SocialSim Proposers Day agenda

Start	End	Duration	Item	
7:45 AM	8:45 AM	1:00	**Registration**	
8:45 AM	8:55 AM	0:10	**Security Briefing** DARPA Security	
8:55 AM	9:25 AM	0:30	**Contracts Management Office Briefing** Mr. Mark Jones, DARPA CMO	
9:25 AM	10:05 AM	0:40	**SocialSim Presentation** Dr. Jonathan Pfautz, Program Manager, DARPA I2O	
10:05 AM	10:50 AM	0:45	**Break/Networking Session**	
10:50 AM	11:05 AM	0:15	**Human Subjects Research Guidance** Ms. Lisa Mattocks, DARPA HSR Action Officer	
11:05 AM	11:25 AM	0:20	**DARPA Q&A**	
11:25 AM	11:30 AM	0:05	**Break**	
11:30 AM	5:30 PM	6:00	**Sidebars** One-on-one with DARPA PM, registration required	**Networking Session** (between attendees) Main conference room will remain available
5:30 PM			**Meeting Adjourns**	

Computational Simulation of Online Social Behavior (SocialSim)

Proposers Day Briefing

Dr. Jonathan Pfautz
Program Manager
DARPA I2O

6 February 2017

Use of the online information environment

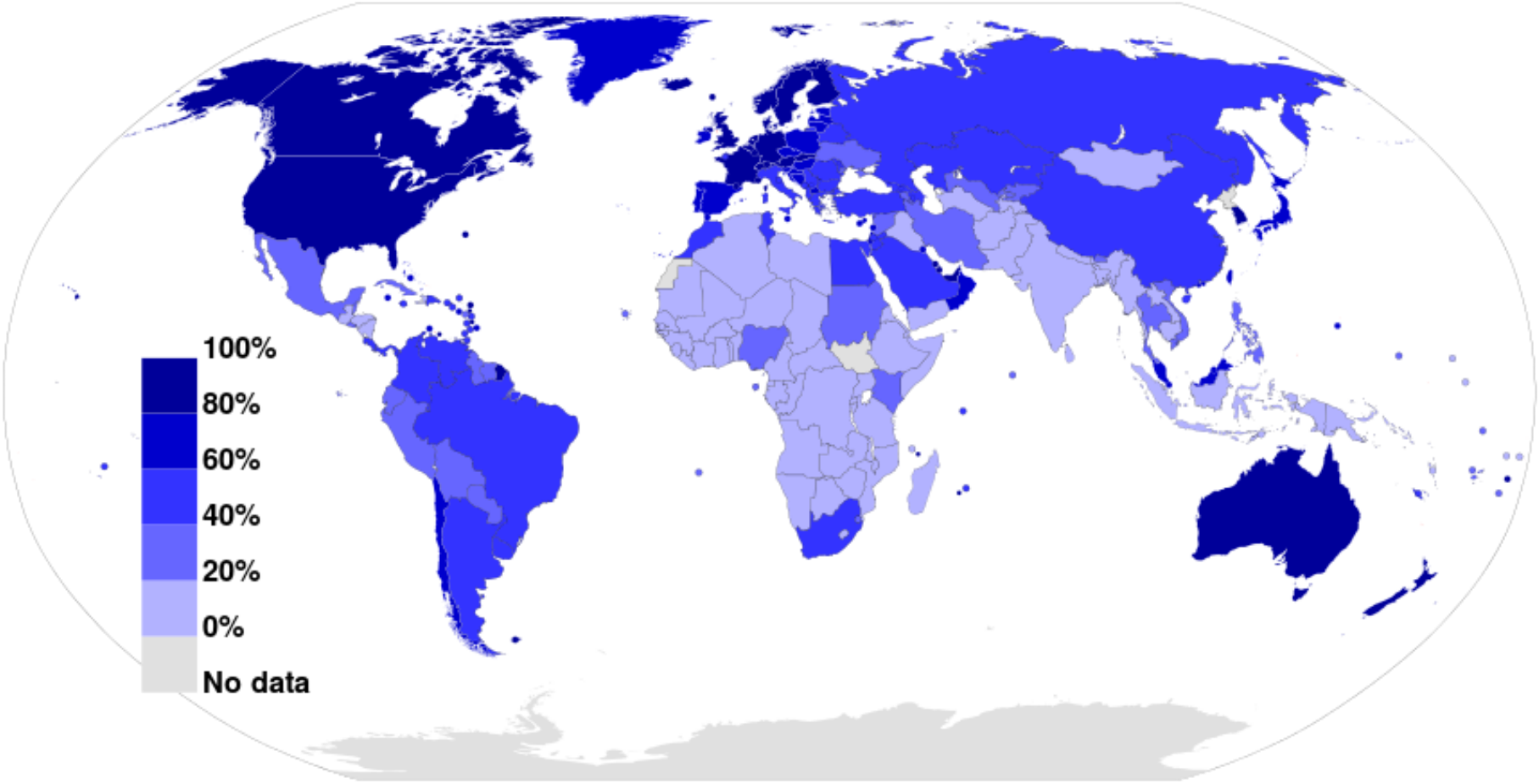
Total human population ~7.4B

(January 2016)

