The Big Data Index

Bringing Data & Communities Together
for the Greater Good

THE ASPEN INSTITUTE

The Aspen Institute initiated an important dialogue about big data.

The Aspen Institute's Socrates Seminar on big data brought together leaders from a variety of sectors and industries to discuss how the information revolution can provide open data to empower individuals, transform industries, and drive economic growth. The group debated big data and data interoperability challenges at the personal, federal, and commercial levels. Together, seminar participants envisioned a platform called the Big Data Index to address these issues.

Thicket Labs documented the conversation about the Big Data Index and designed a mental model that synthesizes the collective vision. This report summarizes our findings about why an intervention like the Big Data Index is needed to improve the current ecosystem of big data, how such a platform could help, and offers recommendations for the future.

Table of Contents

A Complex Ecosystem

The Socrates Seminar on Big Data revealed that we find the acceleration of data collection and sharing practices in turns terrifying and exhilarating. Some believe that the notion of personal privacy is well on its way to becoming an anachronism. Others are fascinated by the potential of the quantified self movement. All participants agree: The big data ecosystem is plagued by serious systemic issues.

Many people and organizations are collecting vast amounts of data with little to no regulation, oversight, or accountability. Those of us at the mercy of this system (including many who are collecting big datasets themselves) are responding to these issues with a fundamental lack of trust and resistence to participation. Big data is not on the path to success — not without an intervention of sorts.

We agreed that big data is driving new advances across the sciences that are essential for our future. But legitimate concerns about personal privacy and security need to be addressed alongside new advances. So how do we start to address those concerns without clamping down on big data's potential?

Seminar moderators Peter Levin and Hari Sreenivasan challenged participants to think beyond existing concerns to design a solution to help strengthen the ecosystem of big data.

Vision for Impact
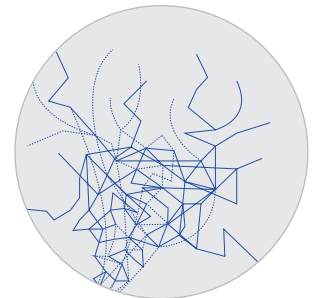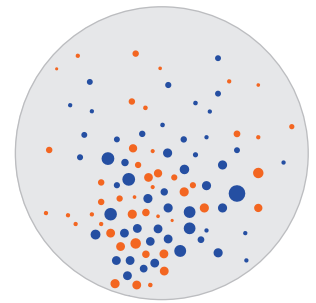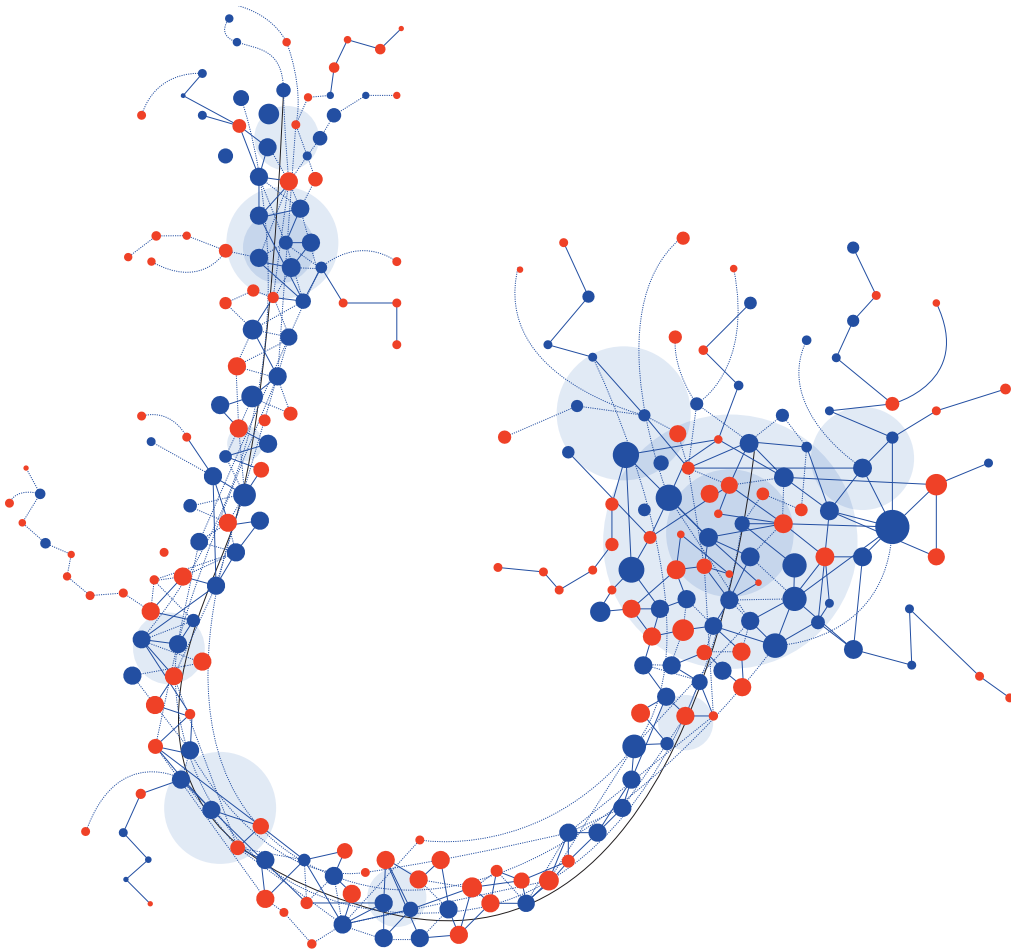
The Big Data Index is a platform to organize, share, and set standards around big data.

