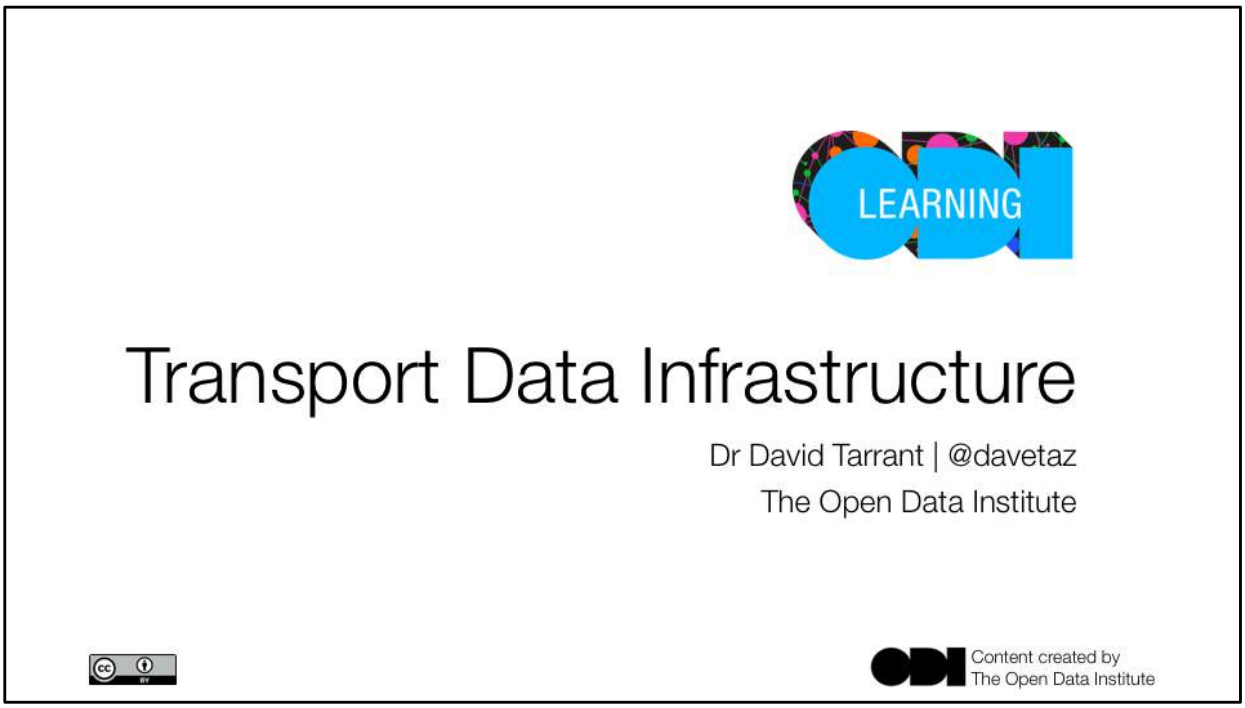


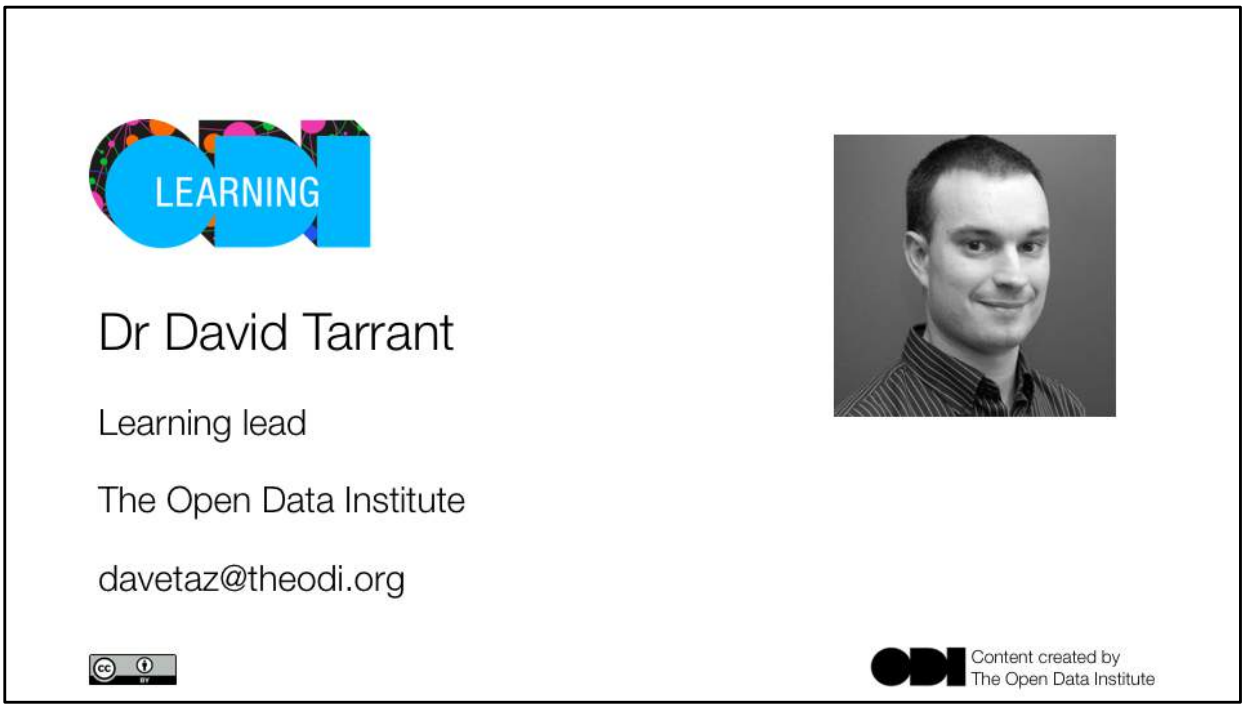
Hello and welcome this weeks training entitled Transport Data Infrastructure.

My name is Dr David Tarrant from the Open Data Institute.

Over the course of this week we are going to be looking at how the UK is attempting to deliver a world leading transport infrastructure using data.

The Open Data Institute has been carrying out a lot of work in this area and has partnered with the top organisations to both deliver services and evaluate impact.

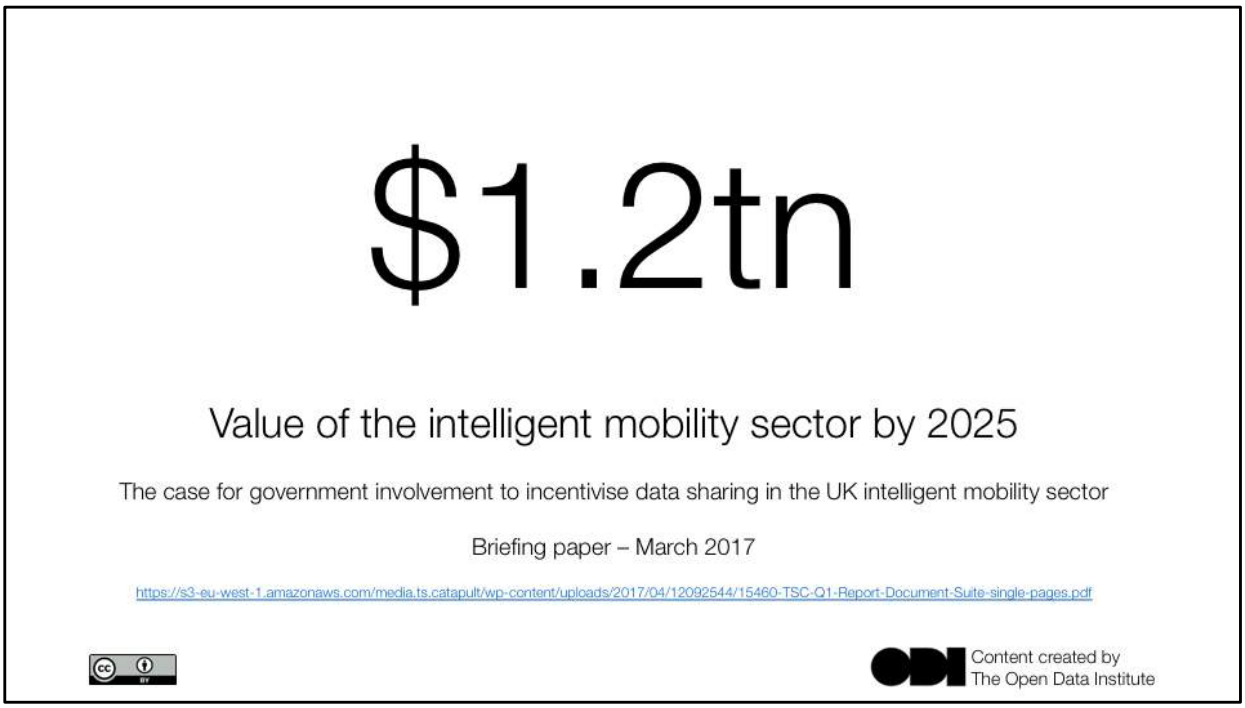
One of the organisations we have been working with is the Transport Systems Catapult (who I believe you may have met). As part of the work we undertook with them we looked at the impact of data in transport and estimate that worldwide there is $1.2tn…



I'm Dr David Tarrant, learning lead at The Open Data Institute.

I joined the ODI from the University of Southampton where I was a Lecturer in the Web and Internet Science Group. I joined with founder Nigel Shadbolt and have been responsible for developing and delivering key educational content that has helped transform governments and unlock over $15m for startups.

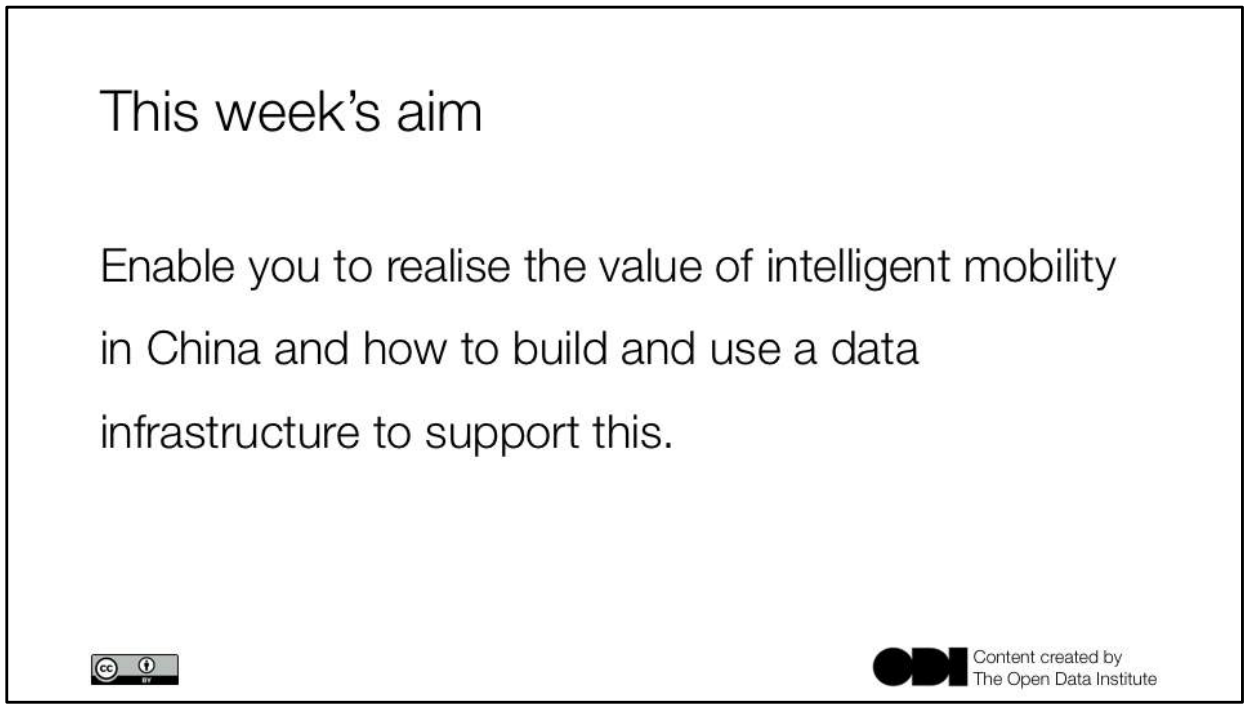
I still live in Southampton, which is at the other end of the country on the south coast from here. As I work in London, I am a specialist with trains and London transport simply because I use it very regularly. So I apologise if you are more into road or water based forms of transport, I have some great content on these as well, but we in the UK are a fairly train obsessed country.



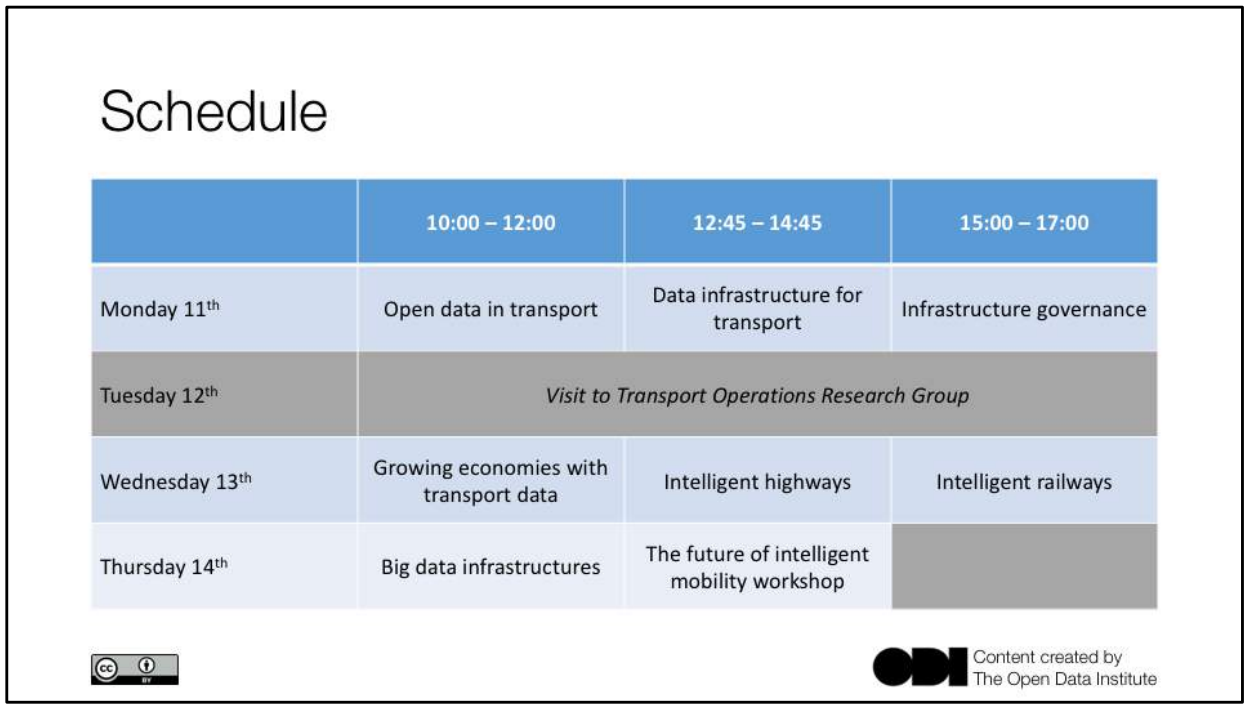
..of value in the intelligent mobility sector to be realised by 2025. All from data.

Over the course of this week we are going to look at the evidence for this and look at how data and its different types can help deliver this huge amount of value worldwide.

Drawing on use cases in the UK, we are going to look at how some of this value has already been realised and how advanced uses of data and already enabling intelligent mobility.



This weeks aim is to enable you to realise the value of intelligent mobility in China and how to build and use a data infrastructure to support this.



To enable this, we have divided the week into a number of key sessions. Each session will last between 1 and 2 hours to allow for translation and questions.

Today we shall be focusing on building and supporting a data infrastructure for transport.

One of the most important aspects of a fair infrastructure for everyone is open data. Our first session will look at what open data is and what other types of data there are.

This afternoon we will be looking at how to establish and govern a integrated transport system and data infrastructure to go with it.

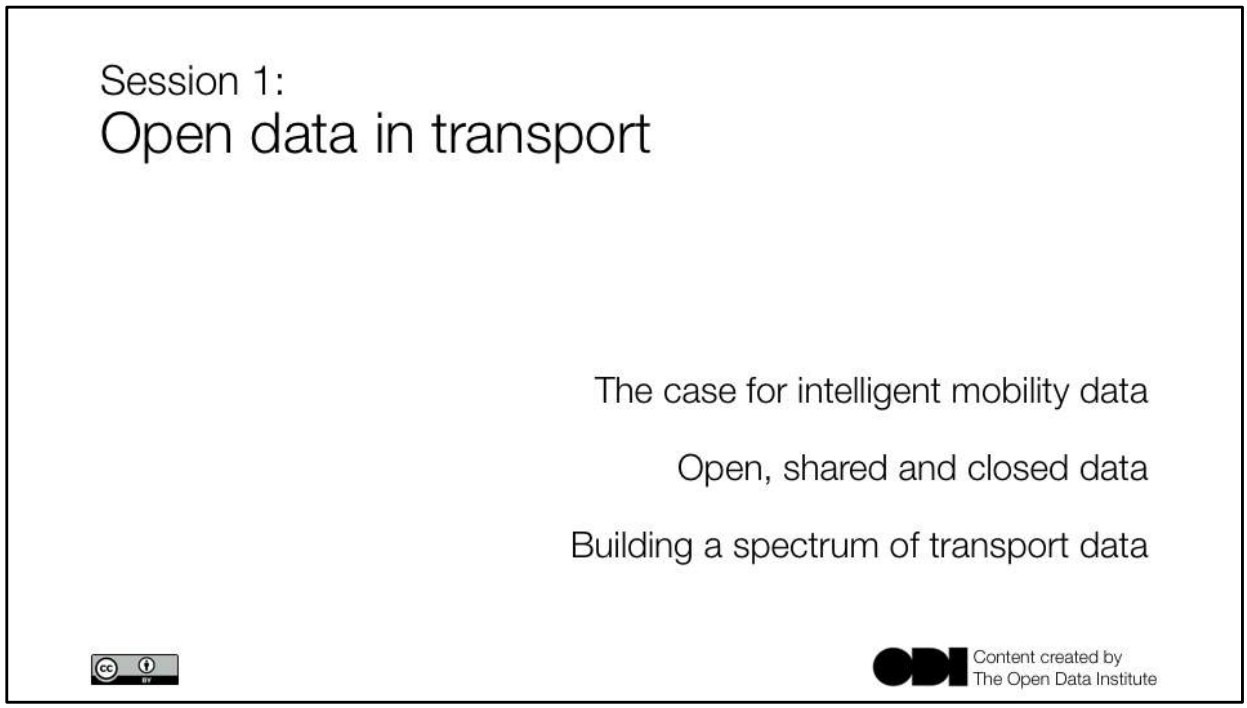
Tomorrow you are visiting the Transport Operations Research Group so will not be with me.

On Wednesday we will be looking at how the infrastructure supports businesses and the emerging economy before spending a couple of sessions looking specifically at highways and railways in the UK and their use of data. This will include bike schemes

in highways.

Finally on Thursday we will look at Big Data and where this fits and the benefits and risks of the big data hubris.

We will conclude the training with a workshop looking at the next steps for a Transport Data Infrastructure in China.



We begin our first session by looking at Open data in transport, a key piece of our infrastructure.

The first part of this session will look at the intelligent mobility report and evidence of why such a infrastructure is so critical.

We will then focus on the different types of data and end this session by building our first spectrum of transport data for a mode of transport in China.

I will also introduce many UK examples as we go and we will help each other to understand the differences so we can both get the most out of the training.

Intelligent mobility

Using emerging technologies to enable the smarter, greener and more efficient movement of people and goods around the world.

https://ts.catapult.org.uk/intelligent-mobility/introduction/

Intelligent mobility is about using emerging technologies to enable the smarter, greener and more efficient movement of people and goods around the world.

Essentially, it is all about the role of technology as an infrastructure that enables better use of the physical infrastructure.

The Open Data Institute views data infrastructures as a key part of a growing and successful economy.

As part of the work we carried out with the transport systems catapult, a lot of research was done into the impact of what was termed intelligent mobility data.

This report estimates that the benefits of a mobility solution are estimated to include:

- Almost 3000 new high skilled jobs (I think this figure is low and hopefully we'll find out why I think that this week)
- £4bn in export value
- Significant contributions to improved productivity and lower costs (I recon somewhere in the £2-£4bn a year)
- Faster journeys and less congestion worth £4bn per year
- Safer roads and fewer incidents worth another £4bn per year
- Improved regional and countrywide connectivity worth £100m per year
- Optimised and more resilient delivery of freight worth £500m per year
- Lower emissions, equivalent to saving £1bn per year

All this adds up to between 14bn and 20bn per year in value.

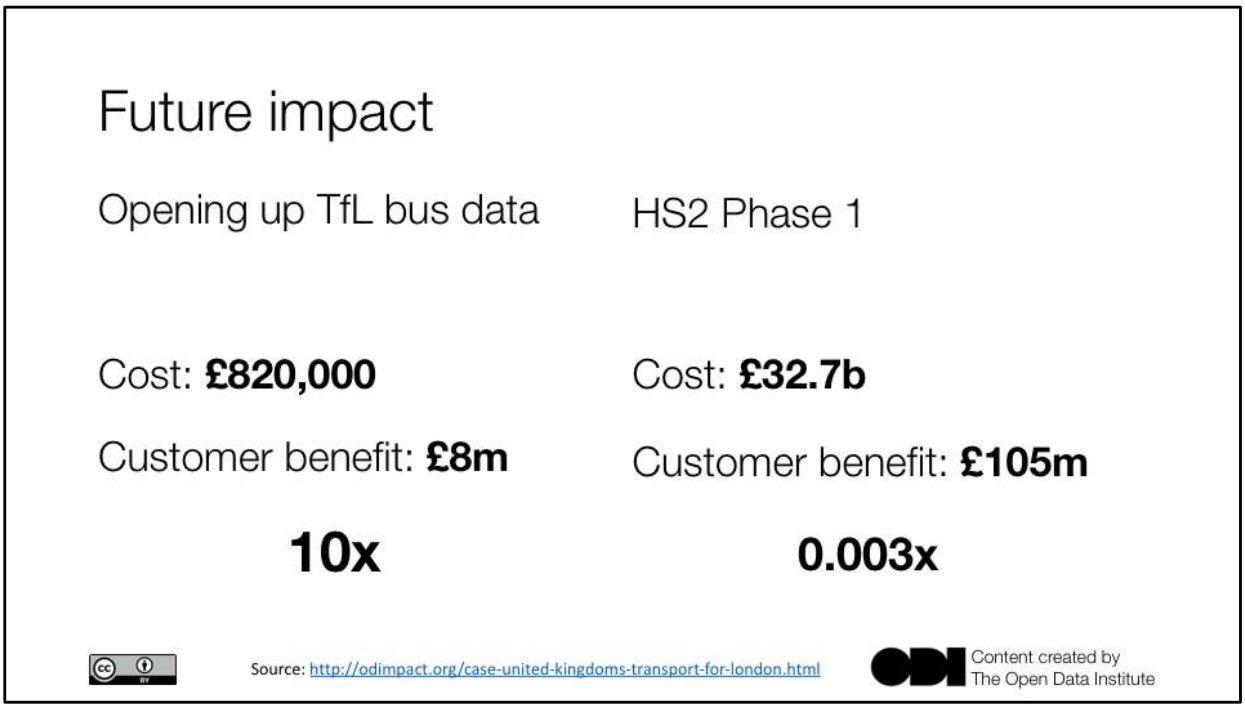
Although these are estimates, there is some good evidence of this already.

One early publisher and supporter of an Open Data Infrastructure is Transport for London (I believe you used them last week) who manage the red busses and famous tube in London).

In 2012 TFL did a study on the potential saving for consumers from using applications to plan travel that source data from their open data infrastructure.

They estimated that the open data used via these applications saved consumers of their services between 15m and 58m a year in total.

This is significant given almost all of these applications are developed and delivered by third parties at near zero cost to TfL.



Lets compare this to another project in the UK.

Here we have another example from TFL involving their recently published real time bus data (something we will look at in more detail on Thursday).

Unlike the previous data, they have published the cost of opening up this dataset as being £820,000 plus recurring cunning costs. Compare this to the estimated customer benefit per year of £8m and they achieve a 10x return in the first year.

Compare this to a physical infrastructure project on the right. HS2 or High Speed 2 is a new high speed rail link between London and Birmingham which will cut one way journey times by 30 mintes (or 33%). The cost in 2010 was of phase one of this route was projected to be 32.7b and the customer benefit (using the same metric as TfL) at £105m per year. Using this metric it would take 311 years to break even.

I would be cautions with both of these metrics as both physical and non-physical infrastructure projects have their benefits and drawbacks. One key issue with data projects in the lack of literacy leading to only data specialists obtaining new jobs. While intelligent mobility estimates 3000 new jobs, HS2 is predicting upwards of

40000 new jobs.

This does show that data infrastructure can lead to fantastic returns that are comparable to physical infrastructure. Something we will look at again on Wednesday in the rail session.

Data infrastructure

Data is infrastructure. It underpins transparency, accountability, public services, business innovation and civil society.

https://theodi.org/what-is-data-infrastructure

A data infrastructure underpins transparency, accountability, public services, business innovation and civil society.

Data such as statistics, maps and real-time sensor readings help us to make decisions, build services and gain insight. Data infrastructure will only become more vital as our populations grow and our economies and societies become ever more reliant on getting value from data.

A data infrastructure consists of data assets, the organisations that operate and maintain them and guides describing how to use and manage the data. Trustworthy data infrastructure is sustainably funded and has oversight that provides direction to maximise data use and value by meeting the needs of society.

Data infrastructure includes technology, processes and organisation.

The ODI believes data is the new raw material of the digital age.

Some people like to say that data is the new oil, but data is in no way like oil.

For a start, data is not a limited resource.

Everyone can have a copy of data without affecting the quality or ability for anyone to use the data for benefit.

The ODI was established in 2012 to help unlock the value of open data, estimated at the time to be $300bn, a figure disputed by some as actually being too low.

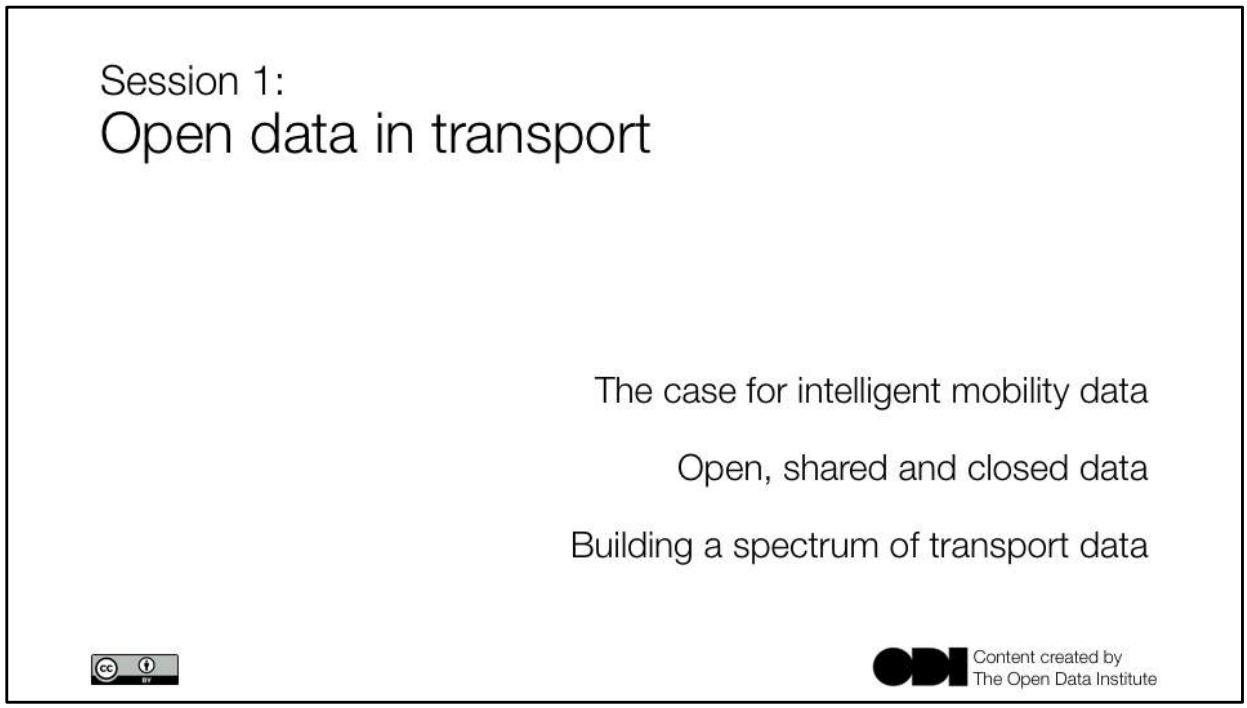
Co-founded by the inventor of the web, Sir Tim Berners-Lee, and renowned Artificial Intelligence professor, Sir Nigel Shadbolt, the ODI works to build a strong, fair and sustainable data economy by helping businesses and governments get data to people who need it, informing decisions, driving efficiencies and creating opportunities using data.

We connect, equip and inspire people around the world to innovate with data

In other words, it is not about what we can achieve. But what we can help you achieve.

Schedule

	10:00 – 12:00	12:45 – 14:45	15:00 – 17:00
Monday 11th	Open data in transport	Data infrastructure for transport	Infrastructure governance
Tuesday 12th	*Visit to Transport Operations Research Group*		
Wednesday 13th	Growing economies with transport data	Intelligent highways	Intelligent railways
Thursday 14th	Big data infrastructures	The future of intelligent mobility workshop	



We have covered part of the case for why data is so essential and what value it can have.

There will be lots more examples as we progress through the week.

As I mentioned, the focus this week in on Transport data infrastructures.

All of the sessions will involve interactive exercises where I will ask you for input and to complete exercises.

On your tables are a number of post-its and pens as well as some paper to help you complete the exercises.

As this is a multi-language session, I am happy for you to discuss answers to the exercises and write answers in your own language.

I would ask that for post-it based exercises that you leave a bit of space for myself or the translator to add a translation.

The exercises are of critical importance to both myself and to you as they will help guide and focus some of the content as appropriate and I thank you in advance. The exercises also provide a bit of fun and are essential in learning, they also provide myself and the translator a little break.

We will start most sessions with an exercise, including this one.

Before we proceed, I want to break down the title of the course into its parts.

Transport. Data. Infrastructure.

We all know what transport is, but what about the other two terms. I would like to start my session with a simple question, which may not have a simple answer.

What is data?

We use the word all the time, but what is data. I would like to give you 10 minutes to write down some words or short phrases on post-its that define data.

In the English language we have another word that gets used a lot alongside and interchangeably with data, and that word is information. For the purpose of this exercise I would like you to avoid the word information and its equivalent and think purely about data.

What is data? and what makes it different from information?

While you are doing this exercise I will visit a number of the groups and get an introduction to each of you so I can get to know who you all are a little better.

You have 10 minutes, to write down words or short phrases that define data for you, examples are welcome.

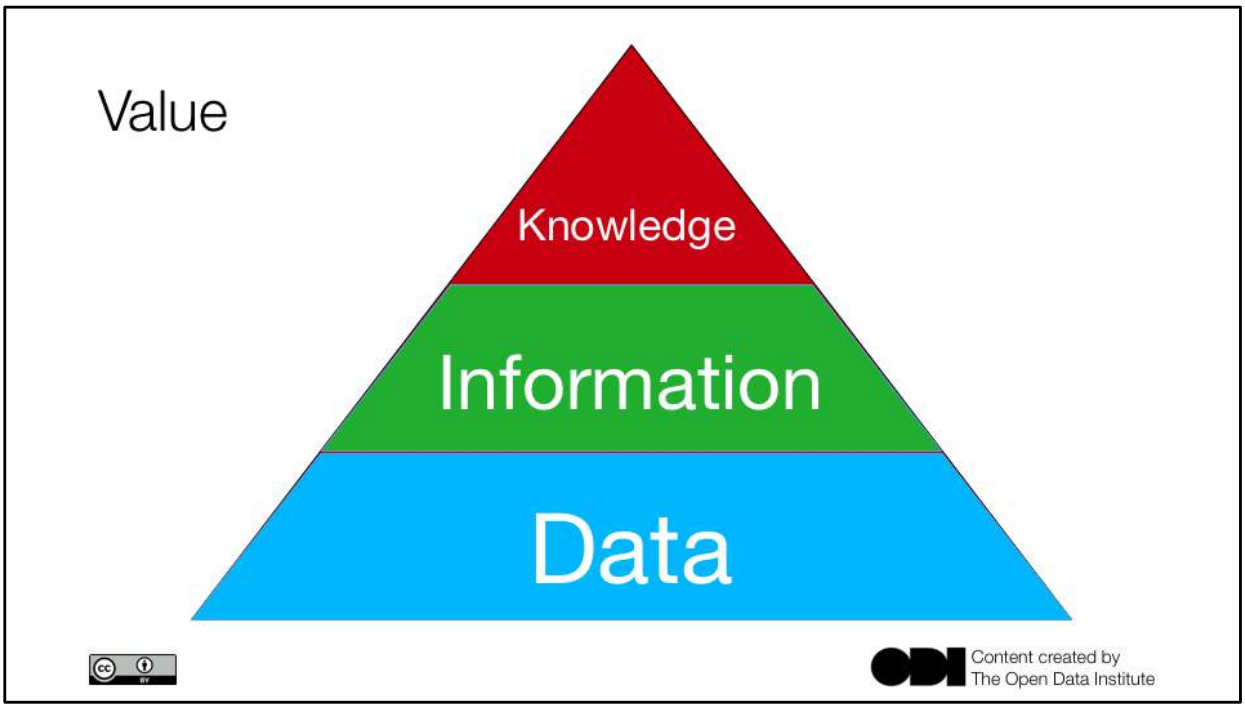
(Collection)

========

I would now like to ask each group for a single post-it to add to our collective board such that we can see if we all agree, I'll start with the group on my right.

When you think of data you think of some key examples, including numbers and words. But data can also include sounds, pictures and places. Multiple pictures make movies, and places are geographic locations or geographic data.

Data doesn't have to be digital, but in order for everyone to benefit from it, we currently have no better platform for sharing data than over the internet.



We define data as the lowest level of abstraction from from information and knowledge is derived.

This allows us to draw this triangle, with data being the biggest segment at the bottom, as it is the largest resource with the most potential.

In order to obtain information from data you need to add context.

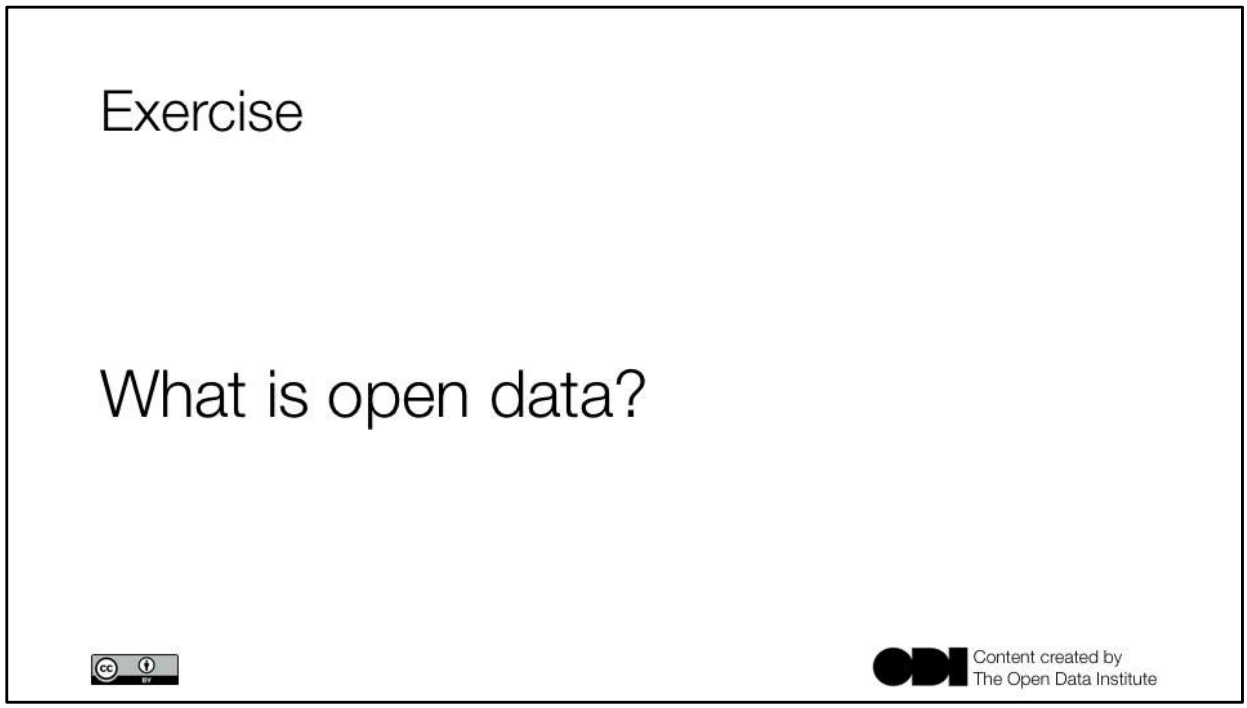
For example you could have a set of random numbers as data, but adding a context might turn these numbers into the ages of males in the first class coach of a train. This is information. Combining this with other information might reveal that the average age of males in first class of a train is younger than the average for the rest of the train.

Not all data has to be factual, in fact I just made that last statement up and have no idea if it is true, so please don't quote me. For one thing I'm not sure there is any data on that available currently.

However in this example it was the knowledge that was interesting. Why are males

younger in first class? That's interesting, the data isn't. Data isn't that exciting, however what can be done with it is. We continually ask ourselves at the ODI why people deny access to data that can be shared easily, and sometimes the reason is to protect knowledge. However we also know that turning data into knowledge is hard and requires subject matter experts to get it right. We will look at what happens when you don't involve an expert in the Big Data session on Thursday.

In the rest of this session we are going to look at different types of data, starting with Open Data.



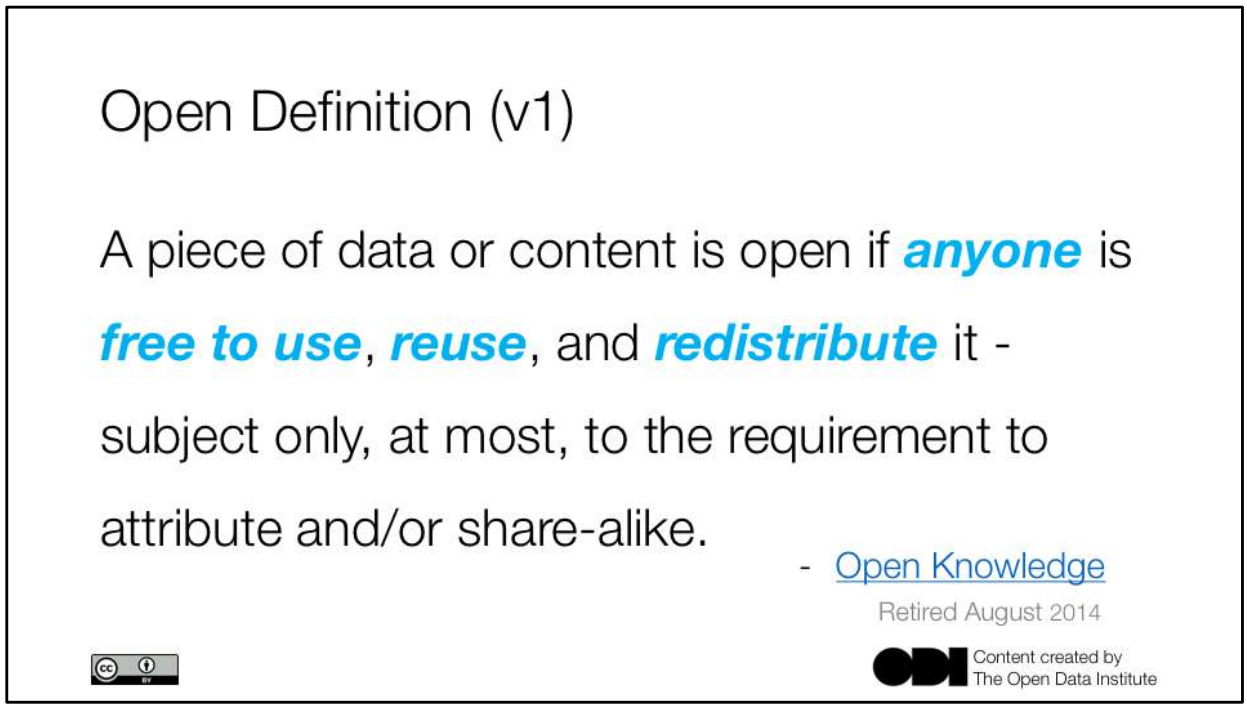
As a part two to this exercise I would like you to define open for me.

We have already defined data, so now we need to look at what makes data open so it can deliver the benefits that TfL have seen.

We shall do the same exercise with post-its and I will come around a meet anyone I haven't yet. We shall also leave 5-10 minutes for this depending on how many people I have yet to meet.

Write down some words or short phrases on what makes data open.

Collection
=======
That is a fantastic set of responses, let's look at a couple of definition of open data before looking at how it applies to transport.



This is the definition of open created by Open Knowledge International. This summary definition was retired in August 2014 however is still widely used today as a measure of something being open.

It states that a piece of data or content is open if anyone is free to use, reuse and redistribute it.

For data to be open, it should have no limitations that prevent it from being used in any particular way. Anyone should be free to use, modify, combine and share the data, even commercially.

This causes contention with many publishers who wish to control what can be done with their data and feel that people might try to harm them or misrepresent the data. One of the main benefits of data being open to anyone is that if one organisation or person does misuse the data then others can easily refer to the source data and evidence the fact this person or organisation is in the wrong. Misuse is worse when you don't have open data as no one can check an orgnaisiton or persons use and thus it is harder to dispute it.

Open data must be free to use, but this does not mean that it must be free to access.

There is often a cost to creating, maintaining and publishing usable data.

Ideally, any fee for accessing open data should be no more than the reasonable reproduction cost of the unit of data that is requested.

This reproduction cost tends to be negligible for many datasets.

Live data and big data can incur ongoing costs related to reliable service provision.

As we have seen already with the TfL bus data, the £820,000 cost of publication, which is far less than buying physical infrastructure, has a much quicker return on investment. Thus keeping the cost for users low or zero will help this return on investment speed up.

Once the user has the data, they are free to use, reuse and redistribute it – even commercially.
Open data is measured by what it can be used for, not by how it is made available. Aspects like format, structure and machine readability all make data more usable, and should all be carefully considered.
However, these do not make the data more open.

This definition from OKI is still one of the best as it contains the word “if”. This word turns the definition into a condition and thus we can actually classify something as open data or not strictly using the conditions set out in this definition.

This definition also sets out the maximum requirement that you can put on a user of the data. You can require that the user attribute you as the publishing organisation if they use your data. This is like saying thank you.

You can also set the requirement that any user of the data who changes the data must republish this changed version as open data for others to use, reuse and redistribute. This is the share-alike condition and it know as the viral condition.

The Open Data Institute has its own definition of open data that follows the same principals as the open definition from OKI but this summary fits in a tweet.

We define open data as data that anyone can access, use and share. Simple, short and avoids mentioning anything about cost.

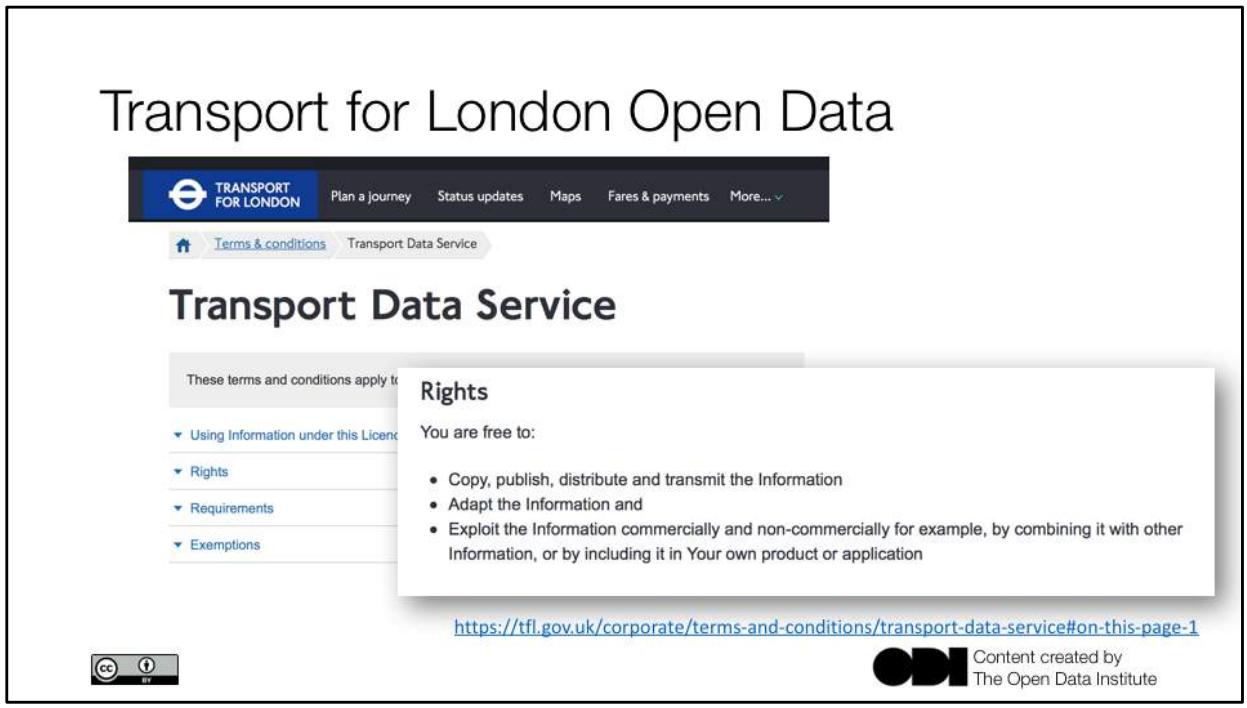
The UK government also has a definition of open data.

This definition adds aspects that specify that the data must be published, which means made available to the public, under a license that allows for free reuse.

The UK government has a specific Open Government Licence that is applied to all open data. This licence explicitly states that users can use, reuse, adapt and modify the data, even commercially.

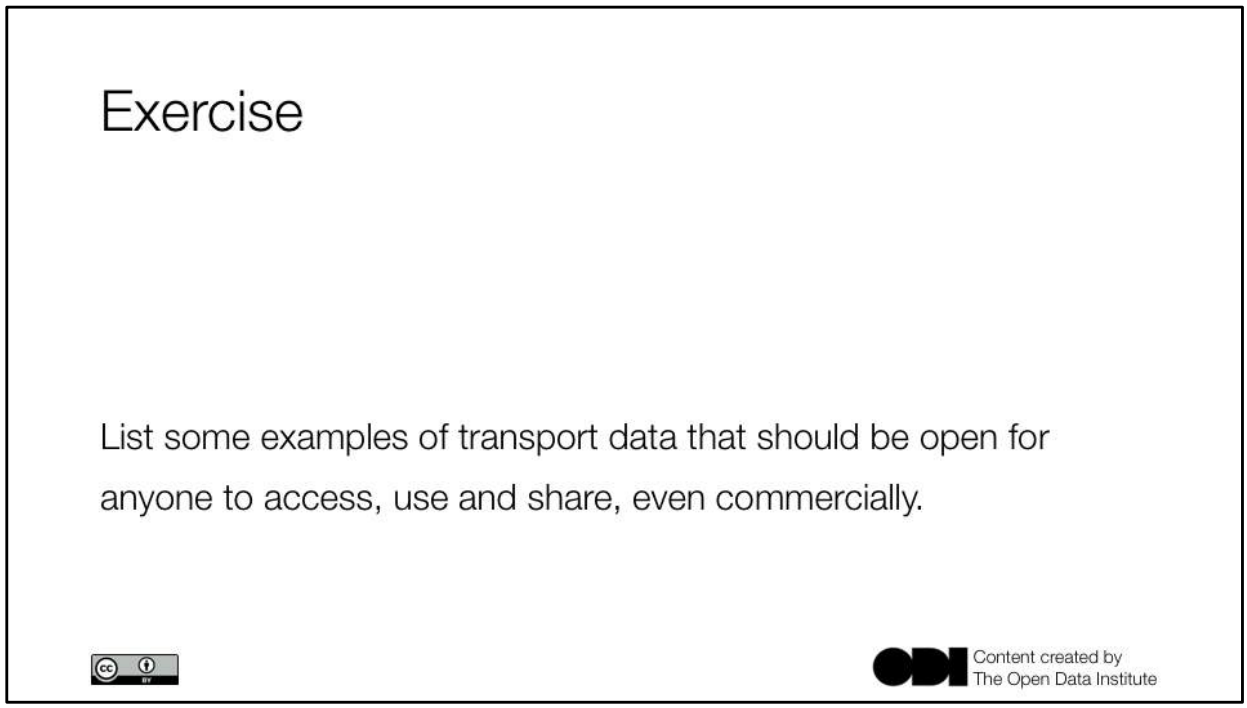
Without a licence, data is not truly open. A licence tells anyone that they can access, use and share your data. Unless you have a licence, data may be 'publicly available', but users will not have permission to access, use and share it under copyright or database laws.

The UK government definition of open data also stipulates that the data must be available in an open, machine readable format which as previously mentioned helps usability but it not a strict requirement for data to be open.



