INM3061/INM430

Principles of Data Science

Week 05

Supporting Analysis through Models (and Prediction)

Aidan Slingsby, giCentre

CITY
UNIVERSITY OF LONDON
EST 1894

Module Schedule

- Week 01: Introduction & Basic Concepts
- Week 02: Data Characteristics & Wrangling
- Week 03: Data Processing & Summarization
- Week 04: Inferential Statistics and the "New Statistics"
- **Week 05: Supporting Analysis using Models and Prediction**
- Week 06: Reading week (no lectures)
- Week 07: Finding structure in data
- Week 08: Analysing text
- Week 09: Networks and Knowledge Representation
- Week 10: Processing data from images
- Week 11: Wrap-up (and writing code in the Real World)

Context

- Models and prediction for supporting analysis
- Regression
  - Continuous: Simple and multiple linear; other types
  - Categorical: Logistic regression; decision trees; SVM
- Validation (and avoiding overfitting)
- Causal thinking
  - Experimental design vs observational
  - Counterfactuals
  - Confounders
  - Causal thinking
  - Know your domain, know your data and *think*

MODELS

What is a model?

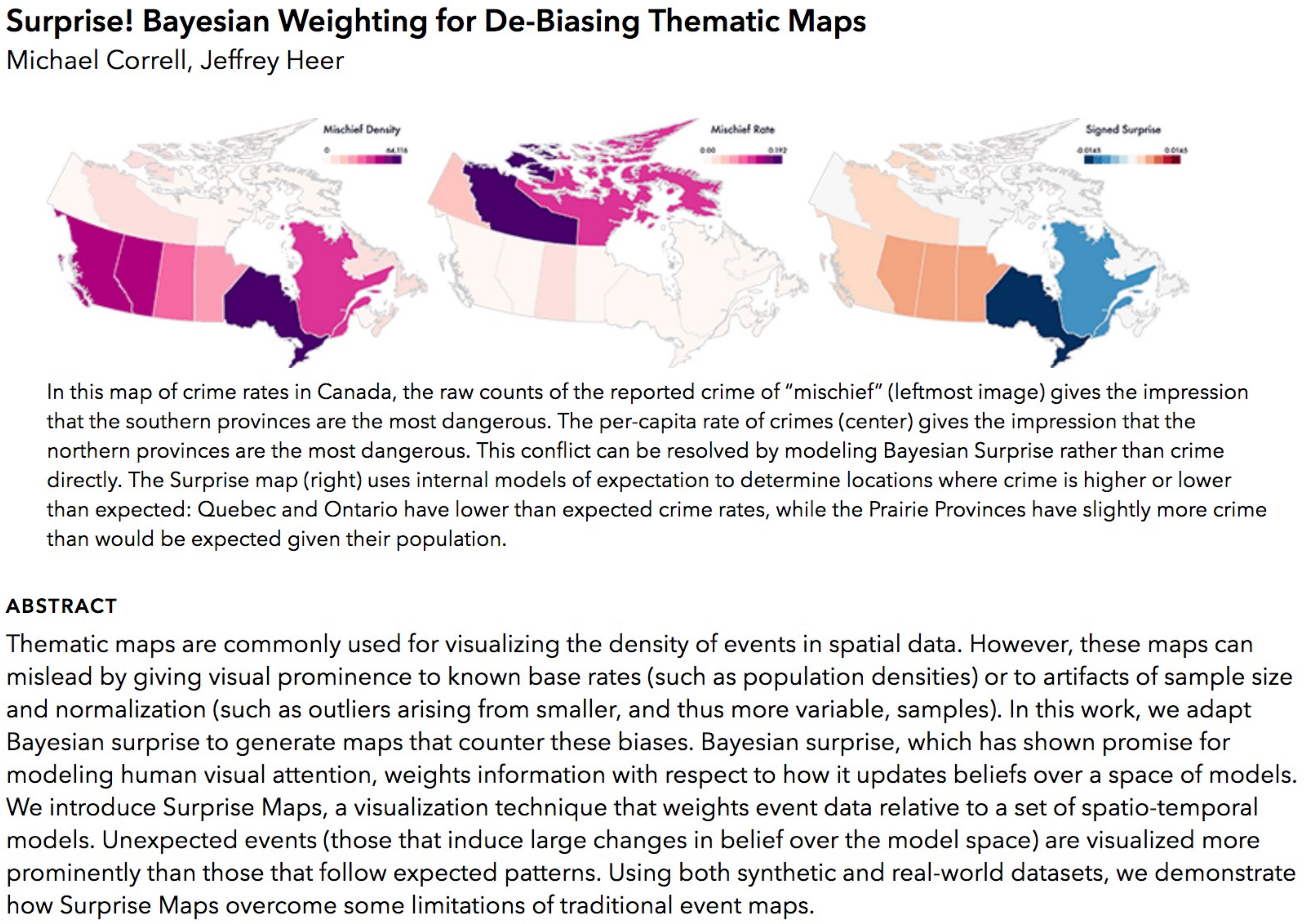
- A representation
  - Can be **physical**, can be **data**, can be **statistical**, can be **mathematical** model
- Mathematical models
  - **process-driven**: explicitly encodes a process using expert knowledge
  - **hypothesis-driven**: compare data to a hypothesis
  - **data-driven**: machine learning (fitting sample data to models)
- Can be used for
  - **prediction**: helping with decisions, what-if scenarios
  - **analysis**: what are the process that operate, how does one or a set of phenomena impact on another

Models: process-driven

- Climatic models based on physics
- Agent-based models (ABMs) or microsimulation models
  - Usually, individual-based models with model of interactions

Models: hypothesis-driven

- Compare with a hypothesis or some kind of expectation
- Null hypothesis testing



Surprise! Bayesian Weighting for De-Biasing Thematic Maps

Michael Correll, Jeffrey Heer

In this map of crime rates in Canada, the raw counts of the reported crime of “mischief” (leftmost image) gives the impression that the southern provinces are the most dangerous. The per-capita rate of crimes (center) gives the impression that the northern provinces are the most dangerous. This conflict can be resolved by modeling Bayesian Surprise rather than crime directly. The Surprise map (right) uses internal models of expectation to determine locations where crime is higher or lower than expected: Quebec and Ontario have lower than expected crime rates, while the Prairie Provinces have slightly more crime than would be expected given their population.

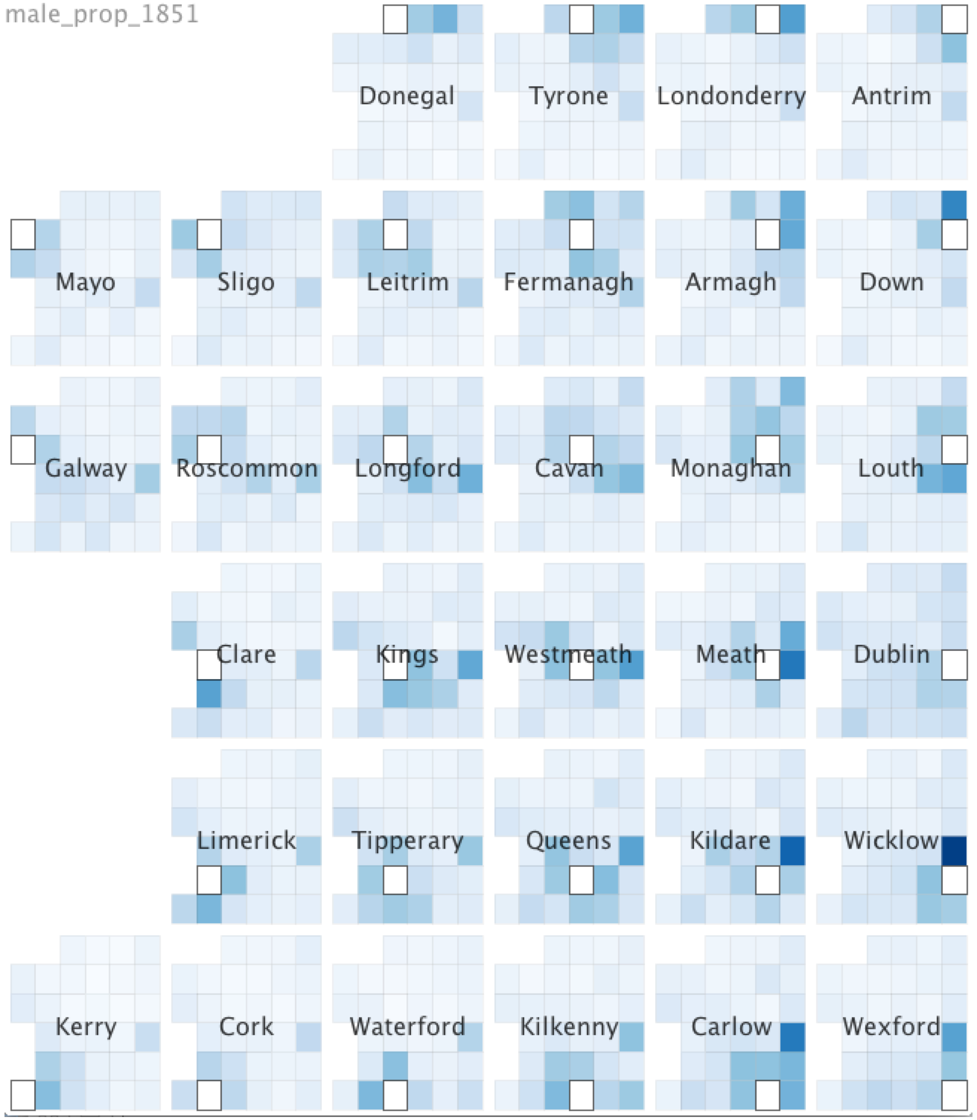
ABSTRACT

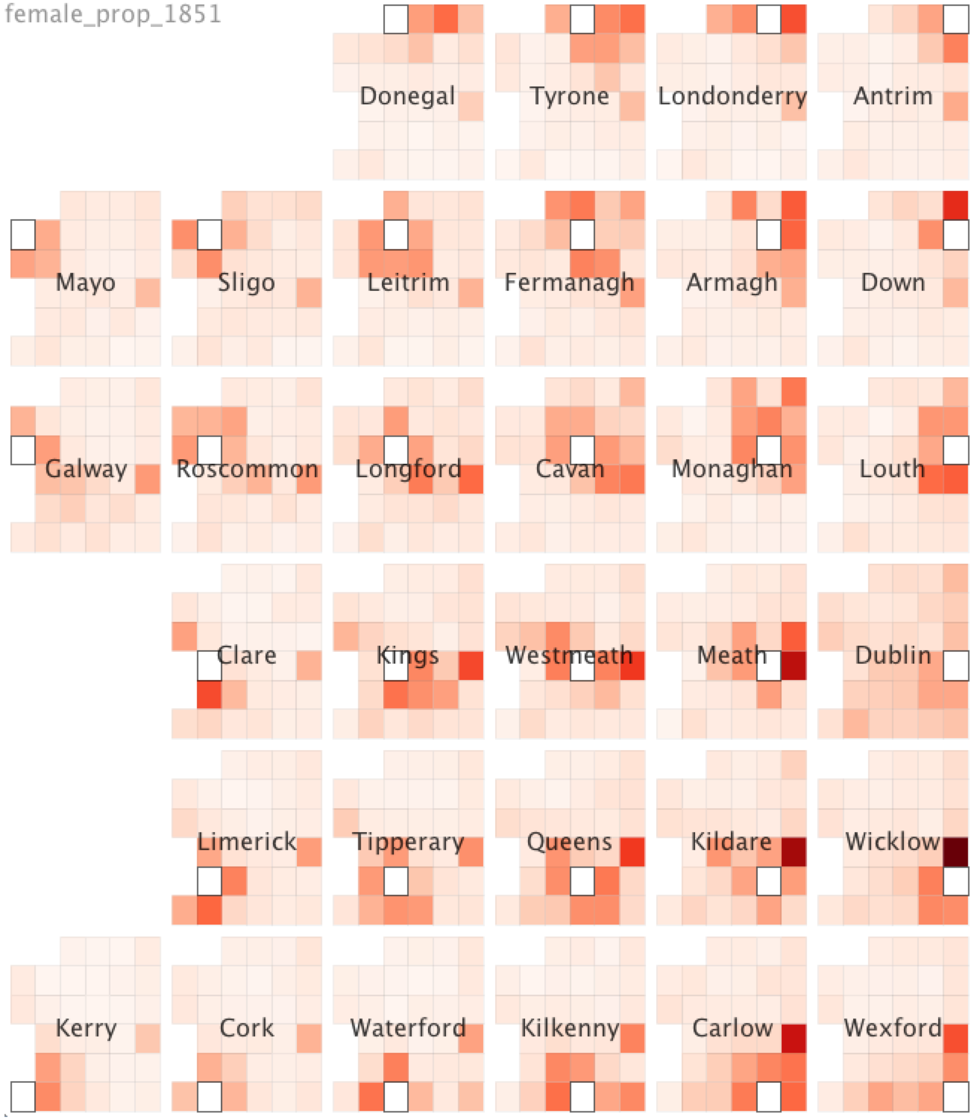
Thematic maps are commonly used for visualizing the density of events in spatial data. However, these maps can mislead by giving visual prominence to known base rates (such as population densities) or to artifacts of sample size and normalization (such as outliers arising from smaller, and thus more variable, samples). In this work, we adapt Bayesian surprise to generate maps that counter these biases. Bayesian surprise, which has shown promise for modeling human visual attention, weights information with respect to how it updates beliefs over a space of models. We introduce Surprise Maps, a visualization technique that weights event data relative to a set of spatio-temporal models. Unexpected events (those that induce large changes in belief over the model space) are visualized more prominently than those that follow expected patterns. Using both synthetic and real-world datasets, we demonstrate how Surprise Maps overcome some limitations of traditional event maps.