	Users		Δ 2015 to 2016	
Internet	3.4B	46%	+330M	+10%
Active Social Media	2.3B	31%	+219M	+10%
Unique Mobile	3.8B	51%	+141M	+4%
Active Mobile Social	2.0B	27%	+283M	+17%

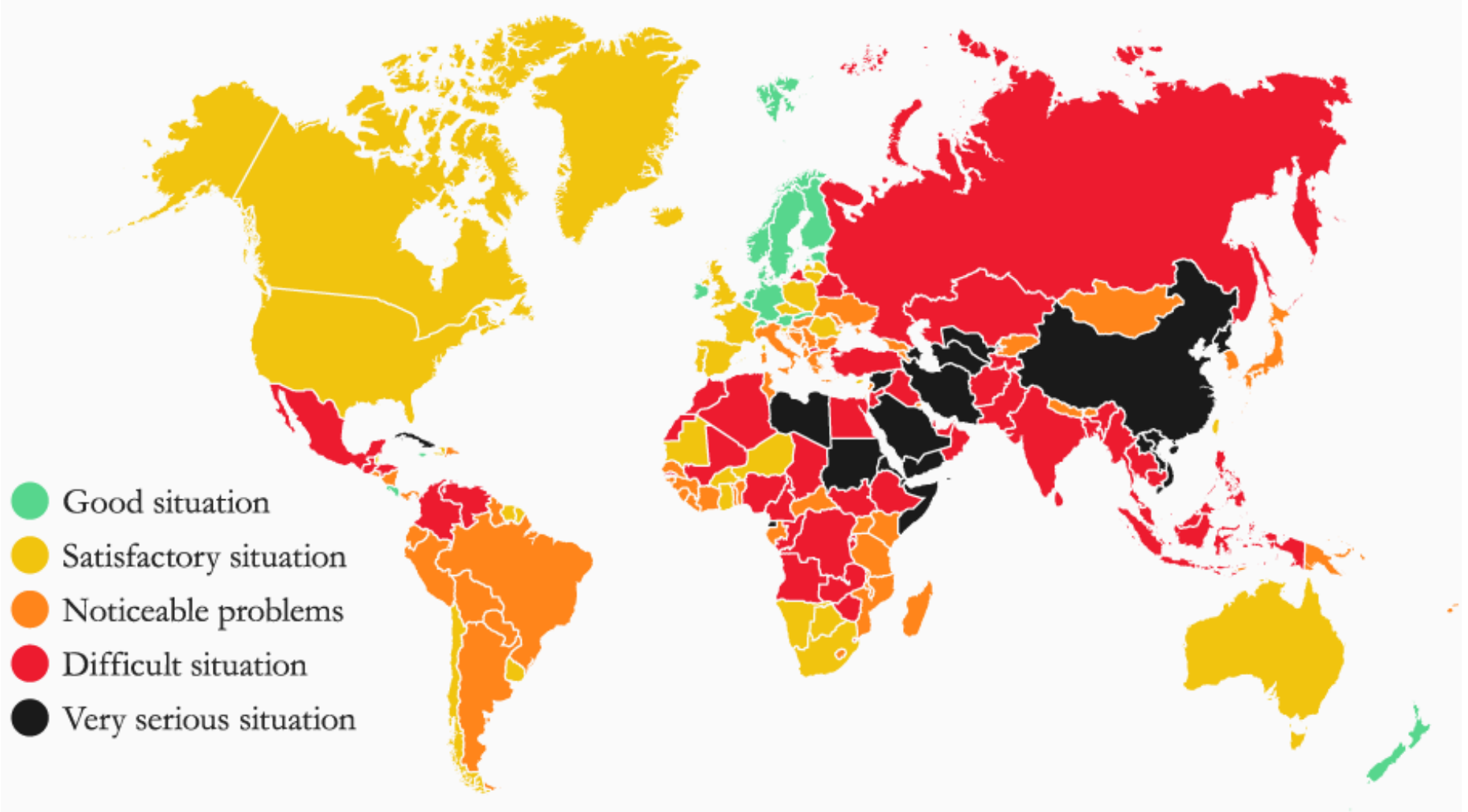
(Chaffey 2016)

Internet users as a percentage of population



https://en.m.wikipedia.org/wiki/Global_Internet_usage#/media/File%3AInternetPenetrationWorldMap.svg (2012)

Global freedom of press



Reports Without Borders – 2016 World Press Freedom Index

Example 1: Systematic intervention in the online spread of information

China

- ~550M internet users
- >1,400 social environments
- Government activities:
  - Traditional media control
  - "The Great Firewall"
  - Automated and manual censorship

(King, Pan, & Roberts 2017)

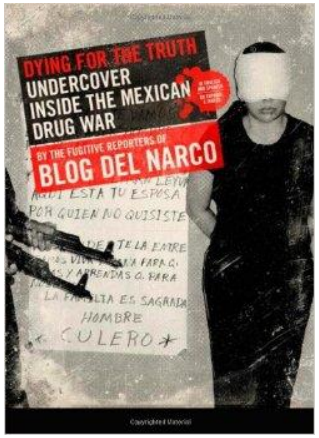
Mexico

The Washington Post

Investigations

Censor or die: The death of Mexican news in the age of drug cartels

By Dana Priest December 11, 2015



Coordinated disruption of online political discourse:

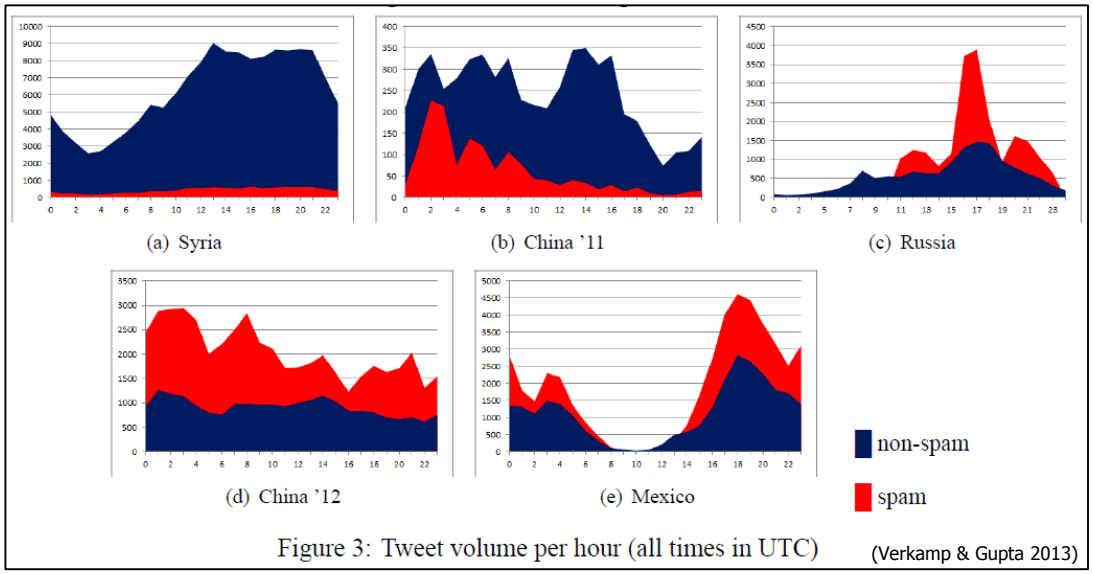


Figure 3: Tweet volume per hour (all times in UTC)

(Verkamp & Gupta 2013)

How do these actions change the spread of online information?

Example 2: The online spread of information during natural disasters

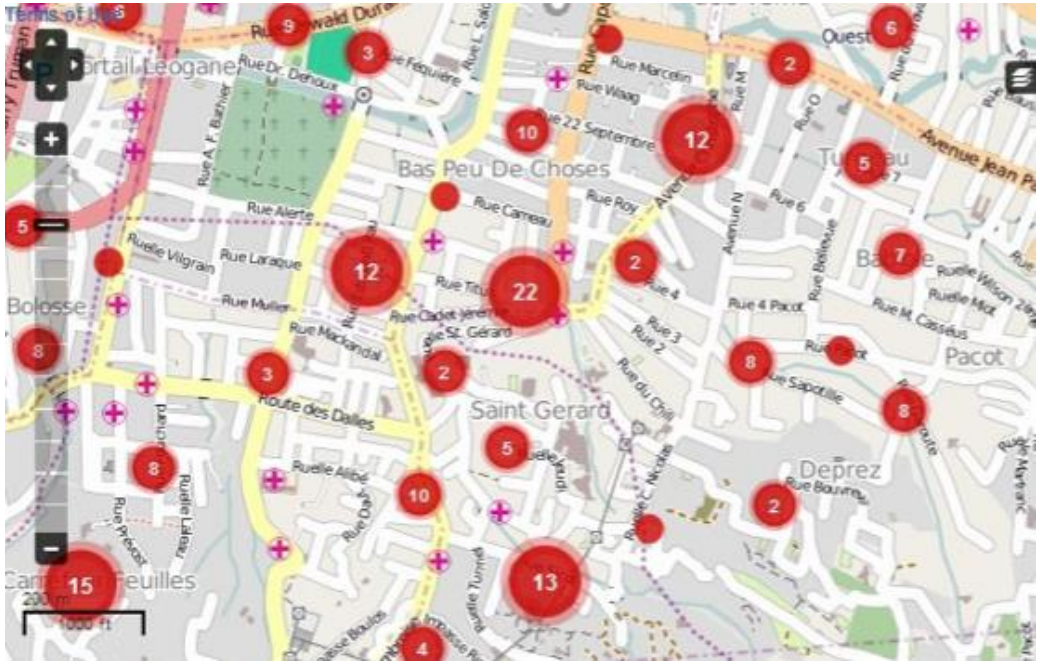
How Twitter's Hashtag Came to Be

By ELANA ZAK
Oct 3, 2013 5:54 pm ET

THE WALL STREET JOURNAL.

"The use of hashtags became mainstream after October 2007, when citizen journalists used them to give updates about a series of forest fires in San Diego."

