

# The Challenges of Crowdsourcing Multidimensional Data

Sean Gorman PhD  
Timbr.io  
@seangorman

# A Traditional Geospatial Workflow



Data Collection



Data Processing  
& Analysis



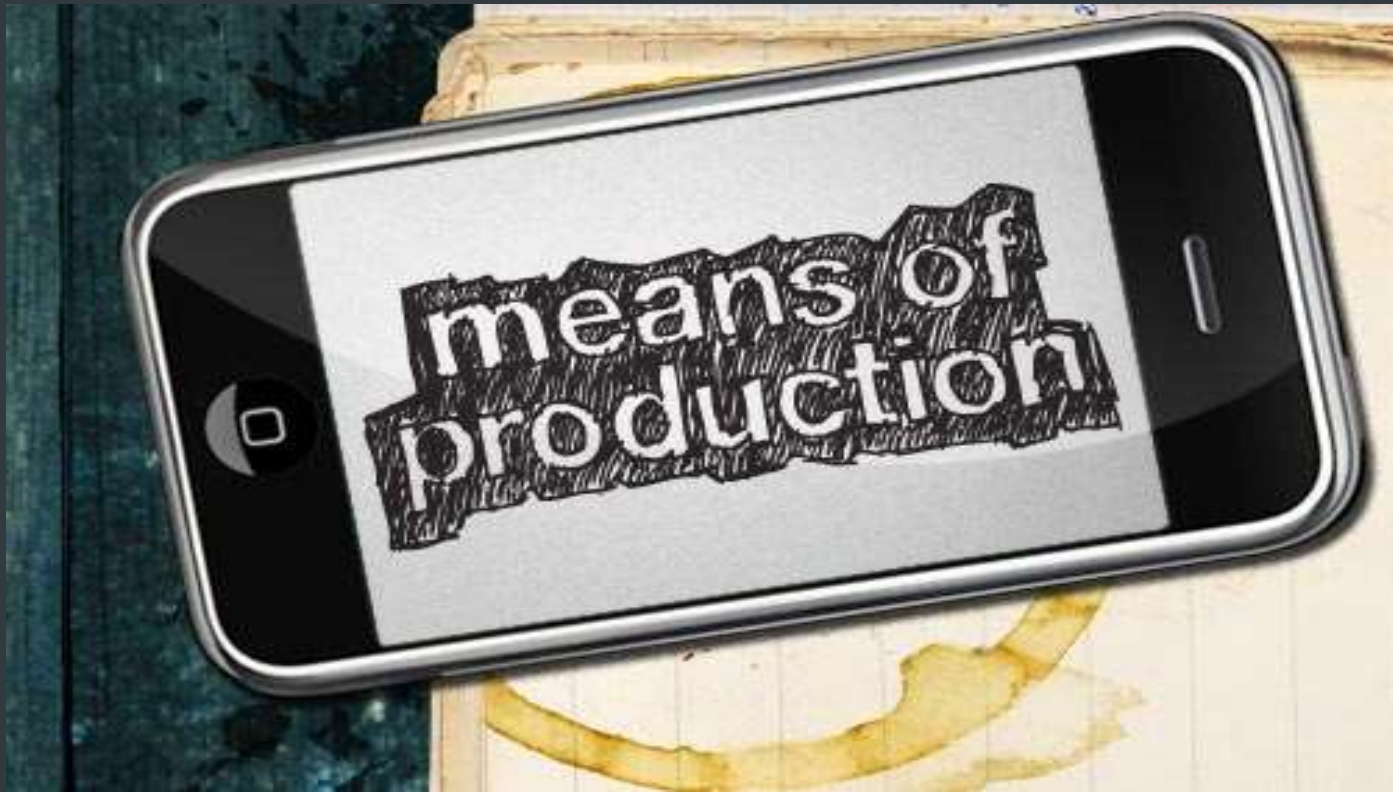
Data Products



