

outline for today:

misuse of data: the ways in which data collection, analysis, and handling go wrong

de-identification vs anonymization: preserving the privacy of individuals

personally identifiable information: what characteristics render a person “discoverable”

differential privacy: techniques for addressing ethical dilemmas

URL handling: accessing online databases via python

Misuse of data

can fall into several **broad** categories

Misuse of data

can fall into several **broad** categories

USING THE WRONG DATA FOR THE PROBLEM

MISHANDLING OF DATA

MISLEADING WITH DATA ANALYSIS

CRIMINAL USE OF DATA

Misuse of data

can fall into several **broad** categories

USING THE WRONG DATA FOR THE PROBLEM

- . improper data content/format
- . insufficient data (granularity, noise, etc.)
- . improper target identification

MISLEADING WITH DATA ANALYSIS

MISHANDLING OF DATA

CRIMINAL USE OF DATA

Misuse of data

can fall into several **broad** categories

USING THE WRONG DATA FOR THE PROBLEM

- . improper data content/format
- . insufficient data (granularity, noise, etc.)
- . improper target identification

MISLEADING WITH DATA ANALYSIS

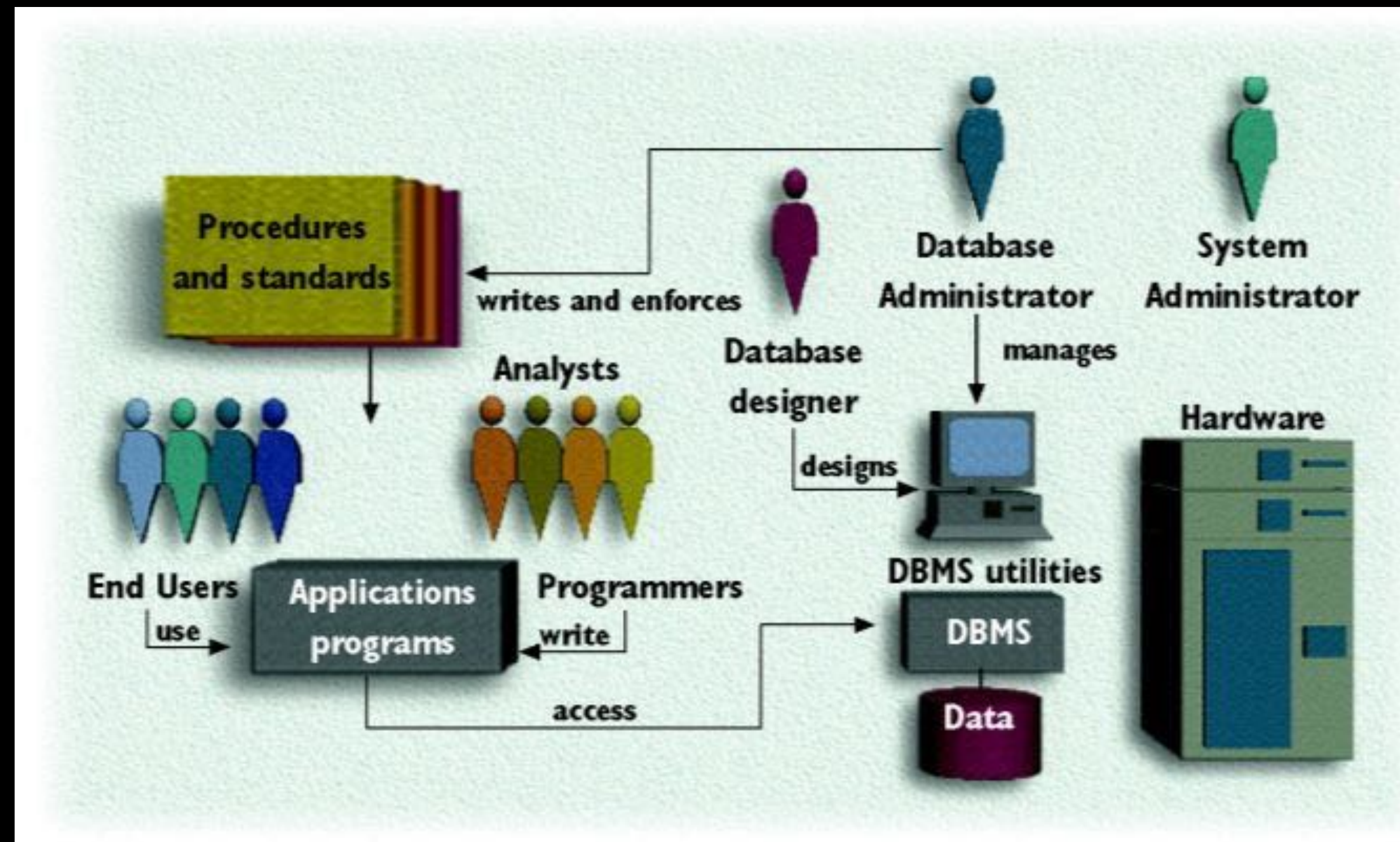
MISHANDLING OF DATA

- . security and access protocols
- . treating non-open data as open (or non-public as public)
- . inadvertent data sharing

CRIMINAL USE OF DATA

Misuse of data

can fall into several **broad** categories



MISHANDLING OF DATA

- . security and access protocols
- . treating non-open data as open (or non-public as public)
- . inadvertent data sharing

CRIMINAL USE OF DATA

<https://slideplayer.com/slide/6932570/>

Misuse of data

can fall into several **broad** categories

USING THE WRONG DATA FOR THE PROBLEM

- . improper data content/format
- . insufficient data (granularity, noise, etc.)
- . improper target identification

MISLEADING WITH DATA ANALYSIS

- . graphical misrepresentations
- . restricted analysis (selective data inclusion)
- . publication bias
- . faulty model selection

MISHANDLING OF DATA

- . security and access protocols
- . treating non-open data as open (or non-public as public)
- . inadvertent data sharing

CRIMINAL USE OF DATA

Misuse of data

can fall into several **broad** categories

USING THE WRONG DATA FOR THE PROBLEM

- . improper data content/format
- . insufficient data (granularity, noise, etc.)
- . improper target identification

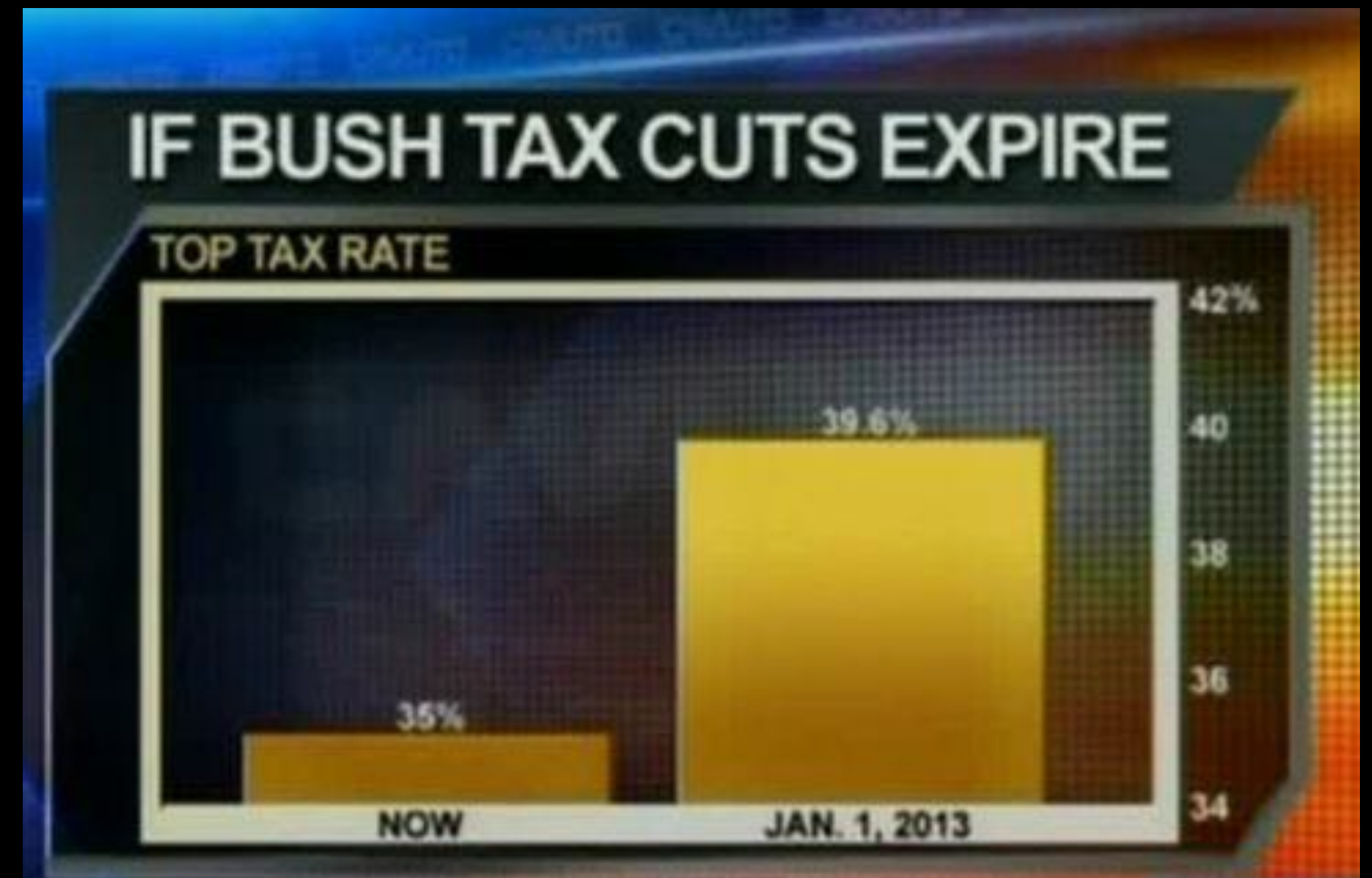
MISLEADING WITH DATA ANALYSIS

- . graphical misrepresentations
- . restricted analysis (selective data inclusion)
- . publication bias
- . faulty model selection

MISHANDLING OF DATA

- . security and access protocols
- . treating non-open data as open (or non-public as public)
- . inadvertent data sharing

<http://www.mediamatters.org>



Misuse of data

can fall into several **broad** categories

USING THE WRONG DATA FOR THE PROBLEM

- . improper data content/format
- . insufficient data (granularity, noise, etc.)
- . improper target identification

MISLEADING WITH DATA ANALYSIS

- . graphical misrepresentations
- . restricted analysis (selective data inclusion)
- . publication bias
- . faulty model selection

MISHANDLING OF DATA

- . security and access protocols
- . treating non-open data as open (or non-public as public)
- . inadvertent data sharing

CRIMINAL USE OF DATA

- . improper access and hacking
- . rogue data collection
- . restricted data collection regimes
- . data theft

Misuse of data

can fall into several **broad** categories

USING THE WRONG DATA FOR THE PROBLEM

- . improper data content/format
- . insufficient data (granularity, noise, etc.)
- . improper target identification

MISLEADING WITH DATA ANALYSIS

- . graphical misrepresentations
- . restricted analysis (selective data inclusion)
- . publication bias
- . faulty model selection

MISHANDLING OF DATA

- . security and access protocols
- . treating non-open data as open (or non-public as public)
- . inadvertent data sharing

CRIMINAL USE OF DATA

- . improper access and hacking
- . rogue data collection
- . restricted data collection regimes
- . data theft

Kyllo vs United States (2001)



<https://civicscourtcases.weebly.com/kyllo-vs-united-states-2001.html>

Kyllo vs United States (2001)



<https://civicscourtcases.weebly.com/kyllo-vs-united-states-2001.html>

Justice Scalia: "Where, as here, the Government uses a device that is not in general public use, to explore details of the home that would previously have been unknowable without physical intrusion, the surveillance is a "search" and is presumptively unreasonable without a warrant."