Coordinated online reporting during Haiti earthquake:

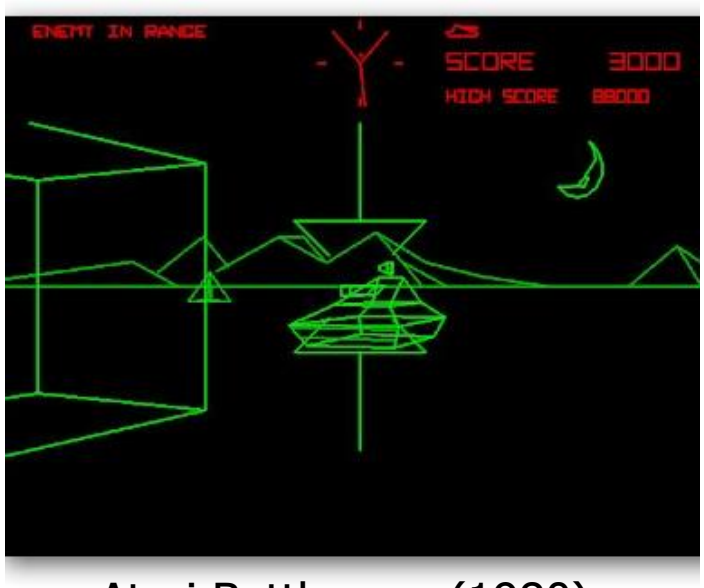


Close up of the Ushahidi Haiti Map. Credit: Ushahidi

See also: Japanese tsunami; Queensland, Victoria floods; Chilean earthquake; Philippines typhoon

How do these actions change the spread of online information?

Benefits of simulation



Atari Battlezone (1980)

DARPA SIMNET (1983)

- Training
- Wargaming
- Forecasting

- Explore possibilities
- Understand space of potential outcomes
- Predict outcomes
- Anticipate others' actions
- Develop new courses of action (COAs)
- Select optimal COAs

… without the potential consequences of operating in the real world

SocialSim program objective

Develop technologies for high-fidelity simulation of online social behavior (the spread and evolution of online information) while rigorously testing and measuring simulation accuracy

Accuracy of simulation

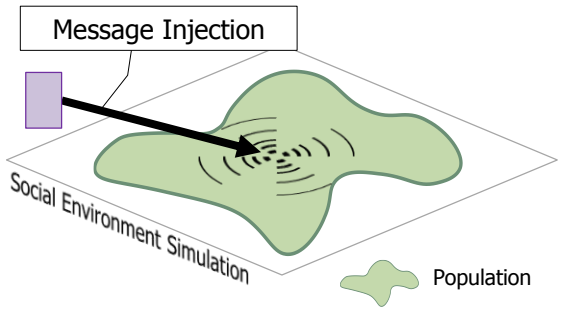
\+ *Scalability* to populations of interest

\+ *Rigor* of testing and measurement

High-fidelity simulation

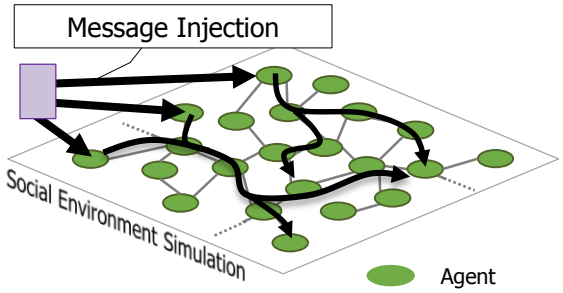
Current computational approaches to social and behavioral simulation

Top-down simulation



- Focus on dynamics of entire population(s)
  E.g., a population's affinity for a group grows over time
- Populations act uniformly → Can limit accuracy
  E.g., information spreads like epidemics
- Larger populations → No effect on run time

