AN ILLUSTRATED GUIDE

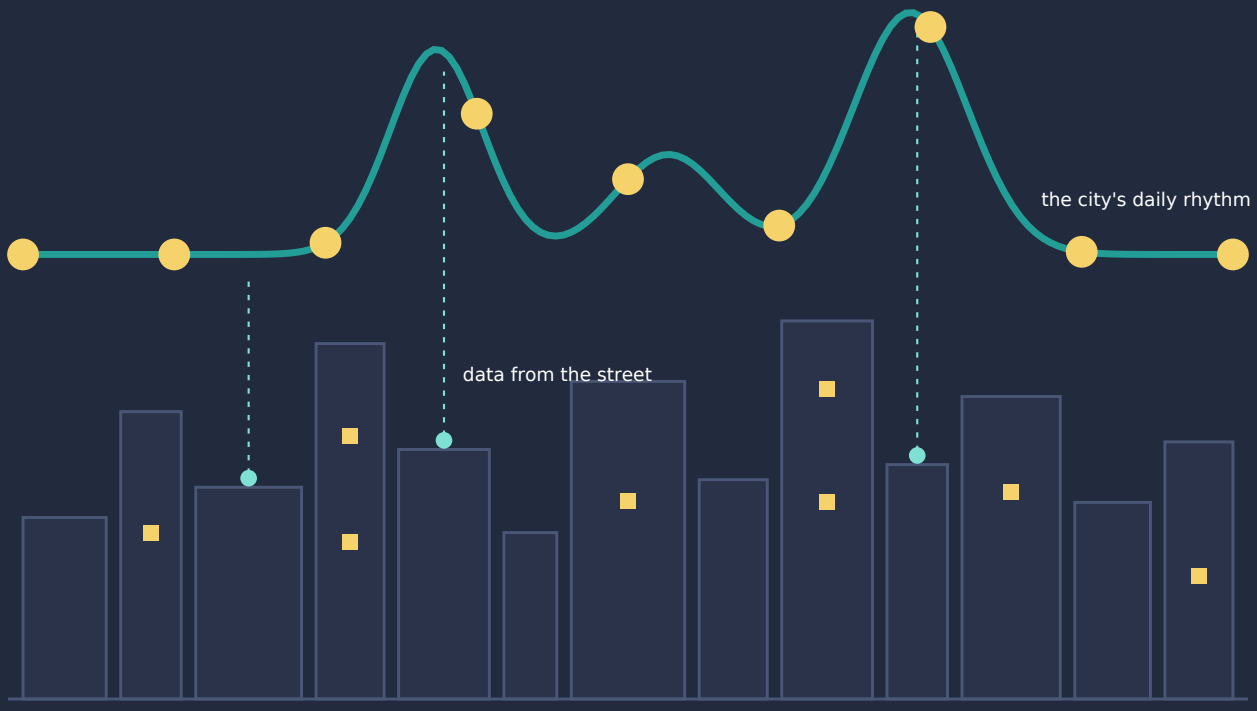
Urban Informatics

A friendly guide to understanding cities with data, computing and people

Plain English, pictures first, real worked examples with real numbers, and a glossary, a quiz and an eight-week plan.

Fourteen short chapters · Quiz · Glossary · Eight-week plan

Prompted and directed by Mahmoud A.
Written and illustrated by Claude (Anthropic)



How to Use This Guide

Read it in order, or jump to what you need.

This guide explains **urban informatics**, which means using data, computing and people's own knowledge to understand cities and help them work better. You need no coding and no maths beyond school level. Every idea starts with a plain explanation and a picture. The worked examples use small made-up datasets, so you can see every number.

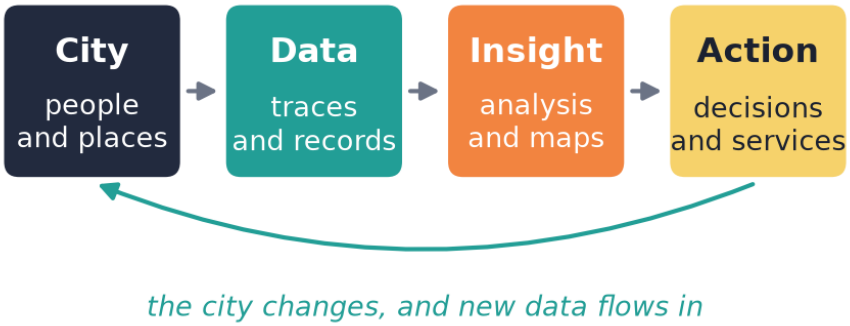


Figure 1. The whole guide in one picture. Data flows from the city to insight and action, and the city changes in return.

Who it is for

- Curious people who hear words like **smart city** or **open data** and want to know what they mean.
- Planners, architects, engineers, designers, public servants and community groups who work with information about places.
- Anyone who has to judge whether a city data project is useful, fair and safe.

How each chapter works

- A **banner** gives the topic in one sentence.
- Plain explanation first, then a picture, then an example.
- Boxes: **TIP** (a shortcut), **SAY IT LIKE THIS** (a phrase you can use out loud), **WATCH OUT** (a common trap) and **IN ONE LINE** (the chapter in a sentence).
- At the back: myths and facts, a quiz, a glossary, a cheat sheet and further reading. Chapter 14 ends with an eight-week plan.

A note on the numbers

Every number in the worked examples was computed by the script that built this guide. The data are **synthetic**, made by a computer from simple rules plus some randomness. Because we know the rules, we can check whether each method finds the truth. Real data are messier, and Chapter 13 lists the main traps.

Contents

Fourteen short chapters and a toolbox at the back.

1 **What Is Urban Informatics?**
The big idea, and how it relates to its neighbours.

2 **The City as a Data Generator**
Every tap leaves a trace, and every source has blind spots.

3 **The Data Pipeline**
Collect, clean, store, analyse, visualise, act.

4 **Formats, Access and Open Data**
CSV, JSON, GeoJSON, APIs and databases in plain words.

5 **Time and Space Together**
Daily rhythms, odd hours and patterns that move.

6 **Indicators and Dashboards**
A few honest numbers that track what matters.

7 **Machine Learning for Cities**
Sorting, grouping and predicting, with honest limits.

8 **Worked Example: Grouping Neighbourhoods**
k-means clustering, built from scratch.

9 **Mobility and Movement**
Origin-destination matrices and trip patterns.

10 **People, Participation and Civic Tech**
Working with residents as well as studying them.

11 **Ethics, Privacy and Data Governance**
Who is watched, who decides, and who benefits.

12 **Communicating with Data and Maps**
Charts and maps that people understand and trust.

13 **Pitfalls and Honest Limits**
Traps that catch beginners and experts alike.

14 **Tools and Try It Yourself**
A short recipe, open tools and an eight-week plan.

At the back: Myths and Facts · Quick Quiz · Quiz Answers · Glossary · One-Page Cheat Sheet · Further Reading

1 What Is Urban Informatics?

The big idea, and how it relates to its neighbours.

Cities are made of people, places and the flows between them. Every day, people travel, work, shop, play and complain. Much of this now leaves a digital record. **Urban informatics** is the field that turns those records into understanding, and understanding into better decisions.

The word *informatics* means the study of how information is collected, organised and used. Urban informatics sits where three things meet (Figure 2). The first is data and computing. The second is people and communities. The third is the physical city of streets, buildings and parks. Take away any one of them and something important goes missing. Data without people misses how life feels on the ground. People without data struggle to see the whole city at once. Both need the physical setting to make sense.

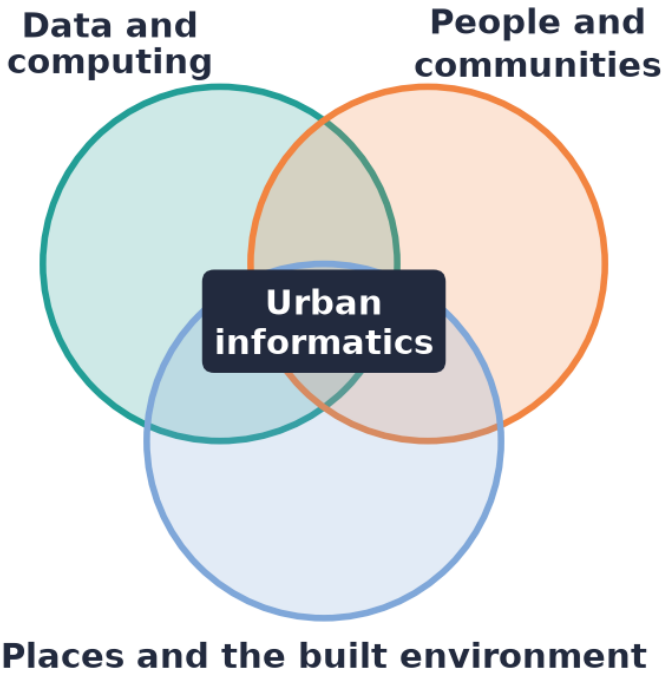


Figure 2. Urban informatics sits where data and computing, people, and places meet.

