



Roger Beecham  
[www.roger-beecham.com](http://www.roger-beecham.com)



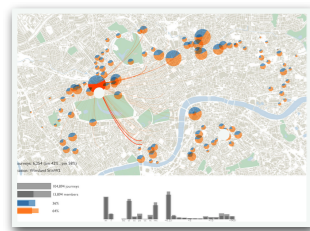
# Roger Beecham

Lecturer in Geographic Data Science  
[www.roger-beecham.com](http://www.roger-beecham.com)

Technical: Data Visualization, New and Computational Statistics  
 Applied: Transportation, Crime Science, Political Science



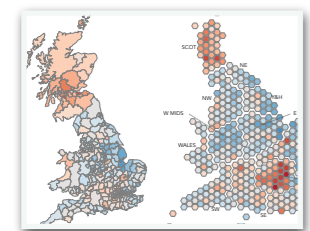
Beecham & Wood  
 Exploring gendered cycling behaviours  
 doi: [10.1080/03081060.2013.844903](https://doi.org/10.1080/03081060.2013.844903)  
 Processing, R | Estimation-based stats



Beecham et al.  
 Studying (inferring) commuter workplaces  
 doi: <https://doi.org/10.1016/j.compenvurbsys.2013.10.007>



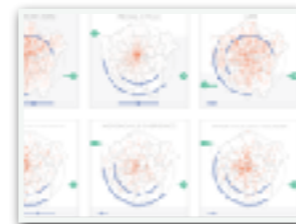
Beecham & Wood  
 Characterising (inferring) group-cycling  
 doi: [10.1016/j.trc.2014.03.007](https://doi.org/10.1016/j.trc.2014.03.007)  
 Java, Processing | Network stats



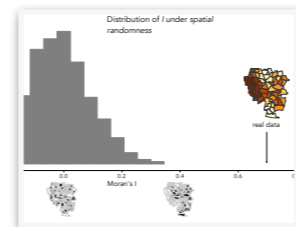
Beecham et al.  
 Locally varying explanations  
 doi: [10.5311/IJOSIS.2018.16.377](https://doi.org/10.5311/IJOSIS.2018.16.377)  
 R | GW stats, penalised regression, permutation



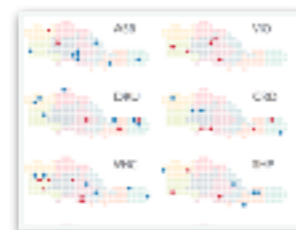
Beecham et al.  
 An update to ecological analysis in Political Science  
 in press  
 R | penalised regression, permutation



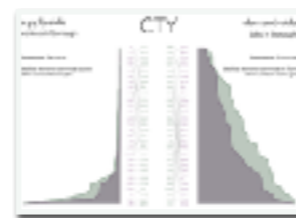
Beecham et al.  
 Multivariate small multiples  
 doi: [10.1111/cgf.12900](https://doi.org/10.1111/cgf.12900)  
 Processing | Circular stats, regression



Beecham et al.  
 Map LineUps  
 doi: [10.1109/TVCG.2016.2598862](https://doi.org/10.1109/TVCG.2016.2598862)  
 R | Spatial stats, regression, permutation



Beecham et al.  
 Statistical Process Control for uncertainty analysis  
 in press  
 Javascript, d3 | Frequentist



Beecham & Slingsby  
 Characterising labour market self-containment  
 doi: [10.1177/0308518X19850580](https://doi.org/10.1177/0308518X19850580)  
 R | Estimation-based stats





Roger Beecham  
[www.roger-beecham.com](http://www.roger-beecham.com)

# Module content and philosophy

## Spatial modelling

- Data mining
- Response modelling
- Microsimulation
- Agent-based modelling

to

simulate and predict  
consumer behaviour

[content]

Research and industry  
case studies

to

evaluate modelling  
techniques in practice

[philosophy]

# Outcomes

By the end of this module you should be able to

1. **explain** and **critically evaluate** the role of spatial analytics in simulating and predicting consumer behaviours

-----

2. **apply** geocomputational modelling and simulation techniques on real data sets

-----

3. **devise** and **employ** spatial modelling tools to address business problems, presenting and justifying recommendations in an appropriate context

# Who's who

Roger Beecham  
Assignment 1 | Convener



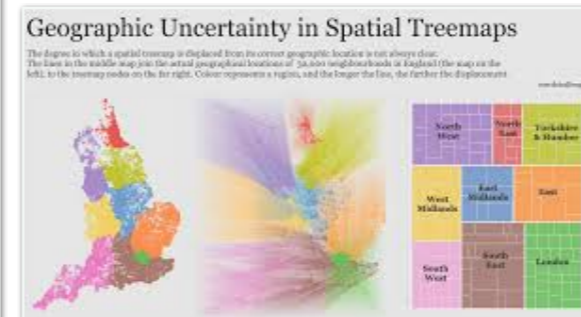
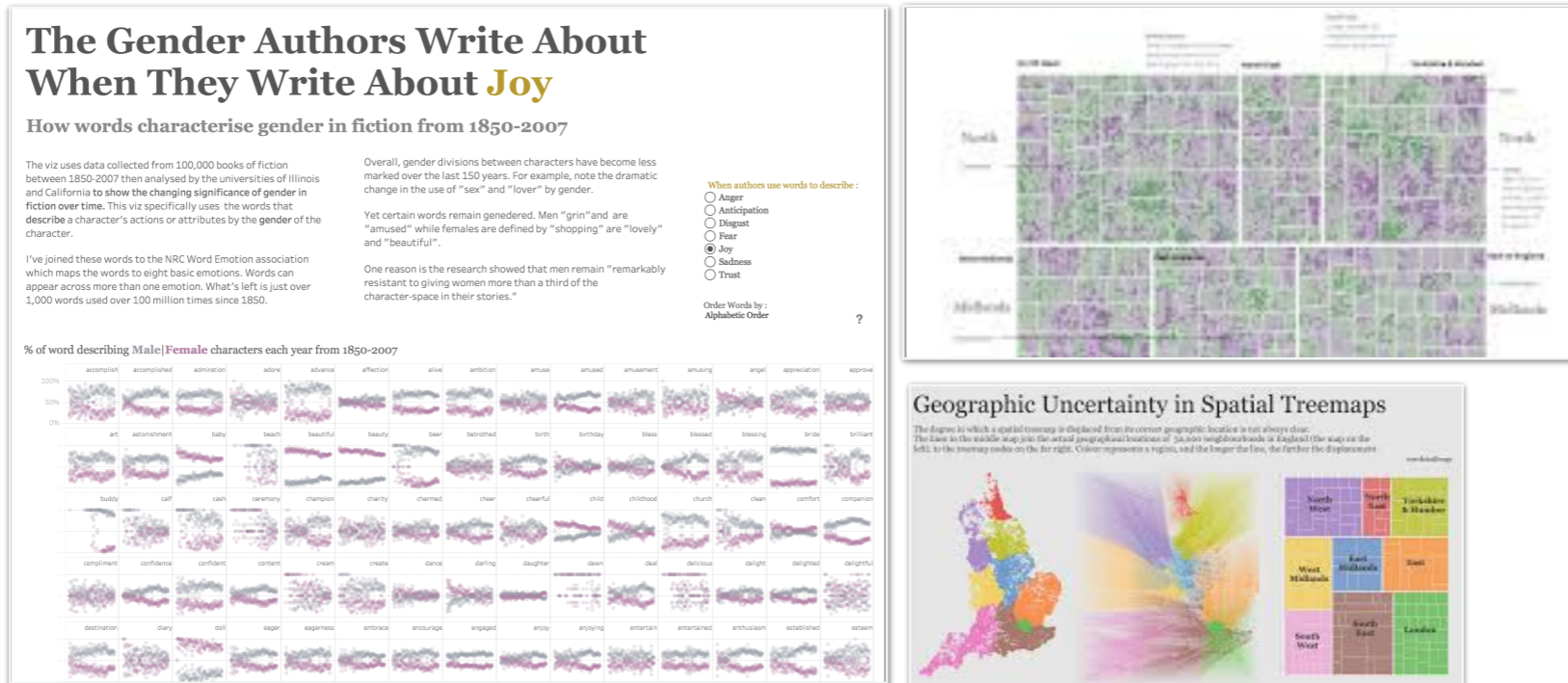
Jaiqi Ge | Nick Malleson  
Assignment 2



