The Challenges of Crowdsourcing Multidimensional Data

Sean Gorman PhD
Timbr.io
@seangorman

timbr.io

A Traditional Geospatial Workflow

Data Collection

Data Processing
& Analysis

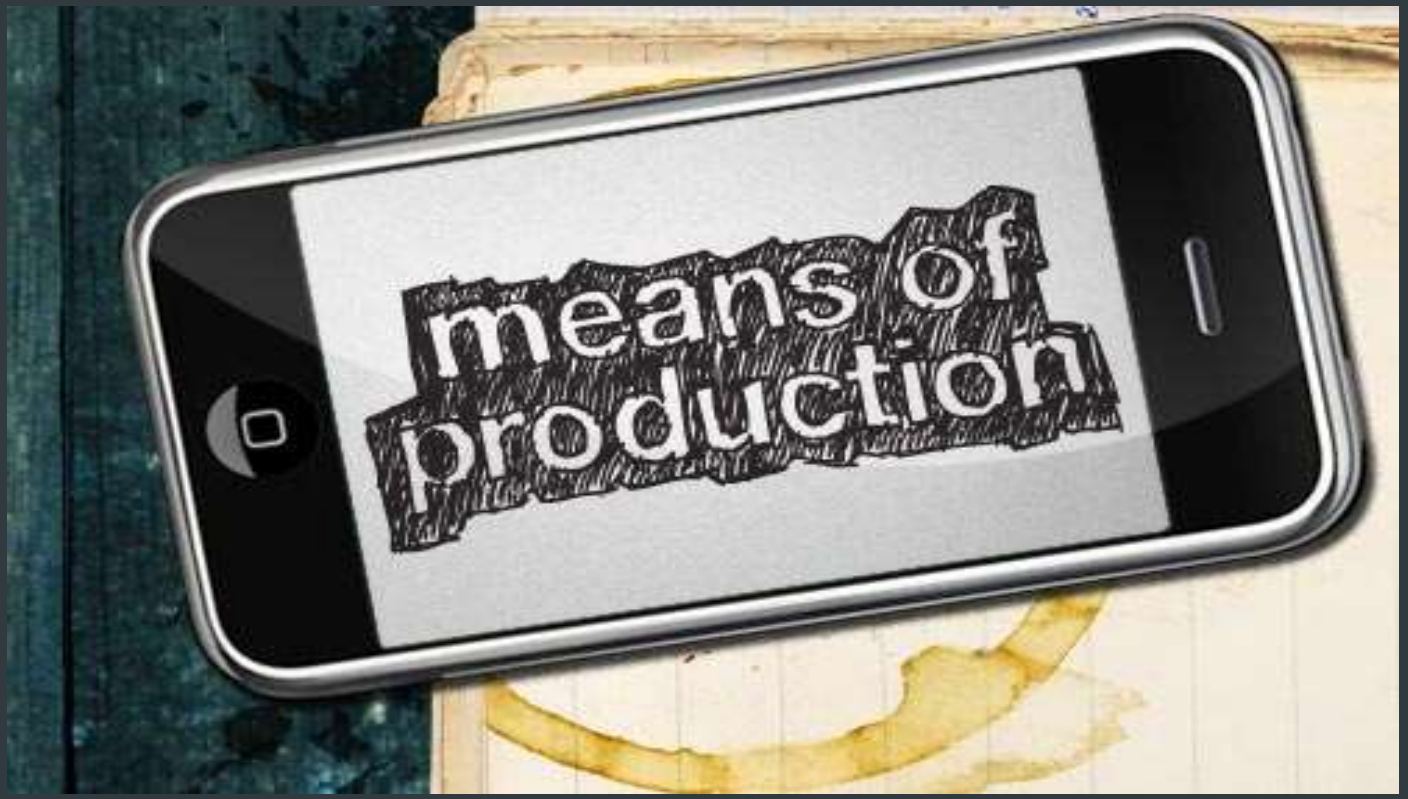
Data Products

Data’s full life cycle was monolithic

