



**Data for Public
Good Campaign**
Twin Cities Innovation Alliance

SIMPLIFYING BIG DATA PRIMER

Supporting community exploration and conversation about the role of Big Data, Predictive Analytics, and Algorithms.



A NOTE FROM THE TWIN CITIES:

Our community has been familiarizing ourselves with the in's and out's of Big Data, Predictive Analytics, Algorithms, and Joint Powers Agreements (JPA) and we want to invite you into this important conversation. This data primer is an attempt to engage everyone in fundamentally understanding the definitions and the role of Big Data, Predictive Analytics and Algorithms in education and everywhere. Here are some useful terms to know and understand for starters.

And remember, No Data About US Without US

Marika Pfefferkorn

Co-Founder Twin Cities Innovation Alliance (TCIA)

Midwest Center for School Transformation (MCST)

**AUTHORS: CHRISTEN PENTEK, MARIKA PFEFFERKORN, &
SUSAN PHILLIPS**

EDITOR: DR. CATHERINE SQUIRES

CONTRIBUTORS: AASIM SHABAZZ, DR. TALAYA TOLEFREE



WHAT IS
DATA?

Data is everything about us and around us. It is everywhere. Data is a collection of facts. Data can be organized as information about a theme or topic. For example, in a group of people, how many identify as women? How many are LatinX? What emotions are present in the space? What objects or artifacts are found in the space? In education, data usually focus on people or their behaviors that might help clarify questions like who, what, when, where, how, and why.

Data can take many forms: it can be

WORDS

colors

NUMBERS

FEELINGS



DATA CAN BE QUALITATIVE

This latte has



a robust smell

frothy milk

strong taste

and is in a green mug

Qualitative data can be used to classify or categorize. For example we have a box of t-shirts that can be sorted by size, or color.



QUANTITATIVE DATA REFERS TO A NUMBER

This latte has

12 oz

is 150 degrees F.

and costs \$4.95



DATA CAN BE USED TO ANSWER A QUESTION OR TO TELL A STORY



A hand-drawn speech bubble with a black outline and a small tail pointing towards the bottom left. Inside the bubble, the text "WHAT IS BIG DATA?" is written in a bold, black, hand-drawn font, arranged in two lines. The background is white.

Big data are large and diverse volumes of information, collected by organizations and created when multiple data sources are combined, that can be mined for information and used in machine learning projects, predictive modeling, and other advanced analytics applications. The data can be collected and used by government systems, academic institutions, health care institutions, the insurance industry, social media, the digital advertising industry, and other corporations.