Users range from academic institutions, governments, companies, and nonprofits to independent researchers, amateur scientists, and other members of the public.

The Big Data Index is designed as an intervention to help mitigate systemic issues of unregulated data collection and application and fragmented or siloed knowledge that limit our current big data practices.

Privacy Considerations

Personal privacy was a topic of contention throughout the seminar. The rise of big data is challenging the notion of privacy, which has long been defined as allowing individuals and groups to keep information that relates to their personal identity out of the public realm.

More and more, data is becoming impossible to keep private in an era when any number of companies, agencies, and organizations collect data about individuals, but individuals have no control over who can access their data or to what end. Data is collected by numerous parties with little consistency and often labyrinthine terms and conditions that leave individuals in the dark about policies governing the sale and use of their data.

As data collection efforts increase and data sets grow larger, anonymity is on the decline. Large companies are a particular target for security breaches leading to data theft. Other types of privacy breaches are also emerging as networked data reduces the possibility of anonymity.

If we accept that most data is going to be publicly available in the future, a more useful priority might be to develop standards around collecting and sharing data that pave the way for individuals to better understand the use of their data and have channels to challenge that use.

For this reason, we deliberately kept the concept of privacy out of the impact metrics for the Big Data Index and instead focused on fostering values of open exchange and dialogue.

Process

Thicket Labs used cognitive mapping to document input from workshop participants in the form of concepts. We took the 109 concepts mapped in the workshop and distilled them into 23 system elements representing the Big Data Index and the big data ecosystem. We used The Possibility Engine, our custom cognitive mapping platform, to synthesize these concepts into a set of related elements that form a mental model representative of the group.

To form the model, we started by identifying key impact metrics that help us gauge system growth, and related them to influential system elements. The four categories of elements are actors within the system, actions they take, conditions that allow us to simulate how the system operates through times, and factors that influence systemic growth.

We then used The Possibility Engine to design a mathematical model that integrates these concepts through fuzzy logic and visualizes possible outcomes. We deployed these system elements in three simulations to demonstrate how the Big Data Index could operate within the larger ecosystem of big data to achieve impact.