Data's full life cycle was monolithic



Creating inherently authoritative data



A fundamental shift in who is creating data

## Authoritative Data

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

## Volunteered Geographic Information

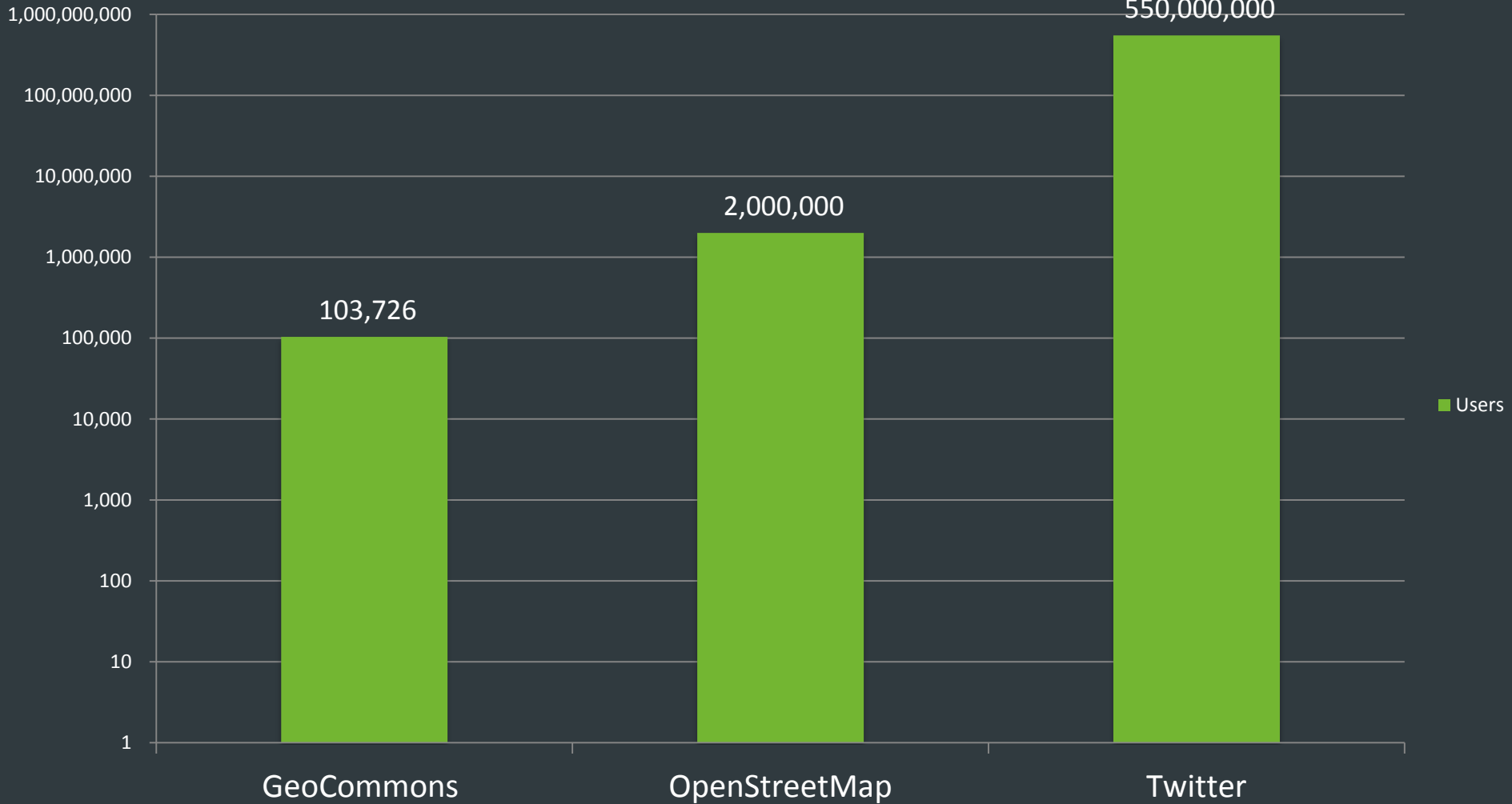
