DATA5207 Notes

GENERAL INFORMATION

Topic	Type	Marks	Due Date
Quiz	Individual	5%	Week 6
Group Work	Group	25%	Multiple Weeks
Research Project	Individual	50%	STUVAC
Research Plan	Individual	20%	Week 9

ASSIGNED READINGS

Data visualization, a handbook for data driven design.

Data visualization is defined as the representation and presentation of data to facilitate understanding. To derive understanding from data we need to see it represented in a different, visual form. This is the act of data representation. The goal of data visualisation is facilitating understanding. There are 3 stages of understanding, these are:

- Perceiving: What does it show? This is the act of simply being able to read a chart.
- Interpreting: What does it mean? This is the act of converting the perceived values into some form of meaning.
- Comprehending: What does it mean to me? This involves reasoning the consequence of the perceiving and interpreting stages to arrive at a personal reflection.

The 3 principles of a good data visualisation design are as follows:

- Good data visualization is trustworthy.
- Good data visualization is accessible.
- Good data visualization is elegant.

Followed by this are a few universal principles of Data visualization:

- Good design is innovative.
- Good design is long lasting.
- Good design is environmentally friendly.

A good data visualization follows the following steps:

- Formulating your brief.
- Working with data.
- Editorial thinking.
- Data representation.
- Data presentation.
- Audience influence:
    - Subject matter appeal.
    - Dynamic of need.
    - Subject matter knowledge.
    - What do they need to know about?
    - Format.
    - Personal tastes.
    - Attitude and emotion.

An elegant design is about seeking to achieve a visual quality that will attract your audience and sustain that sentiment throughout the experience, far beyond the initial moments of engagement. To achieve this elegance, one must follow these steps:

- Eliminate the arbitrary.
- Have a thoroughness.
- Have a style, consistent one.
- Decoration should be additive and not negative.
- Minimalism is the best.

Data Visualization workflow:

i. Formulating your brief
    - Planning, defining, and initiating your project.
    - A brief represents a set of expectations and captures all the relevant information about a task or project.
    - Establishing your project's context involves defining your original curiosity, identifying your project circumstances, people, constraints, consumption, deliverables, resources, and defining the purpose of the project.
    - Following the establishment of project context, the project's vision is to be established. This involves creating the purpose map, experience, tone, harness ideas,

ii. Working with data
    - Going through the mechanics of gathering, handling, and preparing your data.
    - There are 4 different groups of action for this step, these are:
        - Data acquisition: Gathering the raw material.
        - Data examination: Identifying physical properties and meaning.
        - Data transformation: Enhancing your data through modification and consolidation.
        - Data exploration: Using exploratory analysis and research techniques to learn.
    - There are different types of data types available, these are: textual (qualitative), nominal (qualitative), ordinal (qualitative), internal (quantitative), ratio (quantitative), temporal, discrete, and continuous.
    - We need sufficient data to analyse and visualize. There are several different origins and methods involved in acquiring data. A few such methods are:
        - Primary data collection
        - Manual collection and data forging.
        - Web scraping.
        - Extracting from pdf files.
        - Any data issued to you.
        - Downloaded from the web.
        - System reports and exports.
        - Data from third party services.
    - Once data is collected, the data properties are to be understood. The following is a list of properties to know about the data:
        - Data types: Different data types of the attributes.
        - Size of the data: The range of the values for attributes.
        - Condition of the data: Presence of outliers, missing values, null values, etc.

iii. Establishing your editorial thinking

- Defining what you will show to the audience.
- The essence of editorial thinking is demonstrating a discerning eye for what you are going to portray to the audience. A few pointers to keep in mind is the followings:
    - Pick an angle to pitch your ideas from.
    - Why is the angle relevant to your cause?
    - Is one angle sufficient to cover the cause?
    - How do you frame the entire thing together?
    - What do you choose to focus on?

iv. Developing your design solution

- Considering all the design options and beginning the production cycle.
- The production cycle involves a lot of stages, a few common ones are:
    - Conceiving ideas across the 5 layers of visualization design.
    - Wireframing and storyboard design.
    - Developing prototypes or mock-up versions.
    - Testing.
    - Refining and completion.
    - Launching the solution.