This mental model was based on input from a multi-disciplinary group of seminar participants. The Big Data Index does not yet exist, nor is the vision comprehensive. We simplified actors, actions, and other system elements in the spirit of clarity.
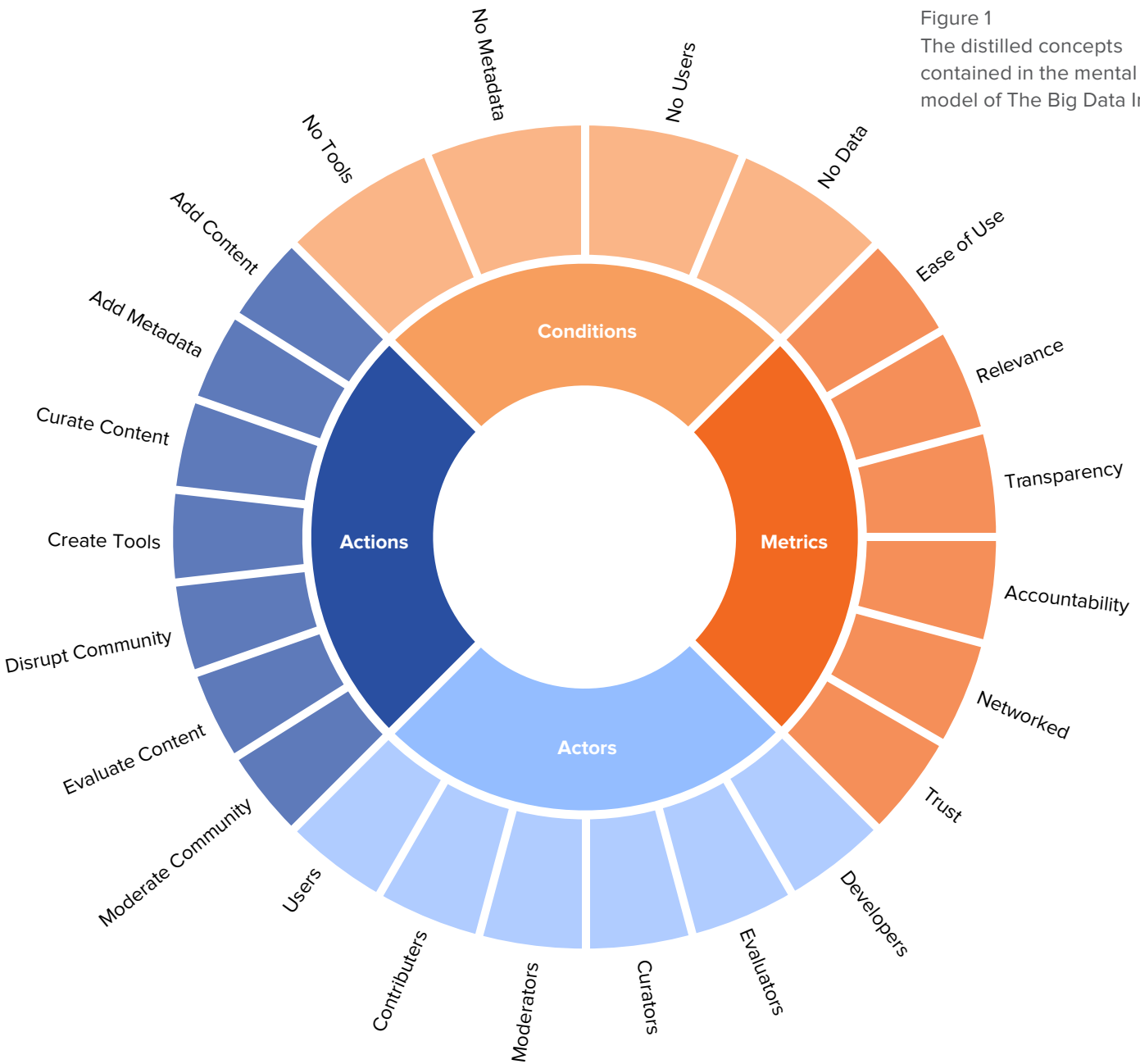
Concepts

Figure 1
The distilled concepts contained in the mental model of The Big Data Index.