Kyllo vs United States (2001)



<https://civicscourtcases.weebly.com/kyllo-vs-united-states-2001.html>

Justice Scalia: "Where, as here, the Government uses a device that is not in general public use, to explore details of the home that would previously have been unknowable without physical intrusion, the surveillance is a "search" and is presumptively unreasonable without a warrant."



De-identification vs anonymization

De-identification vs anonymization



<https://arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cab-drivers-detailed-whereabouts/>

De-identification vs anonymization



<https://arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cab-drivers-detailed-whereabouts/>

[About TLC](#)[Data and Research](#)[TLC Initiatives](#)[Contact TLC](#)

Data

Pilot Programs

Industry Reports

Factbook

Request Data

TLC Trip Record Data

Expand All

Collapse All

The yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). The trip data was not created by the TLC, and TLC makes no representations as to the accuracy of these data.

The For-Hire Vehicle ("FHV") trip records include fields capturing the dispatching base license number and the pick-up date, time, and taxi zone location ID (shape file below). These records are generated from the FHV Trip Record submissions made by bases. Note: The TLC publishes base trip record data as submitted by the bases, and we cannot guarantee or confirm their accuracy or completeness. Therefore, this may not represent the total amount of trips dispatched by all TLC-licensed bases. The TLC performs routine reviews of the records and takes enforcement actions when necessary to ensure, to the extent possible, complete and accurate information.

▶ 2018

<https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

16

De-identification vs anonymization



<https://arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cab-drivers-detailed-whereabouts/>

hash
function



[About TLC](#)[Data and Research](#)[TLC Initiatives](#)[Contact TLC](#)

Data

Pilot Programs

Industry Reports

Factbook

Request Data

Expand All

Collapse All

The yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). The trip data was not created by the TLC, and TLC makes no representations as to the accuracy of these data.

The For-Hire Vehicle ("FHV") trip records include fields capturing the dispatching base license number and the pick-up date, time, and taxi zone location ID (shape file below). These records are generated from the FHV Trip Record submissions made by bases. Note: The TLC publishes base trip record data as submitted by the bases, and we cannot guarantee or confirm their accuracy or completeness. Therefore, this may not represent the total amount of trips dispatched by all TLC-licensed bases. The TLC performs routine reviews of the records and takes enforcement actions when necessary to ensure, to the extent possible, complete and accurate information.

▶ 2018

<https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

17

De-identification vs anonymization



<https://arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cab-drivers-detailed-whereabouts/>

hash
function



About TLC

Data and Research

TLC Initiatives

Contact TLC

Data

Pilot Programs

Industry Reports

Factbook

Request Data

TLC Trip Record Data

Expand All

Collapse All

The yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP). The trip data was not created by the TLC, and TLC makes no representations as to the accuracy of these data.

The For-Hire Vehicle ("FHV") trip records include fields capturing the dispatching base license number and the pick-up date, time, and taxi zone location ID (shape file below). These records are generated from the FHV Trip Record submissions made by bases. Note: The TLC publishes base trip record data as submitted by the bases, and we cannot guarantee or confirm their accuracy or completeness. Therefore, this may not represent the total amount of trips dispatched by all TLC-licensed bases. The TLC performs routine reviews of the records and takes enforcement actions when necessary to ensure, to the extent possible, complete and accurate information.

► 2018

<https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

de-identification

anonymization

De-identification vs anonymization



<https://arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cab-drivers-detailed-whereabouts/>

hash
function

↔



<https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

de-identification: encoding indicators from data that can link an individual and an “object” in a data set

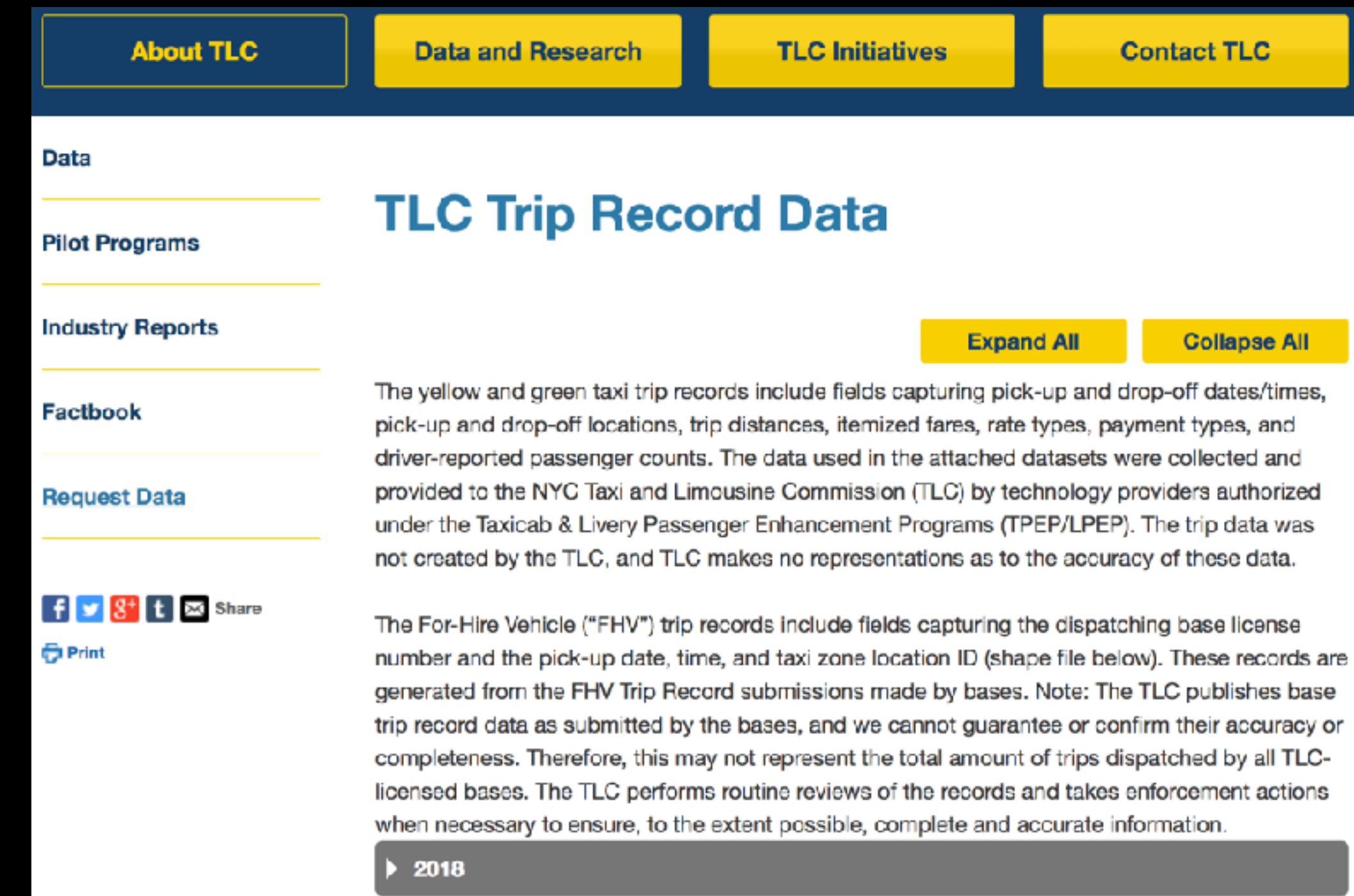
anonymization

De-identification vs anonymization



<https://arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cab-drivers-detailed-whereabouts/>

hash
function



<https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

de-identification: encoding indicators from data that can link an individual and an “object” in a data set

anonymization: permanently removing such indicators from the original data set

Personally identifiable information (PII)

Personally identifiable information (PII)

NIST (National Institute of Standards and Technology)

“Information which can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc.”

Personally identifiable information (PII)

NIST (National Institute of Standards and Technology)

“Information which can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother’s maiden name, etc.”

Full name (if not common)

Face (sometimes)

Home address

Email address

National identification number (e.g., Social Security number in the U.S.)

Passport number

Vehicle registration plate number

Driver's license number

Face, fingerprints, or handwriting

Credit card numbers

Digital identity

Date of birth

Birthplace

Genetic information

Telephone number

Login name, screen name, nickname, or handle

Personally identifiable information (PII)

NIST (National Institute of Standards and Technology)

“Information which can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc.”

Full name (if not common)

Face (sometimes)

Home address

Email address

National identification number (e.g., Social Security number in the U.S.)

Passport number

Vehicle registration plate number

Driver's license number

Face, fingerprints, or handwriting

Credit card numbers

Digital identity

Date of birth

Birthplace

Genetic information

Telephone number

Login name, screen name, nickname, or handle

THINGS TO CONSIDER

1. laws are different in US vs EU vs other
2. PII is minimal requirement for most IRB involvement
3. imagine this data set:

Personally identifiable information (PII)

NIST (National Institute of Standards and Technology)

“Information which can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother’s maiden name, etc.”

Full name (if not common)

Face (sometimes)

Home address

Email address

National identification number (e.g., Social Security number in the U.S.)

Passport number

Vehicle registration plate number

Driver's license number

Face, fingerprints, or handwriting

Credit card numbers

Digital identity

Date of birth

Birthplace

Genetic information

Telephone number

Login name, screen name, nickname, or handle

THINGS TO CONSIDER

1. laws are different in US vs EU vs other
2. PII is minimal requirement for most IRB involvement
3. imagine this data set:

name, zipcode, birth date, binary gender

Personally identifiable information (PII)

NIST (National Institute of Standards and Technology)

“Information which can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc.”

Full name (if not common)

Face (sometimes)

Home address

Email address

National identification number (e.g., Social Security number in the U.S.)

Passport number

Vehicle registration plate number

Driver's license number

Face, fingerprints, or handwriting

Credit card numbers

Digital identity

Date of birth

Birthplace

Genetic information

Telephone number

Login name, screen name, nickname, or handle

THINGS TO CONSIDER

1. laws are different in US vs EU vs other
2. PII is minimal requirement for most IRB involvement
3. imagine this data set:

~~name~~, zipcode, birth date, binary gender

Personally identifiable information (PII)

NIST (National Institute of Standards and Technology)

“Information which can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc.”

Full name (if not common)

Face (sometimes)

Home address

Email address

National identification number (e.g., Social Security number in the U.S.)

Passport number

Vehicle registration plate number

Driver's license number

Face, fingerprints, or handwriting

Credit card numbers

Digital identity

Date of birth

Birthplace

Genetic information

Telephone number

Login name, screen name, nickname, or handle

THINGS TO CONSIDER

1. laws are different in US vs EU vs other
2. PII is minimal requirement for most IRB involvement
3. imagine this data set:

~~name~~, zipcode, birth date, binary gender

how “identifying” is this?

Personally identifiable information (PII)

NIST (National Institute of Standards and Technology)

“Information which can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc.”

Full name (if not common)

Face (sometimes)

Home address

Email address

National identification number (e.g., Social Security number in the U.S.)

Passport number

Vehicle registration plate number

Driver's license number

Face, fingerprints, or handwriting

Credit card numbers

Digital identity

Date of birth

Birthplace

Genetic information

Telephone number

Login name, screen name, nickname, or handle

THINGS TO CONSIDER

1. laws are different in US vs EU vs other
2. PII is minimal requirement for most IRB involvement
3. imagine this data set:

~~name~~, zipcode, birth date, binary gender

how “identifying” is this?

87% of the population of the United States is likely to be uniquely identified by (5-digit ZIP, gender, date of birth)

- Latanya Sweeney (2009)

Differential Privacy

Differential Privacy

In much of the analysis we've seen so far, we have looked at **aggregate quantities**: number of businesses, sea level for whole oceans, total number of people by race, etc...

Differential Privacy

In much of the analysis we've seen so far, we have looked at **aggregate quantities**: number of businesses, sea level for whole oceans, total number of people by race, etc...