Rob Radburn  
Guest lecturer



# Guest Lecture



Rob Radburn  
Leicestershire County Council



# Where & When

## Lectures | Mondays

1400-1600

Roger Stevens LT 25

12.25

-----

## Labs | Wednesdays

0930-1230

Miall computing labs

10.19

1400-1700

Chemical and Process Engineering labs

GR.06



# Outline : Lectures

Roger Beecham	Session 1 Predictive Analytics & Microsimulation
Roger Beecham	Session 2 Response Modelling & Targeted Marketing
Jiaqi Ge   Nick Malleson	Session 3 Behavioural and Agent-based Models
Rob Radburn   Roger Beecham	Session 4 Guest Lecture
Roger Beecham	Session 5 Wrap-up

# Outline : Labs

Roger Beecham	Session 1 Simulating Behaviour
Roger Beecham	Session 2 Targeted Marketing
Jiaqi Ge   Nick Malleson	Session 3 Behavioural and Agent-based Models
Rob Radburn   Roger Beecham	Session 4 Coursework Surgery
Roger Beecham	Session 5 Assignment 2 — Presentations

# Assessment

## Assignment 1

Individual data analysis based on practicals 1 and 2  
1,000 words, 4 figures  
Thursday 16<sup>th</sup> January 2020 by 2pm

## Assignment 2

Group presentations based on practical 3  
Presentations held Wednesday 11<sup>th</sup> December 2019

# Technologies

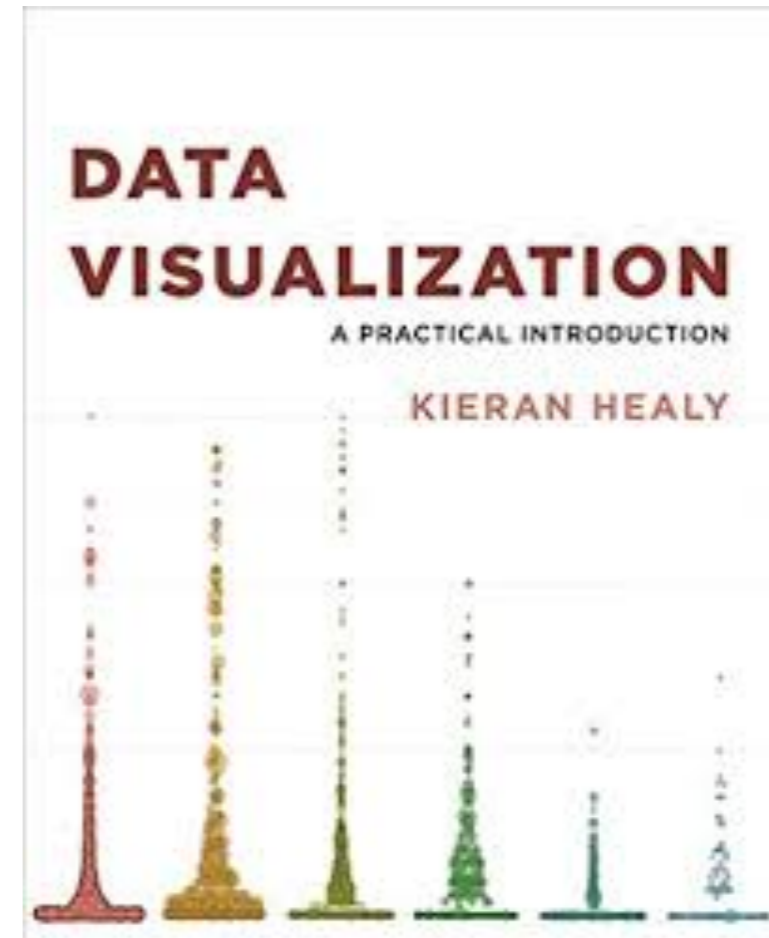
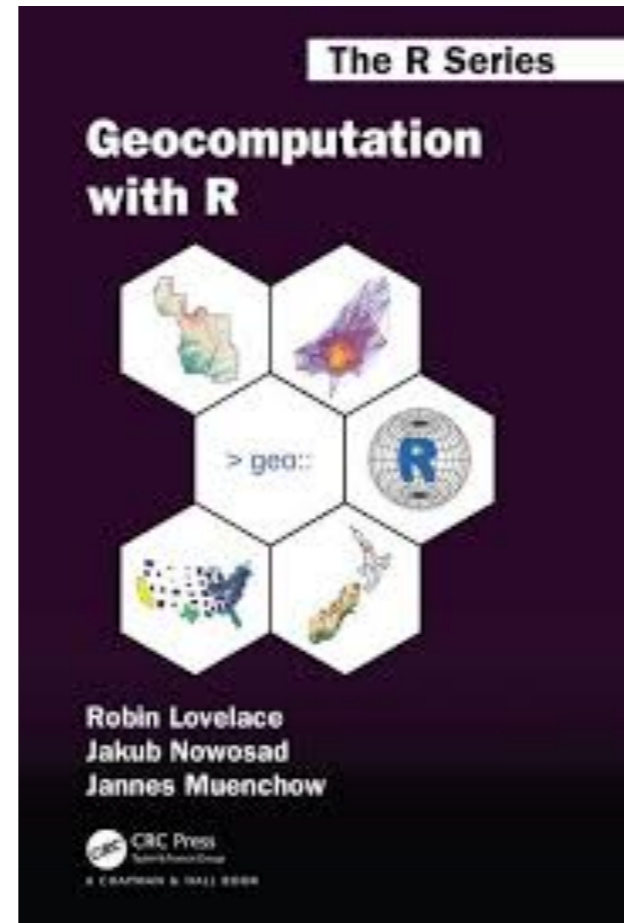
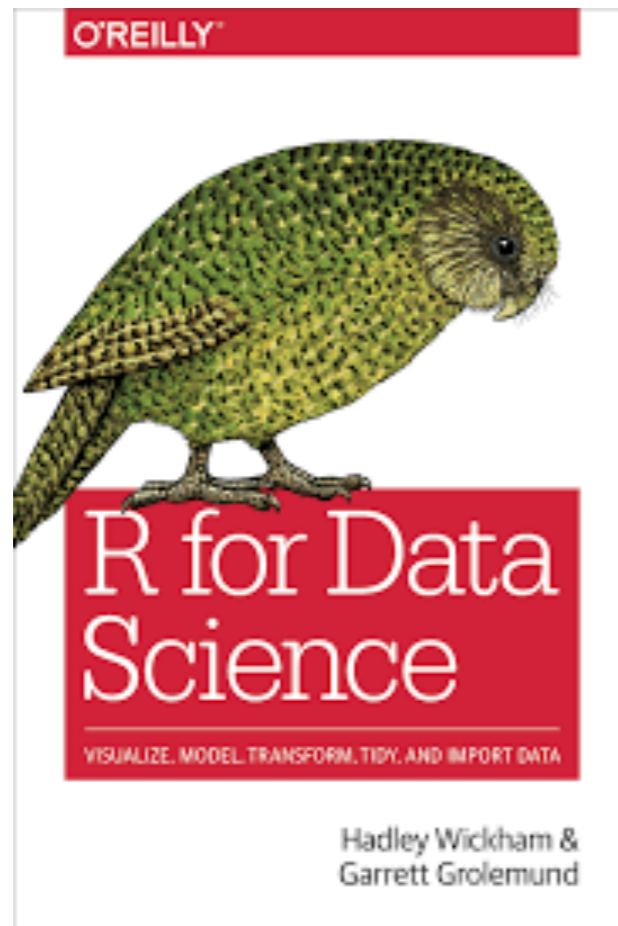


