# R

2/7/15

# intro: hi!

# overview

Module 1: Introduction to R (Paul Paczuski)

Module 2: Graphics And Data Manipulation Using ggplot and dplyr (Paul Paczuski)

Module 3: Statistical Modeling (Matthew Eaton, separate slides)

# details

- hands-on tutorials and exercises
- 10 min break after Module 1 and Module 2

*intro tutorial from James et al*

*custom ggplot2 tutorial*

*vignette from dplyr*

*custom stats modeling tutorial*

# more prep

checklist:

preparation.txt

# 1
# intro to R

"R is a free software environment for statistical computing and graphics"

"The best way to learn a new language is to try out the commands"

**(A) Tutorial**

/programs/1-intro.R

go through Lab 2.3 in James et al, p 42-52

**(B) Exercise**

in James et al, exercise 8 p 54

**(C) Tips, regroup, Q&A, break**

# tips

- set up directory structure ahead of time
- library() calls at top
- document programs in a README file
- use style guide
- etc

# 2
# graphics and data manipulation using ggplot2 and dplyr

R packages are simply add-ons:
for more or better functionality

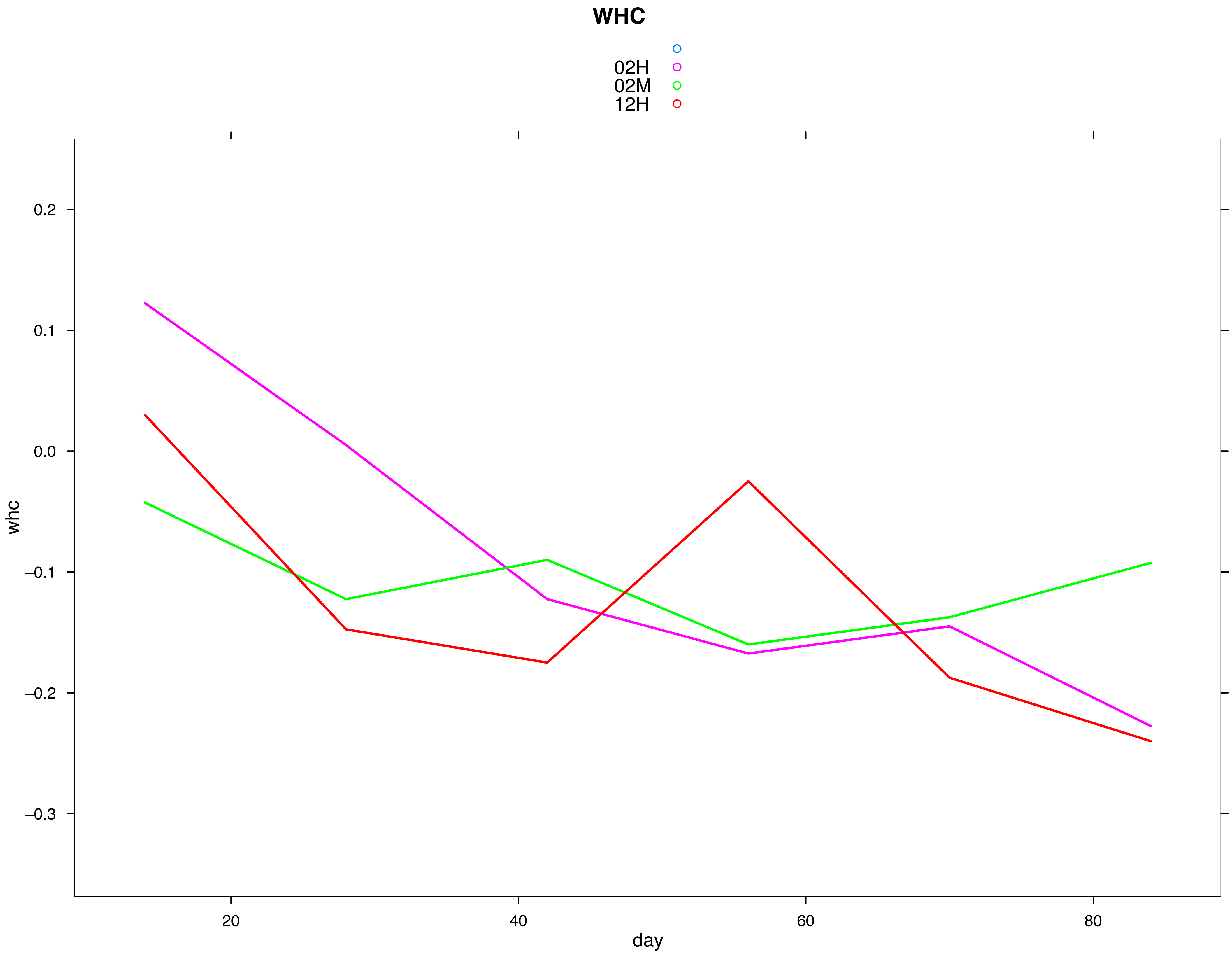
| | |
|---|---|
| ggplot2 | graphics |
| dplyr | data manipulation |