http://idl.cs.washington.edu/papers/surprise-maps/

Migration flow of men & women 1851

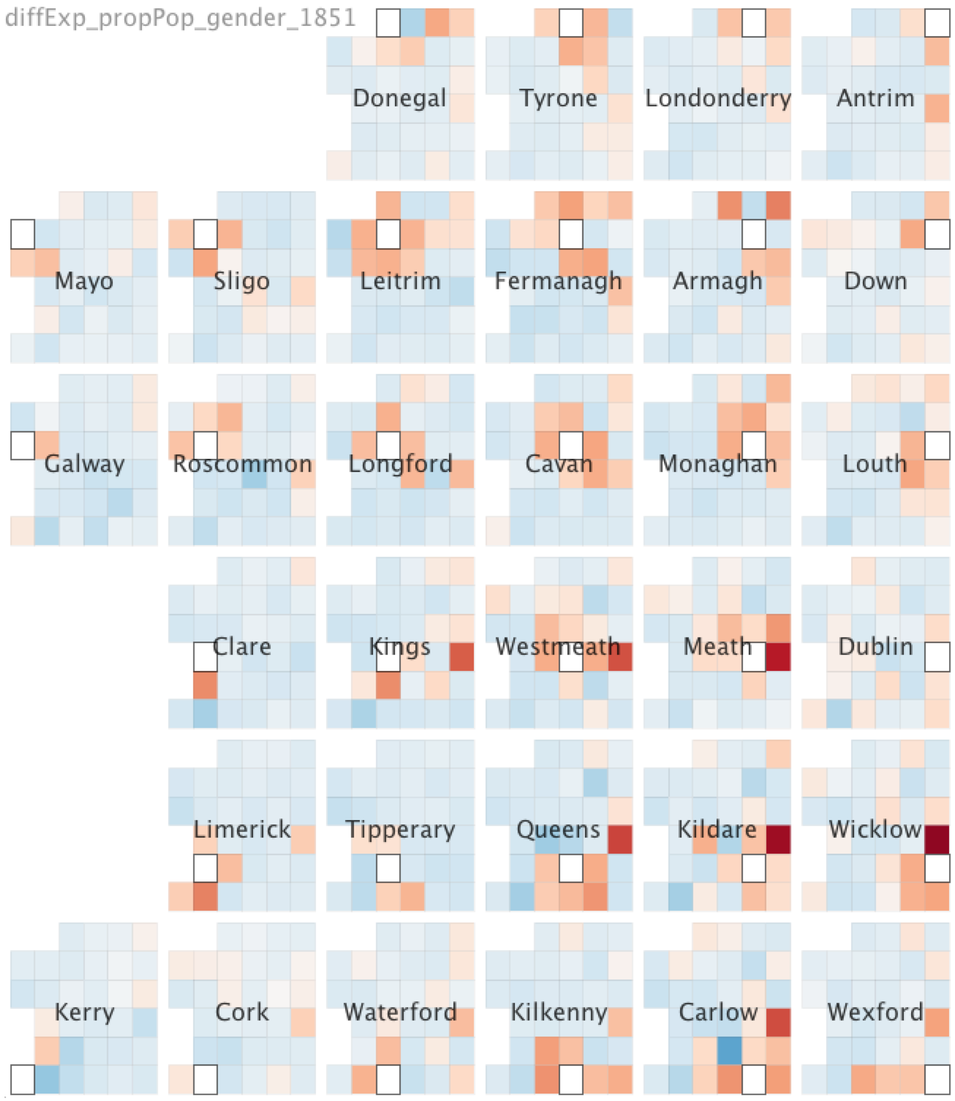
https://openaccess.city.ac.uk/id/eprint/2052/

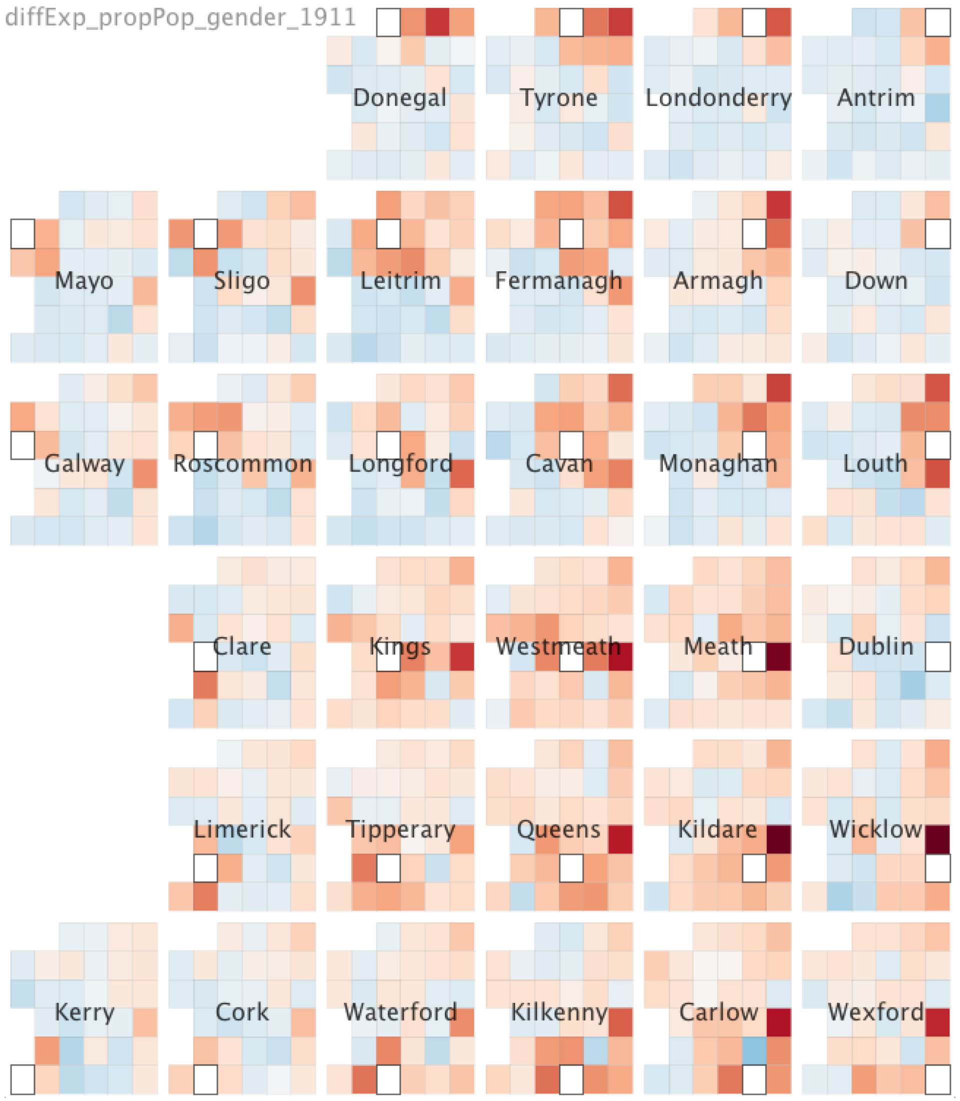




Gender balance residuals in 1851

https://openaccess.city.ac.uk/id/eprint/2052/



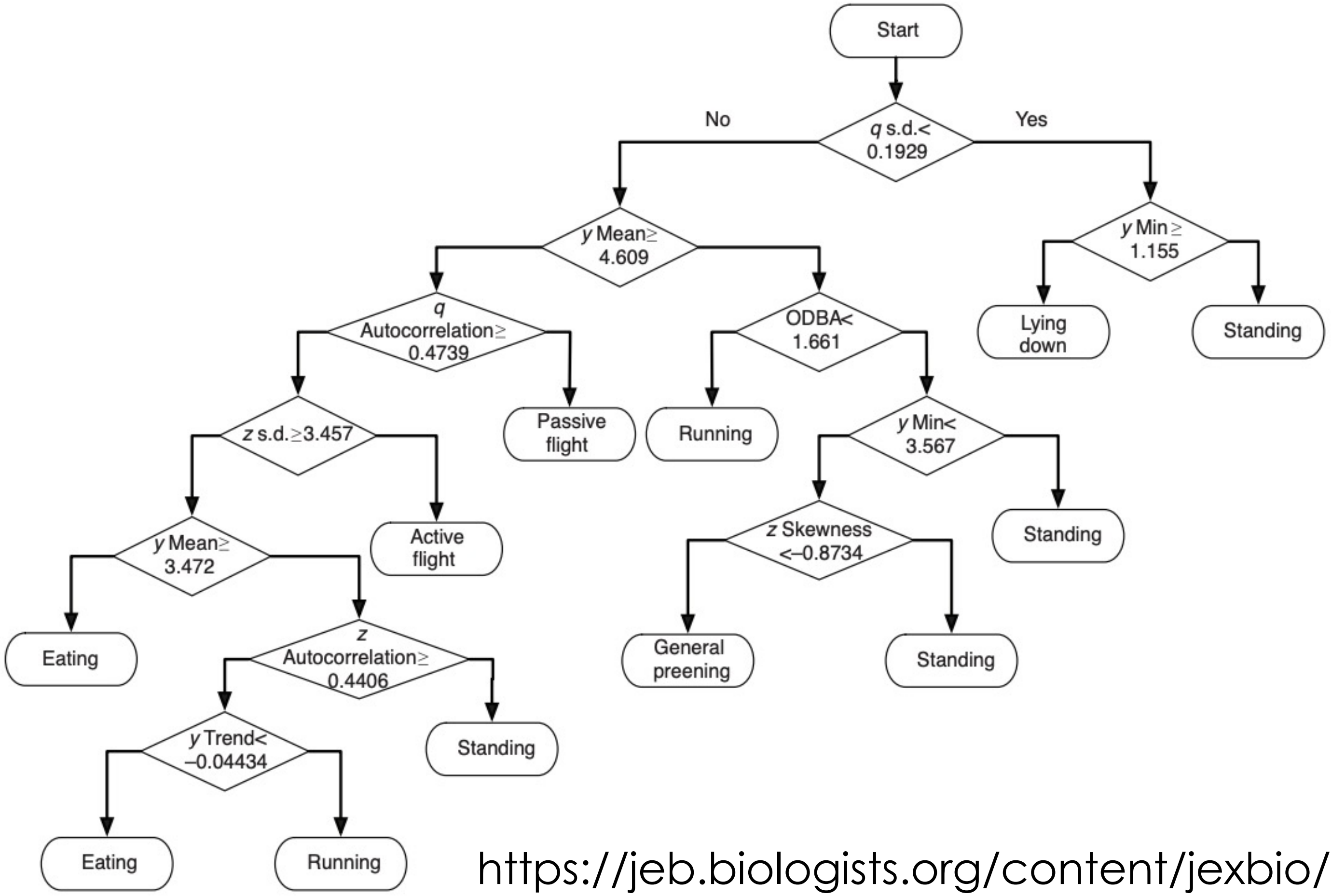


Models: data-driven

- Machine learning
  - Usually **supervised**: we provide data for the relationships to be created
- Different types of model that model:
  - Quantities: regression
  - Categories: classification
    - Scalable vector machines (SVMs), decision trees, logistic regression)

Models: data-driven: classification

- Take labelled data and derive the sequence of thresholds that determine categories with a given probability

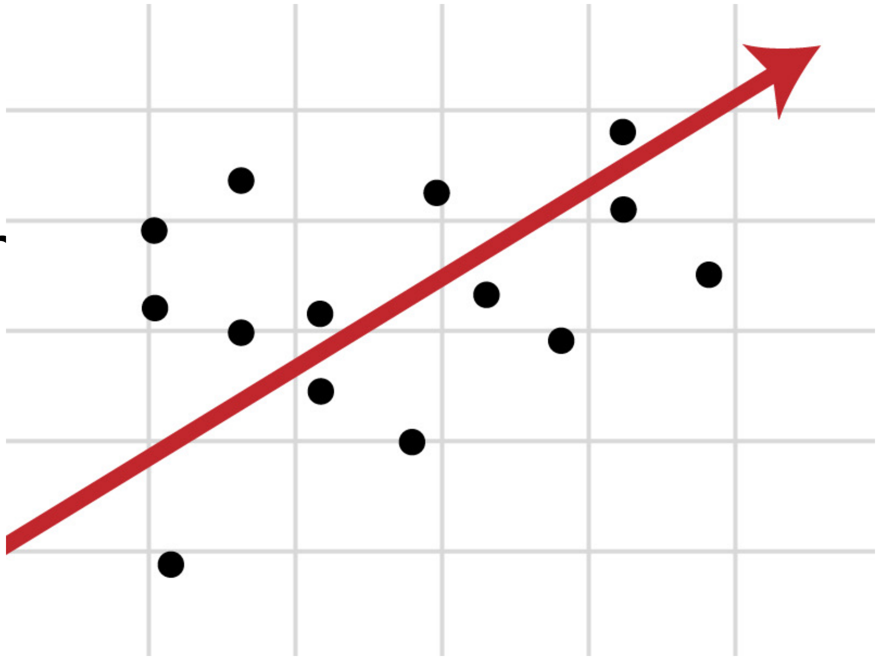


https://jeb.biologists.org/content/jexbio/215/6/986.full.pdf

Models: data-driven: quantities

- Take a model type and derive the **parameters** by fitting sample data to it:
  - dependent variable: the variable you’re predicting
  - independent variable(s): the variable using
- Results in a model that encodes a mathematical relationship between the dependent variable and the independent variable(s)
- Many, many model types
  - Regression, decision trees, neural networks, etc.

$$y = f(x)$$



Models for analysis: Process & hypothesis modelling

- We can set up multiple models
- If our data fits the model output, (maybe) it's a good representation of the process
- If some of the data don't fit can can investigate why not

Models for analysis: Data-driven modelling

- If our data fits the model output, then the estimated parameters are probably valid
  - parameters indicate **which**, **how** and **how strongly** the independent variables can predict the dependent variable
- But
  - needs to be a **good model** for us to do good analysis
    - Good predictive power, even for subdomains
    - Not overfitted, so it’s generalisable
  - the **statistical assumptions** need to be valid e.g. linear/normal
  - independent variables need to be **independent**
  - model should be **interpretable/explainable** (not black-box)
  - we need to **understand how variables relate to the phenomena**

REGRESSION

Regression

- Data-driven statistical processes that **quantify (or classification) relationships** between **independent variable(s)** and a **dependent** one.
  - independent variable X_i
  - unknown parameters β
  - error terms e
  - dependent variable Y_i

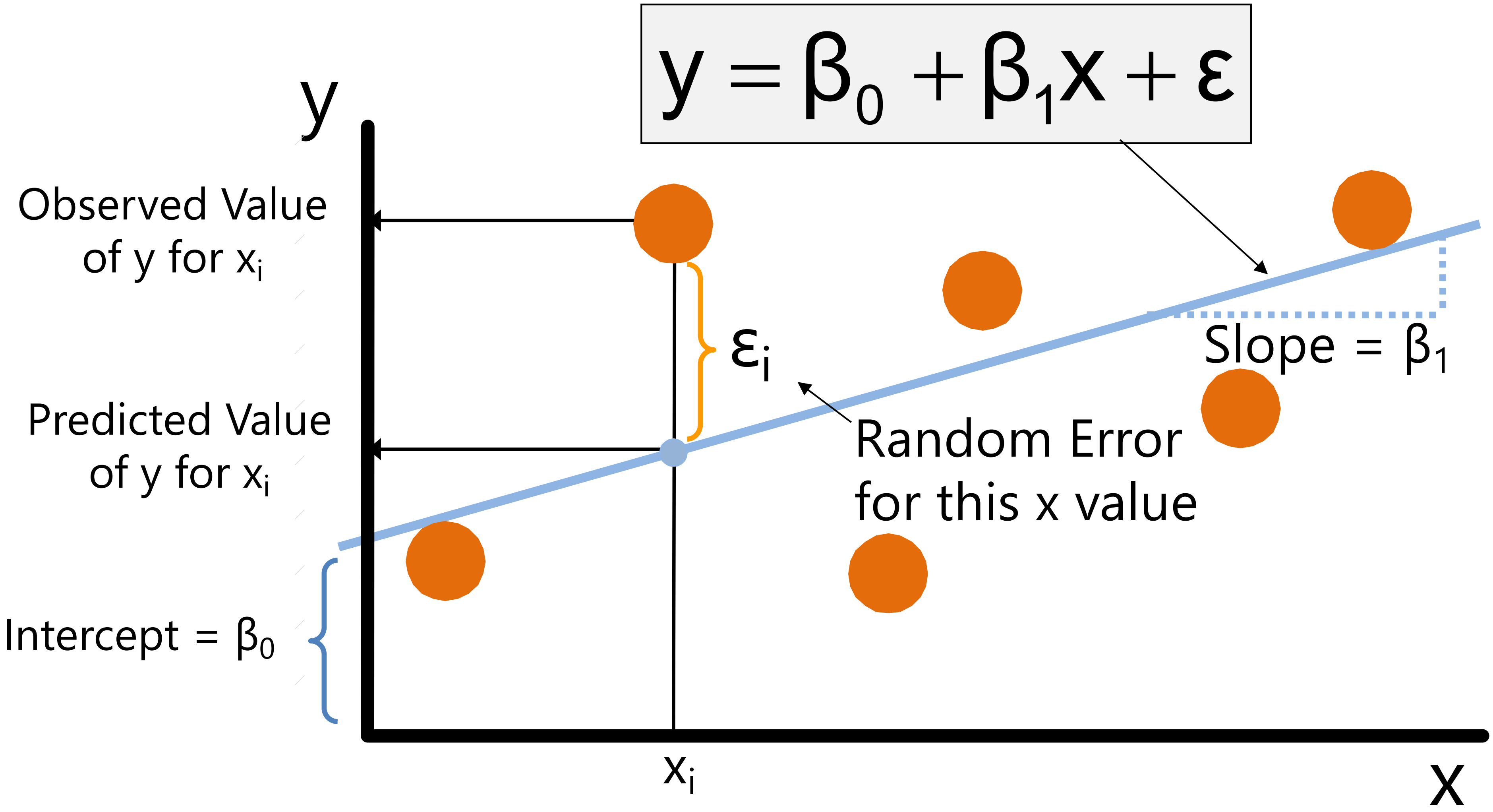
$$Y_i = f(X_i, \beta) + e_i$$

- Assumptions
  - sample is representative of the population
  - independent variables are independent
  - there are no deviations from the model
  - residuals are normally distributed and uncorrelated