# Textbook?

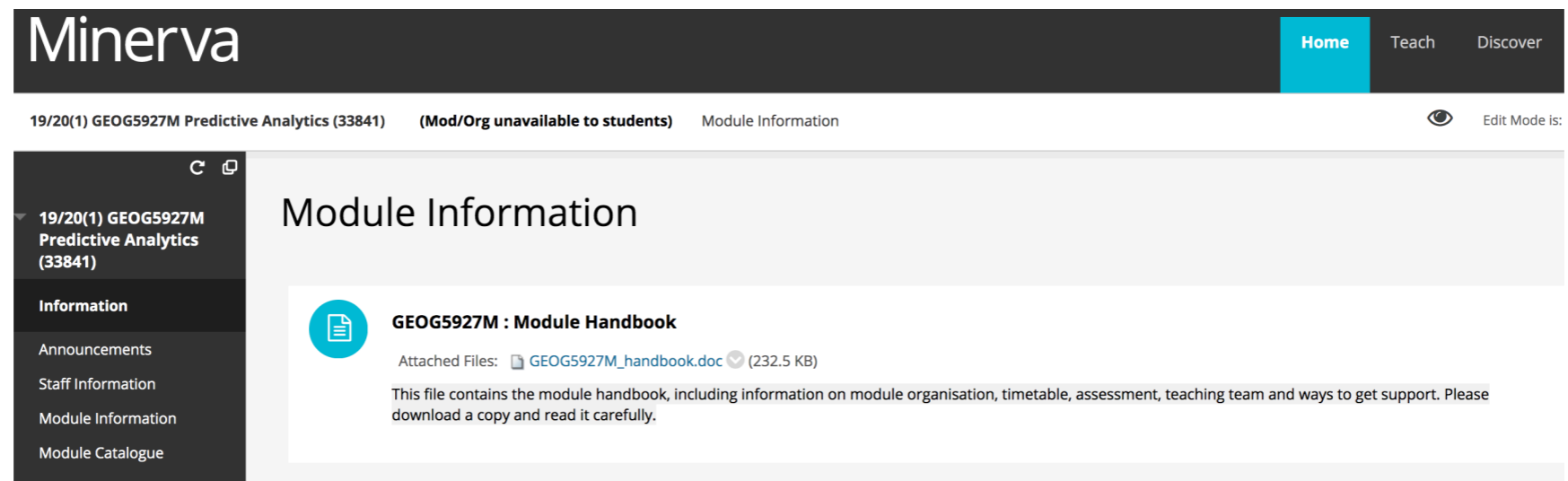
No fixed text book

-----

... But



# Module Resources



Minerva

Home Teach Discover

19/20(1) GEOG5927M Predictive Analytics (33841) (Mod/Org unavailable to students) Module Information

19/20(1) GEOG5927M Predictive Analytics (33841)

Information


Announcements


Staff Information

Module Information

Module Catalogue

## Module Information

 **GEOG5927M : Module Handbook**

Attached Files:  GEOG5927M\_handbook.doc (232.5 KB)

This file contains the module handbook, including information on module organisation, timetable, assessment, teaching team and ways to get support. Please download a copy and read it carefully.

- Module handbook
- Assessment details
- Assessment submission page

# Module Resources

**GEOG5927M: PREDICTIVE ANALYTICS**

SYLLABUS SCHEDULE R

**PREDICTIVE ANALYTICS**

THIS SITE CONTAINS the syllabus, schedule, and assignments for GEOG5927M: Predictive Analytics, held during Autumn/Winter term 2019 at University of Leeds.

**INSTRUCTOR**  
Dr. Roger Beecham  
10.139 Manton  
r.j.beecham@leeds.ac.uk  
@rbeechem

**COURSE**  
Mon (lec) & Weds (lab)  
November 11–December 11, 2019  
2:00pm–4:00pm (lec)  
Roger Stevens LT25 (12.25)

**GEOG5927M Predictive Analytics**

**GEOG5927M: PREDICTIVE ANALYTICS (AUTUMN/WINTER 2019)**  
University of Leeds | School of Geography

Dr. Roger Beecham | r.j.beecham@leeds.ac.uk | Mon (lec) & Weds (lab) | 2:00pm–4:00pm (lec) | Roger Stevens LT25 (12.25)

All content licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.  
This site adapted from the ath-tufte-hugo theme (thanks to Andrew Heiss). Made with blogdown and Hugo.