These impact metrics, system actors, actions, and conditions allow us to model how the platform addresses current challenges, how it creates systemic impact, and how it performs under specific conditions.

Metrics

TRUST
Trusted data sets provided by trustworthy users stand up to peer scrutiny and encourage adoption of the Index.

NETWORKED
A networked system promotes free exchange, in contrast to a fragmented system, which hinders activity.

ACCOUNTABILITY
Holding users account-able to values and standards increases the relevance and utility of contributions, encouraging a wider community base.

TRANSPARENCY
Transparency of pro-cess and information increases trust and reliability, which promotes Index adoption.

RELEVANCE
The relevance of data sets discovered on the Index plays a major role in promoting use and adoption over time.

EASE OF USE
Making the Index easy to use is another critical factor for encouraging adoption and retention over time.

Actors

USERS
Anyone who draws on data sets from the Index.

CONTRIBUTORS
Anyone who posts data sets in the Index for use by the community.

MODERATORS
Members who make sure community inter-actions within the Index falls within agreed-upon standards and values.

CURATORS
Members who relate a collection of different data sets to a topic to promote clarity.

EVALUATORS
Members who vet data sets for authenticity, verifiability, and relevance.

DEVELOPERS
Members who create tools, add-ons, or plug-ins to add functionality and ease of use to the Index for users.

Actions

ADD CONTENT
The act of contributing data sets to the Index.

ADD METADATA
The act of adding additional information to data sets that improves identification, organization, formatting, and sharing of data sets.

CREATE TOOLS
The act of creating tools, add-ons, and plug-ins that make the Index more accessible, functional, more convenient for the community.

CURATE CONTENT
The act of filtering and relating different data sets for ease of use by the community.

DISRUPT COMMUNITY
The act of contributing to the Index in a way that promotes discord within the community.

EVALUATE CONTENT
The act of vetting data-sets for authenticity, verifiability, and relevance.

MODERATE
The act of holding users to standards set by the Index community and resolving disagreements.

Conditions

NO DATA
No data in the Index mimics conditions at Index launch.

NO USERS
No users in the Index mimics conditions at Index launch.

NO METADATA
No metadata in the Index mimics conditions at Index launch.

NO TOOLS
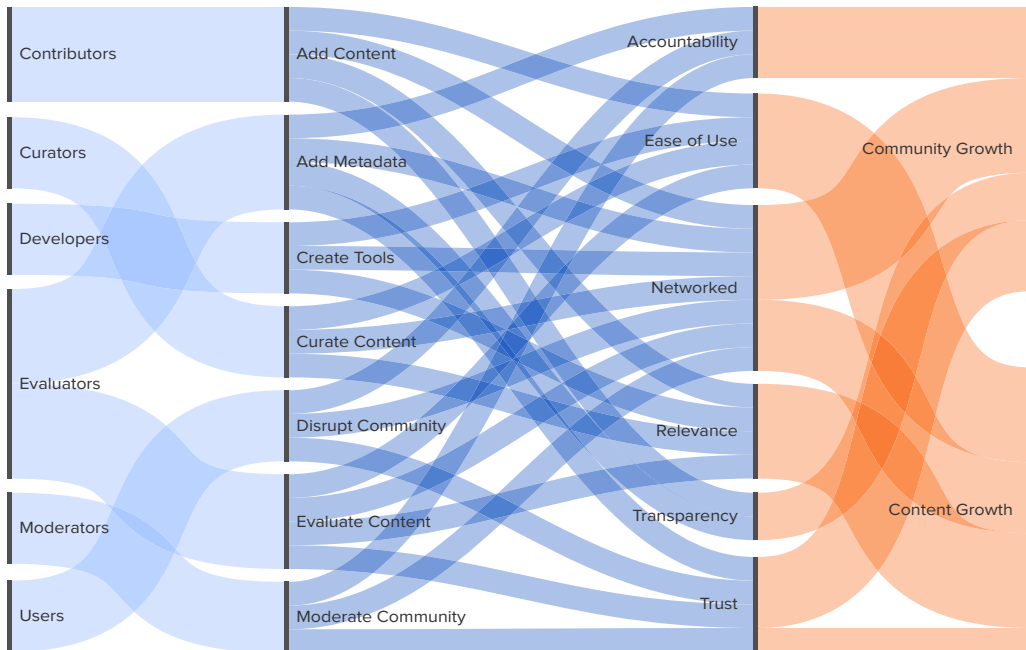No tools in the Index mimics conditions at Index launch.