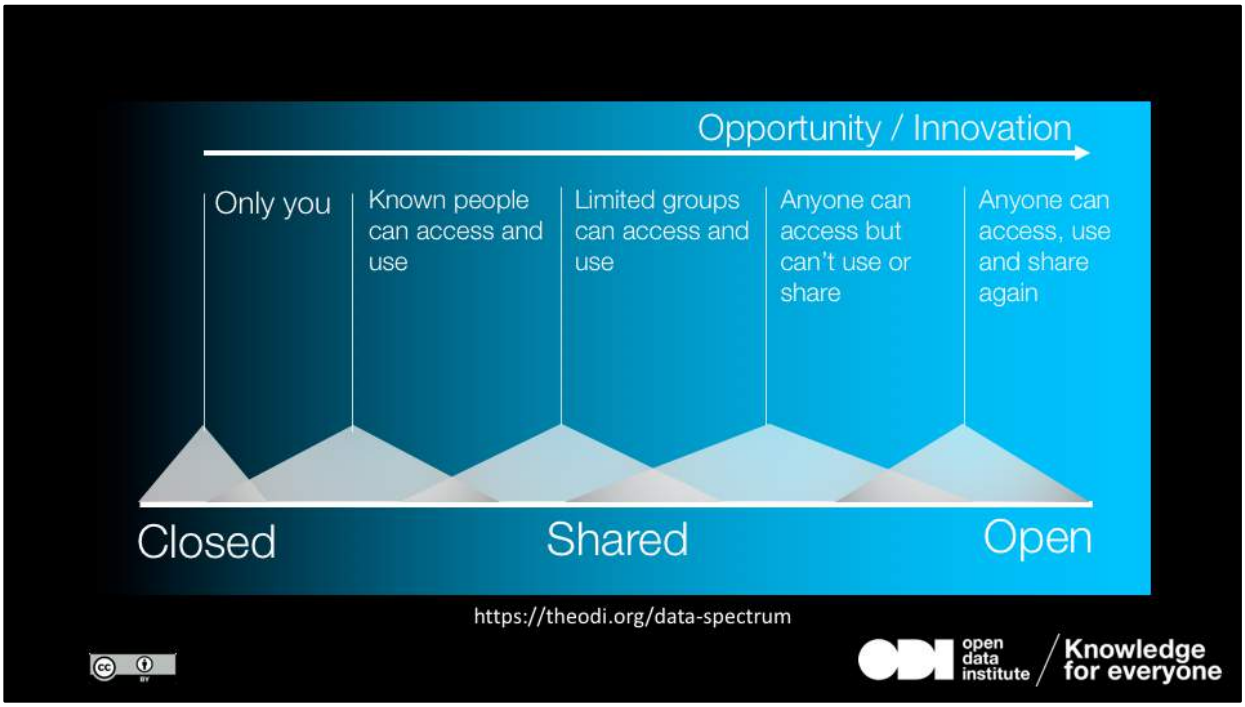
The Transport for London data is available as open data as it is not just available on the public web, but also licensed as open data.

The licence states that anyone is free to:
* Copy, publish, distribute and transmit the Information
* Adapt the Information and
* Exploit the Information commercially and non-commercially for example, by combining it with other Information, or by including it in your own product or application



In your groups can you now write down some example datasets that should be open for anyone to access, use and share, even commercially.

Please write down one dataset per post-it and we will use these after the next section to build a spectrum of datasets.



Not all data is, or should be open.

But this doesn't mean that data that isn't open is closed.

The ODI has put together a data spectrum to help people understand the different levels of access to data and I would like to introduce this as a tool to help you classify the different types of transport data that you have. This will help you classify the access to the data currently as well as point help indicate which datasets should be more (or less) open.

So at one end of the spectrum we have closed data and at the other open.

Closed data is data that only you, as a single individual, have access to. This includes pin codes, passwords and maybe your personal diary. This is data that not even your bank, doctor or family have access to.

At the other end of the spectrum is open data. This is data that anyone can access use and share. Hopefully you should have now written down a number of datasets that exist here, can anyone give me one please?

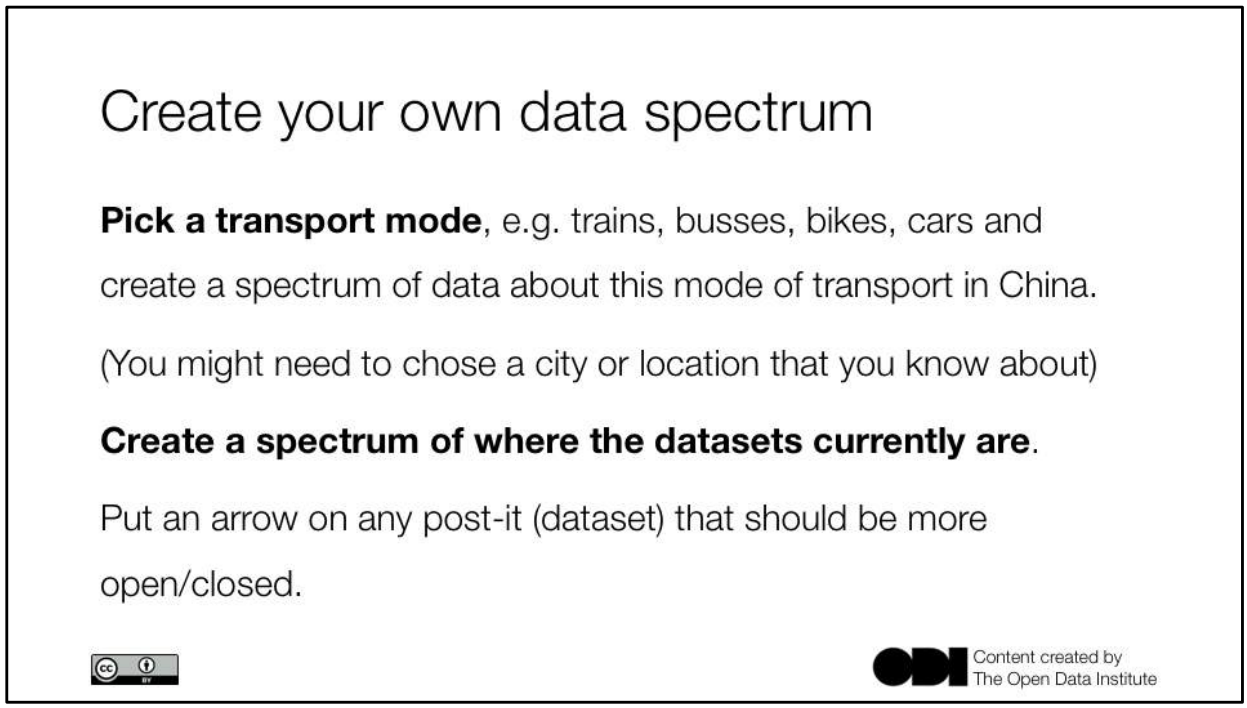
A good example in transport is timetable data for when busses and trains depart and arrive at each station.

This leaves the middle of the spectrum which is all shared data but comes in three different types.

Next to open is public data. This is data that anyone can access but can't use or share due to licensing restrictions. Anything on the web that is not openly licensed falls into this category. Good examples include twitter data, photos on news websites and your personal blog (unless you have openly licensed it).

The next category is data which limited groups can access and use. This includes a lot of medical data which is shared with research organisations for the purposes of creating cures for disease. The key charecteristic of data in this category is that the sharing is with groups and not known individuals, thus the publisher cannot know exactly who has access to the data, just which groups.

The last category is where known people can access and use the data and includes data such as employment contracts and sales reports. In transport this might include operational fuel costs and personnel rotas.



For the last part of this session we are going to create a spectrum of datasets for a transport mode in China, please try and choose different ones per group.

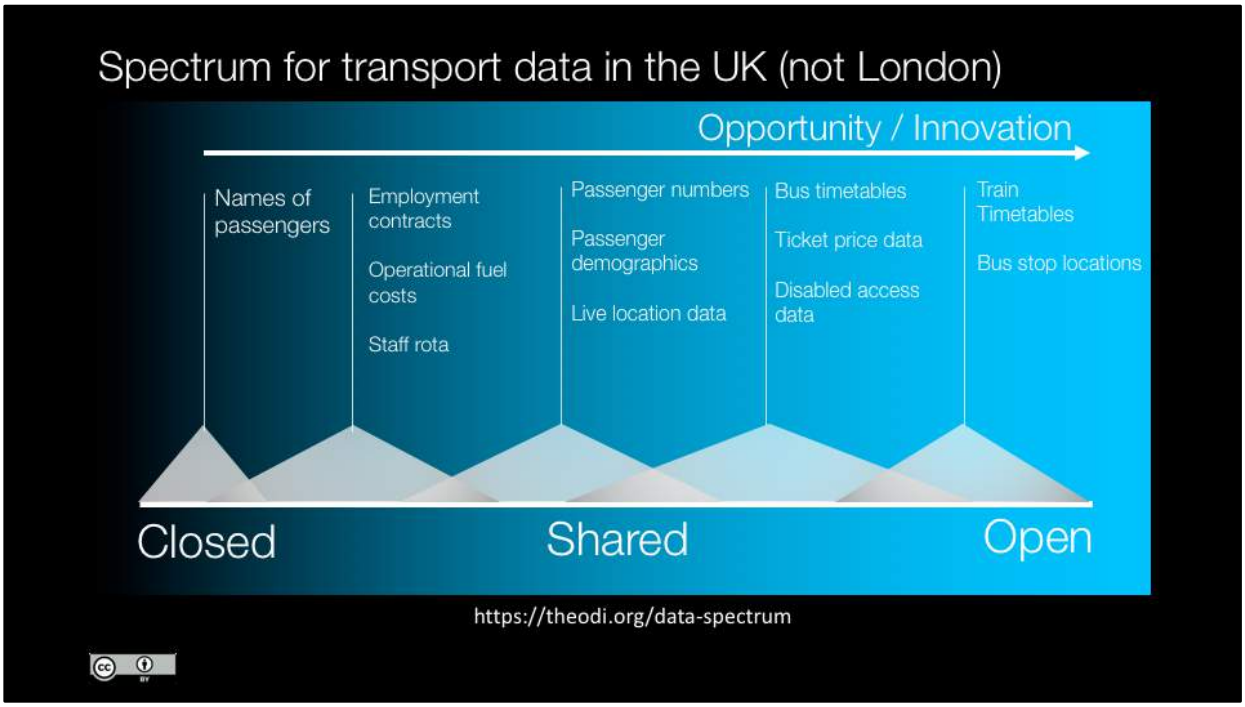
Pick a transport mode, e.g. trains, busses, bikes, cars and create a spectrum of data about this mode of transport in China.

You might need to chose a city or location that you know about.

List one dataset per post-it for that mode of transport and then stick it on your spectrum in the correct place for where that dataset is currently.

Lastly put an arrow on any dataset that should be more open/closed. The arrow indicates the direction the dataset should be shifted towards.

We will be using these in our last session this afternoon when looking at policies and models.



So here is an example from me in the UK.

At the closed end of the spectrum, a train conductor or bus driver might know the names of some passengers.

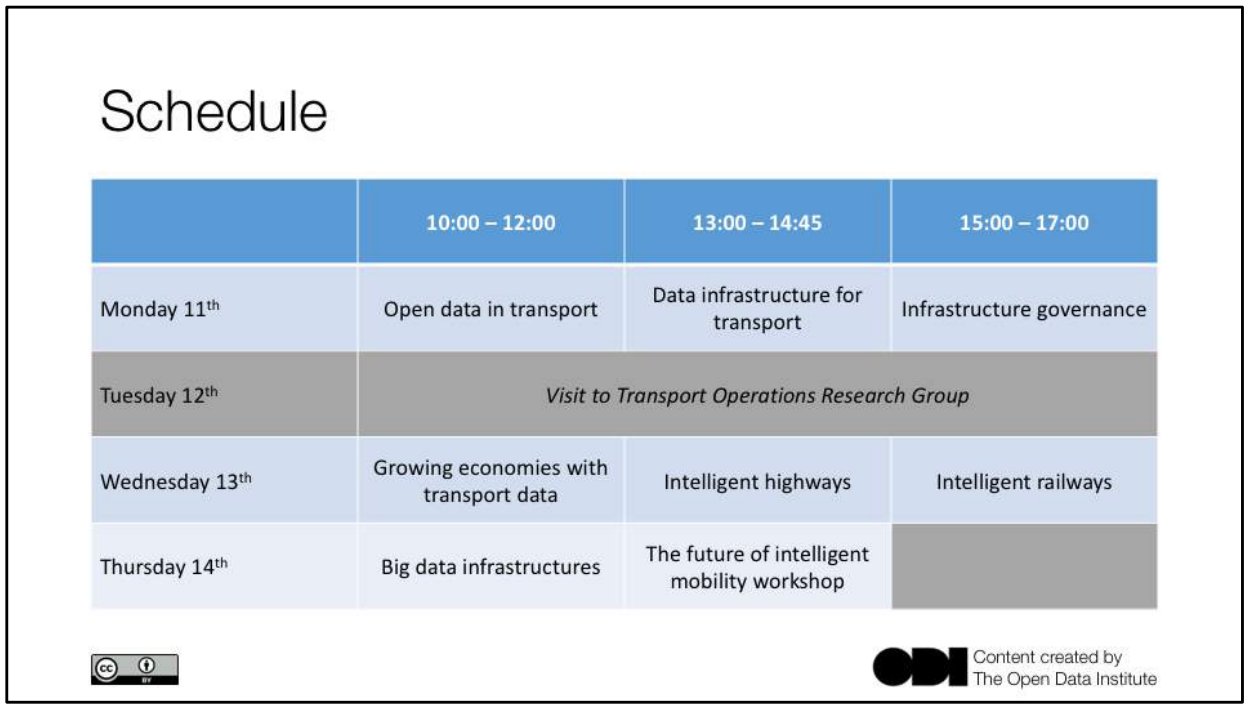
Shared with named people can include employment contracts, sensitive costs and staff rotas.

Shared with groups of people might be aggregate data to do with passenger numbers, demographics and live location data. Many transport providers claim that much of the usage and location data is commercially sensitive.

Much of the transport data exists in the next category including ticket price data, timetables and disabled access data.

Under the open category currently are train timetables and live train position data. Bus stop locations are available as open data, however as can be seen, timetables and live locations are not.

The major push in the UK is to shift the public data into the open data cetegory. London and TFL have got beyond this and are now looking at getting more of the shared data, including passenger statistics open.



I hope you have enjoyed the morning session, this afternoon we will look at what an infrastructure is, the key role that standards play and how data standards and sharing is as important for data infrastructure as for physical infrastructure.

Thank-you

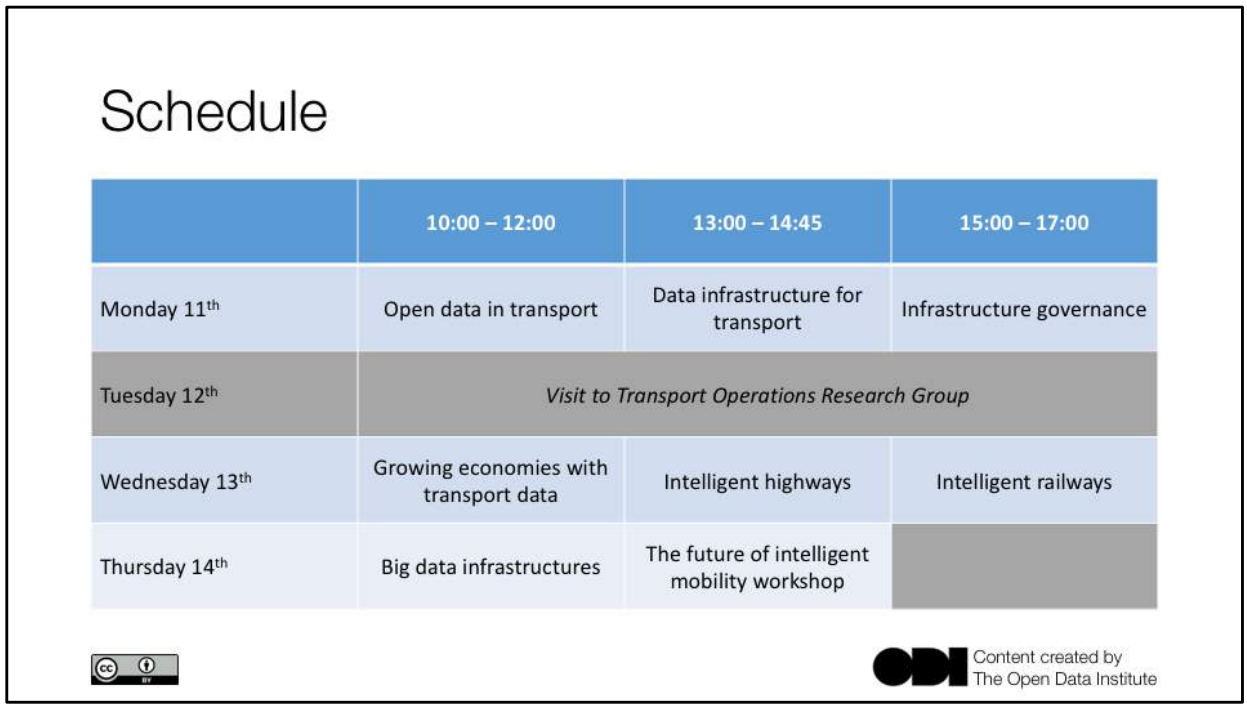
We now have a break for 1 hour for lunch so we can continue to the next session.

=====

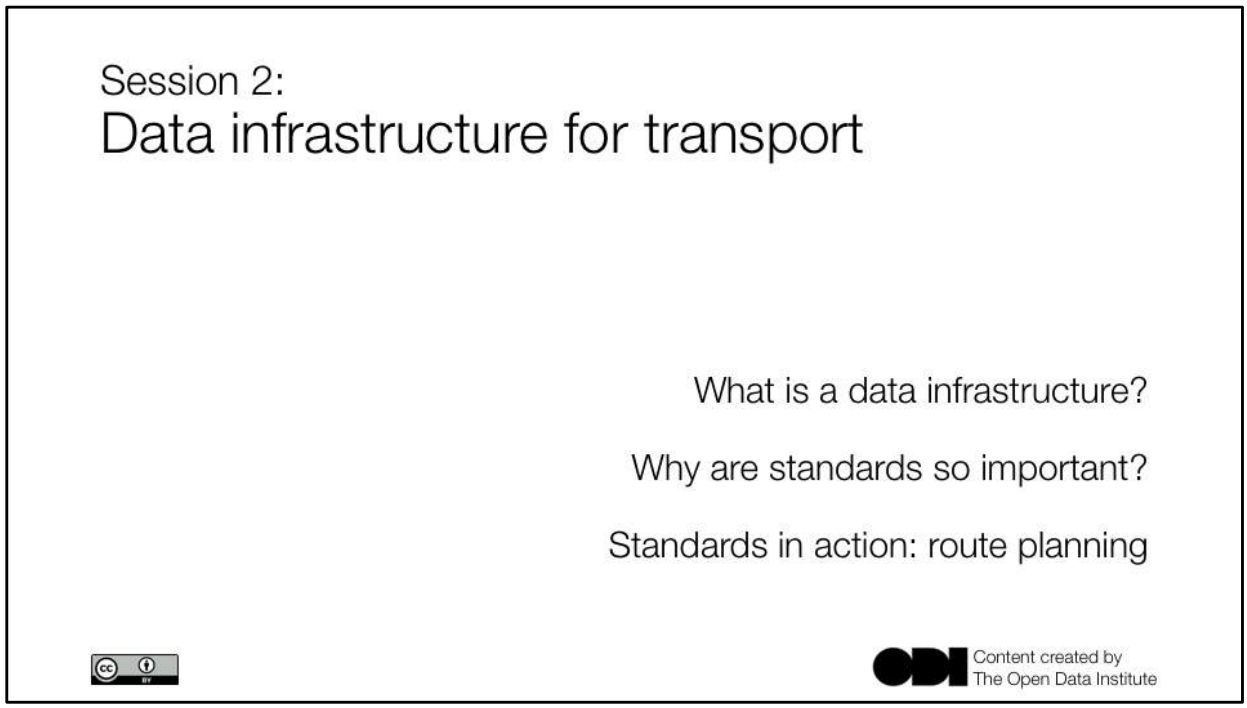
Transport Data Infrastructure

Dr David Tarrant | @davetaz
The Open Data Institute

Content created by
The Open Data Institute



Welcome back,



Having looked this morning at the value that a open transport data infrastructure, such as the one TFL has, can provide. We are now going to look at what an infrastructure is. The key role that standards play and how data standards and sharing is as important for data infrastructure as for physical infrastructure.

We will then look in this session at how difficult it still is to use transport data before looking in the final session about how governance and policy is important in getting it right.

We all know the importance of physical infrastructure. Physical transport infrastructure connects our economy and enables mass movement of people, goods and services.

The invention of the web bought about the information age, allowing the seamless movement of information and digital service across the globe. The internet is now something we cannot now live without and in many countries access to the internet is a right, not a privilege.

Having been through the information age, we are just starting to realise the potential of the web to re-invent itself again as the infrastructure for the age of data. Data that everyone can access, use and share in the same way as our physical infrastructure. The movement and trade of non-physical assets is going to be as important as physical goods.

Data infrastructure

Data is infrastructure. It underpins transparency, accountability, public services, business innovation and civil society.

https://theodi.org/what-is-data-infrastructure

A data infrastructure consists of data assets, the organisations that operate and maintain them and guides describing how to use and manage the data. Trustworthy data infrastructure is sustainably funded and has oversight that provides direction to maximise data use and value by meeting the needs of society.

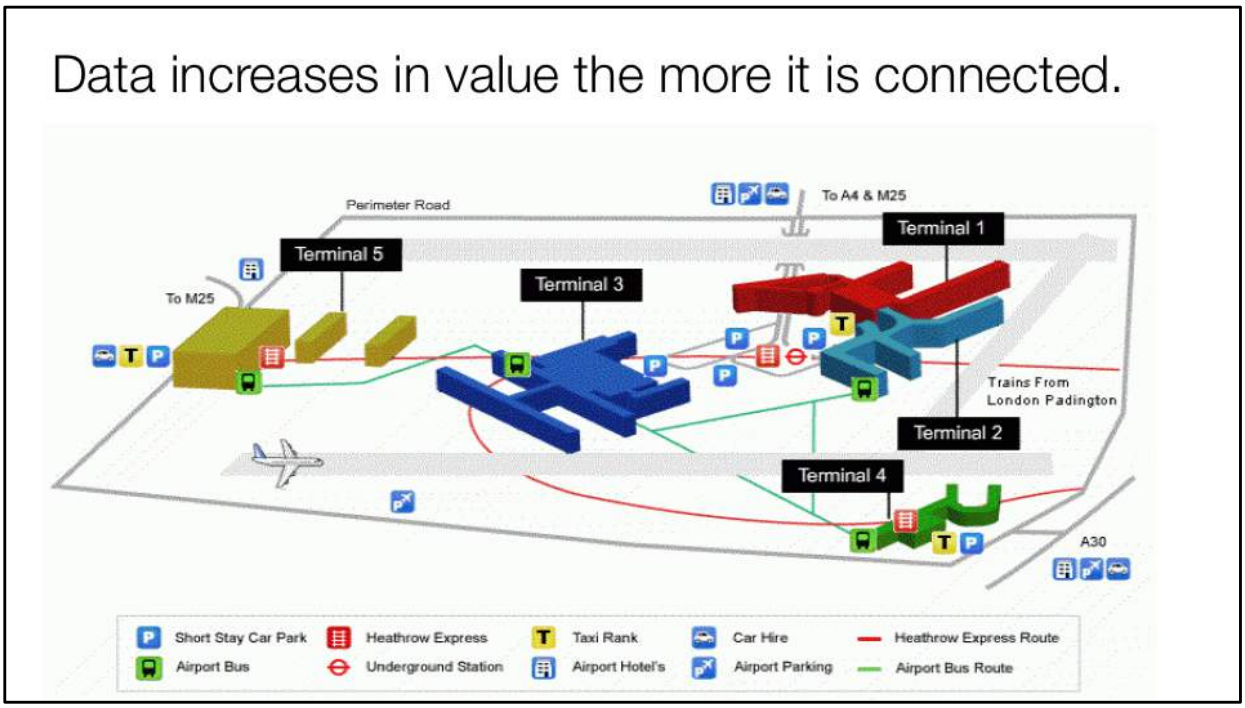
Data infrastructure includes technology, processes and organisation.

Like the data spectrum, not all physical transport networks are open. But we have to be very careful where the boundaries are placed.

In this example, some roads may be closed and only accessible to certain individuals, such as the military. Other roads may require payment of a fee or require users to be the member of a club to access the road.

The majority of roads are open.

That doesn't always mean these roads are managed by the public, some still might be private but allow public open access. This is often the case with business estates and airports. Other roads might be run and managed by the military but allow public access. Getting the access right to physical infrastructure is critical to allow fairness and the opportunity for everyone to trade.



Maximum value comes when everything is connected. Think of a private road like a private database, not connected to anything, not providing any wider value.
As a business, if your physical assets are not **moving**, we know that there is more value to be unlocked.
The same is true for data. It's value is based on the number of connections it has.
You need to do everything you can to **increase the number of connections** your data has. Otherwise you are losing value.
As in economics, reducing transactional friction increases the number of transactions.
Reducing friction in the digital economy is about increasing interoperability. In the world of data, designing for open is the single best bet you can make.
So, stop driving around in your silos.
Unlike physical infrastructure, data can have an unlimited amount of connections, so you need to think on a global, web scale.

Standards drive connection

"I've been here before?"

The aviation industry is one of the prime examples of a global standard.

The aviation industry is one of the prime examples of a global standard. If it wasn't global, then there would be no real aviation industry.

In your groups I would like you to list all the standards you can think of that ensure the aviation industry and airports work effectively for both passengers, businesses and transport operators.

On this slide we can already see the standard for labelling gates (Letter followed by a minimum of two numbers). We can also see the standard for signs that must be labelled in the primary language and English with all gates using the English alphabet.

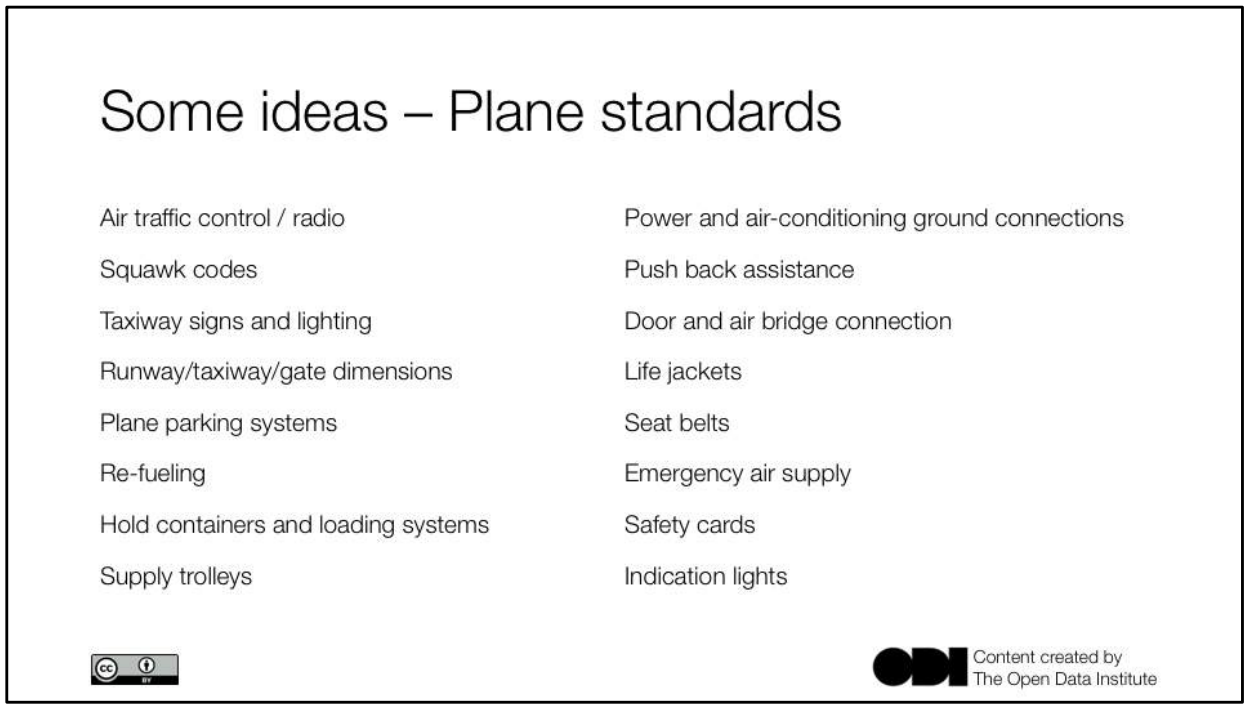
Exercise

Flickr/John Murphy

List all the standards you can think of that ensure the aviation industry and airports work effectively for both passengers, businesses and transport operators.

In your groups I would like you to list all the standards you can think of that ensure the aviation industry and airports work effectively for both passengers, businesses and transport operators.

Please write one idea per post-it



Here are some of the standards that I could think of in relation to the plane and opportator standards

Non-standards

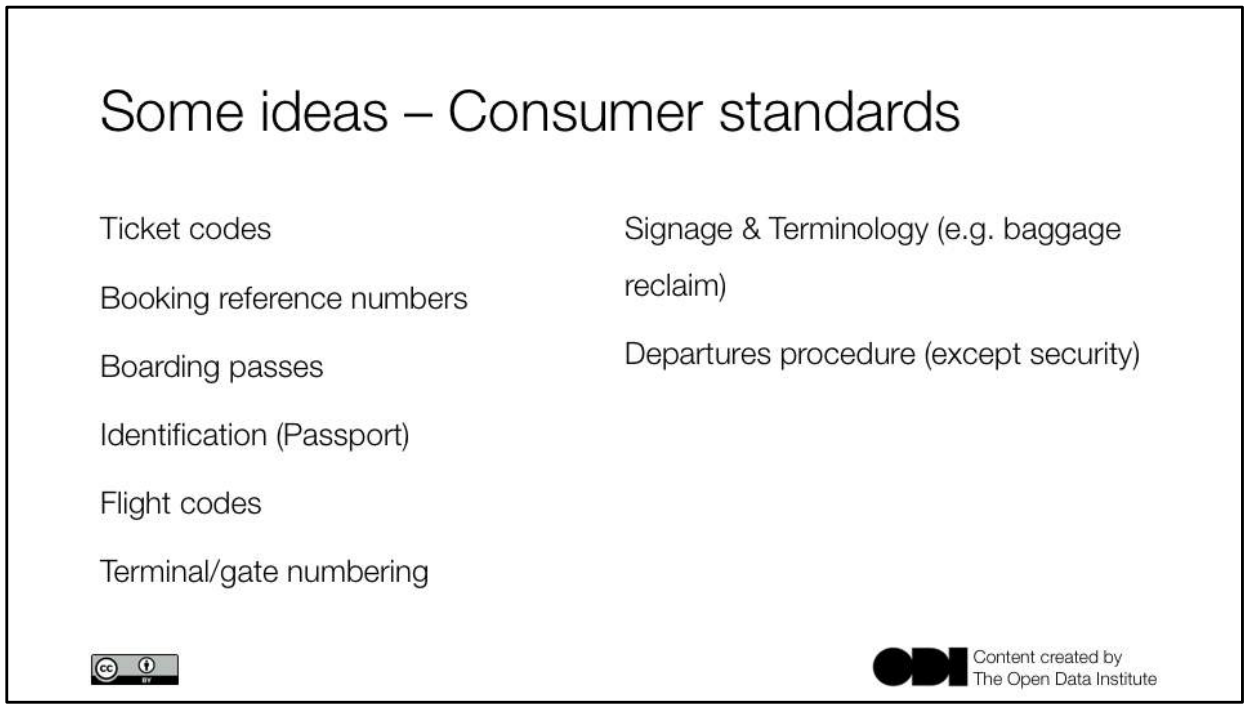
Leg room and seats

Entertainment systems

Luggage policies (even with different operators of the same place)

Cabin layouts

Here are a few of the areas where there are no standards and operators compete, sometimes to the annoyance of passengers.



Here are some of the standards that I could think of in relation to the consumer

Consumer – non-standards

Arrival time advice for departure

Immigration forms

Customs and visa requirements

Arrivals procedure

And some consumer non-standards

While departures is a reasonably consistent process. Arrivals is very much dependent on your nationality and the procedures of your destination. Occasionally arrival procedure can be some complex that it all ends up looking very messy for visitors. Which is not a great first experience.

Surely with all the data we collect today there is a way to make this more efficient?

Standards in Railways

List all the different aspects of a railway network that are required to ensure a fully connected global network.

Sort these into which are globally standardised and which are not.

Let’s do the same exercise again with rail transport systems. This time think of the standards first and then try and identify which are globally standardised or not.

(Non)-Standards in railways

Gauge (of track and stock)

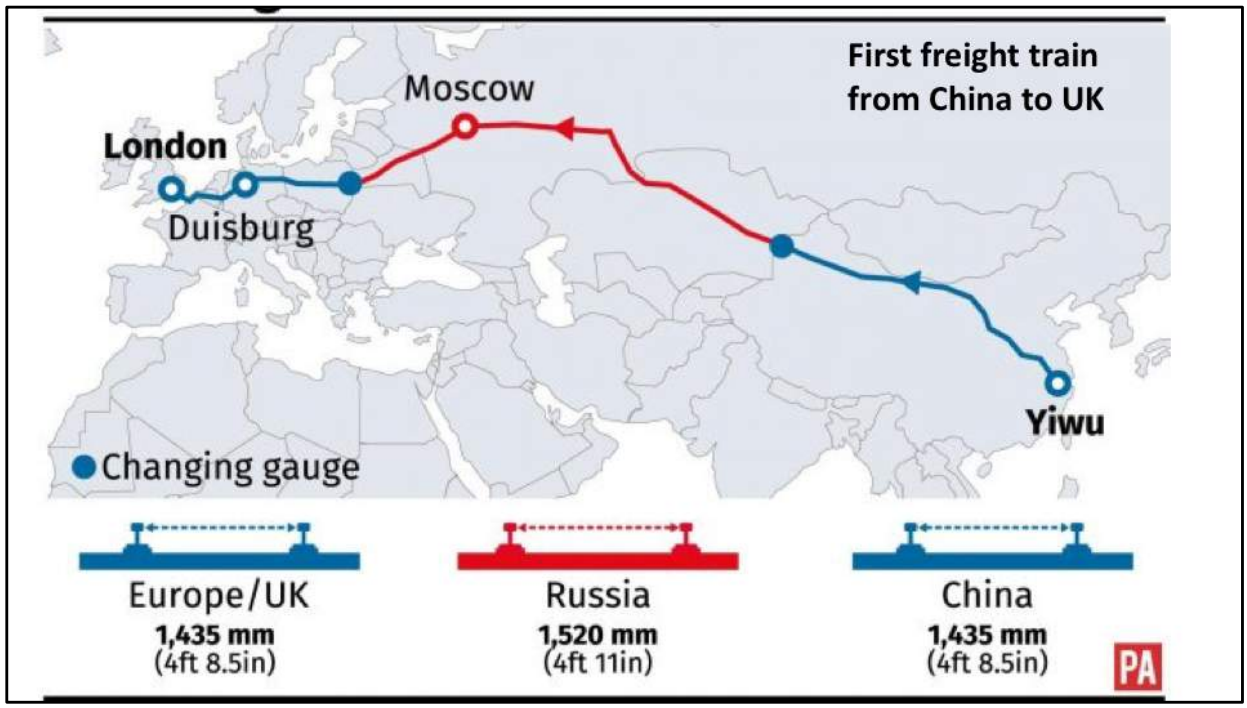
Signals

Power and voltage

Platform height

Classes

Payment system



On January 18th this year the first freight train arrived in the UK from China. This sounds like a straight forward task however was not a simple as driving a single train from the China to the UK. In fact the only part of the train that made it all the way from China to the UK was the containers, which is impressive in itself as they had to fit through the channel tunnel.
The containers were taken through china on a standard gauge railway then transferred onto new wagons for the trip through Kazakstan and Russian which is a Russian guage. The containers were then offloaded onto new wagons again for the journey through Europe to Calais where once again they were offloaded onto wagons purpose built for the channel tunnel.
Finally the wagons arrived in London, ironically being propelled the final mile by a new loco as the loco that came from France was electric and the last mile wasn't electrified. So none of the wagons or locos made it from China to the UK however the containers did.

This is still impressive given that gauging requirements in railways are not just about the track but also about the size and width of carriages and containers to ensure they don't strike other trains, stations or other equipment. Again though, unlike planes, the different systems use different signals on different standards of line and require

drivers with specialist knowledge in each location.

Still, this sort of connection shows that it is possible with minimal changing between origin and destination.

Data standards

Like railways, data also has many standards that affect its global usability.

List aspects of data that are required to ensure its global usability, e.g. format.

Like railways, data also has many standards that affect its global usability.

List as many of the aspects that are important to data standards as possible.

We will use these to develop a canvas of how each is controlled on the web later.

Parts of a data standard

Access

Structure

Format

Description

Rights

There are five main aspects that will make data usable by everyone.

The first is access. How do people access the data? A a provider do you create an API? does the API match others APIs? Is there a standard already set out for access to the type of data you hold?

The second is structure. Not all data is tabular. Recognising the structure of your data is key to picking the right format and set of standard to describe and allow access to your data.

The third is format. Which formats do you provide the data in? In digital we can do more than one. Choose formats that are appropriate for you users and already widely used.

Fourth is description. How is the data described? What column headings or keys are used that mean the data can be easily combined with others data? Are the values and ranges of values in the data also comparable with others so they can be combined?

The final and most important aspect is use. Does the license you provide the data

under maximize the potential benefit or are you creating a road that no-one can use just in case one person abuses it?

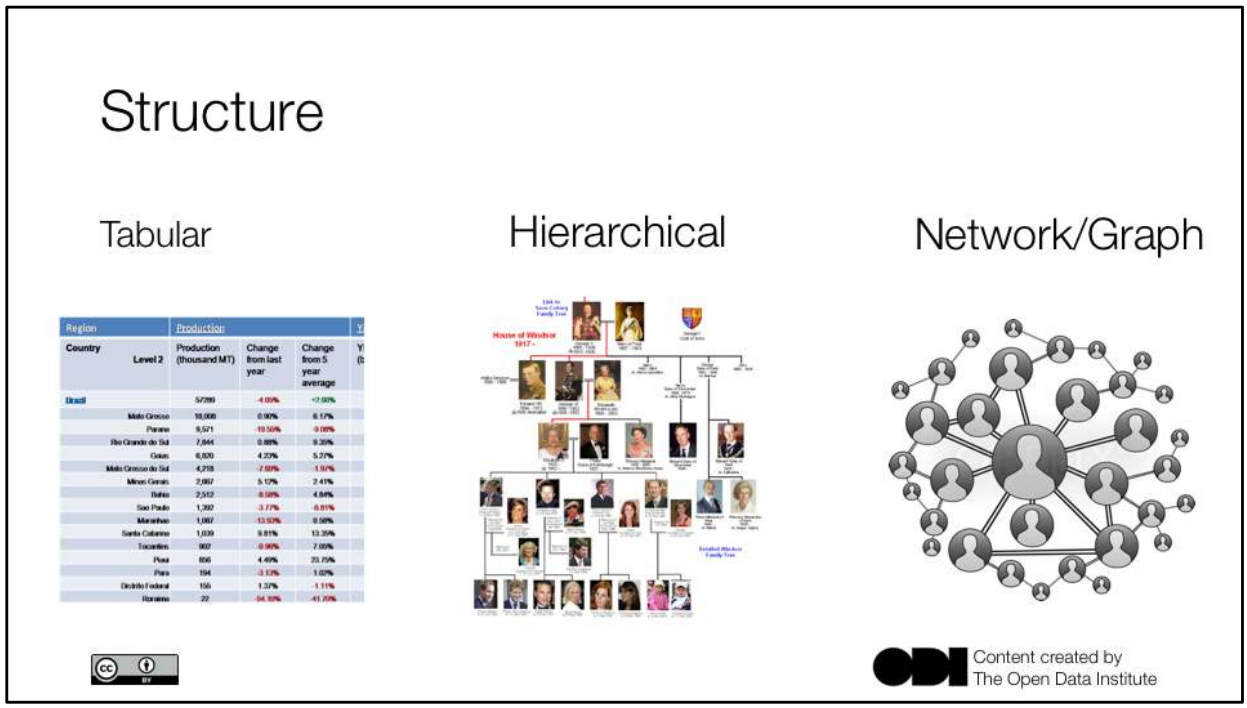
Before we talk about standards, lets talk a little about the platform which led us into the information age, the web.

Here is a picture from the opening ceremony from the London 2012 Olympics. Following your own incredible opening ceremony in 2008, we put the inventor of the web, Tim Berners-Lee in an inflatable house sitting in front of an 80s computer. More importantly was the message that he chose to be displayed around the arena at this time.

“This is for Everyone”

One of the most significant things that Tim chose to do when he invented the web was to not try and protect or commeccialise it. He realised the incredible potential if it was to be remain open, for everyone to use.

Just like a road, the majority of users a good users and respect the rights of others and thus the web continues to evolve for the good of the majority of people. However others have attempted to build areas or technologies on top of the web that have not been for the benefit of everyone.



There are three main structures that data can take.

Tabular
The most common structure for data is tabular. Data is organised into rows and columns listing sequential values, such as expenditure.

Hierarchical
Hierarchical data shows the relationships between data points, such as a family tree or municipalities in each country. In a hierarchical dataset the relationship between parent and child can only exist in one direction.

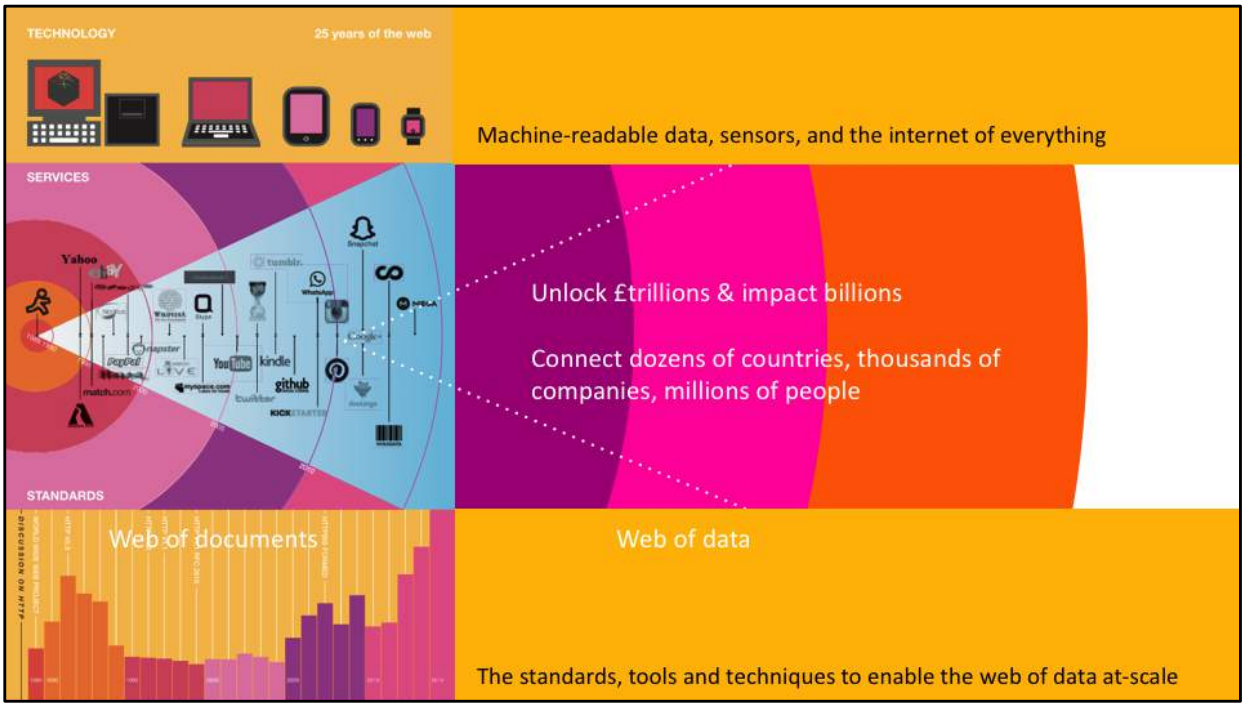
Network
Network structured data allows relationships to exist between any combination of elements in any direction.
A good example of a network data structure is a social network. Think of your network of friends and their friends on Facebook; consider first, second and third degree contacts on LinkedIn.
In a network structure the same relationship can exist between two or more nodes. I can be your friend and you can be my friend. I cannot be you parent if you are my

parent, that would be hierarchical.
The Web is another example of a network data structure, where webpages link to any number of other pages in any direction.

Formats

There are thousands of file formats out there. Each suiting different types of data in different structures and created by different orgnisations and communities.

So how do file and data formats standardise? The answer to this is in the evolution of the web.



Over the past 25 years the web has grown rapidly and is connecting more with our physical world. We have been through the web of documents and social networks and are now in the age of data.

In these years we have seen the birth and death of many proposed standards as well, including flash and Silverlight for video and more recently mp3 audio standard. Technology companies have started a shift away from proprietary technologies for distribution of music, eBooks, pictures and movies in favour of agreed global standards that everyone can purchase on every device. Data and content is now more important than the platform. In a modern connected age we expect access to our content everywhere on every device regardless of the manufacturer and software on that device.

While content and information has seen a settling a standards, with even the inventor of MP3 recommending another's standards, the same cannot yet be said about data.

There are still new and emerging ways of representing and sharing data, each of while requires users to understand a new system. Even though the amount of data is already beginning to dwarf the size of the existing web of information, we are at the

first stages of seeing the emergence of global standards for representing and sharing data.

Standards on the web

You cannot dictate a standard. Standards emerge when a large body of users agree that is the standard.

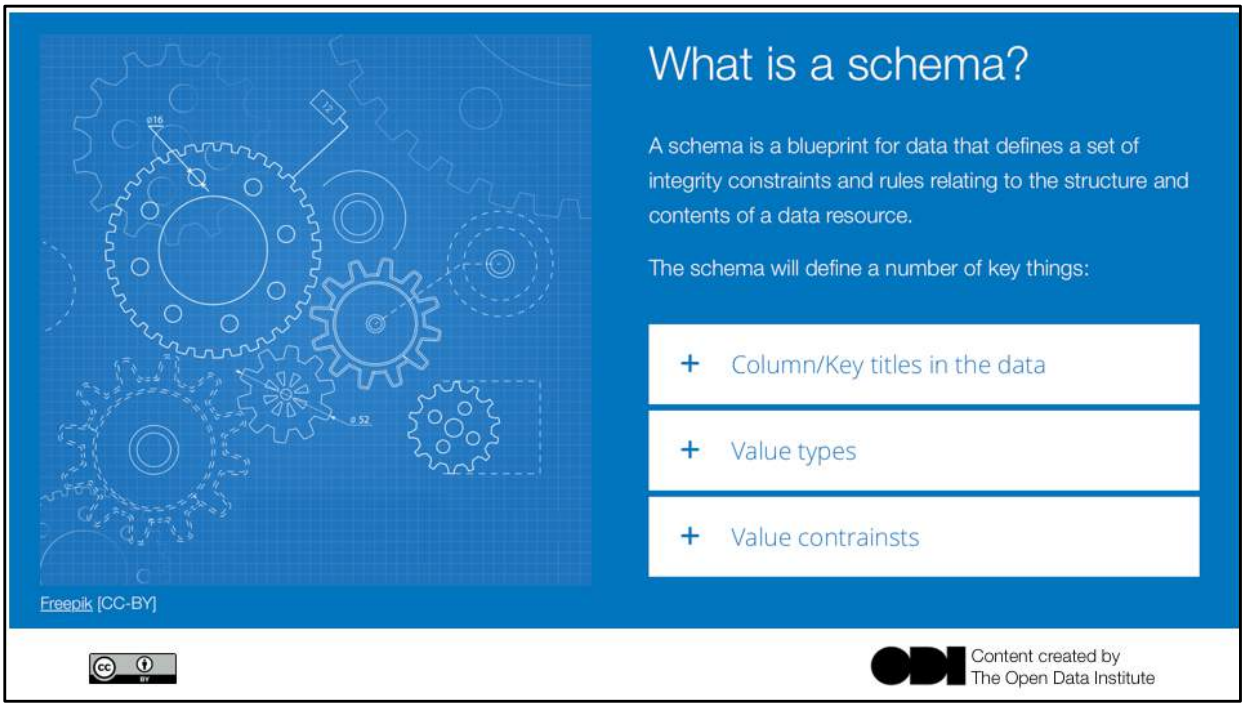
You cannot dictate a standard. Standards emerge when a large body of users agree that is the standard.

In some areas of physical infrastructure it is possible to dictate standards. However, early rail gauges were all over the place so cities could not be connected. It wasn't until a number of large organisations agreed on a standard that others had to forcibly change in order to stay connected.

We are currently seeing the same trend with electric cars. There is not yet one standard for how to refuel these cars thus it is not possible to provide a connected infrastructure efficiently. Until a standard emerges we will not see an explosion in the use of all electric cars. We can also look at phone charging cables to see how we are polluting landfill with cables. We don't have this problem with power sockets in the home, but power sockets on appliances are not standardised.

Data standards also emerge from widespread use and can be regulated once adopted reaches a critical mass. The set of web standards are reviewed every so often by looking at how the web has evolved. This is done by evaluating new technologies the top 1000 websites are using and then adopting the new techniques that are common

across them all into the official standard. The key to successful standards is having a large set of adopters. A standard invented and adopted by a single organisation or person will almost certainly fail.



The next important aspect of a data standard is the description (or schema).

A schema is a blueprint for data that defines a set of integrity constraints and rules relating to the structure and contents of a data resource.