**GEOG5927M: PREDICTIVE ANALYTICS**

SYLLABUS SCHEDULE R

**SYLLABUS**

Combining theory and practical examples, this module introduces Predictive Analytics via two geocomputational techniques in which University of Leeds Geography specialises: spatial microsimulation and agent-based modelling. You will apply these techniques to data analyses highly relevant to consumer analytics domain and using modern data analysis environments.

By the end of this course you should be able to:

- explain and critically evaluate the role of spatial analytics and geocomputational modelling in simulating and predicting consumer behaviours
- apply geocomputational modelling and simulation techniques on real data sets
- devise an analysis strategy for the implementation of sophisticated modelling tools to address business problems, presenting and justifying recommendations in an appropriate context

**INSTRUCTOR**  
Dr. Roger Beecham  
10.139 Manton  
r.j.beecham@leeds.ac.uk  
@rbeechem

**COURSE**  
Mon & Weds  
September 4–December 11, 2019  
2:00pm–4:00pm  
Roger Stevens | Manton

**GEOG5927M: PREDICTIVE ANALYTICS**

SYLLABUS SCHEDULE R

**SCHEDULE**

- Readings
- Practicals
- Lecture slides

Week 7	Simulating behaviours	📖	📄	📄
Week 8	Targeted marketing	📖	📄	📄
Week 9	Behavioural and agent-based models	📖	📄	📄
Week 10	Behavioural analytics (guest lecture)	📖	📄	📄
Week 11	Re-visit	📖	📄	📄

- Schedule — week-by-week overview
- Lecture notes
- Lab exercises

# How to learn

Come to lectures and labs

*try stuff out, engage, ask questions*

-----

Independent learning

*read and explore – be curious*

-----

Coursework

*coursework throughout (labs and lectures)*





# Introduction to Predictive Analytics



EDITOR'S PICK | 9,075 views | Sep 22, 2019, 08:26pm

# A Long View On How Big Data And AI Have Transformed Business Culture

McKinsey & Company



Support The Guardian

Available for everyone, funded by readers

Contribute →

Subscribe →

Search jobs Dating

Sign in

Search

UK edition

# The Guardian

News

Opinion

Sport

Culture

Lifestyle

More

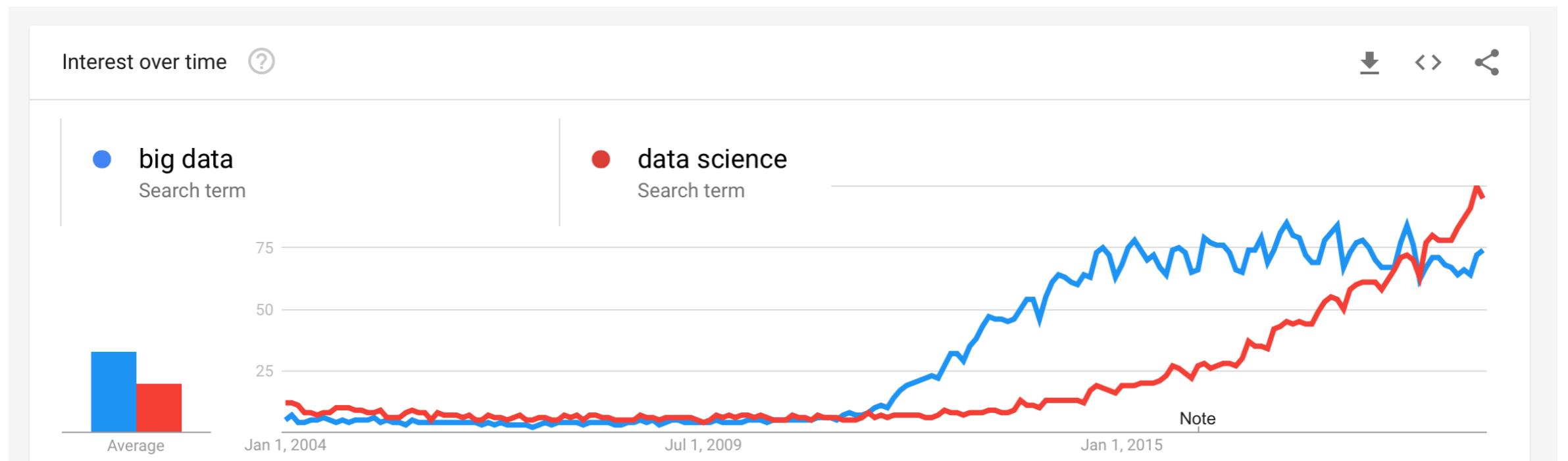
Open data economy  
Media & Tech Network

## All the data you need to power your business is free online

Open data available online is improving and streamlining businesses across Europe. We show you how to make the most of it

Advertisement

# "big data" and "data science" on Google Trends, Oct 2019





Hey, Tony, Stewart Tansley, and Kristin M. Tolle. The fourth paradigm: data-intensive scientific discovery. Vol. 1. Redmond, WA: Microsoft research, 2009.

1000 years ago – **experimental science**  
*description of natural phenomena*

100s years ago – **theoretical science**  
*Newton's laws, Maxwell's Equations*

<50 years ago – **computational science**  
*Simulate complex phenomena*

today – **data-intensive science**  
*Generate knowledge through  
observation (again)*

# THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE



Illustration: Marian Bantjes

*The next generation of scientific discovery will be data-driven as previously unrecognised patterns are discovered by analysing massive and mixed datasets.*

David Willets MP, 2013,  
Then Minister for Universities and Science




**Video:** <http://goo.gl/4ysAmw>

## ▼ nature

Letter | Published: 19 February 2009

### **Detecting influenza epidemics using search engine query data**

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

*Nature* **457**, 1012–1014 (2009) | [Cite this article](#)