REGRESSION: SIMPLE LINEAR REGRESSION

Simple linear regression

- One independent variable
- Linear relationship



$$y = \beta_0 + \beta_1 x + \varepsilon$$

Observed Value of y for x_i

Predicted Value of y for x_i

Intercept = β_0

ε_i

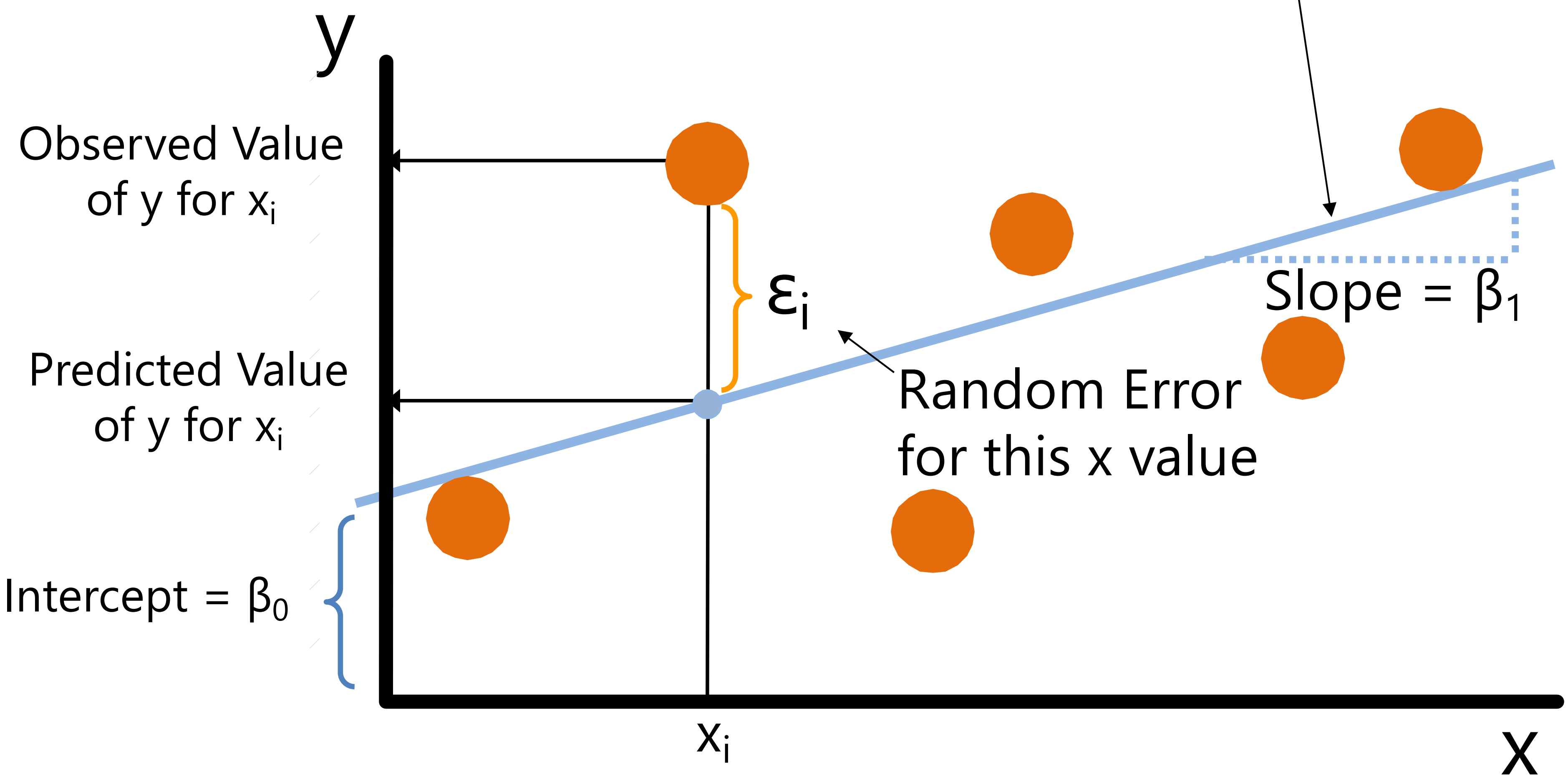
Random Error for this x value

Slope = β_1

Simple linear regression: parameter estimates

- Equation of a straight line
  - β_0: the y-intersept
  - β_1: the slope of the line

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



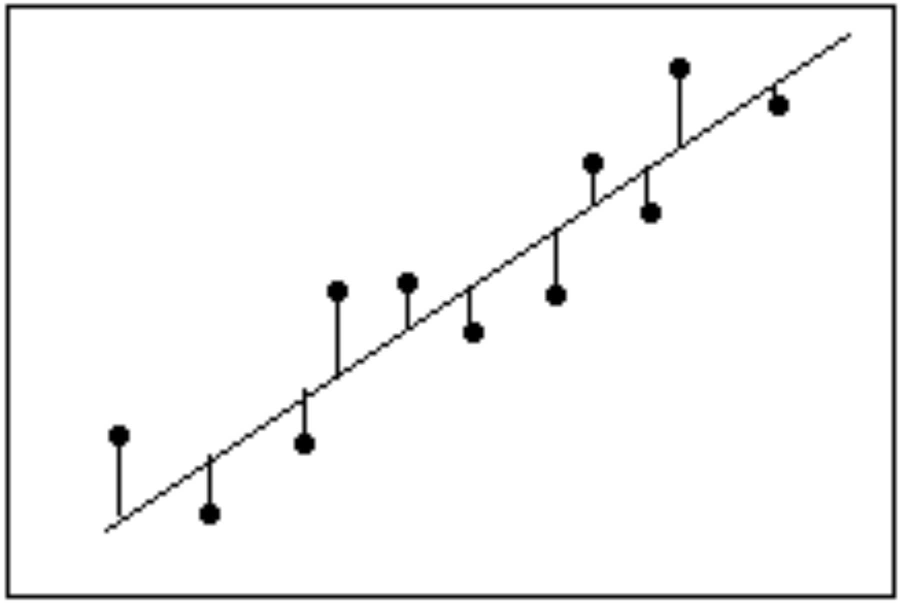
Simple linear regression: interpretation

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- β_0 is the estimated average value of y when the value of x is zero
- β_1 is the estimated change in the average value of y as a result of a one-unit change in x

Linear regression: Ordinary least squares (OLS)

- Method for estimating the unknown parameters in a linear regression model.
- Minimises the **sum of the squares of the differences** between the observed and predicted values



House prices

- What is the relationship between floor area and house price?
  - independent variable (x): floor area (square feet)
  - dependent variable (y): house price (1K$)

House prices: data

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

House prices: interpreting β_0 (intercept)

$$\text{house price} = \boxed{98.24833} + 0.10977 \text{ x (square feet)}$$

- β_0 is the estimated price (*Y*) when the floor area (*X*) is zero
  - – it's price of a house of zero size!
  - – indicates that $98,248.33 of the model variation is not related to floor area

House prices: interpreting β_1 (slope)

$$\text{house price} = 98.24833 + \boxed{0.10977}\ \text{x (square feet)}$$

- β_1 is estimated change in price (*Y*) if the as a result of one-unit of change in floor area (*X*)
  - Here, 0.10977 tells us that the average value of a house increases by 0.10977 (**$109.77)** for each additional one square foot of area

Sum of squares

- Most assessments of regression model quality are based on these

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$$

Total sum of Squares

Variation in the observations (data)

Sum of Squares Error

Difference between observations and estimate

Sum of Squares Regression

Variation in the estimates

$$SST = \sum (y - \bar{y})^2$$

$$SSE = \sum (y - \hat{y})^2$$

$$SSR = \sum (\hat{y} - \bar{y})^2$$

where:

$\bar{y}$ = Average value of the dependent variable

y = Observed values of the dependent variable

$\hat{y}$ = Estimated value of y for the given x value

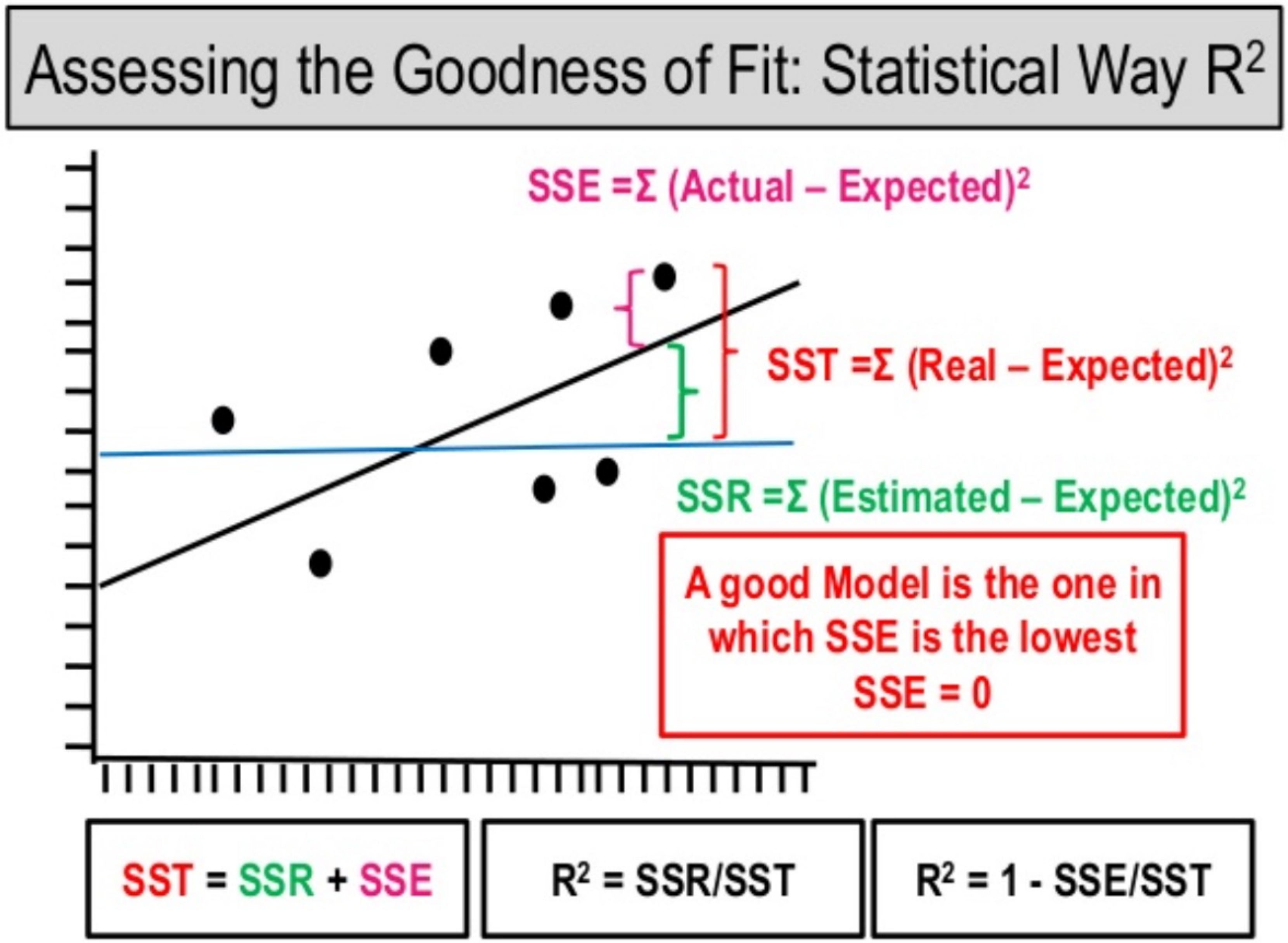
Sum of squares

- Mean square error
- Root mean square error

Explained variance: R^2

$$R^2 = \frac{SSR}{SST} = \frac{\mathit{sum}\text{ of squares explained by regression}}{\mathit{total}\text{ sum of squares}}$$

- Coefficient of Determination
- Quantifies the proportion of variance in the dependent variable (e.g. house price) that is due to the independent variable (e.g. floor area)
  - 1 is perfect fit: all the variation can be explained by inputs
  - 0.5: half of the variation can be explained by inputs
  - 0: none of the variation can be explained by inputs (no better than the mean)
  - negative values: worse than mean model (wrong model)
- **This quantifies how much of the variation in the**