*OSM, WikiMapia, Waze (?)*

## Ambient Geographic Information

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

# CROWDSOURCING

## Users



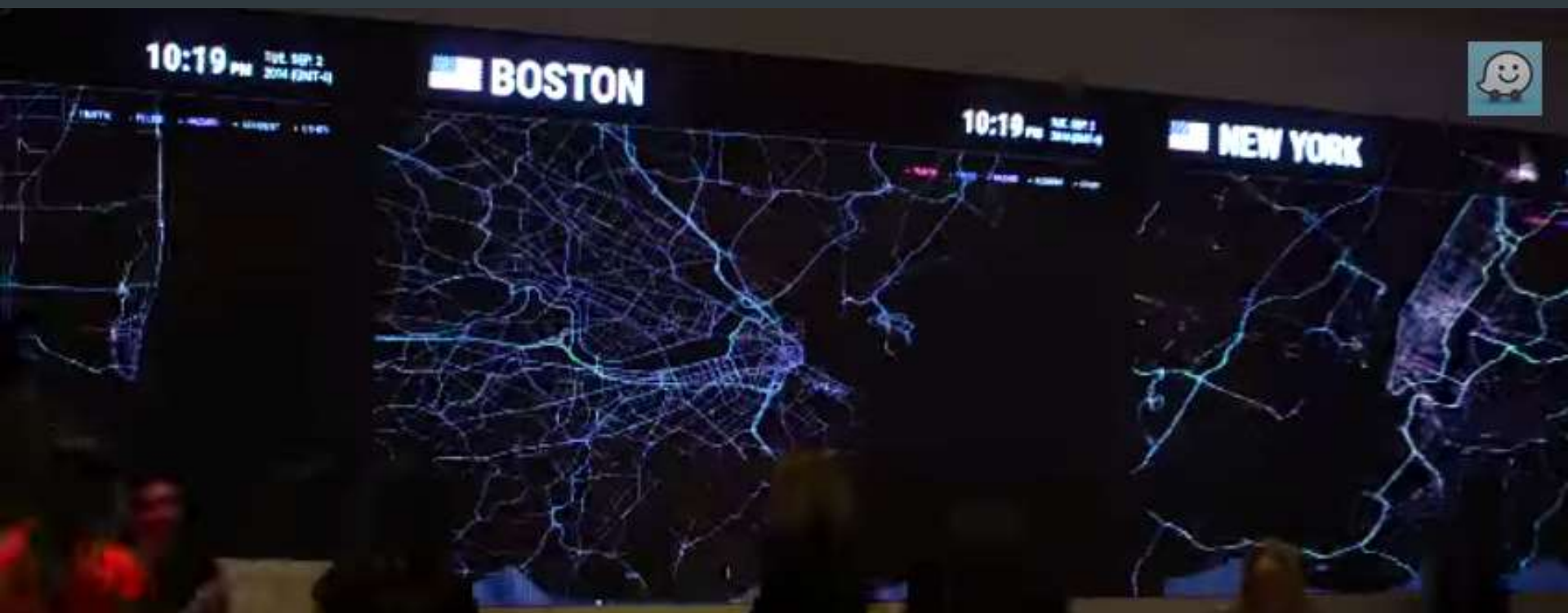


# Strava Metro

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.