Science

Contents ▾

News ▾

POLICY FORUM | BIG DATA

### The Parable of Google Flu: Traps in Big Data Analysis

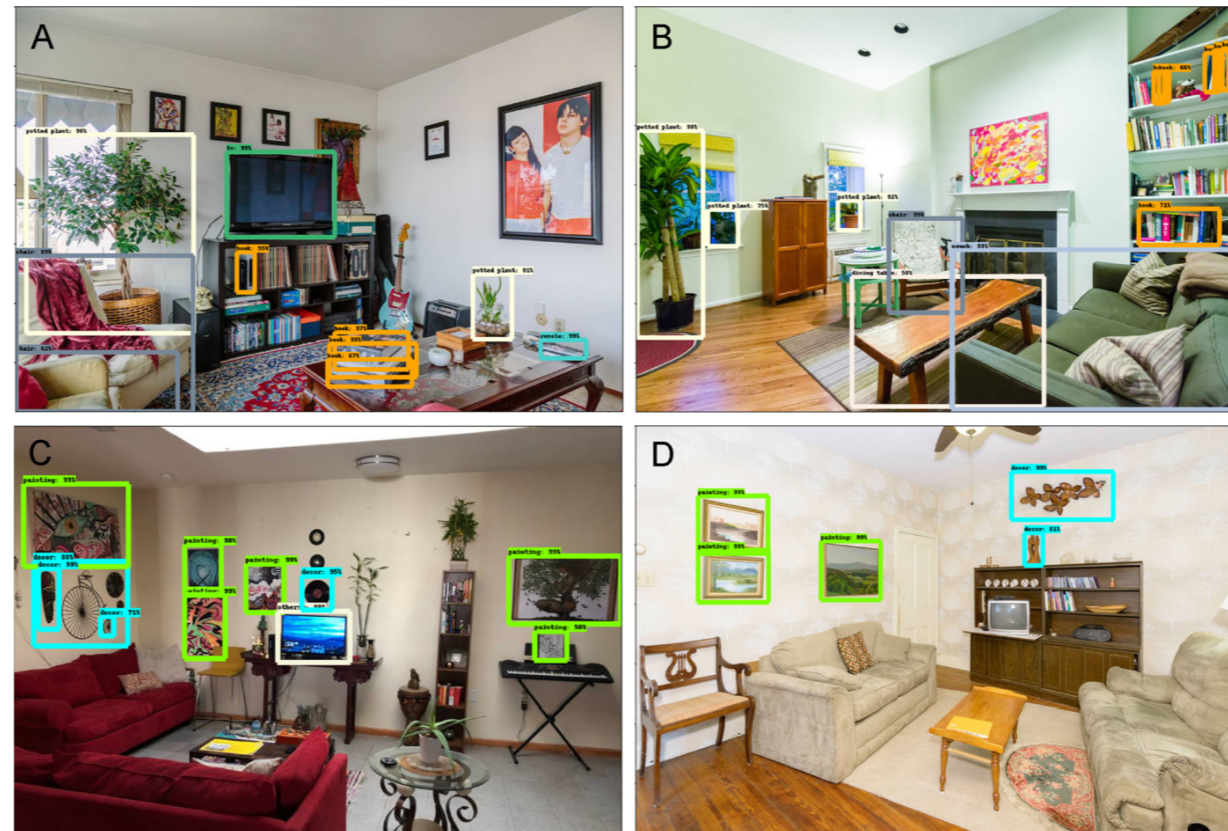
David Lazer<sup>1,2,\*</sup>, Ryan Kennedy<sup>1,3,4</sup>, Gary King<sup>3</sup>, Alessandro Vespignani<sup>5,6,3</sup>

+ See all authors and affiliations



# Inside 50,000 living rooms: an assessment of global residential ornamentation using transfer learning

Xi Liu<sup>1</sup>, Clio Andris<sup>1\*</sup>, Zixuan Huang<sup>2</sup> and Sohrab Rahimi<sup>3</sup>



**Figure 2** Object detection. Object detection examples for living room images. Ivory and orange bounding boxes in (a) and (b) are the model's results for plant and book identification, respectively. Green and blue bounding boxes in (c) and (d) are the model results for wall art and decor identification, respectively



# Should Prison Sentences Be Based On Crimes That Haven't Been Committed Yet?

By [Alex S. Taylor](#), [Betsy Johnson](#), [Dan Sussman](#) and [Drew Harlow](#)  
 Graphics by [Barbara Conner](#), [Sasha Fischer-Bain](#) and [Andy Rossbach](#)  
 Filed under [Criminal Justice](#)  
 Published Aug 1, 2016



## Who Should Get Parole?

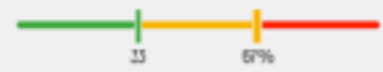
Even the best risk assessments yield probabilities, not certainties. That means they label as "high risk" some people who won't commit another crime and label as "low risk" some people who will. This simulation lets you sort offenders into risk categories based on the results of an assessment. Think we should rarely lock up anyone who wouldn't reoffend? Set the "low risk" threshold high and the "high risk" threshold even higher. Have little tolerance for recidivism? Try the opposite. In the real world, policymakers have to strike a balance. [Read more »](#)

1

The prisoners in this simulation are up for parole. Some will reoffend if released and some won't. They each take an assessment, which estimates the chance they will reoffend.

2

Prisoners are placed in one of three categories based on these estimates. **Move the slider** to change the cutoffs for each category. "Low risk" prisoners will be awarded parole, "high risk" prisoners will be denied.

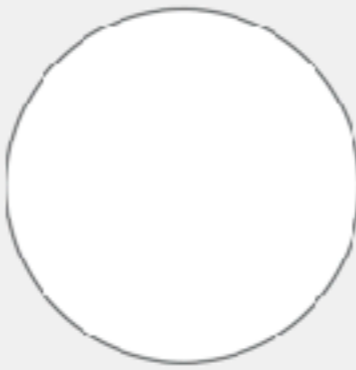


3

Some people you let out reoffended. Some people you left in prison wouldn't have. **Are you OK with the results?**

Start

PAROLE-ELIGIBLE PRISONERS



● - Will reoffend  
 ○ - Will not reoffend

LOW RISK

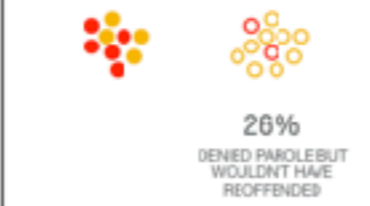
MEDIUM RISK

HIGH RISK

AWARDED PAROLE



DENIED PAROLE



## Data mining and machine learning –

*Detect hidden patterns in data*

## Information Visualization –

*Explore complex structure and patterns in data that are difficult to expose using computation alone*

## Predictive analytics —

*Use these patterns to predict, under uncertainty, what will happen in future*



# Assignments

In *Assignment #1* we'll be generating a large synthetic dataset of customers and looking for behavioural and demographic associations between individuals to better *\*target\** marketing activity.