> **SAY IT LIKE THIS**
> Urban informatics means using data and computing, together with what residents know, to understand how a city works and to help it work better for the people in it.

Its neighbours

Several fields overlap with urban informatics. People use the labels loosely, so treat this table as a rough map.

Field	Main question	Typical output
Urban informatics	How can data, computing and local knowledge help us understand and improve a city?	Datasets, analyses, tools, maps, dashboards
Urban science	Are there general patterns that hold across many cities?	Models and theories, for example how things change with city size
Smart cities	How can digital technology run city services better?	Sensors, platforms, connected infrastructure
GIS (geographic information systems)	Where are things, and how do places relate?	Maps, spatial analysis, location databases
Urban planning	What should the city become, and how do we get there?	Plans, policies, rules and designs
Civic tech	How can technology help residents take part?	Apps, platforms, open tools

Think of GIS as a toolbox, urban science as a search for patterns, and smart cities as a set of technologies. Urban informatics borrows from all three. It adds a strong interest in who makes the data, who uses it and who benefits.

TIP

When someone says *smart city*, ask which problem it solves and for whom. A good answer names a real service and real people.

IN ONE LINE

Urban informatics joins data, computing and people's knowledge to understand cities and improve them.

2 The City as a Data Generator

Every tap leaves a trace, and every source has blind spots.

Picture one ordinary weekday. A **smart meter**, an electricity meter that sends its readings automatically, records how much power your home uses every half hour or so. You tap a card as you board the bus. You walk past a **footfall counter**, a small sensor that counts people walking by. You pay for lunch by card and post a photo tagged with its location. In the evening you report a pothole in an app. Your phone talks to the mobile network all day long. By bedtime you have left dozens of digital **traces**, records that something happened at a certain time and often at a certain place. So has almost everyone else.

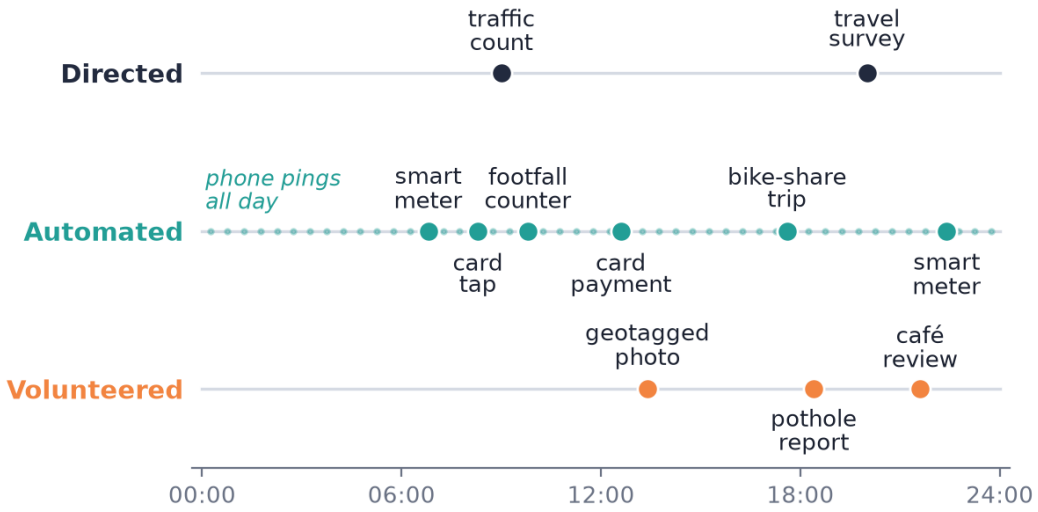


Figure 3. One person's made-up day, sorted by how each trace was produced.

Three ways city data is made

Kind	How it is made	Examples	Strength	Weakness
Directed	Someone collects it on purpose, with a method	Census, travel surveys, inspections, manual traffic counts	Clear design and known coverage	Slow and costly; often years old
Automated	Machines record it as a side effect of a service	Card taps, sensors, phone network records, meters	Fast, detailed, continuous	Made for billing or operations; gaps are hidden
Volunteered	People share it themselves	Map edits, reports, posts, reviews, citizen science	Local knowledge and rich detail	Who takes part shapes what you see

This three-way split is a typology from the geographer Rob Kitchin. People also contrast **designed data**, made to answer a question, with **found data**, which was made for another reason and reused. Found data is cheap and plentiful. But you also inherit someone else's choices about what to record and what to leave out.

Nine sources and their blind spots

Every dataset is a window onto the city, and each window faces one way. The table lists the main sources, what each captures and its usual **blind spot**, meaning the people or places it under-counts or misses.

Source	What it captures	Usual blind spot
Sensors and IoT (the Internet of Things, everyday devices on a network)	Air quality, noise, traffic, footfall, energy, water	Only where sensors sit; outages and drift; cheap sensors can be inaccurate
Open government data	Permits, budgets, service requests, land use, crime reports	Shows what agencies record and choose to publish; definitions change
Transit smart cards	Where and when people board, sometimes where they leave	Cash payers, walkers, cyclists and drivers; often no destination
GPS and mobile phone traces	How devices move across the day	People without phones; shared devices; data often held by private firms
Social media and reviews	Opinions, events, popular places	Skews young and online; bots; people post what feels worth sharing
Street-view imagery	Street design, trees, pavements, building condition	One day's snapshot, often years old; main roads favoured
Satellites and remote sensing (measuring from far away)	Land cover, heat, night lights, growth	Clouds; cannot see indoors or under trees; limited detail
OpenStreetMap (a free world map edited by volunteers)	Roads, buildings, shops, paths	Detail depends on local volunteers, so it varies by area
Surveys and citizen science	Travel habits, opinions, local observations	Who answers; small samples; question wording

A worked example: who is in the data?

Suppose a city wants to know where people spend time and uses **geotagged** social media posts, meaning posts that carry a location. We built a made-up city of 100,000 residents aged 15 and over and gave each age group a different chance of posting. We invented those chances for the example. The script then simulated one month, in which 16,261 people wrote 78,257 posts.

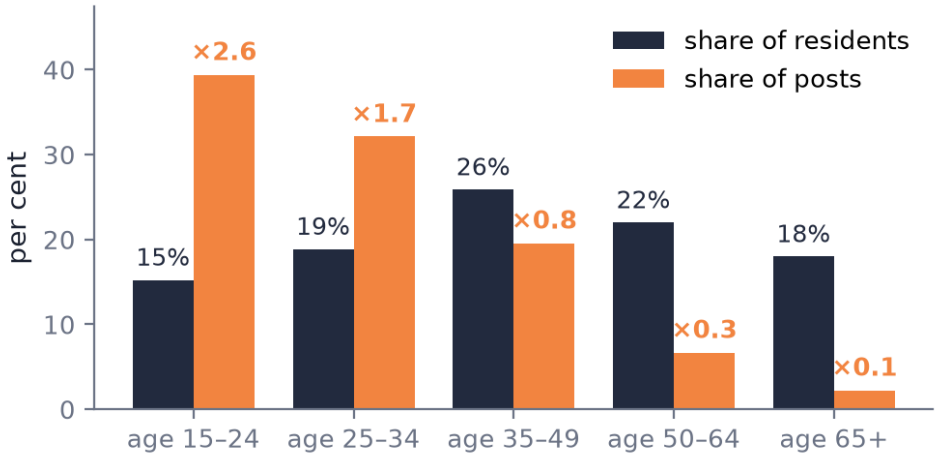


Figure 4. Share of residents and share of posts by age group in the made-up city. The orange label is how over- or under-represented each group is. Computed for this guide.

