Defining Responsible Data Practices - A Community-driven Approach

NATALIE EVANS HARRIS

COO & VP of Ecosystem Development, BrightHive

Former Senior Policy Advisor to US Chief Technology Officer, Obama Administration

Responsible Use of Data

16 years of federal government experience

The Data Cabinet
www.ntis.gov/thedatacabinet

We are a people in a quandary about the present. We are a people in search of our future. We are a people in search of a national community.

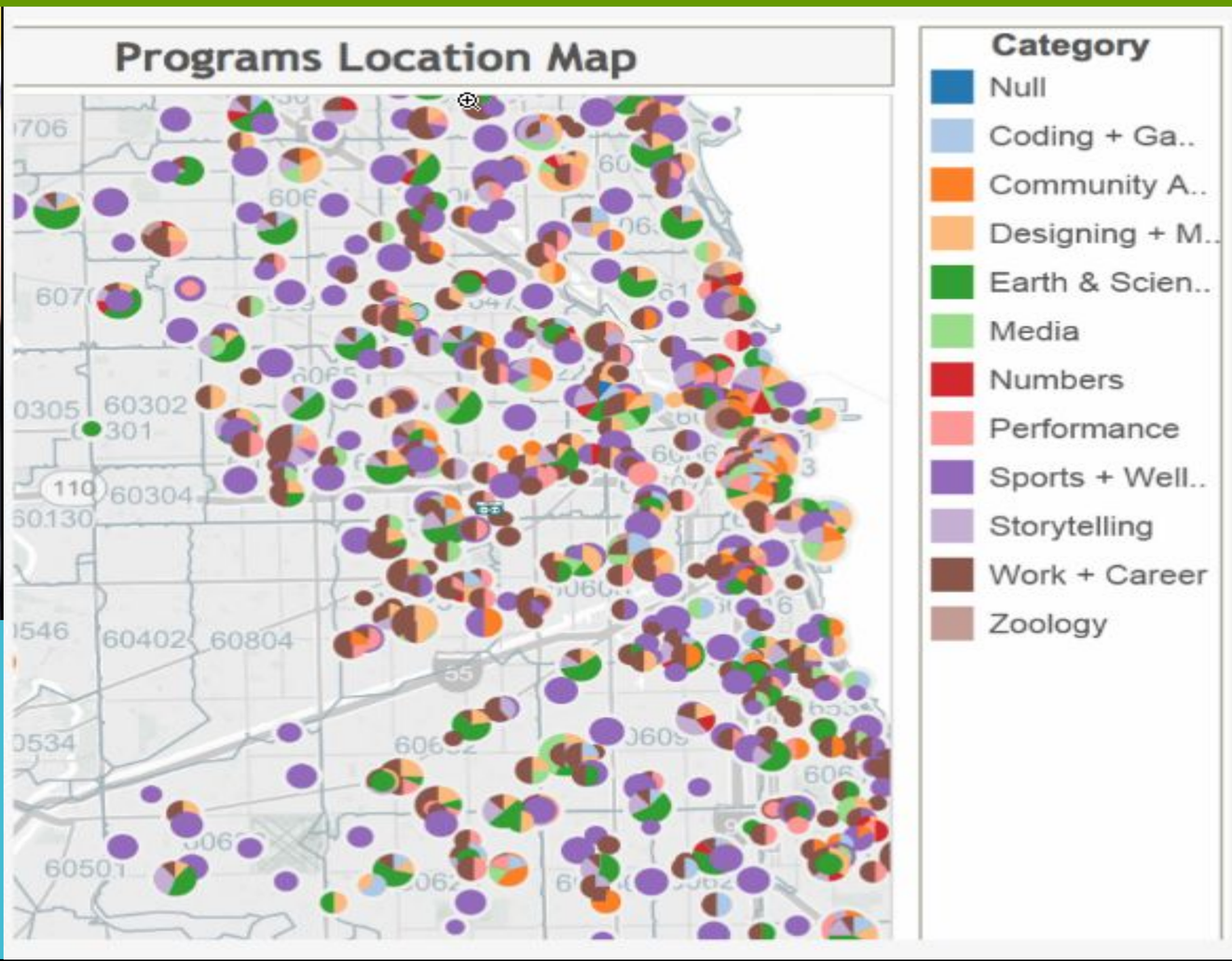
We are a people trying not only to **solve the problems of the present**, unemployment, inflation, but we are attempting on a larger scale to fulfill the promise of America.