It consists of three key aspects:

Column/Key titles in the data
Defining a consistent set of column titles (or keys) for a data set is essential to ensure that datasets of the same type can be merged and analysed easily. Often column titles will change or be abbreviated to save time however this causes a lot of problems when analysing data over long periods of time. Adding column titles is less of a problem but has to be taken into account when analysing data.

Value types
With the column title/keys defined it is important to define the valid data type for the values, e.g. number, text, date, co-ordinate etc.
This will help with a simple datatype validation.

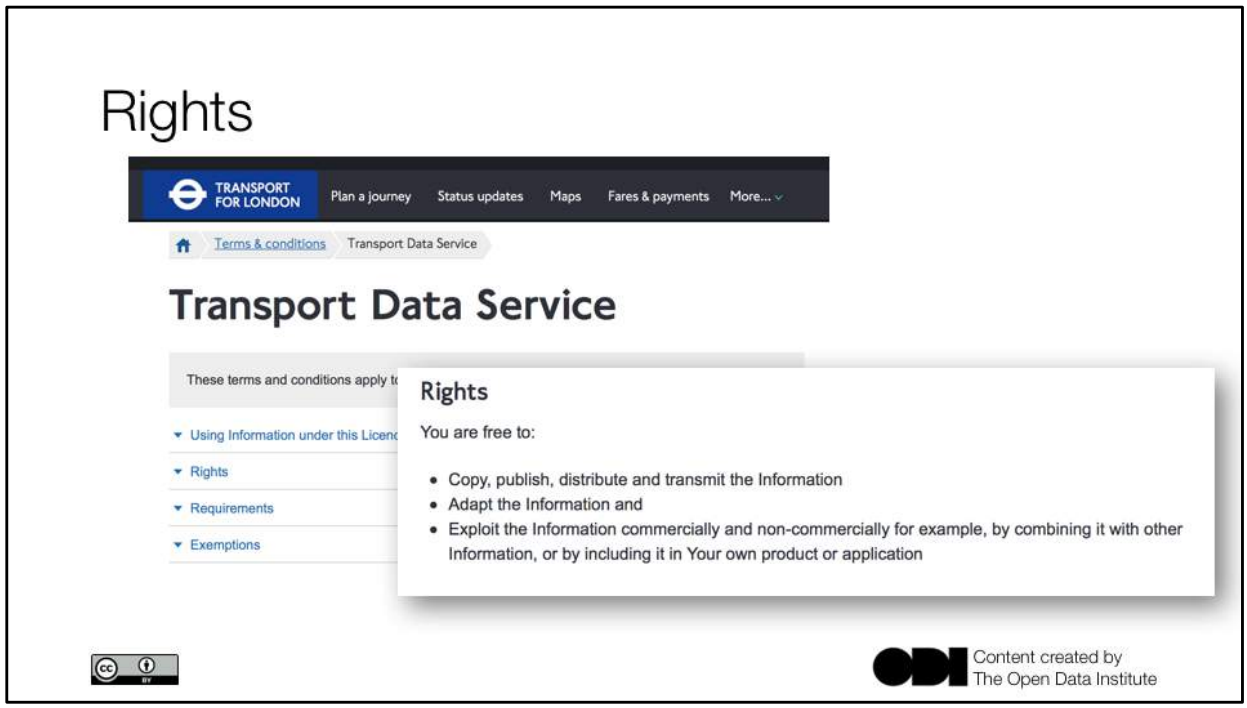
Value constraints

With the value type defined, valid constraints such as being required, need to be unique, being in a certain unit (e.g. Gallons (UK)) or be within a certain range should be defined.

For example a column might be entitled "*Cost (£m)*"; thus any values should be numbers (without commas).

Setting a valid range also help avoid and explain any errors in the data. For example setting a range of 0.001-100 on the "*Cost (£m)*" (if it is known that cost cannot exceed £100m).

Range validation stops people accidentally misreading the column title/units and entering 100000000 instead of 100 for £100m.



Rights for use we have already looked at. This is one of the most important aspects if you want people to be able to gain benefit from your data.

As mentioned before. Transport for London have a very clear rights statement which allows complete use of the available data.

In order to drive standards, we need to realise the demand, get a set of adopters to push policies and standards forward and improve the experience for the consumer.

To close this session lets have a look at standards in action with the most common transport activity, route planning.

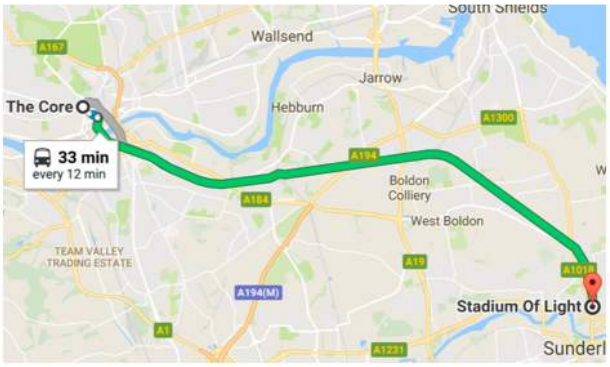
So what does the consumer need to plan a journey from A to B?

Lets have a go at planning some journeys. In your groups see if you can find how to get from point A to point B in the requirements above.

I would like to know how you get there, which forms of transport and how much it will cost. You must use at least one form of public transport.

Bonus points awarded for the cheapest route and the quickest route.

The Core to Stadium of Light



Cheapest

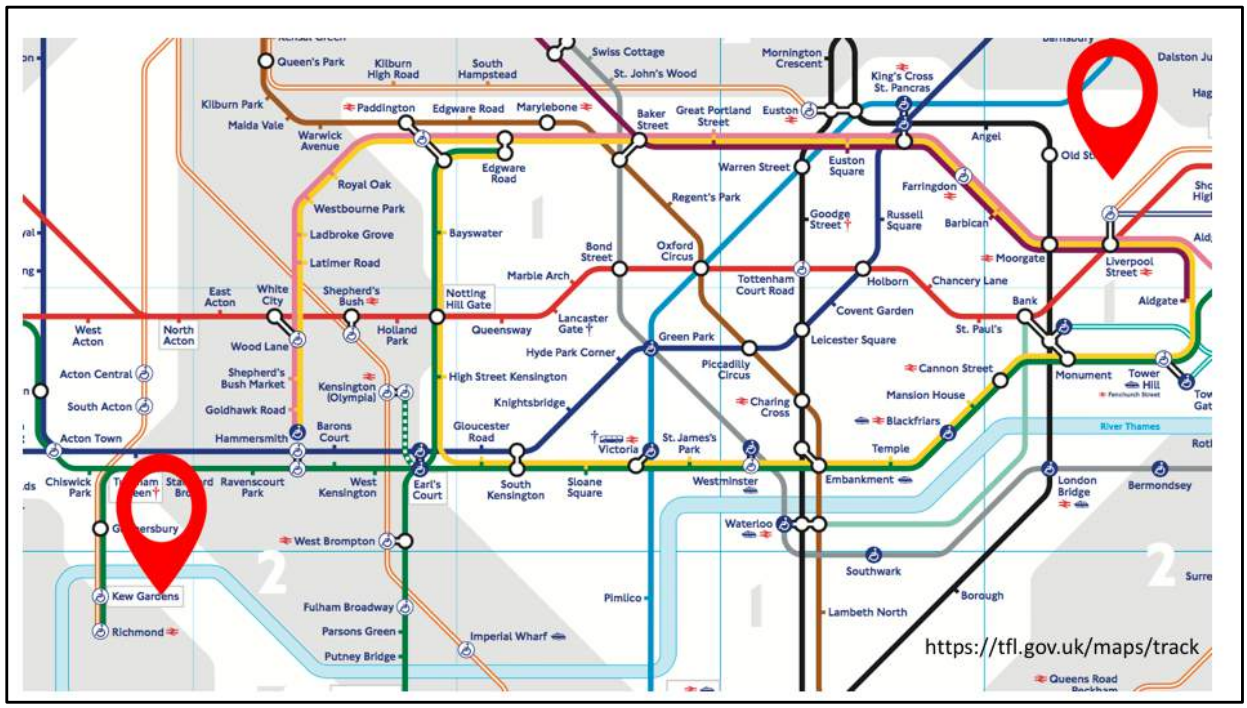
Metro: £3.30

Fastest

Uber taxi: £21-£30

The first one is fairly simple. The fastest option is the metro and the single all zones ticket is £3.30 that you can buy at the ticket machine.

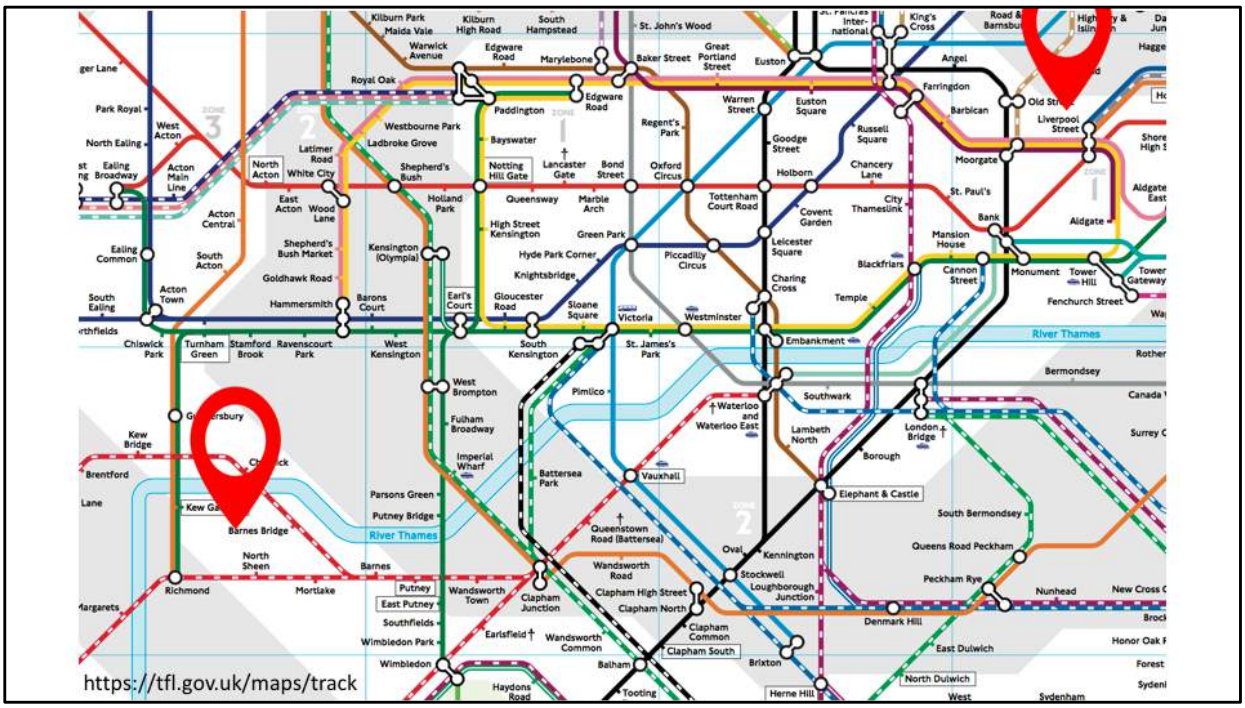
The fastest route is to get an uber taxi which only shows up on the mobile version of google maps, but it does show the price. To find the price of the metro I had to go to the metro website.



Lets have a look at this example is a bit more detail to find out why, even in London, it is not very simple to route plan.

Here is the London tube map.

I have marked the ODI at the top right and the National archives at the bottom left. Here it appears that we need to get to the green district line to Kew Gardens. This would be a journey from zone 1 to zone 3 costing £3.30.

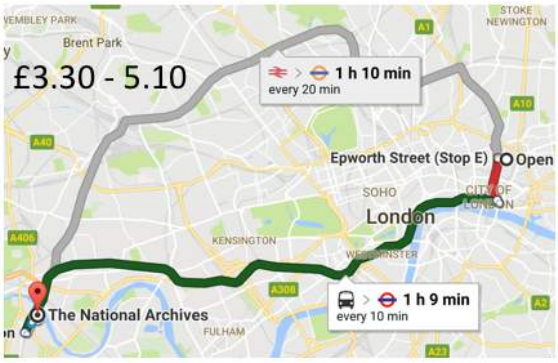


However if we add the overground rail network to the map it gets a bit more complicated.

The district line has many stops between the ODI and National Archives and there is a more integrated route from Liverpool street to North Sheen / Richmond or Kew Bridge using the over ground railway. This route is faster and the same cost.

The time is also the same as the walk or bus to the district line is long near the ODI, but the walk to Liverpool street to get the central line is short. However the walk from Kew Bridge using the second option is 15 minutes but is a very pleasant one along the bank of the Themes.

The ODI to National Archives



13:14 - 14:05 **51** mins

Liverpool Street Station at 13:14
Central line to Bank
1 min
£4.90

Bank DLR Station at 13:22
Waterloo & City line to Waterloo
3 min

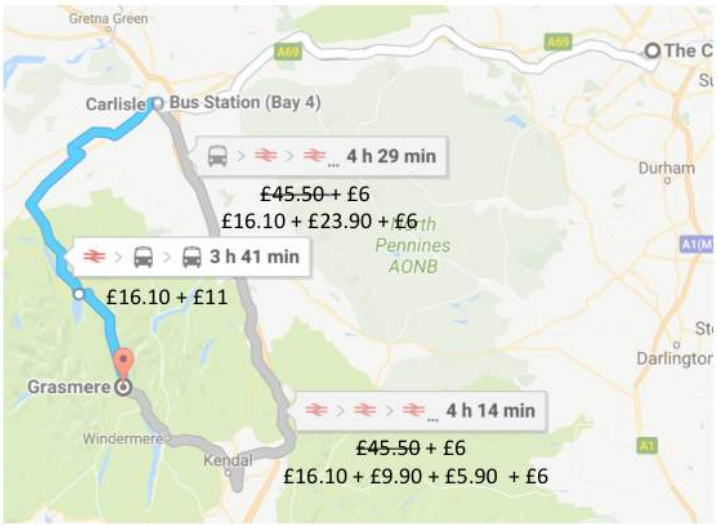
Waterloo Station at 13:37
South Western Railway to Kew Bridge
28 min View stops

Kew Bridge Station at 14:05

https://tfl.gov.uk/

Google says to get on a bus and then the district line. These two forms of transport will charge you twice. E.g. a bus and tube fare. However on the right is a route which just uses rail as the form of transport and will only change you a single fare. Depending on the route you take and where you get off this fare could vary from £3.30 to £5.10 as Kew is also on a zone boundary so getting of one stop earlier and walking can save a fair amount of money. Either way it will take about 1h15 using any of these routes due to the walking either end.

The Core to Grasmere



Here is the final one. Here we have a number of options. At the time of searching taking a train the Carlisle and then two busses for two hours is the fastest and cheapest route.

There is a route that involves lots more trains and takes you via lake Windermere. However, getting the prices for this one is very complex. Buying a single through ticket is £14 more expensive than buying three separate tickets for each train journey in the chain. This happens as you are transitioning between zones of different transport providers and an integrated ticket is more expensive than each providers best offer on the day.

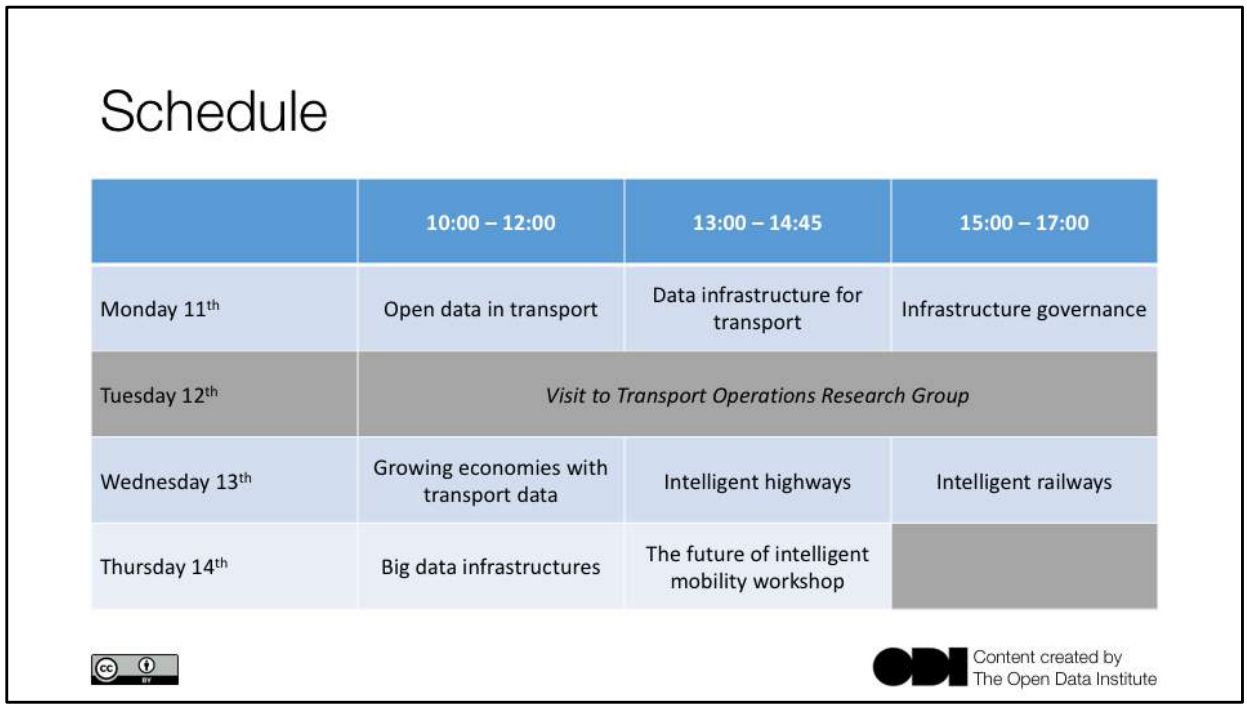
There is a motion tabled that means that in the future ticket machines and online systems much show you the cheapest price, even if it means buying multiple tickets. A lot more data is needed to drive such a system. After the break we will talk about the governance and policies which affect all three of these examples.

Conclusion

Routing data is much easier to get hold of.

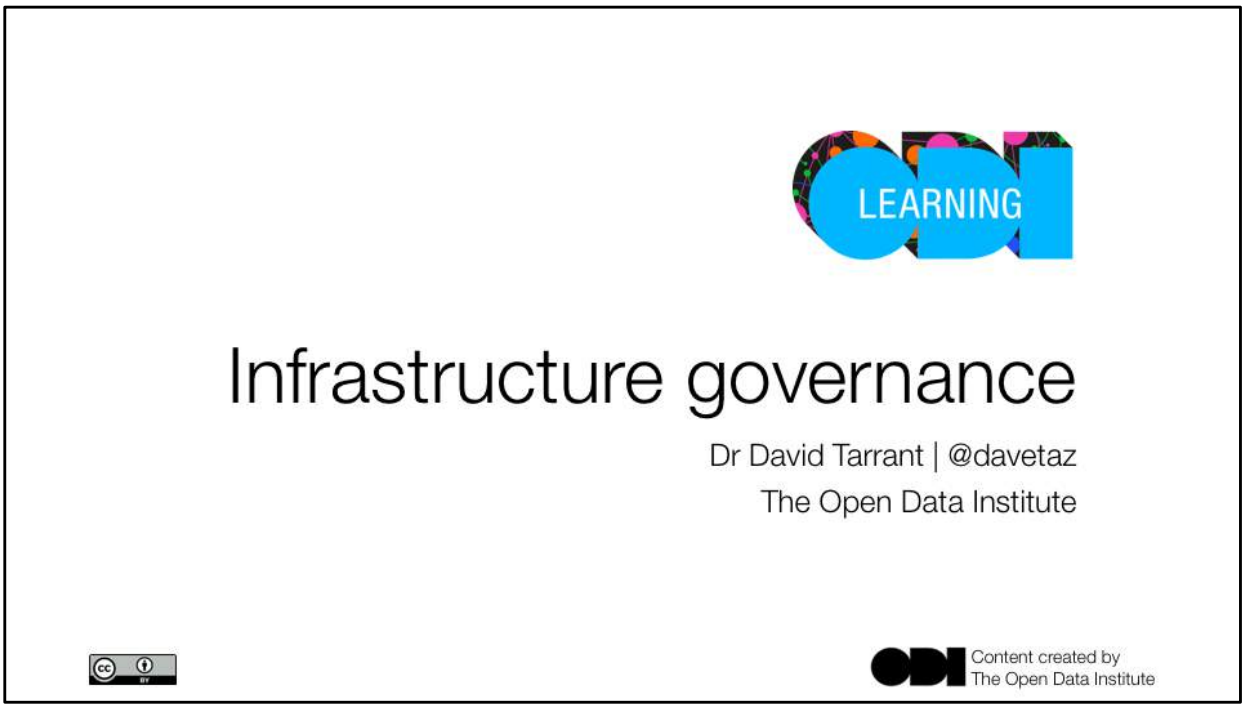
However any other data, price, how to buy tickets and how to get assistance is still pretty difficult. What about when delays happen or other incidents?

This is something we will look at in more depth tomorrow.



It is clear we need a good data infrastructure to support service and apps that support consumers, we'll look at more of them tomorrow. The final session of today looks at governance models and who is responsible for each part of the physical and data infrastructure. This session will map out who you think should run such services before looking at how it works in key parts of the UK.

See you at 15:00



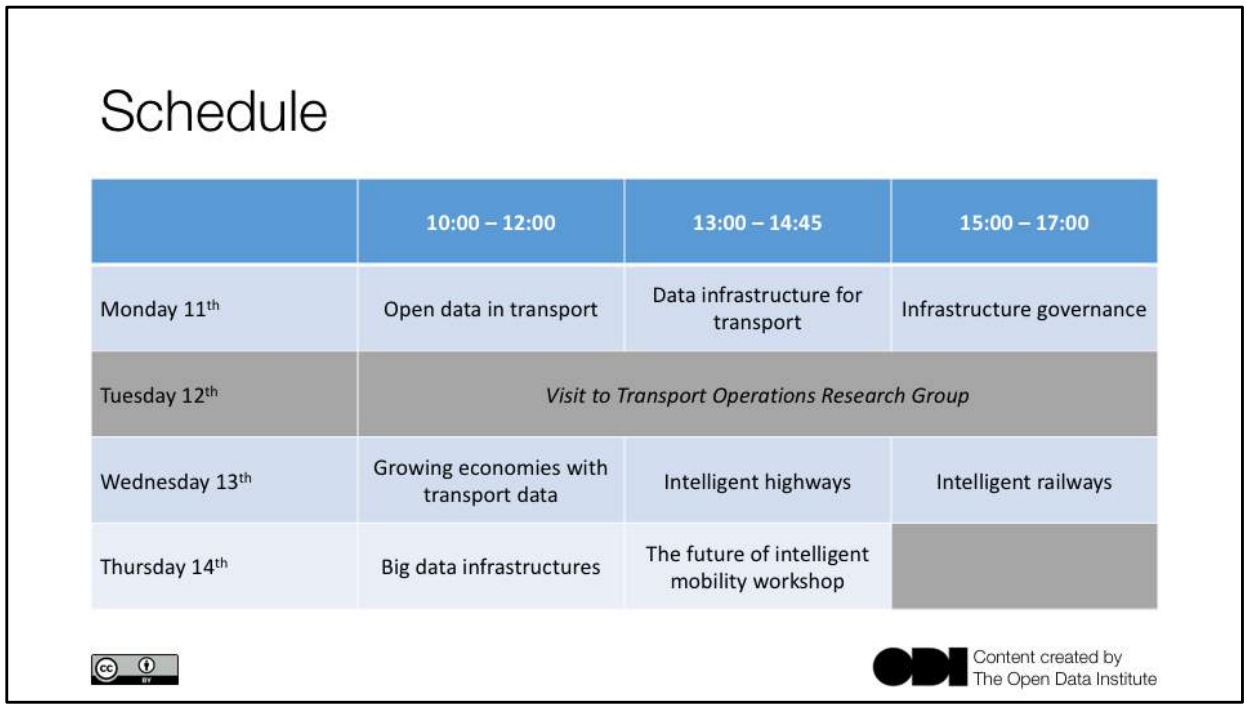
Hello and welcome this weeks training entitled Transport Data Infrastructure.

My name is Dr David Tarrant from the Open Data Institute.

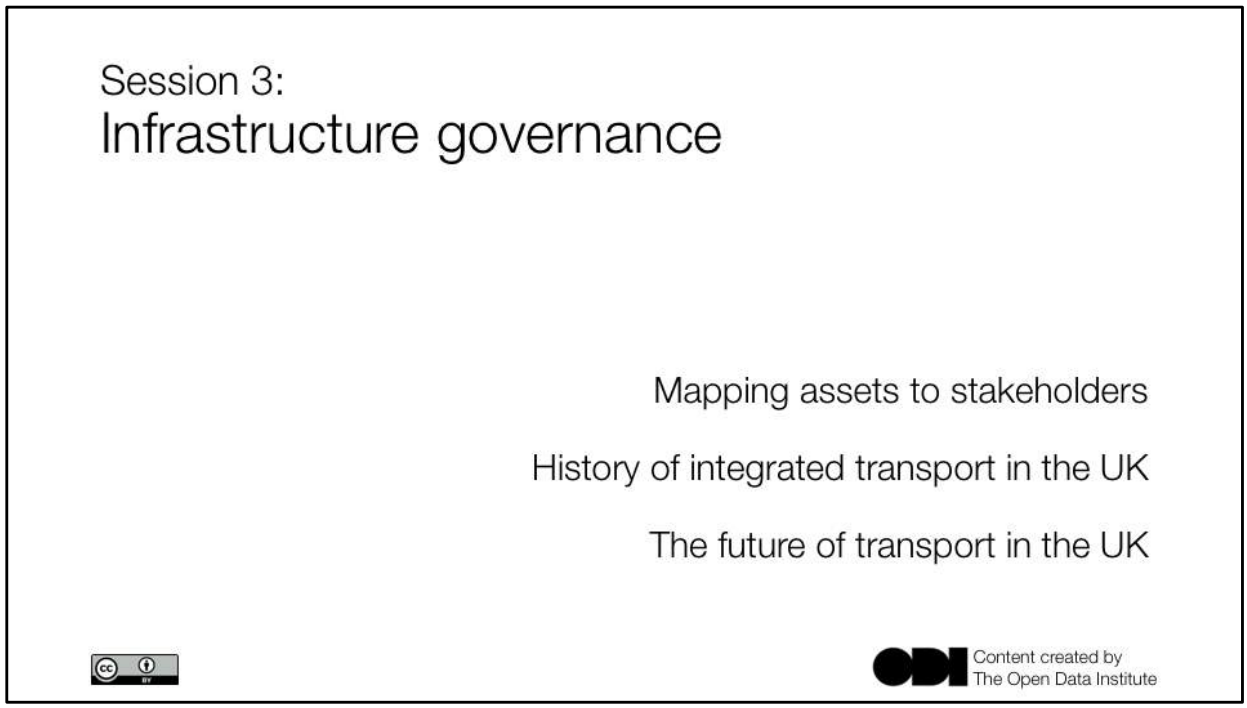
Over the course of this week we are going to be looking at how the UK is attempting to deliver a world leading transport infrastructure using data.

The Open Data Institute has been carrying out a lot of work in this area and has partnered with the top organisations to both deliver services and evaluate impact.

One of the organisations we have been working with is the Transport Systems Catapult (who I believe you may have met). As part of the work we undertook with them we looked at the impact of data in transport and estimate that worldwide there is £1.2tn…



Welcome back,

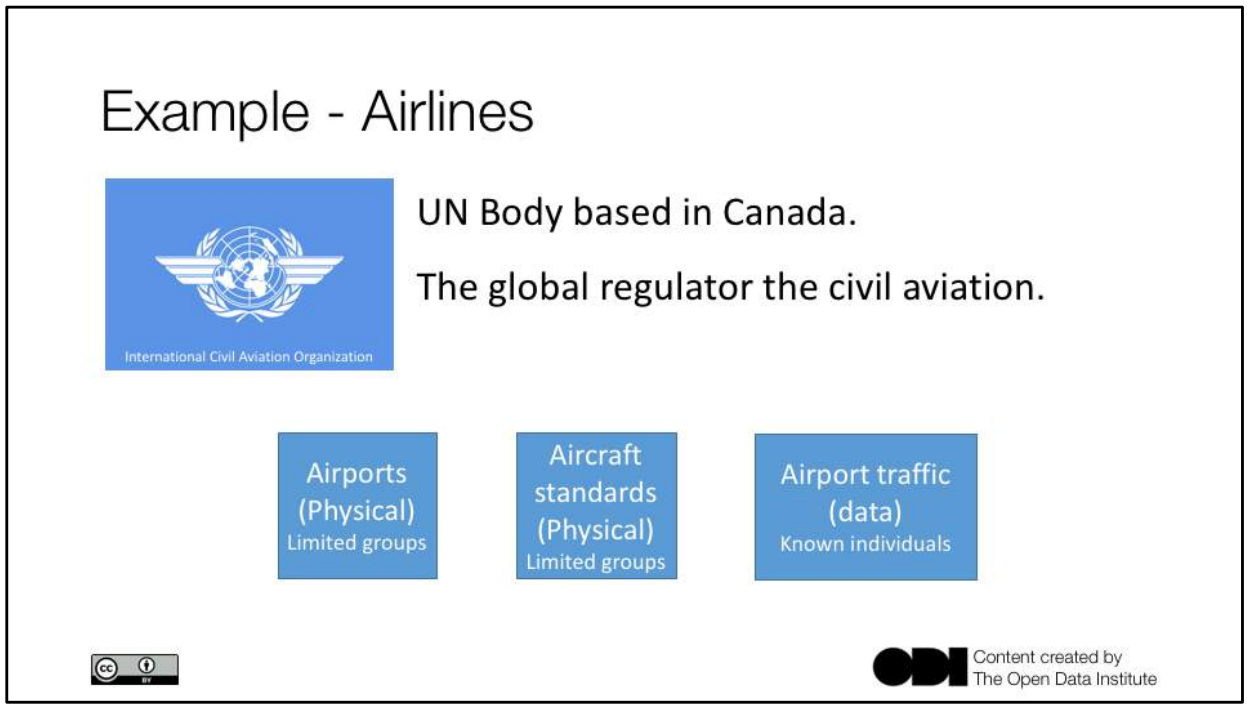


Having looked this morning at the value that a open transport data infrastructure such as the one TFL has can provide we are now going to look at what an infrastructure is, the key role that standards play and how data standards and sharing is as important for data infrastructure as for physical infrastructure.

We will then look in this session at how difficult it still is to use transport data before looking in the final session about how governance and policy is important in getting it right.

Mapping exercise

1) Take your data spectrum canvas from earlier
2) Add to it the **physical infrastructure** for your **transport mode**, e.g. track, signals, station signage for railways. Make sure each physical aspect is also in the right place on the canvas (e.g. open, shared, closed)
3) Now add to the post-its who should control and who should provide each item. e.g. should the government control the timetables or private sector providers. Should the government manage the purchase of busses or the private sector?



Civil aviation is very tightly regulated to help ensure the highest levels of safety. Basic international regulations are set by a United Nations body called the International Civil Aviation Organisation. Individual national regulators then take these regulations implementing and enforcing them in their own country. They may also add to them to further raise safety levels.

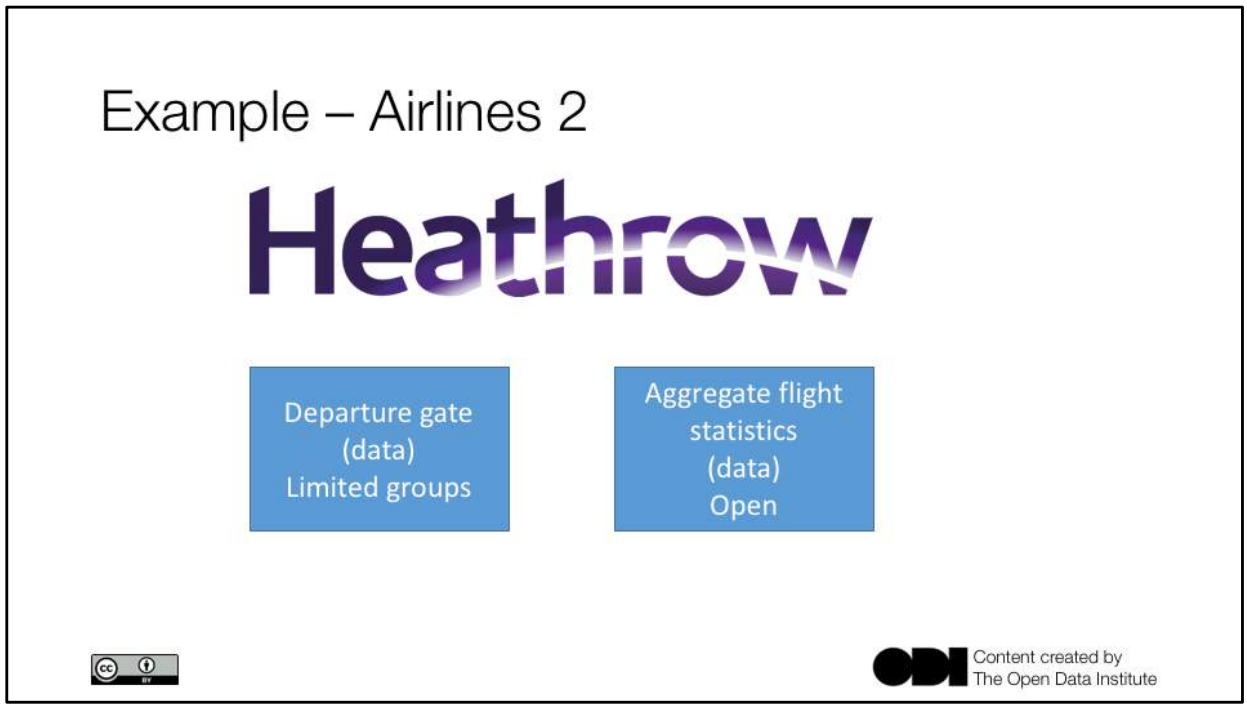
Within Europe much of the safety regulations are set by a European Commission body called the European Aviation Safety Agency. This means there is a common set of requirements across Europe on areas like pilot licensing and aircraft type approvals. National regulators, such as the UK CAA, then use those requirements to regulate civil aviation in their country.

You can download all of the annex's and standards from the ICAO. So can discover what the airport regulations are if you want to open an airport. It is unlikely however that anyone can use an airport (like a road), especially not a civil one.

Airports in the UK themselves are built and owned by private sector organisations, who have to follow the regulations. A number of years ago the monopoly of ownership on the London airports was split up which has led to major investment in

Gatwick in order to compete with Heathrow. Heathrow has also announced new connections with the TfL infrastructure and cheaper pricing to and from the airport, something Gatwick does not have, resulting in slow passenger movement at the station which causes congestion and bad experiences.

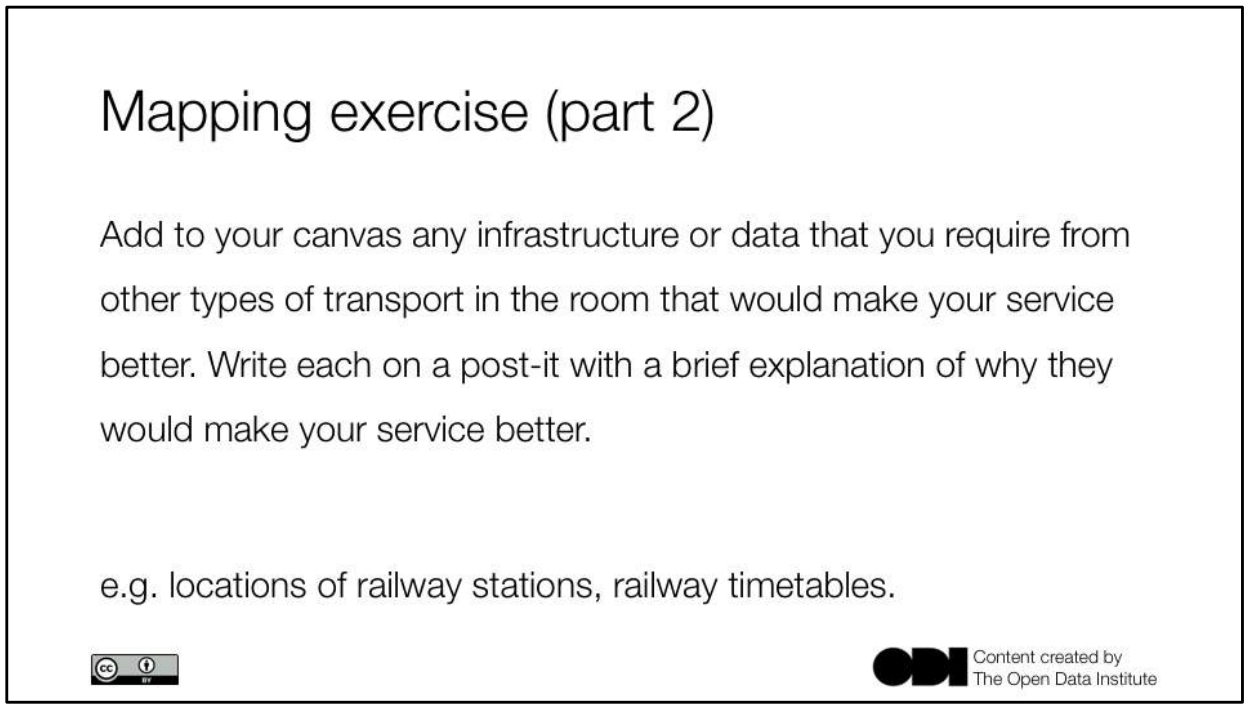
Interestingly the ICAO make 6 large datasets available to users, however these cost upwards of $1,000 a year to access and only come with a single user licence. Thus these are only accessible to known individuals.



Heathrow airport manage live departure gate data and release it to limited groups (such as the airlines) to integrate into mobile applications and websites. The availability of this data varies widely from airport to airport and is why Google can only sometime tell you the departure gate.

On the other hand Heathrow do public high level aggregate statistics about the number of flights be day. While useful information, this is not really raw data and not much can be done with it other than to look at trends over time.

The airline industry is fantastic about regulating standards internationally however when to comes to data, they are well behind other transport sectors with many taking a harsh commercial view over flight schedules, departure gate data and other items of data that would help commuters with a smoother departure and arrival experience. Although services like flight radar exist, the contributors of the data wish to remain in the control of the data and are not happy about commercialisation of their data.



Do a similar exercise for your own data and infrastructure to see who it is that controls or regulates your data.

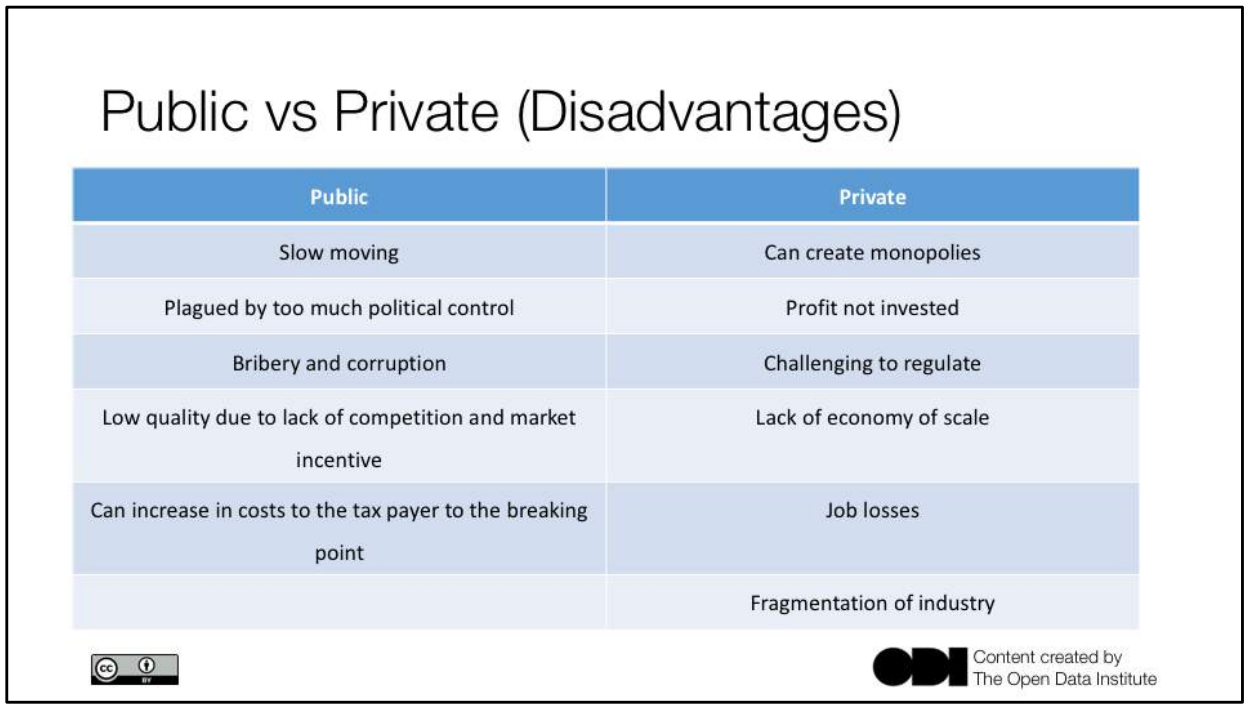
Public vs Private (Advantages)

Public	Private
Non-profit, all money invested	Increased competition
Easier to influence economic activity	Focus on customer service
No abuse of power when monopoly	Improvement to local services
Easier to plan nationally	Improved efficiency
Can be subsidised	Separation from short term political agendas
	Forced to adapt to suit the market

In your groups you might have gone all public/government run or all private or a mixture of the two, just like airports.

There are arguments to be made for both models. Here are a few of the advantages of each model.

One of the key considerations is what is regulated by a public body. This is an essential part of working with the private sector in many cases and can maximise innovation. Sometimes however regulation can get in the way of progress, as I will discuss later.



Here are a number of disadvantages.

How you work with the private sector depends very much on the service being provided.

An industry like telecoms is a typical industry where the incentive of profit can help increase efficiency.

However, if you apply it to industries like health care (such as the UK National Heath Service, which is free at the point of use) or public transport (which is not free) the profit motive is less important.

It depends on the quality of regulation. Do regulators make the privatised firms meet certain standards of service and keep prices low?

Is the market contestable and competitive?

Creating a private monopoly may harm consumer interests, but if the market is highly competitive, there is greater scope for efficiency savings.

The UK has just privatised the Royal Mail, however as part of the sale, the UK

government sold off the addresses database so now all government departments have to pay to access basic data on street addresses for sending official documents. The ODI was not against the privatisation of the delivery of letters, but the data infrastructure relating to locations of businesses and houses should remain a state asset.

To look at how the canvas of physical and data infrastructure is provided in the UK transport sector I'm going to look at the history of road and rail transport systems in the UK. We will build on this over the following sessions when we explore each in more depth. The last part of today is designed to give you the essential knowledge on the structure of transport companies, providers and regulators in the UK.

History of transport in the UK

The railways led the transport revolution in the UK, we invented them to carry goods but they soon carried people as well.

The UK created the worlds first steam powered railway in 1830 between Liverpool and Manchester.

It was the first to be entirely double track throughout its length; the first to have a signalling system; the first to be fully timetabled; the first to be powered entirely by its own motive power; and the first to carry mail.

Other than passengers the railway was primarily used to transport valuable textiles to be exported via boat to international markets in exchange for tea, no surprise there.

Early railways

Early railways lacked regulation.

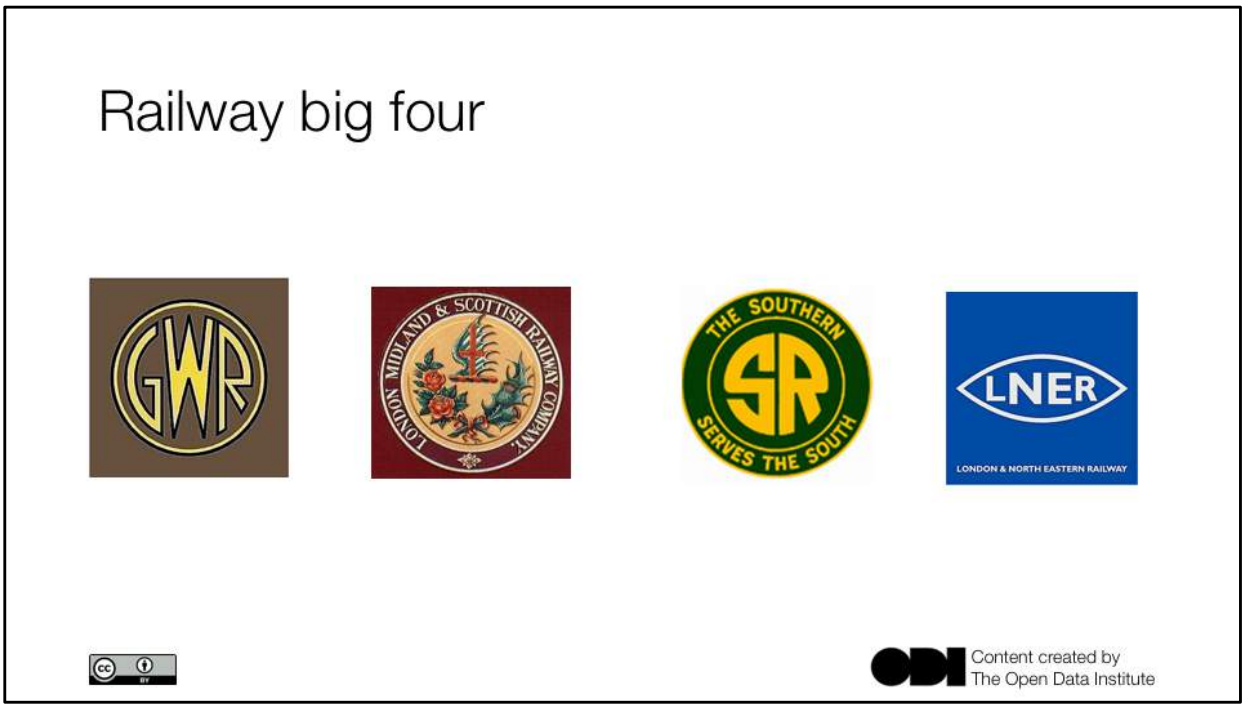
They were built by private investors who went railway mad in a vicious money making frenzy.

In 1846 gauges were regulated but the frenzy continued.

Early railways were all sort of different guages run but investors who maliciously tore up houses and land to build railways.
An act for the regulation of gauge of railways in 1846 put paid to the broad gauge of the Great Western Railway invented by Brunel.
Early railway timetables were written in local times, this made them confusing. In 1847 the Railway Clearing House adopted GMT (now UTC) to standardise timetables. This has been the de-facto time standard ever since.

Railway Nationalisation

As the first world war broke out, the railways were bought under government control.



On 1 January 1923, almost all the railway companies were grouped into the Big Four: the Great Western Railway, the London and North Eastern Railway, the London, Midland and Scottish Railway and the Southern Railway companies.

The "Big Four" were joint-stock public companies and they continued to run the railway system until 31 December 1947.

Road transport

War accelerates innovation and cars were a big part of that.

Initial road transport enabled people and goods to travel short distances to trains.

British rail created

Following the 2nd world war the railways were nationalised completely under the banner of British Rail (BR).

Following the 2nd world war the railways were nationalised completely under the banner of British Rail (BR). There continued to be regional operation, but costs soon started soaring as cars became more popular and railways lost their ‘cool’.

1959 - First motorway built

The M1 opened in 1959, providing an express north to south highway. Political support in the UK shifted towards building roads, they were cheaper and didn't involve making or buying the cars.

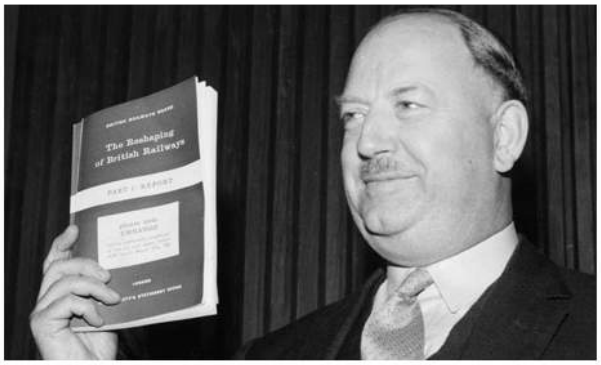
The M1 opened in 1959, providing an express north to south highway. Political support in the UK shifted towards building roads, they were cheaper and didn't involve making or buying the cars.

Early roads, like railways, had no speed limits or regulation. Again leading to deaths and the introduction or white lines and cats eyes by the highways authority.

Roads have always been regulated and managed by the public sector.

1950's Railway 'modernisation plan'

With spiraling costs and the railways in huge financial deficit the network was falling apart. The railways need to react to the market and the 'modernisation plan' and 'Beeching reports' were the solution. This led to the closure of over half of the stations in the UK and removal of 1/3 of the passenger services.



Busses

Busses have always been an alternative to trains.

However they either get caught in congestion in urban areas or are slow in rural areas.

Like trains, busses were originally run by private companies. But after the war these were also nationalised with the Transport Act of 1947.

With the re-shaping of the railways, busses were, and still are today proposed as a solution where a railway doesn't exist. However low demand and low speeds compared to trains mean that rural use of busses ramains low and schedules widely spaced.

Modernisation

1976

2017

Under BR, the network was modernised, with diesel and electric trains being introduced. But BR was never profitable until an ”intercity” brand was introduced with new stock that most people today still prefer to the modern stock.