Assessing the Goodness of Fit: Statistical Way R^2

SSE =Σ (Actual – Expected)^2

SST =Σ (Real – Expected)^2

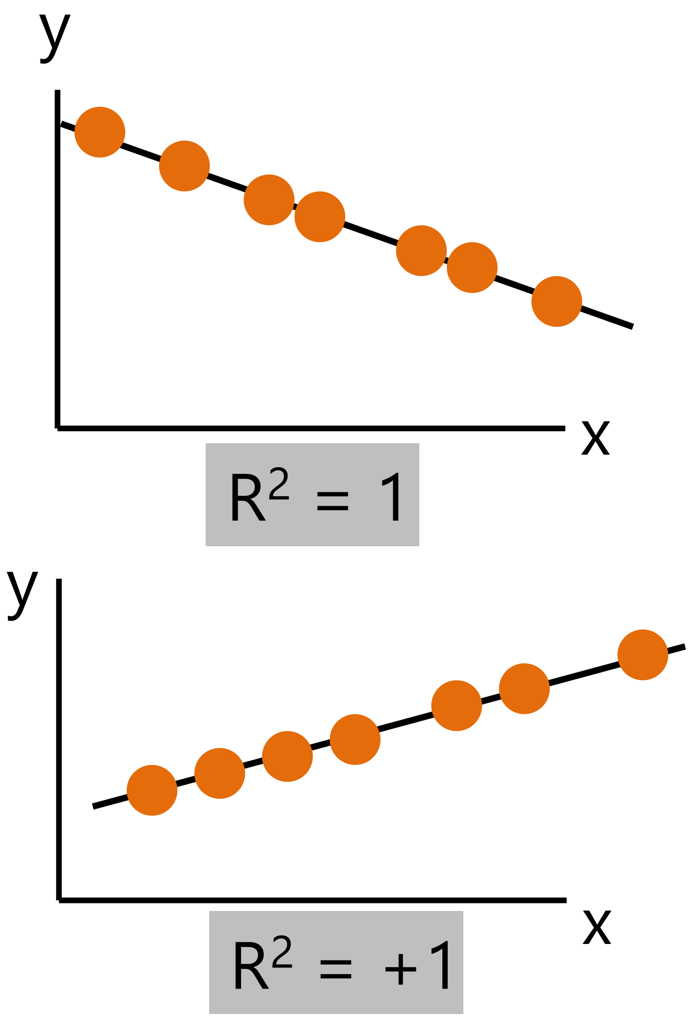
SSR =Σ (Estimated – Expected)^2

A good Model is the one in which SSE is the lowest SSE = 0

SST = SSR + SSE	R^2 = SSR/SST	R^2 = 1 - SSE/SST

https://www.slideshare.net/789667/corrleation-and-regression

R^2 Values

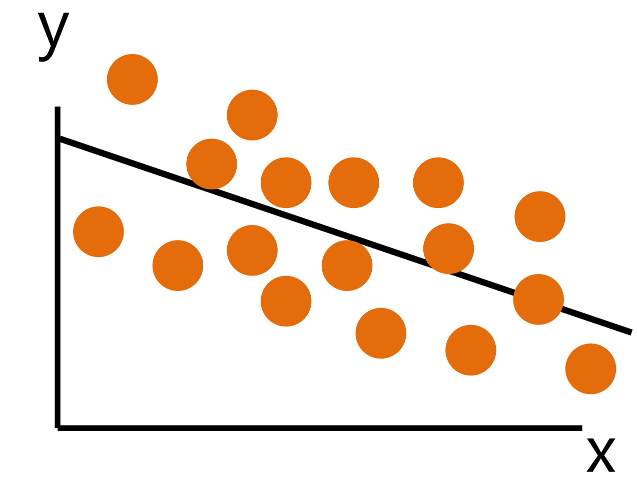


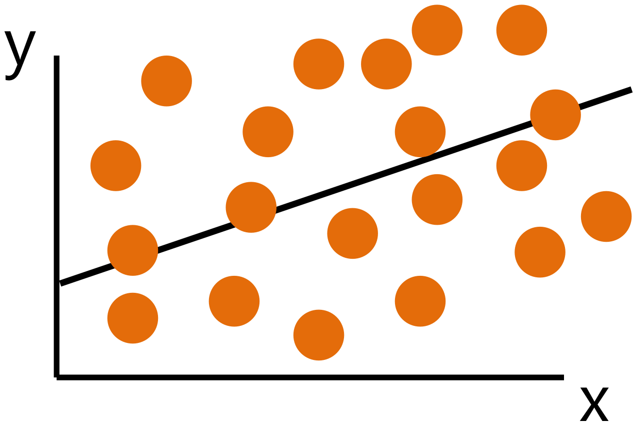
$R^2 = 1$

Perfect linear relationship between x and y:

100% of the variation in y is explained by variation in x

R^2 Values



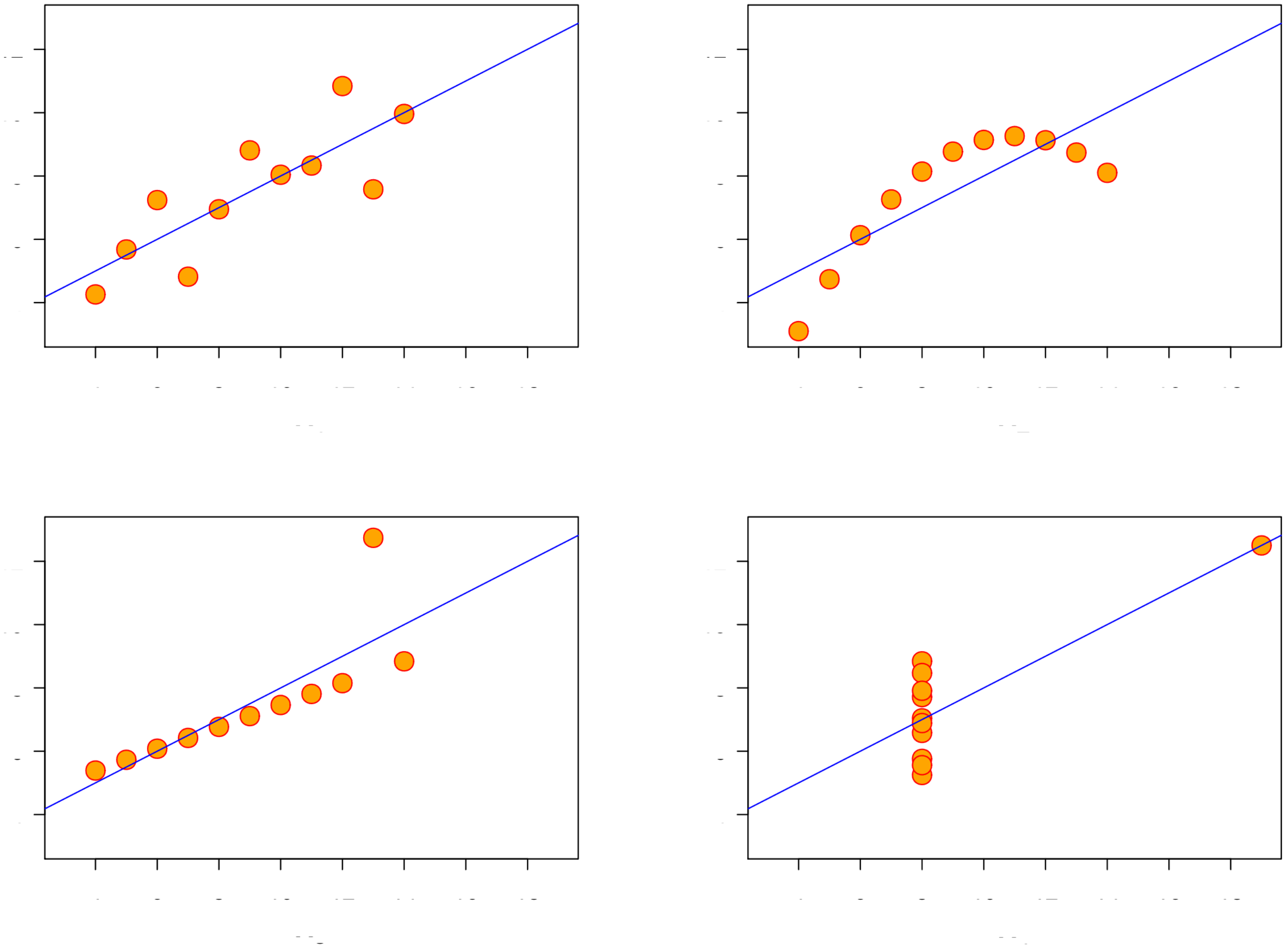


$0 < R^2 < 1$

Weaker linear relationship between x and y:

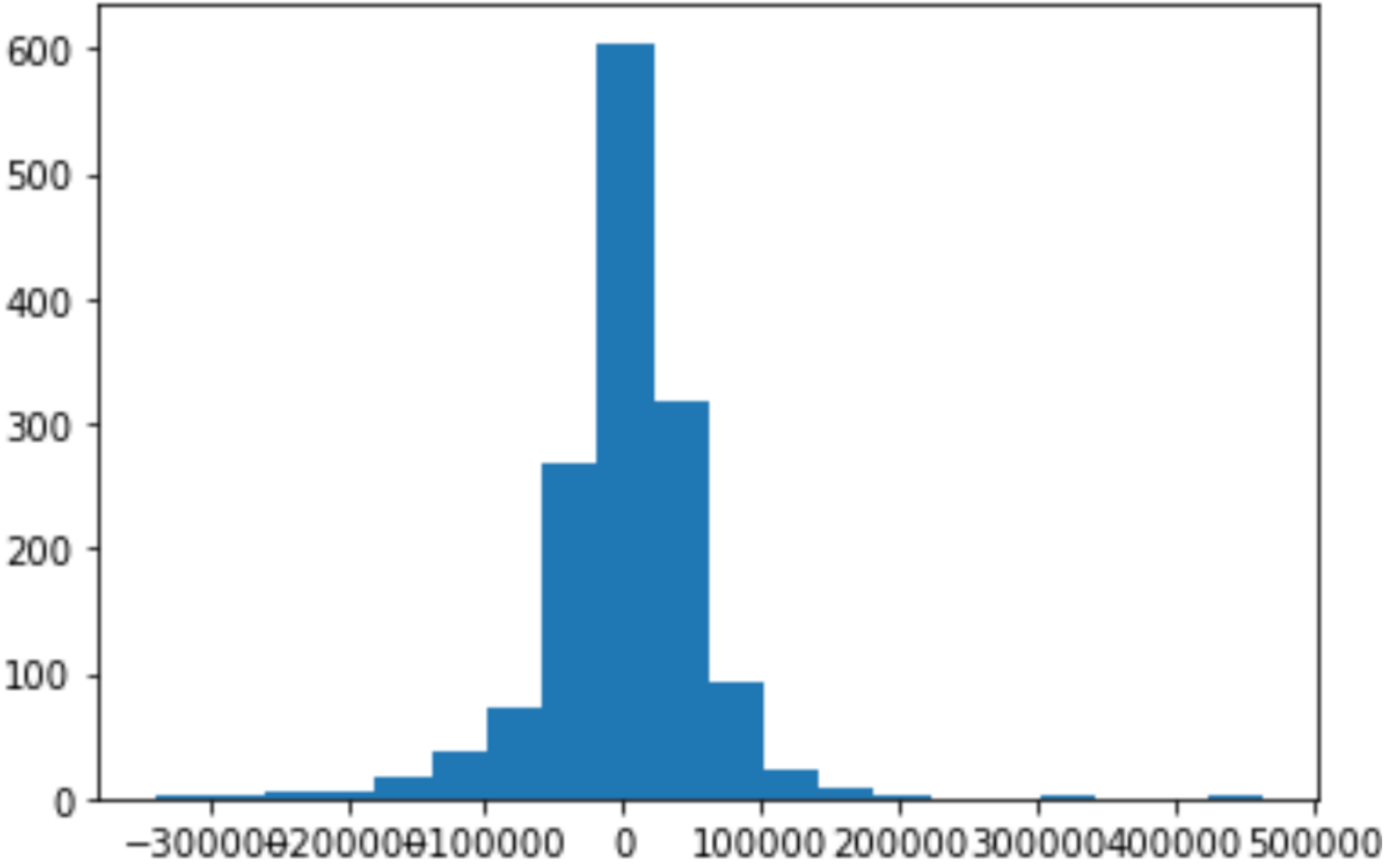
Some but not all of the variation in y is explained by variation in x

Remember that these are global measures



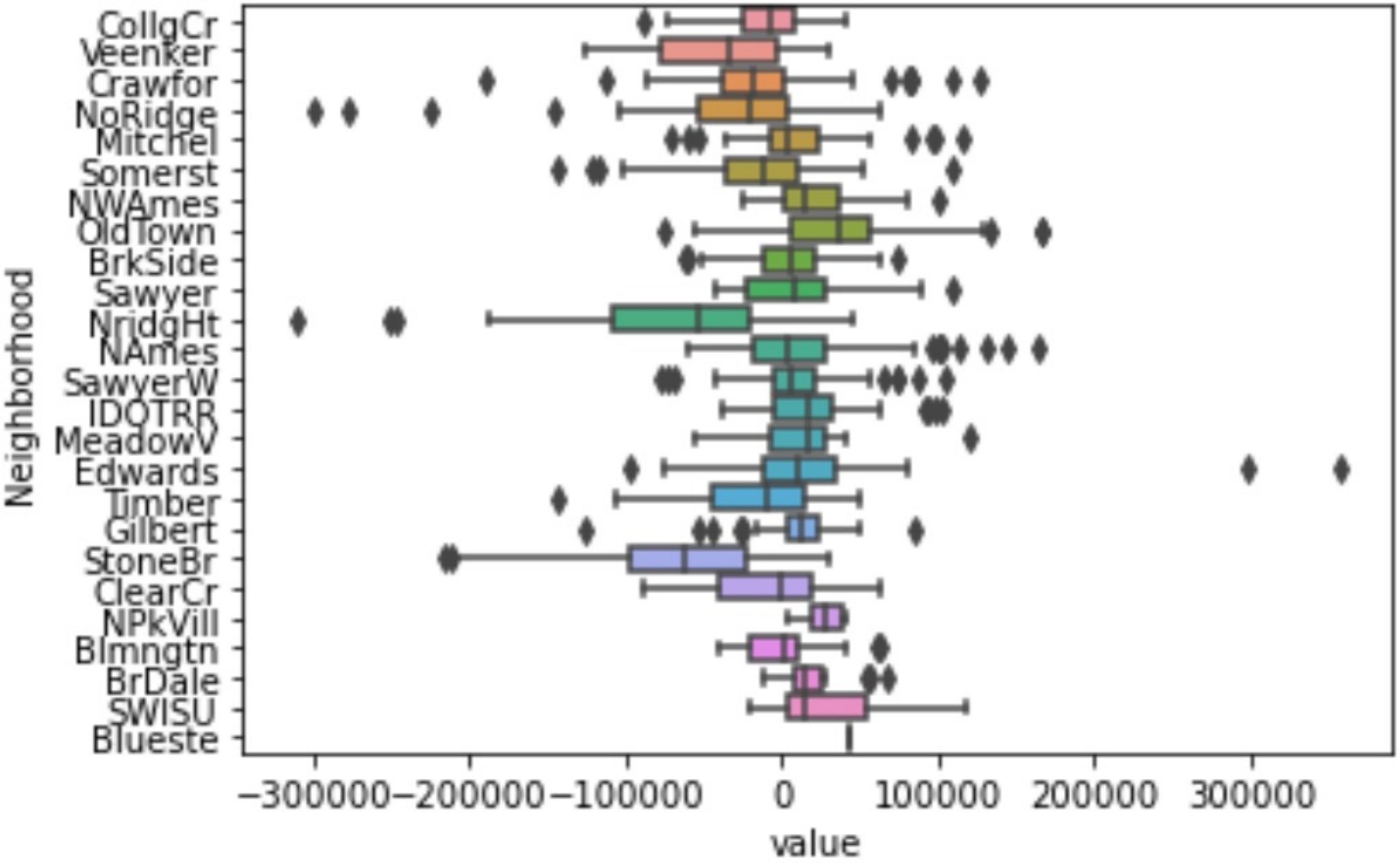
Analyse the distribution of the residuals

- The difference between the predicted and observed
  - this is normally distributed with a mean of zero – as it



Analyse the distribution of the residuals

- The difference between the predicted and observed
  -



Remember the assumptions

- Assumptions
  - Linear relationship
  - Homoscedasticity: variance of residuals is the same for all values of x
  - Independence: observations are independent (sample randomly drawn from a distribution
  - Normality: the residuals should be normally distributed
- Not a deal-breaker if some are missing, but we should be aware when interpreting

Non-linear relationships in linear models

- Use linear models where possible because they are easy to build and interpret
- You can transform any of your variables…
  - log, squared, cubed, square-root
- …and use them in your model
- (There are also non-linear regression methods)

MULTIPLE LINEAR REGRESSION

Multiple linear regression

- A generalization with multiple independent variables. Very common!