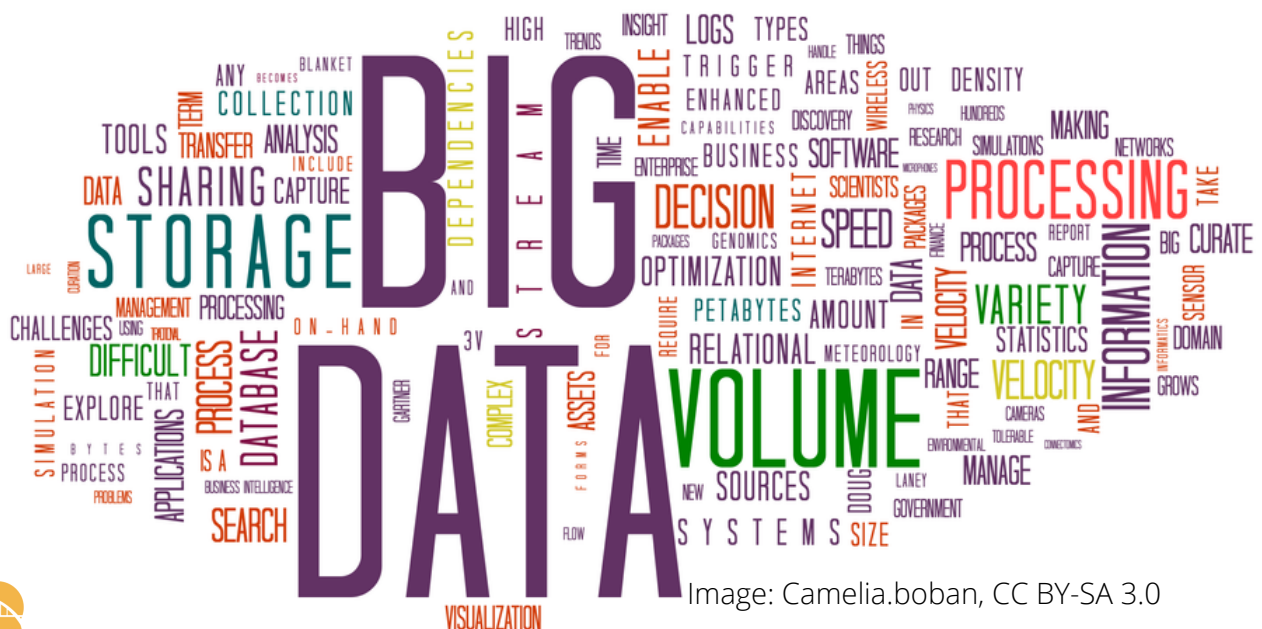


Image: Camelia.boban, CC BY-SA 3.0

HOW ARE DATA COLLECTED?

Data is often 'collected' through observations or direct contact with people. Sometimes data is collected directly from people, through conversations or filling in forms. For example, when a parent registers a kid for summer programming, staff may ask for information about the child. That information is data that the program should use to make decisions and improvements to the program.

Sometimes we aren't aware of the 'observations', like when data about our location, interests and habits is collected through social media and the digital apps on our smart phones and becomes part of a big data set. Or when the surveillance camera in the school hallway records who we speak with and how we walk. Or when we swipe our debit card at Target.





WE USE DATA TO MAKE DECISIONS EVERY DAY

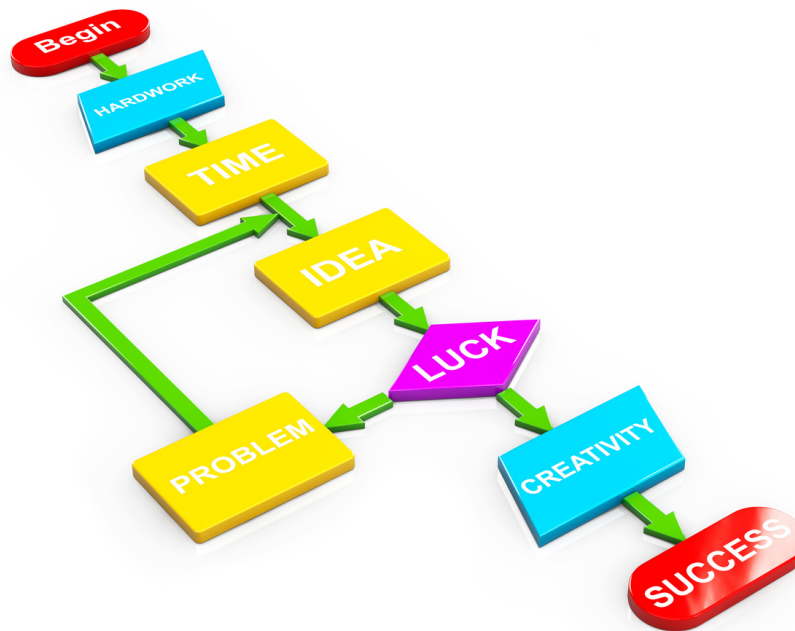


For example, if it is cloudy (data) we will predict it will rain (interpretation) and take a raincoat with us (data informed decision). However, in this example we do not ask the sky, 'is it okay for us to record that you have clouds today?' because we do not consider the rights of the sky in our decision making process. When information about people is gathered, it is important to gain their trust and permission before sharing their story or information about them. This does not always happen, so sometimes people are unaware of exactly when and how data are being collected and used.



ALGORITHMS

Algorithms are a set of steps used to complete a specific task or solve a problem. Digitally, they are a set of mathematical formulas that use data to create outputs and they allow things like computers, smart phones, and websites to function and make decisions. An algorithm decides the best way to get you from point a to point b on Maps or Waze. You might also use an algorithm to make dinner.



PREDICTIVE ANALYTICS

Predictive analytics is the use of data, algorithms, and machine learning to predict future outcomes based on historical data. Predictive Analytics programs look for patterns in behavior and then make predictions about whether certain people are more likely to succeed, buy certain products, be healthy, and so on.