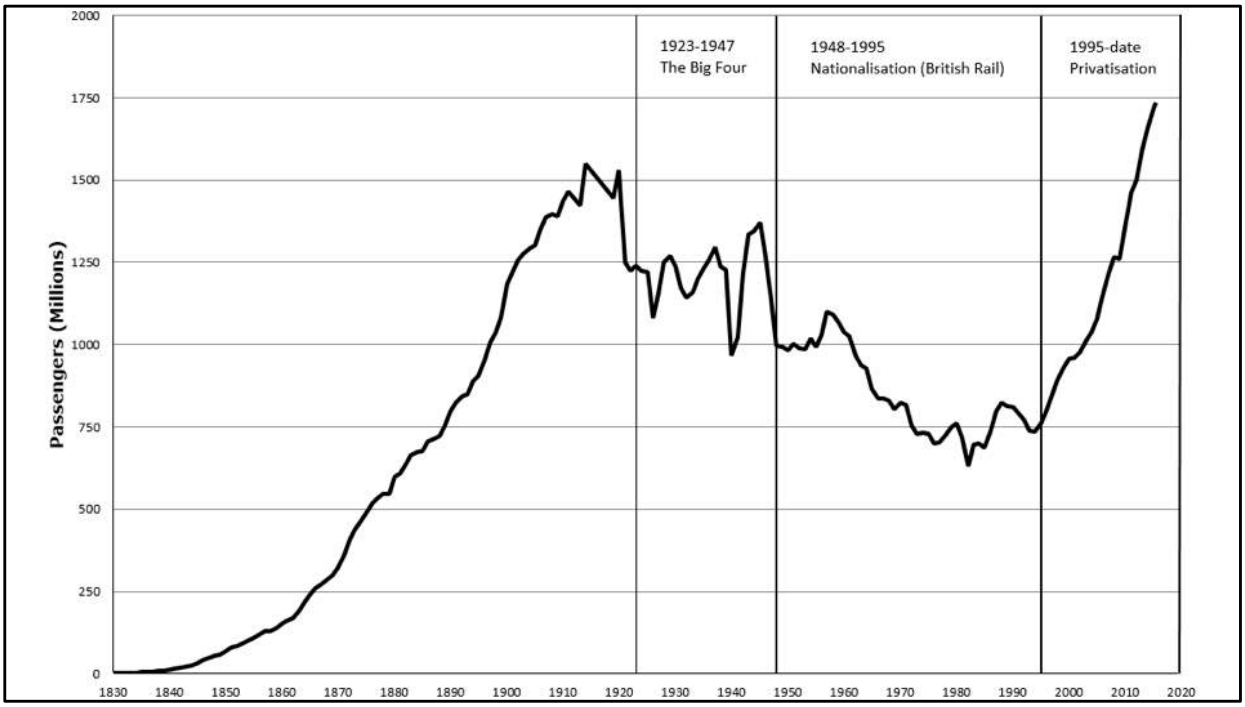
Here is that stock in 1976, and here is it, still running today at 125mph through Newcastle.

Privatisation (Railways)

RAILTRACK

1994 - 2002

In 1994 BR was privatised. Control of the infrastructure was handed to Railtrack and the company floated on the stock market. It was to sell track use to operating companies and invest the money in the network.
In reality it had the monopoly and following two major incidents and lack of investment, control was handed back to the state in 2002.



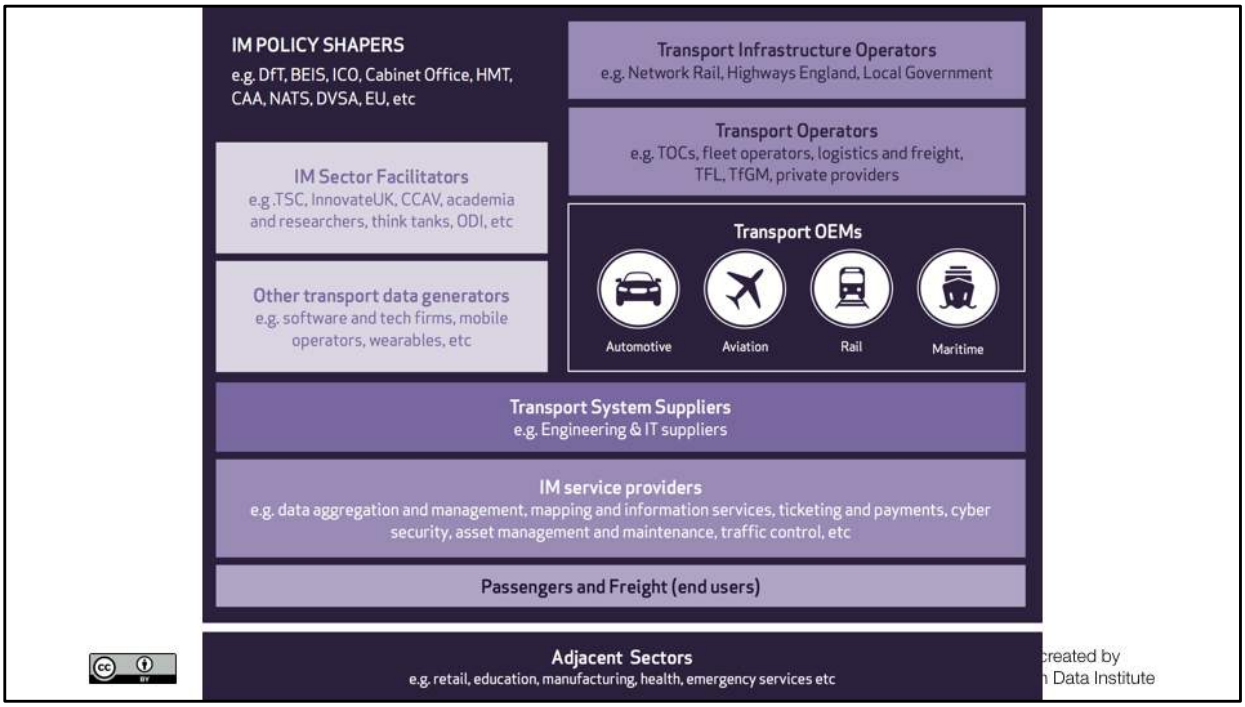
Since the privatisation of the Train Operating Companies (TOCs) passenger numbers have doubled. Investment in trains is at its highest ever levels and the network is the safest in Europe. Ticket prices have risen. Level of service and reliability has improved and regulated fares keep season ticket prices comparable to walk up prices in Europe. However un-regulated fares continue to rise above inflation.

Privatisation (busses)

The Transport Act of 1985 also deregulated and privatised the operation of busses.

This is still the case today.

In 1985, all bus services apart from those in London and Northern Ireland were de-regulated.
This was meant to increase services through competition. Regulations prevented neighbouring state owned companies being sold to the same concern, to create a 'patch-work' distribution of the operating areas. Lack of regulation led to lack of control and confusion for users with bus companies competing in the same area with different routes and fares and even same numbered busses.



This is the situation today.

At the top right we have the infrastructure operators including network rail, highways England and local government, all public.
However there are significant differences between them. While network rail run manage the track and timetabling of trains to ensure fair access, highways england cannot do the same for roads, so timetables for busses and locations of busses are set and controlled by the operators at the next level down. These are where all the train, bus freight and other transport companies sit. Also in this category is Transport for London, who as we already know are different. I'll come to them in a second and tell their history.

Under this are the manufacturers of the transport themselves. However this also gets complicated as very few transport operators actually own their planes, trains and in some cases busses. These are leased from lenders (often the banks) who purchase them, often with the government underwriting the risk of the operator loosing franchise. So that is complex.

Around these main three are the other businesses who provide services and

infrastructure at all levels, such as companies that actually lay the track for Network Rail and the companies that supply station signage etc. There are also many companies who manage and provide data. In all cases careful contracts have to be formed to ensure that the regulators and operators requirements are all met.

Regulation in action on the railwavs

Fares

~~Ticketing~~

~~Timetables~~

Routes

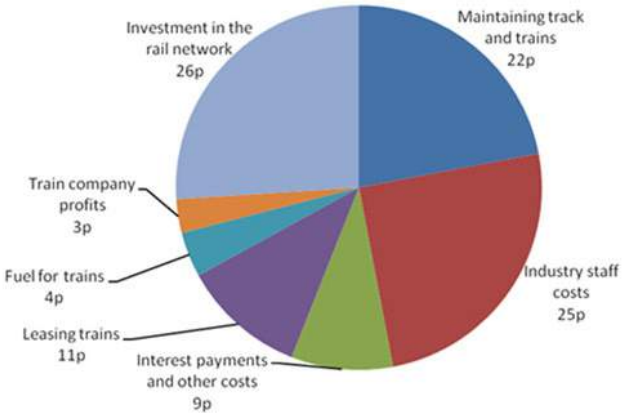
Disabled access

Fair access to network

Heath and safety

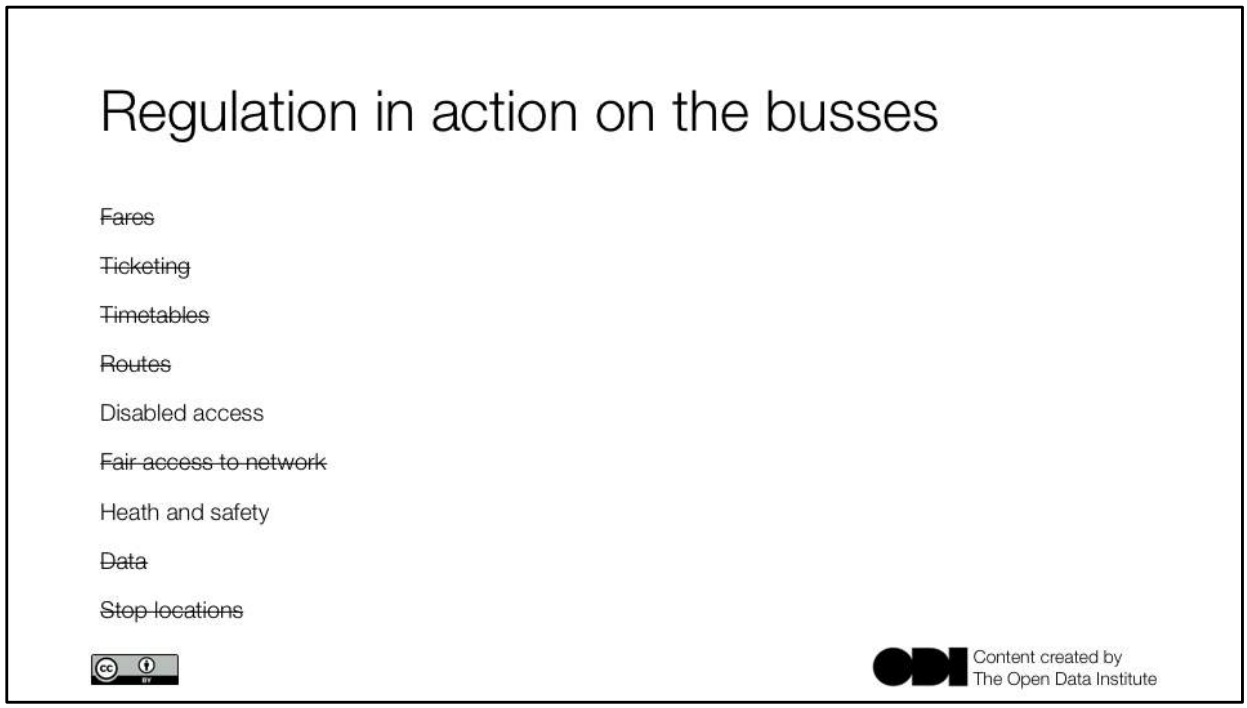
Data

Station locations



http://www.stagecoach.com/media/insight-features/the-facts-about-rail-fares.aspx

When it comes to regulation on the national railways, here is the list of things that are regulated and on the right the distribution of money earnt from fares. Most think that all the money from fares goes to the operating companies, however only 3p per £1 is profit.



On the busses, there is far less regulation which can lead to complexity and confusion even though there can also be greater competition. However this competition tends to be focused on popular inner city routes and not distributed fairly to rural areas.

As I mentioned, London is different.

At the point where railways and busses were privatised nationally, London was not.

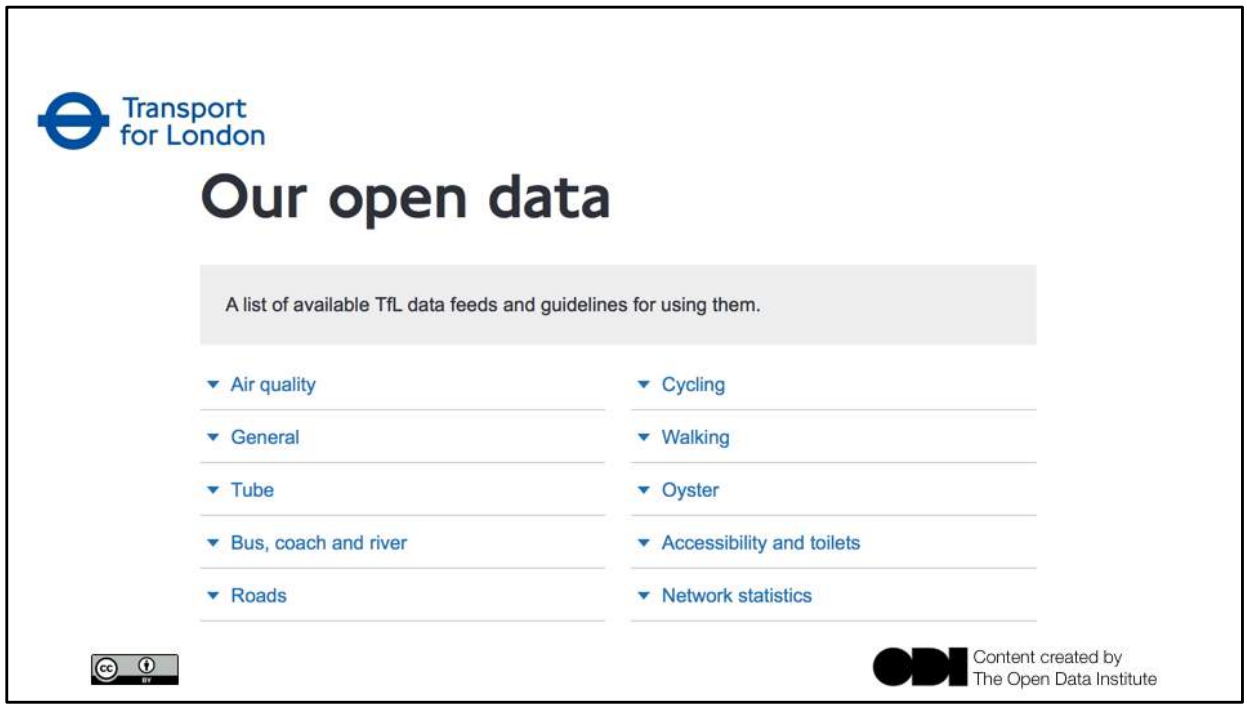
TFL has always been under state control currently run by the Greater London Authority by the Mayor.

TFL has complete control over nearly all network decisions and yet the busses are run by private companies who have to accept the regulation and requirements of TFL. This even includes the requirement about them all being red in colour.

As a result it is an integrated transport system with one payment mechanism which is now contactless bank cards (not even their own).

Many other UK cities are trying to take back control of their transport infrastructure and regulation. Newcastle has just acquired back the metro, which if you have been on is looking rather old.
Manchester is on a mission to do the same. There is also some amendments to the busses bill being proposed to help streamline delivery of bus services with new

regulations, however bus companies are fighting back in court.



We have seen already that Transport for London has an open licence on all of its data. It really is a world leader in this area offering live data feeds on everything from tube to bus locations as well as fare and routing data. There is more to come from TfL but they have made a strong start.

So what is the future of transport in the UK.

Firstly there is movement on the busses bill to bring rail and bus regulation closer together.

There is also a large movement towards more city based control like TfL including here in newcastle.

There is also a big drive towards opening up the data infrastructure more and increase the efficiency of the UK economy.

The transport systems catapult paper also looked at what happens if this doesn't happen.

The Transport systems catapult briefing paper estimates that if something is not done to improve the transport data infrastructure in the UK then there will in fact be a loss of £15bn by 2025 in having to provide other solutions to benefits that can be delivered through a better data infrastructure.

The briefing paper puts this down to three factors:

1. External barriers due to organisations being fearful of breaches in privacy, security and safety
2. Internal barriers due to perceptions that the costs of sharing outweigh the benfits
3. Cultural barriers across the sector leading to siloed thinking and not sharing data beyond organisations' own mode of transport

I would add a forth.

4. Modernize the law to bring more of the data infrastructure under regulation by the state with the aim of creating a more integrated public transport system.

This might help unblock access to key datasets and give the public an easier way to

find and use the right transport at the cheapest fare.

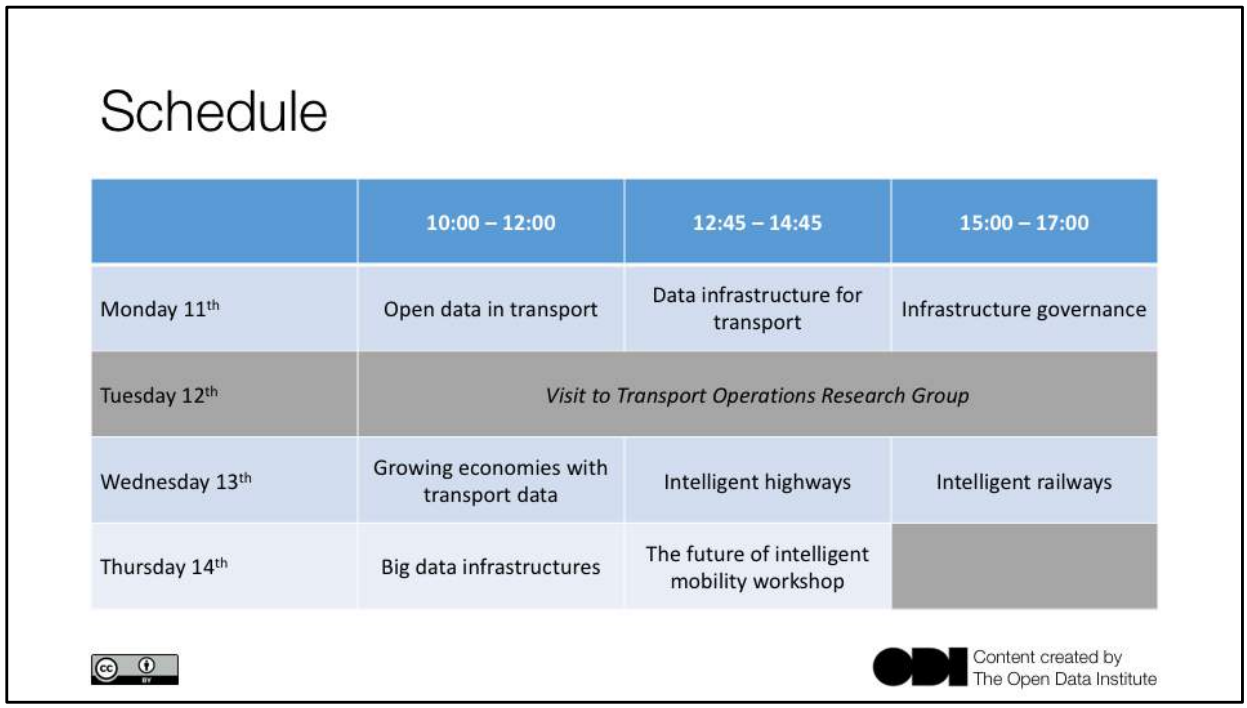
Exercise

What does china need to do to open up its transport data infrastructure?

1 quick action

1 long term action

Spend the next 5-10 minutes reflecting on everything we have looked at today and write down at least two key actions, 1 quick win and 1 longer term action that have come of out today.



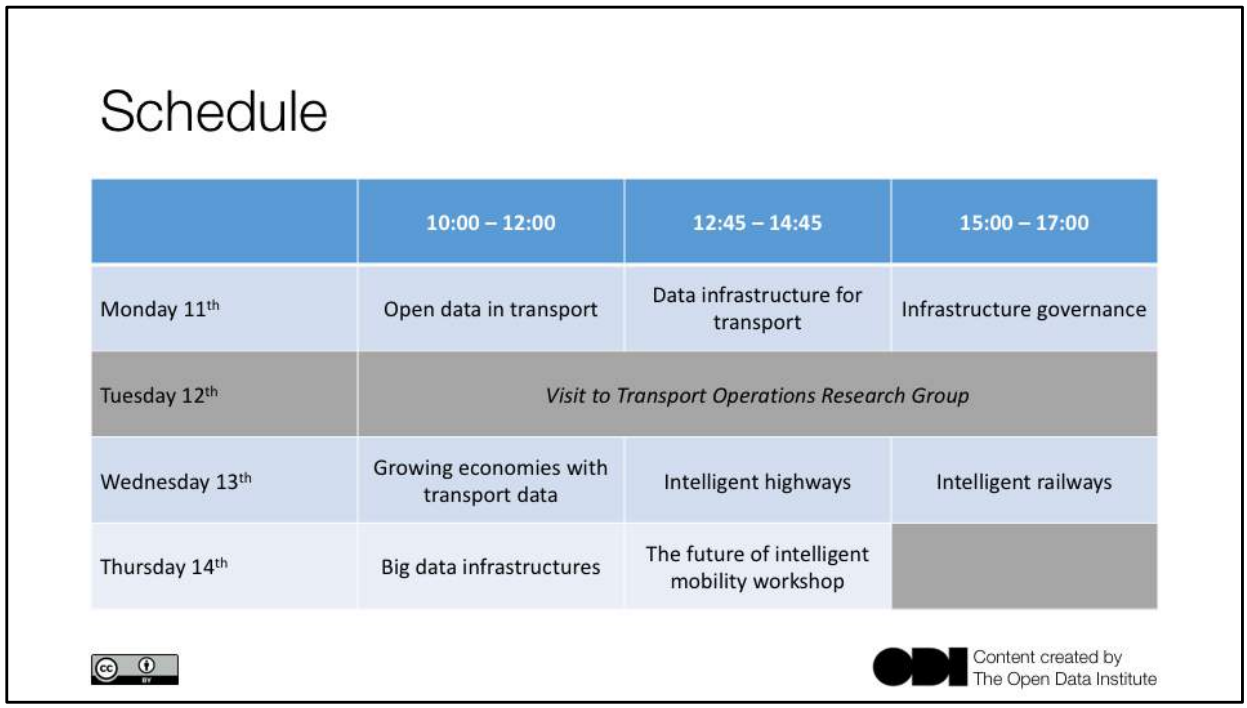
On Wednesday morning we are going to look at the key role of the community around your data and how new economic opportunities emerge around the transport data infrastructure.

Thank you and see you then!

Transport Data Infrastructure

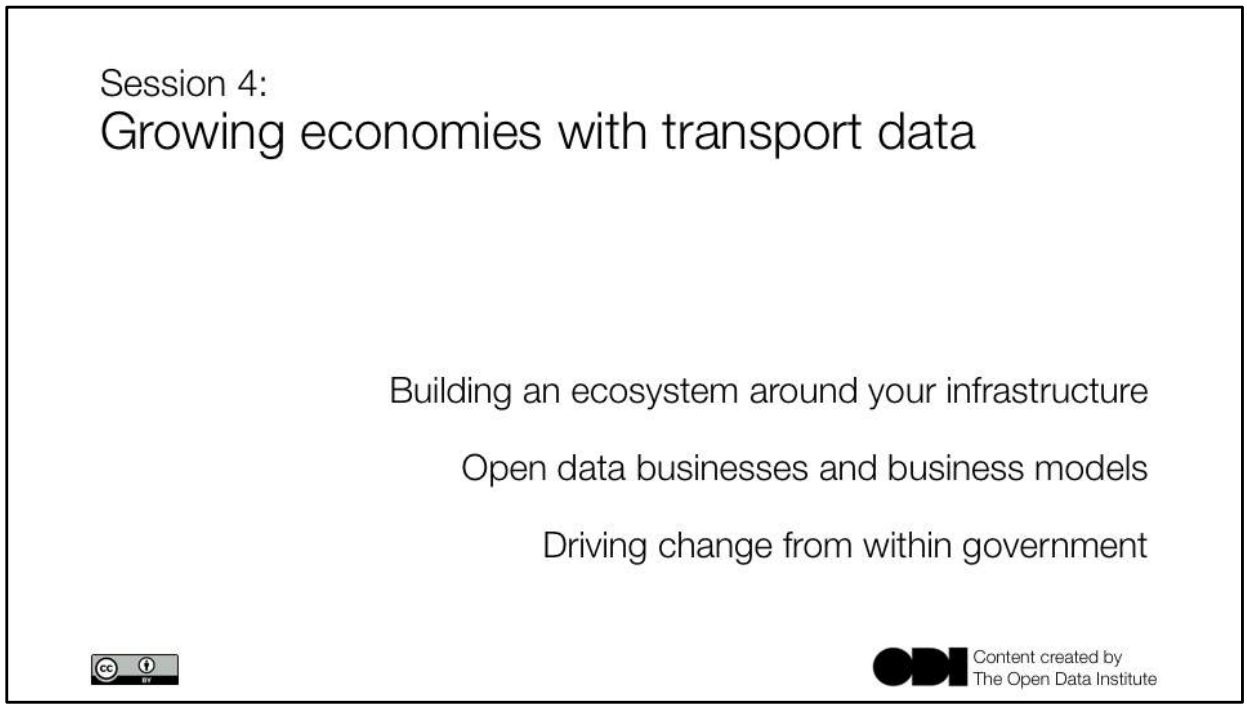
Dr David Tarrant | @davetaz
The Open Data Institute

Content created by
The Open Data Institute



Welcome back. I hope you had a informative time yesterday. Today we are going to dig deeper into how transport data is able to grow economies before looking at some of the innovations an opportunities in the highways and railways sectors. For the last session today we are going to take a short walk to the main station to see the how big data on the railways is delivered to passengers through it's many channels.

We will start our day looking at how open data can build whole economies. The Transport System Catapult estimated that there is a $1.2tn value to be had from intelligent mobility data. This session looks at how this value can be delivered.



The session starts but asking what the key ingredients are for building an ecosystem around your infrastructure, specifically in this case the data infrastructure. We then look at how to deliver a thriving ecosystem around data before looking at a number of businesses and business models that have emerged around the use of transport data.

You have been tasked with building an economy around your transport data. Partly to drive innovation, partly to drive efficiency, partly to prove the worth of the infrastructure. Much like building new roads, spending lots of money on a new road that no-one then uses is a disaster for government.

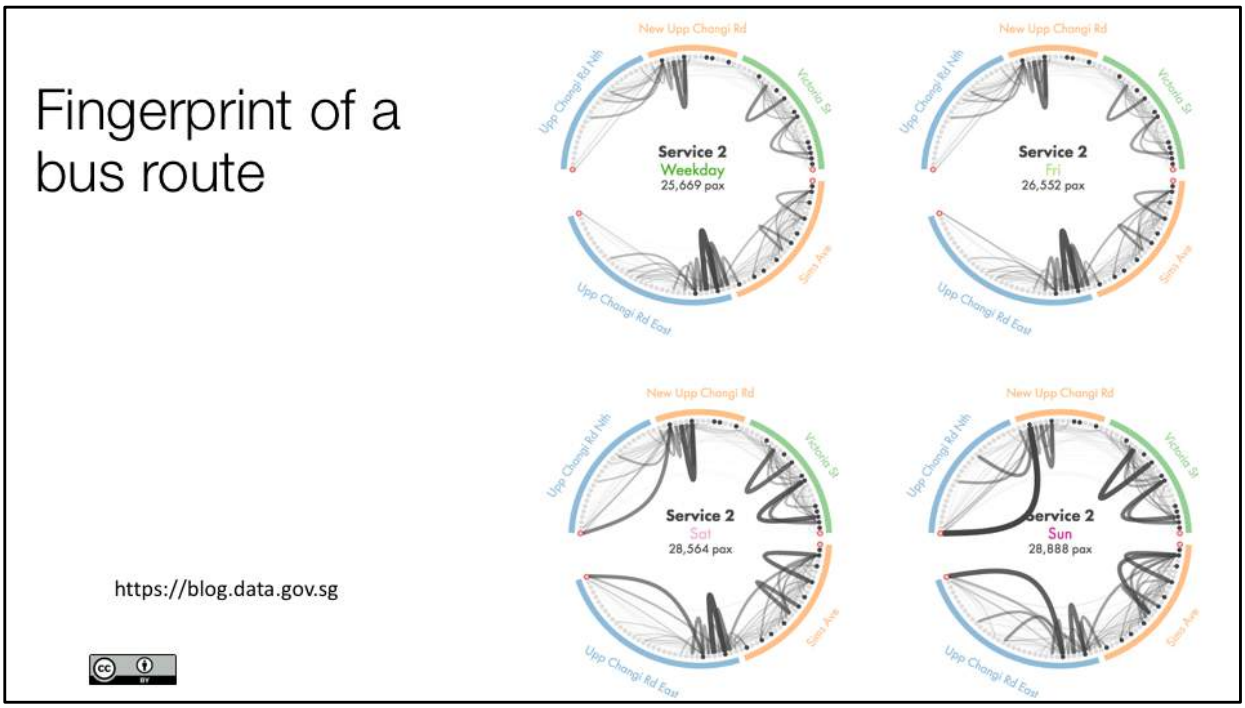
What things might you do to encourage an economy to grow around transport data?

Hackathon

Content created by
The Open Data Institute

A **hackathon** is a design sprint-like event in which computer programmers and others involved in software development, including graphic designers, interface designers, project managers, and others, often including subject-matter-experts, collaborate intensively on software projects.

Although there used to be a healthy ecosystem of hackathons in the US and UK. High sponsorship from companies and 'prizes' resulted in many being used in order to companies to purchase (or steal) Intellectual property. A number of hackathons also saw the release of open data that was later closed with the company internalising the ideas. Today, hackathons are still successful if the data is coming from the public sector, has already been open for some time or is regulated to be open.



This is an example of a piece of analysis done with bus journey data in Singapore during a hack event run within the government.

There are over 280 public bus routes in Singapore with more than 3 million trips made each day on average.

Route lengths and travel times vary widely. Buses on some trunk routes take more than two hours to get from end to end, while most feeder buses make short loops of less than an hour.

How is each bus route here utilised? Using the tap in and tap out data from travel cards, it is possible to analyse this.

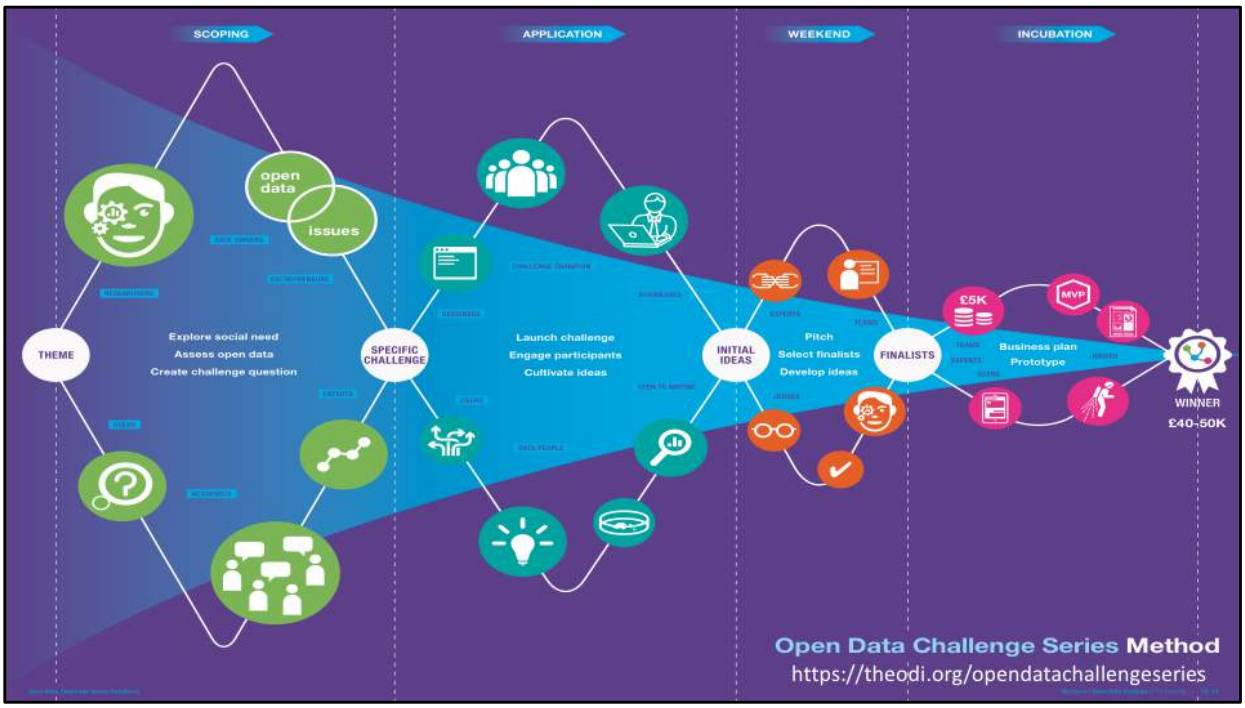
Taking the example of bus service 2, we plot out all the stops of direction 1 as dots, and show the main roads where the stops are located.
We then reposition the dots and lines along a semi-circle arc.

We then add lines to the diagram to show the total number of journeys between every pair of bus stops. The thickness and height of each line represents the number

of trips between these two stops, compared to the rest of the journeys on this bus service.

The fingerprints here show how different the use of busses is at the weekend compared to weekdays. Perhaps a different service pattern is needed.

This particular example was done internally by the ministry of transport data team. They were attempting to demonstrate the value of data science and open data to singapore.



Extending beyond a hackathon, the ODI and NESTA set up the Open Data Challenge series. This longer term challenge was open to startups only and involved, providers and consumers of data from the start.

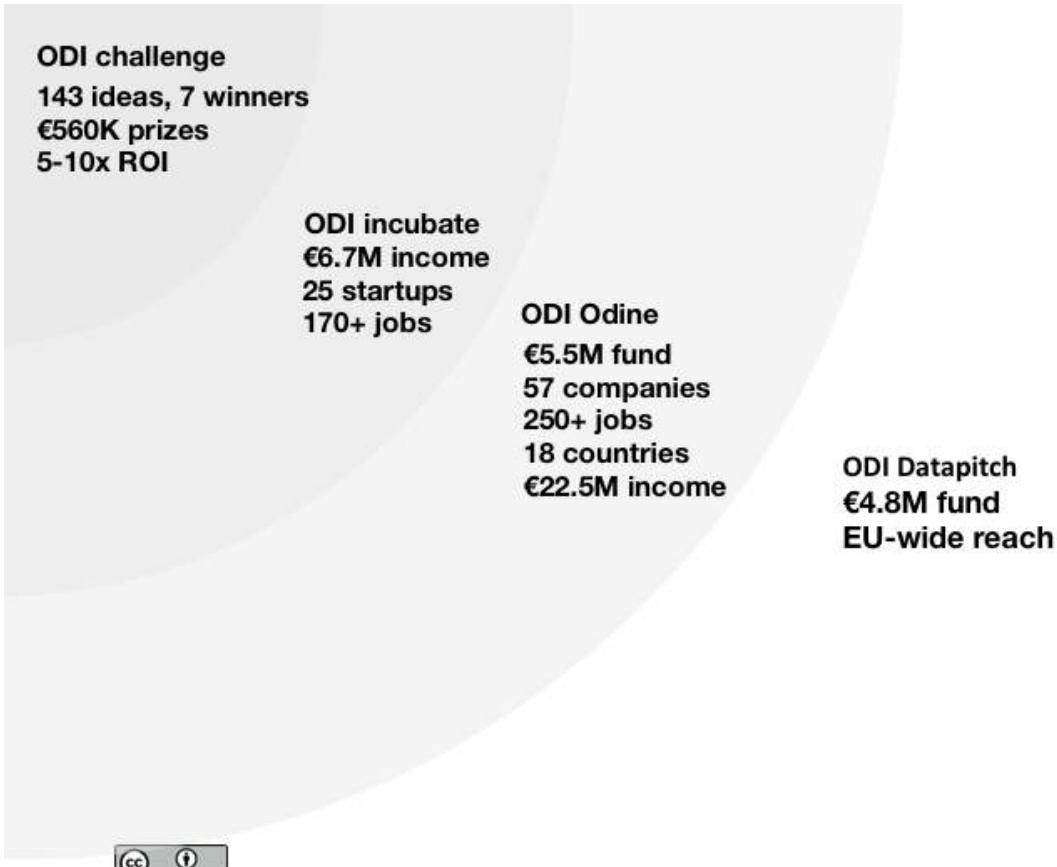
Agreements and contracts would help ensure availability of data and seed funding was awarded as the prize after the nine month process, with finalist each receiving £5000 to complete their entry for the gudges.

There were 7 areas of focus to the challenge series.
Crime and justice,
Food
Housing
Heritage and culture
Education
Energy and environment
Jobs

You will notice there is no transport focus, I'll come to why this is later.

However, the winner of the crime and justice challenge was Check That Bike!. This site allows people wanting to buy a second-hand bike to check whether it has been previously stolen.

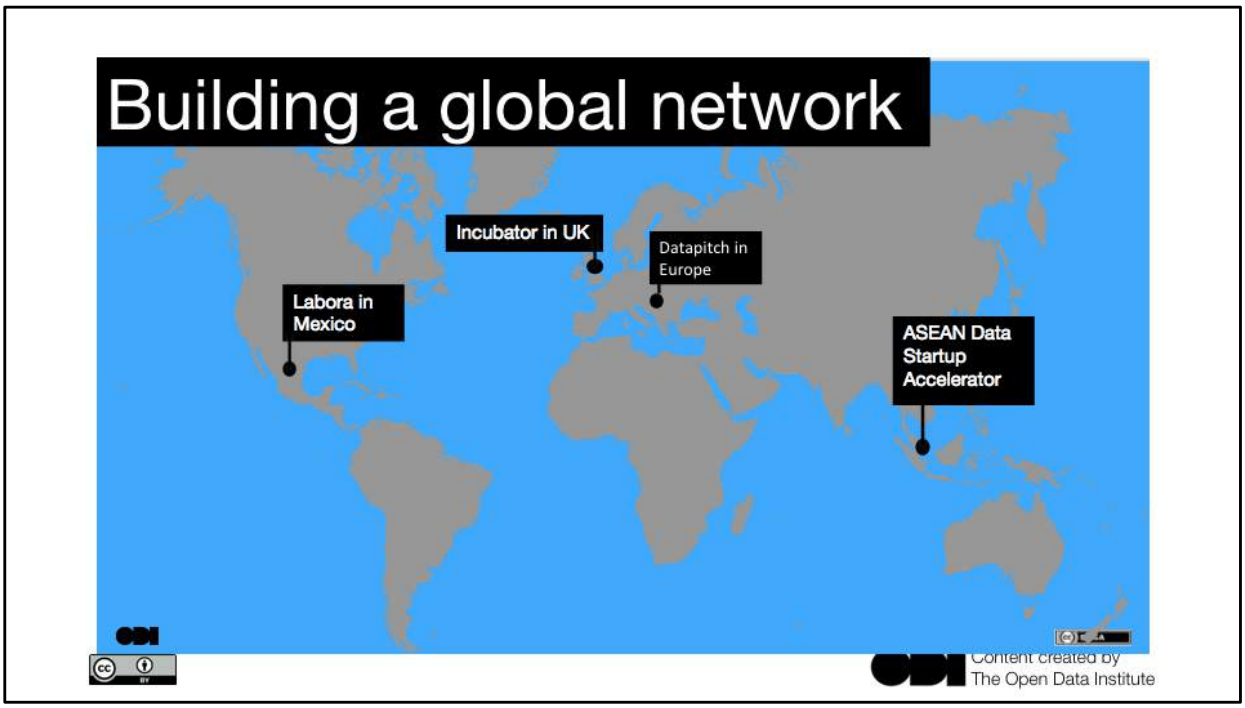
By cross- checking unique frame numbers against Check That Bike!, cyclists can make better decisions about buying second-hand bikes at the time they're actually buying them. The service will also help police forces tackle bicycle theft by disrupting the market for stolen bikes.



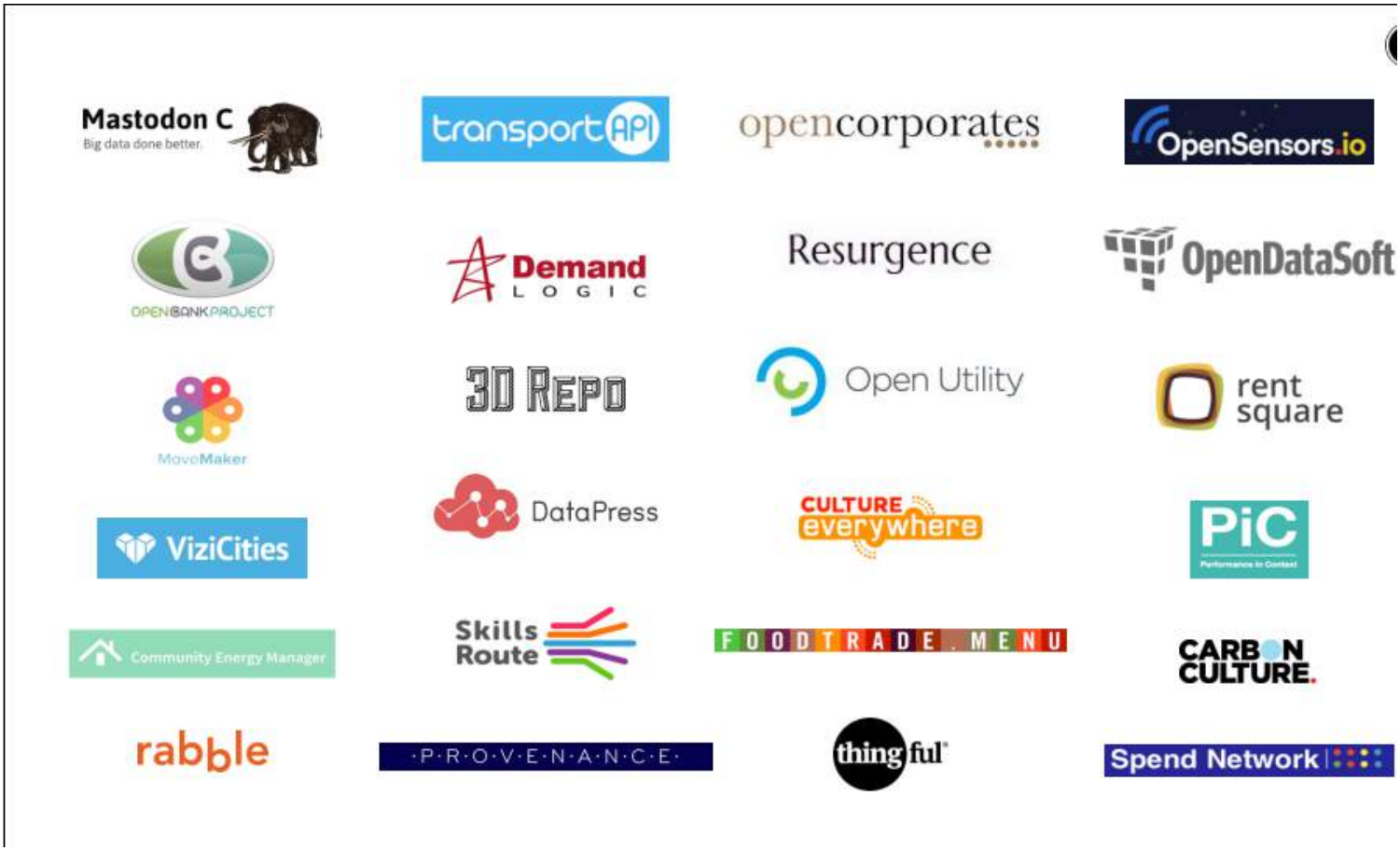
You might go for something bigger again, like a full incubation programme. Alongside the challenge series, the ODI started out with its own incubation programme for London start-ups. On a shoestring budget but an office full of experts we helped unlock **€6.7** million in income for these start-ups which was then followed with one of them securing $13m of seed funding in the US.

We followed this with ODI accelerate, a EU funded programme which has just completed. This has helped get 57 companies developed in 18 countries and unlocked over **€**22.5m in income so far.

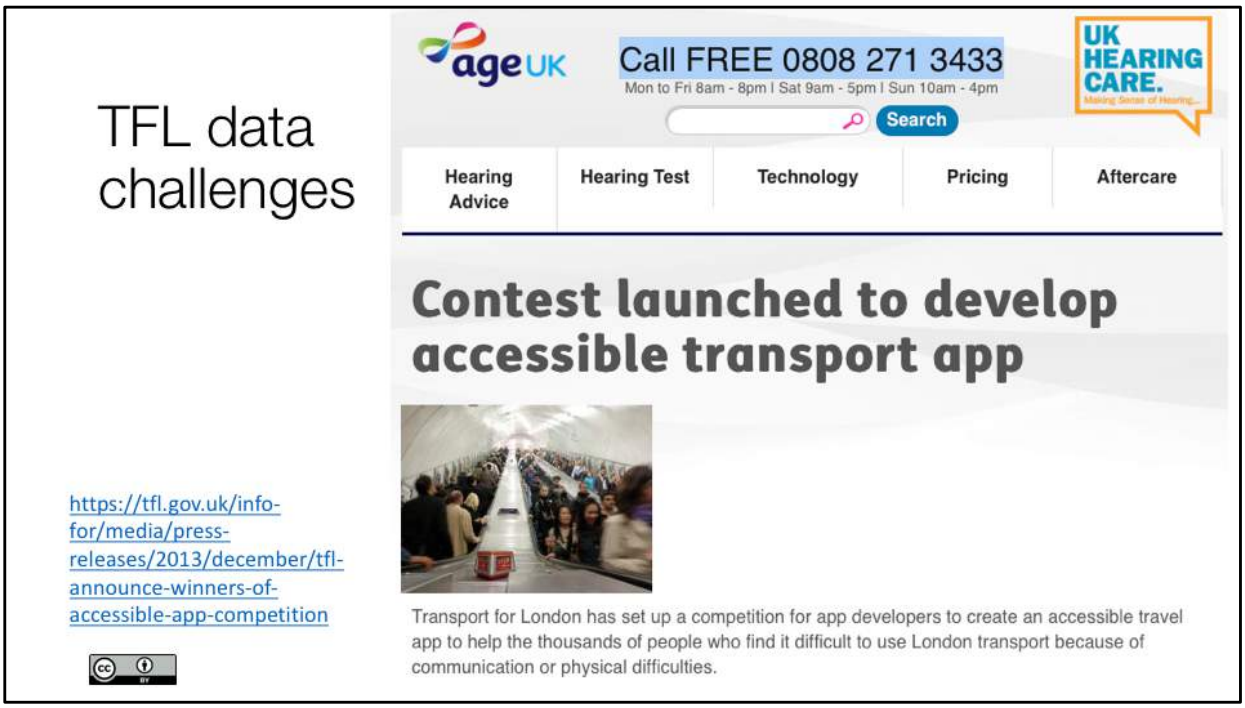
Our latest programme, again EU funded has just started with a similar aim, accept this time the project has a stake in the companies and any money gained back will go into further programms.



We are going to be re-launching a new startup programme in the UK in 2018. As well as the European programme we also have programmes running in Mexico and south east Asia.



Here are just some of the companies that we helped incubate. You will see transportAPI at the top there and I will come to them later.



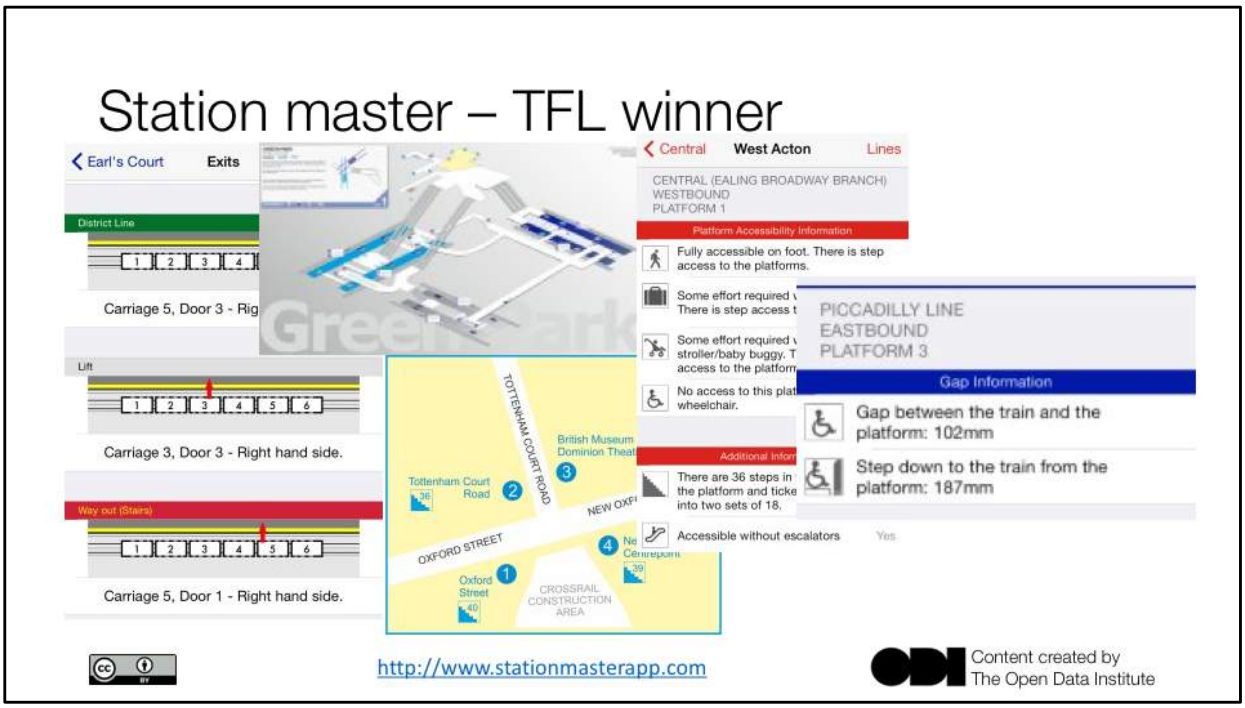
I mentioned earlier that there was no transport sector in the challenge series. This is because TfL were running their own challenge.