There are many different ways of encoding data but these always comprise combinations of two different properties, namely marks and attributes. A mark can be in the form of a point, line, area or form (a 3 dimensional shape). Attribute on the other hand comprises of multiple features such as the position, size, angle, slope, quantity, colour saturation, colour lightness, pattern, motion, symbol, shape, colour hue, connection/edge, and containment.

There are multiple chart types to choose from. A few of them are:

Bar chart	Clustered bar chart	Dot plot	Connected dot plot
Connected scatter plot	Proportional shape chart	Univariate scatter plot	Back-to-back bar chart
Polar chart	Range chart	Box and whisker plot	Word cloud
Histogram	Bubble chart	Pie chart	Waffle chart
Stacked bar chart	Radar chart	Tree map	Pictogram
Venn diagram	Dendrogram	Sunburst	Scatter plot chart
Bubble plot	Parallel coordinates	Heat map	Matrix chart
Node-link diagram	Chord diagram	Sankey diagram	Line chart
Bump chart	Slope graph chart	Area chart	Horizon chart
Stream graph	Connected timeline	Gantt chart	Instance graph
Choropleth map	Isarithmic map	Grid map	Prism map
Dot map	Flow map	Area cartogram	Dorling cartogram
Proportional symbol map			

The essentials of Political Analysis

Researchers observe the sometimes-chaotic political scene and create explanation for what they can see. They offer hypotheses about political relationships and collect facts that can shed light on the way the political world works.

A concept is an idea or a mental construct that represents a phenomena in the real world. A conceptual question is a question which is expressed using ideas, and is usually frequently unclear and thus is difficult to answer empirically. A concrete question on the other hand is a question expressed using tangible properties and is hence answerable empirically.

WEEK 1

Assigned reading: Observing Behaviour

Digital footprint or digital shadow refers to one's unique set of traceable digital activities, actions, contributions, and communications manifested on the Internet or digital devices. In addition to these traces held by the businesses, there are also large amounts of incredibly rich data held by the government. And the combination of these records is called big data. Big data is usually defined with the combination of 3 Vs: Volume, Variety, and Velocity.

Observational data is any data that results from observing a social system without intervening in some way. Big data are created and collected by companies and governments for purposes other than research. Using this data for research therefore requires repurposing. The main source of big data is the online data that has been created and collected by companies whereas the second major source of big data are government administrative records. Big data sources tend to have several characteristics in common some are generally helpful for research while some are problematic. A few common characteristics are listed below:

- Big: Big data literally means large datasets.
- Always on: The data is being constantly updated; more records are added regularly.
- Non-reactive: Any measurement in big data sources is much likely to change the behaviour.
- Incomplete: No matter how big the data is, it will never have a complete answer to the question.
- Inaccessible: Data held by the government and companies is difficult to access by researchers.
- Non-representative: These types of data are bad for out of sample generalizations but quite useful for within sample generalizations.
- Drifting: The data can not be used to study long term trends due to usage drift, population drift, etc.
- Algorithmically confounded: Behaviour in big data is not natural and is driven by the engineering goals of the systems.
- Dirty: Big data is loaded with junk and spam.
- Sensitive: Some of the information held by the companies and government is sensitive.

Based on the common characteristics of the big data sources, there are mainly 3 different strategies to learn from this big data, these are:

- Counting things
- Forecasting things
- Approximating experiments

Big data will be most valuable to social sciences in 3 ways:

- Enable researchers to decide between competing theoretical problems.
- Big data sources can enable improved measurement of policy through now-casting.
- Big data sources can enable researchers makes casual estimates without running experiments.

Assigned reading: Data, Data everywhere.