Bottom-up simulation



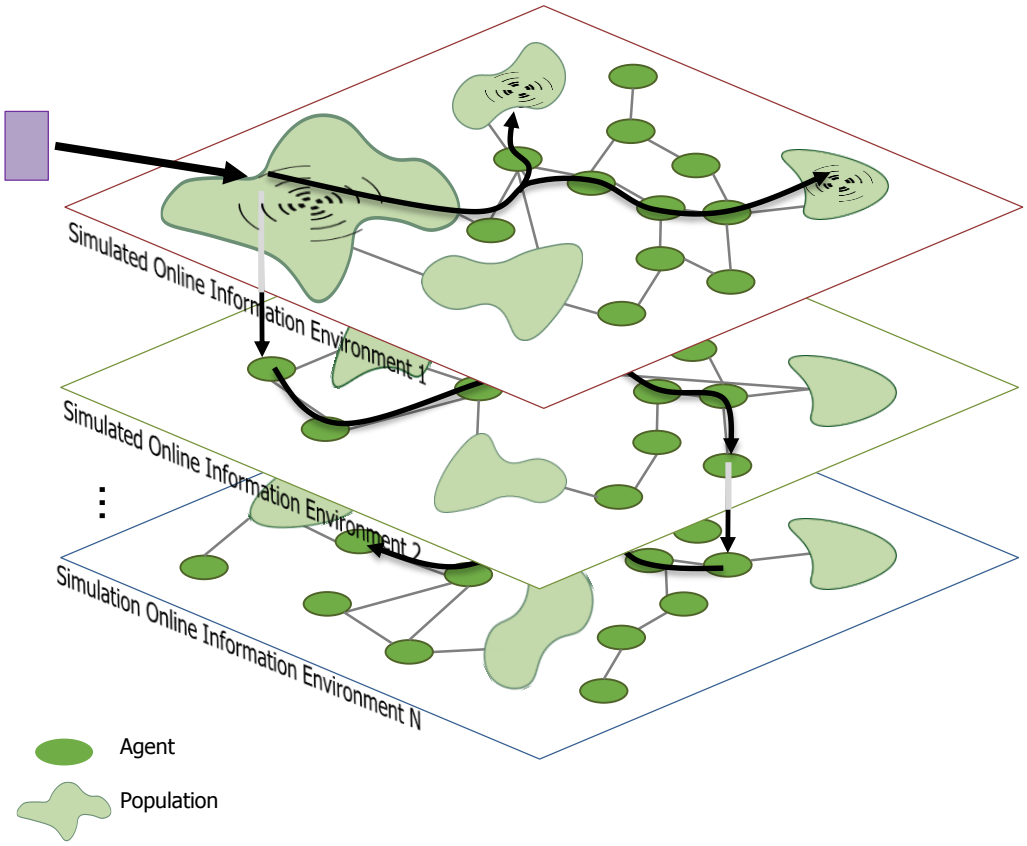
- Focus on diverse activities and interactions
  E.g., the number of people who belong to a group, then tell others they like it, grows over time
- Agents act differently → Can improve accuracy
  E.g., discrete influencers change how a meme spreads
- Larger populations → Longer run time

Many dimensions of computational simulation must be considered to achieve accuracy at scale

E.g., Rahmandad & Sterman, 2008; Lymperopoulos & Ioannou, 2016

SocialSim vision: Multi-resolution simulation

Creatively combining and/or extending existing simulation approaches will lead to breakthroughs that radically increase accuracy and scalability



What approaches are more/less effective under which conditions?

What population, environment, and content factors must be represented? Why?

Which approaches are best for accurately representing which factors?

Multi-resolution simulation in other domains

Work in other domains suggests combining and/or extending existing approaches to achieve multi-resolution simulation can improve accuracy and speed

Domain	Predicted variable	Improvement in **accuracy**	
Urban Planning [1]	Population size	**14x**	14% to 1% error
	Population growth	**40x**	455% to 12% error
Vehicle Traffic [2]	Flow volume	**2x**	0.147 to 0.075 (using Thiel U statistic)

Domain	Simulated dynamics	Improvement in **speed**	
Urban Planning [3]	Population density	**3x**	19,356 sec to 5,964 sec
Vehicle-Based Networks [4]	Road and net traffic	**15x**	180 sec to 12 sec
Computer Networks [5,6]	Packet flow	**400x**	5 sec to 0.0026 sec

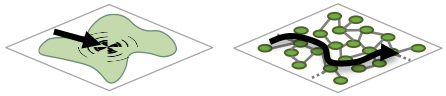
Greater speed enables simulation of larger scale populations, and therefore can increase fidelity

[1] White 2007; [2] Burghout 2006; [3] Torrens 2012; [4] Schiller 2015; [5] Nicol 2003; [6] Yan 2005