TFL ran a competition to develop accessible apps. Those which focus on providing better services for the disabled or those who need assistance.

They had 194 apps put forward with 41 being carried to the competition stage, all using the TfL open data for free.

All of the apps cost less than £3 in the app store with the winners also receiving £5000 to assist in the development of their apps. More importantly for TfL were the connections to the developers who have helped shape data policy and access to data ever since. I can tell you there is a new report coming in the next few weeks from TfL on the impact of their work which updates the 15-58m figure I presented yesterday. I will be very excited to see the new figure.

So who won their competition…



Station Master is the definitive travel reference App for London, with facts and figures collected from every station on the Underground, Overground and Docklands Light Railway networks.

Station master tells you which carriage and door to be at for your exit at the station on every tube and DLR train, at every station.

It gives you 3D maps of every station. Exit maps of every station.

But the best bits are the options for accessibility.

TfL do provide two guides - one is the Step Free Tube guide, and another is the Avoiding Stairs Tube guide - which are great - except that their definition of 'Step Free' refers to wheelchairs users only - which is flawed, because if you have luggage or push chair, a station that has escalator access to platform level (e.g. Walthamstow) is easy for you to use.

Station master breaks assessibility into four categories:
People able to walk on foot

Travellers with luggage
Parents with children in buggies
Wheelchair users

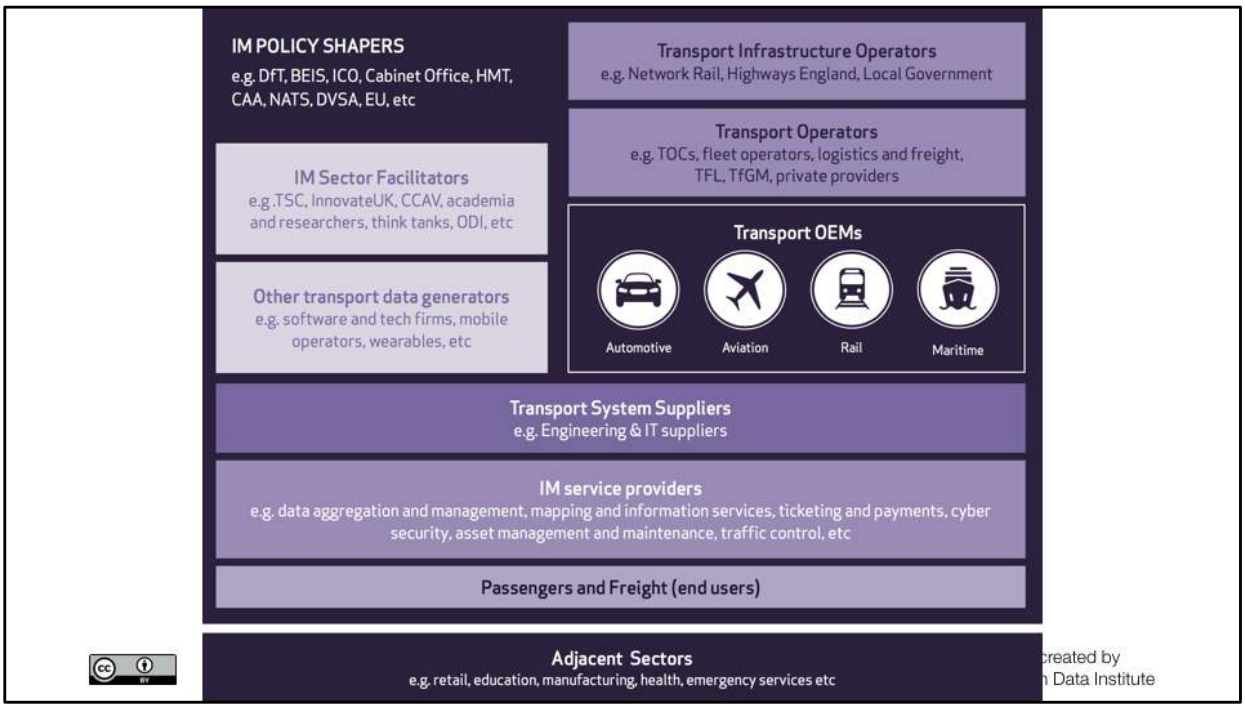
Accessibility means different things to different people
Some people can manage a small number of steps (up to 10), but no more. There are those who can walk down stairs, but not up them. Some people can only manage steps if there is a rail to hold on to - i.e. Accessibility can vary upon your own personal needs, so Station Master presents you with the facts at that station and let's you make your own decisions.

The team behind station master went to every station and counted all of the steps, everywhere.

The last feature is my favorite and shows how obsessed with data the team are.

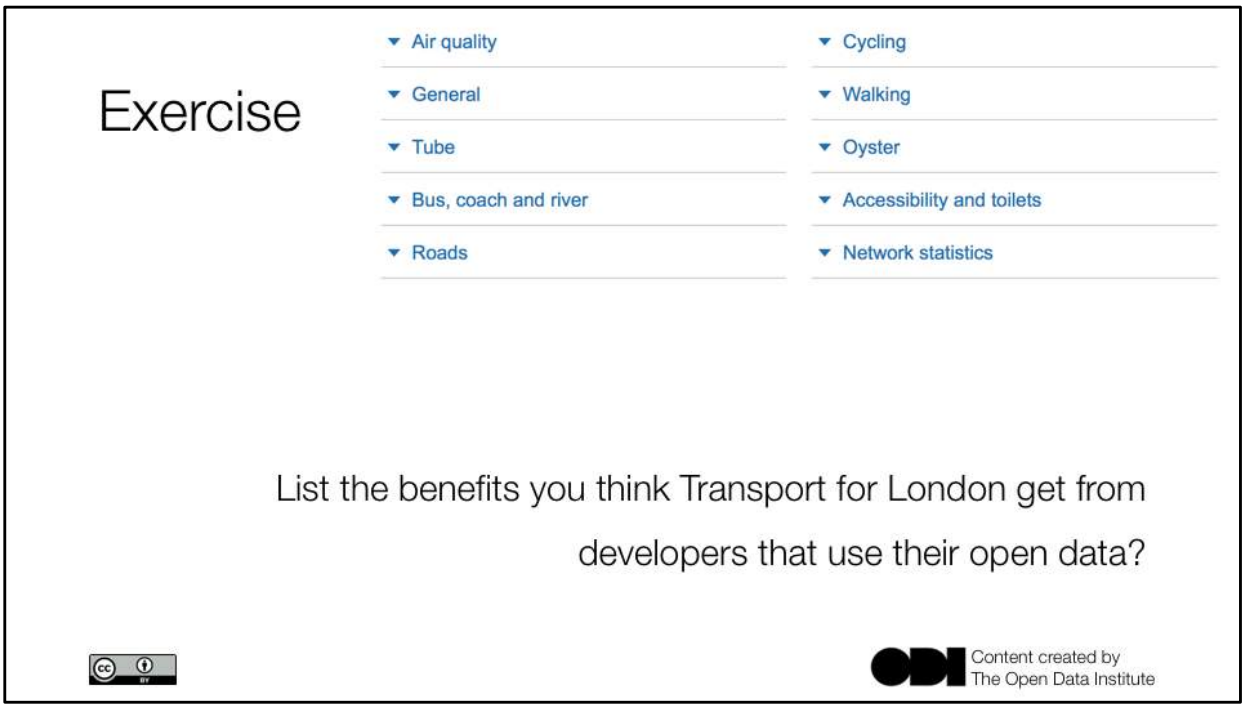
TfL provide gap/step information at station where there is step-free access to the station, but there isn't always level access from the platform to the train. Where they provide these measurements though, they don't say if it's a step-up or step-down from the platform to the train, so they went out and measured them - wheelchair users find this really useful as they can manage a step *down* but not a step *up* when boarding or exiting a train.

Why did they do all this. Well, the team behind station master hold the world record for visiting every London underground station by train in a single day. This is not an easy thing to do and requires 16 hours, 20 minutes and 27 seconds. I know some people who have tried it and failed to even visit them all in a day.



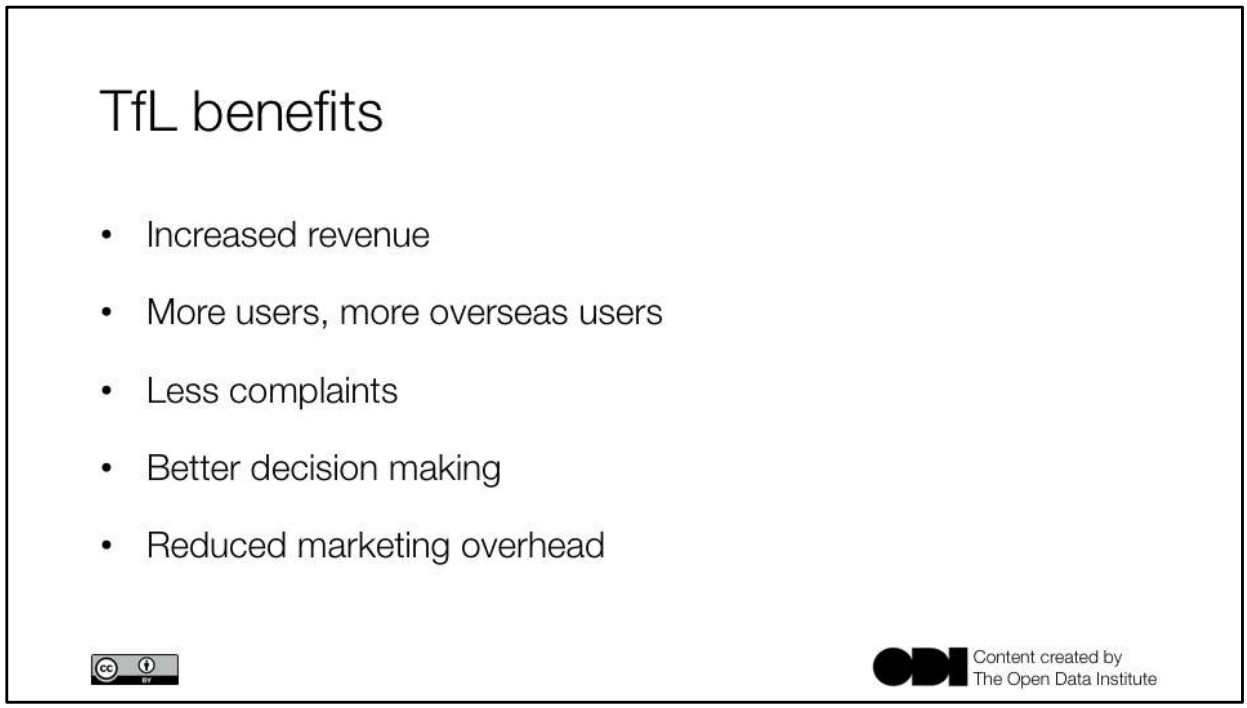
As we found out this morning. Transport systems in the UK are complex, as are the relationships between the companies involved.
TfL are able to move fast as they own their own trains, the regulate as a public body, they own their data and talk directly to end users and developers as a result. Yes other companies run the busses but have to abide by the TfL licensing rules. Some people don't like this monopoly, however this regulated monopoly is working. London is the busiest public transport city in the world.

So how did they establish such good relationships with developers and why do they need them.



I would like you to take a minute to think of all the benefits Transport for London get from developers using their data.

As a reminder I have included reference to what open data TfL provide on this slide however feel free to simply extrapolate which benefits you think they might gain from developers using any transport related data. Maybe you'll even identify a benefit TfL are not yet getting as that data is not available.



Here are a number of the benefits that TfL get from opening up their data.

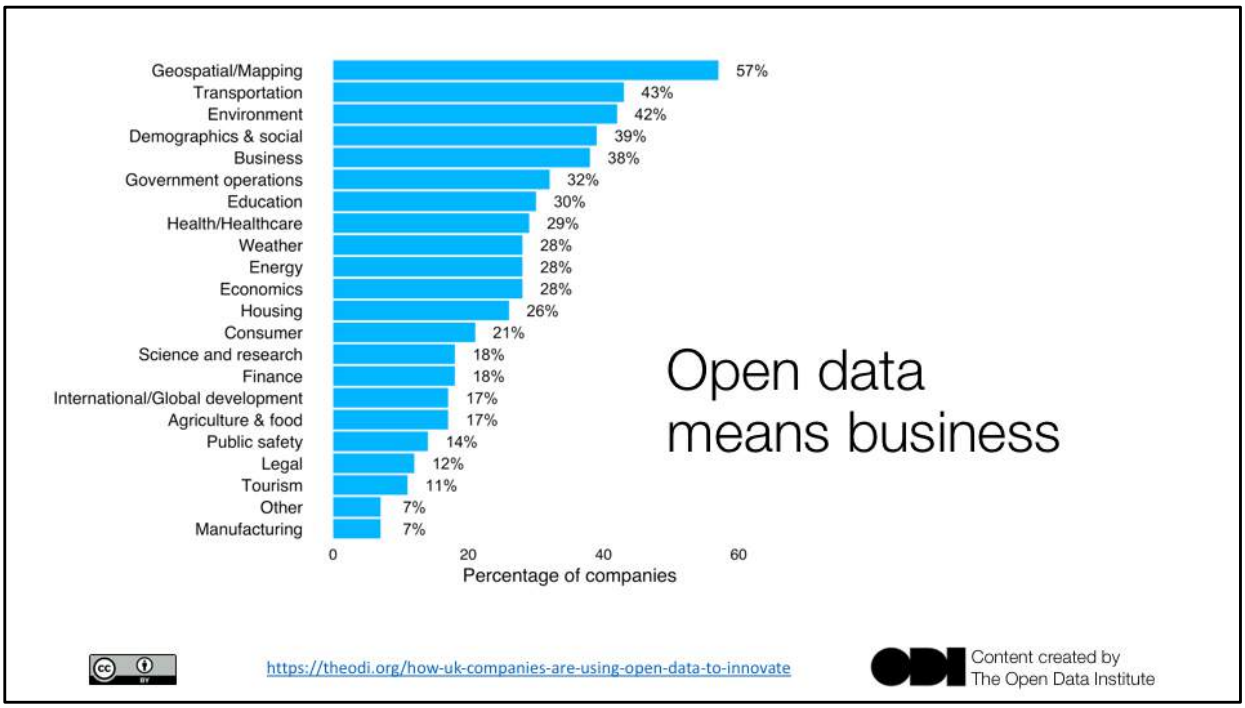
Firstly, increased revenue. The more people who get recommended to travel on TfL the more money they get.

This is especially the case when broadening the market to visitors from overseas who might not otherwise feel confident to use the network. As we have already seen this week, apple and google maps integrate well with transport data in the UK.

Perhaps less obvious is the benefit of less complaints. The more people are informed about the services and incidents, the less likely they are to complain when issues arise. I myself have calmed down angry passengers by explaining that the reason for the delay is due to a medical situation on the train in front of the one we were on. They seemed to accept this, they just needed to be informed.

Better decision making is another benefit. The more people who can become expert though data use means that everyone who works or becomes an employee of TfL makes a better and more efficient system.

As more apps can promote the use of TfL services, the less TfL need to market themselves. How many people have planned public transport by reading what people have said on trip advisor for instance?



So why does this work so well for TfL.

To answer this question we are going to look at the underlying business models and specifically how a three way partnership between the transport provider, private sector and user can be made to work for everyone.

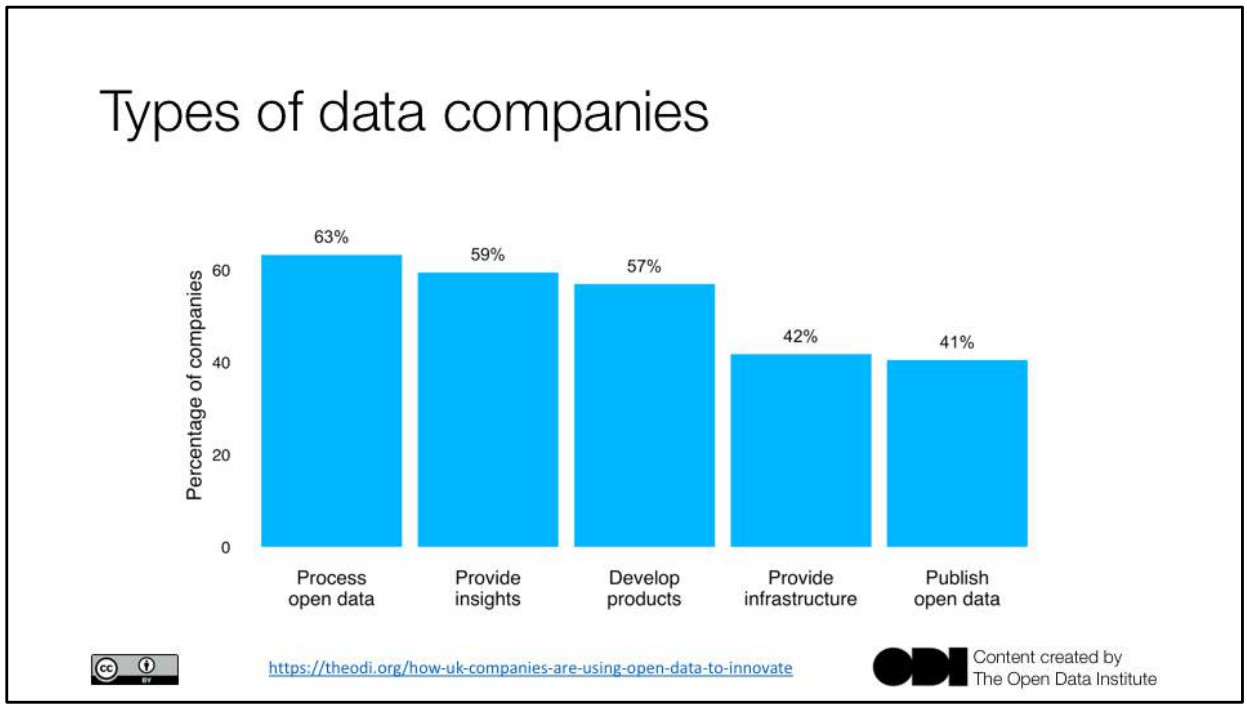
Significantly, transport is the second largest open data sector in the UK.

As part of a ODI study, we identified and analysed **270 companies** that use, produce or invest in open data as part of their business, through desk research, surveys and interviews about their experiences. The open data companies we studied have an **annual turnover of over £92bn**, and over **500k employees** between them. This alone shows the scale of open data's potential value in business.

Transport data was used by 43% of the companies we surveyed, second only to those using geospatial and mapping data.

This finding further demonstrates the importance and massive opportunity of notjust sharing transport data, but opening it up for others to use.

So who are these companies and how does the business model work?

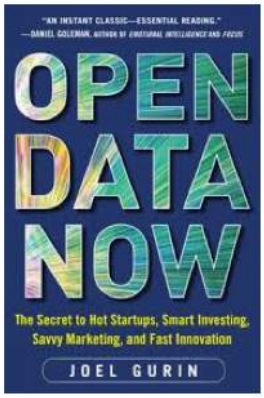


As part of the survey we asked each company what their primary role was in relation to the open data.

As can be seen here, the majority of companies are data processors (63%). 59% provide insights, 57% develop consumer applications, 42% provide infrastructure and 41% assist in data publishing.

Current companies

- Suppliers
- Aggregators
- Enrichers
- Enablers
- Developers



Another categorisation of companies is provided in the book Open Data Now. This book looks at the types of companies emerging around daa infrastructures and finds 5 different types.

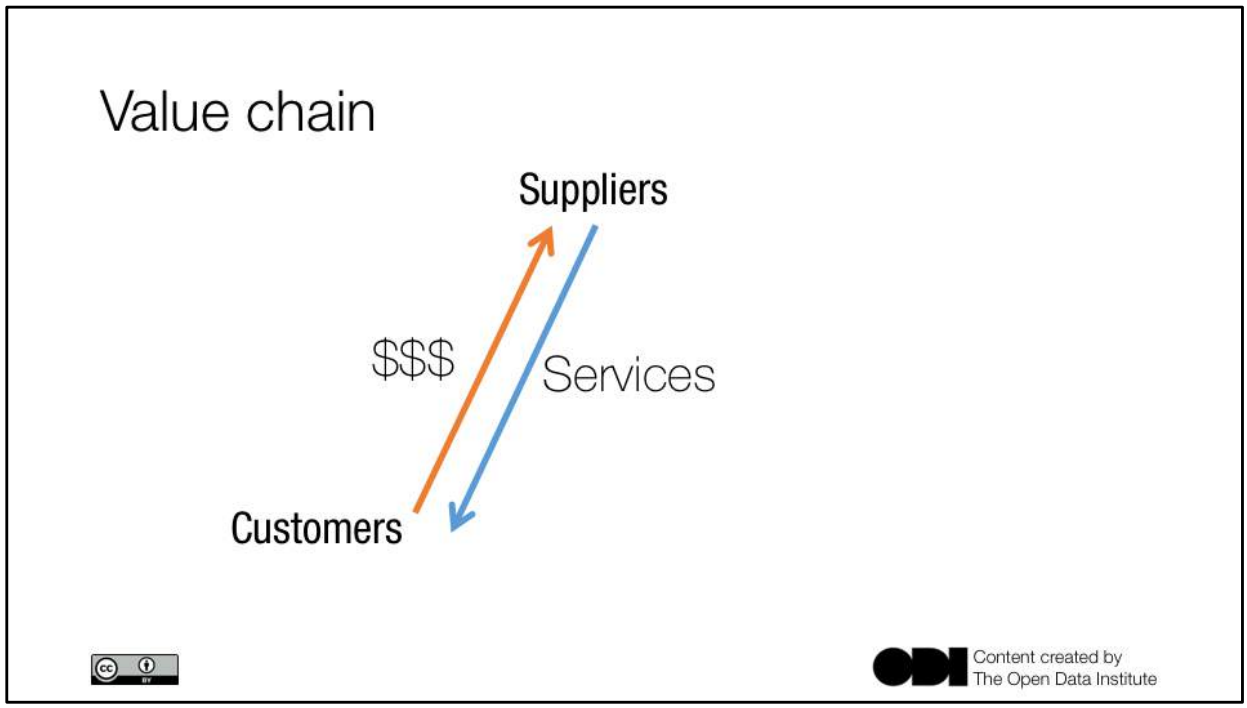
The first, suppliers, are those companies who either own and supply the data, or companies that make services and platforms that allow the supply of data. The government would be an example supplier in this case.

Aggregators take data from many suppliers and allow broader level access to it. Google is a good example on the web and we will look at another example later.

Enrichers take data from suppliers and add value for others. This may include performing analytics or some enrichment or combining with proprietary data or insight.

Enablers build business to business services in the market that speed up the flow of data. Enablers also build many of the platforms and tools used by both suppliers and consumers of data in order to operate more efficiently.

Finally, developers are those which use the data to build consumer services. The station master app that we looked at earlier is a good example of a developer application designed from consumers.



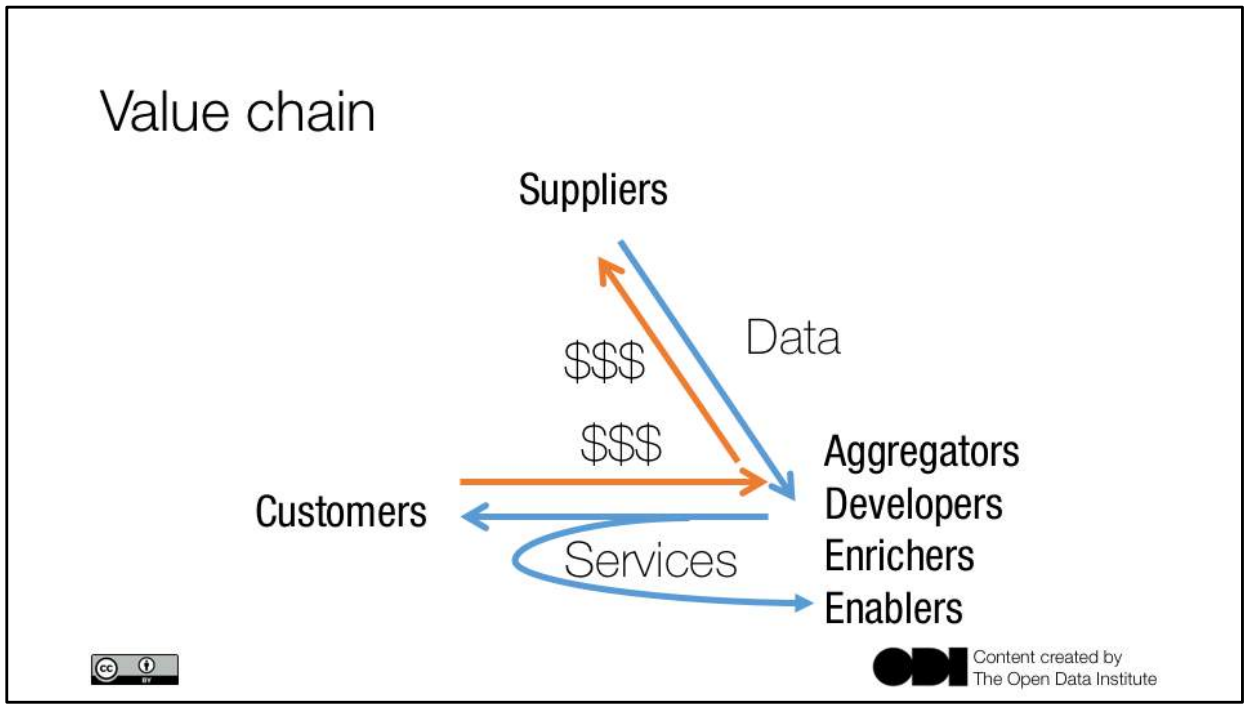
So lets look at how the value chain works in transport and how it breaks elsewhere.

First, lets look at a typical supplier, customer relationship. Here the supplier is directly interfacing with the customer and suppling all services, from the transport links and busses themselves to the route planning applications.

So the supplier interacts directly with the customer, providing services. The customer purchases these services, be it travel tickets or smart phone applications from the supplier.

In this model the supplier has complete control over all services offered using their data. What are the disadvantages of this model?

The main disadvantage of this model is that you are not exploiting services that others can provide. So lets look at a different model.



When data is introduced, suppliers are unlikely to offer this directly to their customers.

Data will be offered instead to the aggregators, enrichers, enablers and developers that build services for their customers.

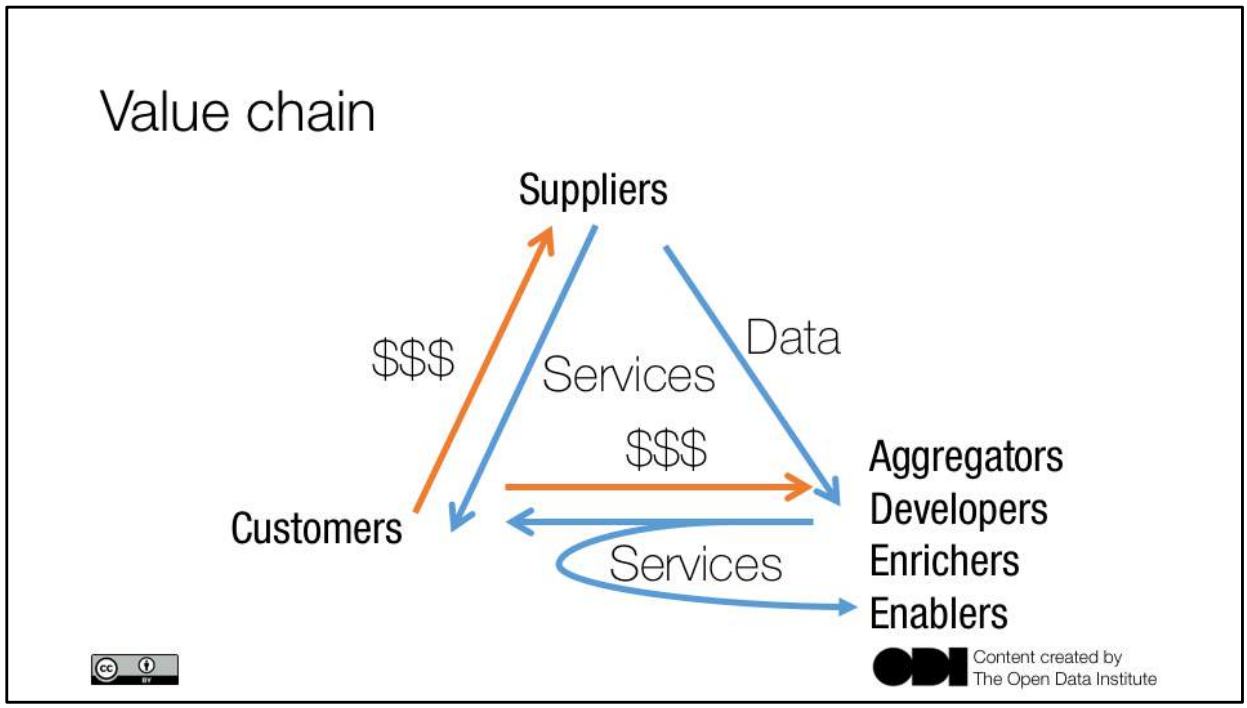
This means that these companies are now interactive with the customers, or each other. In turn the customers are likely to be buying services from these companies, rather than the supplier.

This is typically what is seen as the biggest threat of open data to businesses. Allowing others to supply services you currently offer.

Historically suppliers have controlled the market through the selling of expensive commercial agreements around the data, thus ensuring that the cycle is complete here.

What is the biggest problem with this business model for a government?

The biggest problem with this model is that it often cuts out now startup companies who either can’t afford the data, or can’t get access to it as existing contacts have exclusivity clauses in them. This stifles innovation badly and create monopolies.

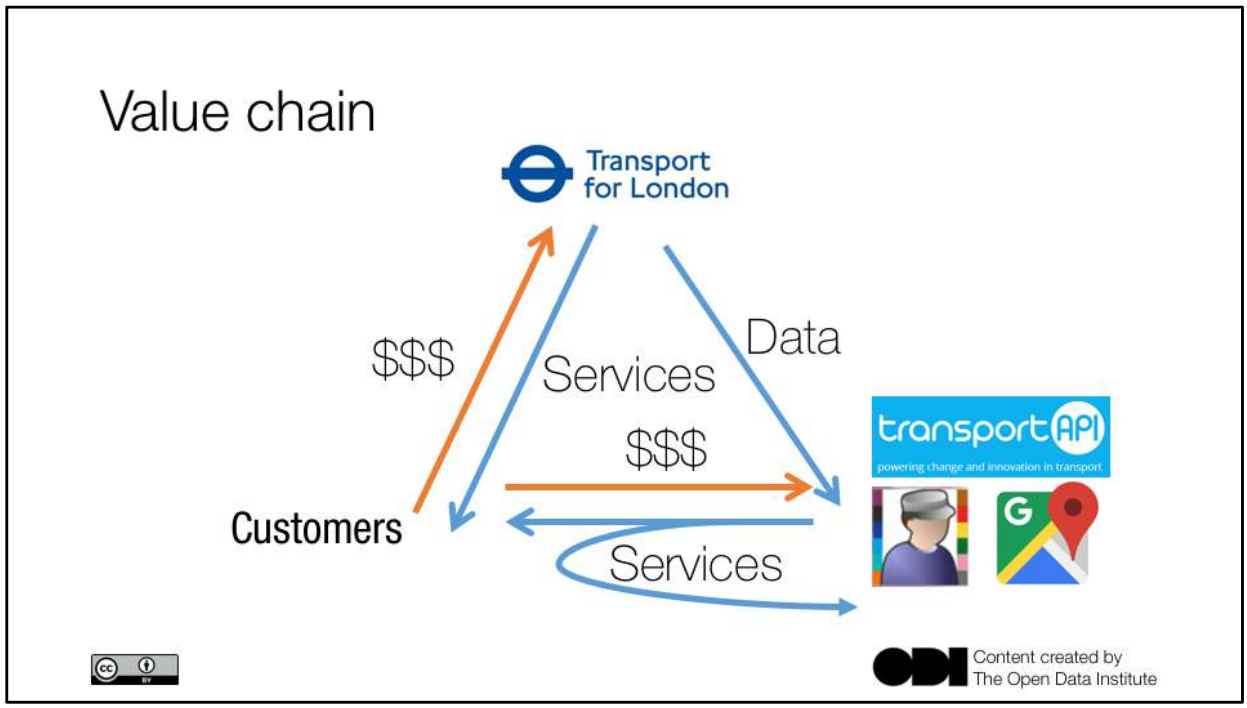


Going back to this model, it is possible to reveal that it is different for transport. Ultimately the goal of public transport is to sell tickets to consumers who use the mode of transport. The more consumers on your method of transport, the bigger your business.

So supplying data to a whole host of companies, regardless of size, will enable them to build services for consumers (or each other), which ultimately ends up with the benefit of more customers on your transport system. If the service they are building is valuable enough then they might even make money thus opening up an economy.

In some cases, what the service providers do might help fulfil the suppliers other goals, like providing better access for the disabled, or even help the supplier with data analysis and efficiency gains.

I still can't believe how some transport companies won't open up their timetable data as they think there is more revenue to be had from the adverting fees on their website than the increase in passengers that comes from everyone in the world being able to discover their services using google maps or citymapper.



To close this session, let's look at how this works for Transport for London.

Transport for London are the supplier, and their open data is made available for everyone to access, use and share.

The companies consuming this data include examples such as transport API, who are an aggregator. Google maps, also an aggregator but also provide consumer facing services, and station master. The winner of the accessibility hack competition.

These are three different companies with three different objectives and backgrounds.

Google, like Baidu are attempting to be a worldwide provider of information, including routing information on maps. They are basically in the satellite navigation business for consumers. B to C

Transport API are also in the routing business, but as a B to B organisation. Their main goal is to aggregate together transport timetable data for the whole of the UK under a single API for others to use. This saves other developers having to work with potentially hundreds of APIs to be able to plan travel across the whole of the UK. We

witnessed earlier just how challenging this can be.

Station master is a B to C organisation looking specifically at the customer niche of those who require better accessibility data. While this is a niche part of the population, it is a significant one and providing fair disability access is mandated in most countries. TfL is no exception and government mandates mean that not only should the physical infrastructure be suited for those with accessibility requirements, but the data and information infrastructures must also be. Opening up their data and working with the team behind station master was a critical step to fulfilling this mandate. TfL saved themselves a huge amount of money in supplying key services to this niche of the population.

The ODI is actively working in the area of service delivery models at the moment. We believe that opening up data opens new markets, allowing others to create services for the broad population freeing up providers time to focus on the gaps for those it must provide services to.

Exercise

What things might you do to encourage an economy to grow around transport data?

To close this session I would like to return to the question we started it with and get your views on what you might be now to encourage an economy to grow around transport data in China?

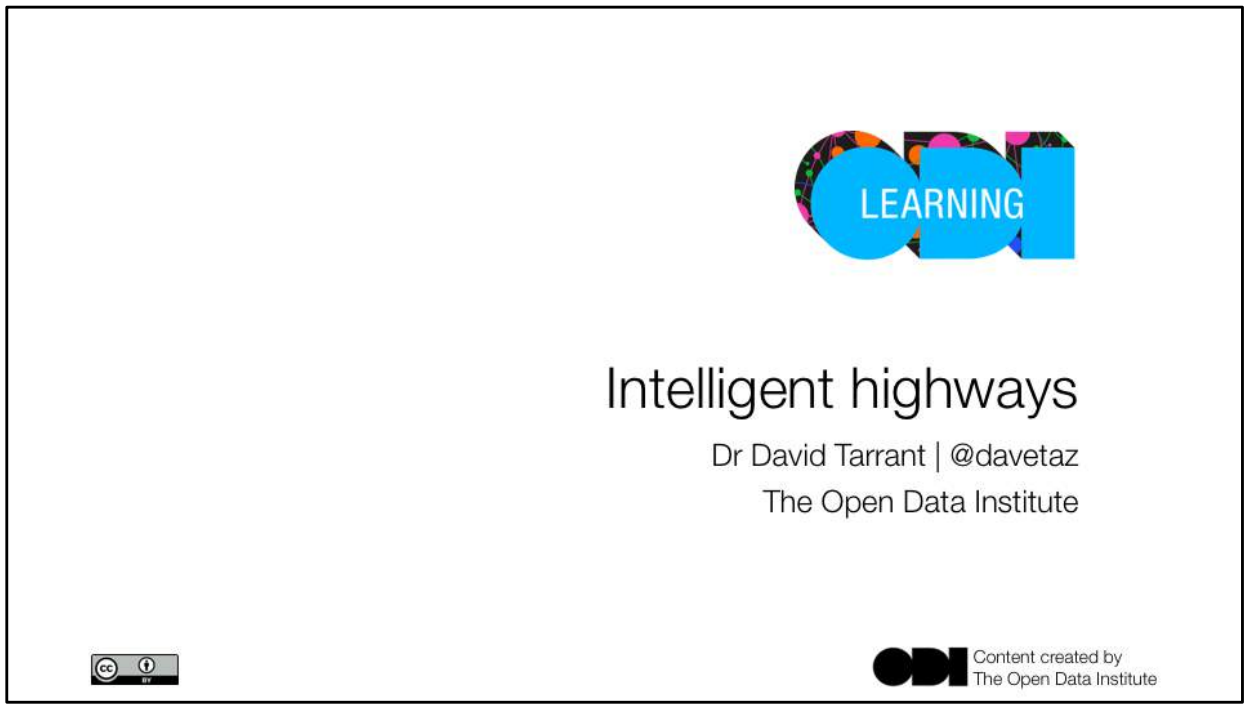
Transport Data Infrastructure

Dr David Tarrant | @davetaz
The Open Data Institute

Content created by
The Open Data Institute

Schedule

	10:00 – 12:00	12:45 – 14:45	15:00 – 17:00
Monday 11th	Open data in transport	Data infrastructure for transport	Infrastructure governance
Tuesday 12th	*Visit to Transport Operations Research Group*		
Wednesday 13th	Growing economies with transport data	Intelligent highways	Intelligent railways
Thursday 14th	Big data infrastructures	The future of intelligent mobility workshop	

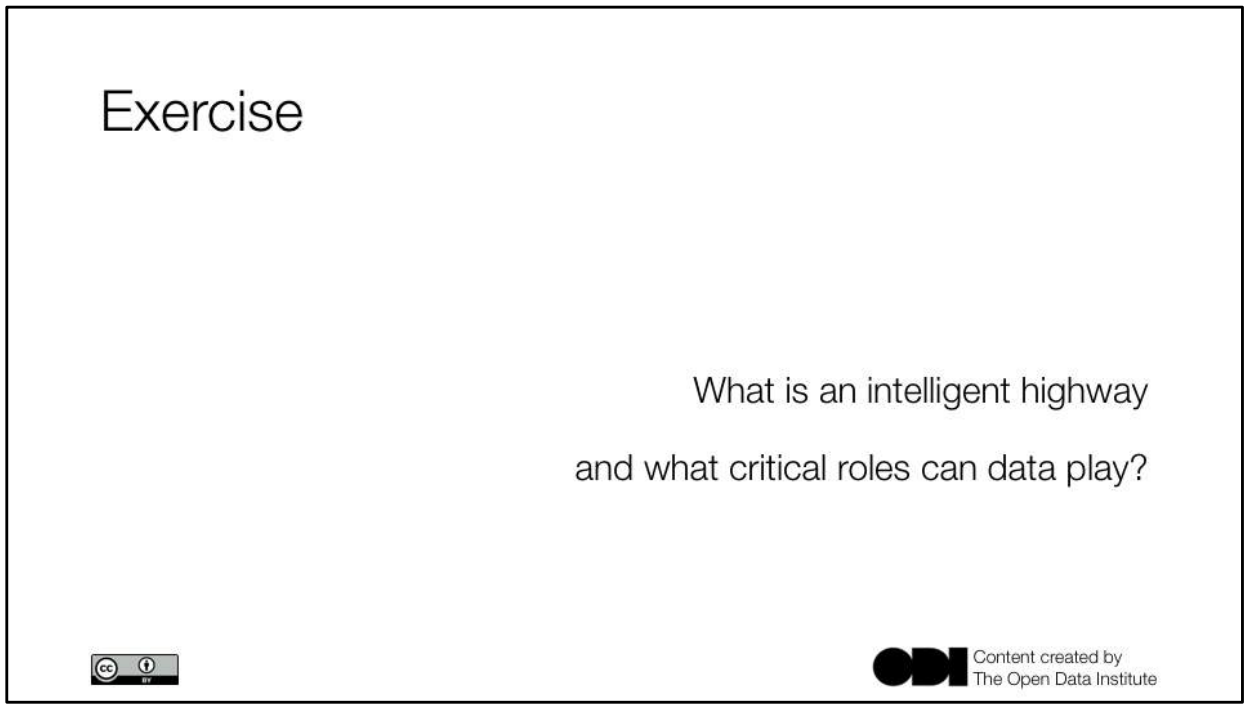


We are going to take a break from public transport with tickets for this session and look at highways.

Highways, like roads and cycle routes enable anyone to access them with the appropriate form of transport. Unlike the railways and the sky where highways are regulated and controlled to keep things from crashing into each other, highways don't always prevent crashes and can become very congested.

In this session we are going to look at three main aspects. Firstly we are going to look at what an intelligent highway is and why they are required.

We will then look at the European priorities for intelligent transport solutions focused on highways before looking at the businesses operating in this emerging system



Intelligent highways have been seen as part of the solution to this problem, but what is an intelligent highway and what critical roles can data play?

As before have a discussion in your groups and we'll then discuss ideas.

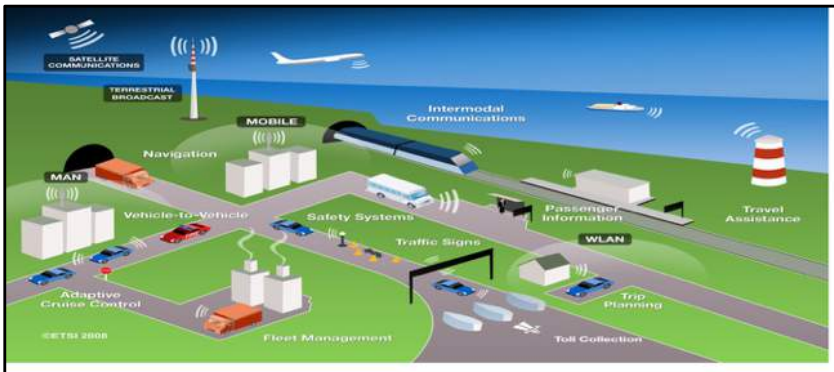
Intelligent highways

A highway without accidents and congestion which minimises pollution

Department for Transport

Report on Information on National ITS actions envisaged over a five year period

https://ec.europa.eu/transport/sites/transport/files/themes/its/road/action_plan/doc/2012-united-kingdom-its-5-year-plan-2012_en.pdf



The European Union has had a large agenda that has been pushing the future of transport technologies. As you can see from this diagram it all depends on the ability for each element of the transport system to communicate with or be aware of other elements.

Back in 2008 the EU created a number of key directives to push forward what it called the intelligent transport system.

Priority Area I: Optimal Use of Road Traffic & Travel Data

Open data

Communication standards

Smart motorways

Regulation for (semi)autonomous vehicles

https://ec.europa.eu/transport/sites/transport/files/themes/its/road/action_plan/doc/2012-united-kingdom-its-5-year-plan-2012_en.pdf

Intelligent transport systems priority area 1 covering the optimal use of road traffic & travel data.

Within this priority area there are a number of agenda items that the UK Department For Transport has chose as their own focus, the most important being open data and role of this data to connect everything.

Priority Area I: Optimal Use of Road Traffic & Travel Data

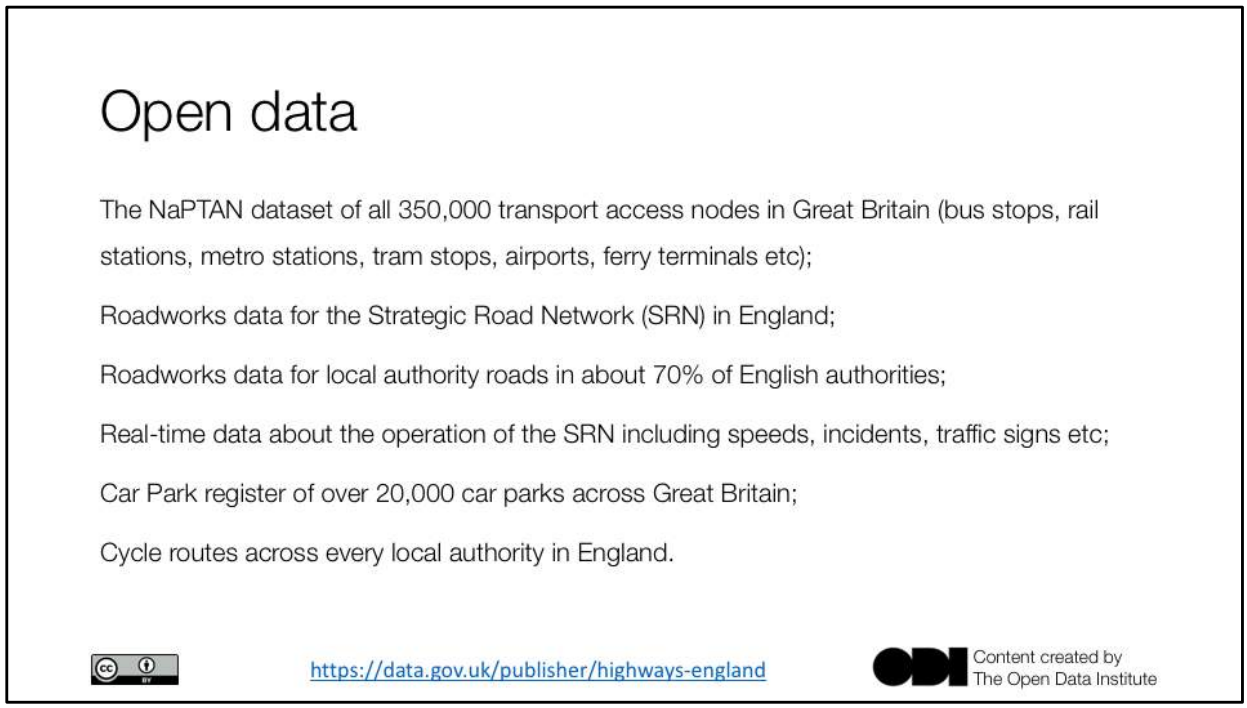
Open data

Communication standards

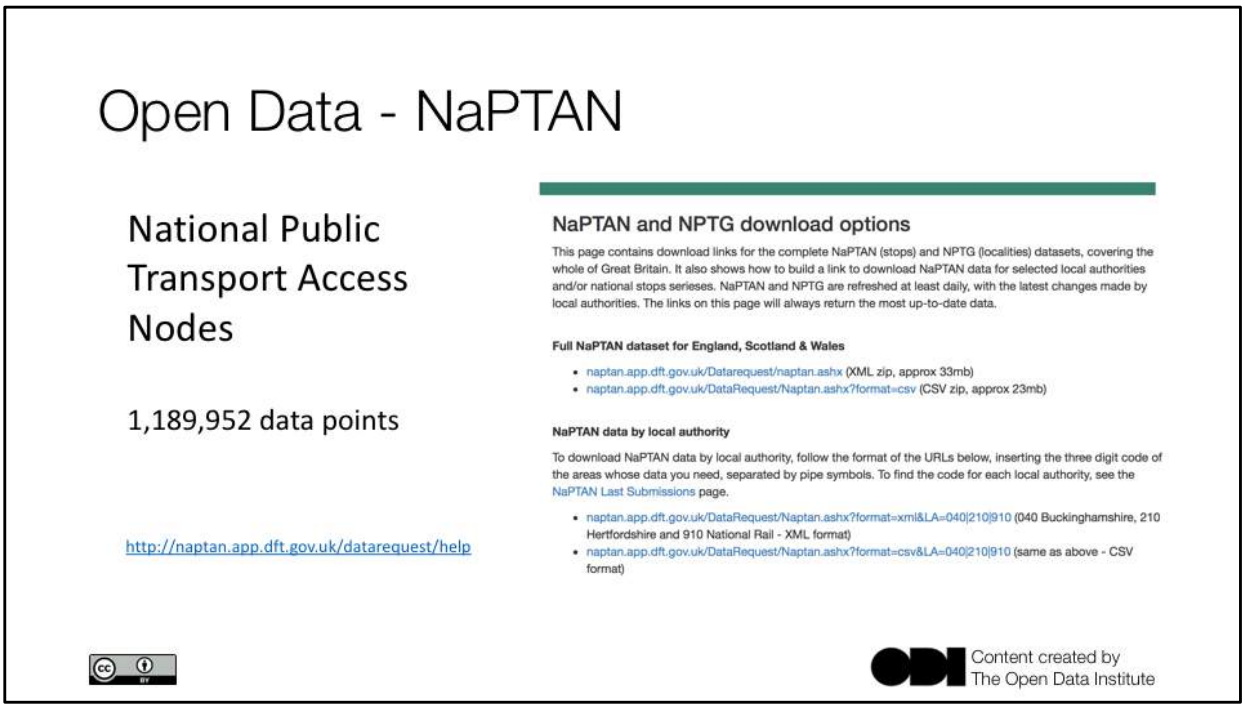
Smart motorways

Regulation for (semi)autonomous vehicles

So let's take a look at the Open Data that is now available in the UK.