The world contains an unimaginably vast amount of digital information which is getting ever vaster ever more rapidly. This makes it possible to do many things that previously could not be done: spot business trends, prevent diseases, combat crime and so on. Managed well, the data can be used to unlock new sources of economic value, provide fresh insights into science, and hold governments to account. But they are also creating a host of new problems. The amount of data being captured exceeds the available storage space. Moreover, ensuring data security and protecting privacy is becoming harder as the information multiplies and is shared ever more widely around the world. As the world is becoming increasingly digital, aggregating, and analysing data is likely to bring huge benefits in other fields as well. Sometimes those data reveal more than was intended. But big data can have far more serious consequences than that. During the recent financial crisis, it became clear that banks and rating agencies had been relying on models which, although they required a vast amount of information to be fed in, failed to reflect financial risk in the real world. This was the first crisis to be sparked by big data--and there will be more.

Introduction

We are in the middle of a data revolution and we are using data science to understand the social world. The main theme of this unit is incorporating an understanding of subject in our analysis and understanding the bias, causality, and confounding factors and lastly, communicating the ideas to a larger audience.

Social Science in the broadest sense means the study of human behaviour, particularly at the level of the society and how people behave in groups, and influence the world. Social Sciences helps us understand how the society works. It is a broad topic which encompasses multiple disciplines such as anthropology, archaeology, economics, human geography, psychology, history, etc. Some idiosyncrasies of the social science research are:

- Human behaviour is a stochastic (random) process.
- Specific issues are encountered when studying human behaviour and the social world.

A good social science theory has the following characteristics:

- It tries to understand the problem.
- An idea of causal mechanism underpins the phenomenon being studied.
- The theory is falsifiable (able to be altered or represented falsely).
- Developed a working theory to guide the model making.

WEEK 2

Introduction

It is important to effectively communicate in data visualization. One of the most effective ways to relay information gained from the data is through visualizations. Data visualization is the representation and presentation of data to facilitate communication and understanding. The reason why we visualize data can be varied, such as communication information to stakeholders, make the audience easy to understand data, decision making through giving a overall view of the data, etc.

Data visualization simplifies complexity. As we dig deeper into data and make the analysis more complex, patterns become harder to interpret and this calls for the need of visualization of data. Table representation of data is very basic but the representation of the data through visuals makes it easy to understand for the audience with lesser effort. Visualizations can play an important role in inductive and deductive approaches to science. Data visualizations generalise the data and the precision and accuracy are lost in the process. Inductive reasoning is to display the data and reveal patterns and anomalies and suggest processes explaining them. Data visualizations allow non-experts to understand patterns.

Data visualizations can be good and bad. A good data visualization should be minimalistic (less is more). A good visual will be easy to read and infer. There are 3 considerations to have when visualizing data:

- What information are you trying to communicate? What are you trying to show?
- Who is your target audience?
- Why is a design feature relevant? What needs does it serve?

This brings us to 3 rules of data visualization:

- Avoid defaults. (Think carefully about what you are trying to achieve and the best way to do this)
- Minimize distractions and focus on key message (Less is more). Plot must be accessible and understandable.
- Form follows function (The goal of the visualization and the audience and medium should dictate the design).

WEEK 3

Additional Readings: Simpson's paradox in Behavioural Data

Observational data about human behaviour is often heterogeneous. Simpson's paradox is one important phenomenon confounding analysis of heterogeneous social data. According to the paradox, an association observed in data that has been aggregated over an entire population may be quite different from, and even opposite to, associations found in the underlying subgroups. Simpson's paradox also affects analysis of trends. When measuring how an outcome variable changes as a function of some independent variable, the characteristics of the population over which the trend is measured may change with the independent variable. As a result, the data may appear to exhibit a trend, which disappears or reverses when the data is disaggregated by subgroups. Multiple examples of Simpson's paradox have been identified in empirical studies of online behaviour, a few are:

- Exposure response in social media.
- Content consumption in social media.
- Answer quality in stock exchange.

A simple test that can help ascertain whether a pattern observed in data is robust or potentially a manifestation of Simpson's paradox. The test creates a randomized version of the data by shuffling it with respect to the attribute for which the trend is measured. Shuffling preserves the distribution of features but destroys correlation between the outcome variable and that attribute. As a result, any trends with respect to that attribute should disappear. This suggests a rule of thumb: if the trend persists in the aggregate data, but disappears when the shuffled data is disaggregated, then Simpson's paradox may be present.