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i$$

- β_0 is the estimated average value of y when the values of x_1 and y_2 are zero
- β_1 is the estimated change in the average value of y as a result of a one-unit change in x_1 with everything else constant
- β_x is the estimated change in the average value of y as a result of a one-unit change in x_x with everything else constant

Tip: don’t use too many variables

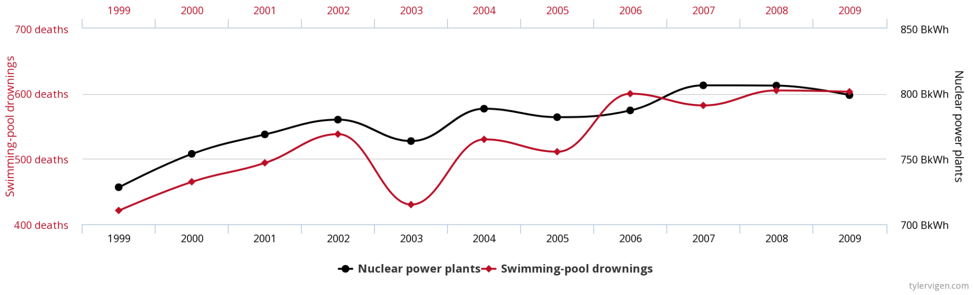
- Artificially inflates R2 (use adjusted R2 – see later)
- Create **parsimonious** models **because**
  - reduces chance of unexpected interactions between variables
  - makes model easier to interpret
- Create **parsimonious** models **by**
  - Only using variables that have explanatory power
  - Only using variables that make sense for your application
  - Techniques that help select variables may help
    - Lasso and Ridge regression include such methods

Tip: use sensible variables

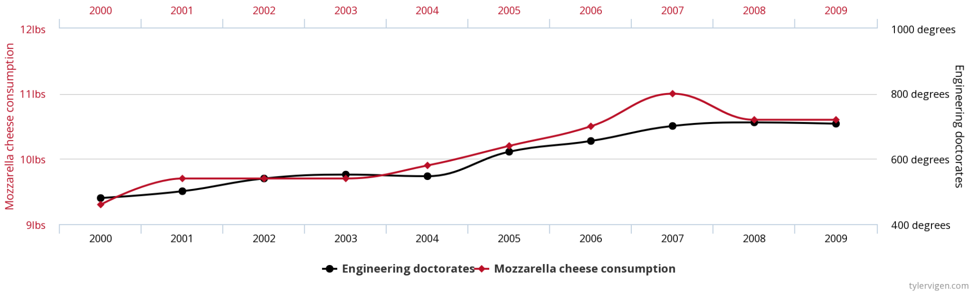
- For models built for analysis ensure the variables make sense for your applications
- If you want make causal inferences, you need to understand how the variable may have a causal effect
  - – directly
  - – a proxy for
  - – a confounder (no direct causal effect)

https://www.tylervigen.com/spurious-correlatic

Number people who drowned while in a swimming-pool
correlates with
Power generated by US nuclear power plants



Per capita consumption of mozzarella cheese
correlates with
Civil engineering doctorates awarded



Tip: avoid variables that correlate to each other

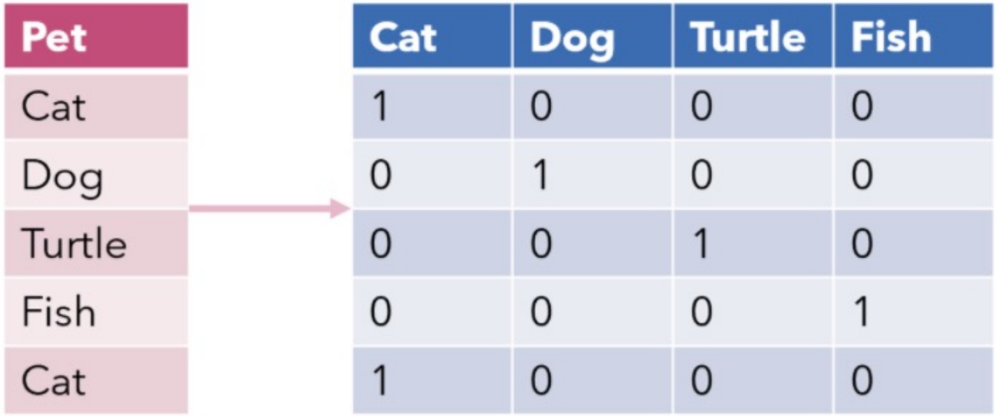
- Independent variables should be **independent**
  - If not, they interact in weird ways, making interpretation more tricky
  - Covarying variables do not meaningfully improve the model because they describe the same variation
- Use correlation to help determine which ones to include

Tip: Consider standardizing your variables

- The variable coefficent slopes are in the same units as the variables, so not directly comparable with with other
- Consider putting all the variables on the same scale so that you can directly compare coefficients. This helps determine which variables explain more of the variation

Tip: dealing categorical independent variables

- **Inputs:** Regression also works with categorical independent variables, but need to transform them to binary ones
  - Dummy variables
  - One **new** binary/Boolean (true/false) variable **for each category**
  - Panda's `get_dummies()` function can create these for you
  - Coefficients either apply or don't



Pet
Cat
Dog
Turtle
Fish
Cat

Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0

- **Split dataset:** split your dataset and build a separate model on each partition
- **Split results**: compare model output by category

Adjusted R^2

- Where you have multiple independent variables, R_2 can get artificially high
  - because it's not affected by poor predictors
- The **adjusted R2** penalises R2 where more variables are added, so better to use for multiple regression, especially for large numbers of variables

P-values

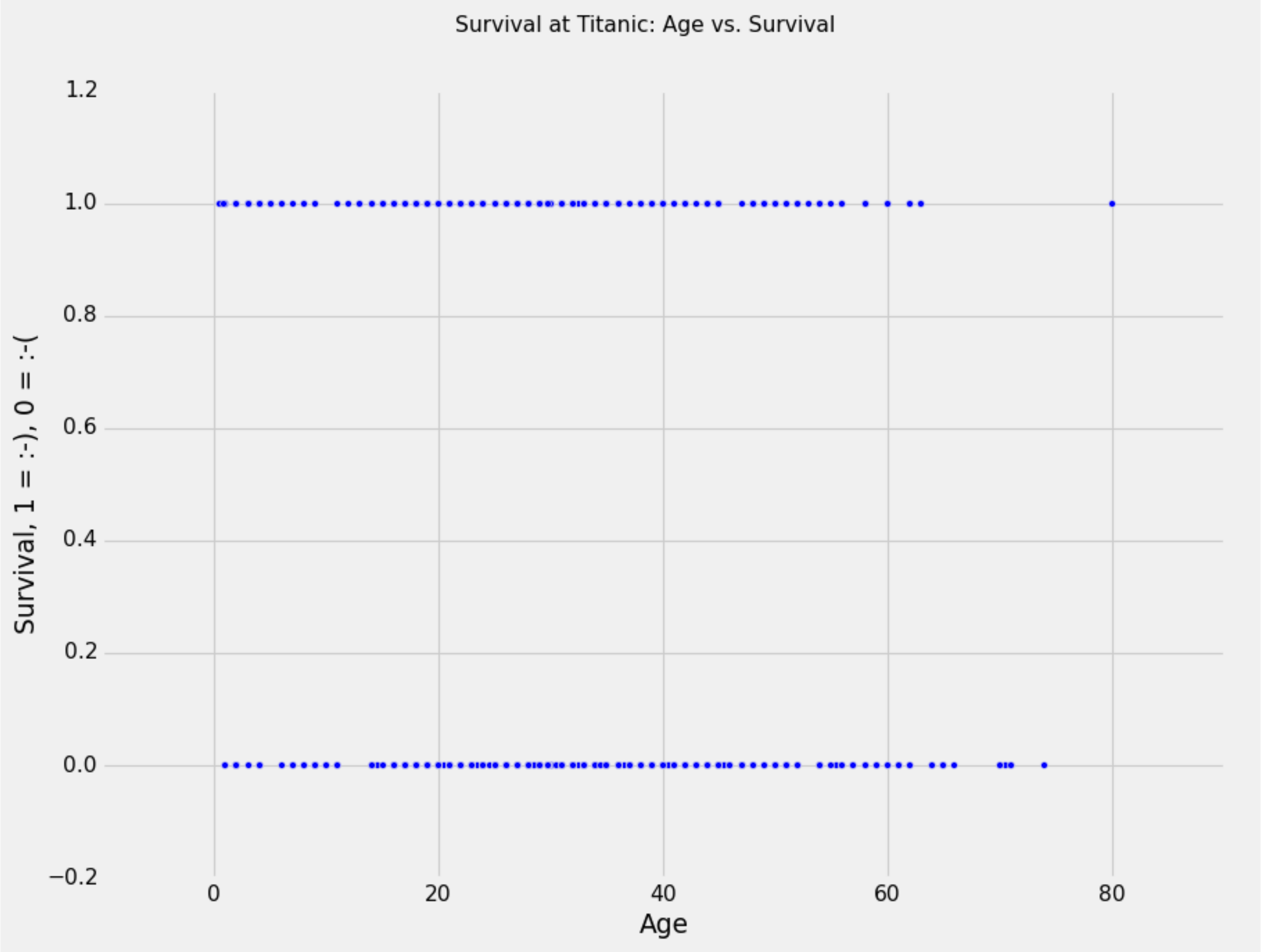
- Each coefficient has an associated p-value
  - whether the relationship is likely to exist in the wider population
- If less than 0.05 (the convention for statistical significance) then it is significant
  - there is sufficient evidence to reject the null hypothesis, meaning that the relationship is likely to occur in the population

LOGISTIC REGRESSION

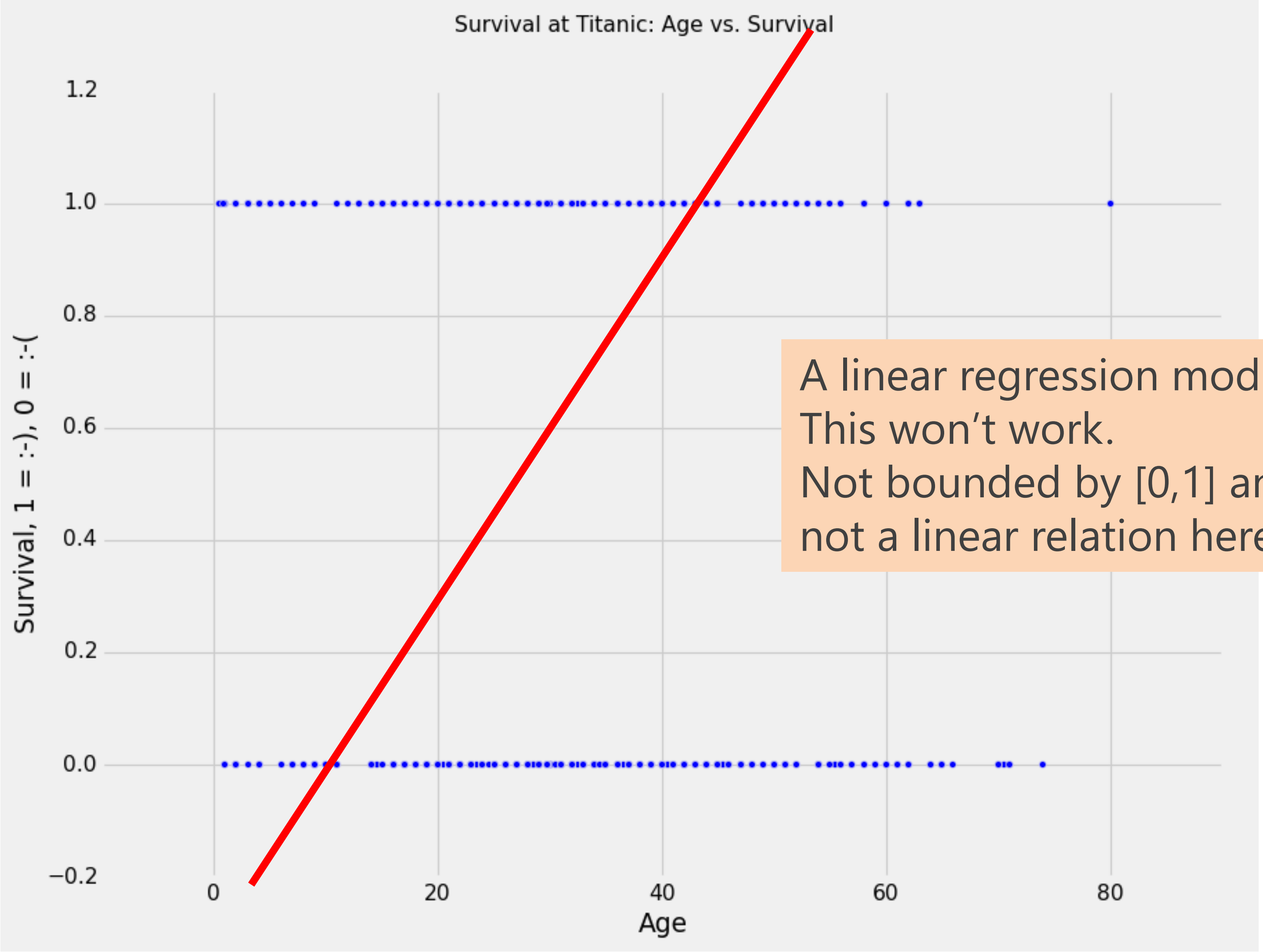
Logistic regression

- Where we have a binary dependent variable
  - yes/no, pass/fail, etc
  - Outcome is a probability 0 or 1
- Multinomial versions (more than 2 categories)

Titanic Survival



Titanic Survival: want to estimate survival



A linear regression model? This won’t work.
Not bounded by [0,1] and not a linear relation here.