The open data agenda presents opportunities to provide better information to the public to inform their travel choice. To date there have been a series of significant transport Open Data releases including:



The national public transport access nodes (NaPTAN) database is a system for uniquely identifying all points of access to public transport in Great Britain. It contains a record for each of around 400,000 bus stops across England, Scotland and Wales, as well as for all other transport terminals such rail stations and airports. NaPTAN is a core component of the national transport information infrastructure and is used by a number of other UK standards and information systems.

NaPTAN consists of:
a standard for identifying and naming access points to public transport
a database of all public transport access points in Great Britain
the XML schema for exchanging data or an alternative CSV exchange format version

The NaPTAN database is managed by the Department for Transport and updated by local authorities. NaPTAN is available under Open Government Licence and can be downloaded in XML or CSV format as a complete national dataset or in separate local authority files of your choosing. See NaPTAN and NPTG download options.

NaPTAN corrections

Size of dataset: 300,000
Error rate: 6% (18,000)

How long for one person to correct?

18,000 * 15 minutes = 270,000 minutes
= 4,000 hours
= 643 working days
= 3.3 years

In 2013 when the NaPTAN dataset was released the Open Street Map community imported the whole dataset into their mapping platform. They soon realised that thousands of the stop locations were incorrect. This led to local activities to correct the data involving both community members as well as those from local authorities. The Department for Transport supported this activity and following the correction, set up a system to automatically update data from the corrections. There is now a process for local authorities to source and update data on a daily basis, keeping the locations data up to date for all service providers.

So how long would it have taken one person to correct all the points. I did a calculation based upon each incorrect point taking 15 minutes to identify and correct. Given there were 18,000 corrections submitted, this equates to a 6% error rate, which is about right on any dataset as a minimum. For a single individual to correct all 18,000 points would take 3.3 years of effort, 7 hours a day for 40 weeks a year.

Not all data is well supported however and one person cannot be expected to maintain the quality of some of the big data we are now generating.

There are 147 providers of the NaPTAN data, so if we assume that each authority

dedicates 1 person to the task of correcting their own data the whole process might only take a week. However the distribution of stops managed is not even, especially in big cities.

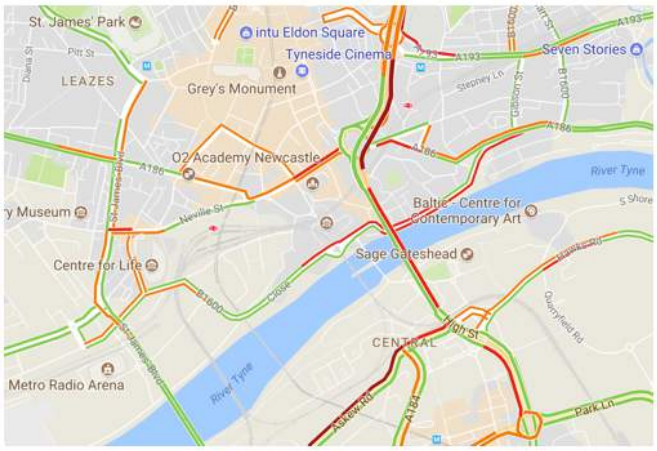
Open data –

Travel times

Highways England has a complex network of cameras and road sensors that work out accurate journey times based upon live data.

This physical infrastructure that detects traffic both in the road and via cameras is expensive and takes ages to install.

Android auto / Apple carplay



Content created by The Open Data Institute

Of course we know that mobile devices now provide a much richer amount of data that can be used to calculate live travel times and show congestion.

Such mobile data can now even show you how busy shops are during the day.

A worldwide system which only requires a mobile data connection. Google and Apple can collect a huge amount of incredible accurate data from the billions of connected devices. This is far more accurate than expensive road based sensors.

Open data (car parks)

http://isdublinbusy.com

Car parking data is another important aspect in big cities and airports. Although google might be able tell you when locations are busy it can't yet tell you exactly which car parking bay to park in. Parking censors can help provide the data and now the most useful sensors are linked to car parking lights that help guide the driver rather than just track them.

This shows and important aspect of gathering transport data. Wherever you put censors, think about how they will enhance the experience for users.

Priority Area I: Optimal Use of Road Traffic & Travel Data

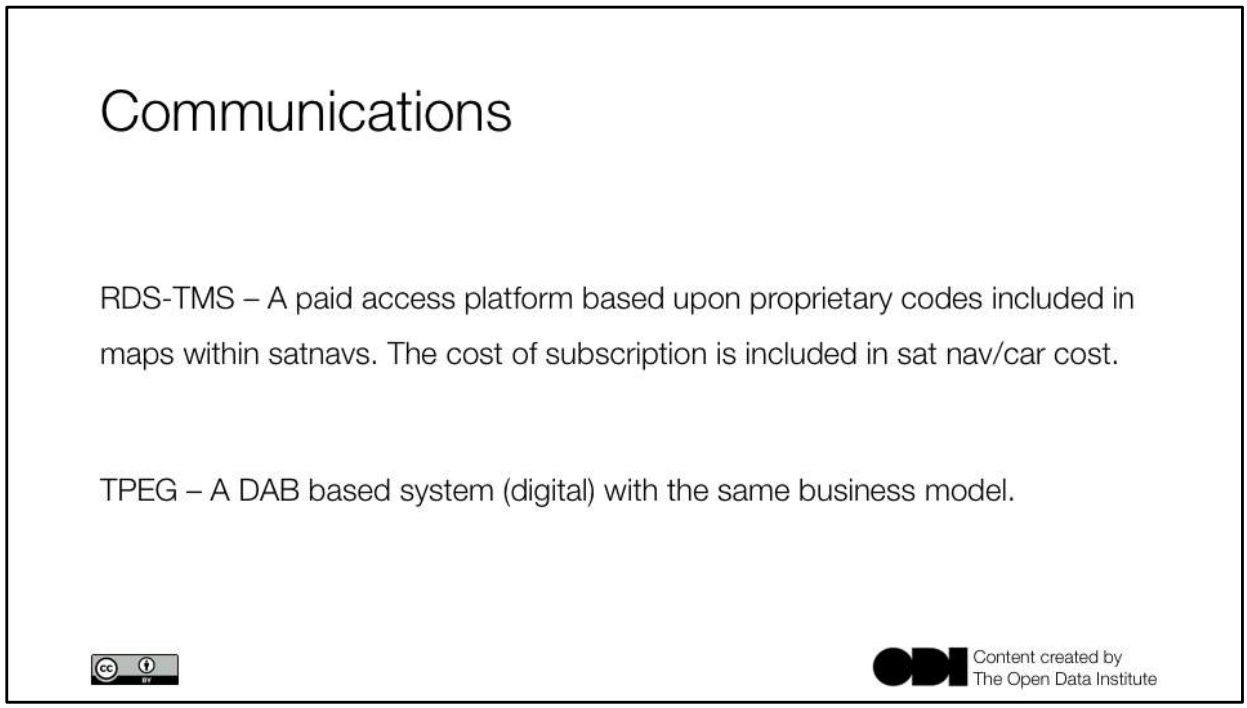
Open data

Communication standards

Smart motorways

Regulation for (semi)autonomous vehicles

Now lets quickly look at the communication standards that help on the highways, some you might have never heard of.



RDS-TMC: There are two competing services provided by private companies: INRIX and Trafficmaster They collect their own journey time data from ANPR, GPS and mobile phones and fuse it with other sources of data , e.g. police, Highways Agency, maintenance information

These are fully EU standard complaint services, which are free to receive for the lifetime of the vehicle or sat nav. RDS-TMC broadcasts have been in place for over ten years. In the UK TMC model, the cost of the service is included in the sat nav or vehicle or an RDS-TMC adaptor can be added for some older units. There is no other subscription or extra cost to the user.

Between INRIX and Trafficmaster, there are now over 4 million UK RDS-TMC sat nav users and this continues to grow as new vehicle makers are adopting TMC all the time. The additional cost of the TMC licence is small compared to the cost of the sat nav or even the vehicle..

A non UK driver or sat nav user with a TMC unit with Europe wide mapping and codes can also receive RDS-TMC and hence traffic information in their own language at no cost. Most vehicle makers provide Europe wide mapping on their vehicles already (BMW, VW) and so the appropriate codes are already available for many visitors to the UK.

TPEG: Both Trafficmaster and INRIX provide DAB TPEG services using enhanced data based on the TMC business model.

Both work throughout Europe if you have the codes.

With roaming charges now abolished in the EU, more people can simply use their devices and the satnav industry is suffering, driving up costs of in-car head units which include sat-nav vs. buying an iPad and sim card.

Priority Area I: Optimal Use of Road Traffic & Travel Data

Open data

Communication standards

Smart motorways

Regulation for (semi)autonomous vehicles

Another UK scheme to control traffic flow and keep things moving is smart motorways.

Smart motorways have a number of key features and the M25 uses all of them.

One of the key technologies in use involves detecting traffic flow. At the time done with expensive sensors under the roads and camera technologies (this can now be done with mobile phone location systems). Upon detecting congestion the system automatically enforces lower speed limits in sections to ease the flow of traffic. A combination of flexible speed limits and speed cameras which enforce this help ensure that traffic keeps moving and the congestion can clear.

Before this technology was in use, people would approach congestion at speed and cause a traffic wave (or traffic shock) where the congestion slowly moves backwards but may not clear until the middle of the night.

Another core feature of smart motorways is the ability for all lane running. This is where the previous hard shoulder, used for break downs and safety is converted into a full running lane. Refugee areas are then installed every so often if people can make

it there. If they are unable to make it then lanes can be closed using gantry signs. Evidence suggest that journey times are improved on smart motorways and safety isn't impacted (either positively or negatively). Safety groups are worried about the lack of a spare lane for emergency services but thus far the evidence is that there has been no significant impact on safety.

Priority Area I: Optimal Use of Road Traffic & Travel Data

Open data

Communication standards

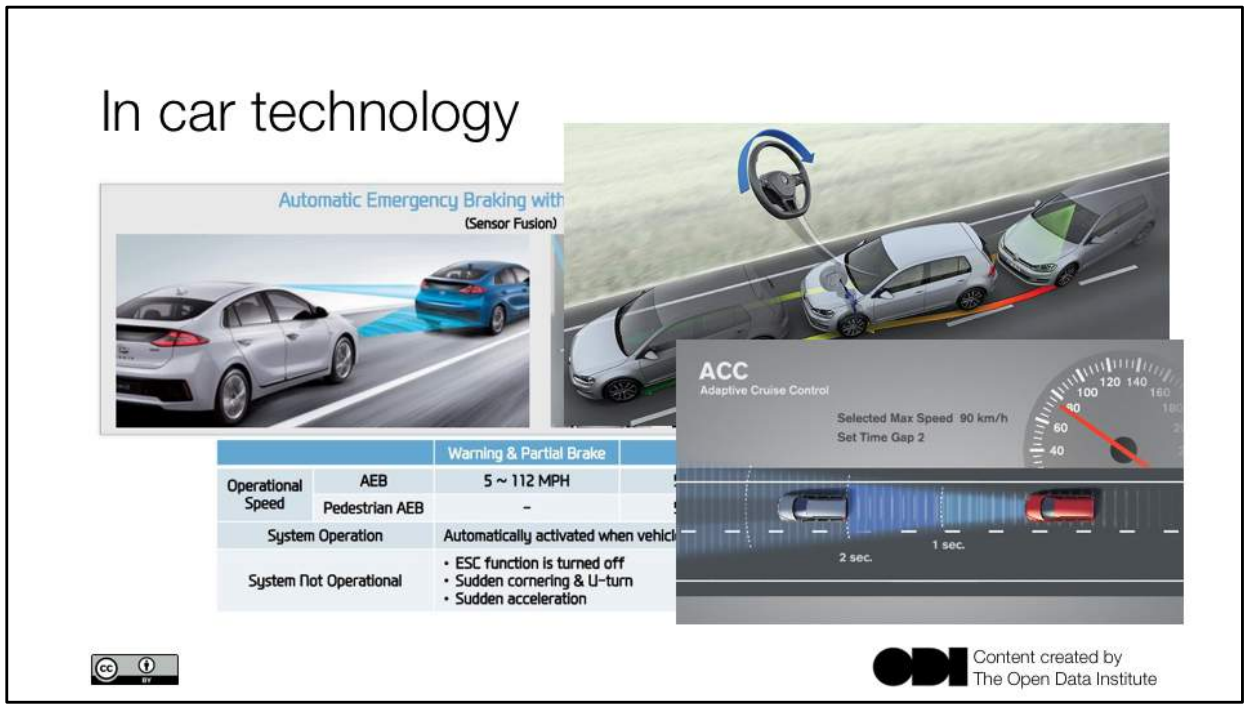
Smart motorways

Regulation for (semi)autonomous vehicles

(semi)Autonomous vehicles

The first autonomous cars didn't look much like cars as the sensors and camera rigs required made them work.

However, the technology of self-driving drives has been evolving hugely over the past few years and we are starting to see some of the key developments in this area make their way into every day cars.



I've recently purchased a new family car that combines safety and efficiency with features. In fact this is my new car, and one of the main safety features is Automatic Emergency Breaking. The car has 5 radars transmitters and two cameras which keep an eye on the environment and watch out for hazards.

Automatic Emergency Braking uses the front facing radar to ensure that you don't hit the vehicle in front of you.

The same radar sensor is also used to give you adaptive cruise control, you set your maximum speed and then let the vehicle take over, if the vehicles in front slow down, the car does also, if they speed up so do you. These are fantastic features that rely on the recording and instant processing of data.

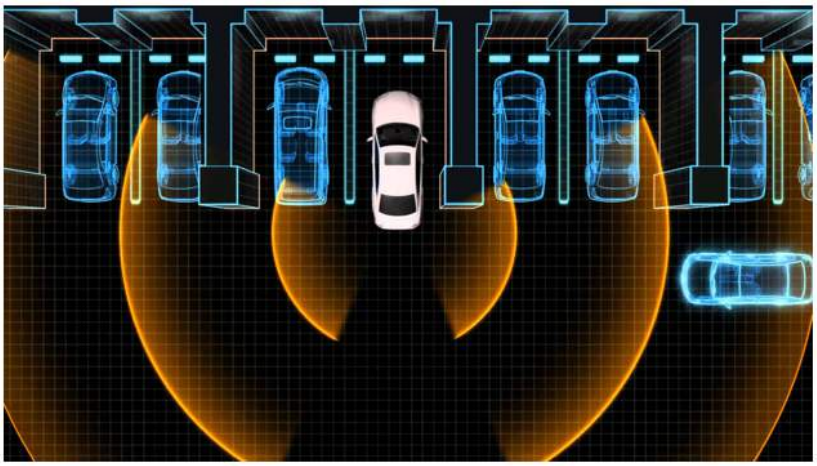
The front facing camera is used to find the lines on the road and on dual carriageways and motorways will automatically return your car to the middle of the lane.

I love the combination or these features, and the imense amount of great data processing.

In car technology

But they didn't stop there, I said there were five radar sensors and two more of them do blind spot detection so you get a warning of traffic to your left or right.

In car technology



The final two radar nodes detect cross traffic when reversing out from a parking bay. Given that parking bays seem to be getting smaller (or cars getting bigger) this is a fantastic feature.

A big part of the move towards autonomous cars is not for safety as road safety is already pretty good, but to ease congestion. But why is this better than the current situation? Well the theory states that if everyone has knowledge of where everybody else is going then a self forming system can form and congestion can be avoided all together.

Solving congestion (the science)

To look at this in action we need to turn to the theory of how planes are boarded.

So here is the plane I presume you flew here on, this is a China Airways Boeing 777-300 EI with first, business and economy cabins. What I'd like you to do for me is to work out the fastest and most satisfactory way for everyone to board the aircraft. You can design your own algorithm and if you need to come up and look at the plane layout on the projector, please do. The group who designs the algorithm which gets everyone one happily on board wins this…

The majority of airlines board their planes in the same way and people think it is pretty slow, lets find out what the best methods are.

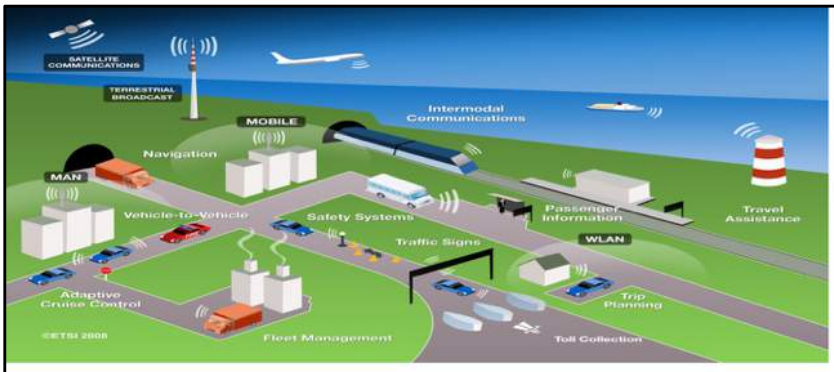
…video…

So there we have it, the back to front method is the slowest. Reverse pyramid is complicated but random is by far the fastest.

Boarding in order of check in is the best method as then it is down to people to assign their number and not have to crowd around the gate. The biggest problem in the UK

is people wanting to get on first so they can get their luggage in the overhead bins. Some even think that the bins in business class are slightly bigger and put their luggage in their before taking a seat in economy.

So random boarding is faster than the current back to front method, but not as pleasing as any of the methods that mean you know where people are going. Thus the conclusion for autonomous vehicles is that if they can communicate with each other than congestion can be avoided and journey times reduced. Of course you have to be careful to regulate the adoption such that the same premium services don't, such as priority boarding, don't start to exist on the road.



It is clear that communication is going to be a big part about easing congestion. Cars will need to know where other cars are going and thus not have to rely on radar and camera inputs. These can then be used to avoid obstacles which are not communicating. We are a long way from this reality however but perhaps it is achievable for traffic lights, speed control and routing.

Outcomes

What is an intelligent highway?

Intelligent transport solutions in Europe

Businesses on the highways

To conclude this session I want to look at some of the emerging businesses using the highways to make a living.

Uber

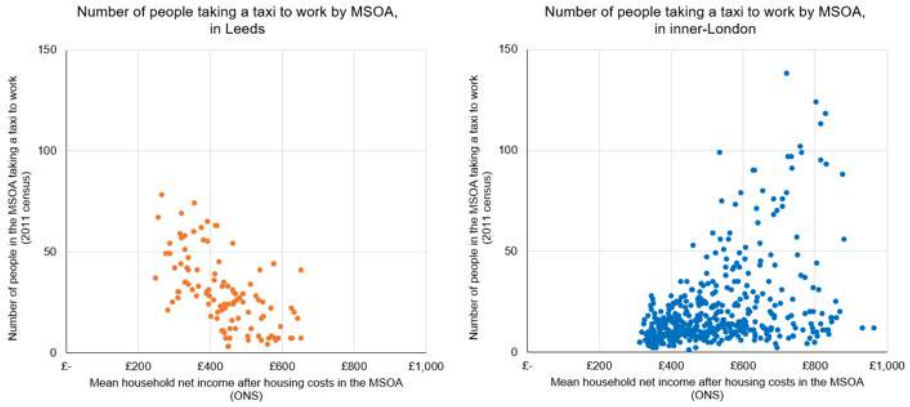
Upfront pricing for Taxi fares bookable via a mobile app.

No cash required.

Their entry into the market wasn't exactly welcomed or clean.

Uber are disruptive in the marketplace in some places unsetting the local monopoly. They offer upfront pricing and mobile payment meaning anyone can use them without having to worry about local currency or the language. The entire trip can be booked and paid via a mobile. Sometimes this can undercut the local services which has led to accusation that UBER are not paying licensing fees. The evidence suggests this is not the case. Although evidence does suggest that there has been some bad incidents of foul play in both directions between Uber and its competitors they now have the former CEO of Expedia as their new leader.

Uber



http://www.tomforth.co.uk/defendinguber/

The best data we have on the demography of taxi use comes from the 2011 census, in the methods of travel to work section. This shows that in London taxis are a luxury used by the rich. But in Leeds they are a connection to employment for the poor. For many, taxis are the only real competition that exists to restrain private bus companies' price rises. Most people in Leeds that for many trips, especially with more than one person, a taxi is just as cheap and much more convenient than the bus.

And so, while good public transport remains an option that is unavailable to England's large cities, I will continue to support Uber. I'm not sure why a multinational chooses to lose money helping poor people in Leeds get to work, but I'm glad that it does.

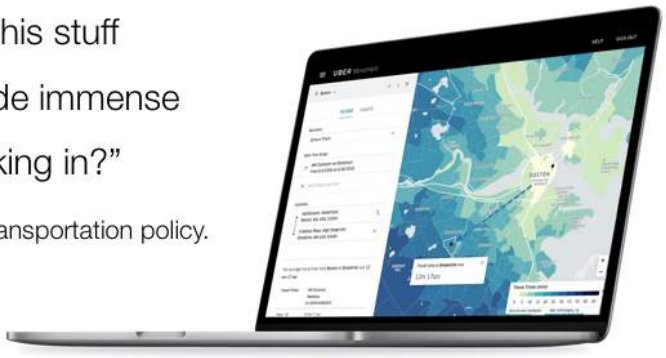
Additionally Uber opens up its data without being asked to do so, which is rare for a private business and a lot of other transport companies don't like this, even though they could use it to compete by supplying cheaper bus routes.

Uber and open data

"We don't manage streets. We don't plan infrastructure. So why have this stuff bottled up when it can provide immense value to the cities we're working in?"

- Andrew Salzberg, Uber's chief of transportation policy.

https://movement.uber.com

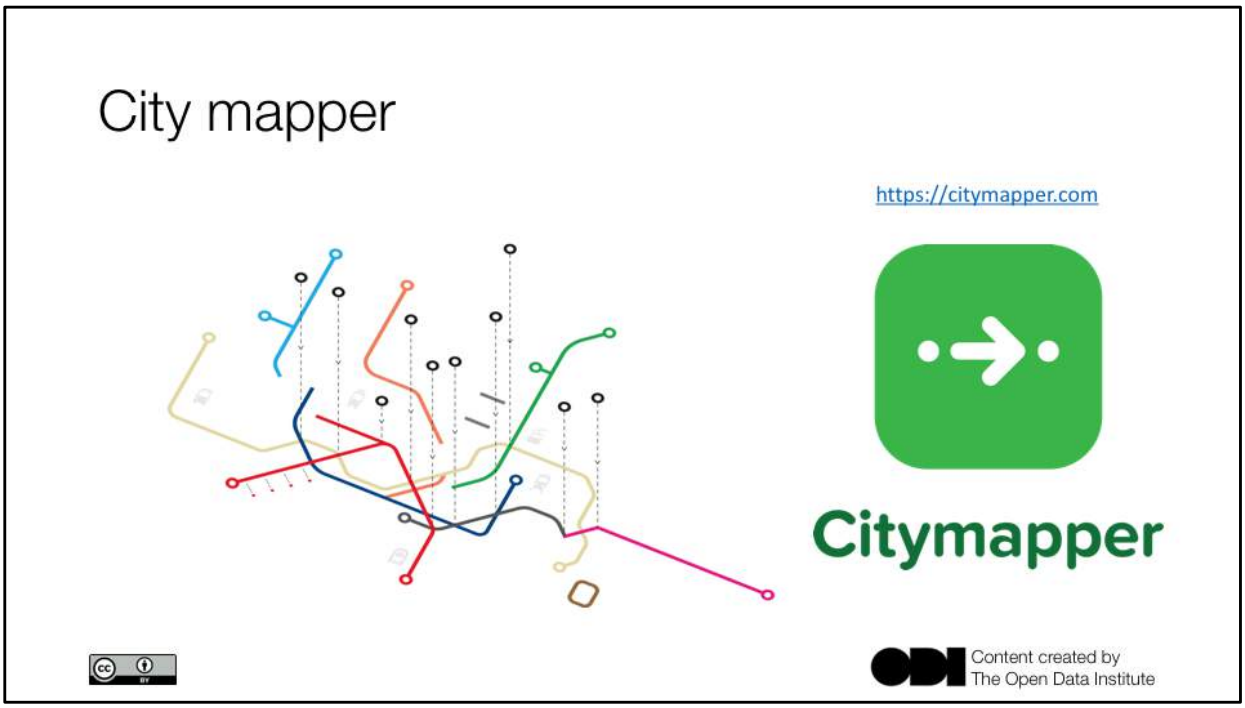


Known as Uber Movement, the website offers data sourced from Uber trips in more than 450 cities. Planners using Uber Movement will be able to search for average trip times between two points for specific times of day, days of the week, and months, information that could help cities improve traffic flow.

Uber Movement's GPS-extracted information will help city planners to examine traffic patterns and monitor how infrastructure changes like road closures can impact congestion. Routinely tracking such behavior could help city planners make decisions about where to place new lanes to compensate for shutdown road lines during renovation periods. Uber says it's also looking at releasing access to the data as an API, but is "trying to figure out how to do it in a performant way" at this stage.

So why aren't competing companies using the uber data to take away their business?

Is it because they operate in a world where data is not a part of the core business practice. A intelligent disruptive new comer is being shunned rather than exploited?



Cities are complicated. City mapper uses the power of mobile and open transport data to help humans survive and master them.

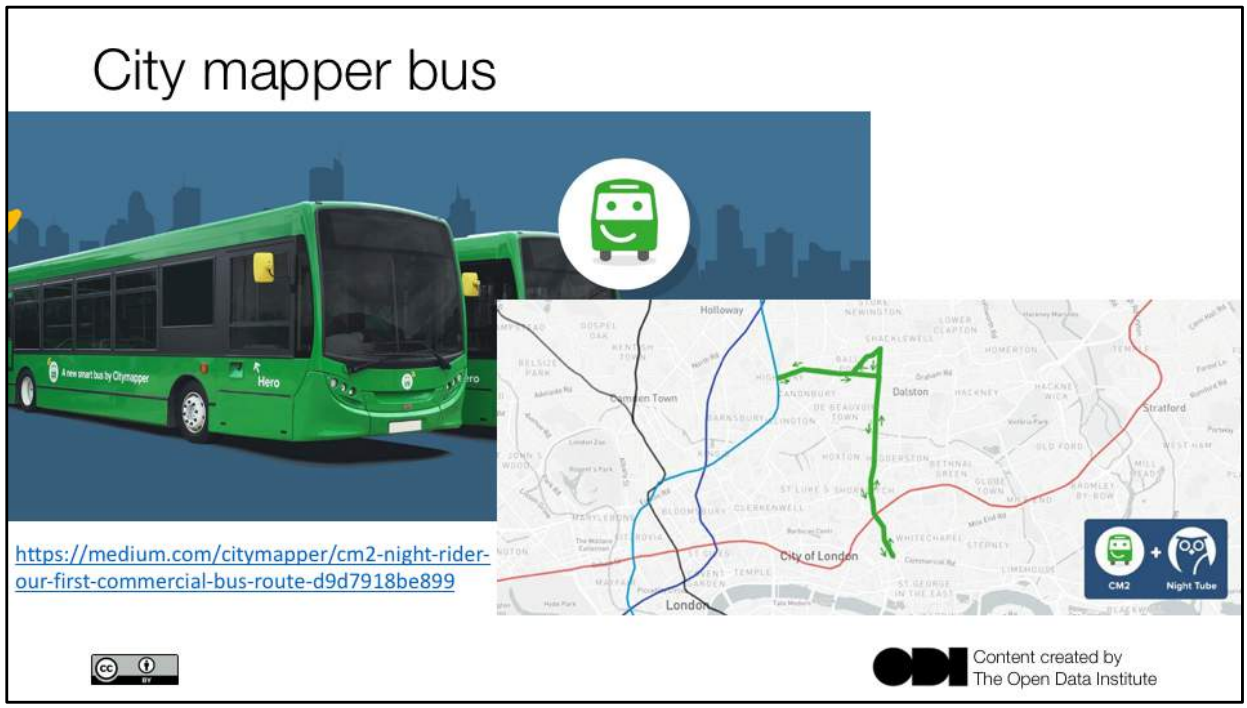
In every city they cover, citymapper has gone on to be the best transport planning application as this is their focus is only on public transport.

City mapper and open data

https://medium.com/citymapper/building-a-city-without-open-data-124356672deb

City mapper will only cover cities that release open data as they believe that the benefit should be for everyone. They don't pay for the data and they don't sell data, yet they have millions in VC funding for their open approach.

I would also recommend reading their open data blog which gives insights in how they also work to benefit the local transport providers and future infrastructure planning.



Citymapper have even decided to try running their own bus in london to cover some of the gaps which the private bus companies don't or won't provide. Again better use of data to identify gaps in the market. This year citymapper obtained a license from TfL and

We found central London fairly well covered during the day by existing TfL services, but we identified bigger gaps in the night network. People in London are staying out later, especially in East London. For example there are more late night destinations on Commercial St, without any night bus support.

The emergence of the Night Tube has also encouraged late night mobility, but also exposed gaps in the supporting night bus network. We found Highbury & Islington Station (an important hub on the night Victoria line) with inadequate bus coverage linking east.

Bike schemes

Dock based

Dock free

We have already discussed bike earlier in the week and again bike companies are starting to exploit data in their cities using GPS trackers on bikes the not only enhance security but allow for complex data analysis.

Dock-based (like stations), not really the point
MoBike (Chineese) – Dockless (launched in Manchester)
GPS, long battery life and cheap data allows the bike to remain connected

London light cycles

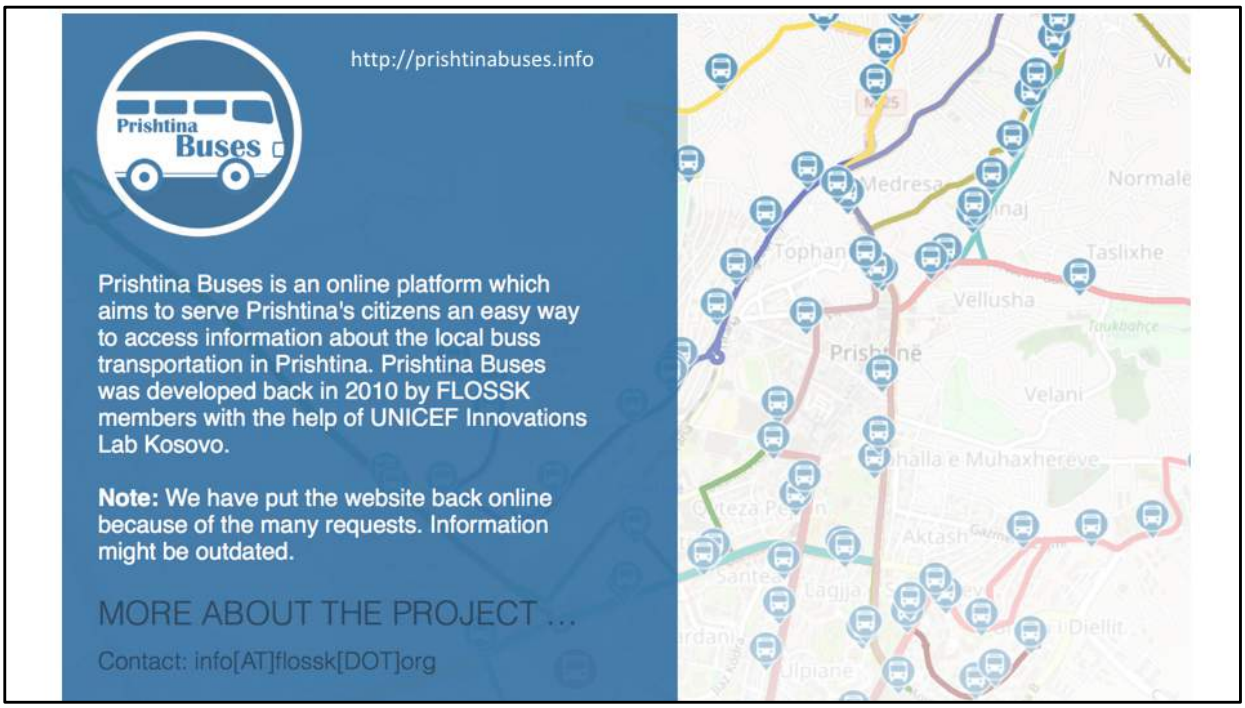
https://vimeo.com/33712288

Content created by
The Open Data Institute

This visulisation shows a summary of five million bicycle journeys made in 2010/11 in central London as part of the Barclays Cycle Hire scheme. Origins and destinations of each journey recorded and animated along a curved trajectory.

This animation shows the effect of changing the length of the 'trail' left by each journey (starts to increase from 15 seconds into the animation). By changing the prominence given to more common journeys (from 45 seconds onwards), structure emerges from the apparent chaos of journeys.

Three major systems become apparent (from 1 minute onwards) - Hyde park to the west, commuting to/from King's Cross St Pancras to the north and Waterloo to/from the City to the east.

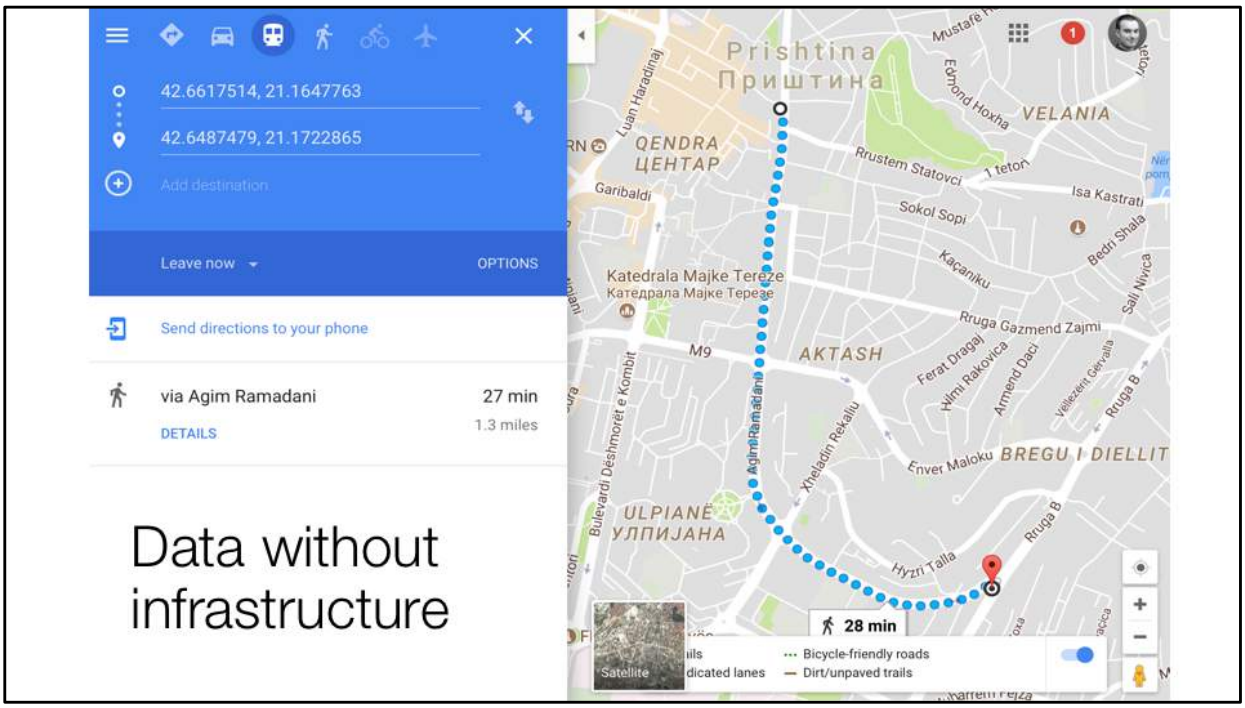


As the capital city of Kosovo, Prishtina has no information system on public bus transportation.

However, there are 16 bus lines operating in Prishtina and in the closest area, but locals as well as internationals are very confused about its operation. This fact, together with our belief that people would use it more often if there is more information, was the main impetus to start this project.

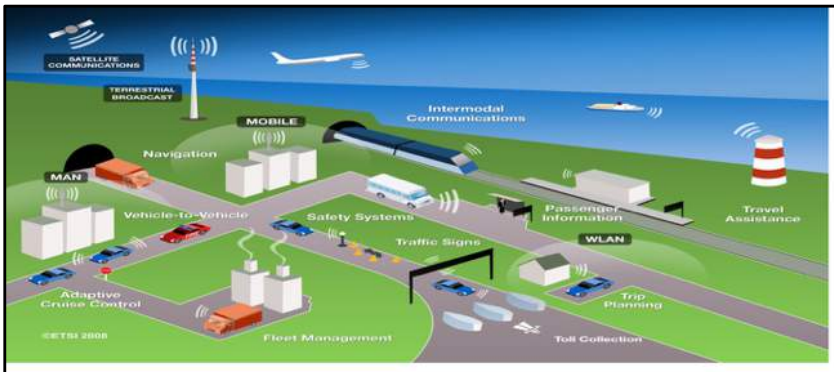
There is now an open data set available detailing all the locations of bus stops and routes that was collected by the community. This activity was carried out in 2010. The website is still online however there is a notice that says that the data is not maintained and likely to be outdated.

- To make a cheap solution you need to create a passionate community to solve a real problem. Not compete with existing solutions.



This is a good example of data without a supported infrastructure. The short term benefits are soon lost.

Here is a google maps route search I did recently in Prishtina as you see, there is no public transport mechanism known.

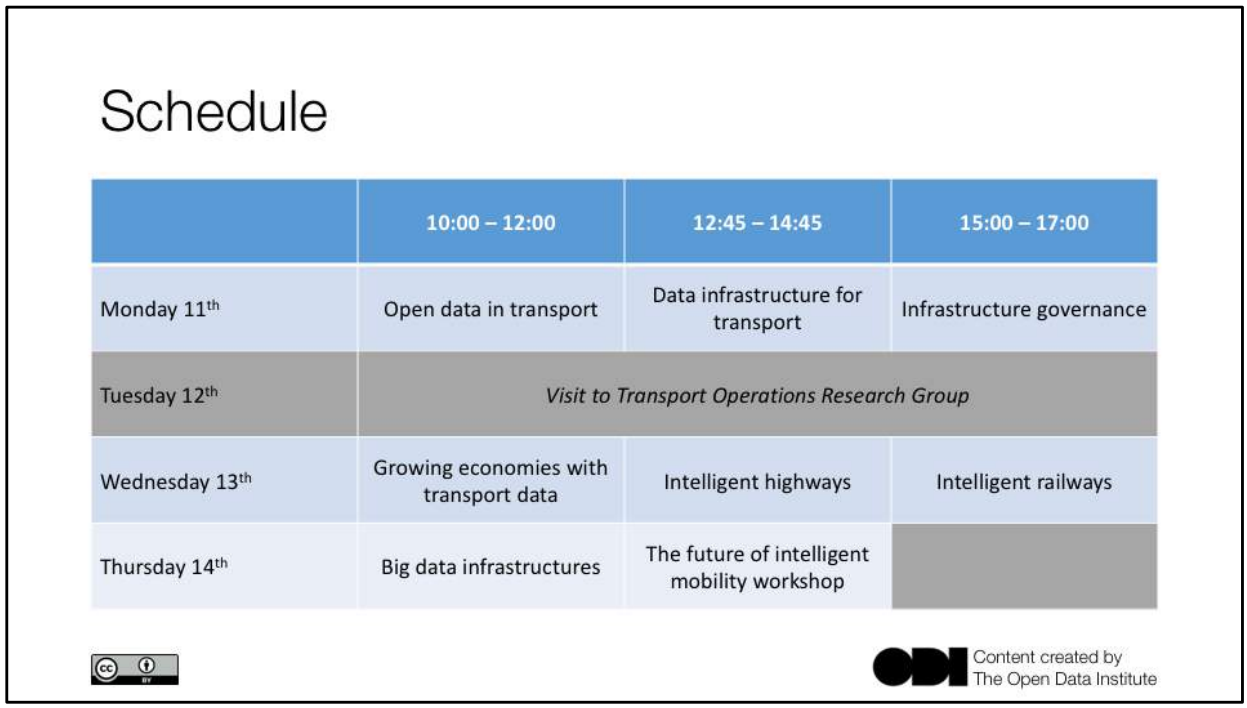


It is clear that communication is going to be a big part about easing congestion. Cars will need to know where other cars are going and thus not have to rely on radar and camera inputs. These can then be used to avoid obstacles which are not communicating. We are a long way from this reality however but perhaps it is achievable for traffic lights, speed control and routing.

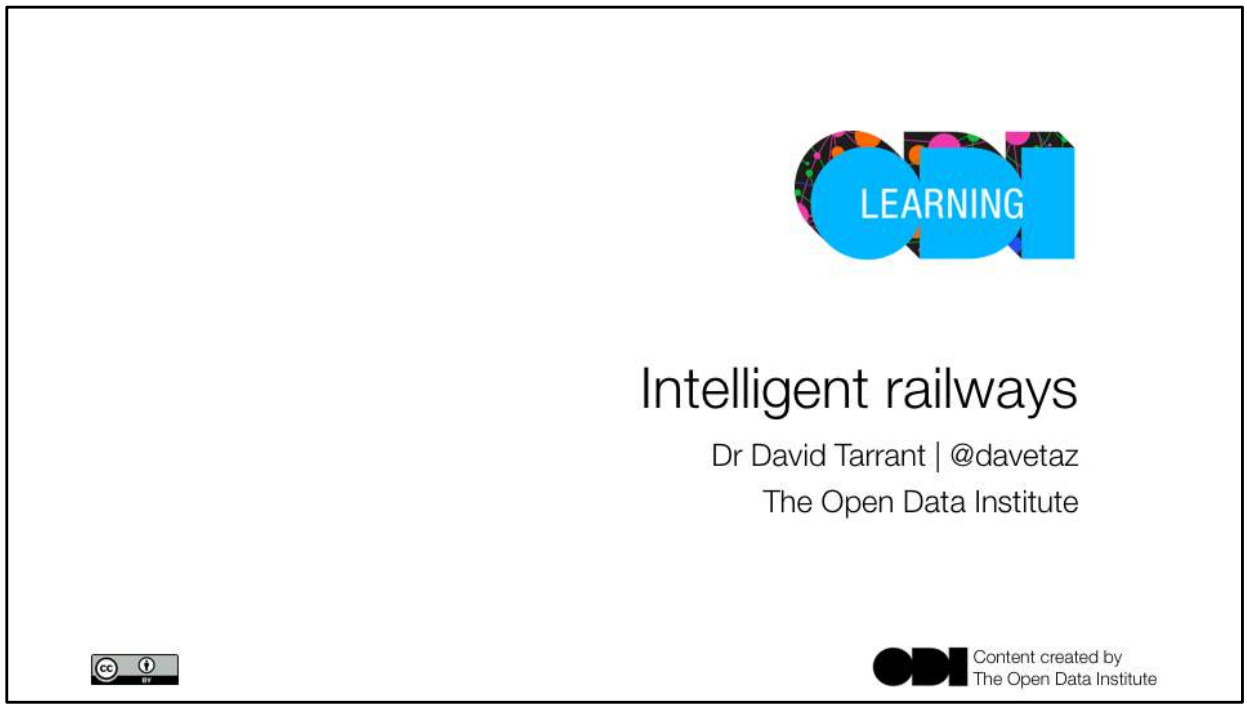
Transport Data Infrastructure

Dr David Tarrant | @davetaz
The Open Data Institute

Content created by
The Open Data Institute



Welcome back. I hope you had a informative time yesterday. Today we are going to dig deeper into how transport data is able to grow economies before looking at some of the innovations an opportunities in the highways and railways sectors. For the last session today we are going to take a short walk to the main station to see the how big data on the railways is delivered to passengers through it's many channels.



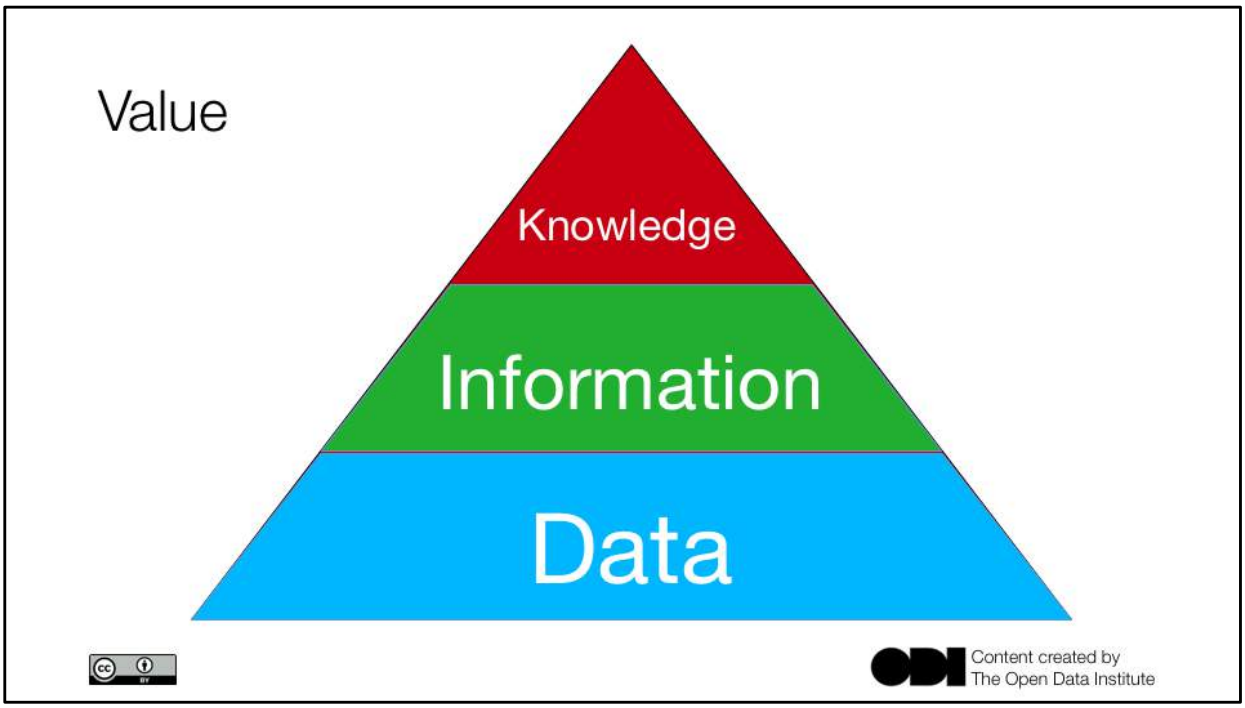
And now we get to talk about my favorite area of transport, railways.

Outcomes

Technologies on the railway

The commuters journey

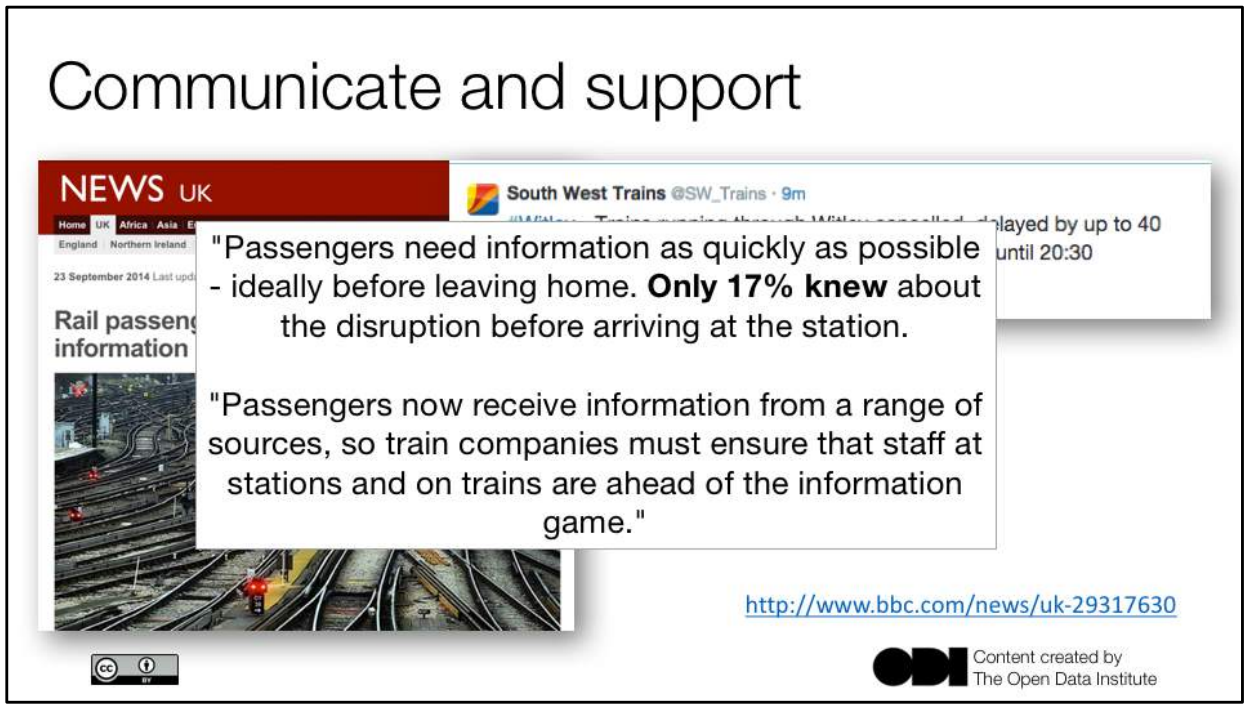
Big data science in railways



The majority of transport is used for commuting to and from work. The expert on your commute is you, no one else has the same knowledge. You might share some common parts of the route with others but not the whole route. Especially if you have a partner who can pick you up from a number of stations which can affect your options.

So in order to make the right decision you need enough data to apply your knowledge to.

This session looks at how I use transport data to save 30 minutes a day and a lot of stress when things go wrong.