... and for many inquiries aggregate quantities provide sufficient information. We do not need detailed knowledge of **individual records**.

Differential Privacy

In much of the analysis we've seen so far, we have looked at **aggregate quantities**: number of businesses, sea level for whole oceans, total number of people by race, etc...

... and for many inquiries aggregate quantities provide sufficient information. We do not need detailed knowledge of **individual records**.

But we may want to generate custom aggregators that take advantage of information from individual records when aggregating (e.g., detailed pair-wise locations of individuals)

Differential Privacy

In much of the analysis we've seen so far, we have looked at **aggregate quantities**: number of businesses, sea level for whole oceans, total number of people by race, etc...

... and for many inquiries aggregate quantities provide sufficient information. We do not need detailed knowledge of **individual records**.

But we may want to generate custom aggregators that take advantage of information from individual records when aggregating (e.g., detailed pair-wise locations of individuals)

Differential privacy adds noise to the raw data that renders individual identification impossible while preserving aggregate characteristics.

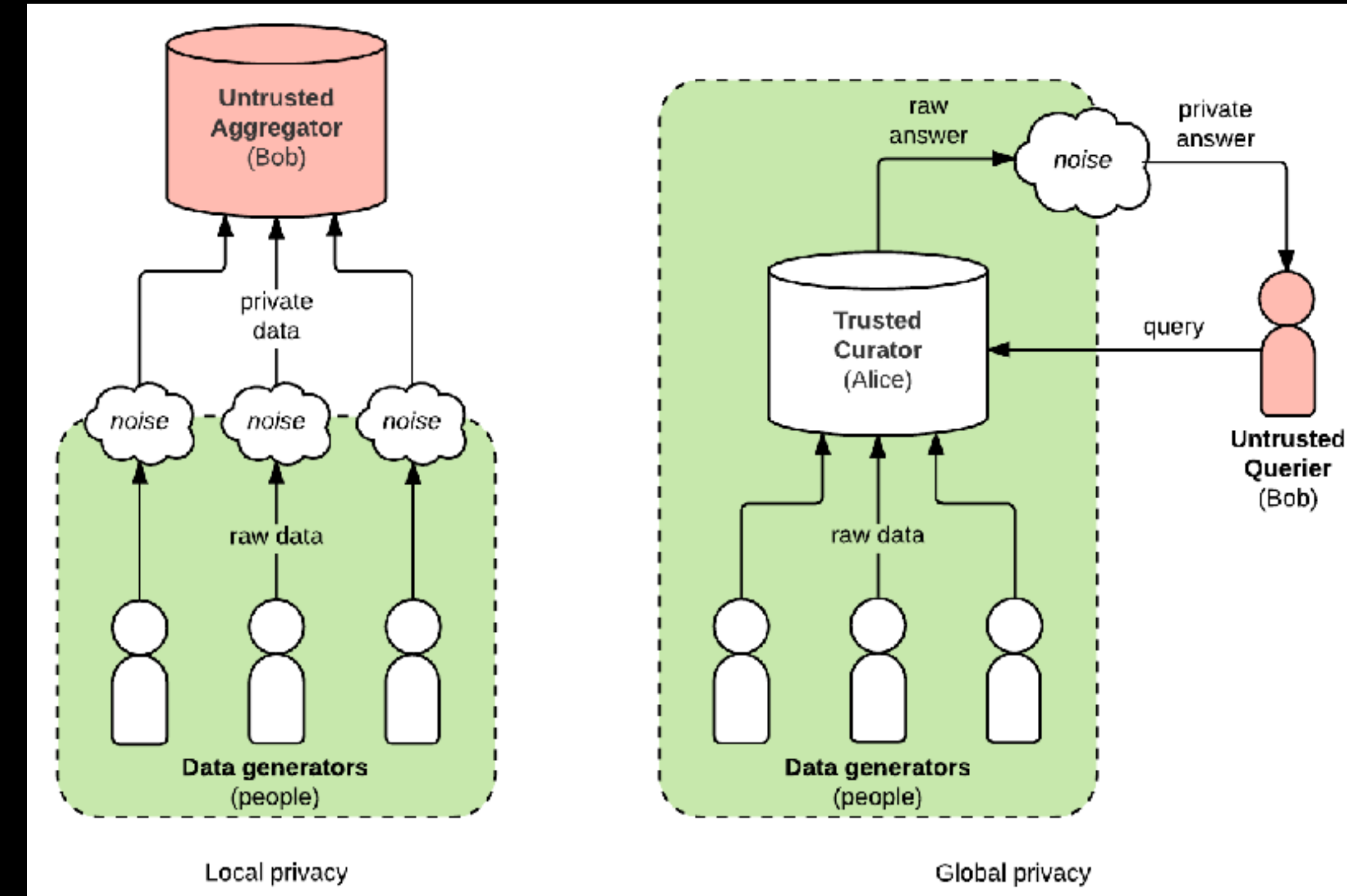
Differential Privacy

In much of the analysis we've seen so far, we have looked at **aggregate quantities**: number of businesses, sea level for whole oceans, total number of people by race, etc...

... and for many inquiries aggregate quantities provide sufficient information. We do not need detailed knowledge of **individual records**.

But we may want to generate custom aggregators that take advantage of information from individual records when aggregating (e.g., detailed pair-wise locations of individuals)

Differential privacy adds noise to the raw data that renders individual identification impossible while preserving aggregate characteristics.



<https://www.accessnow.org/understanding-differential-privacy-matters-digital-rights/>

Open Data

Open Data

sometimes referred to as public data^{*}

^{*} inaccurately... there are several key distinctions

Open Data

sometimes referred to as **public data**^{*}, is data that is:

free

available

redistributable

^{*} inaccurately... there are several key distinctions

Open Data

sometimes referred to as **public data**^{*}, is data that is:

free

no subscription fees or payment are required for access

available

redistributable

^{*} inaccurately... there are several key distinctions

Open Data

sometimes referred to as **public data**^{*}, is data that is:

free

no subscription fees or payment are required for access

available

hosted on a web server that is not restricted

redistributable

^{*} inaccurately... there are several key distinctions

Open Data

sometimes referred to as **public data**^{*}, is data that is:

free

no subscription fees or payment are required for access

available

hosted on a web server that is not restricted

redistributable

there are no ownership rights that prevent sharing

^{*} inaccurately... there are several key distinctions

Open Data

sometimes referred to as **public data**^{*}, is data that is:

free

no subscription fees or payment are required for access

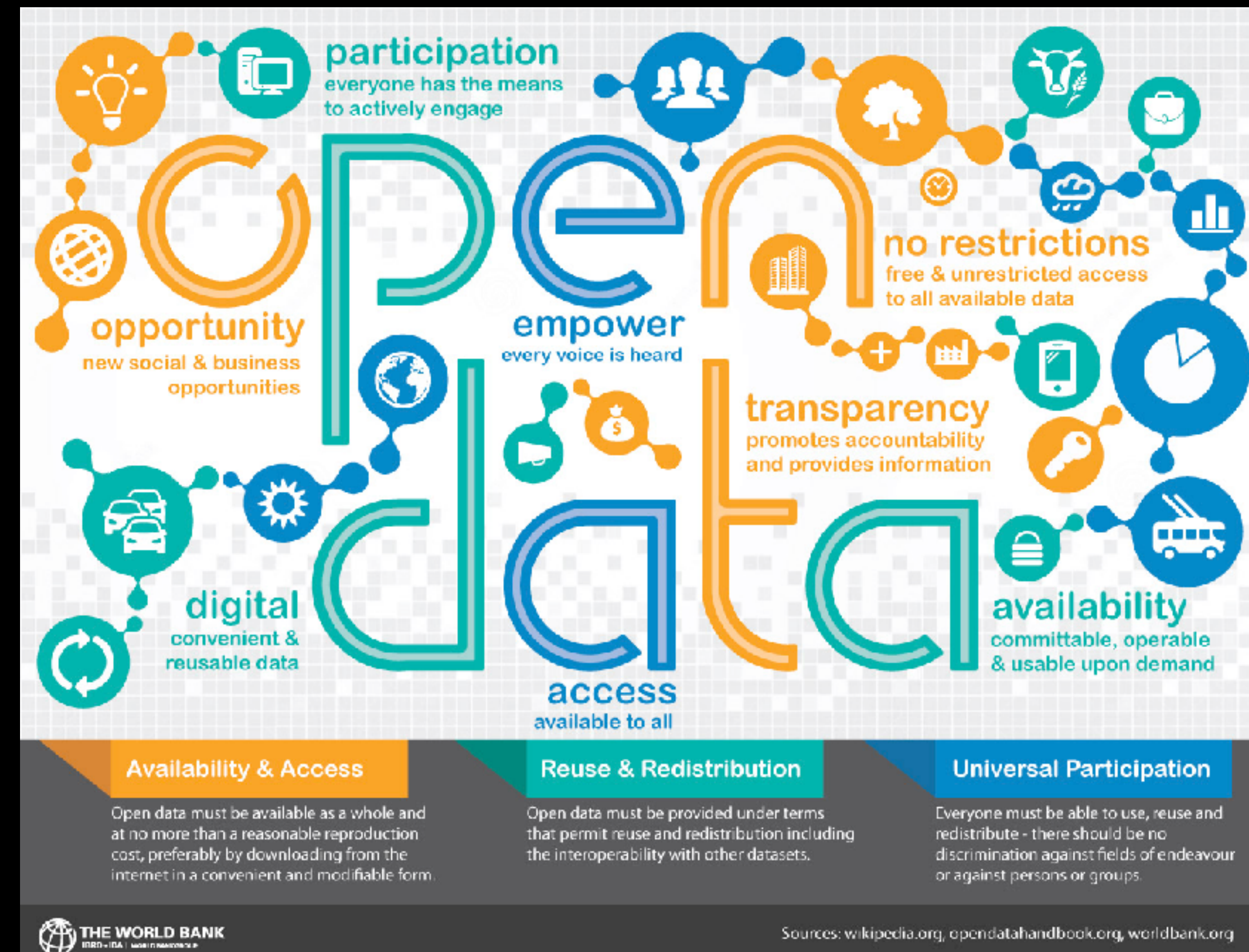
available

hosted on a web server that is not restricted

redistributable

there are no ownership rights that prevent sharing

<http://www.worldbank.org>



^{*} inaccurately... there are several key distinctions

Open Data

sometimes referred to as **public data**^{*}, is data that is:

free

no subscription fees or payment are required for access

available

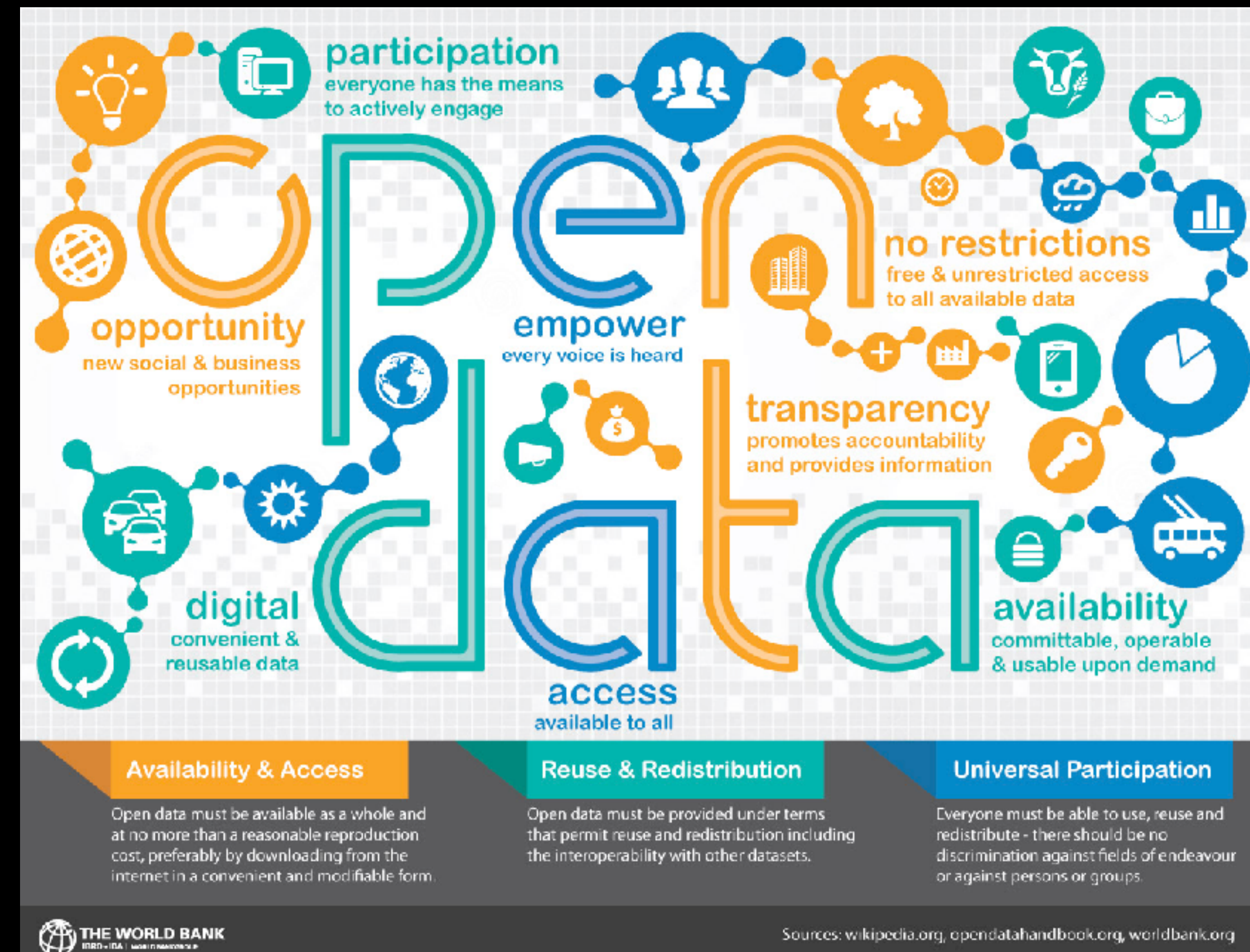
hosted on a web server that is not restricted

redistributable

there are no ownership rights that prevent sharing

but note that “open” ≠ “easy”

<http://www.worldbank.org>



^{*} inaccurately... there are several key distinctions

Open Data

sometimes referred to as **public data**^{*}, is data that is:

free ≠ **transparent**

no subscription fees or payment are required for access

available ≠ **usable**

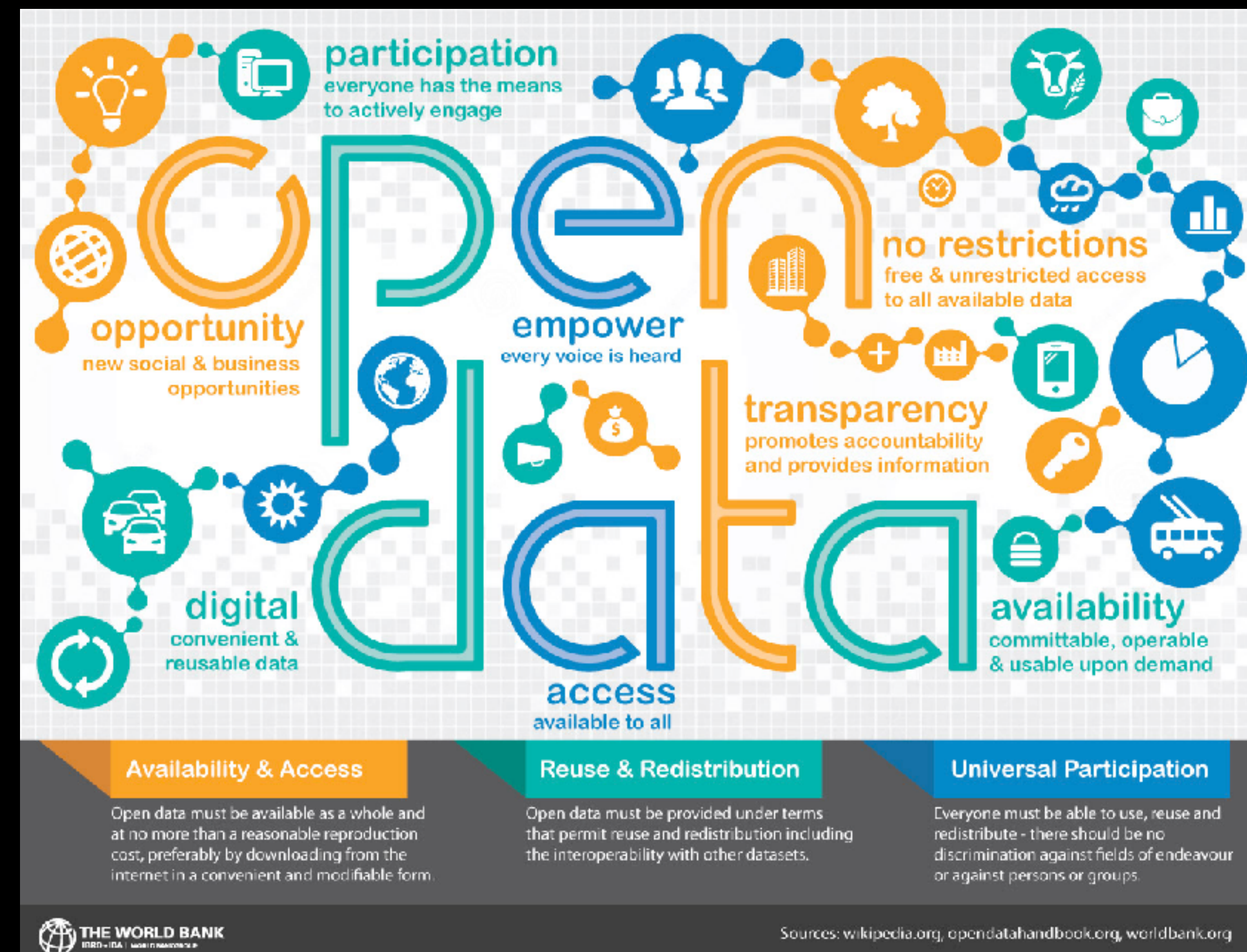
hosted on a web server that is not restricted

redistributable ≠ **sourced**

there are no ownership rights that prevent sharing

but note that “open” ≠ “easy”