In *Assignment #2* we'll be using data and heuristics to explore and predict how customers will behave and respond to different store formats.





break





# GEOG5927M: PREDICTIVE ANALYTICS

[SYLLABUS](#)   [SCHEDULE](#)   [R](#)

## PRACTICAL 1 : SIMULATING BEHAVIOUR

*The aim of this session is to create a synthetic population of households in Leeds. You will use an individual-level survey dataset describing individuals' holiday-making behaviour and apply spatial microsimulation to estimate a population-level dataset of these holiday-making behaviours at the household-level in Leeds. In Practical 2, you will use this population to undertake a data analysis to support a targeted marketing strategy.*

# microdata.csv

15,189 records

Person_ID	OA_GRP	Sex	Ageband	NumberCh	CombinedI	OverSeasAi	UKAirport	OverallHoli	AgeSex	Supergroup
11603	8c	F	a35to49	2	26-30K	LEI	MAN	Excellent	F35to49	Hard-Press
11285	8c	F	a25to34	0	0-10K	IBZ	MAN	Fair	F25to34	Hard-Press
13938	8c	M	a50to64	1	16-20K	LCA	BHX	Fair	M50to64	Hard-Press
10255	8c	F	a25to34	1	26-30K	ALC	LBA	Poor	F25to34	Hard-Press
831	8c	M	a50to64	0	26-30K	AGA	MAN	Good	M50to64	Hard-Press
1754	8c	M	a65over	0	Not Answer	DLM	MAN	Good	M65over	Hard-Press
2330	8c	F	a65over	0	Not Answer	DLM	MAN	Excellent	F65over	Hard-Press
10818	8c	M	a25to34	0	36-40K	KGS	MAN	Fair	M25to34	Hard-Press
8237	8c	M	a65over	2	16-20K	FUE	MAN	Good	M65over	Hard-Press
11508	8c	F	a35to49	2	71-80K	ZTH	LBA	Poor	F35to49	Hard-Press

ageBand	demographics
incomeBand	demographics
oac	geodemographics
originAirport	preference
destinationAirport	preference/attitude
satisfactionScore	preference/attitude







# microdata.csv

## 15,189 records

Person_ID	OA_GRP	Sex	Ageband	NumberCh	Combined+	OverSeasAi	UKAirport	OverallHoli	AgeSex	Supergroup
11603	8c	F	a35to49	2	26-30K	LEI	MAN	Excellent	F35to49	Hard-Pressé
11285	8c	F	a25to34	0	0-10K	IBZ	MAN	Fair	F25to34	Hard-Pressé
13938	8c	M	a50to64	1	16-20K	LCA	BHX	Fair	M50to64	Hard-Pressé
10255	8c	F	a25to34	1	26-30K	ALC	LBA	Poor	F25to34	Hard-Pressé
831	8c	M	a50to64	0	26-30K	AGA	MAN	Good	M50to64	Hard-Pressé
1754	8c	M	a65over	0	Not Answered	DLM	MAN	Good	M65over	Hard-Pressé
2330	8c	F	a65over	0	Not Answered	DLM	MAN	Excellent	F65over	Hard-Pressé
10818	8c	M	a25to34	0	36-40K	KGS	MAN	Fair	M25to34	Hard-Pressé
8237	8c	M	a65over	2	16-20K	FUE	MAN	Good	M65over	Hard-Pressé
11508	8c	F	a35to49	2	71-80K	ZTH	LBA	Poor	F35to49	Hard-Pressé

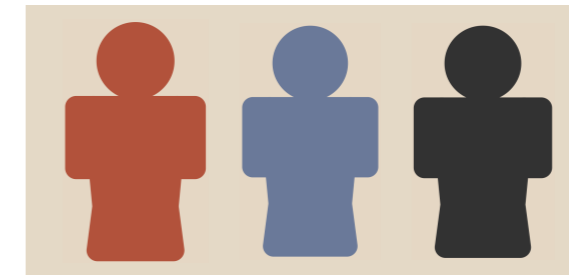


# Spatial Microsimulation

Census data  
high-level | low in detail  
population-level | complete



Survey data  
individual-level | rich in detail  
small-scale | unrepresentative



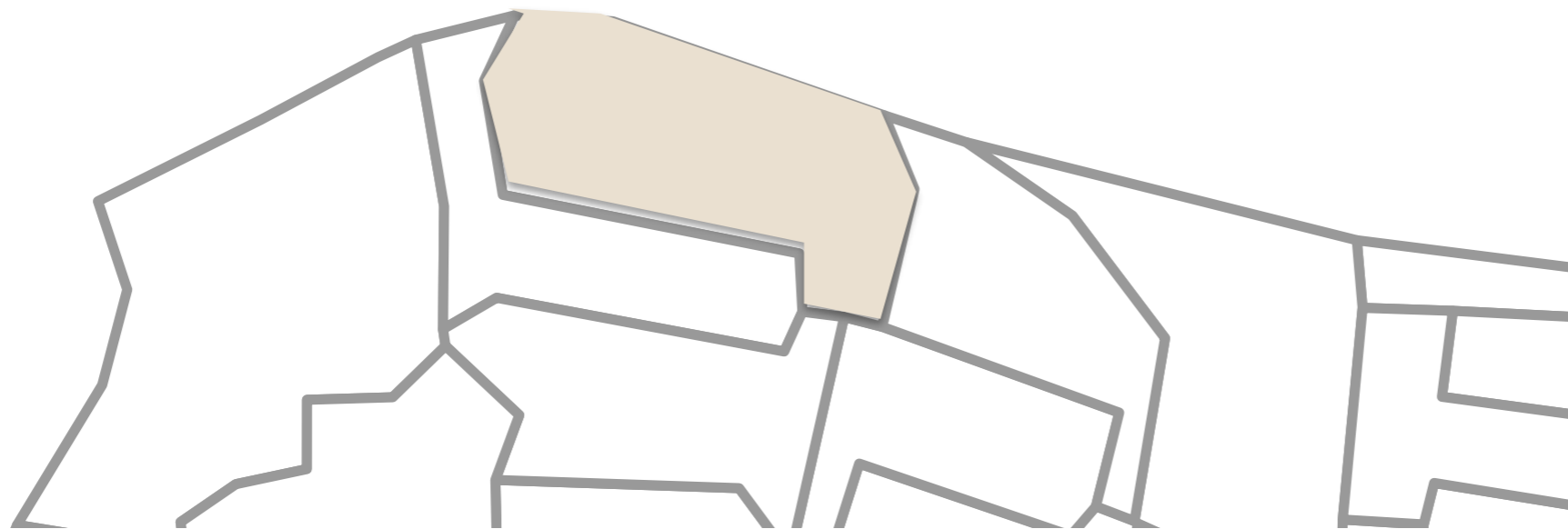
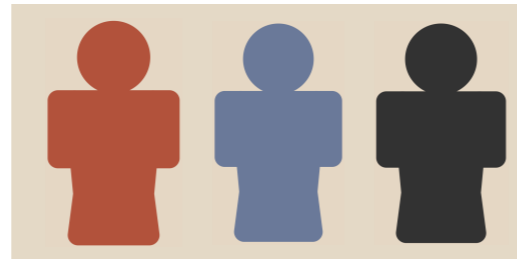
# Spatial Microsimulation