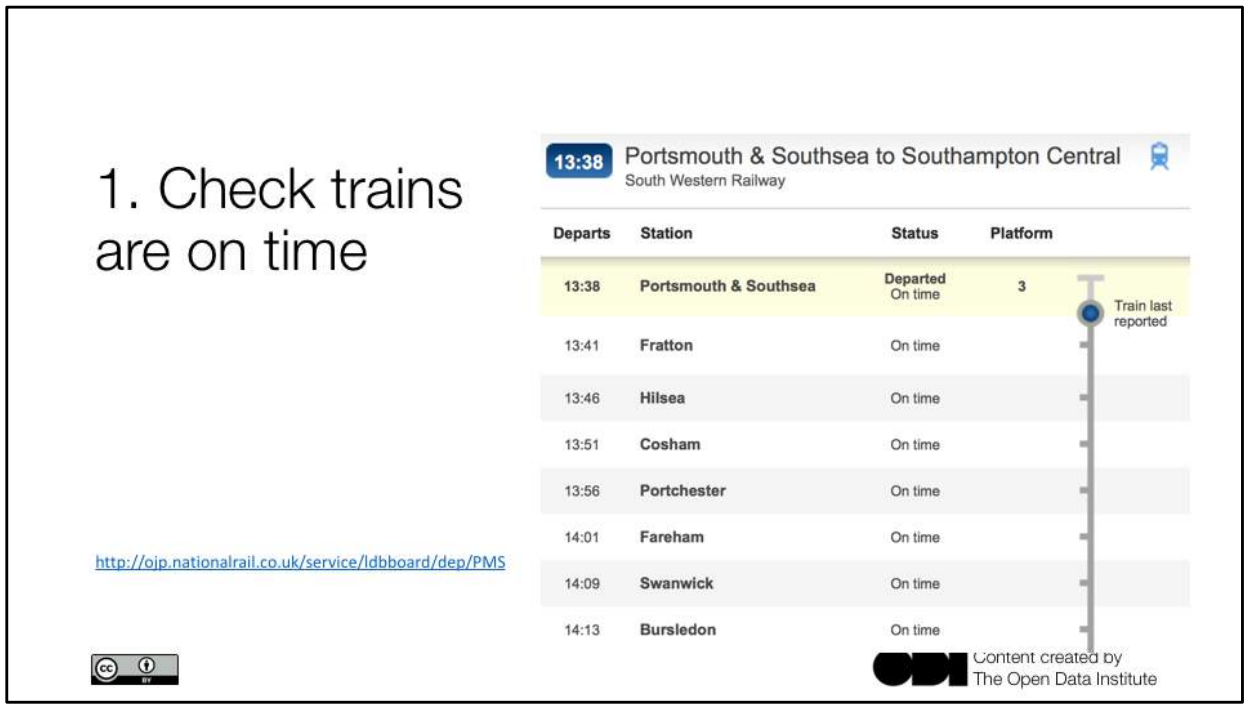
Every year, like clockwork, our news channels are populated with railway news about the unreasonable rise is fares, or delays which always seem to be getting worse or lack of information.

Yes fares are rising but this rise is controlled by the Retail Price Index (RPI) which is a measure of the changing cost of a fixed basket of goods and services over time. Both the RPI and Consumer Prince Index are calculated by combining together around 180,000 individual prices for over 650 representative items. Differences between CPI and RPI arise due to coverage, the population base of the indices and the way in which individual price quotes are combined at the first stage of aggregation.

As rail tickets are a product, they are allow to increase as per RPI, which since the financial crash has been higher than CPI. Ironically the membership fee for Unions is part of the calculation that works out what RPI is and last year the RMT union increased their membership prices by 1.8%.Argments are it should be linked to CPI, not RPI, which is lower currently.

The current system is also much more informative and drive by data however. Rail companies do face challenges when customers are more informed than staff, but this

is slowly changing.

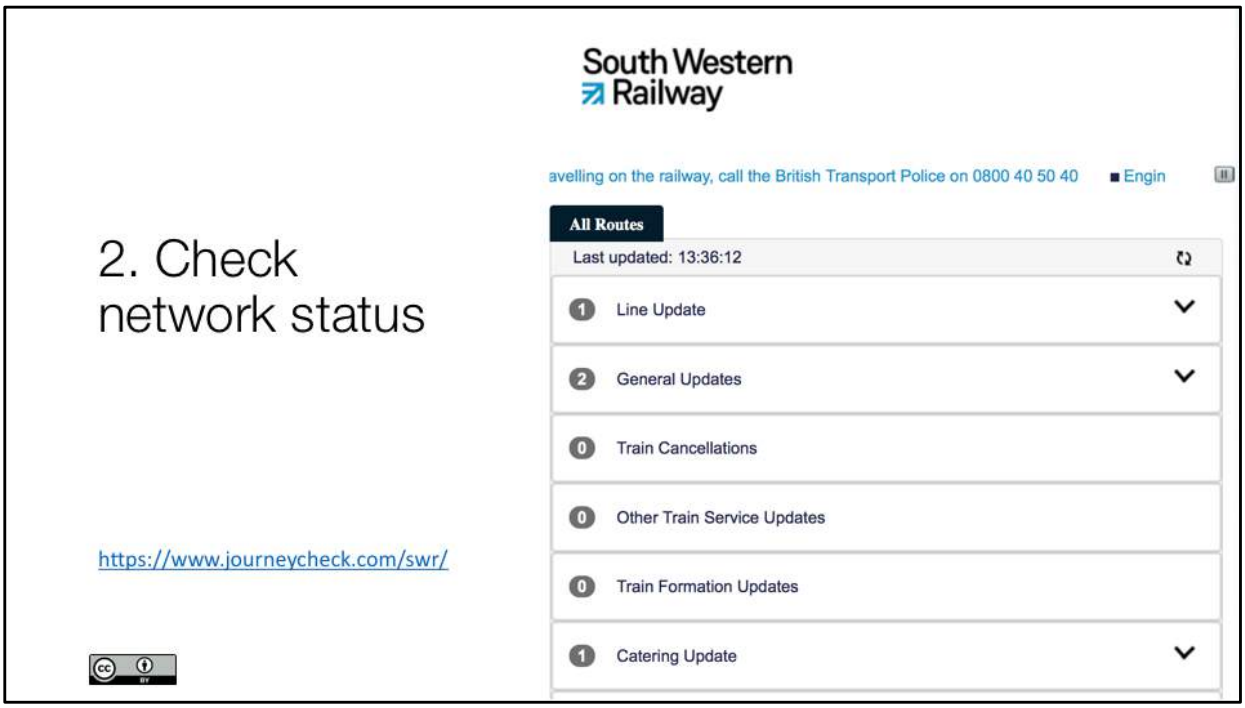


Take my own commute to London.

I set my alarm in the morning at 6:23 am.

This is the time my train departs its origin station to come and pick me up.

So the first thing I check when I wake up is if this first train has left.

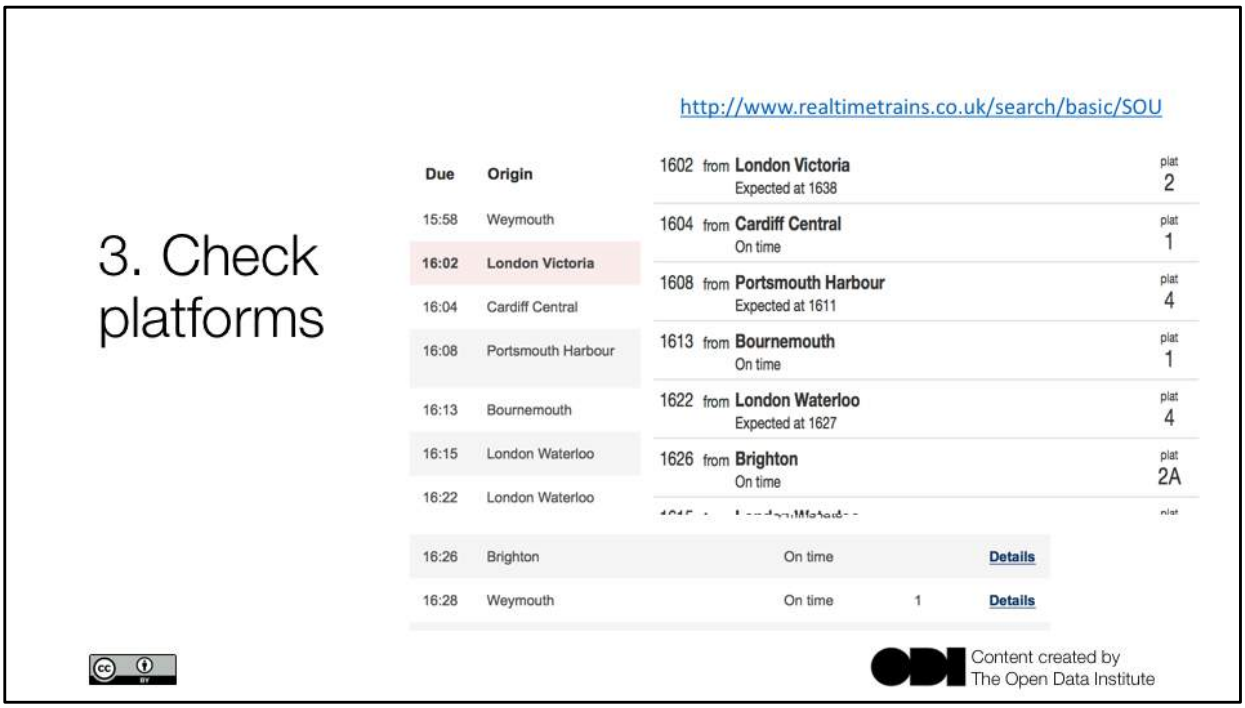


I have to change at a hub station to get into London, so I next check the network status to see if trains are going to be delayed on the network.

Here I'm looking for a line update. These are the worst and affect the whole network. At this time yesterday there was one line update which was a weather warning meaning trains would be running slowly in windy areas.

General updates I ignore, but cancelations and other updates I check to ensure there are no diversions.

Finally I look at the formation updates to see if my train will be running in reverse. This tells me where the free seats will be, front or back.



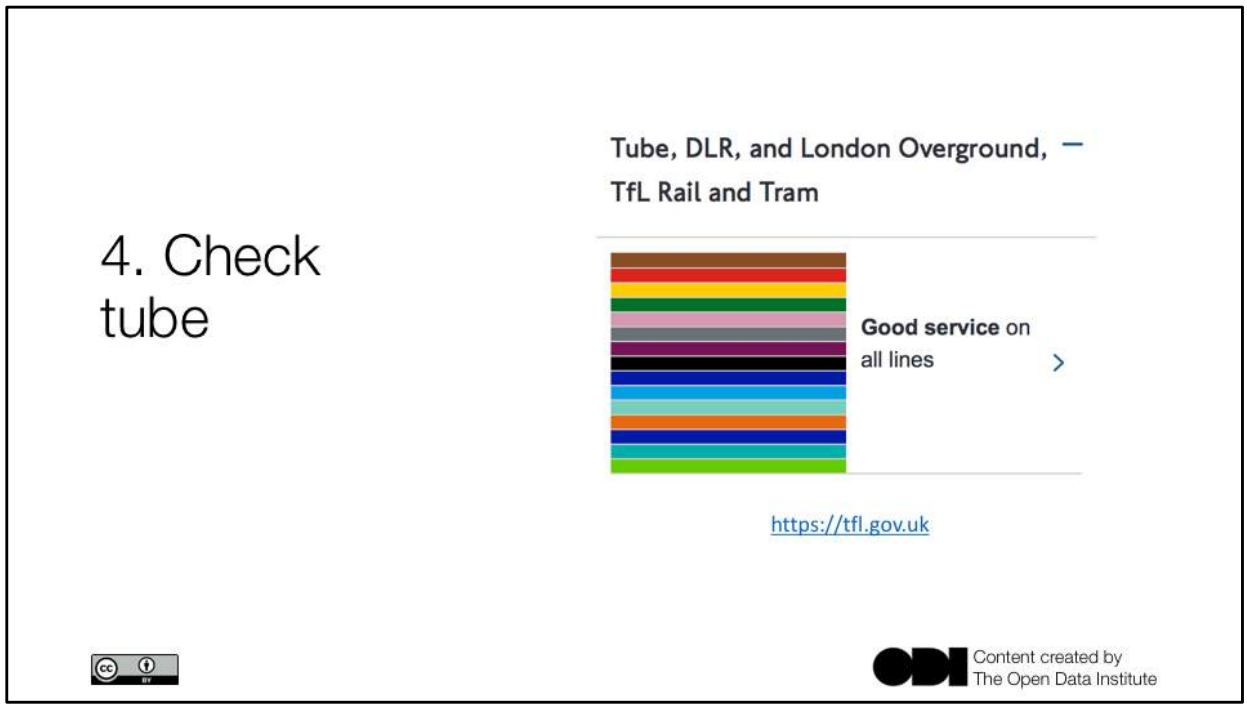
Next I check which platforms my trains will be arriving and leaving from at the hub. The hub station has A and B ends to the platforms and depending which end of the platform your train comes in on could increase walking time by up to 3 minutes, which on a short connection is bad. Sometimes it is best to stand in a busy carriage just to make the connection.

As can be seen here the website of the official provider does not show the A and B data, however the open data available from another application Real Time Trains does.

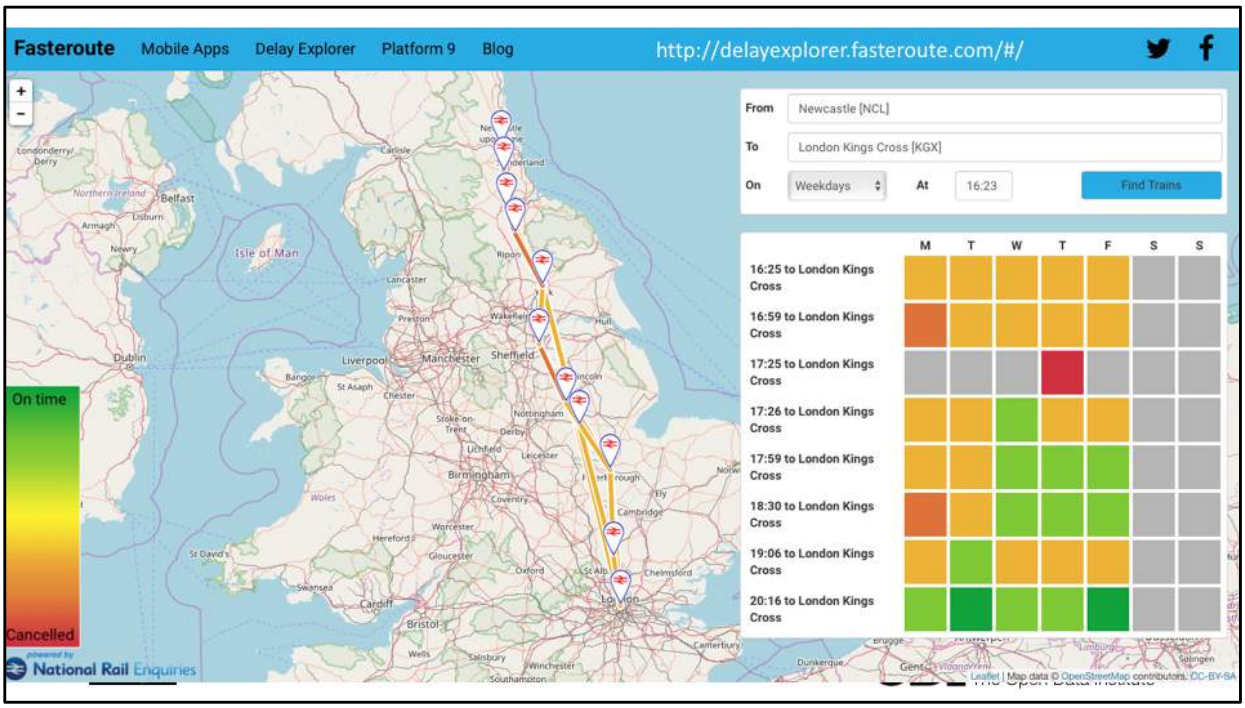
Real Time Trains shows a complete view on the open data including working timetable and real time table for both planned and actual train movement respectively.

RTT also shows you advanced information about the trains. Such as type of train, this helps with planning where to sit, just like a seat layout on a plane.

RTT also allows you to view the historical information for a train, which I use to claim compensation, more on that in a second.



Finally when I get close to London I check the status of the tube to make sure I pick the right route between Waterloo and Liverpool street, where the ODI is.



I mentioned that RTT gives historical train running data. One of the companies that uses this data is fasterroute.

On average, a Londoner spends 18 months of their life travelling to and from work. For commuters all across the country, travelling to work often involves multiple stages, meaning a delay at any stage could potentially be amplified as the journey progresses.

There are now a multitude of apps providing real-time data to help people plan journeys they are about to take. What often isn't included are indications of a particular service's reliability based on historic records of punctuality. Such a tool would be just as useful in planning a new journey, or reviewing a regular commute.

With funding and support from the Open Data Institute, the Fasteroute/Visualising Rail Disruption team built a web application, Delay Explorer, and integrated records of historic punctuality into their app, Fasteroute. With these, users can save significant amounts of time by exploring potential routes and planning journeys by rail using services that are more reliable. They can also avoid risky connections that could be missed due to trains that are often delayed.

Autonomous Trains

Communicate with each other and create virtual sections between trains.

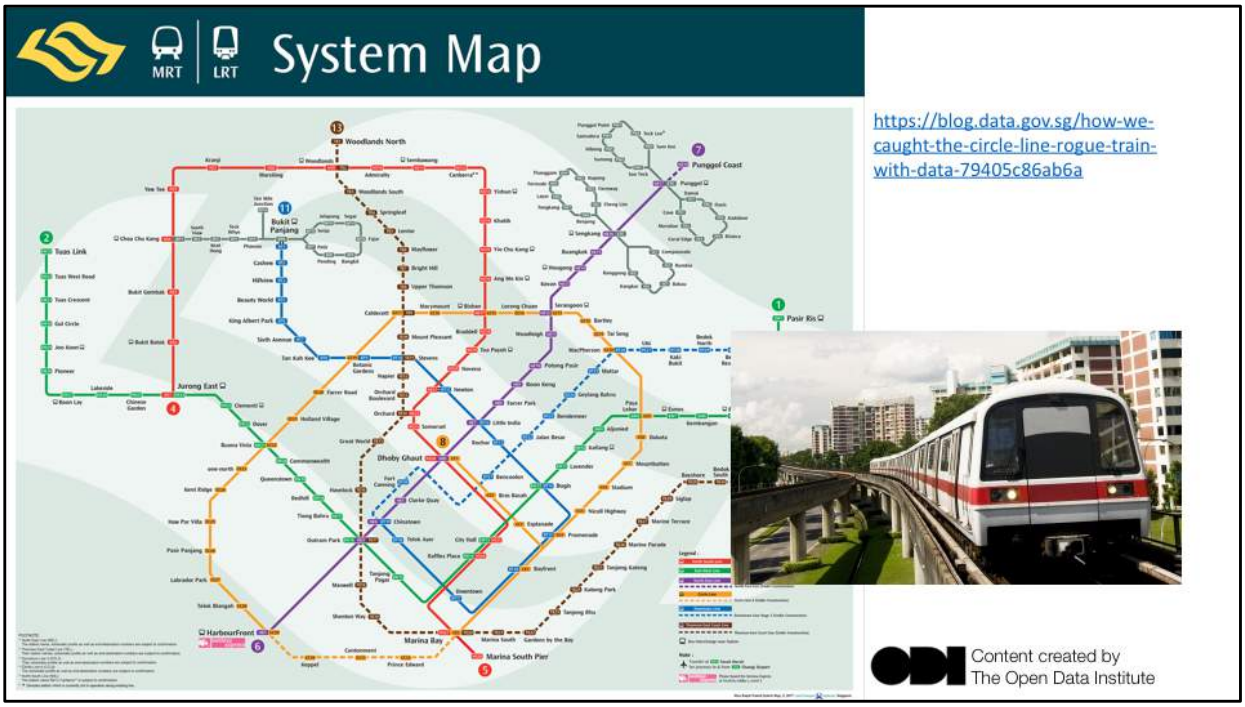
Also know when doors are closing etc.

There are many autonomous train control systems in the world. The most advanced work when the trains talk to each other and know each others positions. In such networks the trains control their entire operation without a driver, including position of doors.

In the UK there is a lot of dispute of the role of guards on trains who have always been responsible for the operation of train doors. However in order for trains to conform to European Train Control System regulations, the train must be able to control the position of its doors automatically. This means trains can run closer together and set their own distances. Such a system means that you can quadruple the numbers of trains within a system.

Additionally such systems save money by removing the need for physical signal systems tied to fixed length sections of track. Instead the length of sections can be dynamic dependent on the need of each train. Thus you can have two passenger trains in a section that is longer currently due to the requirements of freight traffic. This means more capacity without extra track!

However this can still go wrong as Singapore found out.



Singapore's MRT Circle Line was hit by a spate of mysterious disruptions in recent months, causing much confusion and distress to thousands of commuters.

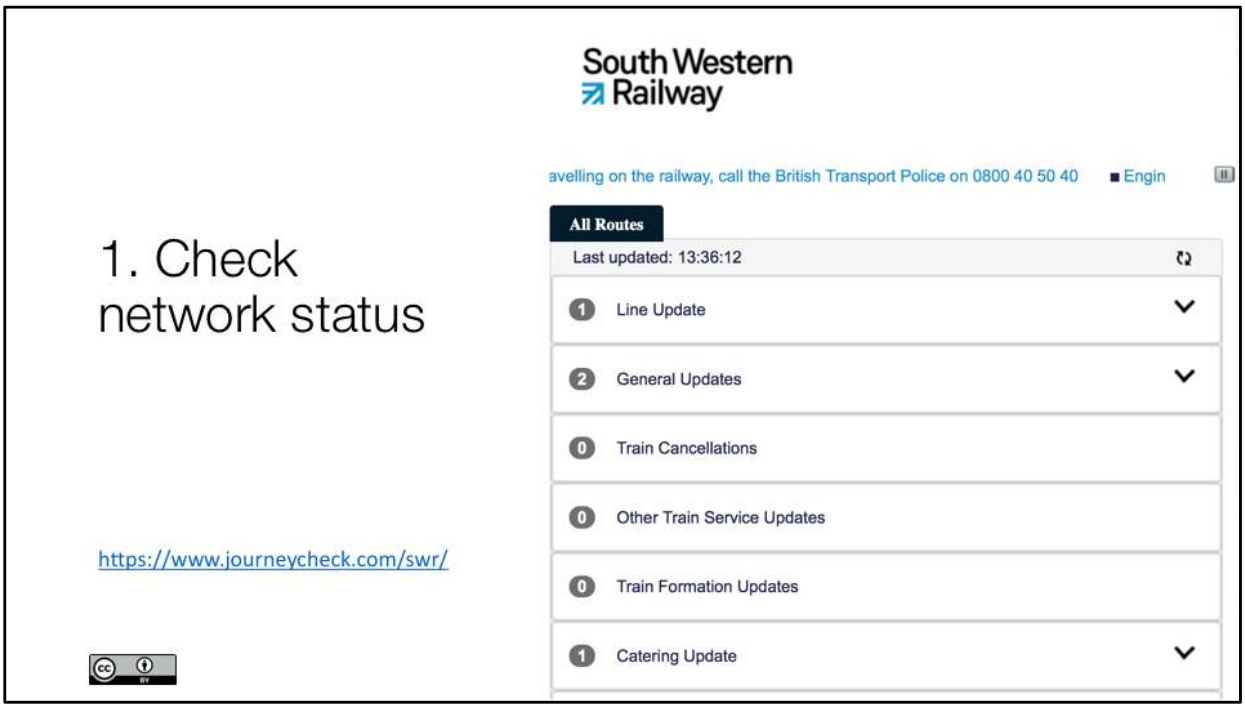
The MRT is a fully automatic train system which is not meant to suffer these kind of disruptions. There are no human drivers, the trains control themselves and all communicate with each other. It is an incredibly safe system, however something was happening that was causing delays, not stoppages on the circle line.

To find out what the culprit was took some decent data analysis.

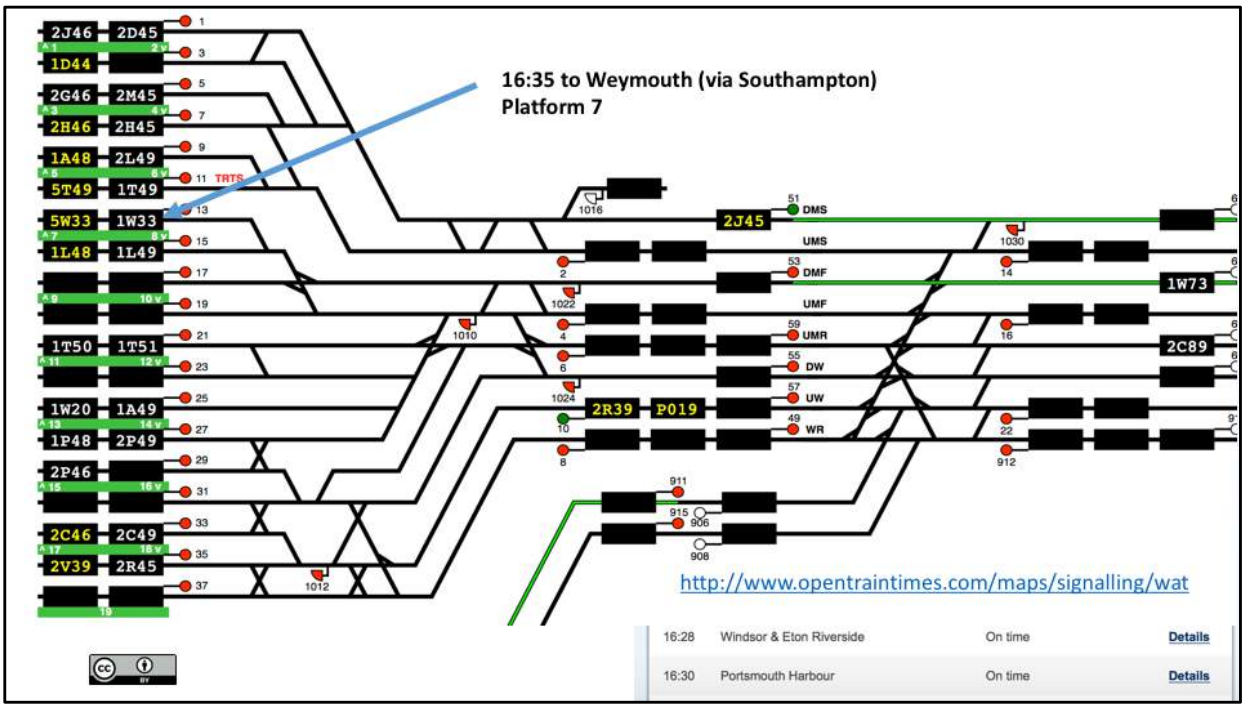
My journey home

John Seb Barber

Interestingly my journey home requires another set of tools, so lets look at these.



As the train starts from Waterloo there is no point checking if it has left. So before I leave work I 1st check the network status as if Waterloo is in a mess I can always go to a different London terminus and take a completely different from Paddington via Reading and Basingstoke. If I need to go to Paddington then a completely different tube journey is required, which is why I check before I leave the office.

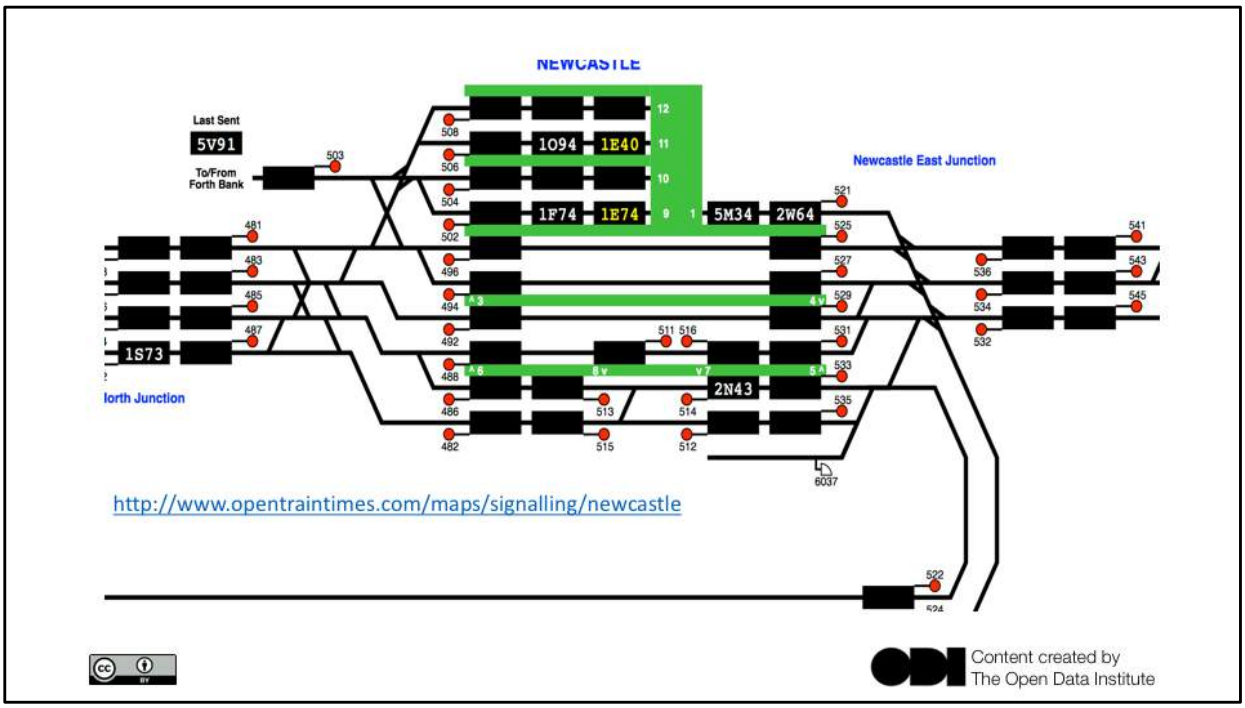


Next I check the platform that my train is on at Waterloo, this means I can head directly to this platform at Waterloo through the tunnels under the platforms rather than having to check platform information signs.

However as you can see, the platforms are only announced 15 minutes before the train leaves and it takes me 20 minutes between the office and waterloo. While I could use RTT I prefer to use the Open Train Times maps as it is more visual. Here we can see a live control room view of Waterloo and my train to Weymouth (via Southampton) is on platform 7. I can easily see trains that go in my direction as the first letter (after the number) will be a B, P or W, representing the destinations of Bournmouth, Poole and Weymouth respectively.

We can also see the train 2R39 outside the station with a green signal approaching platform 19. We can also see 2J45 above it which is deading down the slow line, the green line in front of it representing the routing.

Finally look at platform 8 (above my train) and you can see the train 1T49 which is ready to start (TRTS) so should be receiving a green signal soon.



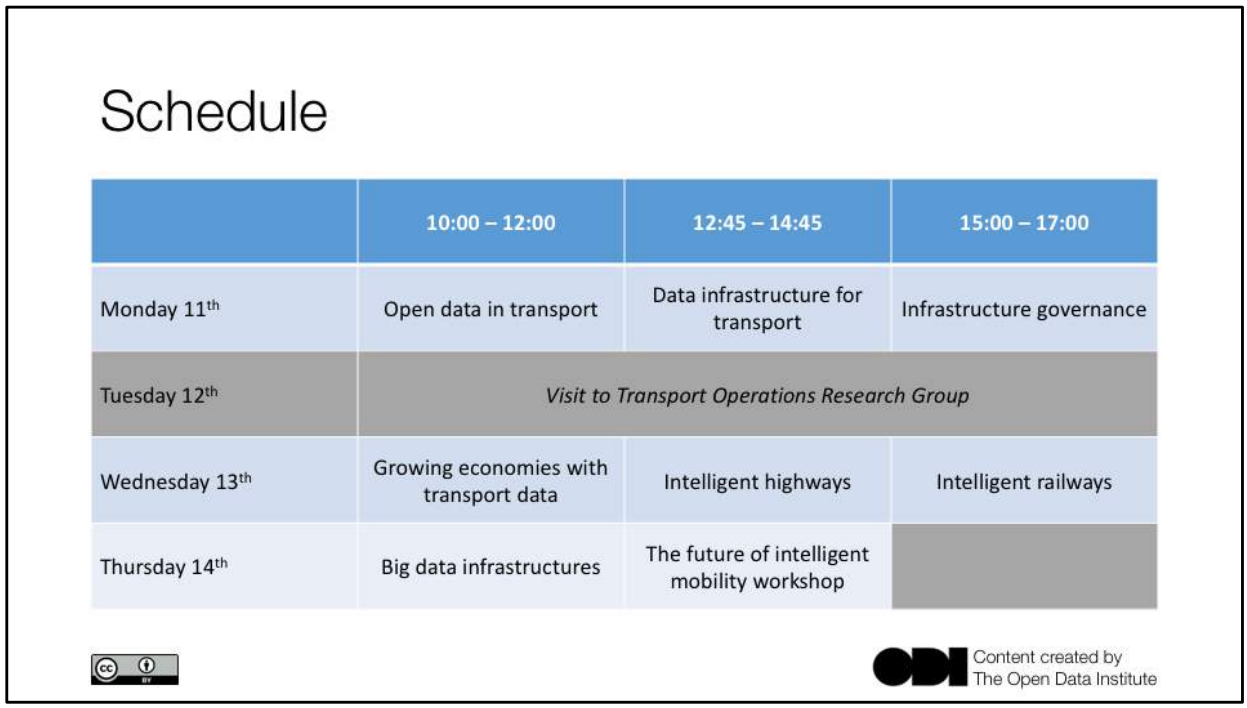
Here is the live view of Newcastle, a through station, we can see the many platforms, including the A and B ends of platforms 3 and 4. Note that 5,6,7 and 8 are not A and B but separate platforms.

For the whole country, the national rail open data represents about 60,000 data points a second including train positions, signal colours and routing data. This is all available for anyone to use and I thought that a nice end for the day might be to take a walk to the station to see this data in action, so please feel free to join me.

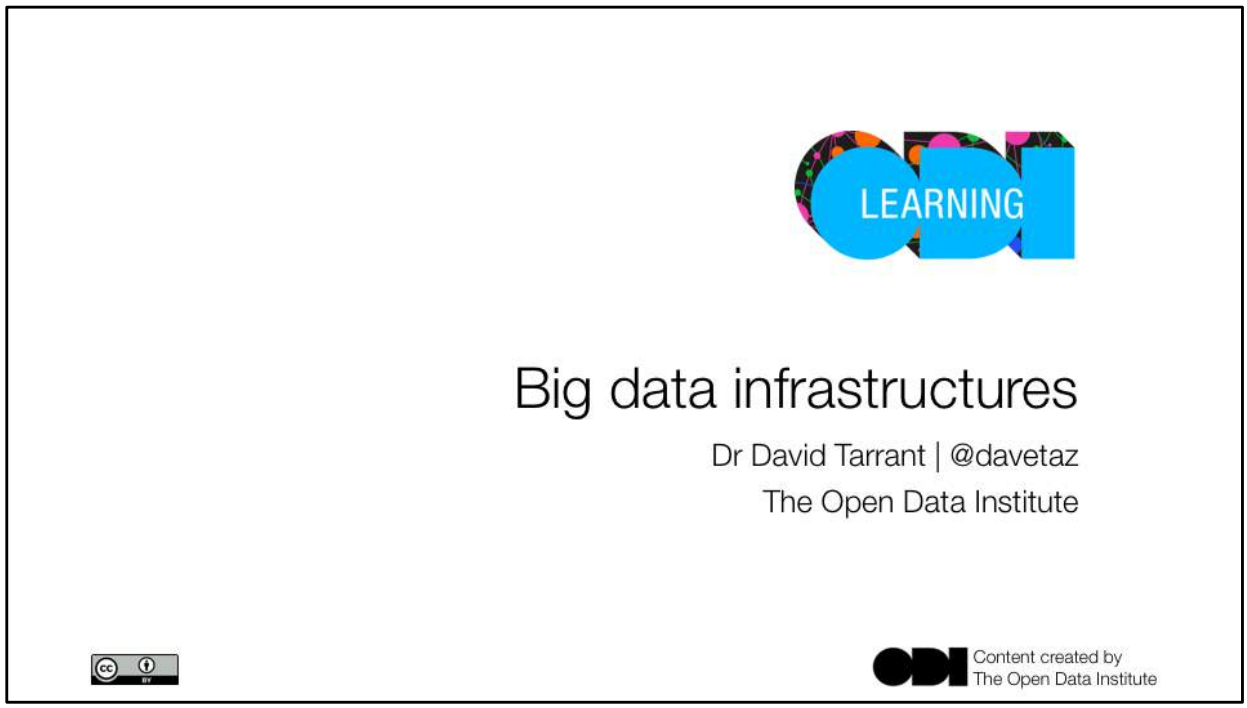
Transport Data Infrastructure

Dr David Tarrant | @davetaz
The Open Data Institute

Content created by
The Open Data Institute



Welcome back. I hope you had a informative time yesterday. Today we are going to dig deeper into how transport data is able to grow economies before looking at some of the innovations an opportunities in the highways and railways sectors. For the last session today we are going to take a short walk to the main station to see the how big data on the railways is delivered to passengers through it's many channels.



And now we get to talk about my favorite area of transport, railways.

Outcomes

What is big data?

Big data in transport

The big data hubris

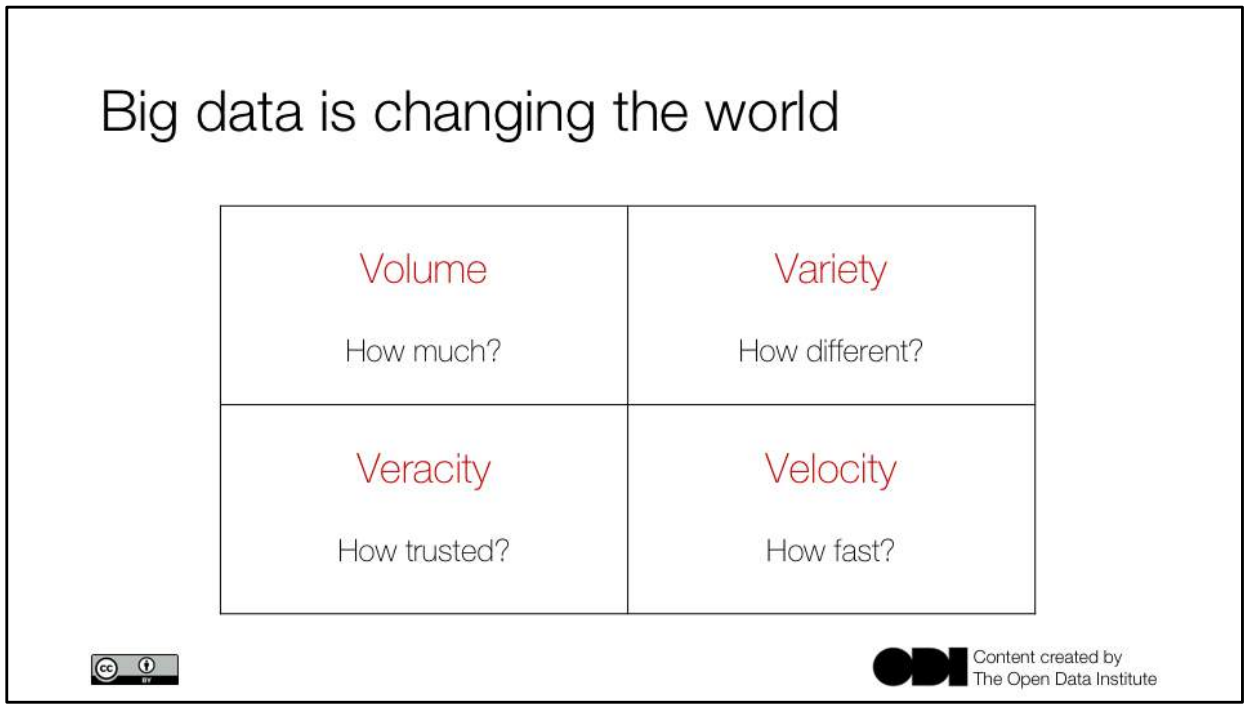
Publishing big data

Content created by
The Open Data Institute

Big data is changing the world

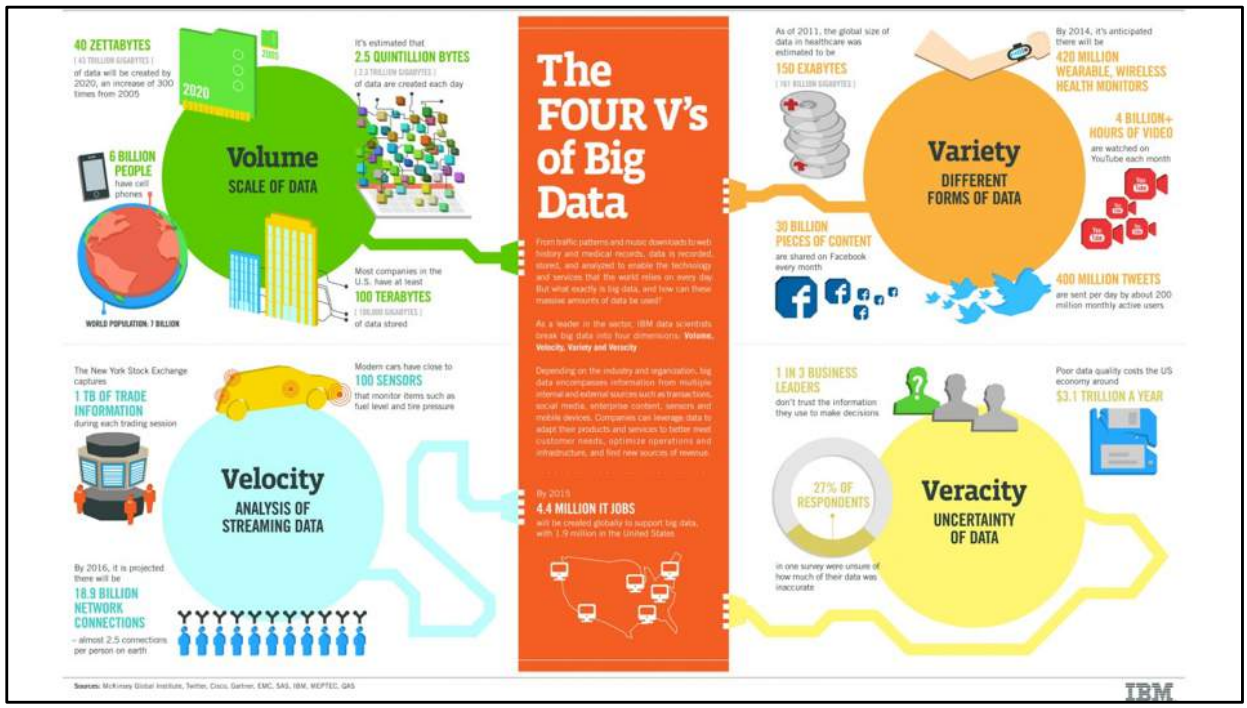
What is big data?

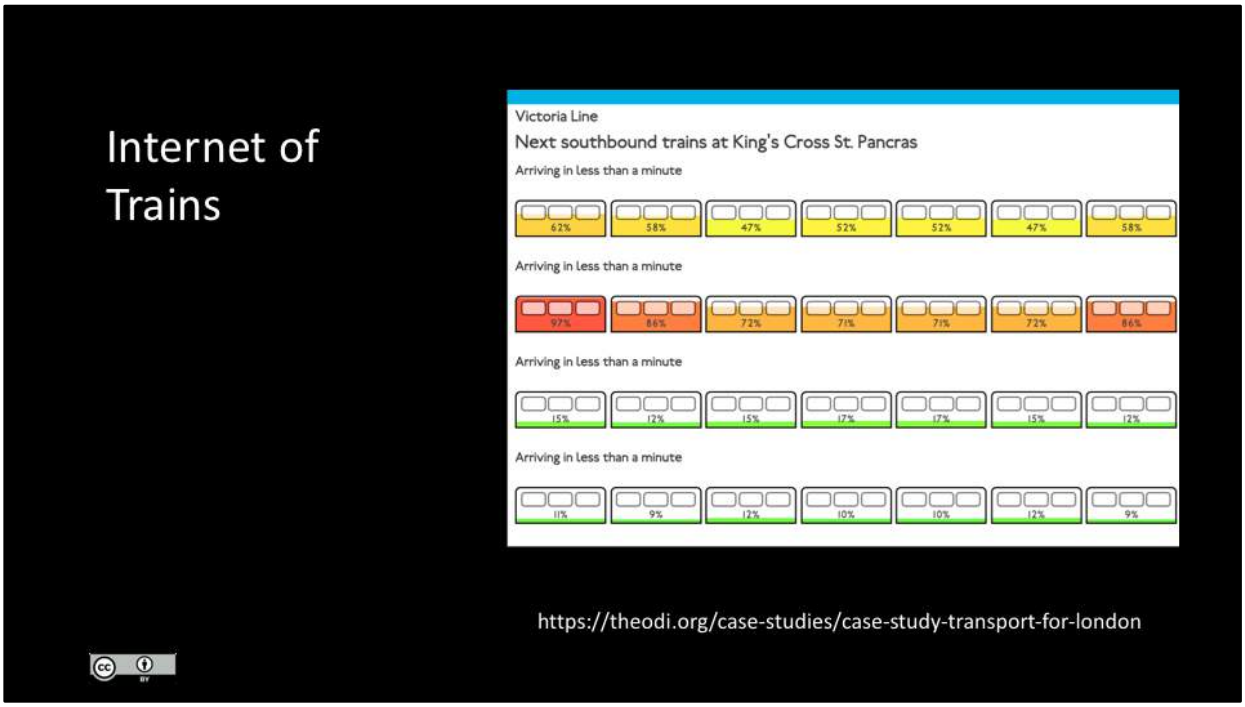
Collect some ideas together on some post-it notes. Do you have any categories emerging?



These are the Vs of big data. Depending on your source there are between 3 and 7 used. I like using 4.

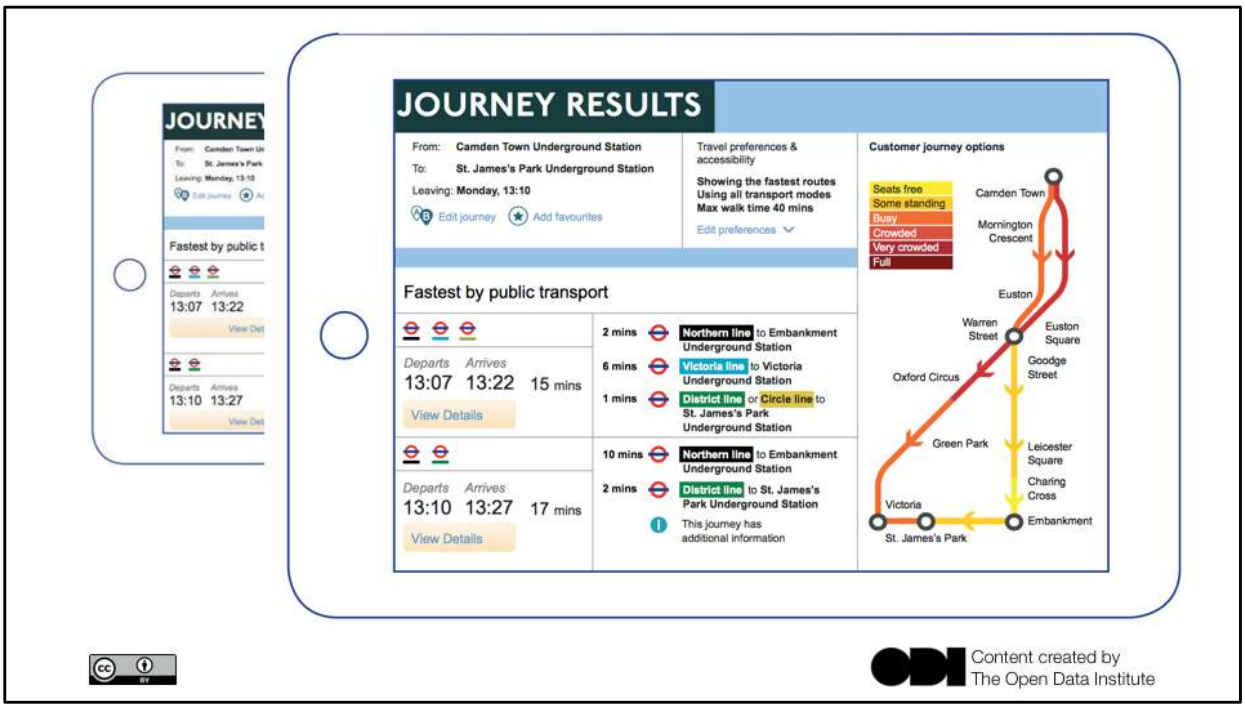
Volume: Representing the large scale of data that we now collect. 90% of the total data has been created in the last few years.
Variety: Data now comes from a huge amount of sources in different formats
Veracity: Data is of varying quality, even from the same source.
Velocity: Data is streamed live at incredible rates making it hard to deal with.


The FOUR V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
Volume
SCALE OF DATA
40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
Variety
DIFFERENT FORMS OF DATA
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
Velocity
ANALYSIS OF STREAMING DATA
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
Veracity
UNCERTAINTY OF DATA
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM



Modern tube trains are packed fully of sensors, which look for faults with the train systems. The ODI worked with transport for London to identify uses for this data beyond train health monitoring.

One of the interesting sensors on a tube train relates to the load on the suspension and axels each end of the coach. This is used to monitor the tracktive force and power required for each coach. We realised that the axel load sensor could be used as a proxy measure for how full a coach was on a train. This allowed us to produce visualisations of the load of each train to help people pick which train to board and where.



TfL have started thinking about how to integrate this into their journey planning services and data to help consumers make a choice both based upon time and comfort. Here we can compare the current interface to the proposed new functionality enabled by this data.