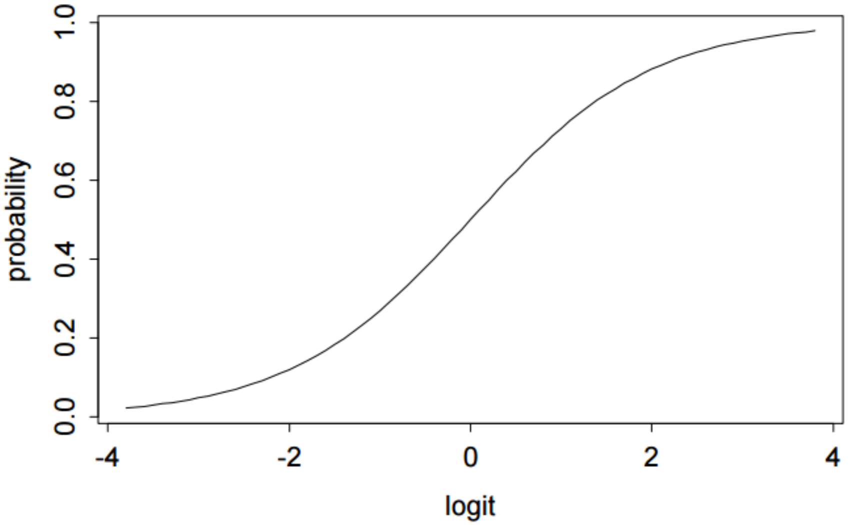
Logistic regression

- Estimate with a logistic function instea $\mathrm{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i}$

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + x \cdot \beta$$

- And solve for p(x):

$$p(x; b, w) = \frac{e^{\beta_0 + x \cdot \beta}}{1 + e^{\beta_0 + x \cdot \beta}} = \frac{1}{1 + e^{-(\beta_0 + x \cdot \beta)}}$$



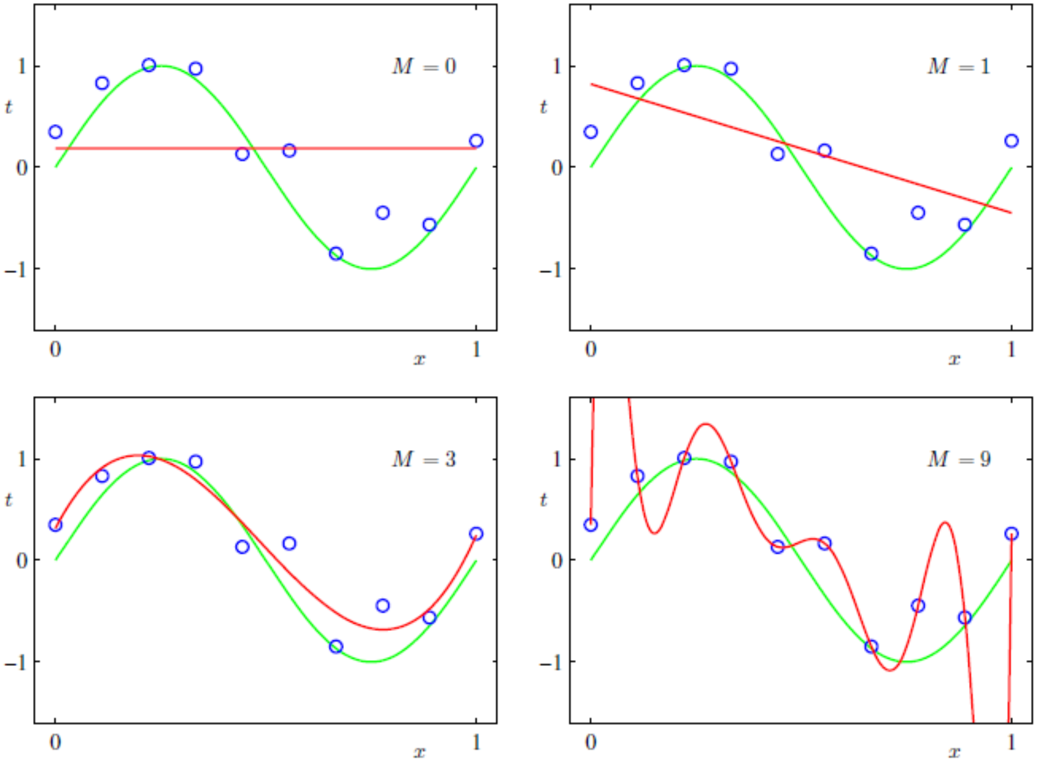
- The logit of the probability π_i, rather than the probability itself, follows a linear model
- You end up with a classifier:

Y = 1 when p ≥ 0.5 and Y = 0 when p < 0.5

OTHER TYPES OF MODEL

Other types of regression

- **Ridge regression:** where independent variables co-vary
- **Lasso regression:** reduces the coefficients of many variables to zero, effectively removing from the models making it easier to interpret
- **Polynomial regressio** Uses an *n*th degree polynomial
  - Be careful not to over-f

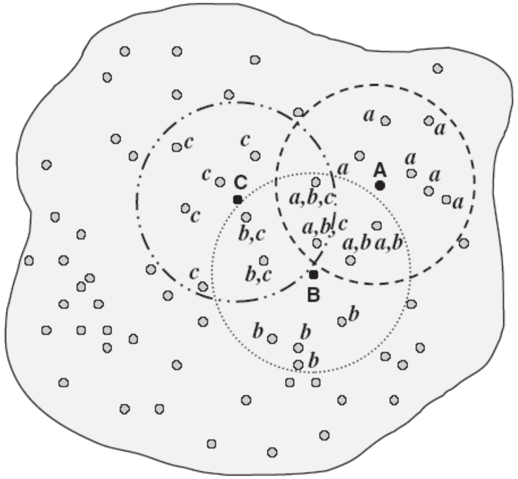


Other types of dependence analysis

- Time-series data
  - Data that has a temporal aspect
  - **Cross-correlation**: Find the correlation between two time series as a **function of time difference** between them

[Chatfield, Chris. *The analysis of time series: an introduction*. CRC press, 2013.]

- Spatial data
  - Relations might vary geographically
  - Spatial auto-correlation
  - Geographically weighted regression



[Lloyd, Christopher D. *Local models for spatial analysis*. CRC Press, 2010.]

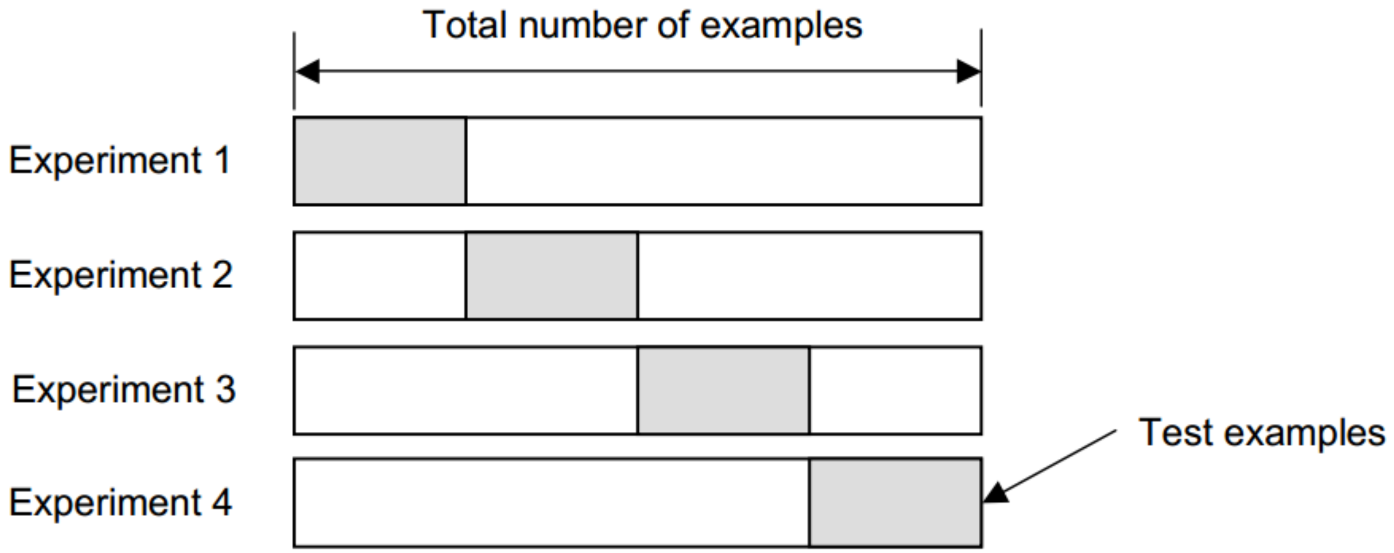
VALIDATING MODELS (AND AVOIDING OVERFITTING)

Cross-validation

- We only have a sample of population data
- We should try to avoid **over-fitting** our model
  - Fit well to our (training data), but not to other randomly sampled data (implication being that it **won't fit to the population** and so it's **not generalisable**)
  - But we only have our sample data to do on
- **Cross-validation** are strategies for helping determine whether a model will generalise to another sample
  - **Holdout:** a **training set** and a **testing set**
  - **K-fold:** partition the data in k random/stratified equally-sized partitions, build k models on each partition, validate with rest

K-Fold

- Choosing k
  - High k: costly to compute
  - 10 is common
- Good tutorial
  - https://machinelearningmastery.com/k-fold-cross-validation/



Ensemble learning

- Use multiple learning algorithms to obtain better performance
- Evaluate the **stability** of the results
- Address overfitting
- Generalizable and robust model
- Parameter optimization

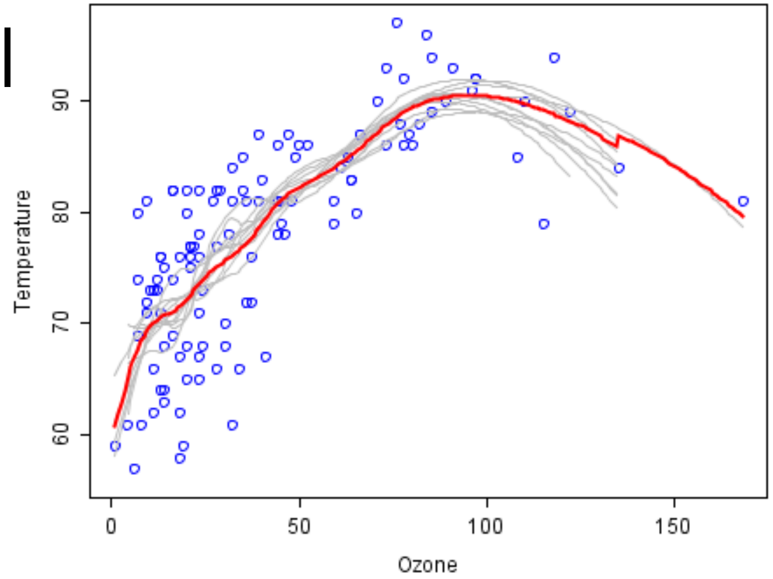
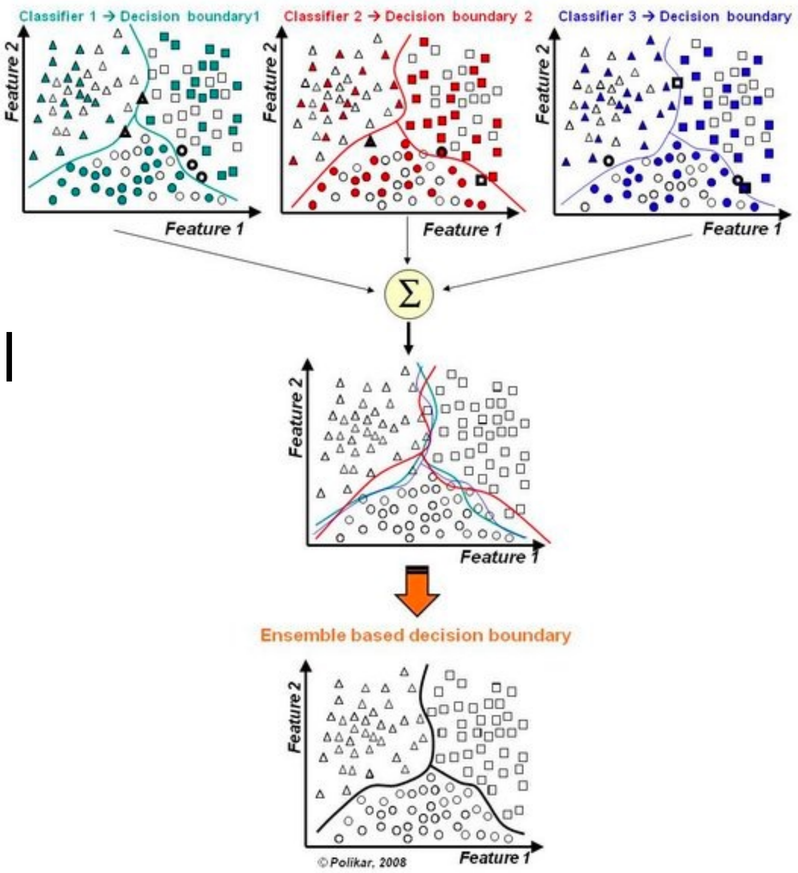


Image from: http://en.wikipedia.org/wiki/Bootstrap_aggregating

Ensemble learning

Models that exhibit diversity in decisions

- Different methods:
  - **Bagging** (Bootstrap Aggregating)
    Create models on samples
    Aggregate models
  - **Boosting**
    - Build a model & evaluate
    - Build another (or more) model where the first fails
    - Merge these "weak" models
  - **Stacking**
    - Layered classifiers


Classifier 1 → Decision boundary1
Classifier 2 → Decision boundary 2
Classifier 3 → Decision boundary 3
Feature 2
Feature 1
Σ
Ensemble based decision boundary
© Polikar, 2008

Finally, combine the results :

- Algebraic ways (mean, product, ..)
- Majority voting

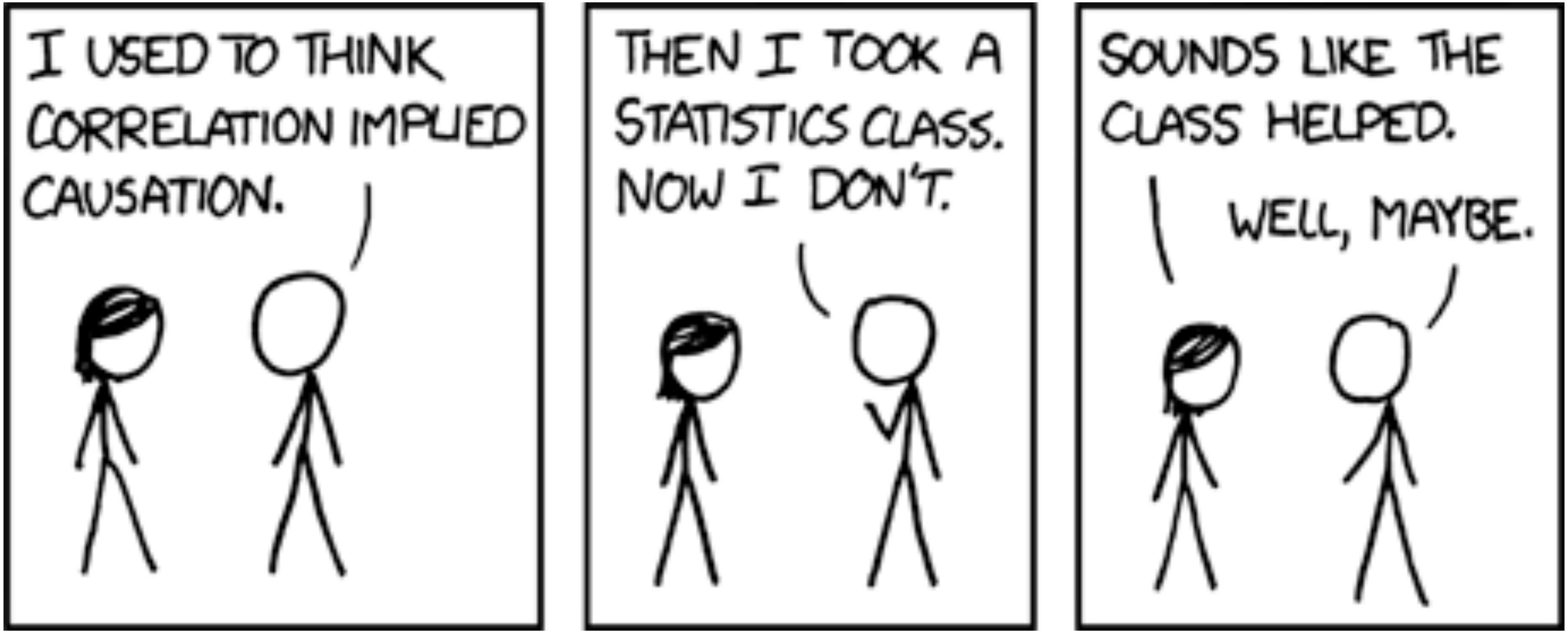
http://www.scholarpedia.org/article/Ensemble_learning