People aged 15 to 34 are 34% of residents but wrote 72% of the posts. People aged 65 and over are 18% of residents but wrote only 2.2%. A map built from these posts would mostly show where young people go. It would be easy to mistake that for where *people* go.

The usual remedy is to ask who is missing, then compare with a source that tries to count everyone, such as a census. You can then **weight** the data, giving each under-represented person a little more say, or combine several sources.

WATCH OUT

Big data is often biased data. A million records from the same kind of person still do not make a fair sample of the city. And a trace shows that something happened, rarely why.

IN ONE LINE

A city produces data all day; learn how each dataset was made, and who it misses, before you trust what it says.

3 The Data Pipeline

Collect, clean, store, analyse, visualise, act.

Raw data rarely arrives ready to use. It flows through a series of steps called a **pipeline**. Each step has its own jobs and its own ways to go wrong.

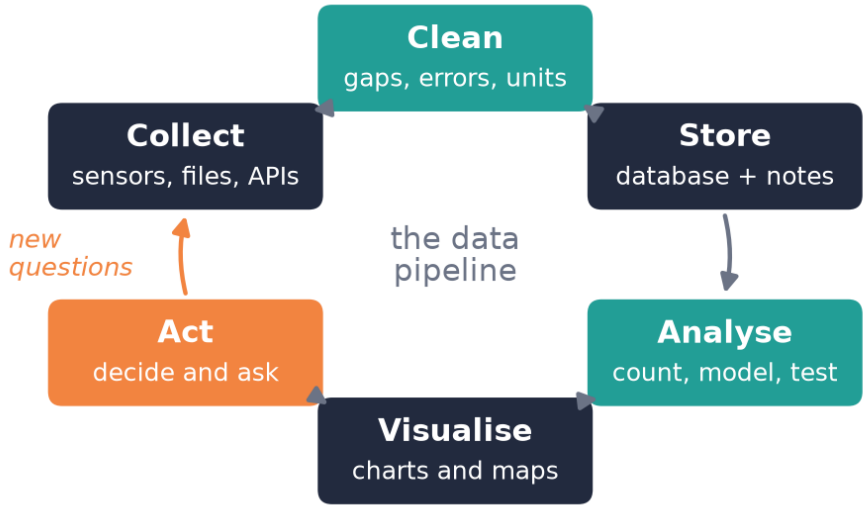


Figure 5. The data pipeline. Acting on results raises new questions, so the loop starts again.

- **Collect.** Gather data from sensors, files, websites and APIs (ways for one program to ask another for data; see Chapter 4).
- **Clean.** Fix or remove errors, gaps, duplicates and mixed units.
- **Store.** Keep data safe and organised, often in a database, with **metadata**: notes on who made it, when, how, and what each column and unit means.
- **Analyse.** Count, compare, model and test ideas.
- **Visualise.** Turn results into charts and maps that people can read.
- **Act.** Make a decision, change a service or ask a better question.

A worked example: cleaning a week of sensor data

We made a week of hourly temperature readings from 10 street sensors, 1,680 readings in all. Then we spoiled them in four common ways. Some rows were sent twice. Some readings were blank. A few showed −999, a code some devices use to mean *no reading*. One sensor reported in degrees Fahrenheit while the others used Celsius. The table shows what the cleaning script found.

Step	What the script found	Rows left
1. Load the raw file	1,760 rows	1,760
2. Remove duplicate rows	80 rows sent twice	1,680
3. Turn −999 codes into blanks	12 codes	1,680
4. Convert the odd sensor to Celsius	1 sensor far warmer than the rest; 162 readings converted	1,680
5. Drop blank readings	57 blanks (including the old codes)	1,623

Before cleaning, the average of every number in the file was 14.2 °C. After cleaning it was 16.9 °C. Because we made the data, we know the true average of all the original readings was 16.9 °C. The −999 codes dragged the raw average down and the Fahrenheit sensor pushed it up. Neither problem shows in the average alone.

TIP

Before any analysis, look at the smallest value, the largest value and the number of blanks in each column. These three checks catch most problems.

IN ONE LINE

Most of the work is cleaning and checking; a small slip early on quietly changes every answer later.

4 Formats, Access and Open Data

CSV, JSON, GeoJSON, APIs and databases in plain words.

Data has to be written down in a **format**, an agreed layout that computers can read. Here is one made-up bus stop written three ways.

CSV (a plain table)

```
stop_id,name,lat,lon,shelter
1042,Market Square,45.6789,10.1234,yes
```

JSON (labels and values)

```
{"stop_id": 1042,
 "name": "Market Square",
 "lat": 45.6789,
 "lon": 10.1234,
 "shelter": true}
```

GeoJSON (JSON with a shape)

```
{"type": "Feature",
 "geometry": {
   "type": "Point",
   "coordinates": [10.1234, 45.6789]
 },
 "properties": {
   "stop_id": 1042,
   "name": "Market Square",
   "shelter": true
 }
}
```

One made-up stop in three formats. GeoJSON puts longitude before latitude.

Format	What it is	Good for	Watch out
CSV (comma-separated values)	One row per line, commas between columns	Spreadsheets, quick sharing	No fixed way to state units or types; stray commas break rows
JSON	Nested labels and values	Web services and APIs	Can be hard to flatten into a table
GeoJSON	JSON with points, lines and polygons	Web maps	Longitude first; big files
GeoPackage, shapefile	Common GIS files for shapes plus attributes	Desktop GIS	A shapefile is several files that must travel together
Database, e.g. PostgreSQL with PostGIS, an add-on that lets the database handle maps	A managed store you query with SQL, a language for asking questions of tables	Large, shared, changing data	Needs setting up and looking after

An **API** (application programming interface) is a door that lets one program ask another for data. You send a request, often a web address with a few options, and get an answer back, usually in JSON. Many transport operators and open data portals offer APIs. Most have usage limits and terms, so read them first.

A place on Earth is usually given as **latitude** (north–south) and **longitude** (east–west) in degrees. Maps also use other **coordinate reference systems**, rules for turning the curved Earth into flat numbers. Mixing them up is a classic mistake

that drops points in the sea.

Open data and standards

Open data is data that anyone can use, change and share for any purpose, at most with a requirement to credit the source or to share it on the same terms. It needs a clear open **licence**, the legal permission to reuse it, and a format machines can read. A popular five-star scale, proposed by the web's inventor Tim Berners-Lee, starts lower. A scanned page earns one star for being openly licensed, and each further star makes the data easier to reuse.

Stars	What it means
★	On the web under an open licence, in any format, even a scanned page
★★	Structured, so a machine can read it, such as a spreadsheet instead of a picture of one
★★★	In an open format that does not need one company's software, such as CSV
★★★★	Uses web addresses to name things, so others can point to them
★★★★★	Links to other people's data to give context

Standards are shared agreements about format and meaning, so data from different places fits together. Common ones in cities include **GTFS** (General Transit Feed Specification) for public transport stops and timetables, GeoJSON for map features, and the open standards of the Open Geospatial Consortium, such as SensorThings API for sensor data and CityGML for 3D city models.

> **IN ONE LINE**
> Choose formats others can open, publish with a clear licence and good metadata, and follow shared standards so your data can join up with other people's.

5 Time and Space Together

Daily rhythms, odd hours and patterns that move.

Cities have a pulse. Streets fill in the morning and empty at night. Measurements taken at regular times, such as every hour, form a **time series**. Reading one starts with its rhythm.

A worked example: a pedestrian counter

We simulated an hourly footfall counter on a shopping street for five weeks. The first four weeks set the usual pattern. For each hour of the week, the script averages that same hour across those four weeks. The result is the **baseline**, what we expect to see. The script also measures how far counts normally wobble around the baseline. In week five, any hour outside three times the normal wobble is flagged as an **anomaly**, something unusual worth checking.