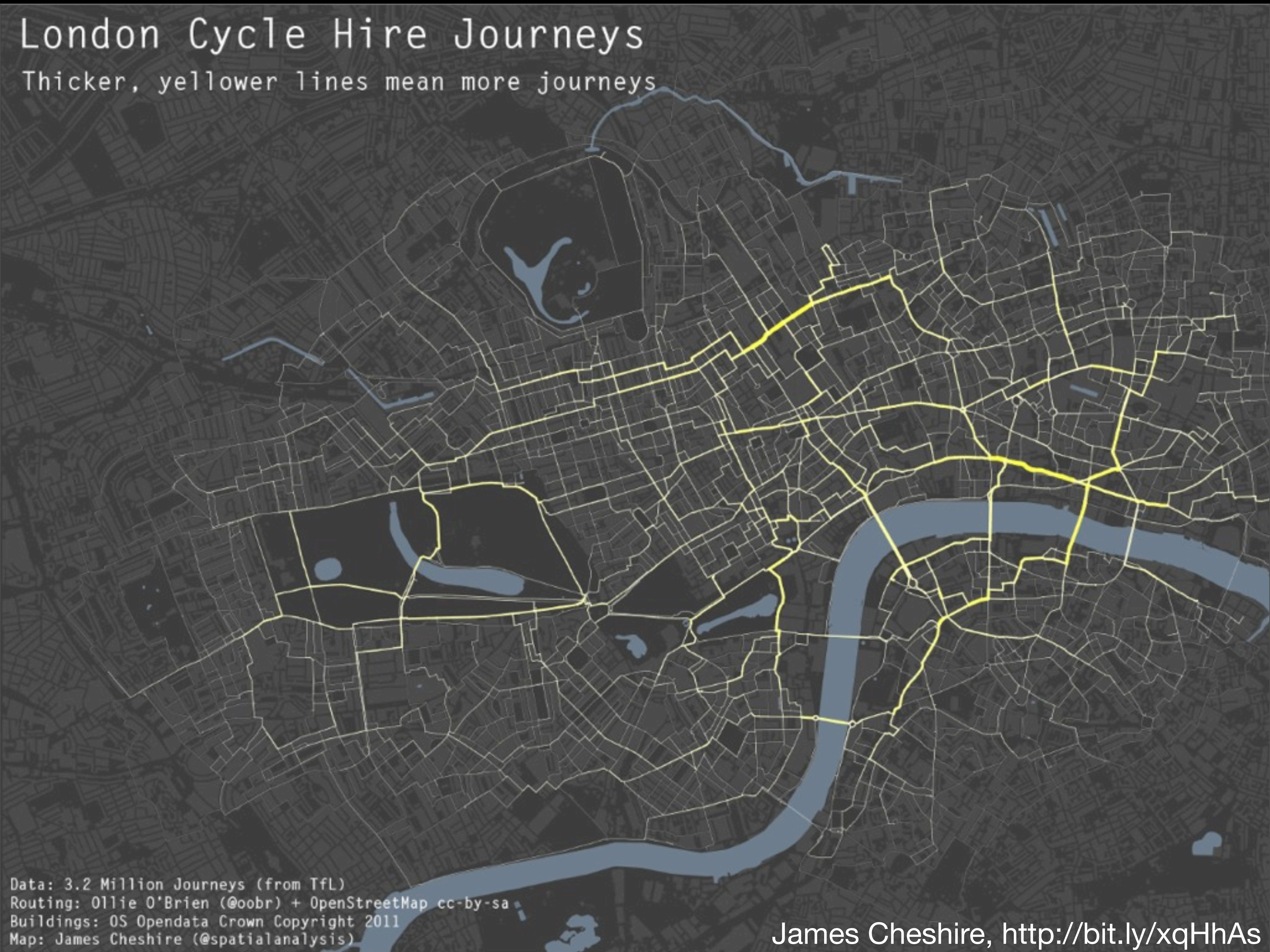
# ggplot2

"ggplot2 is designed to work in a layered fashion, starting with a layer showing the raw data then adding layers of annotation and statistical summaries. [..]"