Simpson's paradox can indicate that interesting patterns exist in data, but it can also skew analysis. The paradox suggests that data comes from subgroups that differ systematically

in their behaviour, and that these differences are large enough to affect analysis of aggregate data. In this case, the trends discovered in disaggregated data are more likely to describe—and predict—individual behaviour than the trends found in aggregate data. Thus, to build more robust models of behaviour, computational social scientists need to identify confounding variables which could affect observed trends.

Additional Readings: Linear regression the basics

Linear regression is a method that summarizes how the average values of a numerical outcome variable vary over subpopulations defined by linear functions of predictors. Introductory statistics and regression texts often focus on how regression can be used to represent relationships between variables, rather than as a comparison of average outcomes. For a binary predictor, the regression coefficient is the difference between the averages of the two groups. Regression coefficients are more complicated to interpret with multiple predictors because the interpretation for any given coefficient is, in part, contingent on the other variables in the model. Typical advice is to interpret each coefficient "with all the other predictors held constant." A few assumptions of the linear regression model are:

- Validity: the data you are analysing should map to the re-search question you are trying to answer.
- Additivity and Linearity: The regression predictors are linearly additive in nature.
- Independence of Errors: The simple regression model assumes that the errors from the prediction line are independent.
- Equal variance of errors: If the variance of the regression errors is unequal, estimation is more efficiently performed using weighted least squares, where each point is weighted inversely proportional to its variance.
- Normality of errors: The regression assumption that is generally least important is that the errors are normally distributed.

Additional Readings: P Values, what they are and what they are not.

P values (or significance probabilities) have been used in place of hypothesis tests as a means of giving more information about the relationship between the data and the hypothesis than does a simple reject/do not reject decision. Virtually all elementary statistics texts cover the calculation of P values for one-sided and point-null hypotheses concerning the mean of a sample from a normal distribution. There is, however, a third case that is intermediate to the one-sided and point-null cases, namely the interval hypothesis, that receives no coverage in elementary texts.

The article shows that one-sided and point-null hypotheses are not two different objects that should never be compared, but rather they are just different versions of the same object of which interval hypotheses are versions as well. For nice data distributions the P value is continuous as a function of the hypothesis. We have also seen that P values cannot be interpreted as measures of support for their respective hypotheses.

Introduction

One of the reasons we use linear regression is because the world is not simple. Basic descriptive statistics is fine for most of the problems but not always adequate for every job such as predictions, or in the presence of confounding factors. In statistics, a confounder is a variable that influences both the dependent variable and independent variable, causing a spurious association. Confounding is a causal concept, and as such, cannot be described in terms of correlations or associations. One way we can control confounding factors is through using a

regression model. Through using it we can control the potential confounding factors and isolate the independent variable and the predictive outcome. The formula for linear regression is:

$$y = \alpha + \beta x + \varepsilon$$

Where y is the dependent variable, x is the independent variable (predictor), ε is the error distribution and α is the constant. When using the linear regression in R we can have multiple predictors as seen below:

```
## lm(formula = earn ~ z.height + race2 + gender + z.age,
##                coef.est coef.se
## (Intercept)   16444.11   758.48
## z.height       4557.52  1427.38
## race2Black    -2632.00  1742.56
## race2Other     1620.40  3590.42
## race2Hispanic -4072.34  2141.91
## genderMale    11240.25  1450.25
## z.age          4252.46  1137.31
## ---
```

A variable can be called statistically significant if the following equation holds true:

$$coef.est > 2 * coef.se$$

General rule of thumb:

i. If a coefficient is 2 standard errors from 0, then it is statistically significant. Basically, the following equation:

$$coef.est > 2 * coef.se$$

WEEK 4

Assigned Readings: Interpreting and using regression.

Statistical theory is a branch of mathematics and can be formulated as a set of symbolic relationships, presented as a series of theorem. All statistical methods depend on assumptions. Assumptions need to be made about the unmeasured forces influencing outcomes – the disturbances. These assumptions jointly – the functional form plus the assumptions about the disturbances – are called the specification. Without correct specifications, conventional statistical theory gives no assurance that the impact of a variable will be estimated correctly.