We are attempting to fulfill our national purpose, to create and sustain a society in which all of us are equal.

Barbara Charlie Jordan, Former House of Representative, Texas
1976 Democratic National Convention Keynote Address

How are summer learning opportunities distributed in my city?

Describe



Which students are at risk of dropping out early?

Predict

Which skills and occupations are changing most rapidly in my area?

Detect

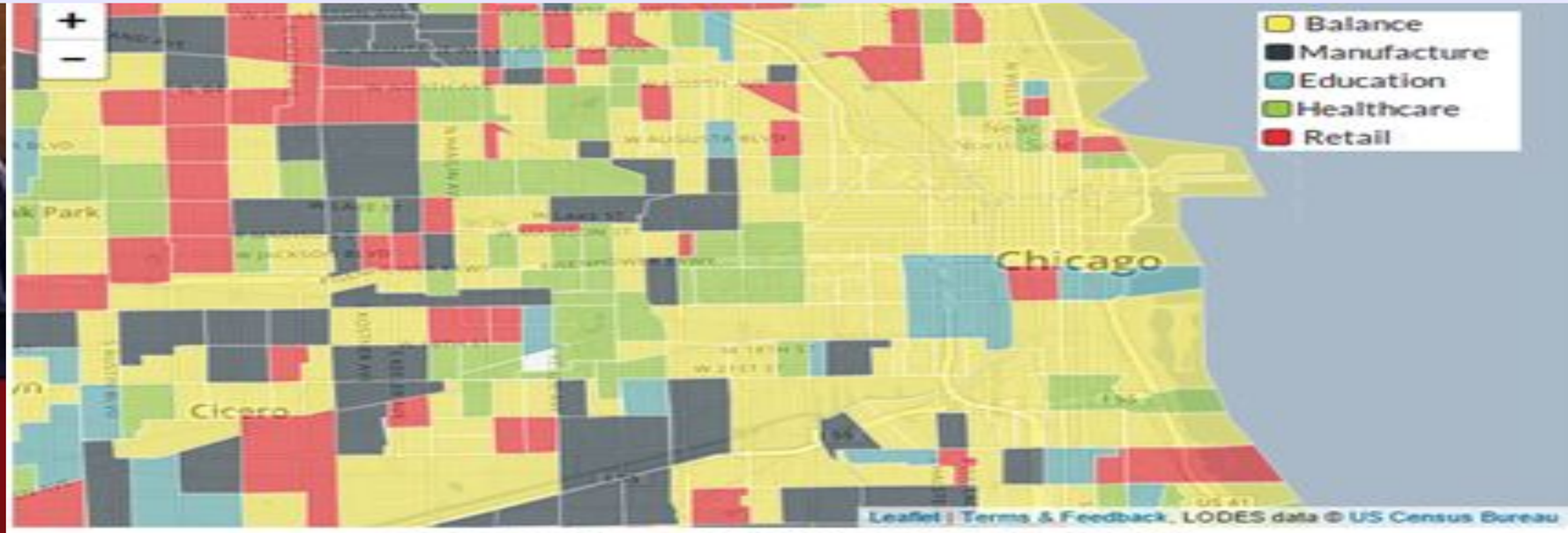
WEST SIDE FORWARD

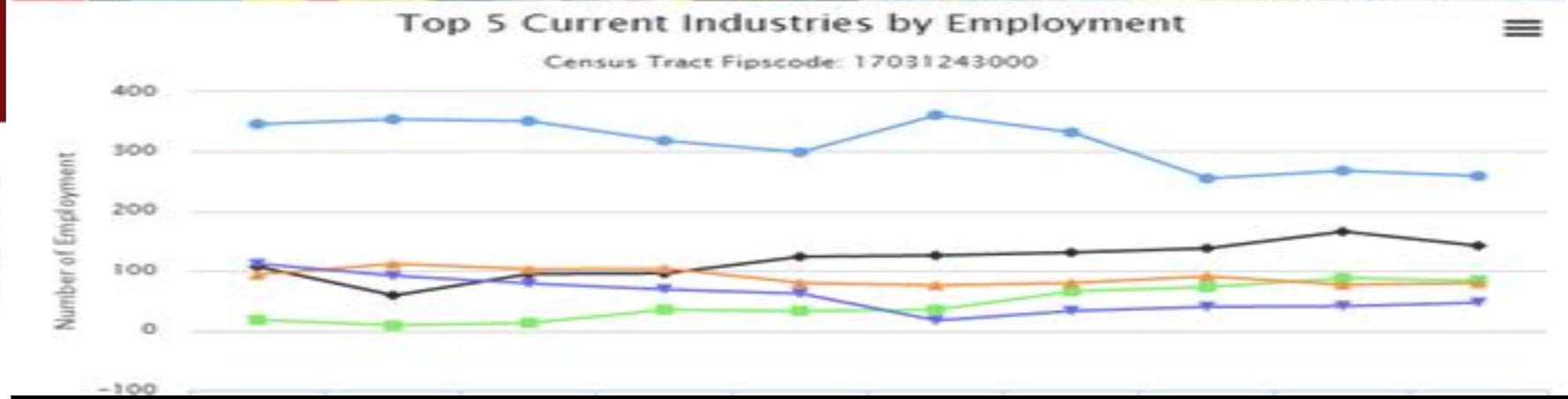
BETHEL NEW LIFE

CITY COLLEGES of CHICAGO

Center for Operational Excellence

Education that Works





How are my training programs affecting future wages and employment?

Evaluate

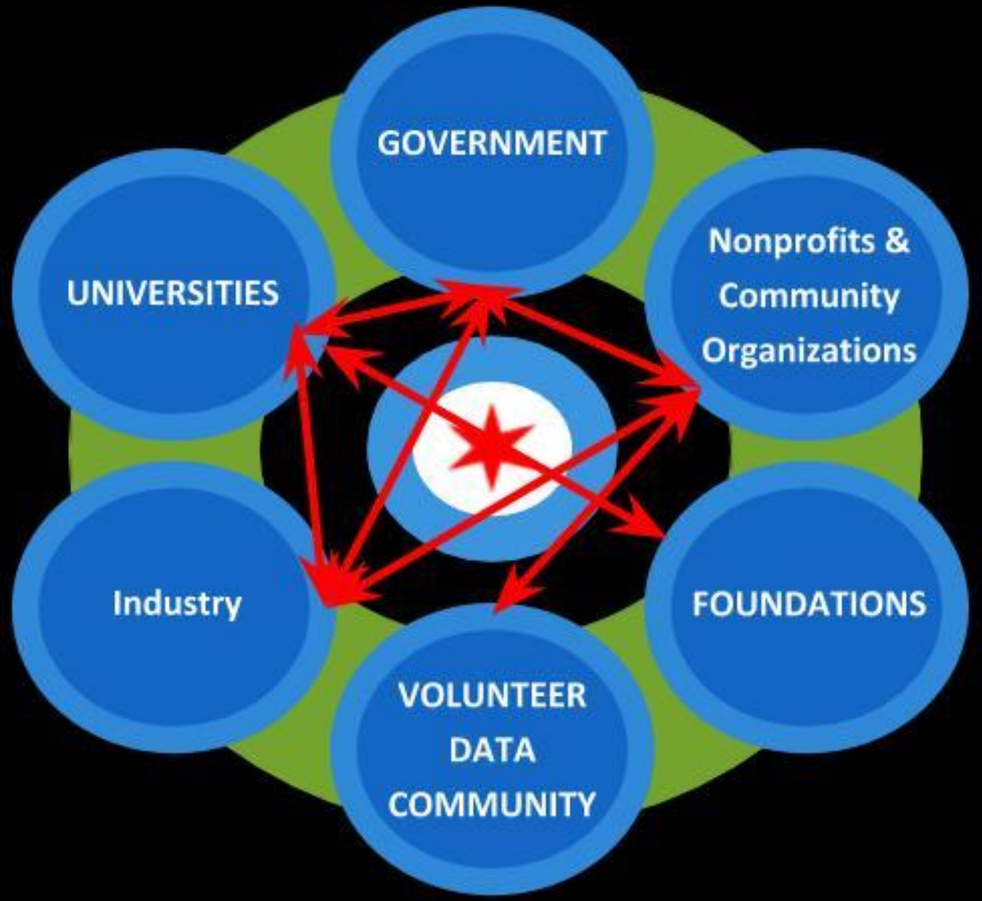
Challenges to Using Data For Social Good