timbr.io

Creating inherently authoritative data

timbr.io



A fundamental shift in who is creating data

Authoritative Data

Geography Network, OSM, GeoCommons, AGOL, ESRI Open Data

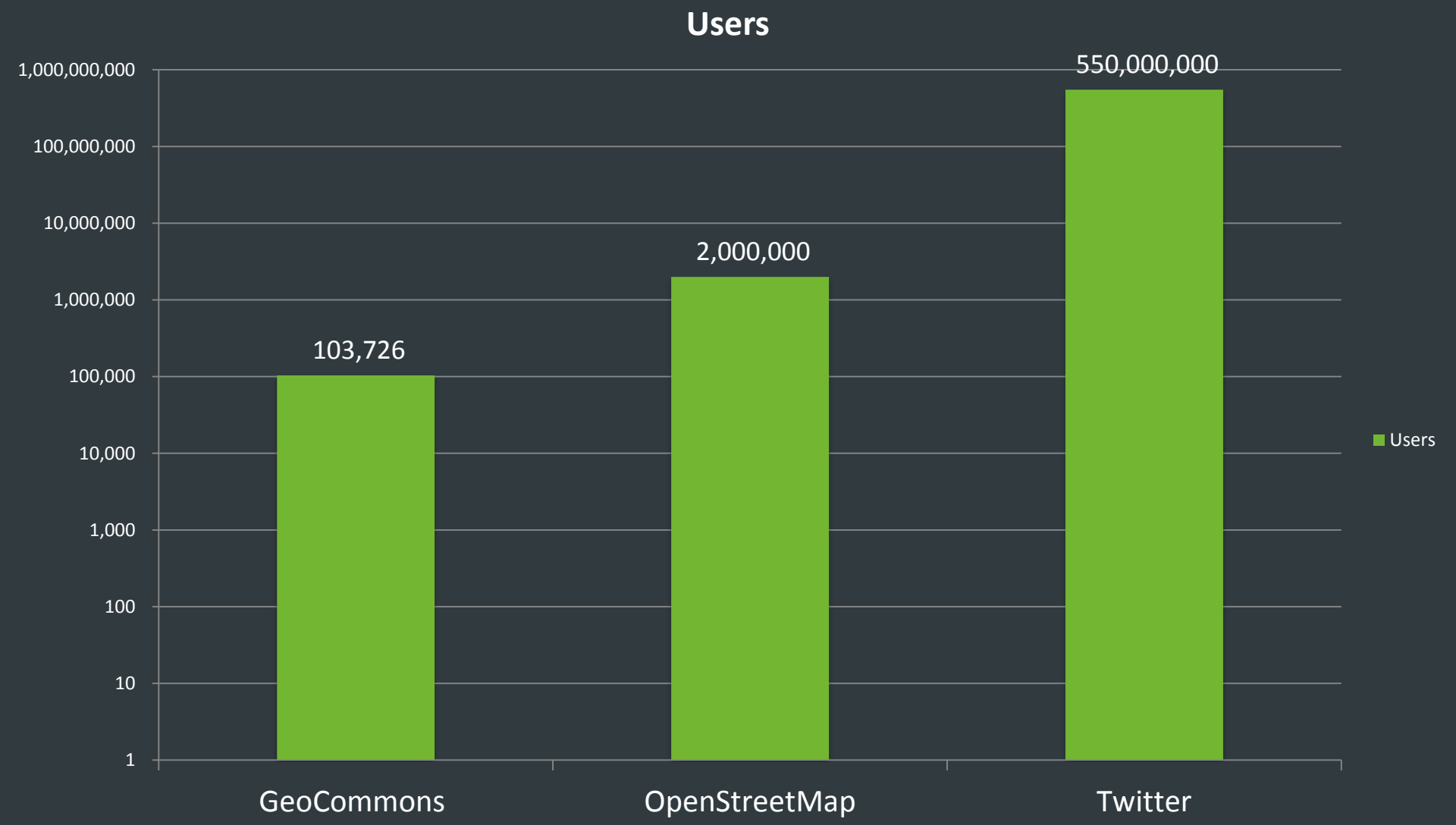
Volunteered Geographic Information

OSM, WikiMapia, Waze (?)