*credits: Hadley Wickham’s slides*

# 2004



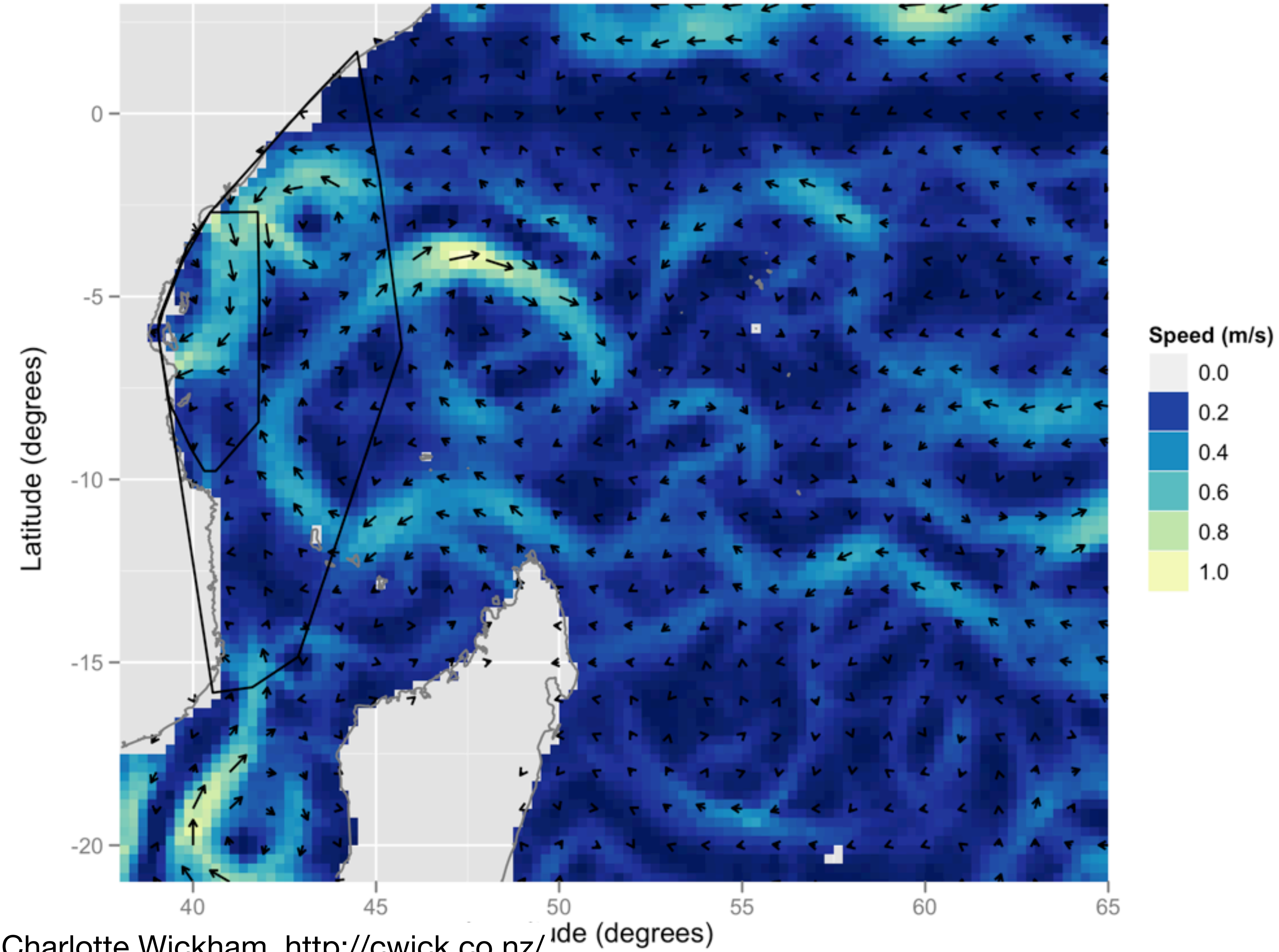


London Cycle Hire Journeys

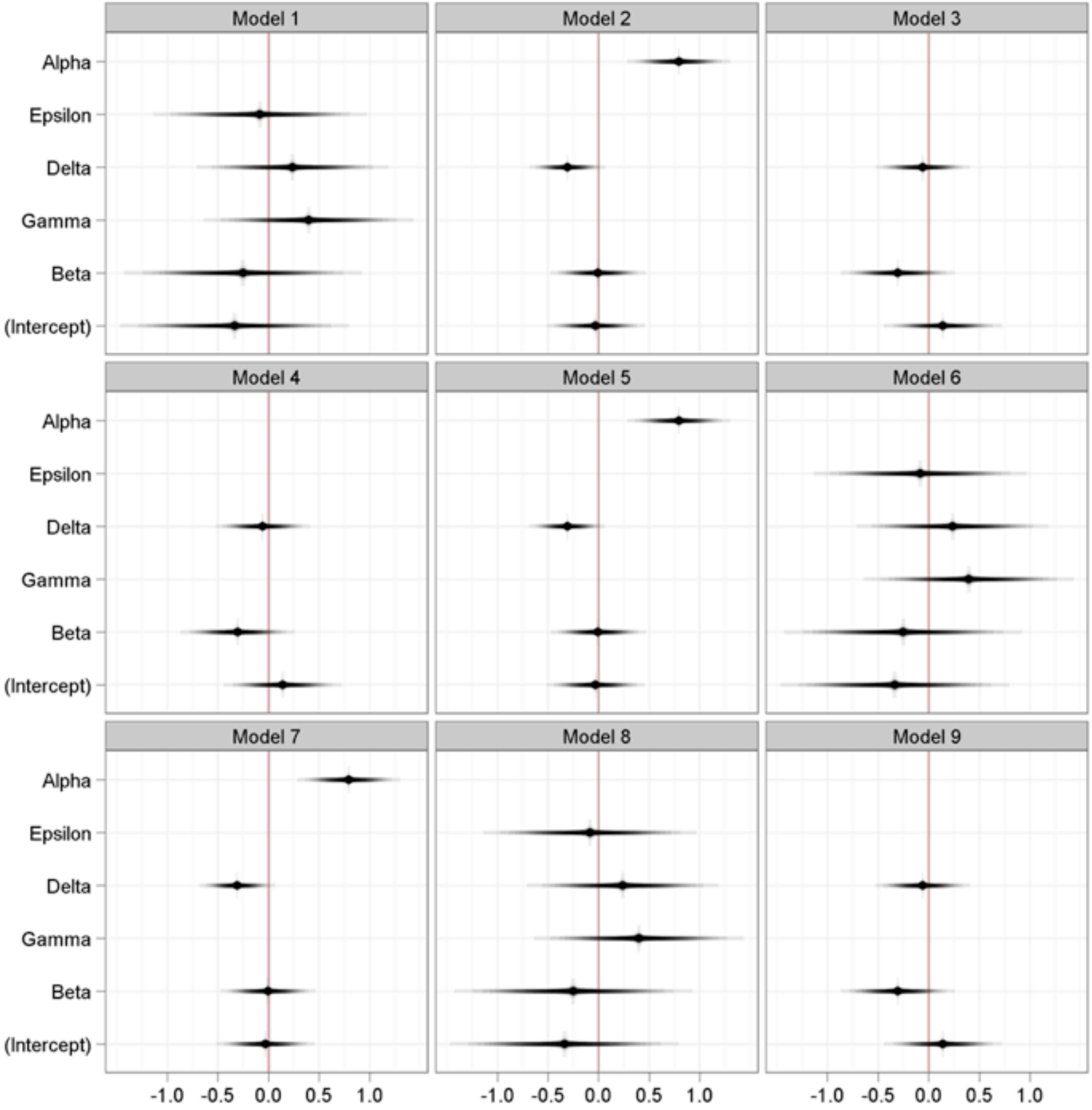
Thicker, yellower lines mean more journeys

Data: 3.2 Million Journeys (from TfL)
Routing: Ollie O'Brien (@oobr) + OpenStreetMap cc-by-sa
Buildings: OS Opendata Crown Copyright 2011
Map: James Cheshire (@spatialanalysis)