*The creation, analysis and modelling of individual-level data allocated to geographic zones.*

*Lovelace & Dumont 2016*

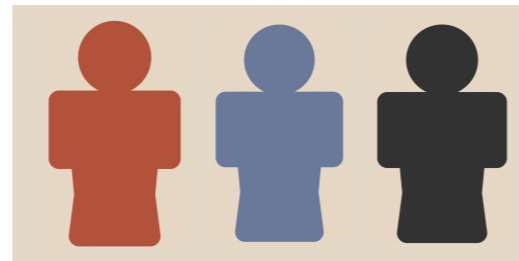
# Spatial Microsimulation

survey data  
individuals



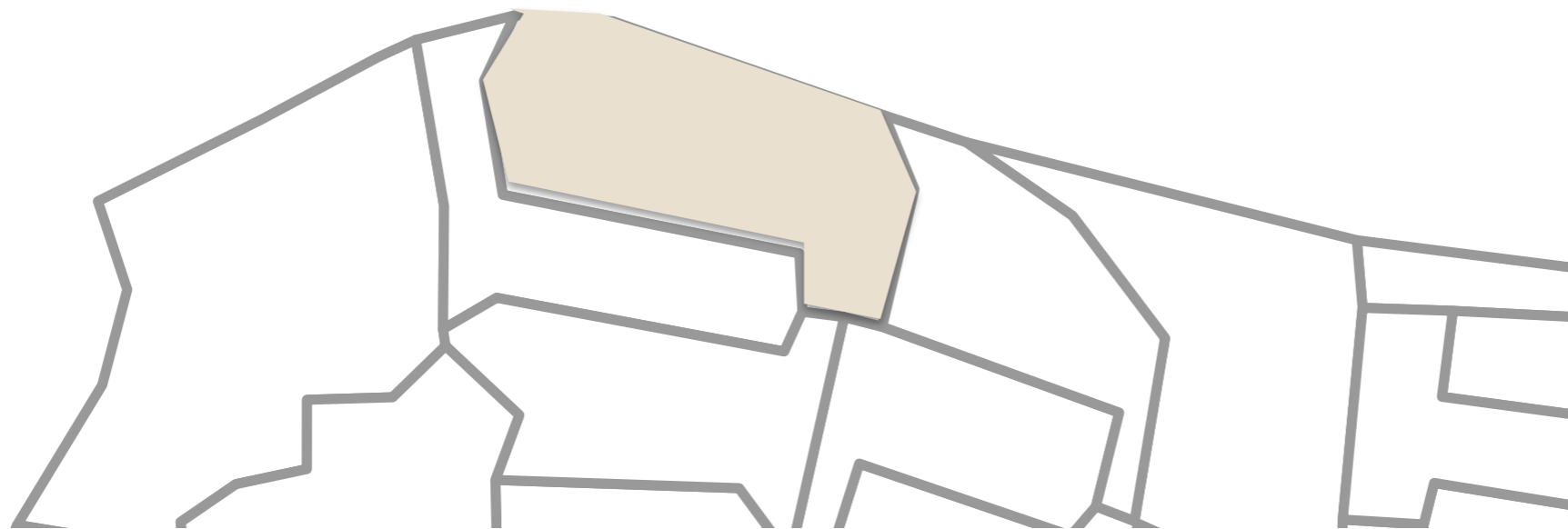
# Spatial Microsimulation

survey data  
individuals

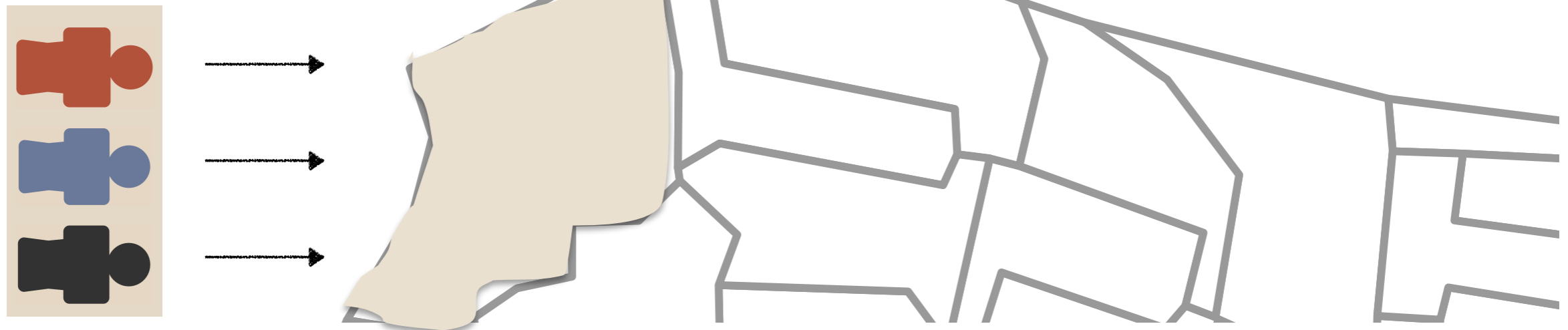


+

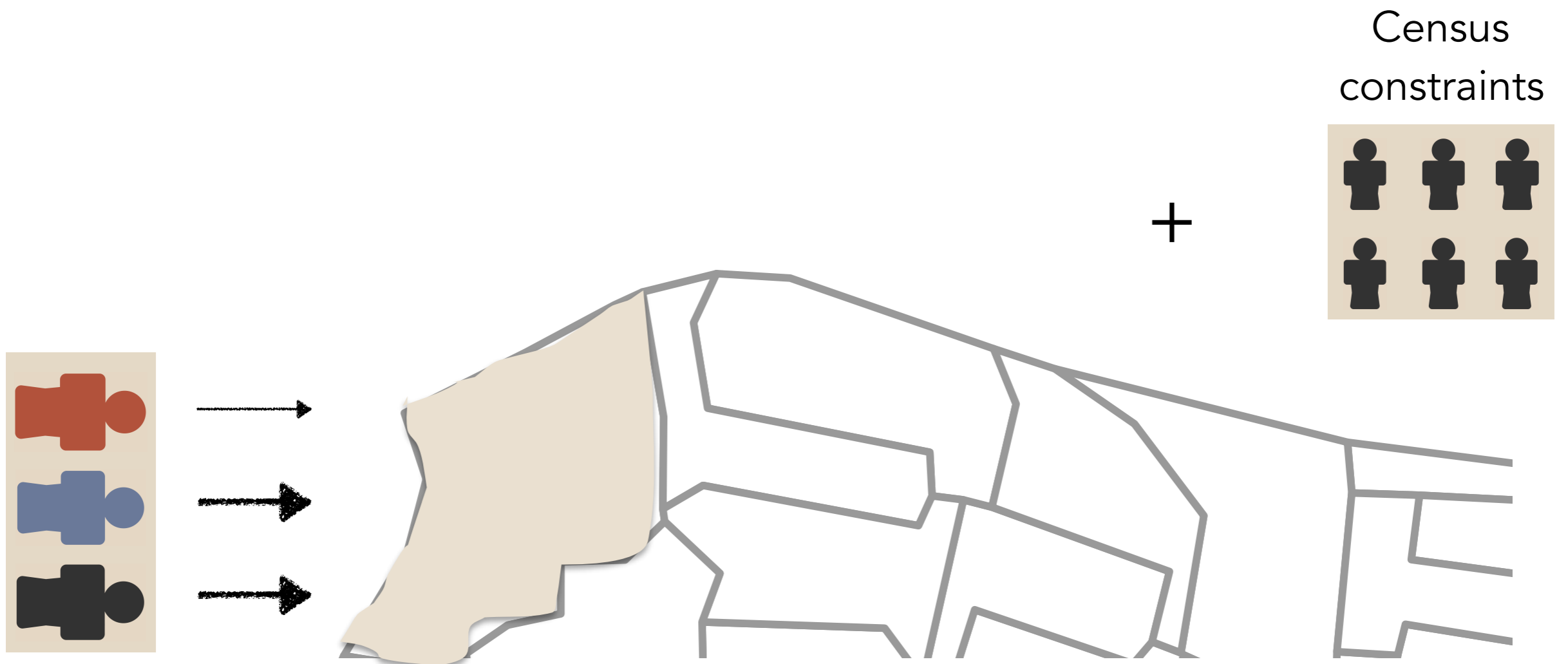
Census  
constraints



# Spatial Microsimulation



# Spatial Microsimulation





*Microsimulation does not  
generate **new** data*

--

*merely copies of existing data*

*Lovelace & Dumont 2016*

# Examples

## Health : smoking [Tomintz et al. 2008]

- **Why?** Reported in individual surveys, but not population-level and not from place-to-place
- **Benefits** : could be used to target/locate smoking support clinics
- **'Benefits'** : could be used by a Tobacco company for targeting investment

# Examples

## Economic policy : evaluation [De Agostini et al. 2014]

- **Why?** Simulate / spread impacts inferred from individual-level data over an entire country
- **Benefits** : quantify (under uncertainty) the impacts of a regressive welfare reform at the country-level
- **Benefits** : evidence-based decision-making

# Examples

Transport : simulating behaviour [Lovelace 2014]

- **Why?** When designing infrastructure, want to know about the distribution of individuals meeting a particular set of characteristics
- **Benefits** : provide evidence around likely winners and losers of a new infrastructure investment

# Assumptions

- Individual-level microdata are representative of the study area
- Target variable is dependent on the constraint variables in a way that is relatively constant over space and time
- Input microdataset and constraints are sufficiently rich and detailed to reproduce the full diversity of individuals and areas in the study region



## GEOG5927M: PREDICTIVE ANALYTICS

[SYLLABUS](#) [SCHEDULE](#) [R](#)