Functionally correct causal specification in social science is neither possible not desirable. A model is said to be saturated when the regression model has independent variables that consist solely of dichotomies plus every possible interaction among them. To summarize the reading, a good social data analysis oriented to theory construction usually begins with a non-functionally specific hypothesis. A suitable data set is found to check the claim and a substantively reasonable statistical description of it is constructed.

Any statistical description must be assessed for accuracy. Assessment of the reliability of estimates is critical to data analysis.

Additional Readings: Practical considerations in applying regression analysis.

There is more to selecting the independent variables than choosing the factors that intuitively appear to be good candidates for explanatory variables. To build the best model as well as understand its potential limitations, it is important to be aware of the potential sources of forecast error. These include:

- Random errors for true population regression
- Random errors in the estimated regression coefficients
- Regression equation may be mis-specified.
    - Due to omission of significant values
    - Maybe because the model is non-linear.
    - Maybe because of a wrong functional form assumed in linear transformation.
    - Error terms are auto related.
- Errors in independent variable values
- Data errors
- Structural changes

Stepwise regression is a highly useful and efficient procedure for isolating and providing summary results for the most statistically interesting equations. There are two basic types of stepwise regression:

- Forward selection
- Backward elimination

A step-by-step regression procedure is given below:

i. Determine the dependent variable.
ii. List all possible choices for explanatory variables.
iii. Choose a subset of these (usually no more than five), taking care to avoid selecting correlated independent variables. Scatter diagrams can be used as an aid in this selection process.
iv. Choose the length of the survey period. Scatter diagrams can also be used as an aid in this step.
v. Apply a stepwise regression program to the selected variables.
vi. Analyse the results by examining the various key statistics: t values, SER, CR2 , F, and DW. If there is any evidence of multicollinearity, check out this possibility and rerun stepwise regression with a different set of variables if necessary.
vii. Generate detail and construct residual plots for the most promising equations in the stepwise regression run.
viii. Check residual plots for outliers. Decide whether outliers should be deleted.
ix. Check residual plots for autocorrelation.
x. If outliers or autocorrelation exist, try to correct through the addition of variables or transformations to achieve linearity.
xi. If autocorrelation is still a problem, try a transformation to eliminate autocorrelation (e.g., first differences).
xii. Check the correlation matrix or R2 values for various combinations of equations based on the explanatory variables in order to verify that multicollinearity is not a problem.
xiii. Repeat steps 3–12 for other selections of explanatory variables.
xiv. Optional: After narrowing the number of possible models to three or less, generate simulations.

Additional Readings: Sample size and power calculations

In a sample survey, data are collected on a set of units in order to learn about a larger population. In unit sampling, the units are selected directly from the population. In cluster

sampling, the population is divided into clusters: first a sample of clusters is selected, then data are collected from each of the sampled clusters. In one-stage cluster sampling, complete information is collected within each sampled cluster. In two-stage cluster sampling, a sample is performed within each sampled cluster. More complicated sampling designs are possible along these lines, including adaptive designs, stratified cluster sampling, sampling with probability proportional to size, and various combinations and elaborations of these. The sample size of a study can be increased in several ways:

- Gathering more data of the sort already in the study,
- Including more observations either in a no clustered setting, as new observations in existing clusters, or new observations in new clusters
- Finding other studies performed under comparable (but not identical) conditions (so new observations in effect are like observations from a new "group").
- Finding other studies on related phenomena (again new observations from a different "group").

Before data are collected, it can be useful to estimate the precision of inferences that one expects to achieve with a given sample size, or to estimate the sample size required to attain a certain precision. This goal is typically set in one of two ways:

- Specifying the standard error of a parameter or quantity to be estimated, or
- Specifying the probability that a particular estimate will be "statistically significant," which typically is equivalent to ensuring that its confidence interval will exclude the null value.