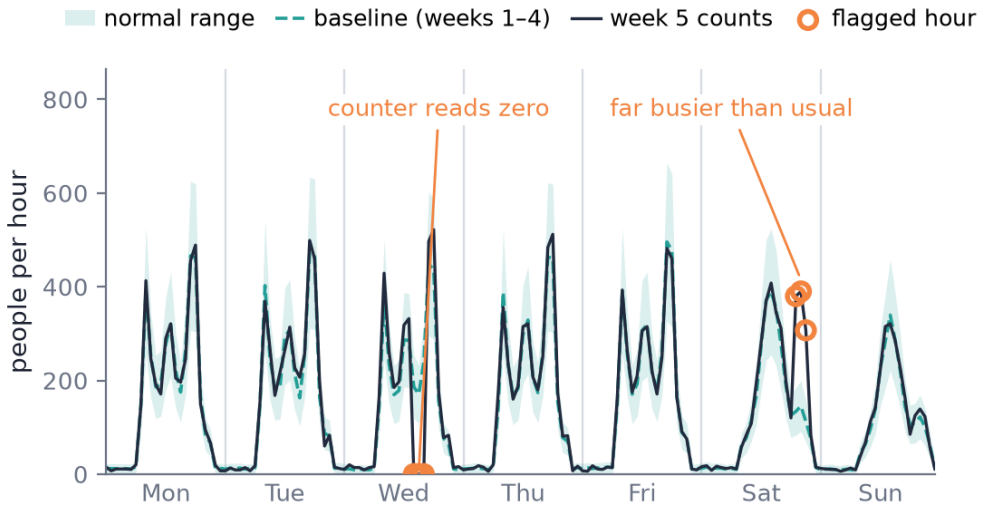


Figure 6. Week five of the made-up footfall counter against its baseline. Orange circles are the hours the script flagged. Computed for this guide.

In the four normal weeks, weekdays peak at about 08:00 with 389 people an hour, and again at 17:00 with 469. Weekends have one broad peak around 14:00. In week five we hid two problems on purpose, a counter failure and a busy evening event. The script flagged 6 hours, exactly those two problems and nothing else.

Hour	Counted	Expected	Likely reason
Wednesday 14:00	0	186	Counter may have failed
Wednesday 15:00	0	171	Counter may have failed
Wednesday 16:00	0	239	Counter may have failed
Saturday 19:00	380	134	Far busier; perhaps an event
Saturday 20:00	390	145	Far busier; perhaps an event
Saturday 21:00	307	112	Far busier; perhaps an event

WATCH OUT

A flag only says *look here*. Someone still has to check whether it was a broken sensor, a holiday, a road closure or a real change.

Adding space

Most urban data also has a place. Tracking both shows patterns that move. Figure 7 uses a simple model of a made-up city of 300,000 people. It shows how many people are present in four rings, from the centre to the edge, hour by hour on a weekday.

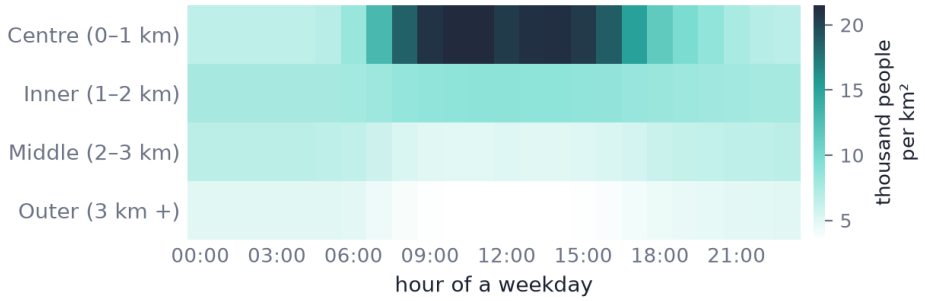


Figure 7. People present per square kilometre, by ring and hour, in a simple model of a made-up city. Computed for this guide.

At 03:00 the centre holds 6% of everyone, and at 12:00 it holds 21%. Its density climbs from about 6.5 thousand people per km² at 03:00 to a peak of 21.5 thousand at 10:00. The outer ring does the reverse, and a single daily average would hide the swing.

Nearby places also tend to be alike, so busy streets sit next to busy streets. Geographers call this **spatial autocorrelation**, and hot spots are one sign of it. It also means nearby readings are not independent: ten sensors on one street tell you less than ten spread across the city. Because of the daily rhythm, compare places at the same hour and on the same kind of day.

IN ONE LINE

Learn the usual rhythm first; then odd hours and moving patterns stand out.

6 Indicators and Dashboards

A few honest numbers that track what matters.

An **indicator** is a number chosen to track something we care about, such as how often buses run on time or how many homes are near a park. A **dashboard** puts a handful of indicators on one screen, often updated automatically. Figure 8 is a mock-up. Every number on it was computed by our script from made-up data.

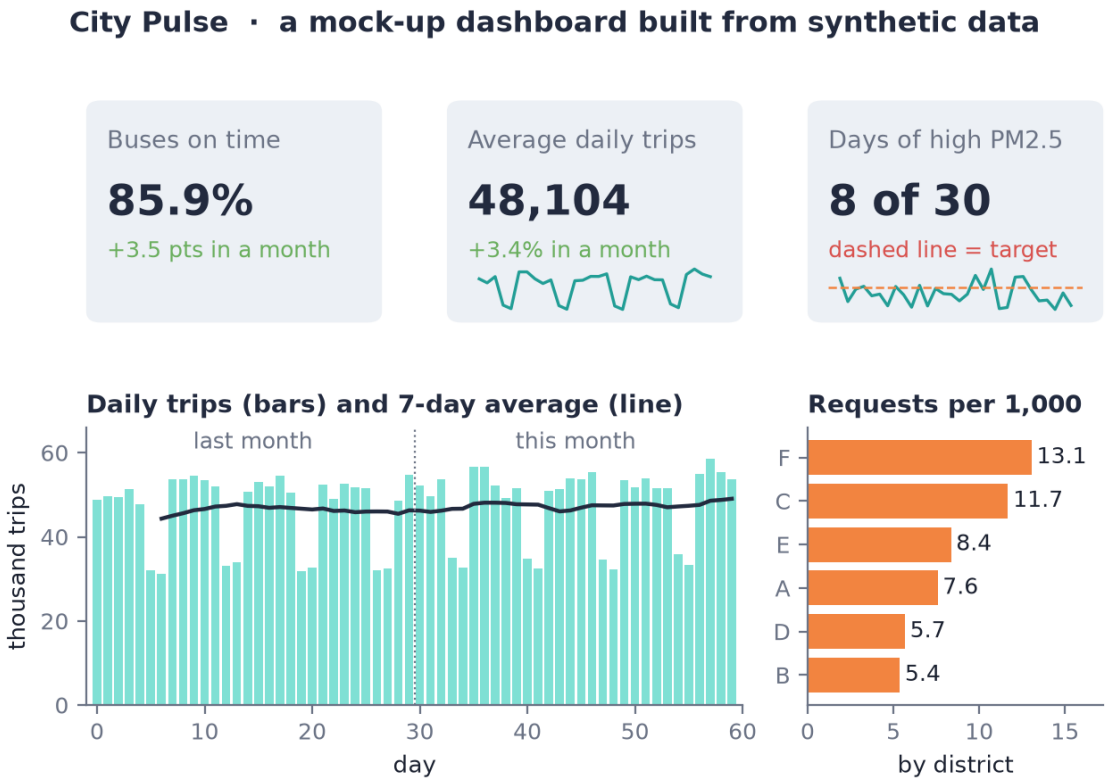


Figure 8. A mock-up dashboard. All values come from synthetic data generated for this guide.

Take the first tile. It counts a bus as on time if it left less than five minutes late. Over the last 30 days that was 85.9% of 36,000 simulated departures, against 82.5% the month before. The tile is only useful because that definition is written down. Change five minutes to three and the number changes too. The third tile counts days when fine dust in the air (PM2.5) averaged more than 15 micrograms per cubic metre, the dashed line on its small chart.

What makes a good indicator

Quality	Ask yourself
Relevant	Does it track something people actually care about?
Clearly defined	Would two analysts compute the same number from the same data?
Reliable source	Is the data stable, and will it still exist next year?
Timely	Is it updated often enough to act on?
Comparable	Can we compare it across places and over time?
Broken down	Can we split it by area, age, income or gender to see who is left behind?
Actionable	If it moves, does someone know what to do?
Honest	Does it show uncertainty, gaps and the date of the data?

Shared frameworks help with definitions. The international standard ISO 37120 lists indicators for city services and quality of life, and the United Nations Sustainable Development Goal 11 covers sustainable cities. Many cities start from lists like these and adapt them.

Keep in mind **Goodhart's law**. It warns that when a measure becomes a target, people find ways to hit the target without improving the real thing. If buses are judged only on punctuality, timetables can be quietly loosened.

> **SAY IT LIKE THIS**
> This tile tells us what share of buses left on time. It does not tell us how crowded they were or who could not get on.