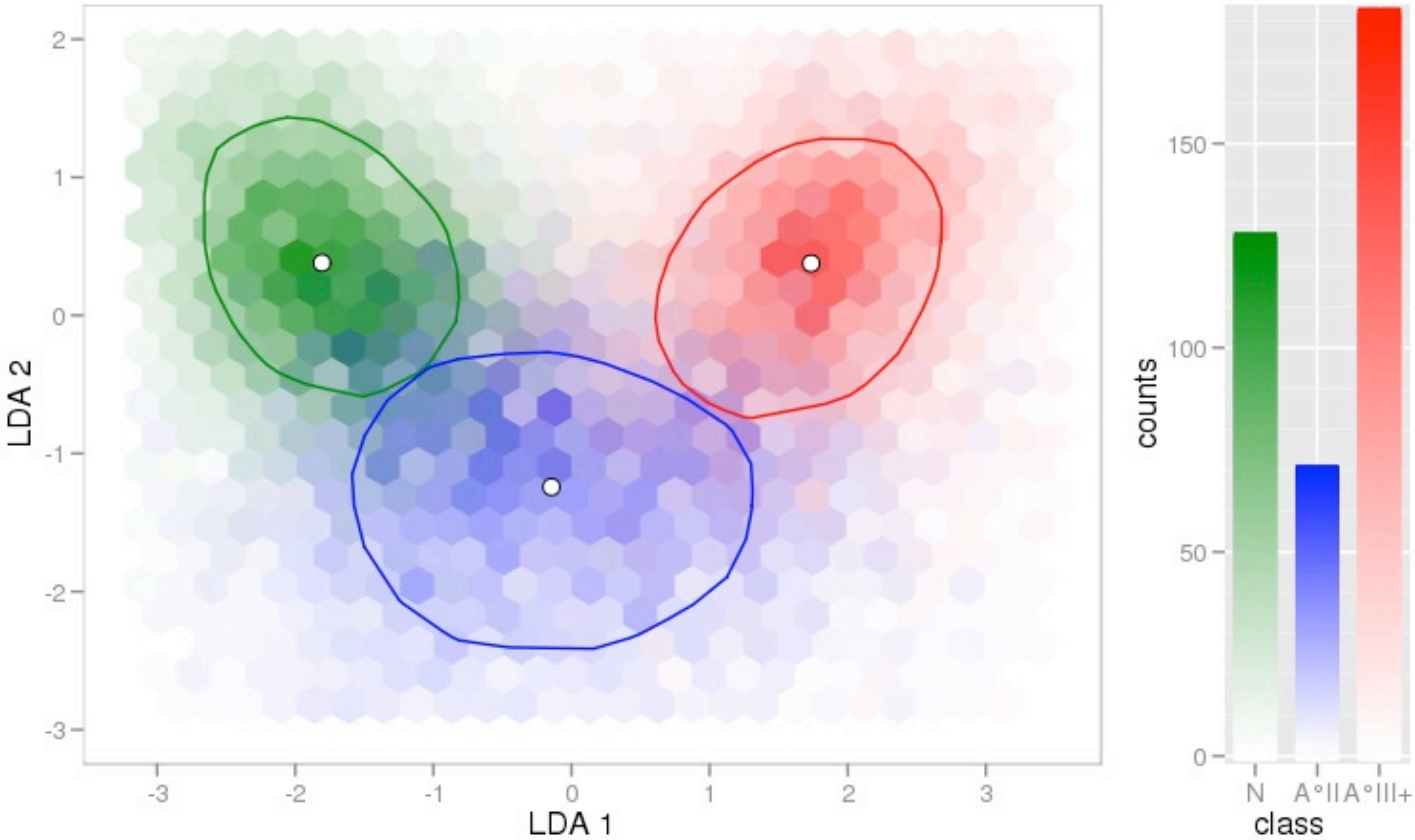
James Cheshire, http://bit.ly/xqHhAs



Charlotte Wickham, http://cwick.co.nz/



David B Sparks, http://bit.ly/hn54NW



Claudia Beleites, http://bit.ly/yNqlpz

# dplyr

dplyr couldn't be easier, with the use of the following verbs as R functions:

**Select** data columns

**Filter** data to select specific rows

**Arrange** the rows in order

**Mutate** your data to add new columns

**Summarise** chunks of your data in some way

# dplyr

Now, for very large data, R may be slow. But dplyr has many of its parts written in C++ which makes it extremely fast

## (A) Tutorials

/programs/2-ggplot2.R

/programs/2b-dplyr.R

## (B) Exercise

explore the college.csv dataset using ggplot2 and dplyr

send results to pavopax@gmail.com

# (extra)

as an exercise, reproduce a graphic from

T. Piketty's “Capital in the 21st century”

https://github.com/pavopax/piketty

also see:

http://simplystatistics.org/2014/06/30/piketty-in-r-markdown-we-need-some-help-from-the-crowd/

# 3
# statistical modeling

(Matthew Eaton)