Technical Barriers

Data Silos Within and Across Verticals

Lack of Internal **Analytics Capacity**

Lack of **Standard** Metrics and Algorithms



Cultural Barriers

Policies Limiting Data Interoperability

Lack of **Talent** Pipeline

Lack of Dedicated **Funding**

Limited Pathways for **Sustainment**


Exposed record
Data breach
Protection
classification
digital
Search
DATA
PRIVACY
law
anonimous
Metadata
Regulations
Patient privacy
Privacy
Sensitive Information
Authentication
RISK
compliance

Community-driven Principles for Ethical Data Sharing (CPEDS)

Bloomberg®

DATA FOR DEMOCRACY

brighthive

- To define values and priorities for overall ethical behavior by data scientists
- To develop a code of conduct that can guide a data scientist in being a thoughtful, responsible, ethical agent

Community-first approach

"By data scientists, for data scientists"

Increases responsiveness to the needs and concerns of data scientists

Better captures the diverse spectrum of interests across the data science community

May facilitate adoption of the code of ethics

We Asked….What Keeps You Up at Night?

(1) **facilitating data sharing** between individuals, domain experts and external data scientists, while

(2) **ensuring compliance** with relevant laws and norms, and, prior to taking action on the results,

(3) **ascertaining the validity of results** (accuracy of predictions and the effectiveness of proposed interventions).

Data Ownership & Provenance
Responsible Communications
Transparency & Openness
Questions & Answers
Bias and Mitigation
Privacy & Security
Thought Diversity

Privacy & Security

Data represent people (or living creatures).

They need to be protected.

Principles for Data Collection
Principles for Data Storage
Principles for Data Processing
Principles for Data Publishing

Privacy and Security – Principles
Data Collection

1. I will clearly state the purpose of the data collection to the person I am collecting/deriving the data from.
2. I will get consent for all that I collect.
3. If use 3rd party data, I will make sure the data has been collected with consent from user/participants or anonymized and respect privacy provisions.
4. When working with administrative data, I will seek to ensure to reduce risks to and create value for the people and communities the data was originally collected from.
5. I will not knowingly bias my data collection efforts in order to support a predetermined conclusion or an agenda.
6. I will properly document the source of my data to make it possible for other people to identify possible bias.
7. I will engage the community in discussion about privacy and security prior to setting up any large passive data collection efforts in that community.

Privacy and Security – Principles
Data Storage

1. I will use best effort to guarantee the safety of the data against loss of data, unauthorized access, destruction of data, misuse of data, tampering / modification of data, disclosure.
2. Discard the data properly when the period mentioned in the consent expires.
3. If there is a breach, I will communicate with the individuals that were affected and share what data was exposed.
4. I will follow best practices, e.g. encrypt data in transit and at rest, apply security patches, use key rotations, avoid data to be store on unsecure laptops.
5. I will keep the data safe, individual privacy should not be affected even in the event of a data breach.
6. I will fully understand the data governance policy for the collected data and abide by them.

Privacy and Security – Principles
Data Processing

1. I will strive to provide transparency and reproducibility for the processing of the data.
2. I will not try to reverse-engineer the data (i.e. de-anonymize) as part of the processing of the data itself.
3. The processing of the data should not create negative privacy side-effects.
4. I will take the same privacy and security care for intermediate results that are part of the processing.
5. In the presence of rules, I will follow them; in the absence of rules, I will adopt some or create new ones. [leave the place cleaner than you found it; leave the process less ambiguous than you found it]
6. I will not de-anonymize data and make it public just to show that this can be done.
7. If I find that a dataset can be de-anonymized, I will work with the individuals or company that shared the dataset to help them fix them problem.

Privacy and Security – Principles
Data Publishing

1. Unless I have consent and a reason, I will remove all PII information before sharing datasets.
2. I will not share data with third parties unless I have the consent from the individuals in the dataset.
3. Publishing of data should be compliant and compatible with the consent that was granted.

Transparency & Openness

Endeavor to provide transparency, in order to build trust, by design, and as default practice, and to do so responsibly, with context, and while preserving others rights to an explanation, recourse, and rectification.

1) Transparency for Trust
2) Transparency by Design
3) Transparency by Default
4) Transparency with Responsibility
5) Transparency with Context
6) Right to an explanation, recourse and rectification

Thought Diversity

Thought Diversity is an important tool to combat individual dogma and groupthink while helping to foster a culture of inclusion, diversity, humility and tolerance.

