

Telling Data Stories

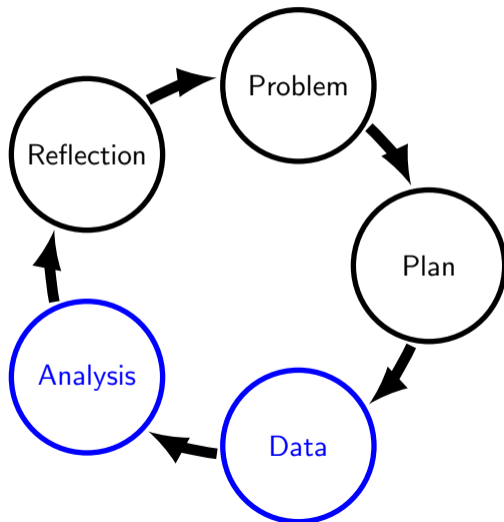
Paul Hewson
texhewson@gmail.com

15th November 2016

Personal transport dis-benefits: injury, lack of exercise, **pollution** . . .



Statistical problem solving cycle



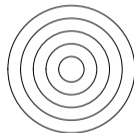
Texas Sharpshooter Fallacy: 1 - Shoot



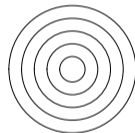
Texas Sharpshooter Fallacy: 2 - Paint Target



Correct order: 1 paint target



Correct order: 2 shoot



Reproducibility “crisis”

Nature asked 1,576 scientists for their thoughts on reproducibility. Most agree that there's a 'crisis' and over 70% said they'd tried and failed to reproduce another group's experiments.

See:

- Nature
- Scientific American report of same

So, using the Problem Solving Cycle

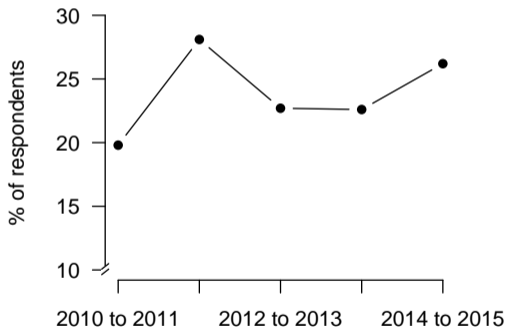
- Problem: Where do we spend £x on cycling promotion
- Plan: What are we trying to achieve, how do we measure it, what existing data can we use
- Data: Are we using all the available secondary data for full understanding, what is the most cost effective way of supplementing this with primary data
- Analysis: How does modern data science let us combine sources that have different information
- Reflections: Assume we have a best choice, how do we measure impact of spend

Where are people cycling

- When do people cycle, **who are they**, how far do they go ([National Travel Survey](#))
- Where (in which LA) do people cycle, **who are they** (Taking Part Survey / Active People Survey)
- Where do we need most improvements in cardio-vascular health (Health Survey for England)
- Where are people talking about cycling (Twitter)

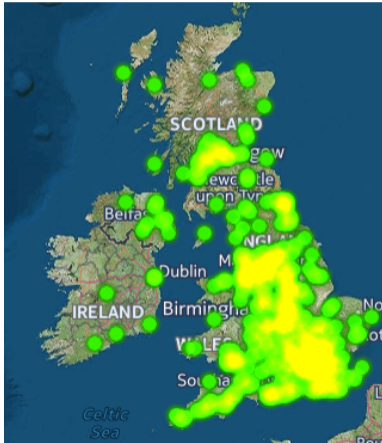
Active Persons Survey: sample headline result

**Active People Survey:
Cycling once per month**



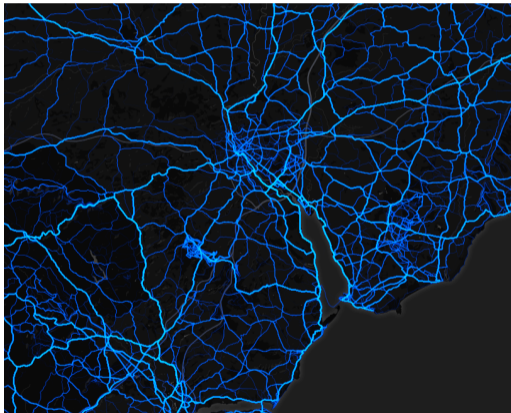
- A lot of people **know** how to cycle
- Why aren't they cycling to work?
- APS/TPS tells us more about commuters / leisure cyclists (socio-demographic features), also tells us a little about geography
- NTS tells us more about trip details (times, purposes)

Social media: Twitter



- People tweeting about Cycle to work day
- Geographical patterns of interest in commuting
- Triangulates with information on utility cycling in official surveys

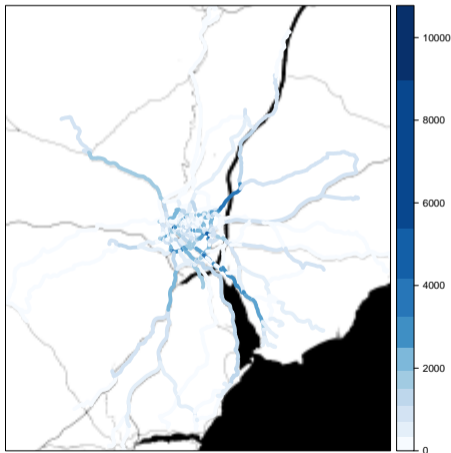
Social media: GPS logs



- Social media (Strava) tells us where cyclists are travelling
- It doesn't tell us why
- It tells us how fast and all kinds of other information
- Contrast this map with demand lines

Census Desire Lines

Google maps predicted routes
from 2011 census



- This tells us **possible** commuter routes
- Why are the short journeys being made by car (use PPDAC cycle - do demographics of APS on cyclists versus census demographics of areas being connected tell us)
- Repeat for cyclists (not shown) - are cyclists actually using their optimal route, if not why not?

Modern Data Science

- Lets us combine complementary data sources to better understand what is happening on the roads, who is doing it and why
- Helps us understand how to better manage the transport system
- Still need to integrate more data to have the information platform for a truly smart city

Harnessing our data to gather insights

Why the Exeter focus?

An increasing population and an expanding travel to work region is attracting more commuters from across Devon, and putting a significant strain on the Exeter regions roads, energy resources and the well-being of its citizens. [Exeter City Futures](#) has set a bold goal to make the region congestion free and energy independent by 2025

- Modern data science will let us build a generalisable data platform (any city, any country)
- If you want to know more about how we can build the data infrastructure to make this happen please contact Dr Paul Hewson texhewson@gmail.com.