Introduction

A few considerations which we should make when building models are:

- The purpose of the regression

  A few things we should think about for the purpose of the regression is whether the regression model is fit for hypothesis testing or prediction. Each of the criteria demands a different model. If we need to make a model fit for hypothesis testing, the model is to falsify or prove a theory whereas to make a model fit for predictive model we need observed data to produce estimates about values in the unobserved data.
- Observational vs causal (treatment effect)

  A data is said to be observational when the data has been collected through observation without being involved in the causal mechanism of it. We can't make causal inference from observational data as we can't control the confounding factors for the data.

Uncertainty

Uncertainty in data and models is the quantitative estimation of error present in the data. Model errors is basically the different between the observed and actual values. These are of 2 types:

- Random errors

  These are naturally occurring errors that are to be expected with any observation. These can be driven by precision limitations of measurements, for example, the sample size of the survey.
- Systematic errors

  Systematic errors are also known as the bias. Results from mistakes or problems in the research design or from flaws in data collection are basically the source of systematic errors. These can also include poorly calibrated instruments, poorly worded

surveys, non-response bias, etc. An example of systematic error that results from a problem in the sample design is called the sample design error. A few types of sample errors include:

- Frame errors: Incomplete or inaccurate sampling frame is used.
- Selection errors: Sampling procedures are incomplete or inadequate or when appropriate selection procedures are not properly followed.

- Measurement Errors

  Various types of errors that may occur happen due to numerous deficiencies in the measurement process. These errors include:

  - Interviewer error: Where the interviewer influence responses.
  - Measurement instrument bias: Problems with the measurement instrument or questionnaire.
  - Response and non-response bias

WEEK 5

Assigned Readings: The Nature of Probability Theory

In each field we must distinguish between 3 aspects of the theory:

- The formal logical content
- The intuitive background
- The applications

The theory of probability is now applied in many diverse fields, and the flexibility of a general theory is required to provide appropriate tools for so great a variety of needs. The original purpose of the theory of probability was to describe the exceedingly narrow domain of experience connected with the games of chance and the main effort was directed to the calculation of certain probabilities.

Assigned Readings: The Sample Space

The mathematical theory of probability gains practical value and an intuitive meaning in connection with real or conceptual experiences. Any theory necessarily involves idealization, and our first idealization concerns the possible outcomes of an experiment or observation.

Introduction

Regression is a commonly used statistical method to investigate the relationships between two or more variables. There are multiple types of regression model such as linear regression, logistic regression, etc. There are many strengths and limitations of linear regression. It is useful for estimating the relationship between 2 variables when there are possible confounding factors involved. It is also useful for predictions, measure the uncertainty in relationships, etc. But one of the major limitations of the model is that it assumes a linear relationship between the outcome and the predictor. Some problems where linear regression may not be appropriate is during classification problems. This is because of the assumption of linearity between the outcome and the predictors.

An alternative for linear regression is logistic regression which assumes that the relationship between he dependent and the predictors are non-linear and usually has a binomial outcome (we can get ordinal or multinomial as well). We can use logistic regression in many cases such as to understand if a person will default in his credit card bills or not, or to examine the relationship between a person's age and their likelihood of voting for political parties. It is to be noted that Logistic regression is important and useful when measuring discrete outcomes.

The equation for logistic regression is given as:

$$Prediction\ (y = 1) = logit^{-1}(x_i\beta)$$

The logistic regression model will always give a curve and the values will never go beyond 1 and less than 0. The output of binomial logistic regression model is the likelihood of the dependent variable being 1 or 0 where the values can be interpreted as the likelihood of either outcome occurring given the values of the independent variables. Interpreting coefficients of linear regression is fairly straightforward as the coefficient value tells us the change in value of the dependent variable when an independent variable increases in value by 1 unit.

The coefficient of logistic regression are typically measured on a logarithmic scale and are referred to as log odds or logits. It is difficult to model a variable which has a restricted range such as probability and logit transformation maps this probability ranging between 0 and 1 to log odds ranging from negative to positive infinity. A binomial logistic regression model allows us to estimate the relationship between a binary outcome and a set of predictors There is an easier way to understand the results from logistic regression and these are called predicted probabilities. This is calculated by using the inverse logit function to transform the output from the logit scale to probabilities. The formula is given below:

$$p = e^{\frac{x}{1+e^x}}$$

We use logistic regression in any phenomena where we are looking at the probability of discrete outcomes occurring.