Simulations

Figure 2.1
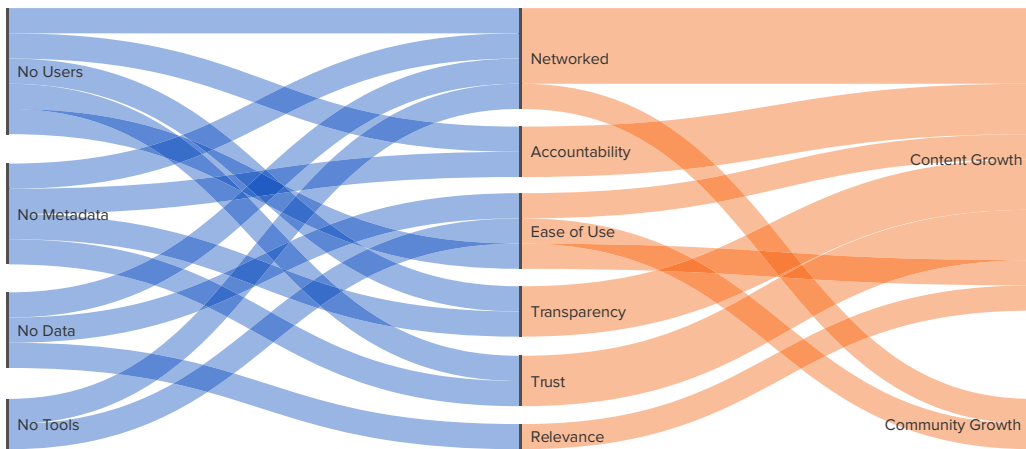How actors and actions influence metrics and growth concepts in the model.

Figure 2.2
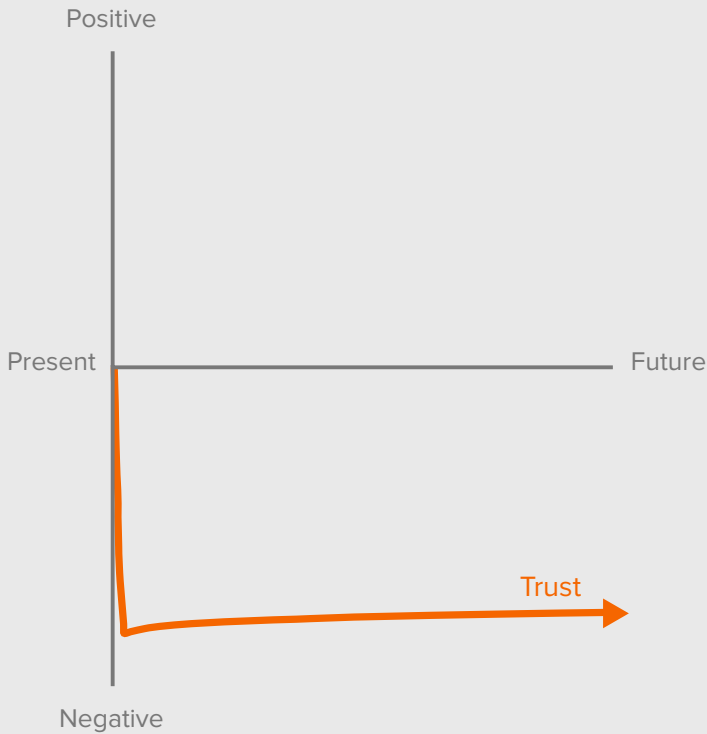How conditions influence metrics and growth concepts in the model.

