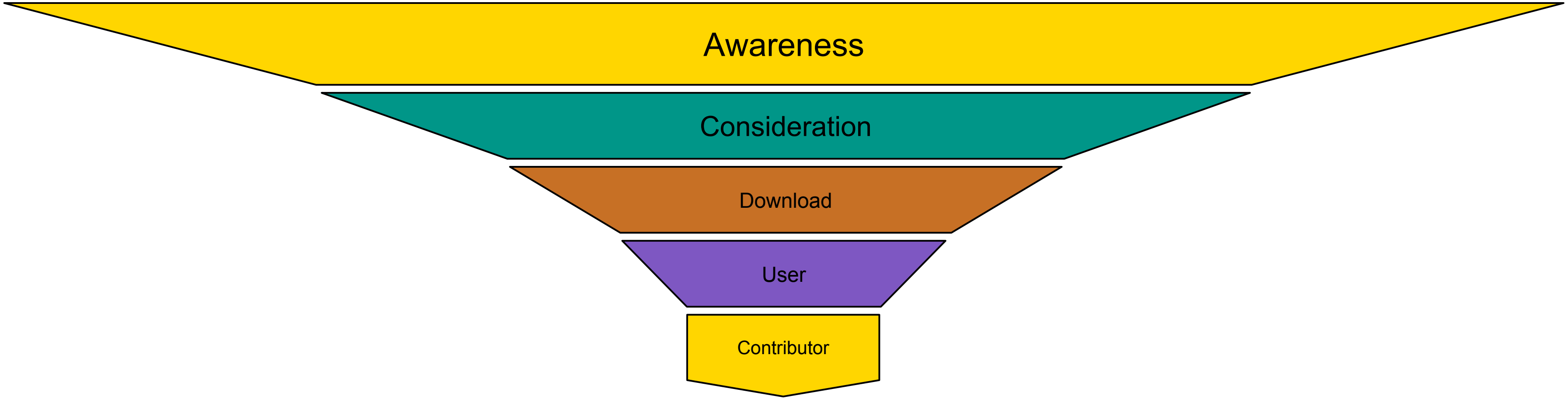
Download

User

Contributor

Q2-Q4 2017

Hooray!



What’s Next

MAHOUT

Building Algorithms Framework

Framework in place which encourages users to contribute algorithms (already paying off)

Seeking to grow “pre-canned” algorithms collection between now and v 0.14.0

Eventually a “CRAN” like repository of algorithms for Mahout.

More work on GPUs

We consider our GPU support a HUGE differentiator among ML packages native to distributed engines (MLLib, FlinkML, etc).

Still opportunities for optimization-

Recent benchmark on unoptimized (still technically PR) CUDA bindings show “significant” speed up on sparse Matrix multiplication

300%

Speed up On Sparse-Sparse Matrix Multiplication on AWS GPU enabled Spark Cluster

Kind of a big deal.

MAHOUT

More Engine Bindings (Tensorflow)

Create template engine bindings- even if not optimized.

Tutorials on writing new engine bindings.

We feel this is also a huge differentiator.

Possibly “Community” supported engine bindings, not officially supported- but in the trunk, attract “drive-by” contributions from other projects.

MAHOUT

Getting over version lock...

Some issue had us version locked on Scala 2.10 (Spark 1.6).

Huge refactor of the POMs

“Map-Reduce” is “Community” now.

- We’re “soon” (maybe) going to be taking MR PR’s again (have not accepted them for 3.5 years)

Conclusion

MAHOUT

Awakening the Giant

Mahout has quietly undergone huge transformation from Map Reduce / Java based Machine Learning to Mathematically Expressive Scala / Engine Neutral / GPU Accelerated