Program phases and technical areas

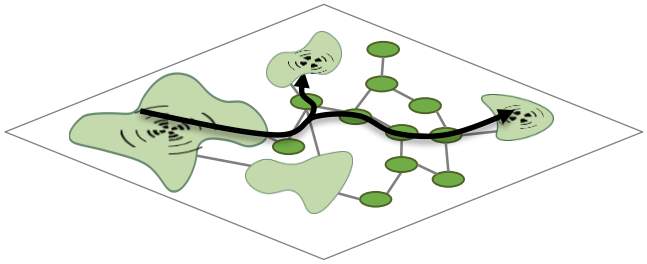
Phase 1 | **Phase 2**

TA1: Simulation

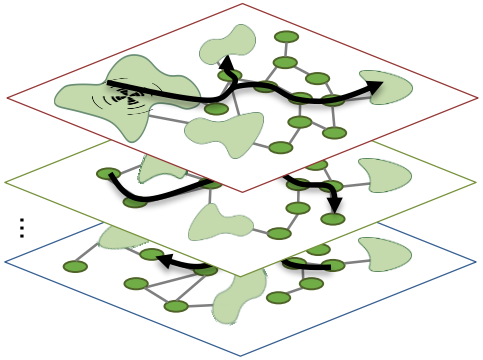


Enhance existing simulation approaches

Existing simulation approaches



Simulate dynamics of a single online information environment



Simulate within and across multiple online information environments

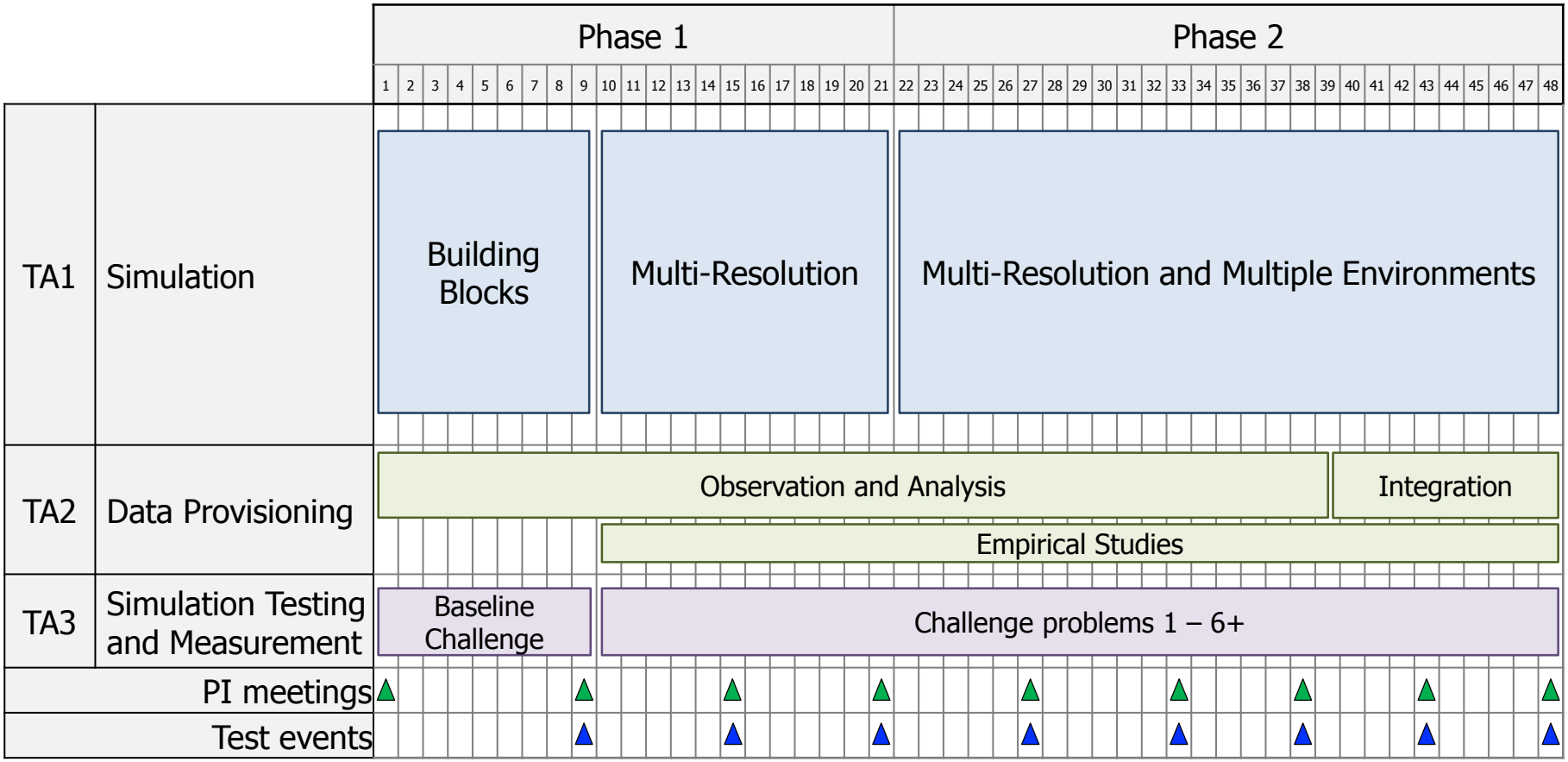
TA2: Data Provisioning

- Gather data to support simulation development and testing
- Develop efficient and robust methods for gathering data

TA3: Simulation Testing and Measurement

- Establish measures of simulation fidelity
- Issue challenge problems to measure program progress