We modeled the system under varying conditions to demonstrate the value of the Big Data Index and several potential risk factors.

Simulation A examines how the current lack of regulation and oversight results in a fragmented ecosystem over time.
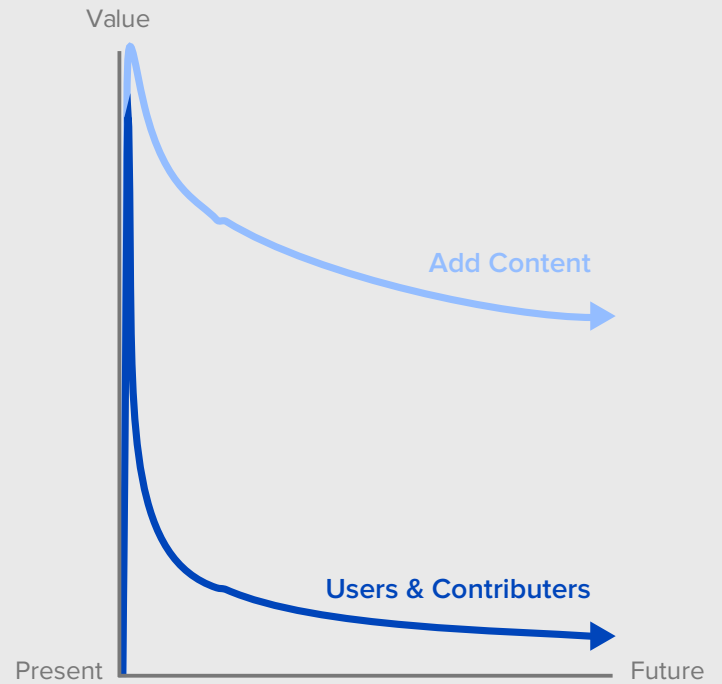
Simulation B models how the Big Data Index intervention could impact this ecosystem over time.

Simulation C zooms in on the launch of the Big Data Index to extract some potential risk factors.

Big Data Ecosystem Impacts Over Time

Positive

Present — Future

Trust

Negative

Big Data Ecosystem Dynamics Over Time

Value

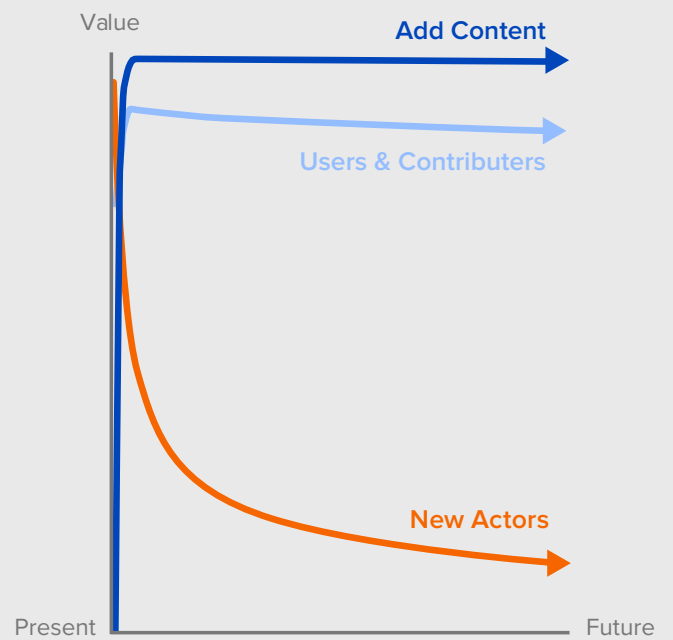Add Content

Users & Contributers

Present — Future

Figure 3.1 (left)
Dynamics of trust in the current big data ecosystem.