A culture such as this provides space that fosters creative and innovative thinking by enabling problem-solving to flourish.

1) Diversity for Accessibility
2) Diversity for Inclusivity
3) Diversity for Equality in Representation
4) Diversity for Openness
5) Diversity for Mitigating Bias

Bias

- Datasets can replicate or amplify social bias
- Especially risky when used in algorithmic decision making
- Outputs can potentially be unfair, misunderstood, and damaging to people

1) Provide specifications and rationale for data collection

2) Disclose how data is processed

3) Disclose how data is stored

4) Dealing with specific protected attributes

5) Check and control for bias in training/test sets

6) Deal with bias in algorithms

7) Disclosure of results

8) Acting on results/output

9) Issue caveats

Provenance

Clear data provenance promotes trust, aids transparency, and reduces the barriers to sharing data.

Collecting, propagating, storing, and curating provenance incurs costs, but these costs improve the value of the data and the decisions that the data inform.

Maintaining data provenance is a key responsibility, and it should be done in a way that is mindful of the privacy trade-offs.

#1: Provenance as a responsibility
#2: Provenance should promote trust
#3: Ignorance is not an excuse
#4: Provenance should be fit to purpose
#5: Provenance can evolve
#6: Ownership

Responsible Communications

Responsible Communications focuses on the **direct relationship between the Data Scientist and the Public**. In every situation, there is a balance point between the public and the data scientist.

Ethical considerations should drive where that balance point exists and support a **two-way, open flow of written, visual and spoken information** between the data science and the public. This should be broader than a specific industry or discipline.

accountability to the public

deciphering black-box algorithms

accessibility to lay audiences

education of the public

Questions & Answers

- Are your efforts directed at the right target?
- Are you addressing the right questions in the first place?
- Are you thinking through the broader implications of the answers and solutions you reach?

#1: I will recognize that everyone involved in a data science project has responsibility for thinking about the “right questions”
#2: I will not actively do harm to others by using data for an aim likely to harm people’s health, liberty, or wellbeing
#3: I will act to minimize the risk and impact of indirect harm
#4: I will consider my responsibility to solve problems of consequence to people and society
#5: I will consider the limits to the available data and what I can reasonably expect to find from it
#6: I will continuously re-evaluate the guiding question of the project
#7: I will engage with others who have more knowledge and experience
#8: I will engage with stakeholders, acknowledge their perspectives, and make efforts to bring them to discussions where possible
#9: I will accept and value multiple perspectives
#10: I will abide by the fundamental legal obligations of my field and profession
#11: I will take reasonable measures if aware of wrongdoing


D4GX SF - Ethics Principles
Poll locked. Responses not accepted.
12%
12%
8%
8%
6%
6%
6%
6%
6%
5%
4%
4%
Practice humility and openness
Be responsible for maximizing social benefit and minimizing harm
Bias will exist. Measure it. Plan for it.
Exercise ethical imagination
0%
2%
4%
6%

As data practitioners and data consumers, we aim to…

Consider (if not collect) informed and purposeful consent of data subjects for all projects, and discard resulting data when that consent expires.

Make best effort to **guarantee the security** of data, subjects, and algorithms to prevent unauthorized access, policy violations, tampering, or other harm or actions outside the data subjects' consent.

Make best effort to **protect anonymous data subjects**, and any associated data, against any attempts to reverse-engineer, de-anonymize, or otherwise expose confidential information.

Practice responsible **transparency as the default** where possible, throughout the entire data lifecycle.

Foster diversity by making efforts to ensure inclusion of participants, representation of viewpoints and communities, and openness. The data community should be open to, welcoming of, and inclusive of people from diverse backgrounds.

Acknowledge and mitigate unfair bias throughout all aspects of data work.

Hold up datasets with **clearly established provenance as the expected norm**, rather than the exception.

Respect relevant tensions of all stakeholders as it relates to privacy and data ownership.

Take great care to **communicate responsibly and accessibly**.

Ensure that all data practitioners take responsibility for **exercising ethical imagination in their work**, including considering the implication of what came before and what may come after, and actively working to increase benefit and prevent harm to others.