TfL Wi-Fi device tracking

Transport for London

54 stations

6,000,000 devices

42,000,000 journeys

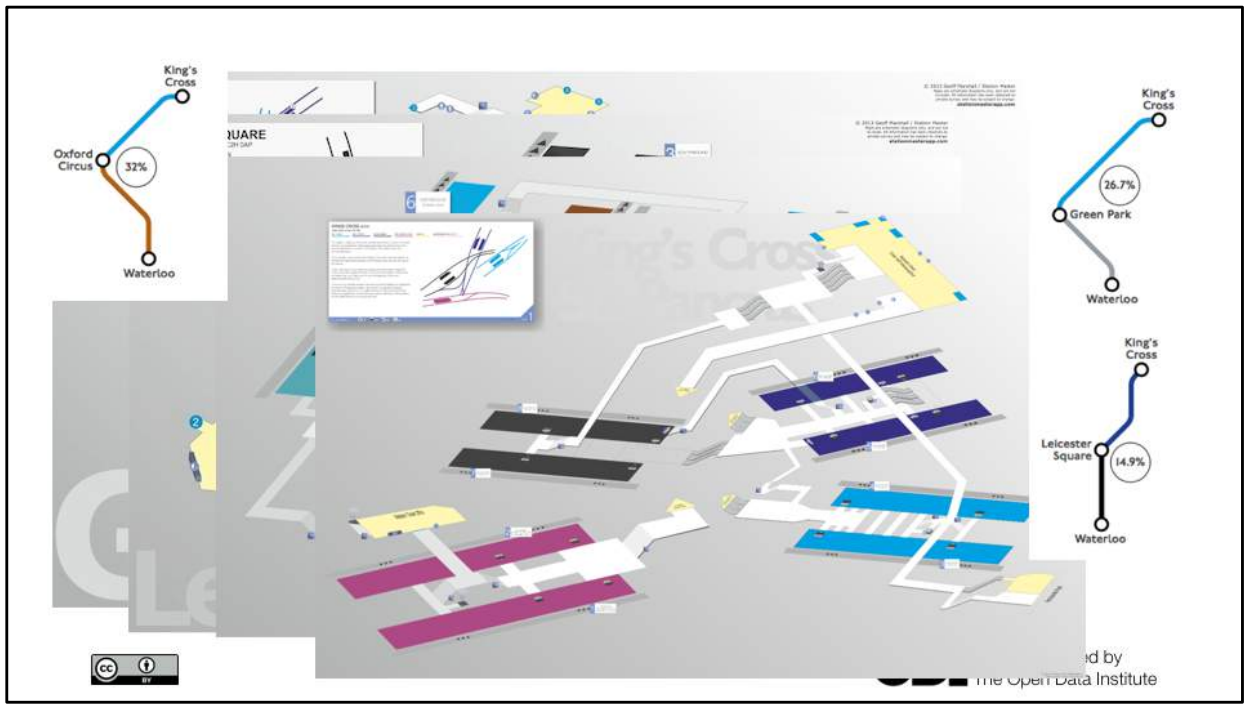
500,000,000 data points

http://content.tfl.gov.uk/review-tfl-wifi-pilot.pdf

Although you are required to tap in and out of the tube to be charged the correct fare, as we saw yesterday and earlier, the choice of routes you can take in vast. The proposal to enhance the route maps should help customer choice. TfL however also wanted to know what routes people chose through not only the network, but also their stations.

To find this out they used Wi-Fi association data from 54 stations in central London in December last year to track nearly six million mobile phones, and gathered over 500 million pieces of data related to 42 million journeys on the mass transit network.

Note that the Wi-Fi is owned and managed by a private company.



This is the route as a transport expert I get asked about the most. Which one do you take?

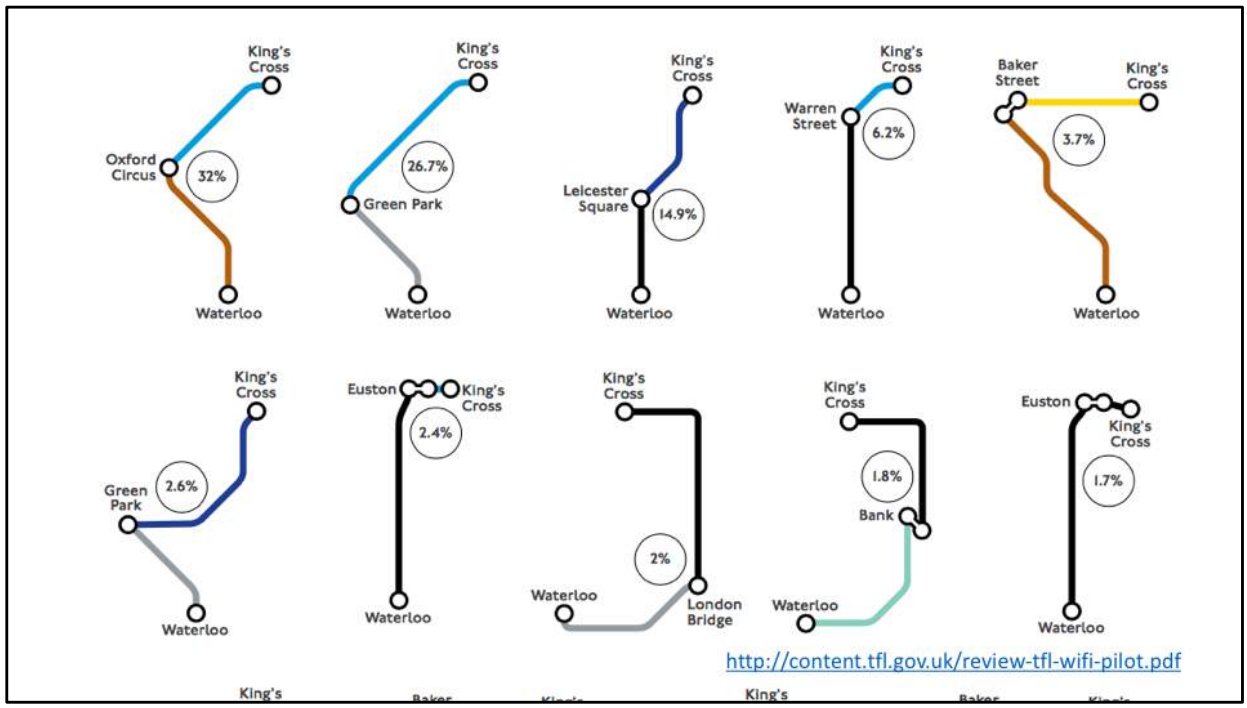
Use the maps I gave out the other day to work out which is the best route between waterloo and kings cross. You can also use your phones to see what they suggest. You have a lot of the data, but do you have the knowledge?

Which routes did you get?

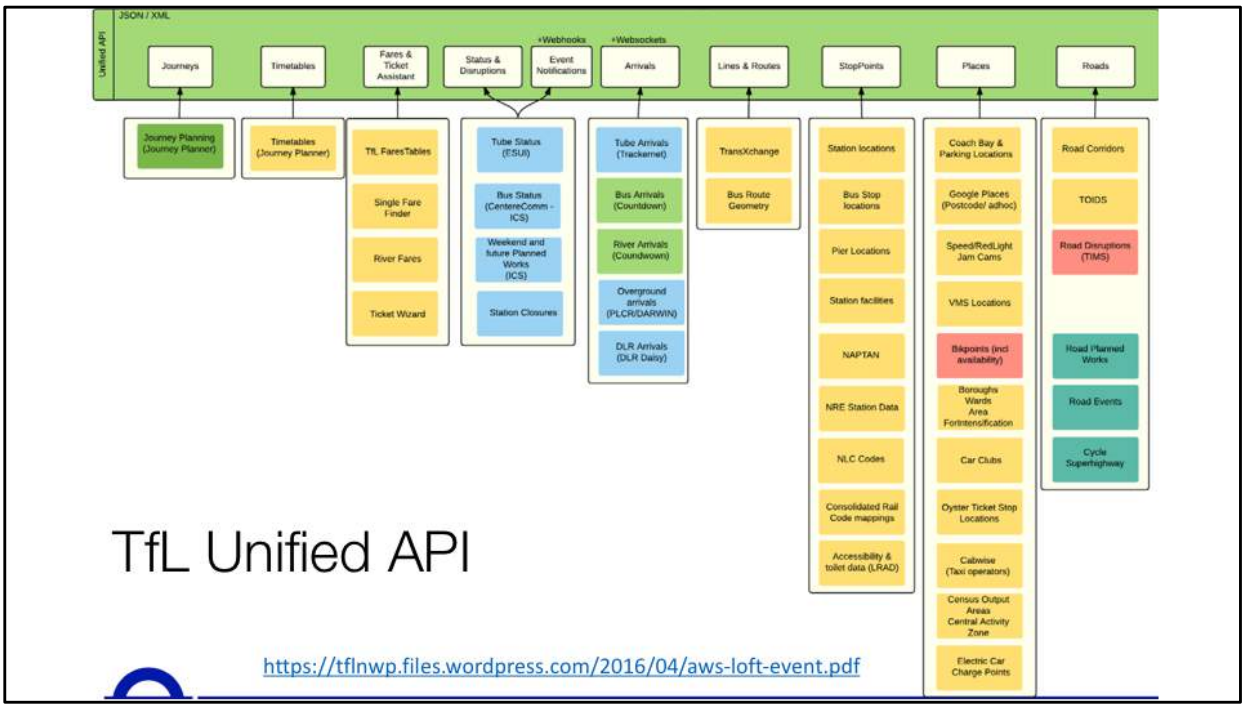
Who picked Waterloo to Kings cross using the Jubilee and Victoria lines changing at Green Park. If you did, you picked the same route as 26.7% of users. Firstly the Jubilee line is the longest walk from the platforms at waterloo, so you have lost time already, lets look at green park…where you need to go up an escalator and over a platform to change lines.

Who picked northern line to Leicester square and then Piccadilly link? If you did, you match 14.9% of users. Northern line is closer to Waterloo platforms than the Jubilee, and you need to make sure you are at the back of the train at Leicester square to change lines. Neither route so far has been good for disabled.

Anyone pick Bakerloo line to Oxford circus then Victoria line? 32% of commuters take this route even though google and apple don’t show it! Citymapper does however, why. Well look at oxford circus, it is the easiest and quickest change, a simple walk to the next platform opposite. If you do arrive at Kings cross on the Victoria line however you need to be at the back of the train and take this exit otherwise it is a very long walk to Kings Cross station! Same in reverse!



Here is the data collected from the Wi-Fi trail for this route. One route that is not tracked is taking the tube to Euston and walking to Kings Cross along the road for 2/3rds of a mile. But 1.7% of people do take the tube to Euston and then walk 1/3rd mile to the northern line bank branch and get this to kings cross, rather than get the Victoria line which is a shorter walk.



Due to the complex organisational structure and diverse use of technologies across TfL's technology estate, the previous Open data provision spanned a large spectrum of quality, accuracy and data formats making it complicated for application developers to be multi-transport-mode applications. TfL's new unified API aims to make accessing the key public information across all modes of transport simpler.

Its aims are:

- Unification of the the data for modes of transport into a common format and structure (common (canonical) data model)
- Live & Web scale – The Unified API is designed for applications to use in real-time and at high volume
- Low latency
- Support common web and data formats – The Unified API supports output in both XML and JSON format
- Supportive of future change whilst minimising end-user (developer) impact.
- Metered and managed

The diagram above shows the simple version of the data model for the TfL unified

API.

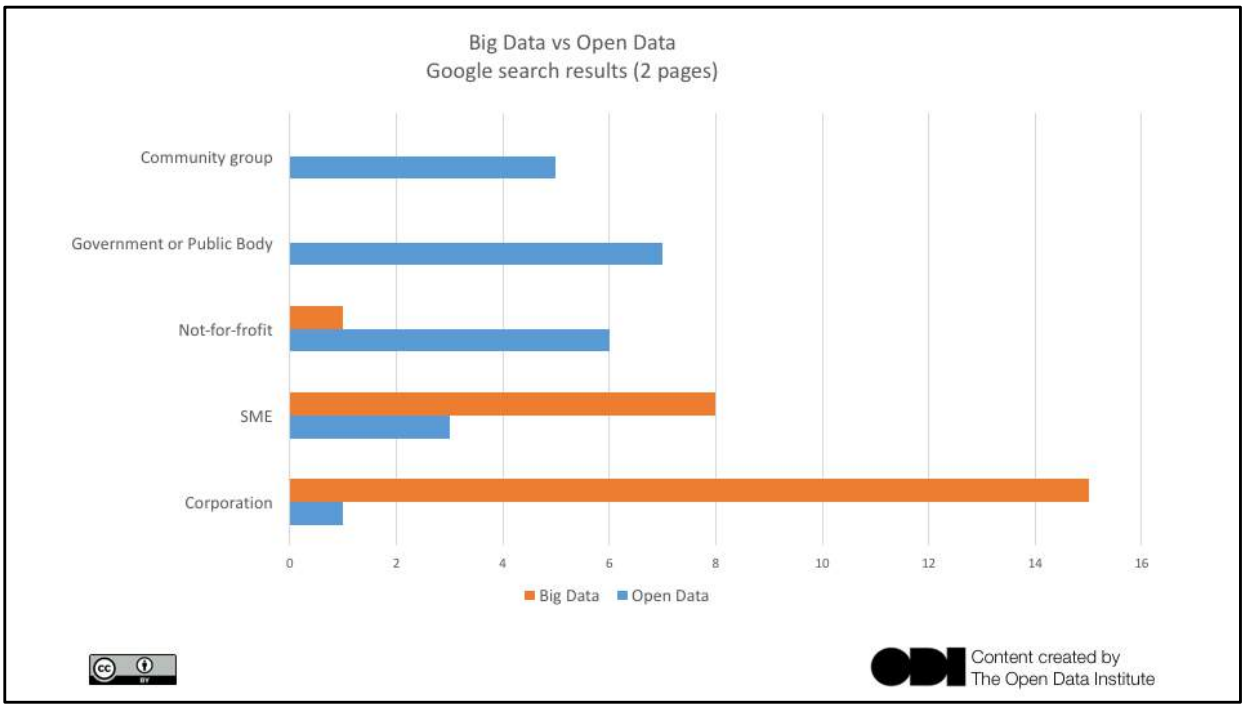
TfL makes use of Amazon Web Services to scale up the provision of their data to users. So much so that earlier in the year the partnership saw Amazon rebrand the whole of Westminster tube station to Webminster.

All this shows how trains are all part of the internet of things, along with the people who work with them.

The last example shows what can result when IOT data and data literate working come together.

Rail technology is booming worldwide and being applied.

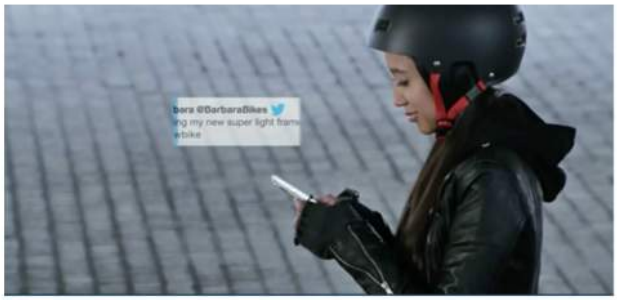
The rail network is also benefiting from the internet of things as TFL have once again led the world in.



However there is still a lot of industry hype surrounding big data, a term which still does not have a definition, but rather a description.

Looking at occurrences of open data and big data in the first two pages of a google search it is possible to observe that open data is talked about by Governments, communities and not for profit organisations, linking it to the social good. Conversely big data is overwhelmingly touted by large corporations wanting to sell analytics and infrastructure for data that might not be that useful.

Big data: It's about you?



Content created by
The Open Data Institute

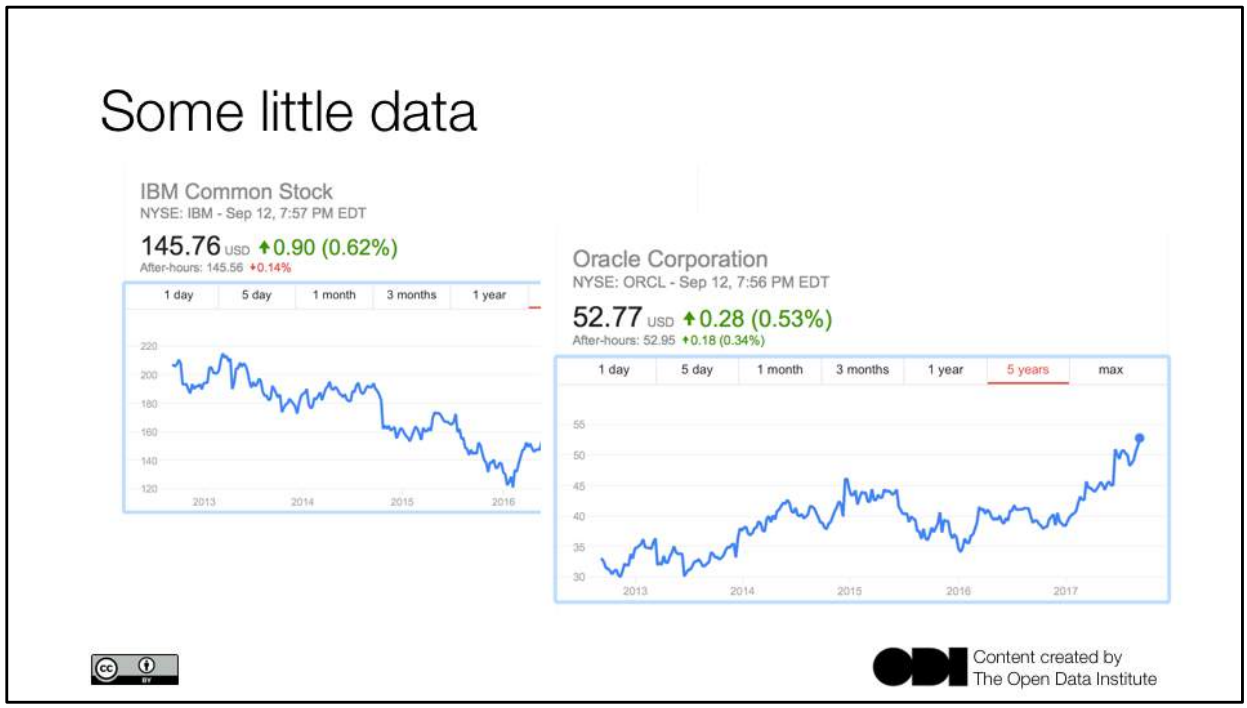
One of those companies is IBM. Take a look at one of their recent advert about big data.

IBM advert transcript:

"Barbara just bought a bike.
She wrote a tweet about it.
You can't learn much from that.
But take data from millions of tweets, combine that with your companies supply chain and sales data. Apply IBM analytics and expertise and all of a sudden you can learn which bikes to build, what to make them from, where to sell them.
Because Barbara and the world just told you.
There's a new way to work and it's made with IBM."

TAKE data from millions of tweets, millions of conversations. This is stuff of spy agencies and snoopers. Barbara didn't tell you. She told her followers. This intrusion is not what big data should be about. Next we'll be recording conversations which are heard by our mobiles with their microphones and using these as big data. Scary IBM.

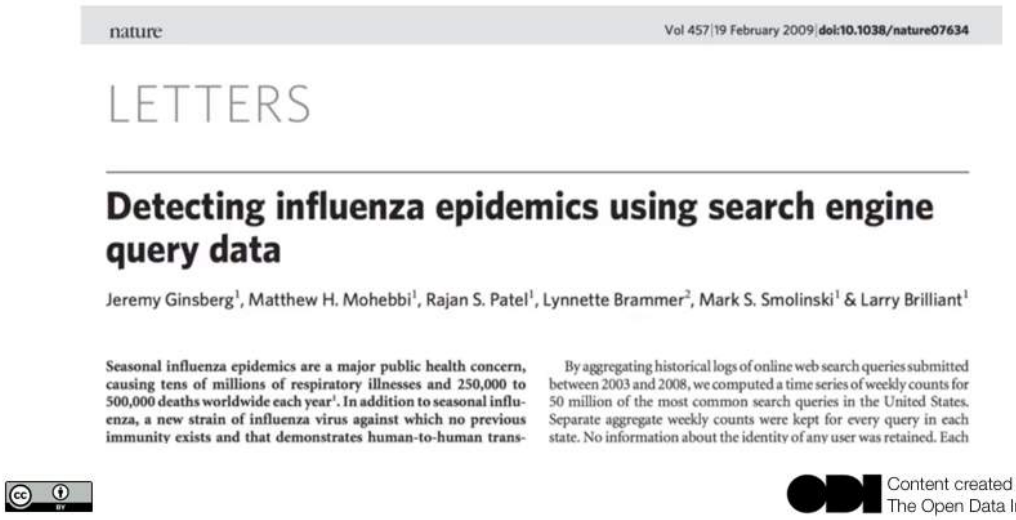
Oracle are the other big corporation who talk about big data, but they don’t touch on the type of data in their adverts, just the infrastructure they can sell you to manage it and the software Hadoop in Java which can process it.



Fortunately we can easily examine the differences in these approaches by simply looking at a 5-year share index of these two companies. While IBM has been struggling recently to remain a closed, patent trading company. Oracle appear to be doing well.

So why am I talking about the risks of big data? Is it to do with big data washing?

Big data is changing the world



nature Vol 457|19 February 2009|doi:10.1038/nature07634

LETTERS

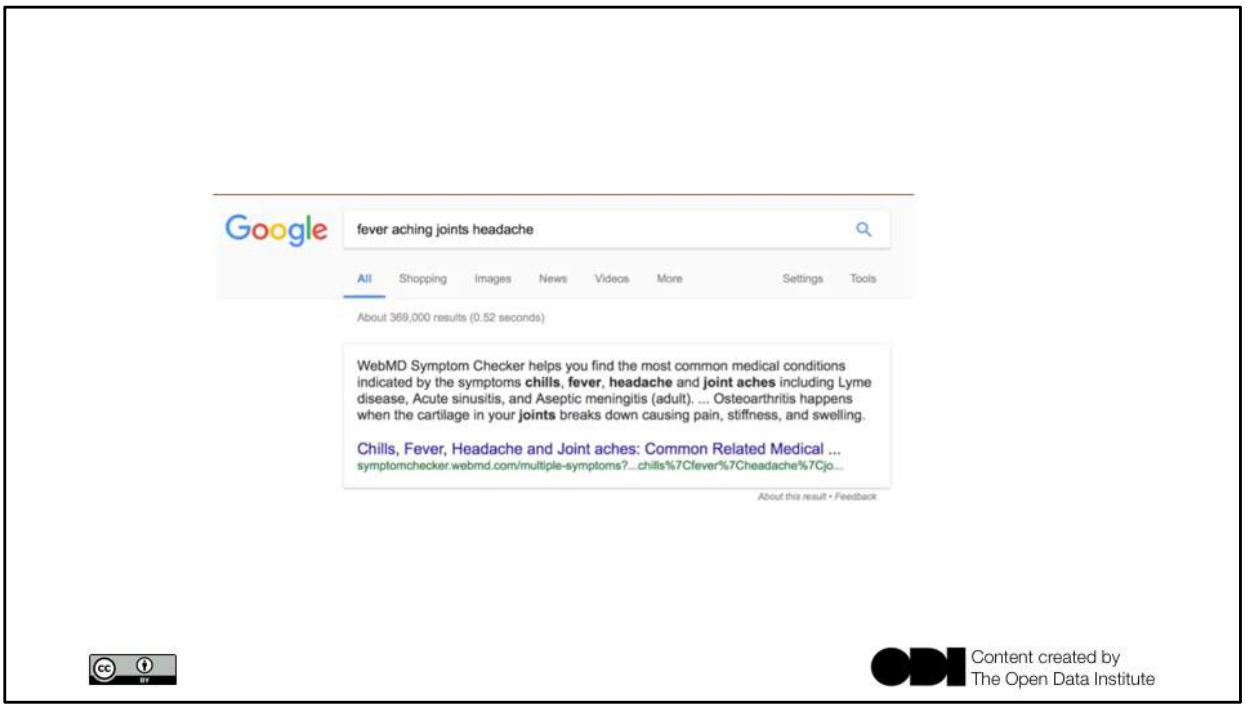
Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year[1]. In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists and that demonstrates human-to-human trans-

By aggregating historical logs of online web search queries submitted between 2003 and 2008, we computed a time series of weekly counts for 50 million of the most common search queries in the United States. Separate aggregate weekly counts were kept for every query in each state. No information about the identity of any user was retained. Each

Big data is changing the world, and the hype has been around for almost a decade now. The hype was clear but big data needed a poster child, something to shout about. And in February 2009 it got it. Google, along with some big data experts published a breakthrough lead article in Nature, a leading journal that stated they could use Google search data to predict flu outbreaks. This would allow the Centre for Disease Control to distribute vaccines more efficiently to prevent outbreaks reaching large numbers and spreading. The initial Google paper stated that the Google Flu Trends predictions were 97% accurate comparing with CDC data.

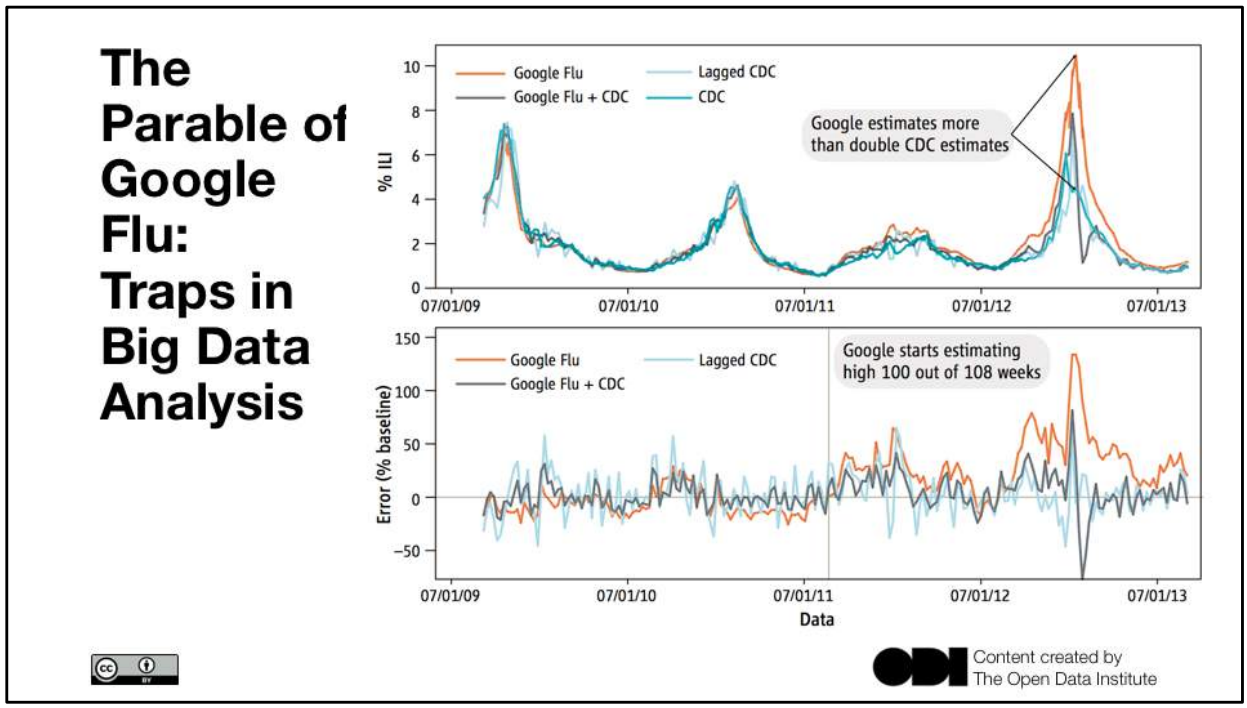


So here is how it works. The system collections search term data related to defined terms and then uses these to map occurrences of these terms across the country. Historical data was then trained against data from the CDC using ‘advanced’ machine learning algorithms to work out the potential correlation.

They even put a web service online to show the live data and where the out brakes are occurring.

Google Flu Trends is an example of collective intelligence that can be used to identify trends and calculate predictions. The data amassed by search engines is significantly insightful because the search queries represent people's unfiltered wants and needs. "This seems like a really clever way of using data that is created unintentionally by the users of Google to see patterns in the world that would otherwise be invisible," said Thomas W. Malone, a professor at the Sloan School of Management at MIT. "I think we are just scratching the surface of what's possible with collective intelligence."

Now I prefer to use the term 'collective ignorance'. We cannot possibly know about everything, if you could diagnose flu yourself then why would you search for it? What people might search for might be the wrong thing and not flu at all. In fact what happened was that the numbers of people visiting doctors actually rose. The whole thing was a disaster for big data… why?

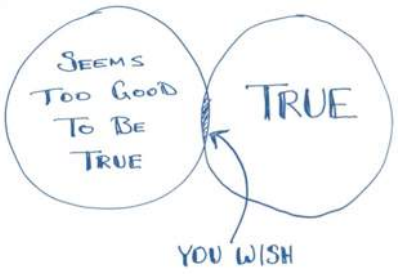


One of the main reasons it failed is down to simple statistics, which tells you that correlation doesn't always equal causation. Take a look at the Orange line and the light blue line in the charts here. Orange is the google flu prediction while the light blue is the actual. They broadly follow the same pattern, however in 2011 the orange line significantly overestimates the actual data, in fact over the two year period it only underestimates flu for 8 weeks out of the 108. This over estimation means more money spent on flu-prevention than less as it thinks there are more cases than there is.

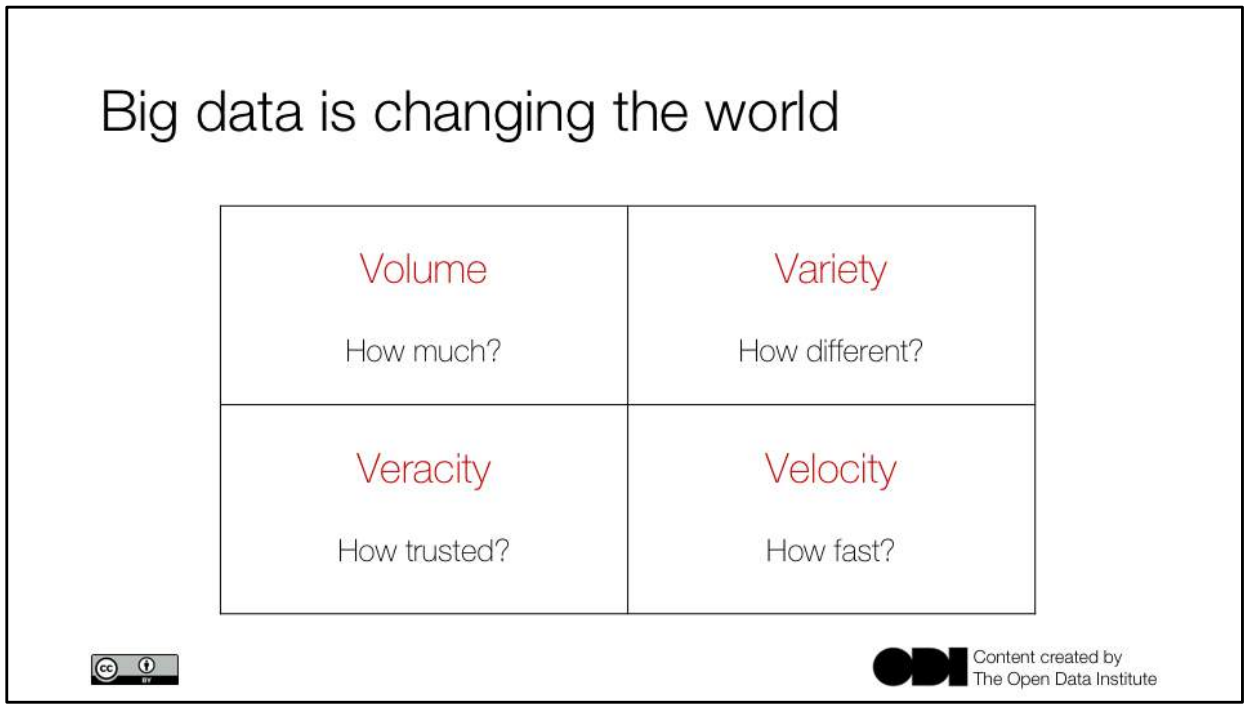
It gets worse however, remember I mentioned that correlation doesn't always equal causation. The group behind this work looked at other seasonal variations and found that temperature (not big data) was a better indicator. And while 45 search terms such as flu and influenza appeared relevant you could have also used the term college basketball to make the same prediction. Flu is seasonal and outbreaks occur at the start of terms when kids go back to school. You don't need big data to tell you that!

Their main recommendation was to complement rather and supplement. The Big Data Hubris is that we only need the data, we don't need experts or other traditional methods. Big data will be useful if used to complement the methods of experts who

are already working in the area.



The main message here is that if some big data analysis seems too good to be true, it probably is. Experts and analysts with years of experience in a domain are likely to know what opportunities there are. Ask them what data would help improve their work rather than just switching to big data analysis.



One tip, if your data has more than two of these characteristics then the analysis is going to be almost impossible as you have too many problems to solve.

Overfitting

Plotted here are 10 points in a series that have been generated using a function.

If another 10 points were plotted of the same function which point (A,B,C or D) would they tend towards. Add a line of best fit that goes through one of these points.

One of the main problems that faced the Google Flu trends work was that of overfitting.

To demonstrate what overfitting is complete the exercise here.

Average

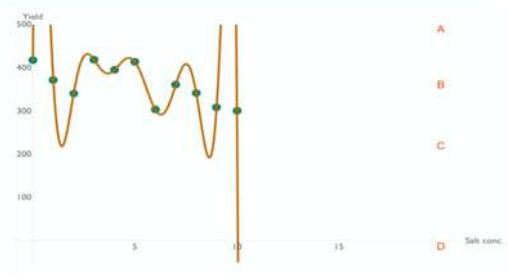
Linear

Quadratic

Quartic

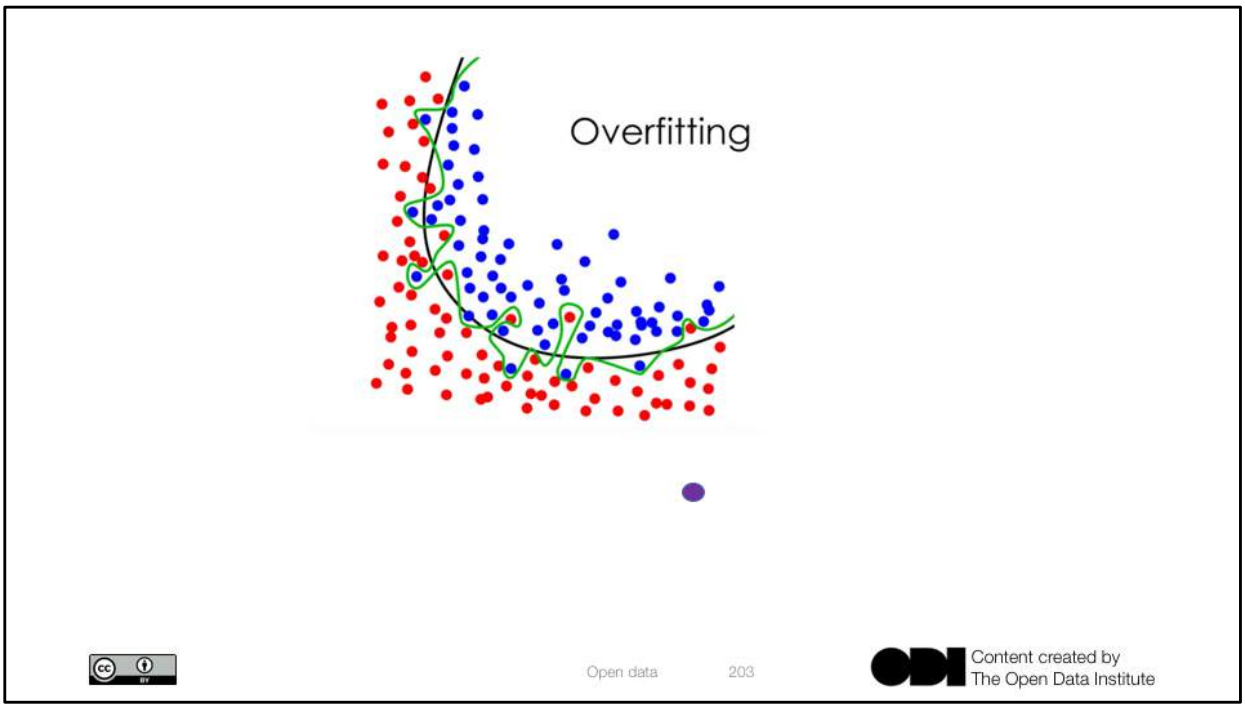
If you take a simple average of the points you end up at B.
A linear line of best fit tends to C.
A quadratic curve of best fit tends to D while a quartic (power of 4) sees a large over-estimation on the last point.

Which one do you think they used in Google Flu trends?



Polynomial

None of them. Google appeared to use a polynomial line of best fit. This is an equation that goes through every point, very complex. However as we can see it has no idea what happens after the last point so cannot possibly predict.

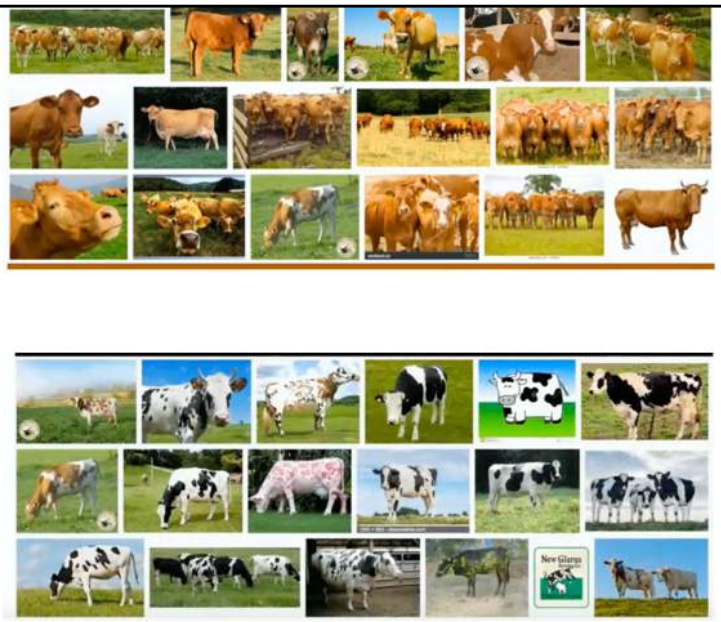


Here is another example of overfitting.

There blue and red dots represent different things and the idea is to be able to predict the colour of a new dot if the colour is not already known.

The black line is a quadratic line of best fit, while the green is the perfect fit.

Given this is the new purple one likely to be blue or red?



How does this apply to transport?

Autonomous cars is the answer. For a car to be fully autonomous it is going to need to recognise more than just other cars, so here is a simple example using cows.

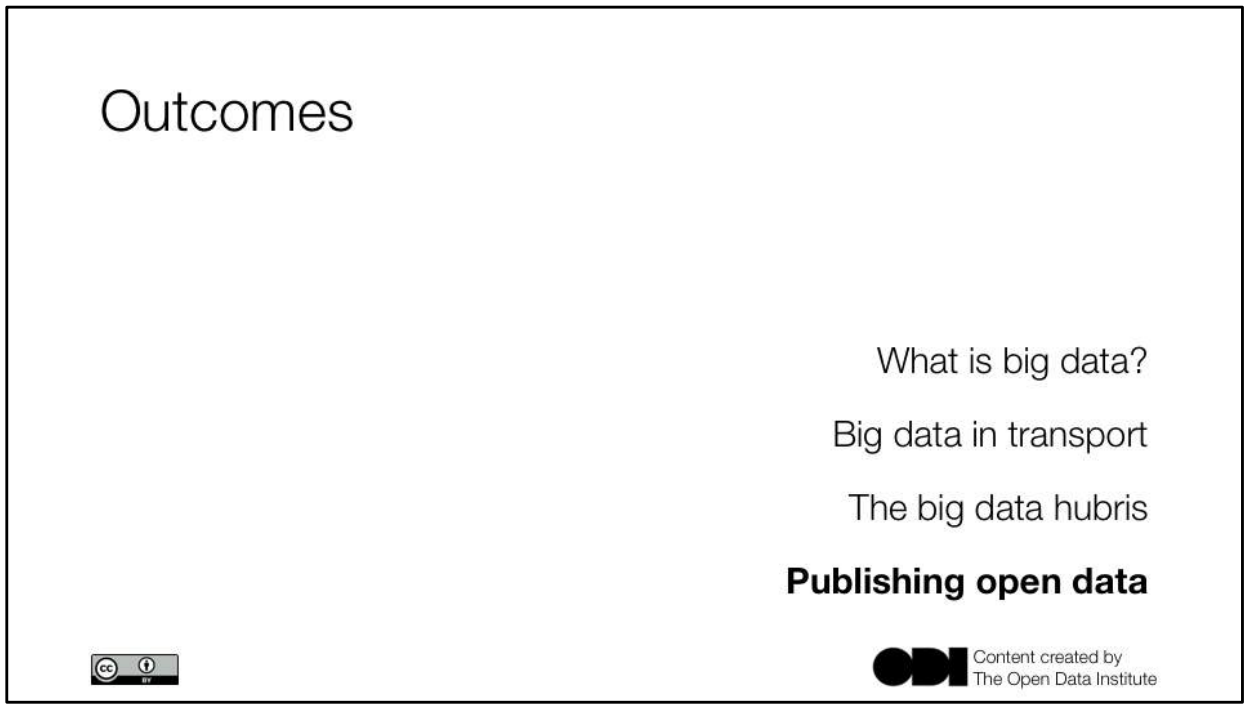
Above the brown line at the top are a set of images of cows used to train a machine learning algorithm what a cow is.

The algorithm has been trained using brown cows, so will it recognise any black and white cows, probably not. The is an obvious example but proves that it is not easy. Apple's new iPhone X has a FaceID recognition system and to train that they used 1 billion faces and even trained it to not recognise professional grade Hollywood face masks, that's a big set of training data! As a result they recon that the chances of someone other than you being able to unlock your phone with their face is 1:1,000,000, unless that person is your twin…

Why did the flu trends example fail?

- Big Data Hubris (complement not supplement)
- Huge veracity problem!
- Used nth degree polynomial (overfitting)
- Solely relied on this data over scientific method

For more on big data search YouTube for "calling bullshit on big data" Excellent lecture series from the University of Washington



To finish the course I've added a couple of extra slides on publishing open data with a little history of the web.

History of web (1)

1994

1995

The web started in 1994 with portals that contained links to other pages and sites that you would click on until you got lost in hyperspace. At which point you would either close the browser and start again, hang up and redial or press the home button if you knew it existing.

A year or so later people started discovering the address bar and worked out there were far more sites on the web than those linked by your providers home page!

History of web (2)

1997

1998

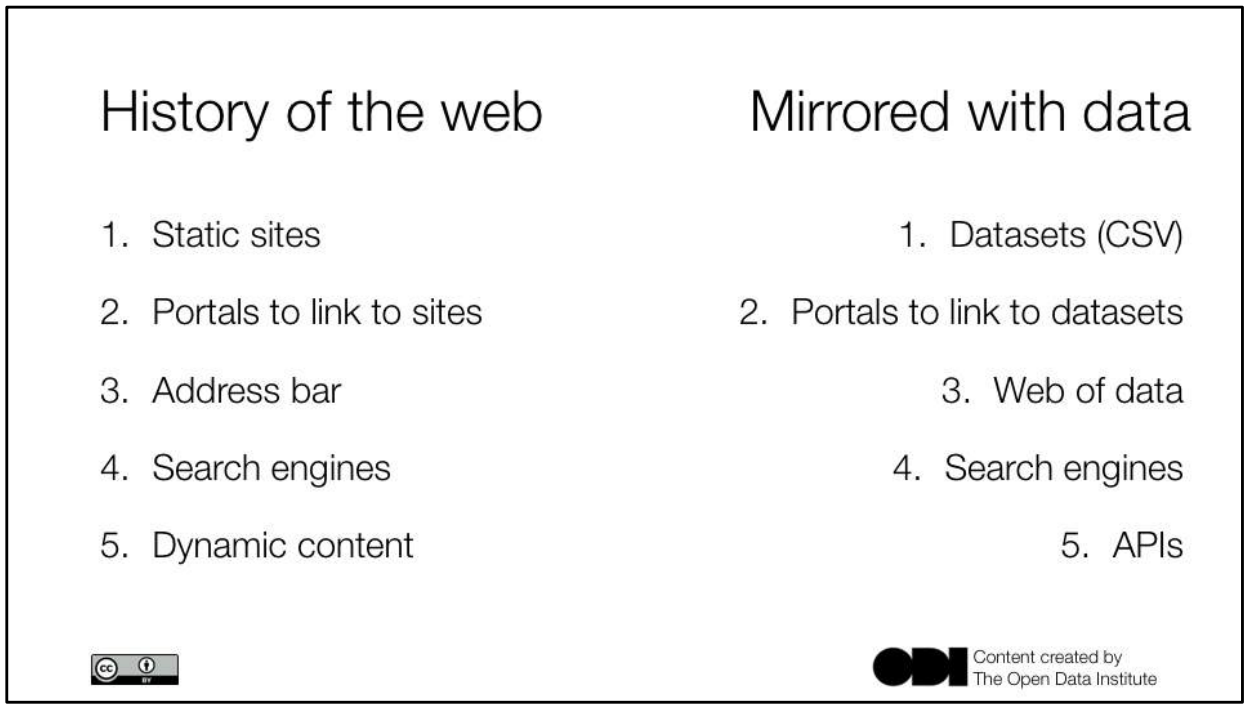
Another year on and the first search engine started to appear, soon followed in 1998 by Google. Note that if you search Google for ”Google in 1998” you get this page which is how Google actually looked. It is so retro! But Google revolutionised the way we find all content.

History of the web (3)

2004

2005

The dotcom boom and bust meant that innovation waned for a few years, there were improvements in audio streaming, but the next key technology came with the launch of Gmail in the browser. Dynamic and custom content was possible. This was followed a year later by Google Maps which suddenly demonstrated the web's capability when it came to data.

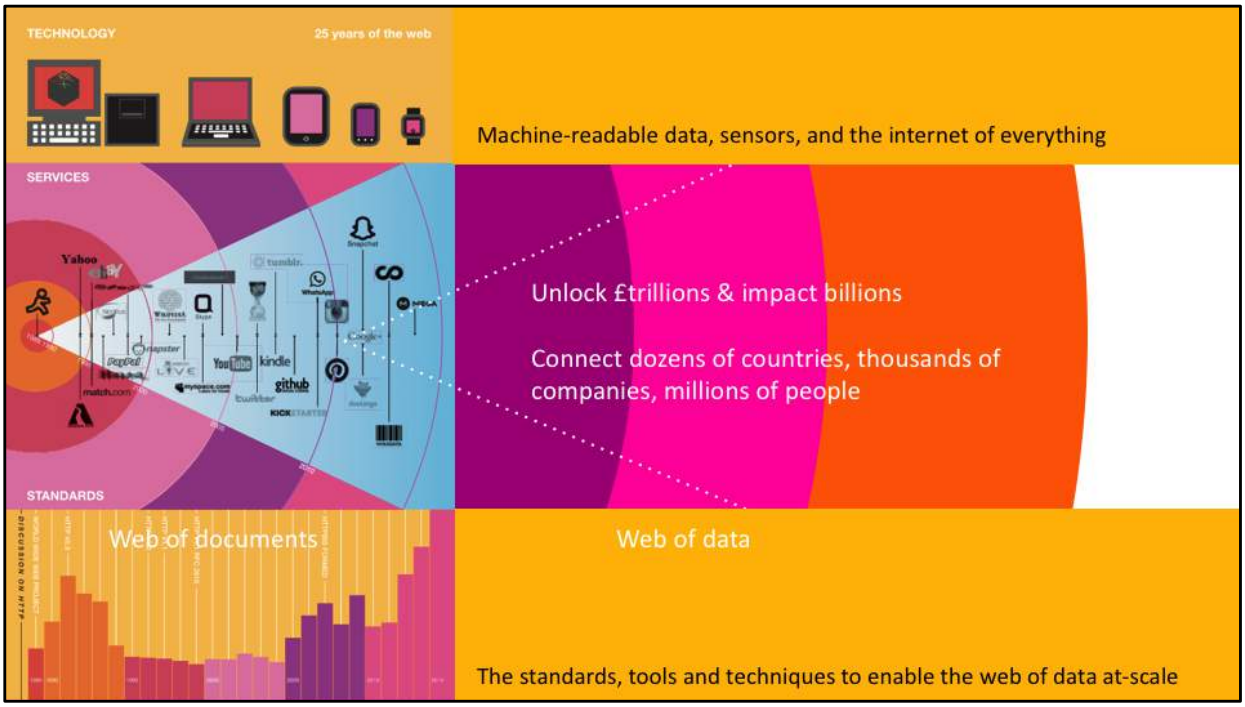


So there we have it, a very brief and incomplete history of the web but it does cover five key aspects which are mirrored in data.

I get asked all the time if we have a portal that lists all the datasets out there and I constantly think about the web in 1994, do we want that?

Saying that, we do. We have static datasets on website you can download, just like early websites. We have portals that link to these datasets. We have the address bar which can be used to address data, but people really haven't discovered that functionality yet! We have the same search engines and it is amazing what data you can find if you put the word data in your query and publishers have optimised their sites to expose data. Finally we have APIs which already power the web's dynamic content anyway. We are just exposing these for others to use openly.

We still have a long way to go but the technology has been there years. The use of the technology to build a stable web of data has not yet emerged.



The web of data is coming, what exactly it is going to look like is still unclear, but I'm excited.

Thank-you!

Dr David Tarrant | @davetaz
The Open Data Institute

Dr David Tarrant

Learning lead

The Open Data Institute

davetaz@theodi.org