Figure 3.2 (right)
Dynamics of content sharing and participation in the current big data ecosystem.

SIMULATION A

By populating our system with only users and contributors freely collecting and sharing data sets, we see that a wholly unregulated system encourages a lack of trust, which in turn decreases participation and content sharing over time.

Trust plummets precipitously before recovering to a low equilibrium. Similarly, participation may start high but quickly drops off as contributors and users become highly fragmented. Data collection and sharing decreases over time but not as dramatically, indicating a highly fractured environment where many data sets are being collected but there is little conversation and exchange.

Big Data Index
Impacts Over Time
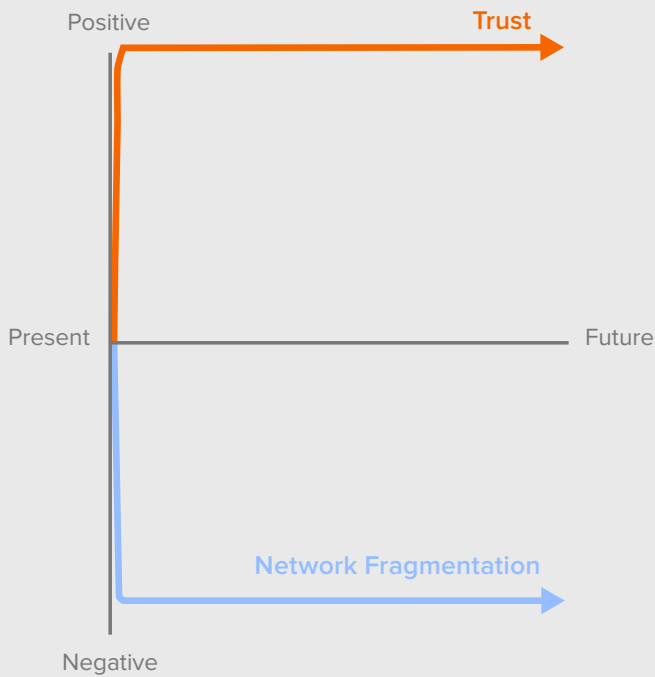
Big Data Index
Dynamics Over Time

Figure 3.3 (left)
Dynamics of trust and network fragmentation after the Big Data Index is introduced.