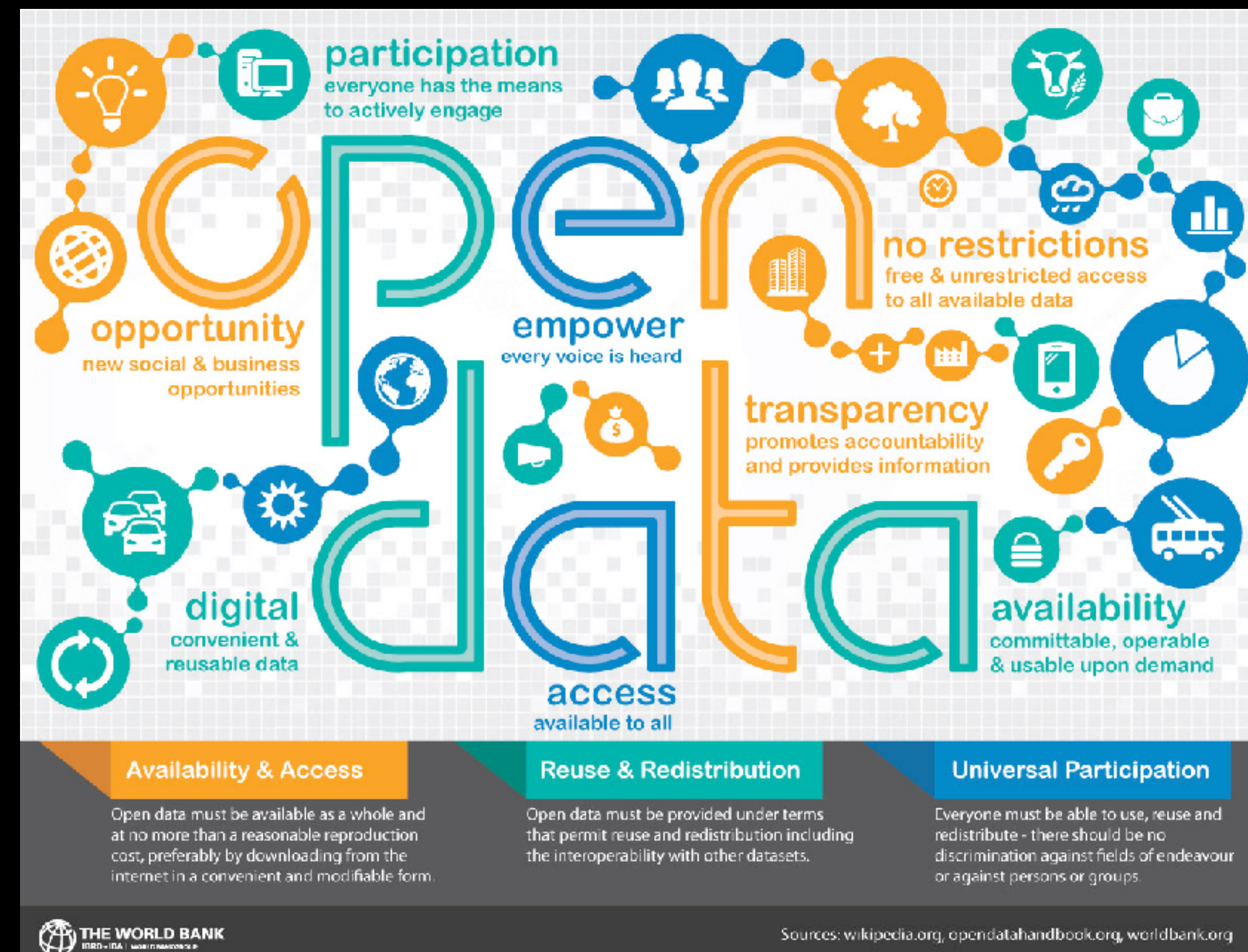
<http://www.worldbank.org>



^{*} inaccurately... there are several key distinctions

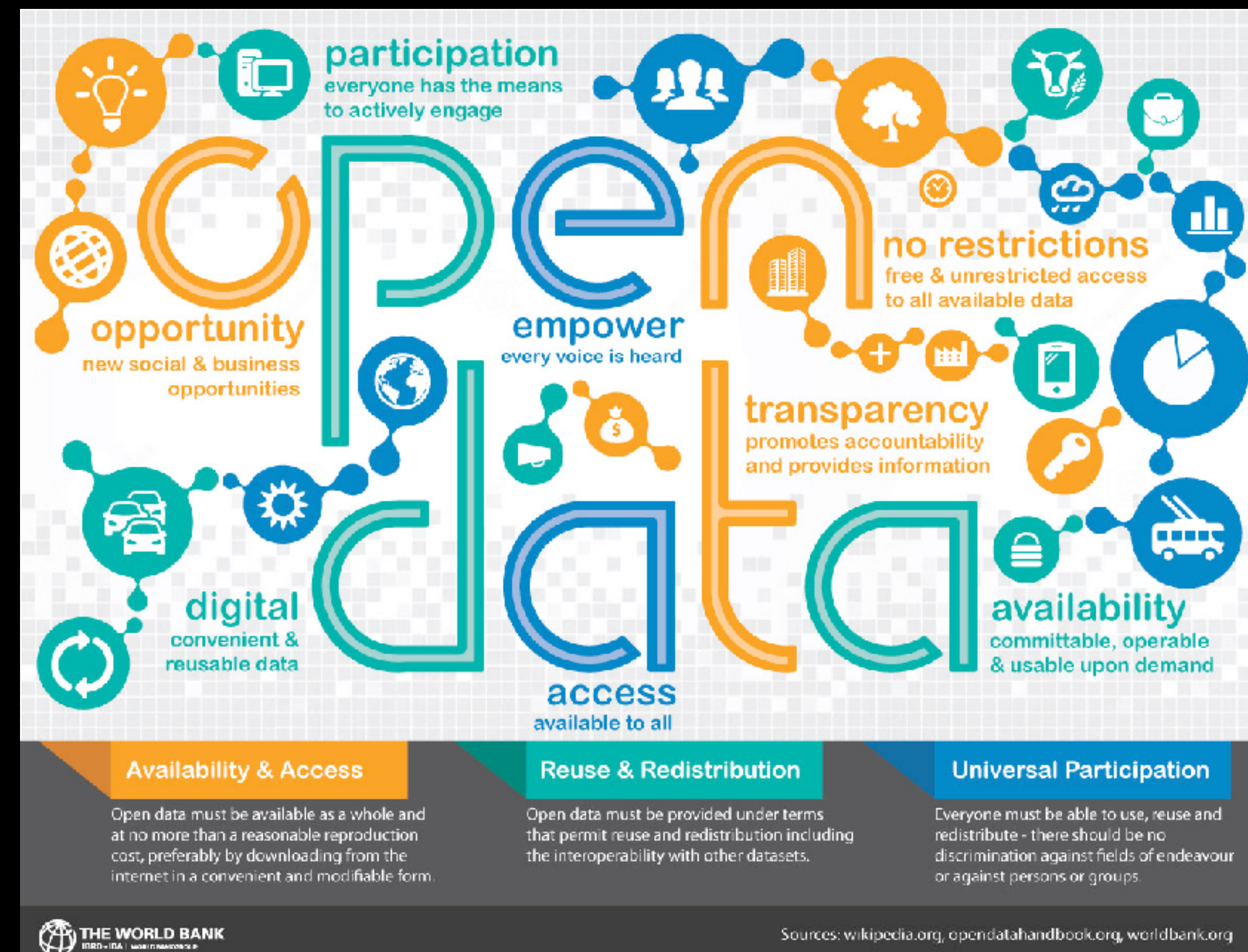
Open Data

<http://www.worldbank.org>



Open Data ...why???

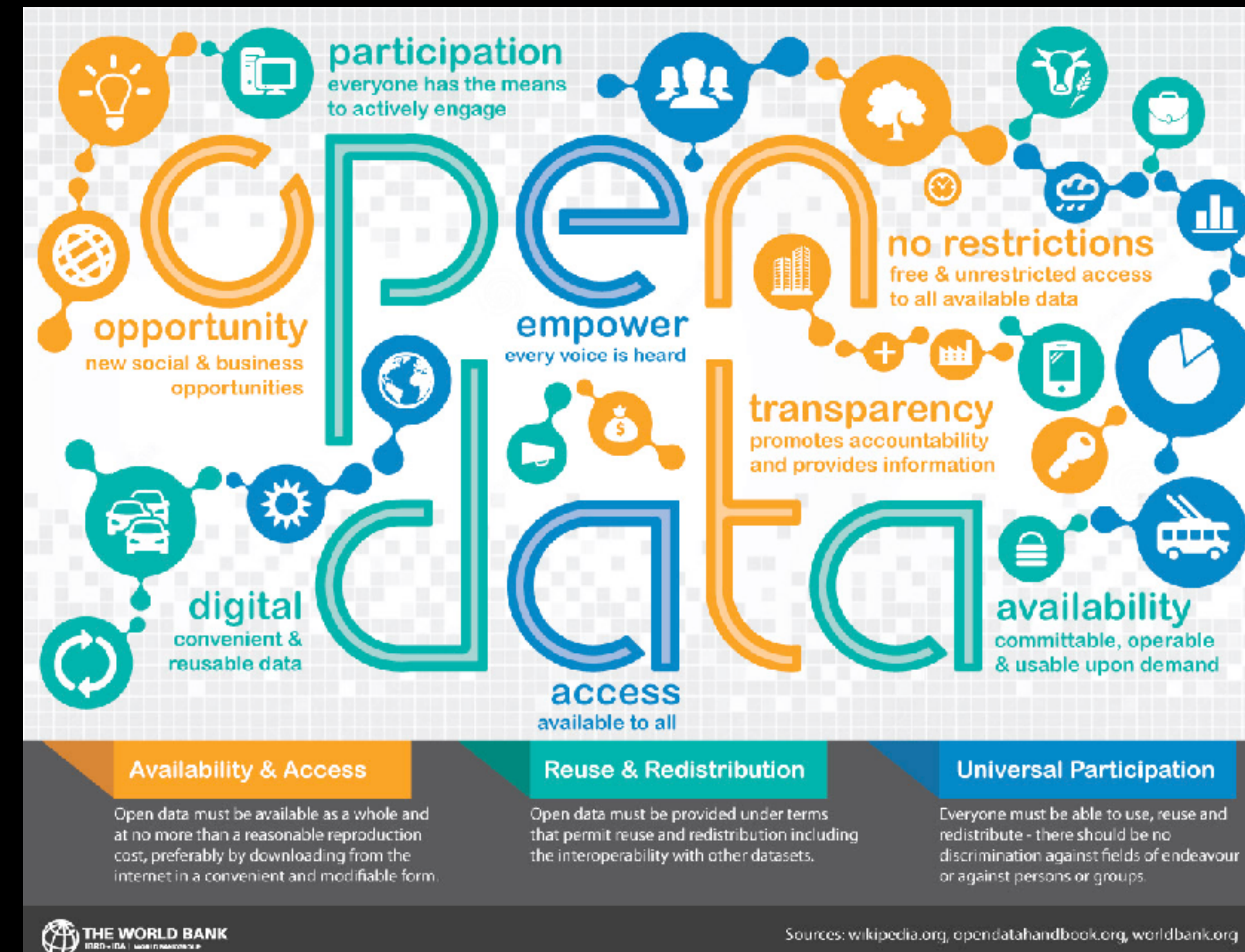
<http://www.worldbank.org>



Open Data ...why???

the open data paradigm has a variety of benefits that motivate data holders to open up access to their data:

<http://www.worldbank.org>



Open Data ...why???

the open data paradigm has a variety of benefits that motivate data holders to open up access to their data:

TRANSPARENCY

REPRODUCIBILITY

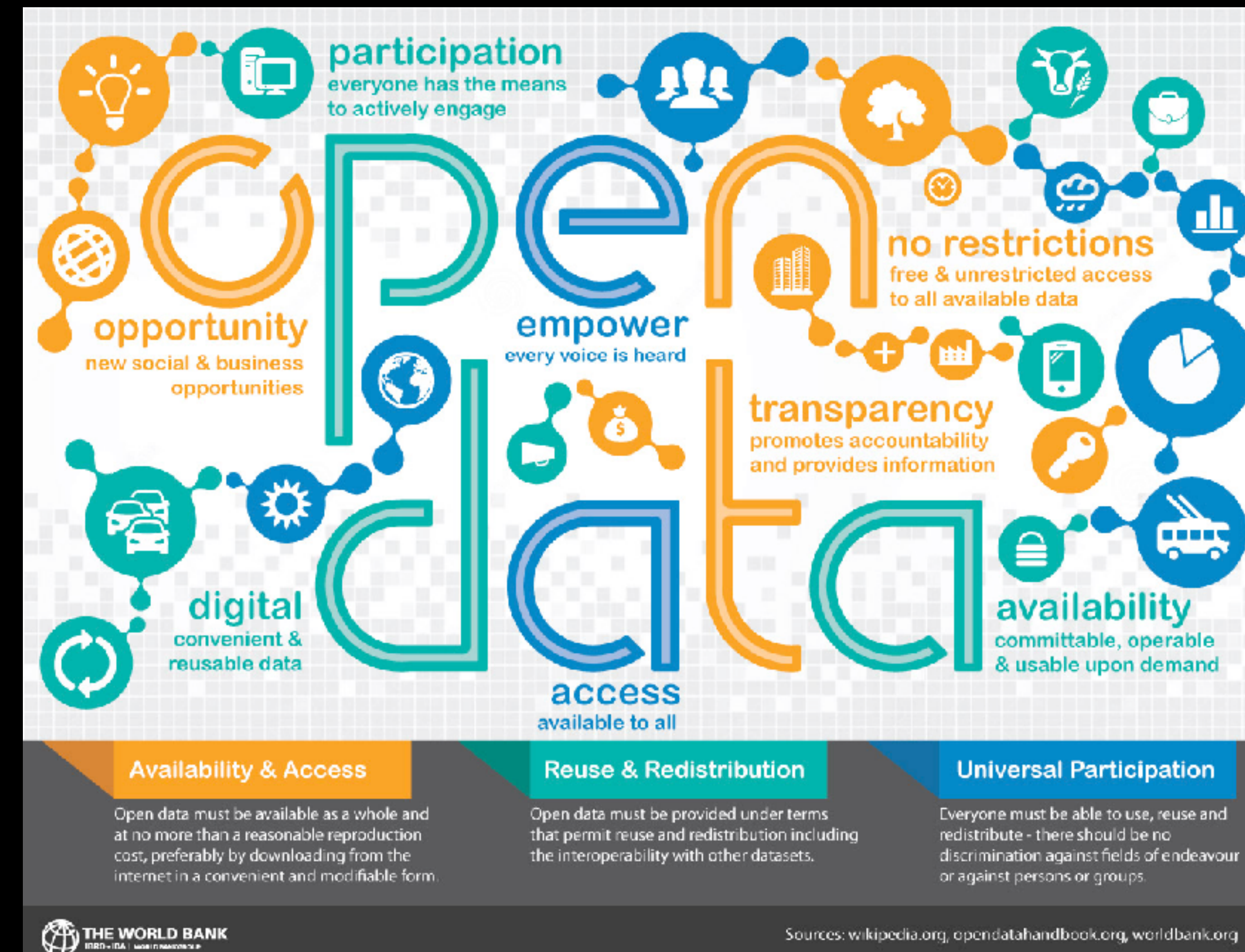
CROWD-SOURCED SOLUTIONS

FAIRNESS/OWNERSHIP

PUBLICITY

DISCOVERY

<http://www.worldbank.org>



Open Data ...why???

the open data paradigm has a variety of benefits that motivate data holders to open up access to their data:

TRANSPARENCY what's being collected?