> **IN ONE LINE**
> Pick a few indicators that are clear, fair and actionable, show their limits, and watch for gaming.

7

Machine Learning for Cities

Sorting, grouping and predicting, with honest limits.

Machine learning lets a computer find patterns from examples instead of being given the rules. In cities it does three main jobs.

Job	Plain meaning	City example	Honest limit
Classification	Put each item in a category, learned from labelled examples	Spot potholes or missing pavements in street images; label land use from satellite pictures	Only as good as its labels; may fail in places unlike its training images
Clustering	Group similar items when no labels exist	Find types of neighbourhood from census-style profiles (Chapter 8)	Groups depend on the variables chosen and how many groups you ask for
Regression (predicting a number)	Estimate a number from other numbers	Forecast bike-share demand, energy use or traffic	Learns the past; struggles when conditions change

A worked example: predicting bike-share trips

We made a year of daily bike-share trips that depend on temperature, rain and weekends, plus random noise. The script set aside about one day in five at random, 67 days, as a **test set**: data the model never sees while learning, so we can check it honestly. The model is a simple formula, a curve for temperature plus adjustments for rain and weekends. It is fitted by **least squares**, which picks the numbers that make the squared errors as small as possible.

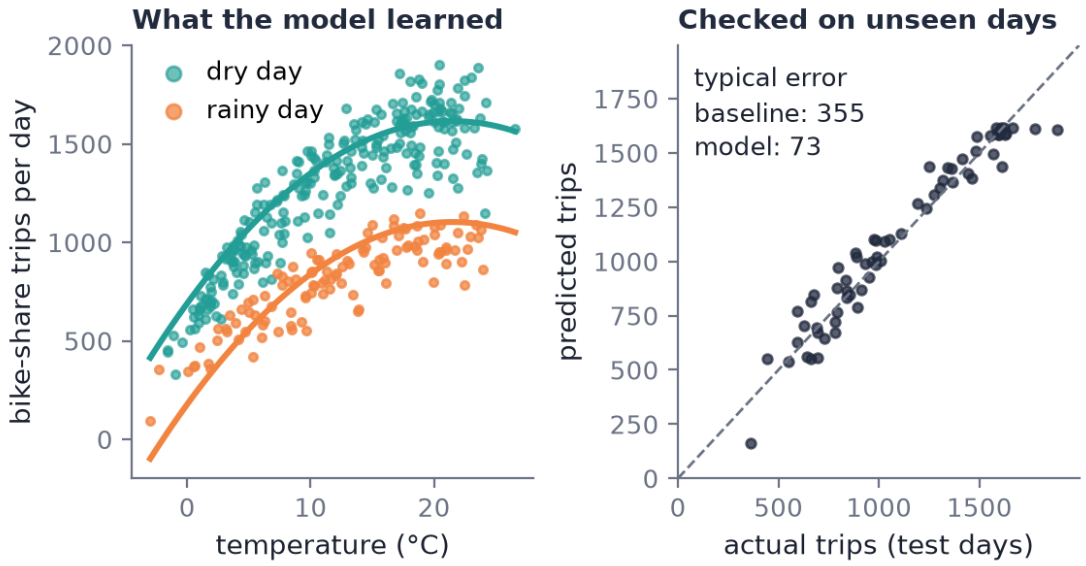


Figure 9. Left: the year of made-up data and the curves the model learned. Right: predictions for the test days against what happened. Computed for this guide.

Always compare with a **baseline**, the simplest sensible guess. Here the baseline predicts the average day, every day. On the test days it was off by 355 trips a day on average. The model was off by 73. That is a real improvement, and we can trust its size because it was measured on unseen days. The model also learned that rain cuts about 512 trips a day.

Honest limits

- The model learned one made-up year. A new bike lane, a strike or a heatwave hotter than anything in the data could break it.
- It finds patterns. Warm days and busy days go together, but the model cannot say why.
- Image models trained in one city often do worse in another, where streets, light and buildings look different.
- Errors are not shared equally. An area with fewer training examples may get worse predictions.

WATCH OUT

A high score on the training data means little. Judge a model on data it has not seen, and against a simple baseline.

IN ONE LINE

Machine learning sorts, groups and predicts from examples; test it on unseen data and remember it only knows the past.

8 Worked Example: Grouping Neighbourhoods

k-means clustering, built from scratch.

Imagine 150 neighbourhoods, each described by four numbers. They are population density in people per hectare (a hectare is 100 m by 100 m), median age, the share of homes that are rented, and the share of commuters who walk, cycle or take public transport. Can a computer sort them into a few types without being told what the types are? **k-means** is the classic way to try.

How k-means works

1. Choose **k**, the number of groups.
2. Drop k starting points, called **centres**, at random among the data.
3. Give each neighbourhood to its nearest centre.
4. Move each centre to the average of its neighbourhoods.
5. Repeat the last two steps until nothing changes.

First the script **standardises** each variable, which means it subtracts the average and divides by the spread (the standard deviation). Otherwise density, which runs into the hundreds, would drown out the percentages. k-means can also get stuck in a poor answer, so the script restarts ten times from different random centres and keeps the tightest result.

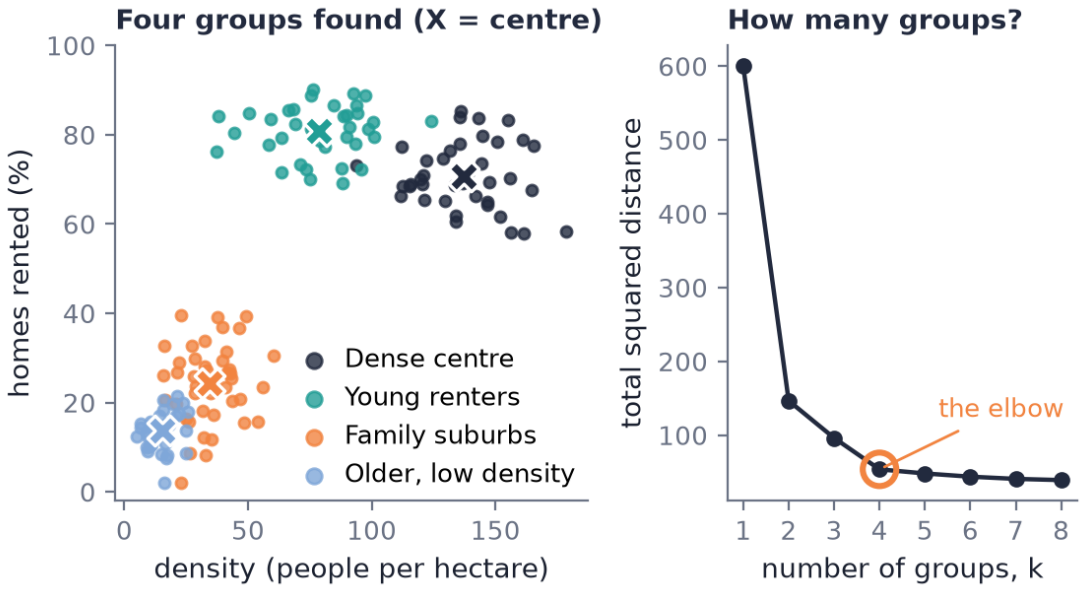


Figure 10. Left: the four groups found, shown on two of the four variables. Right: total squared distance from each neighbourhood to its centre for k = 1 to 8. Computed for this guide.

How many groups? The right panel shows the total squared distance from each neighbourhood to its centre. It falls steeply up to k = 4, then flattens. That bend, called the **elbow**, suggests four groups. The table lists the average profile of each group.

Group (our name)	Places	Density (people/ha)	Median age	Rented (%)	Walk, cycle, transit (%)
Dense centre	39	137	33	71	72
Young renters	36	79	27	81	60
Family suburbs	47	34	39	24	31
Older, low density	28	15	53	14	17

We made the data from four hidden types, so we can check the answer. 97% of neighbourhoods landed in the group made up mostly of their own hidden type. The names in the table are ours. The algorithm only returns numbers, and people decide what each group means.