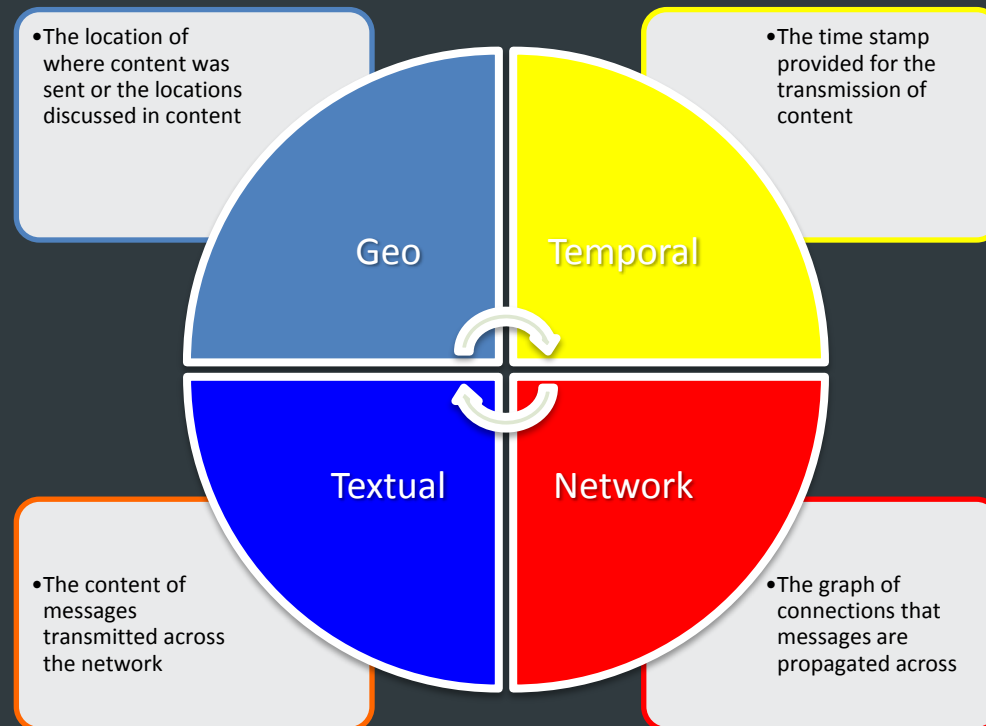


# Waze Connected Citizens Program



# Ambient Geographic Information

# Data Dimensions





How do we deal with data diversity?

Method	Github Repo Link	Wikipedia Description Link
Naive Bayesian Classifier	<a href="https://github.com/harthur/classifier">https://github.com/harthur/classifier</a>	<a href="https://en.wikipedia.org/wiki/Naive_Bayes_classifier">https://en.wikipedia.org/wiki/Naive_Bayes_classifier</a>
Distributed stochastic gradient descent	<a href="https://github.com/MrChrisJohnson/Cc">https://github.com/MrChrisJohnson/Cc</a>	<a href="https://en.wikipedia.org/wiki/Stochastic_gradient_descent">https://en.wikipedia.org/wiki/Stochastic_gradient_descent</a>
Decision trees	<a href="https://github.com/chris-taylor/aima-ha">https://github.com/chris-taylor/aima-ha</a>	<a href="https://en.wikipedia.org/wiki/Decision_trees">https://en.wikipedia.org/wiki/Decision_trees</a>
Belief propagation graph models	<a href="https://github.com/gilesc/factor-graph">https://github.com/gilesc/factor-graph</a>	<a href="https://en.wikipedia.org/wiki/Belief_propagation">https://en.wikipedia.org/wiki/Belief_propagation</a>
Dynamic Bayesian networks	<a href="https://github.com/danielkorzekwa/bay">https://github.com/danielkorzekwa/bay</a>	<a href="https://en.wikipedia.org/wiki/Dynamic_Bayesian_network">https://en.wikipedia.org/wiki/Dynamic_Bayesian_network</a>
Conditional random fields	<a href="https://github.com/chokkan/crfsuite">https://github.com/chokkan/crfsuite</a>	<a href="https://en.wikipedia.org/wiki/Conditional_random_field">https://en.wikipedia.org/wiki/Conditional_random_field</a>
TF-IDF similarity	<a href="https://github.com/bbcrd/Similarity">https://github.com/bbcrd/Similarity</a>	<a href="https://en.wikipedia.org/wiki/Tf%E2%80%93idf">https://en.wikipedia.org/wiki/Tf%E2%80%93idf</a>
Jaro Winkler Distance	<a href="https://github.com/sunlightlabs/jellyfish">https://github.com/sunlightlabs/jellyfish</a>	<a href="https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance">https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance</a>
State Vector Machine (SVM)	<a href="https://github.com/karpathy/svmjs">https://github.com/karpathy/svmjs</a>	<a href="https://en.wikipedia.org/wiki/Support_vector_machine">https://en.wikipedia.org/wiki/Support_vector_machine</a>
K Nearest neighbor (k-NN)	<a href="https://github.com/luispedro/milk">https://github.com/luispedro/milk</a>	<a href="https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm">https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm</a>
Single Value Decomposition (SVD)	<a href="https://github.com/willcannings/Ruby-S">https://github.com/willcannings/Ruby-S</a>	<a href="https://en.wikipedia.org/wiki/Singular_value_decomposition">https://en.wikipedia.org/wiki/Singular_value_decomposition</a>
Latent Dirichlet Allocation	<a href="https://github.com/constadar/farco_lda">https://github.com/constadar/farco_lda</a>	<a href="https://en.wikipedia.org/wiki/Latent_Dirichlet_Allocation">https://en.wikipedia.org/wiki/Latent_Dirichlet_Allocation</a>
Hierarchical Latent Dirichlet Allocation	<a href="https://github.com/josephreisinger/lvm">https://github.com/josephreisinger/lvm</a>	<a href="https://en.wikipedia.org/wiki/Latent_Dirichlet_Allocation">https://en.wikipedia.org/wiki/Latent_Dirichlet_Allocation</a>
Gibbs sampling	<a href="https://github.com/bgamani/bayes-stac">https://github.com/bgamani/bayes-stac</a>	<a href="https://en.wikipedia.org/wiki/Gibbs_sampling">https://en.wikipedia.org/wiki/Gibbs_sampling</a>
Markov Models	<a href="https://github.com/emilmont/Artificial-Intelligence">https://github.com/emilmont/Artificial-Intelligence</a>	<a href="https://en.wikipedia.org/wiki/Markov_models">https://en.wikipedia.org/wiki/Markov_models</a>