REPRODUCIBILITY

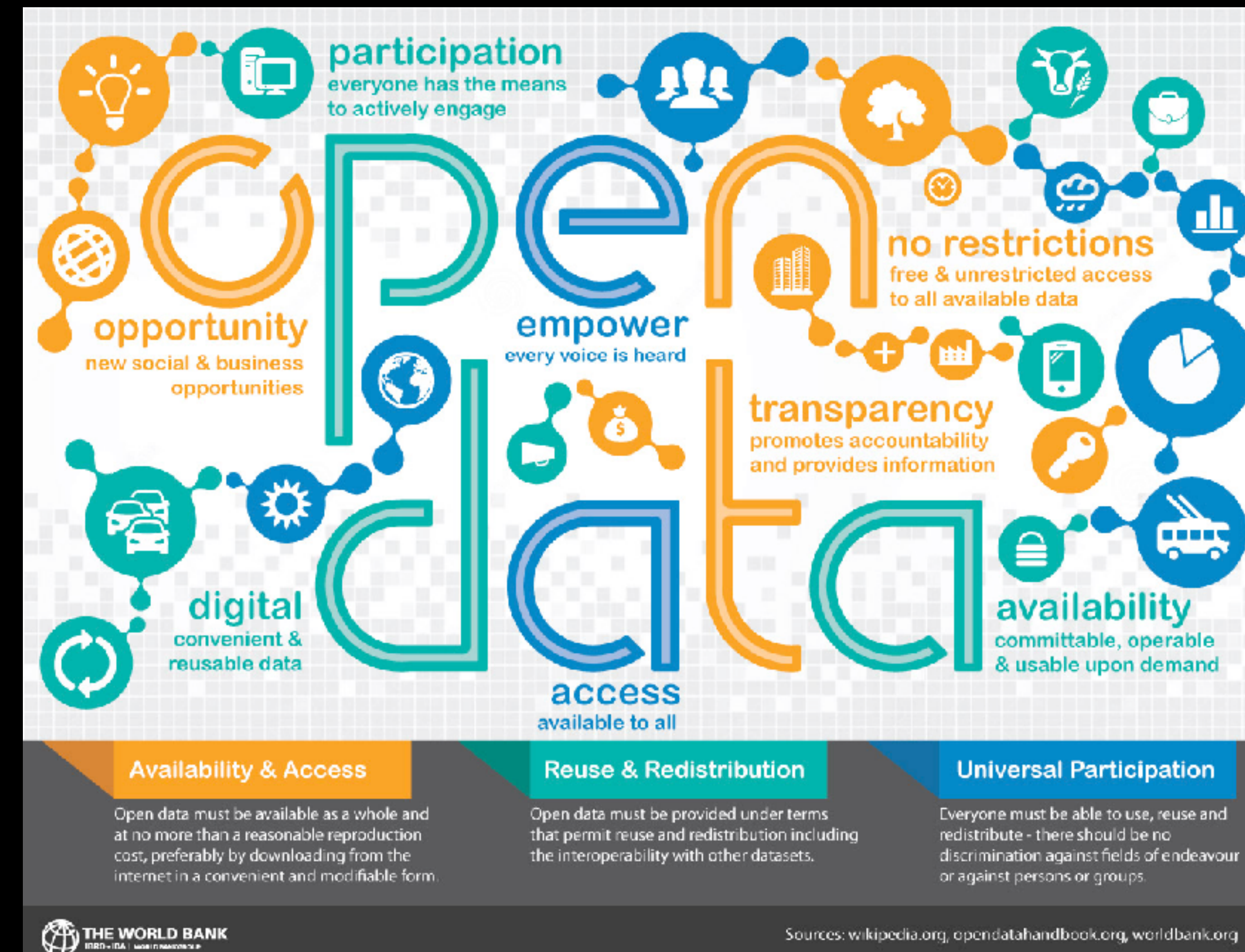
CROWD-SOURCED SOLUTIONS

FAIRNESS/OWNERSHIP

PUBLICITY

DISCOVERY

<http://www.worldbank.org>



Open Data ...why???

the open data paradigm has a variety of benefits that motivate data holders to open up access to their data:

TRANSPARENCY what's being collected?

REPRODUCIBILITY can results be validated?

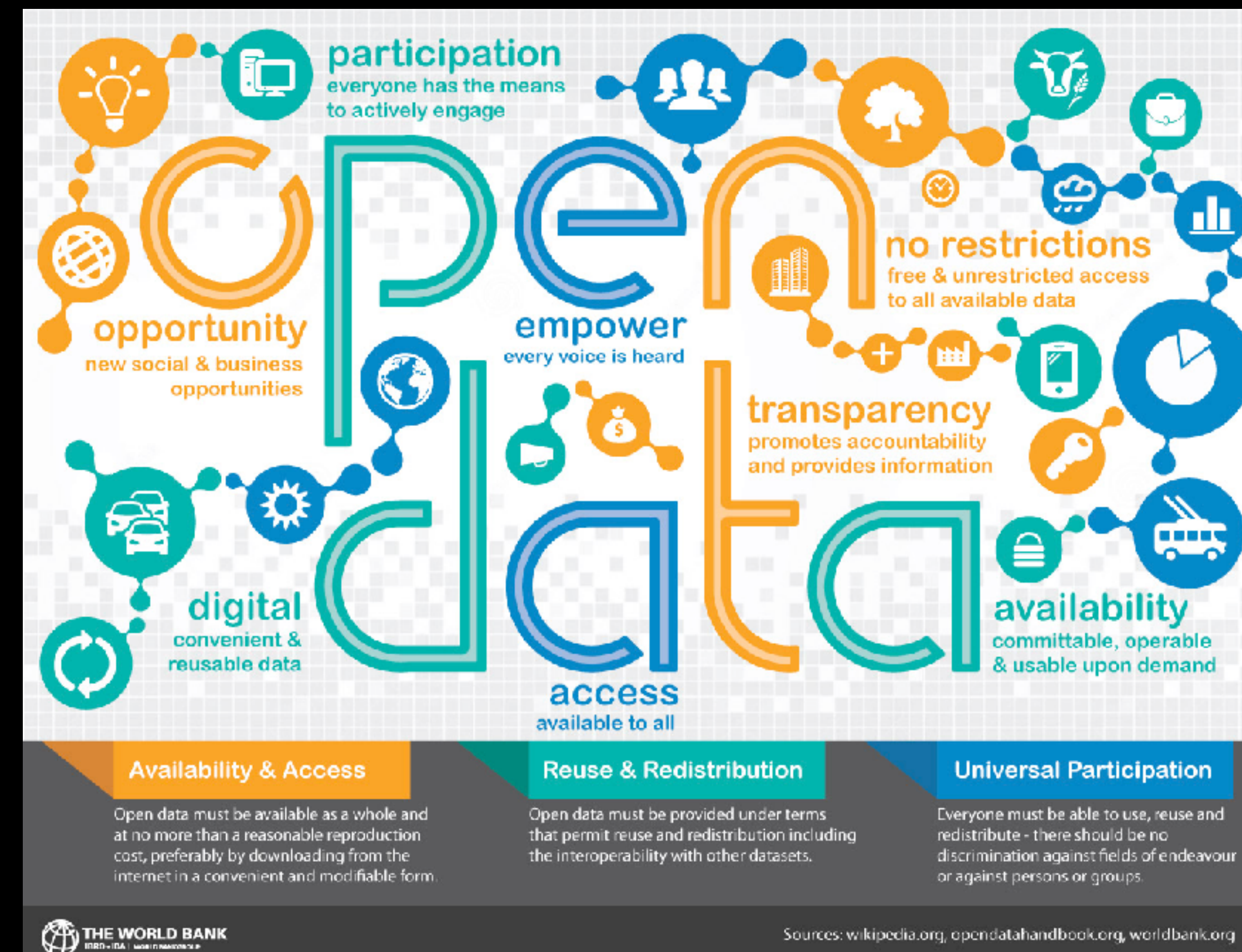
CROWD-SOURCED SOLUTIONS

FAIRNESS/OWNERSHIP

PUBLICITY

DISCOVERY

<http://www.worldbank.org>



Open Data ...why???

the open data paradigm has a variety of benefits that motivate data holders to open up access to their data:

TRANSPARENCY what's being collected?

REPRODUCIBILITY can results be validated?

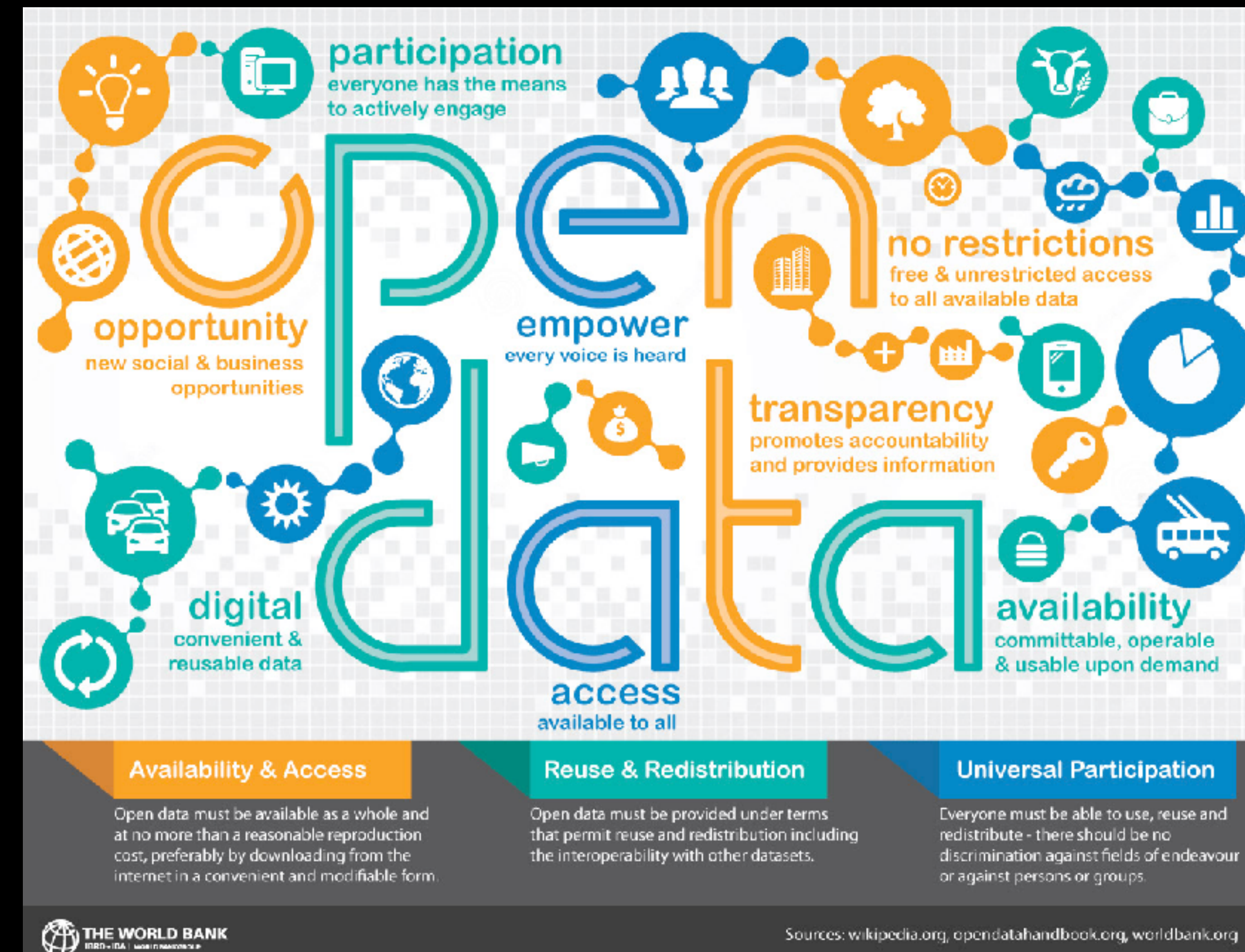
CROWD-SOURCED SOLUTIONS who has an answer?

FAIRNESS/OWNERSHIP

PUBLICITY

DISCOVERY

<http://www.worldbank.org>



Open Data ...why???

the open data paradigm has a variety of benefits that motivate data holders to open up access to their data:

TRANSPARENCY what's being collected?

REPRODUCIBILITY can results be validated?