Here is the core of the method. The script also restarts it ten times and handles groups that end up empty.

```
import numpy as np

def kmeans(X, k, steps=100, seed=0):
    rng = np.random.default_rng(seed)
    centres = X[rng.choice(len(X), k, replace=False)]
    for _ in range(steps):
        dist = ((X[:, None, :] - centres[None]) ** 2).sum(axis=2)
        label = dist.argmin(axis=1)                  # nearest centre
        new = np.array([X[label == j].mean(axis=0) for j in range(k)])
        if np.allclose(new, centres):
            break                                    # nothing moved
        centres = new
    return label, centres
```

WATCH OUT

k-means always returns groups, even for random data. Check that the groups stay put when you change the seed or drop a variable, and ask local people whether they make sense.

IN ONE LINE

k-means groups similar places by moving centres until they settle; people then decide what each group means.

9

Mobility and Movement

Origin-destination matrices and trip patterns.

Mobility analytics studies how people and goods move. Its most useful single table is the **origin-destination matrix**, or OD matrix. Each row is a place where trips start, the origin. Each column is a place where trips end, the destination. Each cell counts the trips from one to the other.

A worked example: a morning in six zones

We split a made-up city into six zones and simulated 11,640 morning trips with a **gravity model**, an idea borrowed loosely from physics. Each zone sends out trips in proportion to its residents. Each trip picks a destination with a chance that rises with the jobs there and falls with distance. The script then drew every trip at random from those chances.

from \ to	Centre	North	East	South	West	Harbour	total
Centre	845	45	56	63	58	133	**1,200**
North	1,390	639	235	63	135	238	**2,700**
East	747	114	625	57	28	529	**2,100**
South	1,244	44	125	690	113	784	**3,000**
West	1,237	117	70	138	647	191	**2,400**
Harbour	66	1	11	20	2	140	**240**
total	**5,529**	**960**	**1,122**	**1,031**	**983**	**2,015**	**11,640**

Morning trips between six made-up zones, simulated for this guide. Rows are origins, columns are destinations. The diagonal holds trips that stay inside a zone. Darker cells carry more trips between zones.

The biggest flow between two zones ran from North to Centre, with 1,390 trips. 48% of all trips ended in the Centre, which holds 41% of the jobs. The Harbour has few residents but many jobs, so 2,015 trips arrived there while only 240 began there. 31% of trips stayed inside their own zone. The average trip, measured between zone centres and counting trips inside a zone as 1.2 km, was 3.7 km.

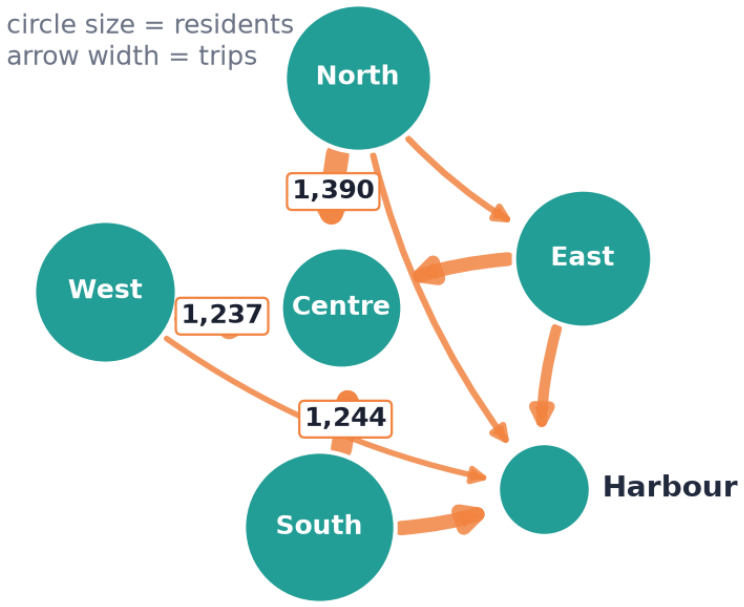


Figure 11. The nine largest flows between zones in the same simulation. Computed for this guide.

Where real OD data comes from

Real OD matrices are estimated from travel surveys, smart card taps, phone records, GPS traces from apps and vehicles, and traffic counts. Each has gaps. Many card systems record only where people board, so analysts must infer where they got off. Phone data show where a phone was, so the analyst must guess which stops were real destinations. Most OD matrices are therefore estimates, and good ones say how they were built. Planners use them to design routes, find where demand outgrows capacity and test what a new bridge or bus line might change.

TIP

Read the matrix two ways. Along a row: where do people from here go? Down a column: where do people who come here start? The column view shows who depends on a place.

IN ONE LINE

An OD matrix counts trips from each place to each other place; it is the backbone of mobility analysis.

10 People, Participation and Civic Tech

Working with residents as well as studying them.

Data about a city is also data about people. Residents know things no sensor sees, such as which crossing feels unsafe, which bus stop floods, where the shade is. Urban informatics works best when residents help shape the questions, the data and the decisions.

How much say do people have?

In 1969 Sherry Arnstein described a *ladder of citizen participation* with eight rungs. At the bottom is manipulation, which is not participation at all. Informing and consultation sit in the middle as token gestures. Citizen control is at the top. A widely used later framework, the *Spectrum of Public Participation* from the International Association for Public Participation (IAP2), has five levels (Figure 12). Each level gives residents more influence.

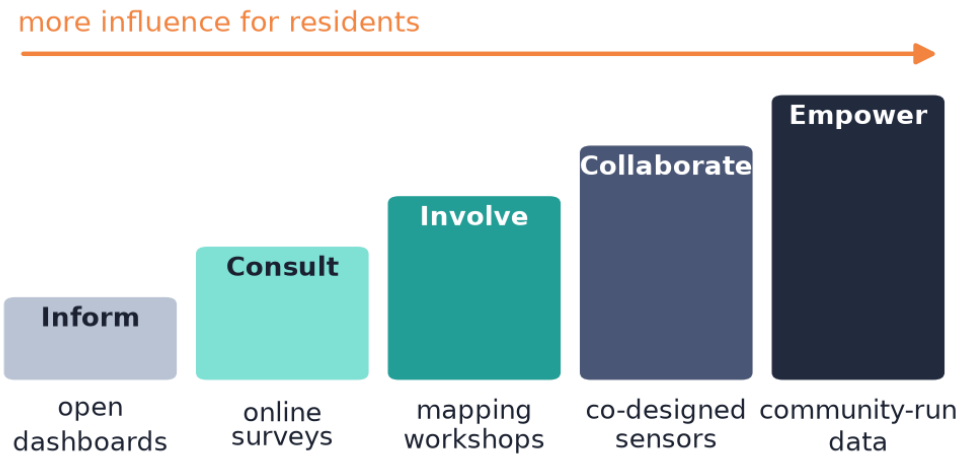


Figure 12. Five levels of public participation, with a data example for each.

Civic tech

Civic tech means technology that helps people take part in public life. Common kinds include:

- **Report-a-problem apps**, where residents send a photo and location of a pothole or broken light and can follow its progress.
- **Participatory budgeting platforms**, where residents propose and vote on how to spend part of a public budget. Decidim, an open-source platform first built for Barcelona, is a well-known example.

- **Community mapping**, where volunteers add local detail to OpenStreetMap, sometimes for disaster response through groups such as the Humanitarian OpenStreetMap Team.
- **Citizen science**, where residents measure air quality, noise, trees or wildlife and share the results.
- **Open data portals**, where anyone can download public data, and **hackathons**, short events where people build tools together, often on that data.

Participation has its own bias. People with spare time, confidence and internet access take part more. Good projects reach out on purpose, with paper forms, local languages, meetings at convenient times and places, and payment for people's time where possible.

> **SAY IT LIKE THIS**
> We would like your help to decide what we measure, as well as to tell us what you see.

> **IN ONE LINE**
> Treat residents as partners who shape the questions, see the data and influence the decisions.

11 Ethics, Privacy and Data Governance

Who is watched, who decides, and who benefits.

City data can improve services. It can also be used to watch people, sort them unfairly or sell their habits. **Ethics** here means asking, at every step, whether a use of data is fair, safe and agreed.

Location data is personal

Removing names does not make location data anonymous. A well-known 2013 study of 1.5 million mobile phone users (de Montjoye and colleagues; see Further Reading) found that four places and times were enough to single out 95% of them. To see why, we placed 300,000 made-up people in a 20 km by 20 km city, each with a home and a workplace. Then we counted how many are **unique**, meaning nobody else shares their exact combination, when locations are rounded to grid squares of different sizes.