Ambient Geographic Information

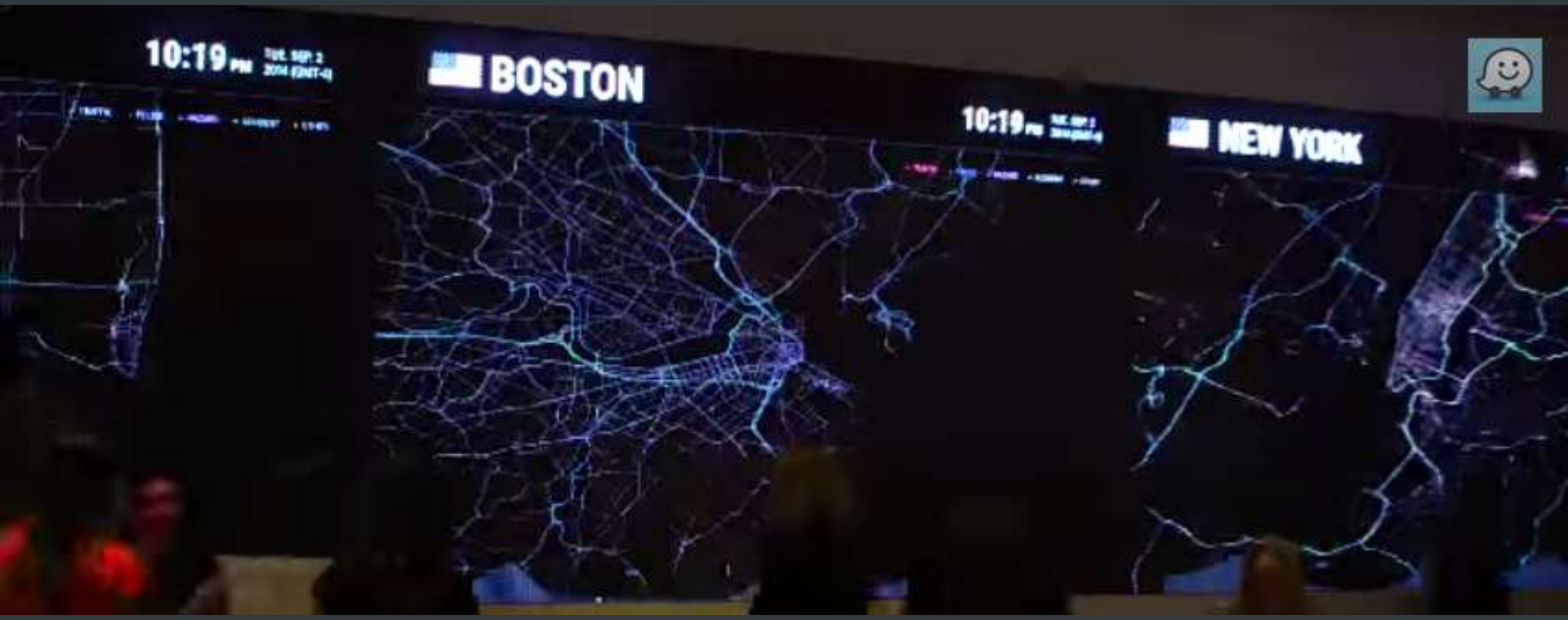
Twitter, FourSquare, Instagram, Fitbit, Strava, Google Traffic

CROWDSOURCING


Users
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
103,726
2,000,000
550,000,000
GeoCommons
OpenStreetMap
Twitter
Users

Strava Metro

Strava Metro is a data service providing "ground truth" on where people ride and run. Millions of GPS-tracked activities are uploaded to Strava every week from around the globe. In denser metro areas, nearly one-half of these are commutes. These activities create billions of data points that, when aggregated, enable deep analysis and understanding of real-world cycling and pedestrian route preferences.