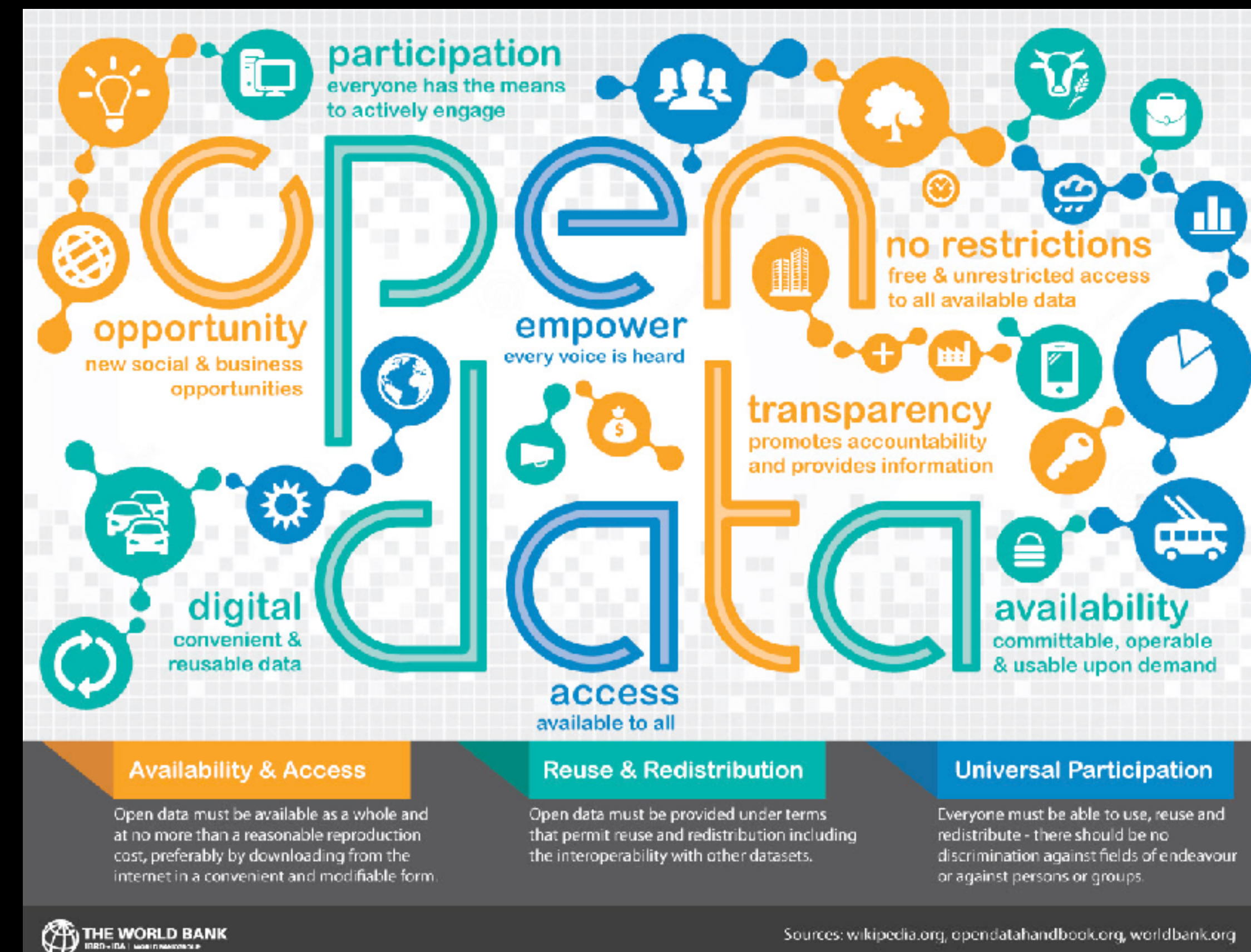
CROWD-SOURCED SOLUTIONS who has an answer?

FAIRNESS/OWNERSHIP whose data is it?

PUBLICITY

DISCOVERY

<http://www.worldbank.org>



Open Data ...why???

the open data paradigm has a variety of benefits that motivate data holders to open up access to their data:

TRANSPARENCY what's being collected?

REPRODUCIBILITY can results be validated?

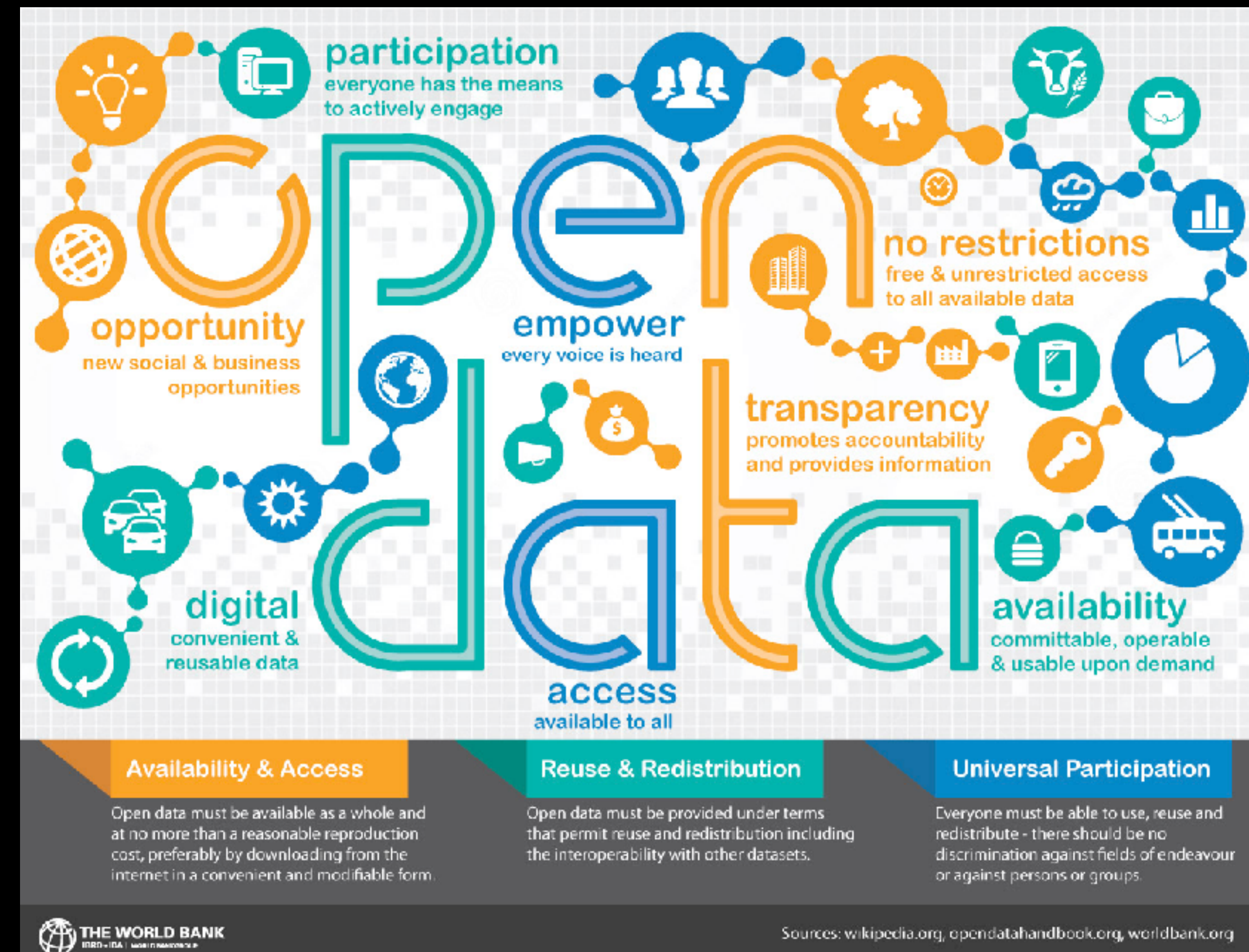
CROWD-SOURCED SOLUTIONS who has an answer?

FAIRNESS/OWNERSHIP whose data is it?

PUBLICITY can we get the community engaged?

DISCOVERY

<http://www.worldbank.org>



Open Data ...why???

the open data paradigm has a variety of benefits that motivate data holders to open up access to their data:

TRANSPARENCY what's being collected?

REPRODUCIBILITY can results be validated?

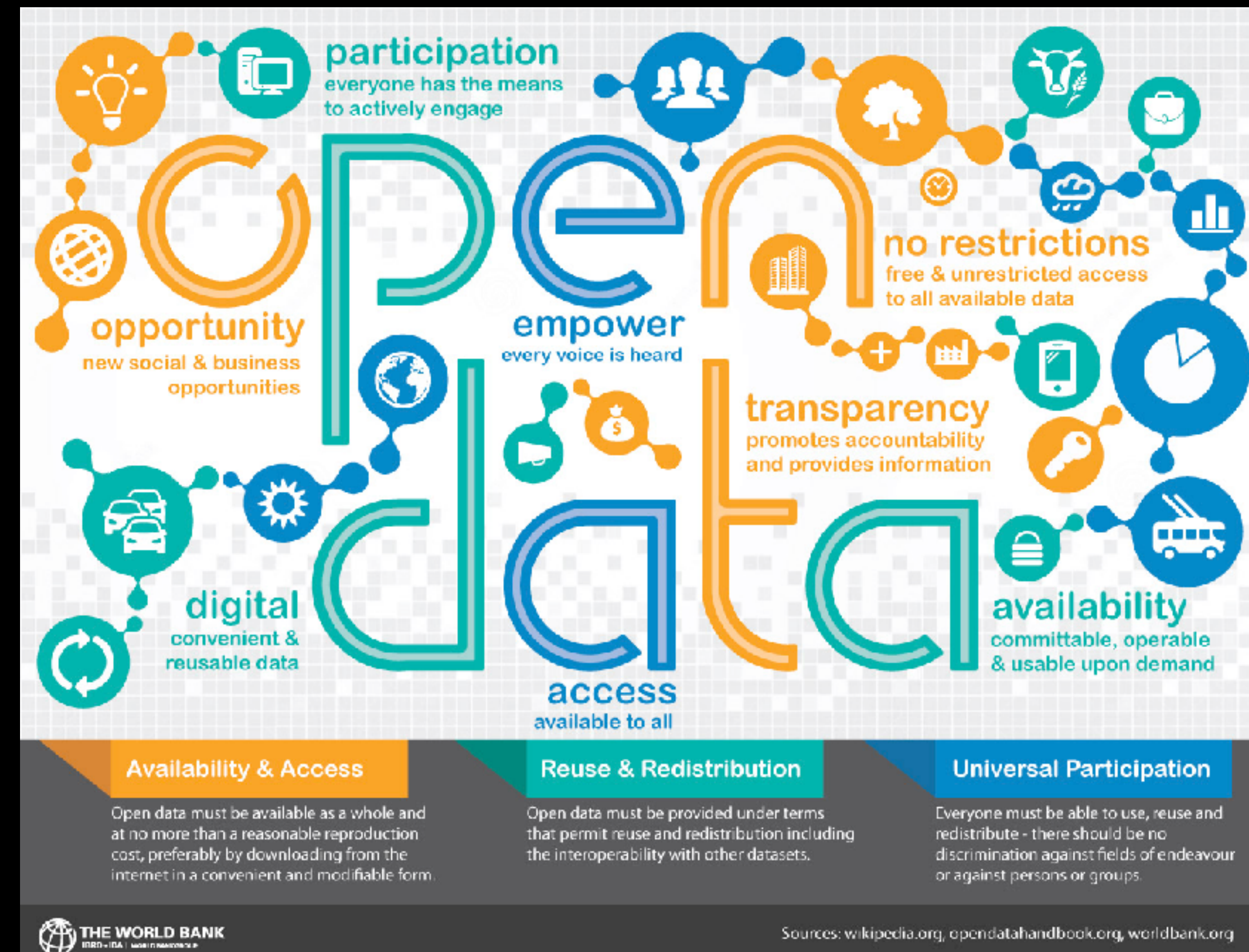
CROWD-SOURCED SOLUTIONS who has an answer?

FAIRNESS/OWNERSHIP whose data is it?

PUBLICITY can we get the community engaged?

DISCOVERY what have we missed?