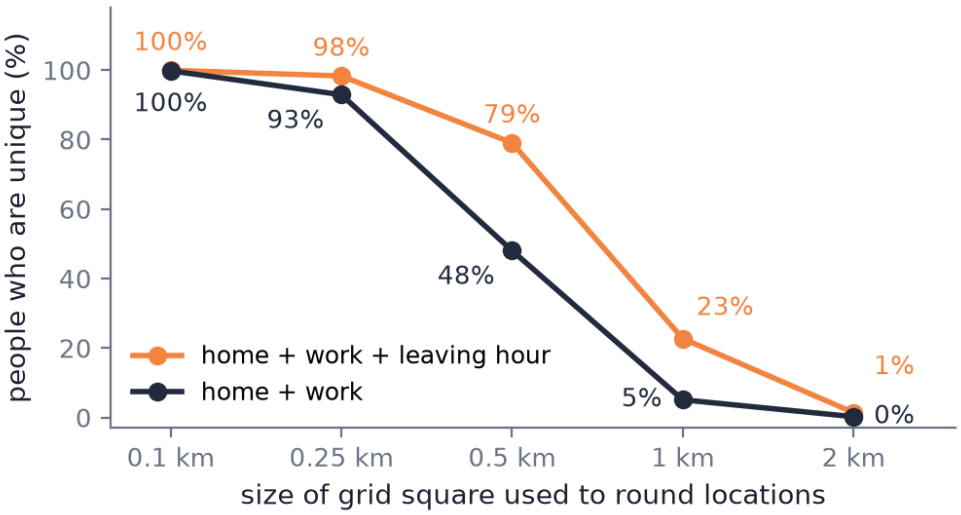


Figure 13. Share of made-up people who can be singled out, by how coarsely locations are rounded. Computed for this guide.

With 250 m squares, a few streets across, 93% of people have a unique home-and-work pair. With 500 m squares, 48% still do, and adding just the hour each person leaves home lifts that to 79%. With 1 km squares the share falls to 5%, or 23% with the hour, and at 2 km almost everyone blends in. How easily people are singled out also depends on how many share each square. With only 5,000 people in the same city, 78% would still be unique with 1 km squares. Every extra detail makes people easier to find.

So the safest release is often **aggregated**, meaning counts per area and per hour, with very small counts hidden. Two more tools help. **k-anonymity** makes sure each person is hidden in a group of at least k people who look the same in the data.

Differential privacy adds carefully measured random noise, so that no single person changes a published result by much.

Bias, fairness and surveillance

Bias enters through who is in the data (Chapter 2), what gets labelled and how results are used. A model that sends inspectors where complaints are highest sends them where people complain most, which may not be where problems are worst. Cameras with face recognition, licence-plate readers and phone tracking can follow individuals across a city, and some cities have limited or banned some of these uses. Before collecting anything, ask whether it is necessary, whether it is proportionate and whether a less intrusive way exists.

Data governance

Data governance is the set of rules, roles and habits that decide how data is collected, stored, shared and deleted.

Principle	In practice
Purpose	Collect for a stated reason; ask again before reusing data for something new.
Minimisation	Collect only what you need, at the coarsest detail that works.
Transparency	Publish what you collect, why, and for how long.
Security	Protect against leaks; limit who can see raw records.
Retention	Delete raw data once it is no longer needed.
Accountability	Name who is responsible and give people a way to complain.
Shared benefit	Make sure the communities who provide data also gain from it.

Two short sets of principles are worth knowing. **FAIR** asks that data be Findable, Accessible, Interoperable and Reusable. **CARE**, written for Indigenous data, asks for Collective benefit, Authority to control, Responsibility and Ethics.

WATCH OUT

"We removed the names" is not a privacy plan. Ask how someone could re-identify people by joining your data with other data.

IN ONE LINE

Collect less, aggregate more, explain clearly, and let communities share in the decisions and the benefits.

12 Communicating with Data and Maps

Charts and maps that people understand and trust.

An analysis only helps if people understand it. A very common map mistake is to map raw counts. Big districts have more of everything simply because more people live there. Figure 14 shows a year of made-up bike-theft reports in twelve districts.

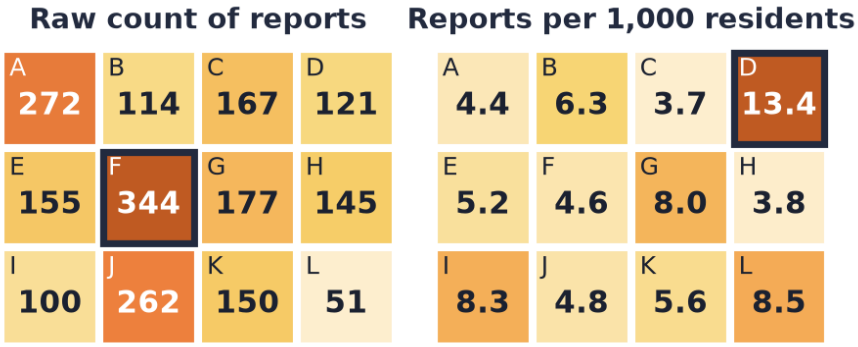


Figure 14. The same made-up reports mapped two ways. The navy outline marks the 'worst' district on each map. Computed for this guide.

By count, district F looks worst, with 344 reports. But it is also home to 75,000 people, so its rate is 4.6 per 1,000. Per 1,000 residents, district D is worst, at 13.4, from just 121 reports among 9,000 people. Dividing by population turns a count into a **rate**. A map that shades areas by a value is a **choropleth** map.

Simple rules

- Lines show change over time, bars compare groups, maps show place.
- Use colours that colour-blind readers can tell apart, light to dark for low to high.
- Label units, source and date, and show uncertainty, such as small samples.
- Make the title a one-sentence takeaway. Test it on a fresh pair of eyes.

SAY IT LIKE THIS

This map shows reports per 1,000 residents over the last twelve months. Grey areas had too few reports to show safely.

IN ONE LINE

Map rates, label sources and dates, show uncertainty, and say the main point in plain words.

13

Pitfalls and Honest Limits

Traps that catch beginners and experts alike.

Most mistakes come from misreading what the data can say. These traps come up again and again.

Pitfall	What happens	What to do
Biased sample	The data over-represents some people or places.	Ask who is missing; weight or combine sources.
Ecological fallacy	Assuming what is true of an area is true of each person in it.	Talk about areas when you only have area data.
MAUP (the modifiable areal unit problem)	Results change when zones are drawn differently.	Try more than one zoning; report the one you used.
Correlation is not causation	Two things rise together, and one gets wrongly blamed.	Look for a mechanism; compare before and after.
Counts instead of rates	Big areas look worst.	Divide by population, area or trips.
Stale data	Decisions rest on old maps or old images.	Check dates and show them on every output.
Overfitting	A model memorises its training data and fails on new data.	Test on unseen data; keep models simple.
Dashboard theatre	Numbers look impressive but change nothing.	Tie each indicator to a decision and an owner.
Solutionism	Reaching for technology before understanding the problem.	Start with residents and the problem; choose technology last.
Lock-in	A city depends on one supplier's closed system.	Prefer open standards and data you can export.

WATCH OUT

A beautiful map can make a weak analysis look certain. The neater the picture, the more carefully you should check what lies behind it.

IN ONE LINE

Ask who is missing, check boundaries and dates, and never mistake a pattern for a cause.

14

Tools and Try It Yourself

A short recipe, open tools and an eight-week plan.

The recipe

1. Pick one small, real question about a place you know, such as which bus stops lack a shelter.
2. Find one or two open datasets: an open data portal, a transit feed in GTFS, or OpenStreetMap.
3. Load the data in a notebook (such as Jupyter, which mixes code, results and notes) and look: size, columns, smallest and largest values, blanks.
4. Clean it, and write down every change you make.
5. Make one map and one chart.
6. Write three sentences: what you found, how sure you are, and who is missing.
7. Show it to someone who lives there and ask what you got wrong.

Open tools worth knowing

Tool	What it is	Good for
Python with **pandas**	A programming language and its table library	Cleaning and analysing data
GeoPandas	pandas with shapes and coordinates	Joining data by location, quick maps
Jupyter	Notebooks that mix code, results and notes	Learning and sharing analyses
QGIS	A free desktop GIS	Making maps without code
PostgreSQL with **PostGIS**	A database with spatial functions	Large, shared spatial data
Kepler.gl	A web tool for large point and flow maps	Exploring movement data
OSMnx	A Python package for street networks from OpenStreetMap	Street and walkability analysis
R with **sf**	A statistics language and its spatial package	Statistics with maps
DuckDB	A small, fast database that runs inside your notebook	Big files on a laptop

TIP

Tools change quickly. Check each project's own page for current install steps.