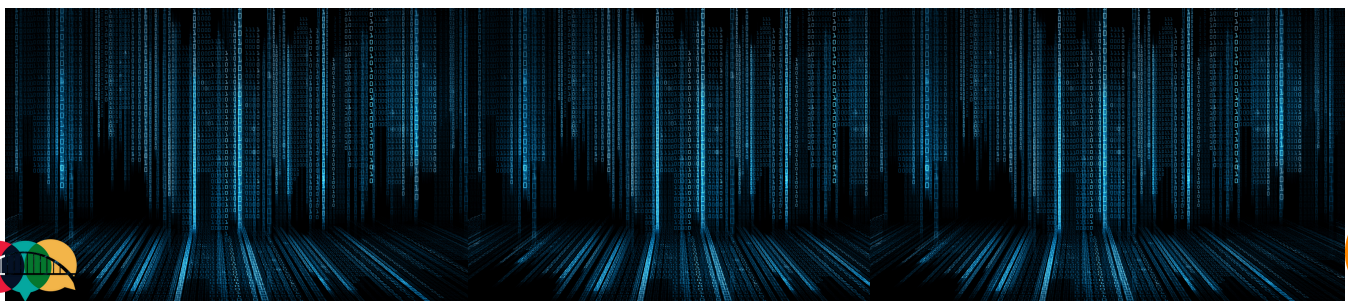
MACHINE LEARNING

Machine Learning is getting computers to act without being explicitly programmed. It is the branch of artificial intelligence in which a computer generates rules and predictions based on the raw data that it has been fed. The data used to train the computer to learn can be incomplete or biased, leading to biased outcomes.



MANY THINGS CAN 'GO WRONG' WITH MACHINE LEARNING

We think that technology is the way past human bias in our systems: machines can't be prejudiced, and data can't lie. Algorithms and the analytical tools are built by folks with bias and the historical data is full of bias as well. Researchers have identified three categories of bias in AI: algorithmic prejudice, negative legacy, and underestimation.



ALGORITHMIC HARM

Algorithmic prejudice occurs when there is a statistical dependence between protected features and other information used to make a decision. For example, early predictive policing algorithms did not have access to racial data when making predictions but the models relied heavily on geographic data (e.g. zip code), which is correlated with race in our geographically segregated society, ZIP codes and other location data are a common proxy for race.



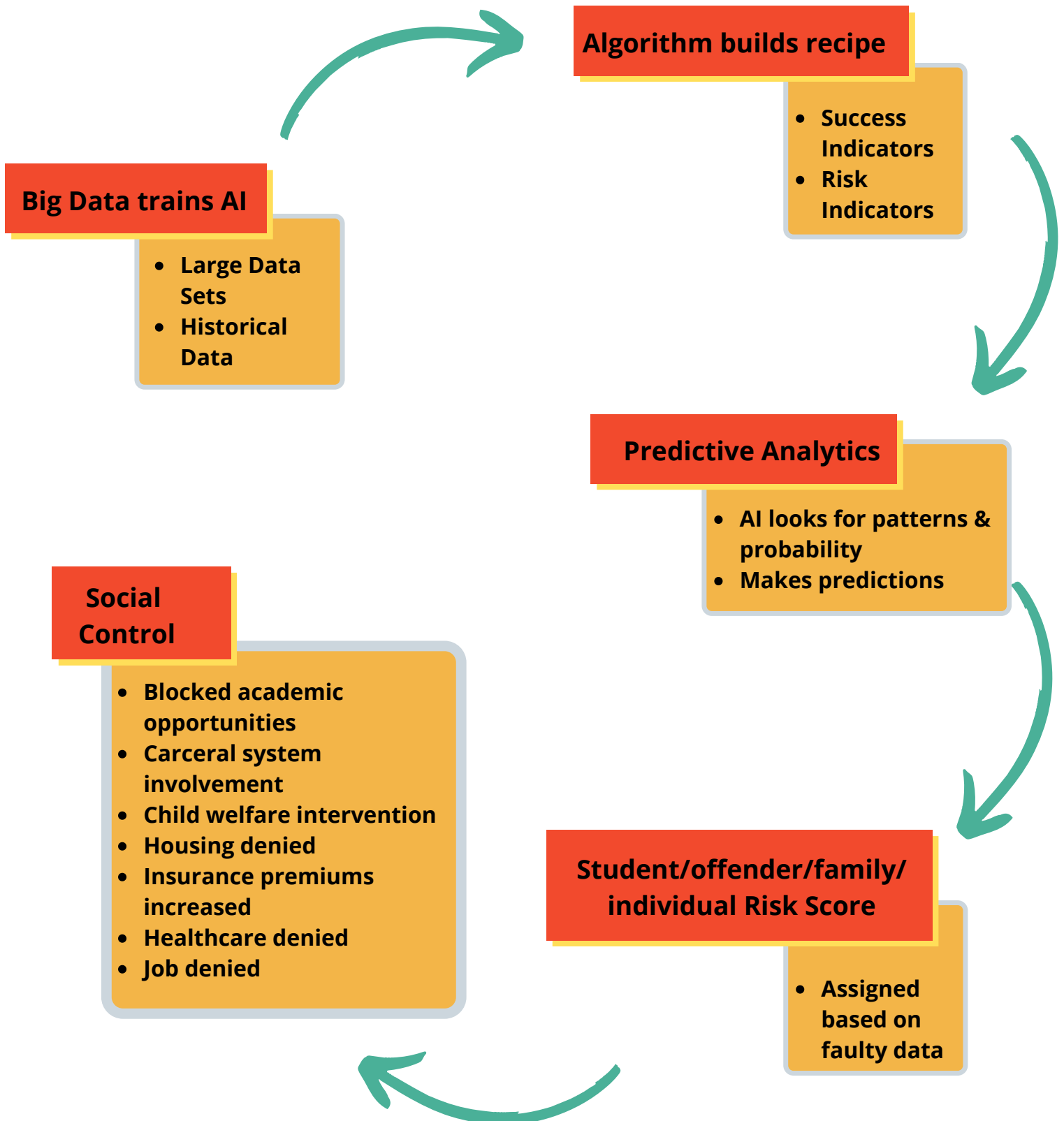
Negative legacy refers to bias already present in the data used to train the AI model. Bias is inherently built into the dataset that the machine learning model learns from. As such, the model can codify years of systemic bias against a population. Redlining, for example, or systematically denying loans to people based on where they live, has biased loan approval datasets towards whites. This bias in the data then leads to biased behavior of the AI model.