<http://www.worldbank.org>



Open Data Purveyors

there are four distinct categories of **open data**

Public Sector

Private Sector

Academia

Individuals

Open Data Purveyors

there are four distinct categories of **open data**

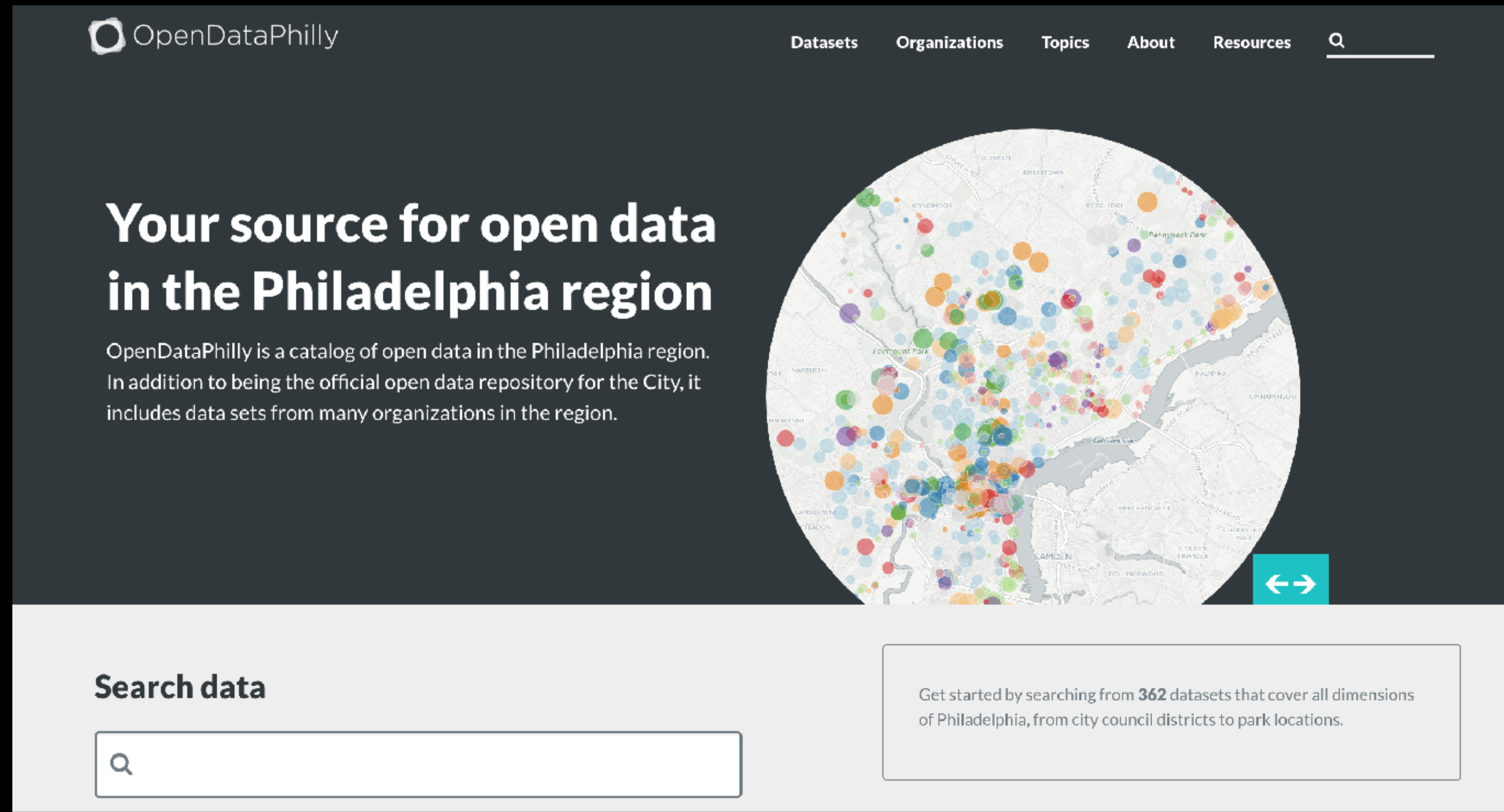
<http://www.opendataphilly.org>

Public Sector

Private Sector

Academia

Individuals



The screenshot shows the OpenDataPhilly website. The header includes the logo and navigation links: Datasets, Organizations, Topics, About, Resources, and a search icon. The main content area features the title "Your source for open data in the Philadelphia region" and a description: "OpenDataPhilly is a catalog of open data in the Philadelphia region. In addition to being the official open data repository for the City, it includes data sets from many organizations in the region." To the right is a circular map of Philadelphia with numerous colored dots representing data points. Below the map is a search bar with the placeholder text "Search data" and a search icon. A callout box on the right side of the page states: "Get started by searching from 362 datasets that cover all dimensions of Philadelphia, from city council districts to park locations."

Open Data Purveyors

there are four distinct categories of **open data**

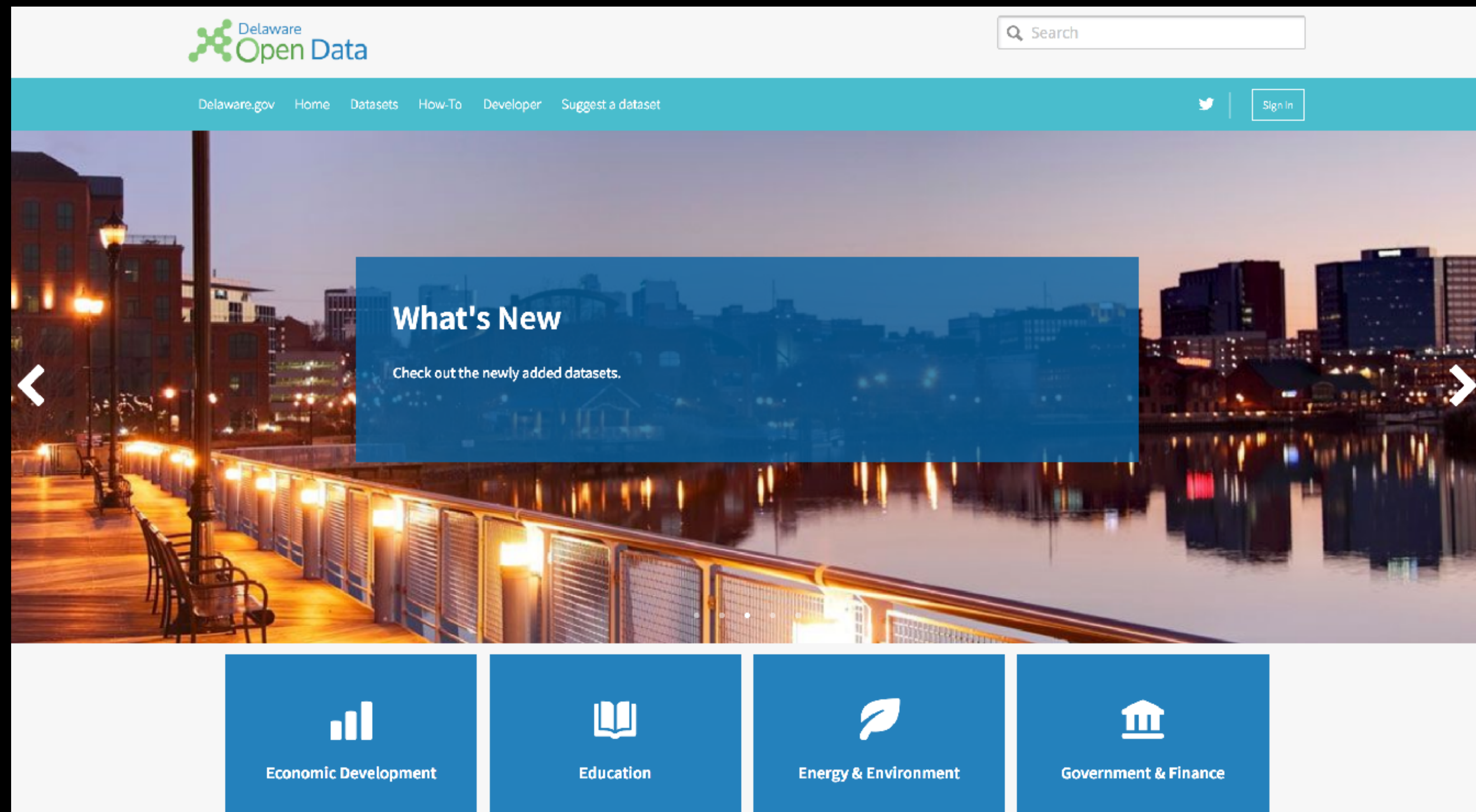
<http://data.delaware.gov>

Public Sector

Private Sector

Academia

Individuals



Open Source

Open Source

shares a similar ethos with the open data movement,
but is focused on software:

Open Source

shares a similar ethos with the open data movement,
but is focused on software:

free

available

reusable

Open Source

shares a similar ethos with the open data movement,
but is focused on software:

free

no licensing fee, but licenses

available

reusable

Open Source

shares a similar ethos with the open data movement,
but is focused on software:

free*

no licensing fee, but licenses

available

reusable

* but there are a lot of hairs to be split here...

Open Source

shares a similar ethos with the open data movement,
but is focused on software:

free*

no licensing fee, but licenses

available

hosted on a web server that is not restricted

reusable

* but there are a lot of hairs to be split here...

Open Source

shares a similar ethos with the open data movement,
but is focused on software:

free*

no licensing fee, but licenses

available

hosted on a web server that is not restricted

reusable

there are no ownership rights that prevent personal use

* but there are a lot of hairs to be split here...

Open Source

shares a similar ethos with the open data movement,
but is focused on software:

free*

no licensing fee, but licenses

available

hosted on a web server that is not restricted

reusable

there are no ownership rights that prevent personal use

with the similar **benefits:**

TRANSPARENCY

REPRODUCIBILITY

CROWD-SOURCED SOLUTIONS

FAIRNESS/OWNERSHIP

PUBLICITY

DISCOVERY

* but there are a lot of hairs to be split here...

Open Source

shares a similar ethos with the open data movement,
but is focused on software:

free*

no licensing fee, but licenses

available

hosted on a web server that is not restricted

reusable

there are no ownership rights that prevent personal use

with the similar **benefits:**

TRANSPARENCY

REPRODUCIBILITY

CROWD-SOURCED SOLUTIONS

FAIRNESS/OWNERSHIP



PUBLICITY

DISCOVERY

* but there are a lot of hairs to be split here...

Handling URLs as a means of accessing data

Accessing data from URLs has several **benefits**:

Handling URLs as a means of accessing data

Accessing data from URLs has several **benefits**:

- . no need to store data locally if not necessary
- . ability to run real time data analysis updates
- . sharing analysis between collaborators without sharing data
- . ability to remotely query large databases

Handling URLs as a means of accessing data

Accessing data from URLs has several **benefits**:

- . no need to store data locally if not necessary
- . ability to run real time data analysis updates
- . sharing analysis between collaborators without sharing data
- . ability to remotely query large databases

... but also several **disadvantages**:

Handling URLs as a means of accessing data

Accessing data from URLs has several **benefits**:

- . no need to store data locally if not necessary
- . ability to run real time data analysis updates
- . sharing analysis between collaborators without sharing data
- . ability to remotely query large databases

... but also several **disadvantages**:

- . changing data and/or meta data formats
- . server downtimes
- . throttling
- . multiple access paradigms

Handling URLs as a means of accessing data

Accessing data from URLs has several **benefits**:

- . no need to store data locally if not necessary
- . ability to run real time data analysis updates
- . sharing analysis between collaborators without sharing data
- . ability to remotely query large databases

... but also several **disadvantages**:

- . changing data and/or meta data formats
- . server downtimes
- . throttling
- . multiple access paradigms

There are many ways to access data in this way with jupyter; three of the most common are:

- . wget (and curl)
- . urllib
- . API's (application programming interface)

Handling URLs as a means of accessing data

Accessing data from URLs has several **benefits**:

- . no need to store data locally if not necessary
- . ability to run real time data analysis updates
- . sharing analysis between collaborators without sharing data
- . ability to remotely query large databases

... but also several **disadvantages**:

- . changing data and/or meta data formats
- . server downtimes
- . throttling
- . multiple access paradigms

There are many ways to access data in this way with jupyter; three of the most common are:

- . wget (and curl)
- . urllib
- . API's (application programming interface)

let's start with wget, the most basic method...