Figure 3.4 (right)
Dynamics of content sharing and participation after the Big Data Index is introduced.
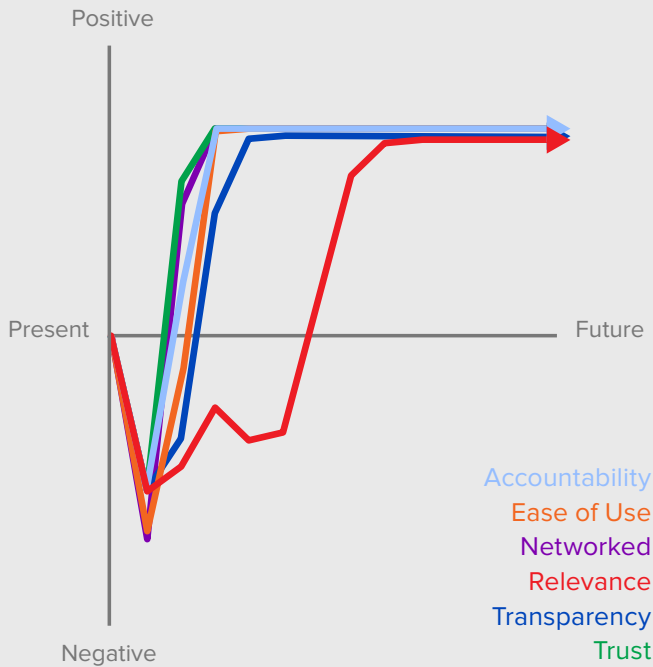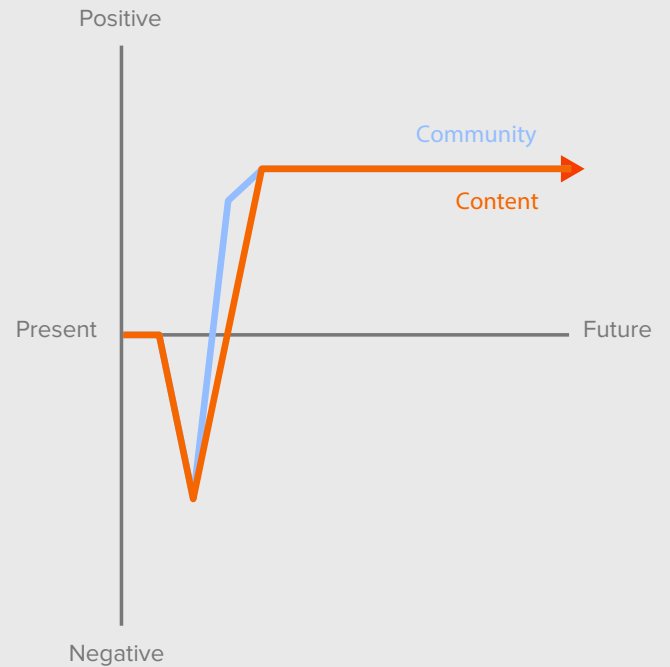
SIMULATION B

The Big Data Index introduces four new actors and their corresponding activities into the environment: moderators, evaluators, curators, and developers. Their actions promote a healthier environment for the controlled exchange of data and information.

We see that the introduction of standards, policies, and tools improve the sharing experience and promote adoption, which results in a much healthier environment for the big data community. Trust, data sharing, and participation all skyrocket before levelling off at a very high level. Network fragmentation decreases, improving our ability to share data sets and new discoveries with increased relevance and efficiency.

Impact Metrics
At Launch

Positive

Present — Future

Accountability
Ease of Use
Networked
Relevance
Transparency
Trust

Negative

Growth Rate
At Launch

Positive

Community

Content

Present — Future

Negative

Figure 3.5 (left)
Dynamics of all impact metrics as impacted by the Big Data Index under launch conditions.

Figure 3.6 (right)
Dynamics of content and community growth as impacted by the Big Data Index under launch conditions.

SIMULATION C

Interventions are particularly impactful to a system due to the existing fragility of the system. Here, we set the conditions to no data, users, tools, or metadata in order to simulate launch conditions. We see all of our impact metrics nosedive before they recover and accelerate to positive impact.

The early days of a tool like the Big Data Index are critical to its long-term success. The role of developers creating new tools to make it easier for new community members to join and share data is particularly critical to a successful launch effort. Moderation and evaluation efforts should be gentle as new standards are developed and introduced so as not to discourage participation before a critical mass of members has been achieved.

Conclusions

The vision contained in this mental model and resulting simulations was developed through vigorous debate and exchange among a group of individuals who deal in data every day. These participants are not only data professionals, but, just as critically, members of a public who are at a strong disadvantage in the current big data environment.

The power asymmetries affecting this system are evident in our simulation outcomes. The potential for an intervention like The Big Data Index is clear; and we hope this report will sharpen further efforts to promote a healthier environment for our data collection and sharing practices.

Thicket Labs is very pleased to have been able to take part in and document this conversation about big data and its impacts on society. We developed the mental model of the Big Data Index as a tool to communicate a new idea to a wider audience and catalyze a community around it. If the Big Data Index moves forward, we will enlarge this mental model with further input from diverse stakeholders.

With a critical mass of stakeholder engagement and support, we can transform this mental model into an empirically driven system model. Informed by observable, real-world data, a system model can facilitate the design of the Big Data Index and provide a framework for measuring system impact over time.