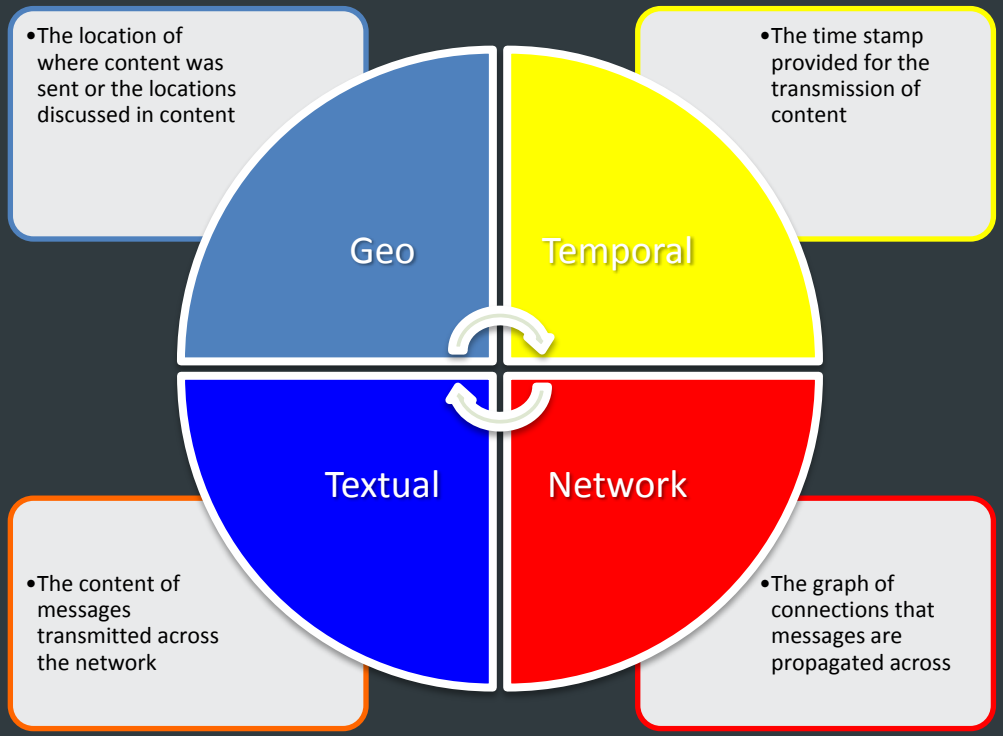


Waze Connected Citizens Program

Ambient Geographic Information

timbr.io

Data Dimensions




How do we deal with data diversity?

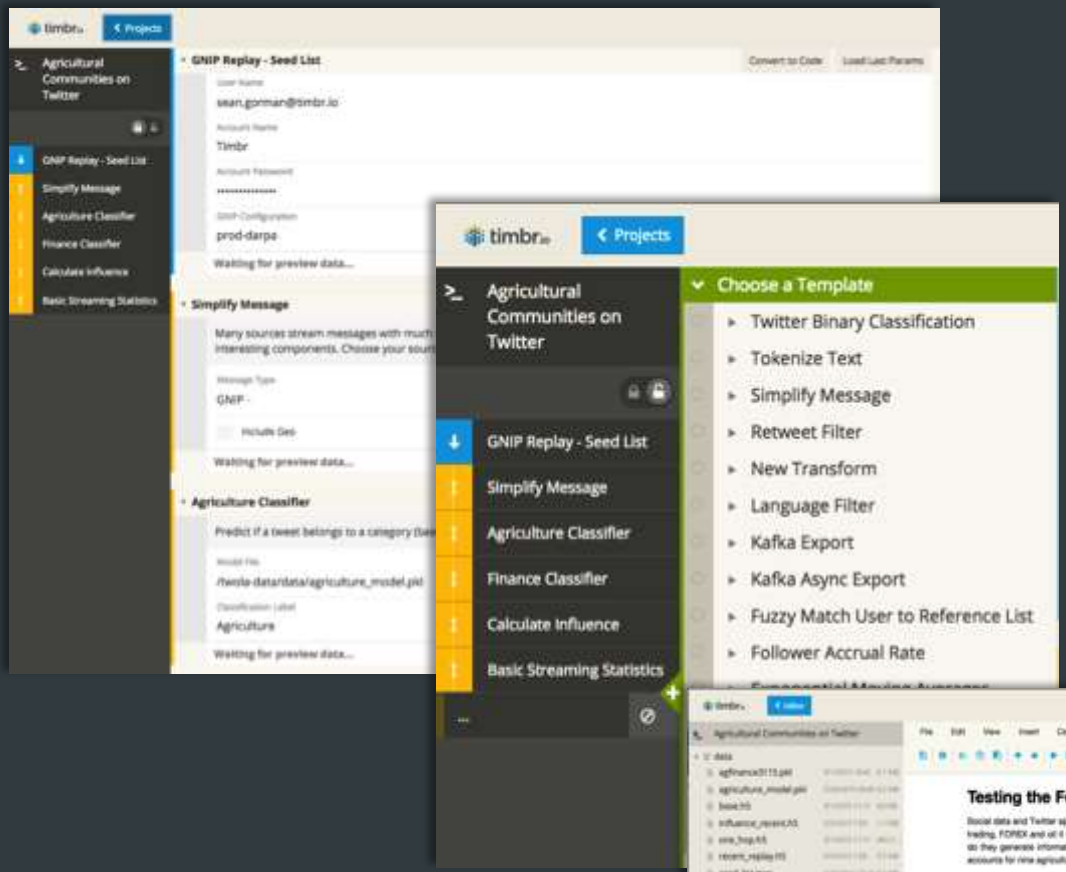
Method	Github Repo Link	Wikipedia Description Link
Naive Bayesian Classifier	https://github.com/harthur/classifier	https://en.wikipedia.org/wiki/Naive_Bayes_classifier
Distributed stochastic gradient descent	https://github.com/MrChrisJohnson/Cc	https://en.wikipedia.org/wiki/Stochastic_gradient_boo
Decision trees	https://github.com/chris-taylor/aima-ha	https://en.wikipedia.org/wiki/Decision_trees
Belief propagation graph models	https://github.com/gilesc/factor-graph	https://en.wikipedia.org/wiki/Belief_propagation
Dynamic Bayesian networks	https://github.com/danielkorzekwa/bay	https://en.wikipedia.org/wiki/Dynamic_Bayesian_netv
Conditional random fields	https://github.com/chokkan/crfsuite	https://en.wikipedia.org/wiki/Conditional_random_fiel
TF-IDF similarity	https://github.com/bbcrd/Similarity	https://en.wikipedia.org/wiki/Tf%E2%80%93idf
Jaro Winkler Distance	https://github.com/sunlightlabs/jellyfish	https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkle
State Vector Machine (SVM)	https://github.com/karpathy/svmjs	https://en.wikipedia.org/wiki/Support_vector_machine
K Nearest neighbor (k-NN)	https://github.com/luispedro/milk	https://en.wikipedia.org/wiki/K-nearest_neighbors_alg
Single Value Decomposition (SVD)	https://github.com/willcannings/Ruby-S	https://en.wikipedia.org/wiki/Singular_value_decompo
Latent Dirichlet Allocation	https://github.com/shuyo/iir/blob/	https://en.wikipedia.org/wiki/Latent_Dirichlet_Allocati
Hierarchical Latent Dirichlet Allocation	https://github.com/josephreisinger/lvm-	https://en.wikipedia.org/wiki/Latent_Dirichlet_Allocati
Gibbs sampling	https://github.com/bgamari/bayes-stac	https://en.wikipedia.org/wiki/Gibbs_sampling
Markov Models	https://github.com/emilmont/Artificial-In	https://en.wikipedia.org/wiki/Markov_models