Validation metrics

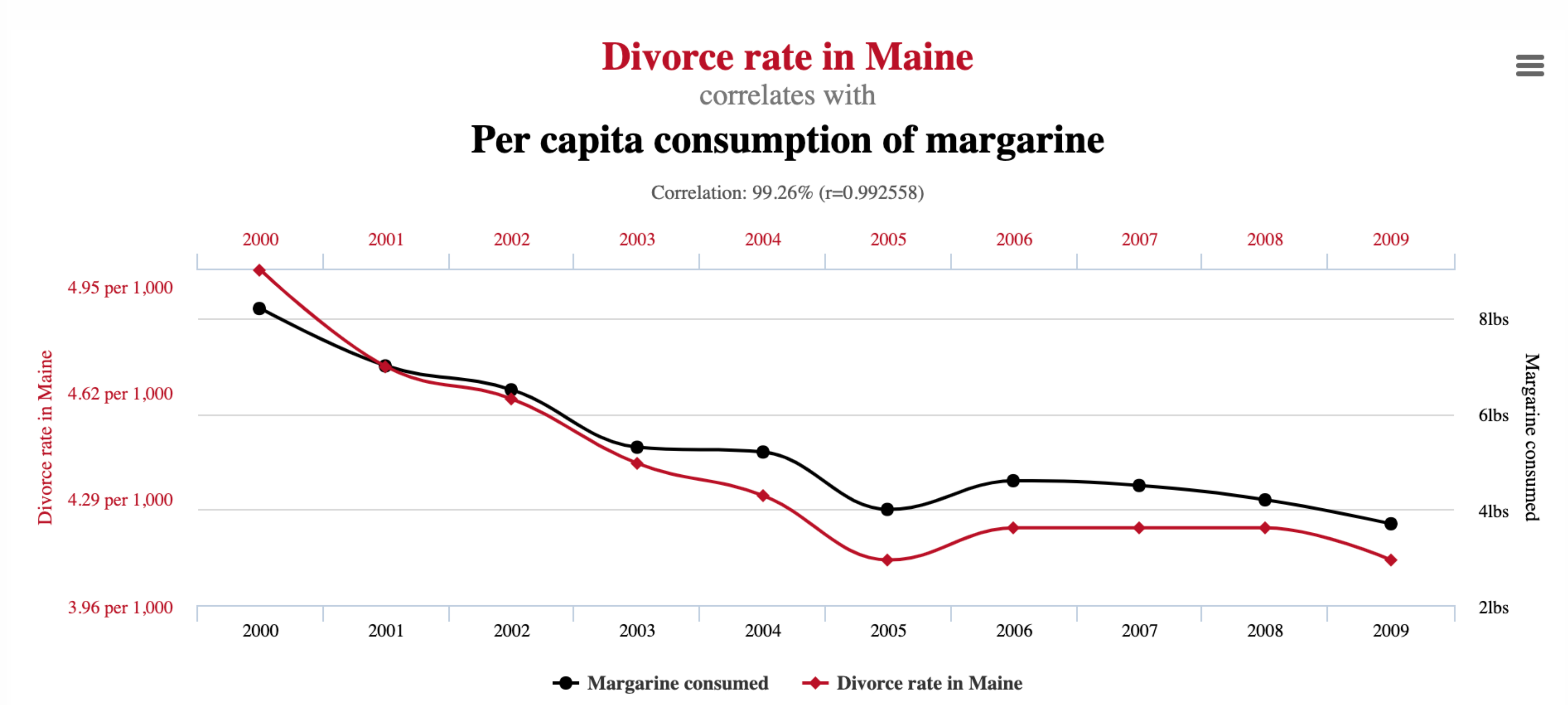
- Continuous dependent variables
  - model fit
  - Analysis of residuals (including stratified analysis)
- Categorical
  - Accuracy
  - Accuracy by model sub-space (confusion matrix)
  - Precision, F1, ROC, etc
- More in Machine Learning!

CAUSAL THINKING

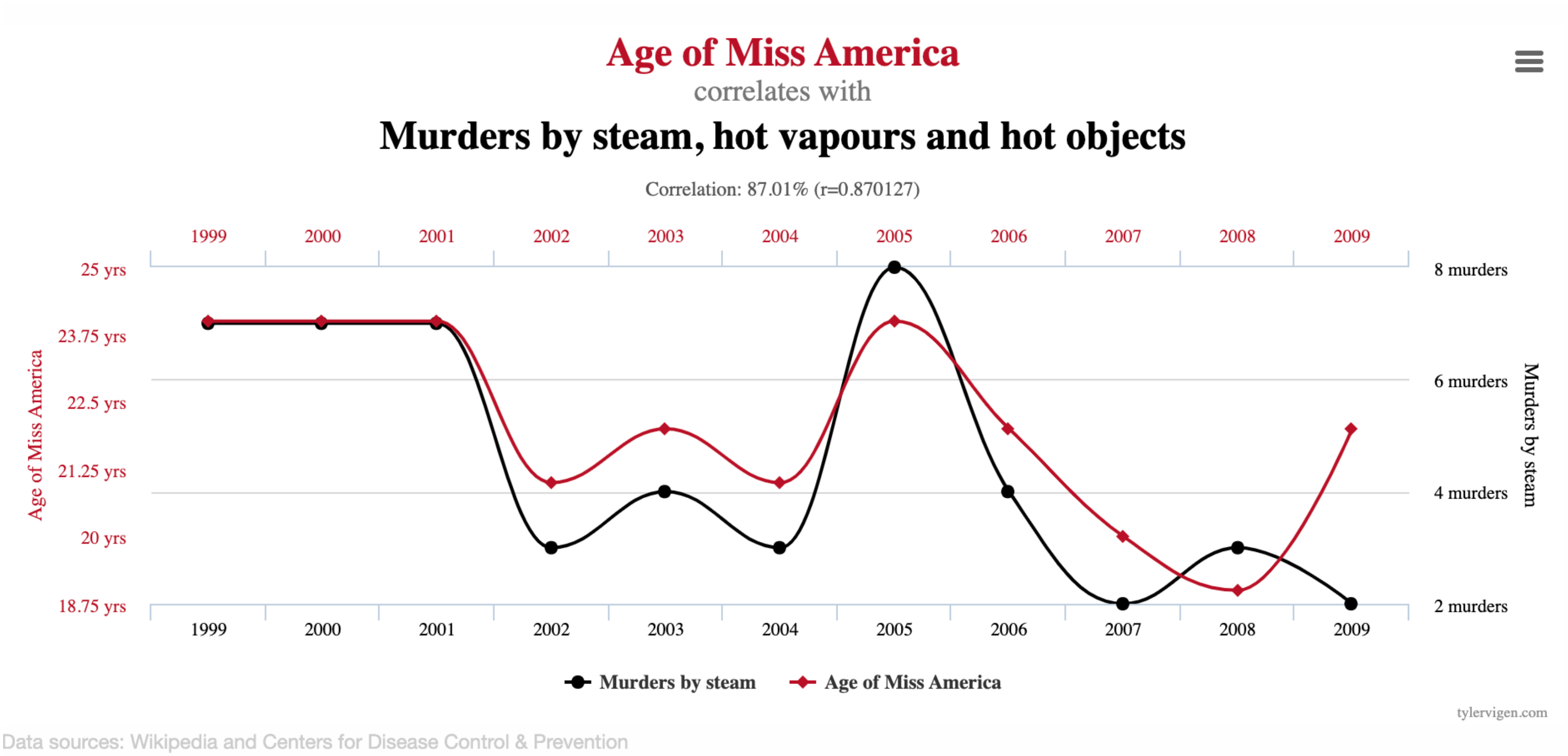
Correlation does not necessarily imply causality


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

https://m.xkcd.com/552/



https://www.tylervigen.com/spurious-correlations



Age of Miss America

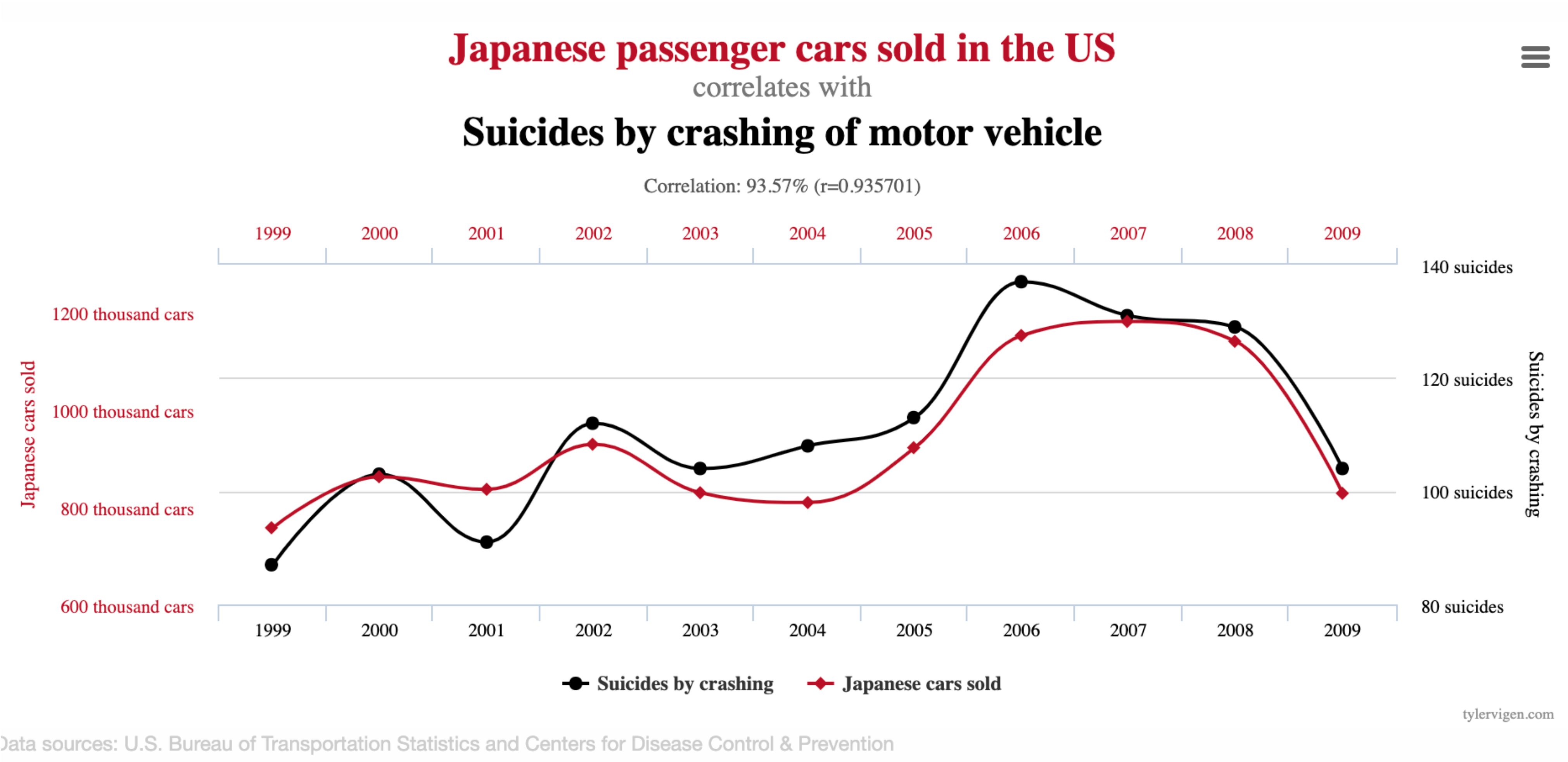
correlates with

Murders by steam, hot vapours and hot objects

Correlation: 87.01% (r=0.870127)

Data sources: Wikipedia and Centers for Disease Control & Prevention

https://www.tylervigen.com/spurious-correlations



Japanese passenger cars sold in the US

correlates with

Suicides by crashing of motor vehicle

Correlation: 93.57% (r=0.935701)

Data sources: U.S. Bureau of Transportation Statistics and Centers for Disease Control & Prevention

https://www.tylervigen.com/spurious-correlations

Experimental design vs observational studies

- Good experimental design
  - Tight control over aspects of the experiment. Can account for known factors (e.g. stratification) or unknown factors (e.g. randomization) to get unbiassed sample
  - Analysis easier and easier to help us isolate the effects of interest
  - E.g. Drug trials
- Observational studies (more common in Data Science)
  - Just observing what happens
  - Need large sample sizes
  - Can test things you couldn’t ethically design
  - But needs more complex modelling

Stratified sampling

- May consider stratified sampling or sampling on some external factor
  - Sampling people based on their background
    - Ensure gender/age/ethnicity representation
  - Sampling records based on their location
    - Ensure all part of a field (which may have areas that flood etc) are represented

Counterfactuals

- 2 states
  - intervention
  - no intervention
- Usually can’t observe both in an individual so difficult to establish if there’s a causal effect
  - Though sometimes we might; e.g. at different times
- We can observe **average causal effect**
  - Observe differ people with same characteristics
  - But are the characteristics really the same?