SocialSim schedule and milestones



Challenge problems

Challenge problems drive progress in the SocialSim program

Accurate = Simulation results replicate real world data

Accuracy is a function of metrics

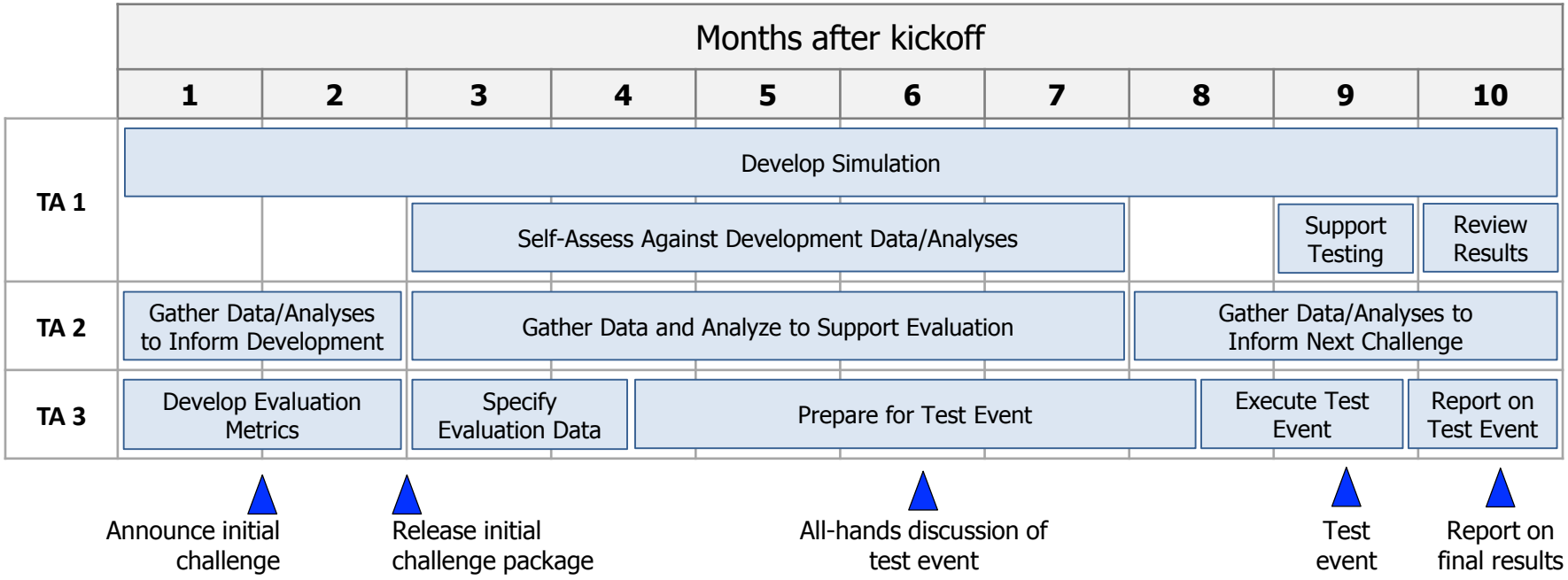
- E.g., Simulation of the likelihood and size of a recurrent cascade may be accurate (70-80%), yet the simulated time between bursts may be less accurate (<60%) (Cheng 2016)
- A key objective of SocialSim is to develop relevant and robust metrics

	Example 1	Example 2
Online behavior to simulate	• Cascades of political messages	• Non-traditional gatekeepers emerge and influence spread rates during natural disasters
Data sources	• Twitter historical data analytic results	• Anonymized email traffic data from researchers • Large-scale, empirical studies about trusted sources of online information
Metrics	• Number of tweets of relevant hashtags over time • Number of unique users that tweet relevant hashtags over time	• Number of gatekeepers that emerge • Rate and reach of aid-related message spread in population

SocialSim baseline challenge problem

The baseline challenge problem:

- Establishes initial simulation accuracy and scale across bottom-up and top-down behaviors
- Defines an initial set of metrics
- Establishes processes for subsequent challenges



Progression of challenge problems

- Initial challenge will focus on previously studied online behavioral phenomena, e.g.,

Phenomenon	Description
Cascades	Accelerated resharing of information[1-3]
Recurrence	Renewed bursts of activity for existing information[4-6]
Gatekeepers	Key influencers change how information spreads[7-9]
Persistent Minorities	Small, committed groups can change rate of spread[10-12]

- Subsequent challenge problems *will* target:
  - Improvement in accuracy and scale
- Subsequent challenge problems *may* target:
  - Additional metrics
  - Additional populations
  - Additional types of information (content, form)
  - Additional behavioral phenomena (e.g., evolution)
  - Different information environments

Proposers across all TAs are encouraged to discuss existing data gathering capabilities, data sets, metrics, and/or potential challenge problems

[1] Cheng 2014; [2] Cha 2008; [3] Adar 2005; [4] Cheng 2016; [5] Yu 2015; [6] Kwon 2013; [7] Meraz 2013; [8] Bakshy 2011; [9] Gruhl 2004; [10] Bastos 2013; [11] Wang 2012; [12] Thomas 2012

SocialSim privacy plan

Online phenomena such as information spread and evolution arise as a result of social interactions within a population

To develop simulations, DARPA anticipates that performers may:

- Gather publicly available information (PAI)
- Encounter personally identifying information (PII)

SocialSim will create and maintain a Privacy Plan and all performers will be active participants and supporters

SocialSim Privacy Plan goals:

1. Respect and protect individual privacy
2. Gather and analyze data only as needed, and in a legal, ethical, and societally responsible manner

SocialSim privacy plan

Expected procedures:

- All performer personnel will undergo program-specific privacy training
- Only data provided by, or reviewed and approved by, DARPA will be used
- Technical means will be applied to eliminate as much PII as possible
- Sensitive data will be:
  - Partitioned and with restricted access
  - Retained only for strictly limited periods of time

All proposals must address:

- Plans for maintaining compliance
- Specific technical means to protect privacy

Any proposals that aim to target specific individuals or reconstitute identifying behaviors will be considered out of scope

Human subjects research (HSR) in SocialSim

Proposals must describe any activities that must adhere to Human Subjects Research (HSR) regulations

All HSR activities will be subject to independent review by both:

- Appropriate Institutional Review Boards (IRBs)
- Department of Defense Headquarters

TA1: Simulation

Goal: Develop technologies that can accurately simulate online information spread and evolution at scales representing populations of interest (i.e., thousands to tens of millions)

Technical Requirements:

- Determine properties and behaviors of populations to represent information spread and evolution
- Determine properties of information environments to represent
- Determine properties of information to represent (e.g., form, content)
- Capture interactions between the population, the information, and the online environment
- Develop innovative computational modeling methods
- Support rigorous testing and measurement of simulation accuracy and scale
- Demonstrate significant improvements in accuracy and scale over the course of the program

TA1 highlights

- Approaches *must be innovative*
- Approaches must be amenable to testing and measurement
- Proposals must provide a clear and compelling explanation of:
  - Underlying theory and/or data and analyses used to identify and select properties of populations, environments, and information – and to define their interaction
  - Translation of theory and/or data into computational simulation
- Approaches that only rely on commercial, off-the-shelf hardware (vs. high-performance computing) are preferred

TA2: Data Provisioning

Goal: Develop efficient and robust methods for providing data to support simulation development, testing, and measurement

Technical Requirements:

- Gather data and perform analysis of online information spread and evolution
- Provide program-wide, managed access to data, analyses, and/or analytic tools
- Gather and analyze data on relevant population properties, information properties, and information environment properties
  - At multiple levels of resolution
- Apply methods to improve veracity and quality of observational data
- Structure data sets, align data, characterize data provenance and uncertainty
- Provide analytic techniques to richly describe the dynamics of information spread and evolution
- Apply technical means to remove PII, protect sensitive data, and preserve privacy
- Accommodate rapid shifts in focus to support different challenge problems

TA2 highlights

- SocialSim is not a "big data" program - extreme data management is out of scope

- SocialSim is not seeking to develop new text/speech or image/video analysis methods
  - Leveraging existing techniques for new purposes is in scope

- Approaches of interest could include none, one, or more of:
  - Traditional data science
  - Empirical studies
  - Development of new methods for obtaining observational and/or empirical data

- Proposals must provide a clear and compelling explanation of:
  - Extensibility to new challenge problems
  - Flexibility to changes in the global information environment
  - Privacy protection measures

- Dedicated physical locations or remotely accessible compute environments for test events are of interest, but not required

TA3: Simulation Testing and Measurement

Goal: Develop rigorous methods and metrics for quantitatively assessing the accuracy and scalability of simulations of online information spread and evolution

Technical Requirements:

- Develop challenge problems and evaluation metrics to assess program progress
- Develop a baseline challenge problem to assess initial accuracy and scale
- Identify a range of behaviors that are fundamental to online information spread
- Develop multiple measures of online behaviors and evaluate simulation accuracy
- Develop additional follow-on challenge problems that build on the baseline challenge problem
- Organize and conduct test events, execute simulations, assess against gold-standard data
- Deliver formal evaluation reports on test events

TA3 highlights

- Proposals must demonstrate a deep understanding of:
  - Current scientific understanding of online social behavior
  - Validation, testing, and measurement of models and simulations
  - Relevant national security challenges
  - Privacy protection requirements

- Proposals must provide a clear and compelling explanation of potential challenge problems, progressions of challenges, and their feasibility
  - Opportunities to accelerate the schedule of challenge problems are of interest

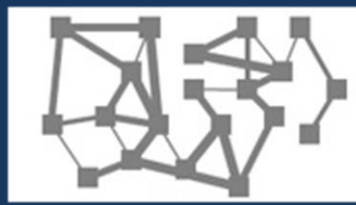
- Solely advancing the state of the art in modeling and simulation toolkits, integration platforms, or APIs is out of scope

- Dedicated physical locations or remotely accessible compute environments for test events are of interest, but not required

Collaboration across technical areas

- Describe data required for simulation development
- Share existing data sets

TA1: Simulation



Support development of:
- Metrics
- Challenge problems

- Provide data, analyses, and analytic tools
- Identify and describe available data

TA2: Data Provisioning

- Issue challenge problems
- Run test events

TA3: Testing and Measurement

General proposal guidance

- Read the BAA carefully
- Highlight innovative ideas
- Give a clear rationale and compelling evidence for all aspects of your proposal
- Provide an appropriate level of technical detail
  - Assume your reader has a Ph.D. in a field other than your own
- Demonstrate an understanding of the requirements of all technical areas
- Clearly and logically separate tasks and their associated cost
- Clarify intellectual property rights for software and data

- Abstracts are not required, but are highly encouraged
  - Recommendations will be given as quickly as possible
  - Teaming need not be complete, but required skill sets and roles should be identified

Anticipated funding

Expected available program funding: $42M

Distribution of funds across TAs depends on quality of submissions

Bidding to multiple TAs:

- Address only one TA per submission
- TA3 performer not eligible to perform under TA1

Closing comments and pointers

After today, questions about the program should be directed to:

SocialSim@darpa.mil

The SocialSim website may be updated with additional materials

https://www.schafertmd.com/darpa/i2o/socialsim/pd/

www.darpa.mil