### PRACTICAL 1 : SIMULATING BEHAVIOUR

*The aim of this session is to create a synthetic population of households in Leeds. You will use an individual-level survey dataset describing individuals' holiday-making behaviour and apply spatial microsimulation to estimate a population-level dataset of these holiday-making behaviours at the household-level in Leeds. In Practical 2, you will use this population to undertake a data analysis to support a targeted marketing strategy.*



# Dataset

Person_ID	OA_GRP	Sex	Ageband	NumberCh	CombinedF	OverSeasAi	UKAirport	OverallHoli	AgeSex	Supergroup
11603	8c	F	a35to49	2	26-30K	LEI	MAN	Excellent	F35to49	Hard-Pressé
11285	8c	F	a25to34	0	0-10K	IBZ	MAN	Fair	F25to34	Hard-Pressé
13938	8c	M	a50to64	1	16-20K	LCA	BHX	Fair	M50to64	Hard-Pressé
10255	8c	F	a25to34	1	26-30K	ALC	LBA	Poor	F25to34	Hard-Pressé
831	8c	M	a50to64	0	26-30K	AGA	MAN	Good	M50to64	Hard-Pressé
1754	8c	M	a65over	0	Not Answer	DLM	MAN	Good	M65over	Hard-Pressé
2330	8c	F	a65over	0	Not Answer	DLM	MAN	Excellent	F65over	Hard-Pressé
10818	8c	M	a25to34	0	36-40K	KGS	MAN	Fair	M25to34	Hard-Pressé
8237	8c	M	a65over	2	16-20K	FUE	MAN	Good	M65over	Hard-Pressé
11508	8c	F	a35to49	2	71-80K	ZTH	LBA	Poor	F35to49	Hard-Pressé

**microdata.csv**

**15,189 records**

ZoneID	Person_ID	OA_GRP	Sex	Ageband	NumberCh	CombinedF	OverSeasAi	UKAirport	OverallHoli	AgeSex	Supergroup
E00056750	11603	8c	F	a35to49	2	26-30K	LEI	MAN	Excellent	F35to49	Hard-Pressed Living
E00056750	11285	8c	F	a25to34	0	0-10K	IBZ	MAN	Fair	F25to34	Hard-Pressed Living
E00056750	13938	8c	M	a50to64	1	16-20K	LCA	BHX	Fair	M50to64	Hard-Pressed Living
E00056750	10255	8c	F	a25to34	1	26-30K	ALC	LBA	Poor	F25to34	Hard-Pressed Living
E00056750	831	8c	M	a50to64	0	26-30K	AGA	MAN	Good	M50to64	Hard-Pressed Living
E00056750	1754	8c	M	a65over	0	Not Answer	DLM	MAN	Good	M65over	Hard-Pressed Living
E00056750	2330	8c	F	a65over	0	Not Answer	DLM	MAN	Excellent	F65over	Hard-Pressed Living
E00056750	10818	8c	M	a25to34	0	36-40K	KGS	MAN	Fair	M25to34	Hard-Pressed Living
E00056750	8237	8c	M	a65over	2	16-20K	FUE	MAN	Good	M65over	Hard-Pressed Living
E00056750	11508	8c	F	a35to49	2	71-80K	ZTH	LBA	Poor	F35to49	Hard-Pressed Living

**simulated\_population.csv**

**320,596 records**