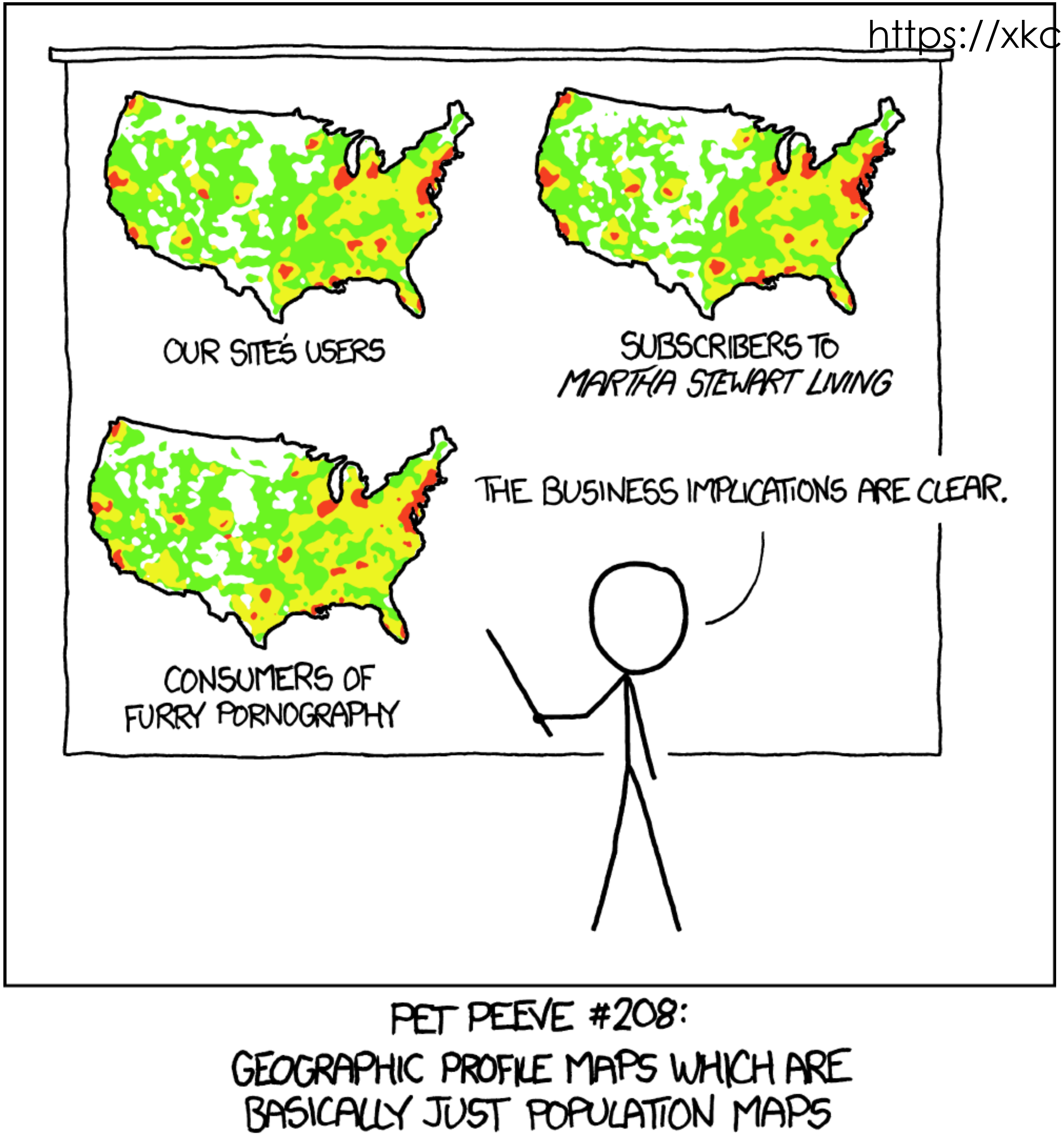
Causal thinking

- We often can’t verify a causal effect, but we can do **causal thinking**
  - reason whether the effect is may be causal
    - Shoe size correlates with literacy
  - investigate further
    - Shoe size correlates to age – more likely to be casual
  - correct
    - Control shoe size for age

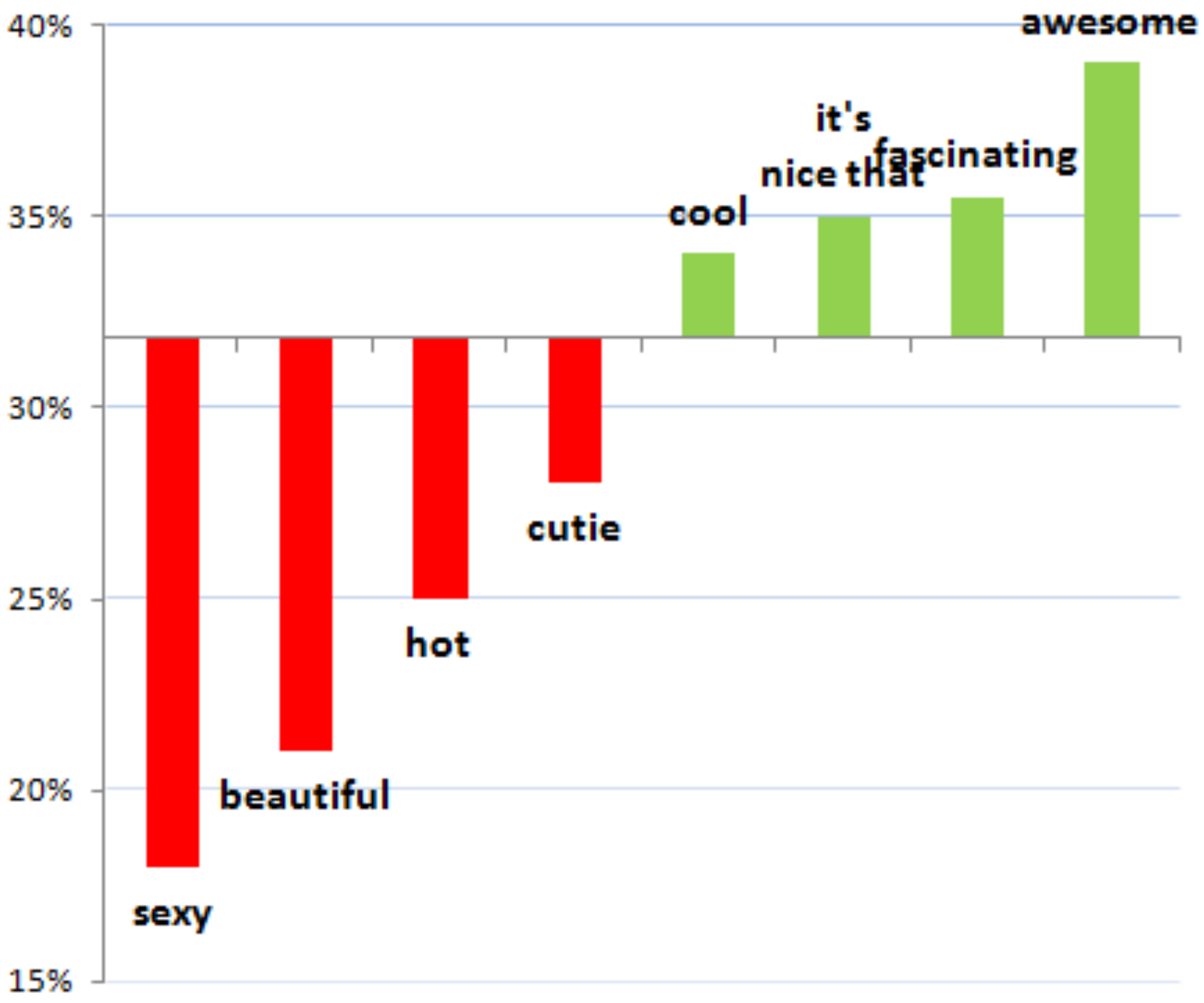
Causal thinking

- If you want make causal inferences, you need to understand how the variable may have a causal effect
  - Directly
    - *Perfect!*
  - a proxy for something you can't measure
    - *That's probably OK, as long as we know*
  - a confounder (no direct causal effect)
    - *That's dangerous!*

https://xkcd.com/1138/

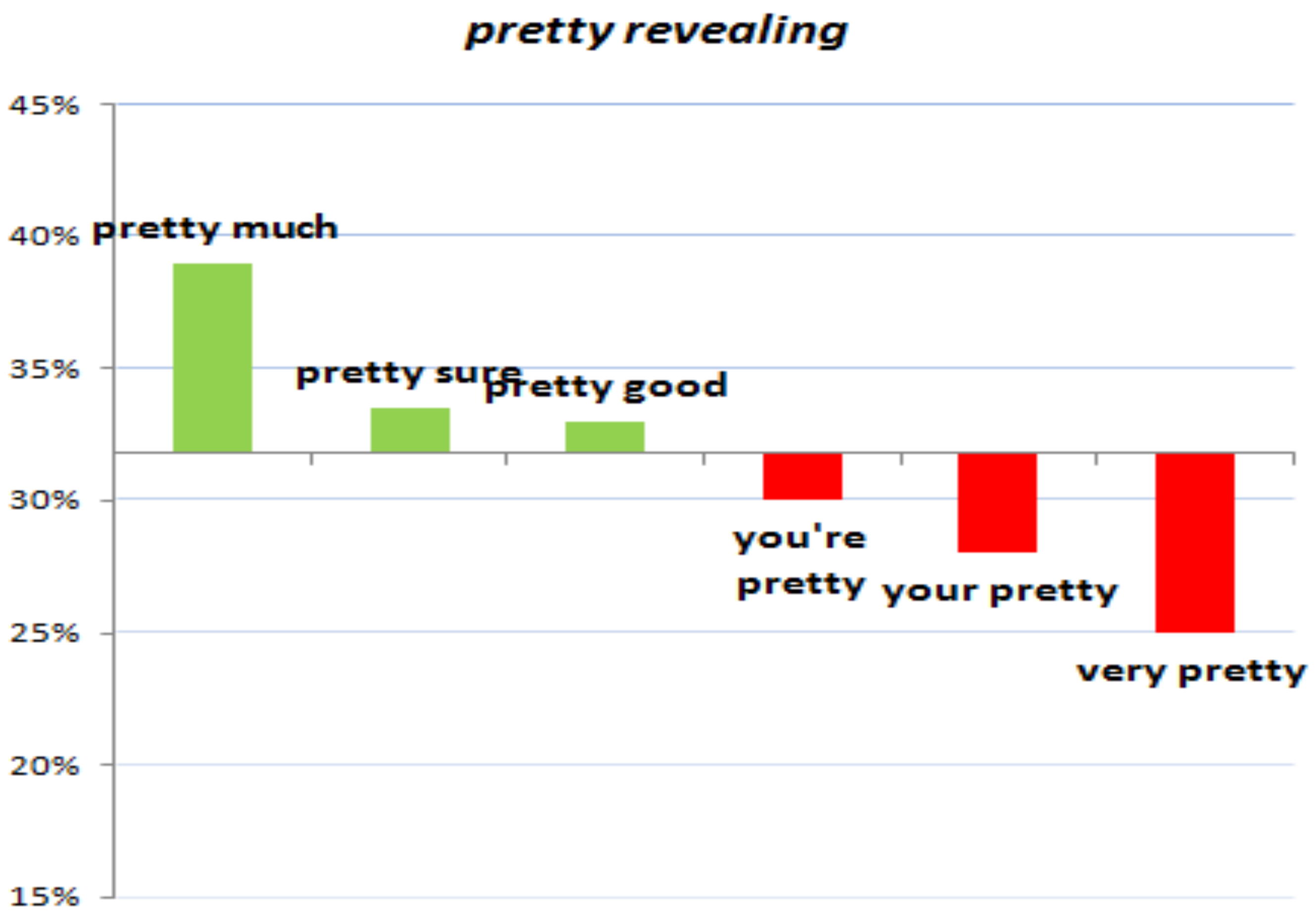

OUR SITE'S USERS
SUBSCRIBERS TO *MARTHA STEWART LIVING*
CONSUMERS OF FURRY PORNOGRAPHY
THE BUSINESS IMPLICATIONS ARE CLEAR.
PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

OK Cupid: "Avoid physical compliments"



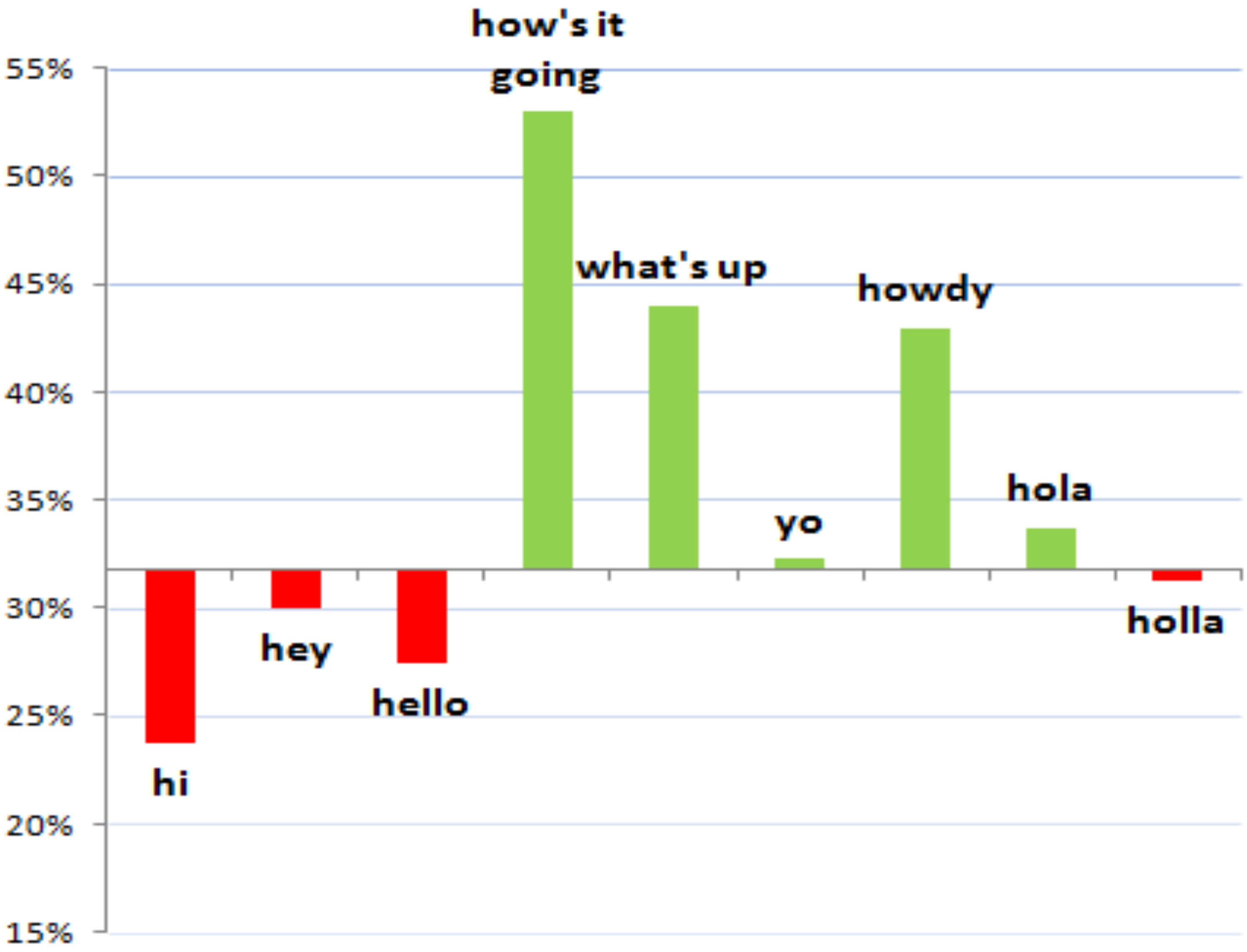
https://theblog.okcupid.com/exactly-what-to-say-in-a-first-message-2bf680806c72

OK Cupid: “Avoid physical compliments”



https://theblog.okcupid.com/exactly-what-to-say-in-a-first-message-2bf680806c72

OK Cupid: "Use an unusual greeting"



https://theblog.okcupid.com/exactly-what-to-say-in-a-first-message-2bf680806c72

Methods for helping determine causality

- Time-lag

Advice for causing thinking

- Know your data so you know what assumptions are reasonable
- Figure out what you want to know and investigate that
- Don’t blindly accept model results. *Think*.

Conclusion

- Models and prediction for supporting analysis
  - Focus on their use for analysis
  - not only data-driven
- Regression
  - Continuous: Simple and multiple linear; other types
  - Categorical: Logistic regression; decision trees; SVM
- Validation (and avoiding overfitting)
  - Split your sample into training and testing; validate
- Causal thinking
  - Are independent variable likely to be directly causal, indirectly causal (proxies) or not causal (confounder)
  - Know your domain, know your data and *think*

Reading

- **Statistical assumptions:** Four assumptions of multiple regression that researchers should always test.
- Doing Data Science Straight Talk from the Frontline (available ONLINE at the library)
  - Chapter 5: Logistic regression
  - Chapter 7: Extracting meaning from data
  - Chapter 11: Causality
- Coursea course: Real-life-data-science
  - Experimental design and observational studies
  - Causality 1
  - Causality 2