What Does This Mean For You

More details can be found at:
https://datapractices.org/community-principles-on-ethical-data-sharing/

The academic community builds the ethos of the data community

★ How ethics are incorporated into curriculums influences how ethics are considered in practice
  - *Are you creating ethical imaginations and teaching to ask questions?*

★ When aligned with academia this community can be used to increase accountability, guidance and development of technical standards
  - *How can you collaborate with the community to endorse, review, understand impact of ethical frameworks?*

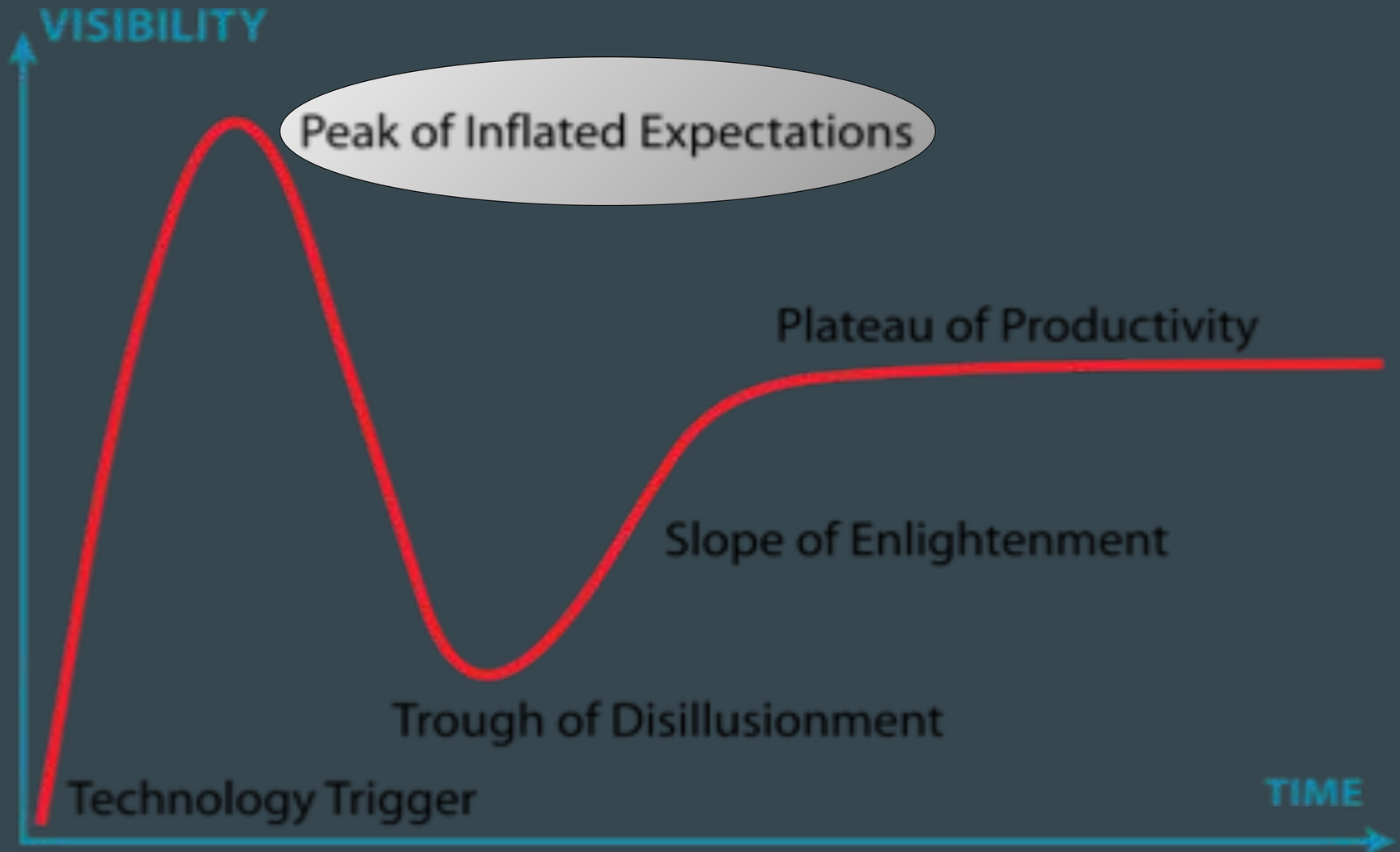
There is no executive order; there is no law that can require the American people to form a national community. This we must do as individuals, and if we do it as individuals, there is no President of the United States who can veto that decision…

we must define the "common good" and begin again to shape a common future.