WEEK 6

Assigned readings: Ethical, Political, Social, and legal concerns.

Data are generated and employed for many ends, including governing societies, managing organisations, leveraging profit, and regulating places. There is a fine balance then between using data in emancipatory and empowering ways and using data for one's own ends and to the detriment of others, or in ways contrary to the wishes of those the data represent. The generation of data and the work these data do are inherently infused with ethical, social, and political concerns. Such concerns have long been recognised and debated within scientific and public fora, leading to the creation of a raft of professional ethical guidelines and legislation that delimits how data are produced, managed, shared, and employed.

Collectively, data footprints and shadows provide a highly detailed record of an individual's daily life: their patterns of consumption, work, travel, communication, play, interactions with organisations, and their thoughts and interests. Dataveillance is a mode of surveillance enacted through sorting and sifting datasets in order to identify, monitor, track, regulate, predict and prescribe. Privacy is a condition that many people expect and value. It is considered a basic human right and is enshrined in national and supranational laws in various ways.

Information that was previously considered private is being more freely shared, such as résumés (via LinkedIn), family photographs and videos (via Flickr, Instagram, and YouTube), personal and family stories (via Facebook and blogs), and personal thoughts (via Twitter, chat rooms and online reviews). At present, privacy legislation is largely constructed around personal rights and consent regarding the generation, use, and disclosure of personal data.

Given the value of data, especially personal data that can facilitate identity theft, or commercial data that can be pirated or used to gain competitive advantage, data security has become an important aspect of data protection. As the data revolution unfolds, and more and more devices produce, share and utilise data, it seems that security issues are going to multiply, not lessen. In turn, this is going to exacerbate crimes such as identity theft, undermine trust in data systems, and raise a series of legal questions concerning responsibility and liabilities in protecting systems when data are mishandled, misappropriated, and stolen.

On the surface predictive profiling looks to be a win-win situation for customers and vendors – customers receive personalised treatment and vendors gain sales and reduce churn. However, predictive profiling can be used to socially sort and redline populations, selecting out certain categories to receive a preferential status and marginalising and excluding others. One of the foundations of privacy and data protection policy in the European Union and North America is the concept of data minimisation. This stipulates that agencies and vendors should only generate data necessary to perform a particular task, that the data are only retained for as long as they are required to perform that task (or if legal considerations dictate), and that the data generated should only be used for this task.

The speculative harvesting of vast quantities of data, much of it captured without individuals' knowledge or understanding, and then being put to secondary uses, clearly raises ethical questions concerning not only privacy and data protection, but also governance. One of the clearest examples related to governance is control creep. Control creep is where the data generated for one form of governance is appropriated for another. Beyond control creep and anticipatory profiling, the data revolution has several potential impacts with regards to the organisation and operation of governance. One of the ways in which governance is being transformed through data-driven technologies is by making it more technocratic in nature.

While such practices have benefits for governments, companies, and citizens, they also have differential and negative consequences. Given how rapidly the data landscape is changing, keeping track of developments, determining their potential implications, and thinking through appropriate social and legal responses is a challenge.

Assigned readings: Research training for social scientists.

Among the first questions that any researcher should ask are: Why am I doing this? Why am I researching this topic? These are the kind of questions that are considered when discussing whether research can ever be value free. After considering one's own motivation, another question to address is: Why should I be concerned about ethical and legal issues?

The term 'empowerment' is often used in relation to various kinds of social action. Research has the potential to empower people if it gives them the benefit of knowledge that will enable them to control their own destinies. But it is necessary to recognize that research also has the capacity for disempowerment. Perhaps the most extreme form of abuse of power that can be perpetrated by researchers is data fraud. The extent of such fraud is unknown and probably unknowable, but there are clear instances of it happening. It is, of course, possible to exaggerate the power of research.

Predictive Models

Predictive modelling is a process used to predict future outcomes by analysing historic (or training) data. It determines the most likely outcomes based on the past (known) outcomes.