A first look in pandas

This sketch loads an hourly count file, cleans two common problems and draws the daily rhythm. The file name and columns are placeholders for your own data.

```
import pandas as pd

df = pd.read_csv("counts.csv", parse_dates=["time"])
df = df.drop_duplicates()                              # rows sent twice
df["count"] = df["count"].mask(df["count"] == -999)  # 'no reading' -> blank
print(df.describe())                                   # smallest, largest, average
hourly = df.groupby(df["time"].dt.hour)["count"].mean()
hourly.plot()                                          # the daily rhythm
```

An eight-week plan

Week	Focus	Do this
1	Seeing the city as data	List ten traces you leave in a day. Sort them into directed, automated and volunteered.
2	Finding data	Browse an open data portal and OpenStreetMap for a place you know. Note the licence and date of each dataset.
3	Formats and cleaning	Load a CSV and a GeoJSON file in a notebook. Find blanks, odd values and units.
4	Maps	Make a map in QGIS or GeoPandas. Map one thing as a count and again as a rate.
5	Time	Plot an hourly or daily series. Find its rhythm and one odd day.
6	Grouping and predicting	Run k-means on a few area variables. Fit a simple prediction and compare it with a baseline.
7	Movement	Build a small OD matrix from a transit feed or survey sample. Draw the biggest flows.
8	Ethics and sharing	Write a one-page data note: sources, gaps, privacy risks. Show your map to someone who lives there.

IN ONE LINE

Start with one real question, one open dataset and one honest map, then build up.

Myths and Facts

Common misunderstandings, cleared up.

Myth	Fact
More data always gives better answers.	Size does not fix bias. A million posts from one kind of person describe that kind of person. A smaller, fair sample can beat a huge, skewed one.
Data is neutral and objective.	People choose what to measure, where to put sensors and how to label things. Those choices shape the answers.
A smart city is mainly about technology.	Technology is a tool. The useful questions are which problem, for whom, and who decides.
Removing names makes location data anonymous.	Home and work locations, known to within a few streets, can single out most people. Aggregation and coarser detail protect far better.
A real-time dashboard shows the truth.	It shows a few chosen indicators, with delays, gaps and definitions. It is a summary that needs context.
If a model predicts well, it has found the cause.	Prediction finds patterns. Causes need a mechanism and, ideally, an experiment or a before-and-after comparison.

Quick Quiz

Ten questions. Answers are on the next page.

1 What three things meet in urban informatics?

2 Name the three ways city data is made.

3 Why can a map of geotagged posts mislead you about where people go?

4 What does −999 often mean in a sensor file, and why does it matter?

5 In GeoJSON, which comes first, latitude or longitude?

6 What is a baseline in anomaly detection?

7 Why must a model be tested on data it has not seen?

8 Why standardise variables before running k-means?

9 What does one cell of an OD matrix show?

10 Why should you usually map rates instead of counts?

Quiz Answers

No peeking until you have tried.

1. Data and computing; people and communities; and places, the physical city.
2. Directed (collected on purpose), automated (recorded by machines as a side effect) and volunteered (shared by people).
3. People who post are not a fair sample. In our example, people aged 15 to 34 were 34% of residents but wrote 72% of posts.
4. It is a code for 'no reading'. Treated as a number, it drags averages down. In our example the −999 codes alone pulled the average from 21.4 °C down to 14.2 °C.
5. Longitude.
6. What you expect to see, for example the average of the same hour in previous weeks. Hours far from it get flagged.
7. Scores on training data are too optimistic. Unseen data shows whether it learned a real pattern or just memorised.
8. So that variables with big numbers, such as density, do not drown out variables measured in percentages.
9. The number of trips from one origin zone to one destination zone in a given period.
10. Big areas have more of everything. Dividing by population or area lets you compare places fairly.

Glossary

Plain definitions of the words in this guide.

Term	Meaning
Anomaly	A value far from what is expected, worth a closer look.
API	Application programming interface. Lets programs ask for data.
Baseline	The simplest sensible expectation, used for comparison.
Choropleth map	A map that shades areas by a value, ideally a rate.
Civic tech	Technology that helps people take part in public life.
Clustering	Grouping similar items without being told the groups.
CSV	Comma-separated values. A plain-text table, one row per line.
Data governance	Rules and roles for collecting, sharing and deleting data.
Differential privacy	Adding measured noise so no one person sways a result.
Ecological fallacy	Assuming what is true of an area is true of each resident.
GeoJSON	A JSON format for map shapes. Longitude comes first.
GIS	Geographic information system. Tools for mapping places.
Gravity model	More trips between big places, fewer between distant ones.
GTFS	A shared format for public transport stops and timetables.
IAP2	International Association for Public Participation.
Indicator	A number chosen to track something that matters.
IoT	Internet of Things. Everyday devices on a network.
k-anonymity	Each person hides in a group of at least k look-alikes.
k-means	Clustering by moving k centres until they settle.
MAUP	Modifiable areal unit problem. Zones change the answer.
Metadata	Data about data, such as source, date, units and meaning.
OD matrix	A table of trips from each origin to each destination.
Open data	Data anyone may use and share under an open licence.
Spatial autocorrelation	The tendency of nearby places to be alike.
Time series	Measurements of one thing taken at regular times.
Urban informatics	Data, computing and local knowledge used to improve cities.

One-Page Cheat Sheet

Pin it next to your screen.

Idea	In short
Urban informatics	Data, computing and people's knowledge, used to understand and improve cities.
Three kinds of data	Directed (on purpose), automated (a side effect), volunteered (shared by people).
Blind spots	Every source misses someone. Ask who, and compare with a census.
Pipeline	Collect, clean, store, analyse, visualise, act, and repeat.
First checks	Smallest, largest, blanks, duplicates, units, dates.
Formats	CSV for tables, GeoJSON for map features (longitude first), APIs for live data.
Open data	Open licence, machine-readable format, good metadata, shared standards.
Time	Learn the rhythm (the baseline), then flag what falls outside it.
Indicators	Few, clear, broken down, actionable, honest about limits.
Machine learning	Test on unseen data and beat a simple baseline.
Clustering	Standardise first. The algorithm returns numbers; people name the groups.
Mobility	OD matrix: trips from each origin to each destination.
Privacy	Removing names is not enough. Aggregate, coarsen, minimise.
Maps	Map rates, label source and date, show uncertainty.

Further Reading

Where the ideas come from.

- Batty, M. (2013). *The New Science of Cities*. Cities as networks and flows; a foundation of urban science.
- Townsend, A. M. (2013). *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*. A readable history of the smart city idea and its critics.
- Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. A careful, critical look at what data can and cannot do.
- Foth, M. (Ed.) (2009). *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*. An early collection that helped define the field.
- Kontokosta, C. E. (2021). Urban informatics in the science and practice of planning. *Journal of Planning Education and Research*, 41(4), 382–395 (online 2018). How data science fits into planning practice.
- Shi, W., Goodchild, M. F., Batty, M., Kwan, M.-P. and Zhang, A. (Eds.) (2021). *Urban Informatics*. A large handbook of methods and applications.
- Kitchin, R. (2014). The real-time city? Big data and smart urbanism. *GeoJournal*. A short, influential critique of data-driven city management.
- Greenfield, A. (2013). *Against the Smart City*. A sharp critique of top-down smart city plans.
- D'Ignazio, C. and Klein, L. F. (2020). *Data Feminism*. Power, bias and who gets counted in data.
- Arnstein, S. R. (1969). A ladder of citizen participation. *Journal of the American Institute of Planners*, 35(4), 216–224. The original eight-rung ladder.
- de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M. and Blondel, V. D. (2013). Unique in the crowd: the privacy bounds of human mobility. *Scientific Reports*. Why location data is so hard to anonymise.
- Boeing, G. (2017). OSMnx: new methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*. A practical tool for street network analysis.
- Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*.
- Carroll, S. R. et al. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*.

About this guide

Written and illustrated by Claude (Anthropic), prompted and directed by Mahmoud A. This booklet is a general introduction and not professional advice. Every worked example uses synthetic data computed for this guide. The results show how the methods behave. They say nothing about any real city.