Underestimation occurs when there is not enough data for the model to make confident conclusions for some segments of the population. Amazon trained a machine learning model to screen applicants in its hiring process, but like many other tech companies, Amazon has a disproportionately male workforce. This data imbalance made its AI model more confident when evaluating men, leading to stronger recommendations for male applicants.



ALGORITHMS + BIG DATA + PREDICTIVE ANALYTICS



UNPACKING DATA DRIVEN DECISIONS

Data is used to inform decision making. Whether you look at the five-day forecast and decide to pack an umbrella, examine test grades to determine if smaller class sizes help students learn better, or use google search records to customize ads in social media, data-driven decisions are usually meant to influence behavior or drive change.

All too often, data is used to perpetuate social violence. The numbers may look “objective,” but when you dig deeper into the assumptions that drove which indicators were used, or what data was used to train the machine, it is easy to see how the numbers can tell a distorted story that harms people. Data can be misused in lots of ways in any part of the data collection, analysis, and decision making, and storage processes.

EXAMPLES:

The algorithms that ride-hailing companies, such as Uber and Lyft, use to determine fares create a racial bias: they charge a higher price per mile for a trip if the pickup point or destination is a neighborhood with a higher proportion of ethnic minority residents than for those with predominantly white residents.



An Optum algorithm was intended to help hospitals identify high-risk patients, such as those who have chronic conditions, to help providers know who may need additional resources to manage their health. But because the algorithm used health costs as the measure for health needs, many black and brown patients missed being identified for care since less money is spent on Black and brown patients in the United States.



MISUSE OF DATA IN RESEARCH

Unexplained. Data is misused when people talk about data without explaining how the data was collected, and what analysis process led to the results.

Extrapolation. Data is misused when an unrepresentative sample is used to represent the experience of a large population.

Non-consensual. Data is misused when institutions or individuals use the data for analysis after date agreed upon by consenting parties.

Dishonest or inappropriate methods of analysis. Sometimes people recycle data sets in ways that use assumptions that do not fit the original data set or collection circumstances.

Interpretation without consultation. Often, researchers or institutions produce analyses without input from representatives from the community. Data sets are gathered and interpreted with intent to delegitimize people's point of view or disqualify them for resources.




NO DATA ABOUT US WITHOUT US!

Our rights to privacy – the condition of being free from being observed or disturbed by other people – are protected by the 14th Amendment to the US Constitution. Privacy is the ability of an individual or a group to choose what to share about themselves publicly. Consent is normally how we give permission for our information – data – to be collected and used, when it will be collected, and who it will be shared with.