# A Multitude of Algorithms for Multidimensional Data

Can we crowdsource algorithms  
and match them to data structures?

timbr.io < Projects

**GNIP Replay - Seed List** Convert to Code Load Last Params

User Name: sean.gorman@timbr.io

Account Name: Timbr

Account Password: \*\*\*\*\*

GNIP-Configuration: prod-darpa

Waiting for preview data...

**Simplify Message**

Many sources stream messages with much interesting components. Choose your source

Message Type: GNP-

Include Dev:

Waiting for preview data...

**Agriculture Classifier**

Predict if a tweet belongs to a category (like

Model File: /twilio-dataroot/agriculture\_model.pkl

Classification Label: Agriculture

Waiting for preview data...

timbr.io < Projects

**Choose a Template**

- Twitter Binary Classification
- Tokenize Text
- Simplify Message
- Retweet Filter
- New Transform
- Language Filter
- Kafka Export
- Kafka Async Export
- Fuzzy Match User to Reference List
- Follower Accrual Rate

# Our current work

timbr.io < Home

Agricultural Communities on Twitter

- data
- agfranco3113.pkl
- agriculture\_model.pkl
- base.py
- influence\_replay.py
- raw\_tag.py
- revert\_replay.py
- seed\_tag.py
- testbook.py
- tw\_analyse.py
- tw\_tag\_analyse.py

**Testing the Feasibility of Agricultural Communities on Twitter as Financial Trading Signals**

Social data and Twitter specifically have been used with success in academia and the private sector to advance financial trading decisions. While Twitter has been used as a data source for stock trading, FOREX and all it was unclear if could be useful for agricultural commodity trading. Specifically are there communities of farmers and the ecosystems that support them on Twitter. If so do they generate information on Twitter that would be useful for financial decision making in regard to agricultural commodities. To test this Timbr.io worked with a customer to identify seed accounts for nine agricultural related categories. We then collected five days worth of data to study the structure of these users' conversations on Twitter.

Number of Tweets

Farmer  
Govt  
IntGroup  
Market  
Media  
Research  
Solutions  
Supply  
Trader

Category	Number of Tweets
Farmer	~100
Govt	~100
IntGroup	~100
Market	~100
Media	~100
Research	~100
Solutions	~100
Supply	~100
Trader	~100

# THANKS

Sean Gorman

sean.gorman@timbr.io

202-321-3914

@seangorman