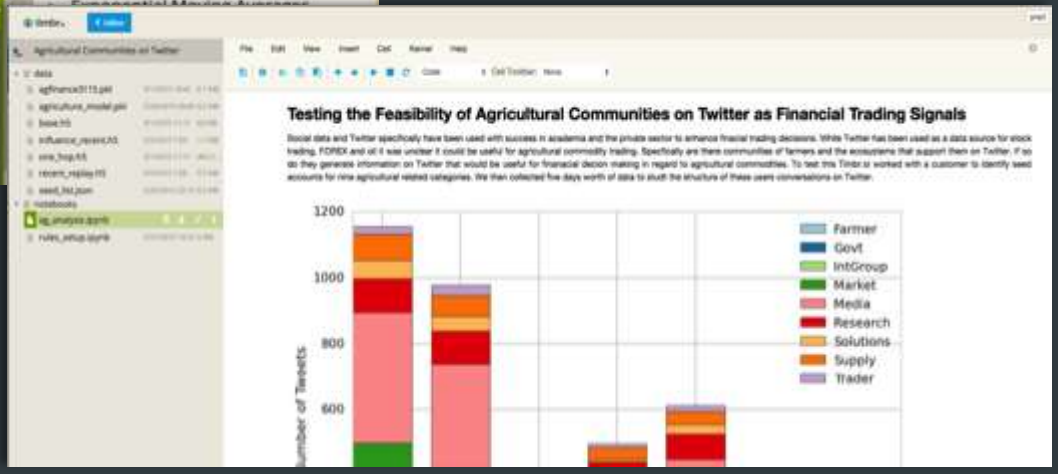
A Multitude of Algorithms for Multidimensional Data

timbr.io

Can we crowdsource algorithms and match them to data structures?

timbr.io





Our current work

timbr.io

THANKS

Sean Gorman
sean.gorman@timbr.io
202-321-3914
@seangorman

timbr.io