Barbara Charlie Jordan, Former House of Representative, Texas
1976 Democratic National Convention Keynote Address

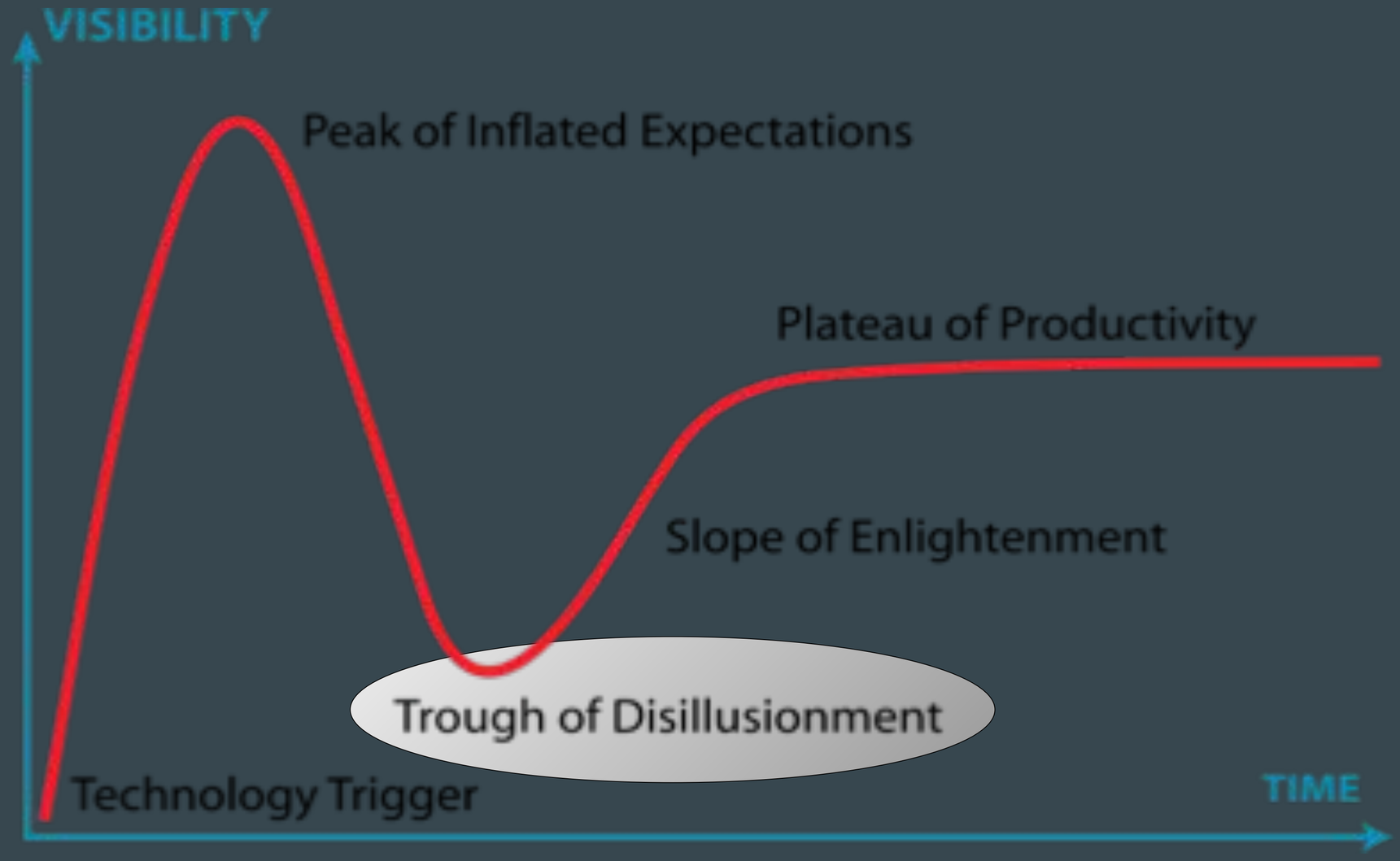
While Data Science can’t do the impossible, ***if we get this right***, it can make the possible more doable

“DATA SCIENCE CAN DO THE IMPOSSIBLE!”



Gartner

“DATA SCIENCE IS IMPOSSIBLE TO DO!”



Gartner