When we register our children for school we give permission for the district to collect and share identifying information with other 'school officials'. What we don't get told clearly is that often 'school officials' includes the police officer in the building and even the company who built the math App used in the classroom.

Each time we download an application for our smart phone without reading and customizing the privacy policies we are giving consent for data about us to be collected and likely sold. Each purchase we make with a debit card is another collection of data moment.

An illustration of a person with dark skin and short dreadlocks, wearing an orange t-shirt, blue jeans, and red sneakers. They are holding a red rectangular sign above their head with both hands. The sign has the text "NO DATA ABOUT US WITHOUT US!" in white, bold, sans-serif capital letters. The person is also holding a pink and white striped bag.

**NO DATA
ABOUT US
WITHOUT US!**

Because big data sets supposedly exist without identifying information, they don't need our permission to be shared. Unfortunately, it is easy to separate the data sets and eventually identify a person, and these AI systems could negatively impact your life.

An illustration of a person with short purple hair, wearing a red long-sleeved shirt and a red skirt. They are sitting in a black wheelchair and holding a large orange rectangular sign in front of them with both hands. The sign has the text "DATA JUSTICE NOW!" in black, bold, sans-serif capital letters.

**DATA
JUSTICE
NOW!**





It is time for us to come together and demand algorithmic accountability and the protection of privacy!

GET EDUCATED! Watch documentaries, read articles and books, follow leaders in this ecosystem to deepen your own understanding of data violence.

GET ENGAGED! Pass information on to your neighbors and young folk, do your own research to understand how data violence is showing up in your community- examine policies, ask officials questions.

GET ACTIVATED! Organize with your community, build power, and take action to push back on bad policies / tech design and get those most affected at the planning tables.



GLOSSARY

The following terms may be useful for exploring and understanding data and data systems:

ARTIFICIAL INTELLIGENCE: the theory and development of computer systems able to perform tasks that normally require human intelligence.

DATA BASE: A collection of data. Databases are most often digital; storage and access to this collection of information about people or things can be done electronically.

DATA ENTRAPMENT: the constant reinforcement of systems to perpetuate race-based myths and misinformation.

DATA VIOLENCE: the process of people being harmed by data misuse.

SURVEILLANCE: the monitoring of behavior, many activities, or information for the purpose of information gathering, influencing, managing or directing.

THREAT ASSESSMENT: an evaluation of events that can adversely affect operations and/or specific assets. (Threat assessment for schools is a fact-based process developed by the U.S. Secret Service and U.S. Department of Education that helps schools evaluate and assess potentially threatening students or situations.)

CRIMINALIZING TECHNOLOGY: the way in which digital technology has expanded the wars on crime and drugs, enabling our current state of mass incarceration and further entrenching the nation's racialized policing and punishment.

DIGITAL PANOPTICON: As a work of architecture, the panopticon allows a watchman to observe occupants without the occupants knowing whether or not they are being watched. The digital panopticon refers to the surveillance tendencies of disciplinarian societies.

PRIVACY: information that is not shared. (Fourth Amendment: Protects the right of privacy against unreasonable searches and seizures by the government.)

CONFIDENTIAL: information that is shared only with a small, select group of people. For example, a nurse may collect confidential information that will only be shared with the healthcare team.

CONSENT: to give permission. Consent should be:

- **Freely given.** Doing something with someone is a decision that should be made without pressure, force, manipulation, or while incapacitated. In technology, if an interface is designed to mislead people into doing something they normally wouldn't do, the application is not consentful.
- **Reversible.** Anyone can change their mind about what they want to do, at any time. In technology, you should have the right to limit access or entirely remove your data at any time.
- **Informed.** Be honest. For example, if someone says they'll use protection and then they don't, that's not consent. Consentful applications use clear and accessible language to inform users about the risks they present and the data they are storing, rather than burying these important details in e.g., the fine print of terms & conditions.
- **Enthusiastic.** If someone isn't excited, or really into it, that's not consent. If people are giving up their data because they have to in order to access necessary services and not because they want to, that is not consentful.
- **Specific.** Saying yes to one thing doesn't mean they've said yes to others. A consentful app only uses data the user has directly given, not data acquired through other means like scraping or buying, and uses it only in ways the user has consented to.
- **Reparations.** If your privacy is violated, what is the way in which harm will be repaired?

ENGINEERED CONSENT: the conditions that an unprecedented range of technologies, combined with an intentional taking advantage of short attention spans and lack of knowledge, that inhibit and eliminate freely given, reversible, informed, enthusiastic, and specific consent.