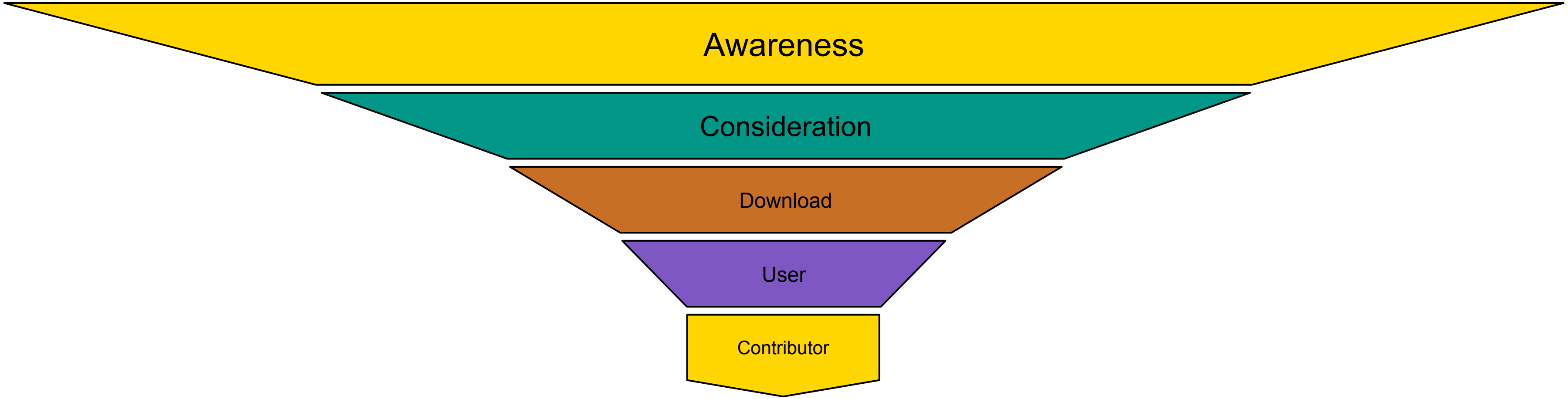
Need to let everyone know- I mean you- go tell your friends and tweet and write a blog.

MAHOUT

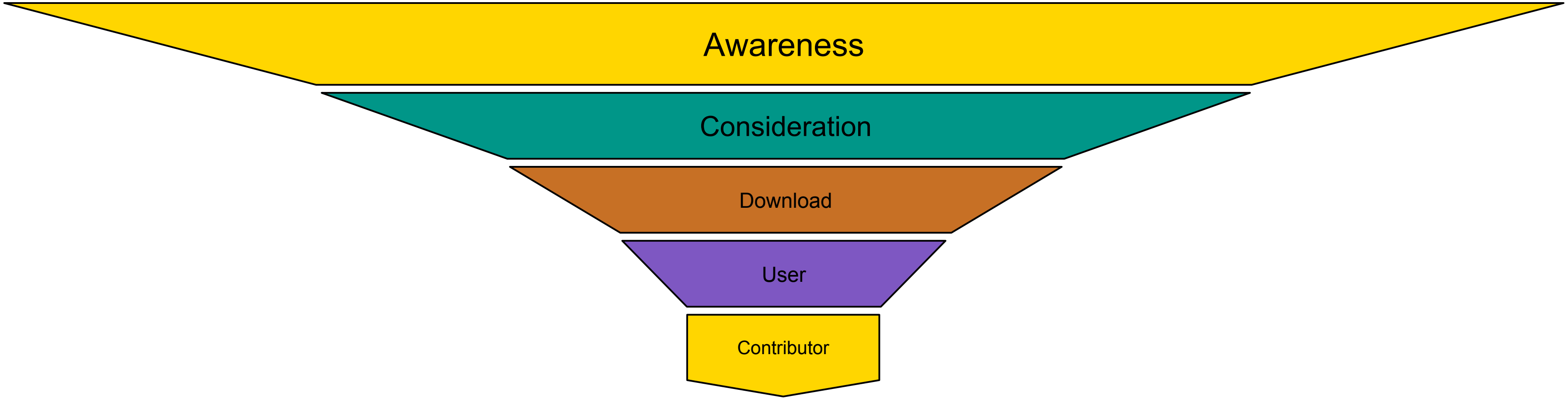
“I want you so badly” - The Beatles

MAHOUT

Remember the funnel for your project.



People can only download if they are aware.



Awareness

Consideration

Download

User

Contributor

MAHOUT

Questions?

MAHOUT