Data-Informed Decision Making

We live in a world dense with data, computational power, and connectivity. This creates an expectation that decision-making be rigorous, rational, and empirically grounded rather than being based purely on intuition or expertise. This course introduces you to this expectation and how to meet it. We begin with the power and importance of data-informed decision making (DIDM) and how to recognize opportunities to leverage the power of data. We learn how to communicate the stories told by data, how to rigorously assess the accuracy and validity of those stories, and how to identify, collect, and analyze the data needed to generate them.

1. Data Science for Decision Making

Data-informed decision making (DBDM) is the process of making decisions based on the analysis of empirical data rather than intuition or ideology or materially motivated preferences. In this unit we explore the elements of this definition, describe a generic workflow for doing DBDM, unpack the benefits of DBDM, clarify when to and when not to use, and explore the many types and variants of DBDM. This overview of data-informed decision making motivates the following unit: how data inform through the stories they tell.

1.1. Introduction to data-informed decision making

Welcome, set the right tone, put students at ease.

Lesson Goal(s): 1. Understand what the course is about and how it will proceed and the kind of things one should be able to do at the end of the course. 2. To be excited about what is to come. 3. Be able to describe arc of the course, expectations, etc. 4. Able to distinguish data driven from data informed. 5. Be able to name several reasons to prefer DIDM to intuition or expertise based decision making.

1.2. Lesson 2: Types of Decisions and Types of Analytics

In this class we ground the class in the big picture of decision making as a serious human challenge and the conditions of endless data, computational power, and connectedness of the world (which makes it more and more important for organizations to make good decisions in competitive markets) generate an imperative for decision-making be rigorous, rational, and empirically grounded rather than being based purely on intuition or expertise.

Lesson Goal(s): 1. Deeper understanding of decision making as a fundamental problem and the way different kinds of analytics informs different kinds of decision making problems.

1.3. Lesson 3: Descriptive Analytics and DIDM Workflow (Part 1)

In this lesson we examine the wide range of things that come under the heading of descriptive analytics - mostly familiar - and we learn the basics of a vocabulary of descriptive statistics (central tendency, dispersion, trend, for example) and of descriptive visualization (generically what pie and line and bar charts are used to show, for example).

Lesson Goal(s): 1. Understand how "descriptive" is distinguished from other kinds of analytics. 2. Ability to read bar charts, pie charts, line graphs, tables to describe "what happened?" or "what is happening?" 3. Ability to critique the use of descriptive statistics and graphics relative to rhetorical intent.

1.4. Lesson 4: Predictive Analytics and DIDM Workflow (Part 2)

Predictive analytics (PA) is about making a model of data for the purpose of making predictions about the future. They can be used to manage inventory, develop marketing strategies, and plan resource deployment. In this lesson we focus on identifying decision situations that can benefit from PA, describing the workflow necessary to yield actionable insights, and the limitations associated with the tool.

Lesson Goal(s): 1. Able to articulate predictive decision situations around risks, opportunities, forecasting, etc. 2. Able to walk through scenarios of how, say, a staff deployment decision question might drive a data analytical process. 3. Able to connect the utility of data analytics to dealing with cognitive biases encountered in other courses.

1.5. Lesson 5: Prescriptive Analytics and DIDM Workflow (Part 3)

This lesson introduces the idea of how data analysis can inform decisions about what should be done and the idea of optimization. Examples drawn from commerce and optimizing for customer behavior.

Lesson Goal(s): 1. Able to articulate the utility of prescriptive analytics; 2. Able to walk through scenarios of how an optimization question might drive a data analytical process.3. Able to connect utility of data analytics to dealing with cognitive biases encountered in other courses.

1.6. Lesson 6: Variants of DIDM and Conclusion

Data-informed decision making both overlaps and exists alongside numerous other approaches (such as Cost-Benefit Analysis, Rule-Based, Multi-Criteria, Heuristic, Simulation, Optimization, Risk-Based, and Group Decision Making). In this lesson we put DIDM in this context and discuss pros and cons relative to eliminating decision maker bias, amplifying social bias, and optimally serving organizational needs.

Lesson Goal(s): 1. Able to competently describe at least one alternative to DIBM and compare them in terms of different objectives and resources that might be available in a given situation.

1.7. Lesson 7: Synthesis

Lesson Goal(s): 1. Review the Unit

2. The Story Within Your Data

Every method of DBDM links the taking of a decision with the telling of a story. In this unit we ask what are the kinds of things a dataset can say? We talk about measures of central tendency and dispersion, and patterns such as trends and association. The unit includes exploratory data analysis (EDA), basic data visualization, and how to present and write about the stories that data tell. Our focus on the story that data tell motivates the next unit where we ask what makes a data story believable.

Each class in this unit has a mini-presentation type assignment. Typically this is just a storyboard and a script.

2.1. Lesson 8: Distributions, central tendency, dispersion: Telling the simplest data stories well

In this lesson we make the counterintuitive point that just being able to say something about the most common value or how varied values are is a story about the data - at once added information and a condensation of information.

With this we set the stage for what a story about data is. Along the way we'll learn about some measures of central tendency, dispersion, etc., introduce the idea of a distribution and frequency, and some ways of visualizing one dimensional data.

Lesson Goal(s): 1. To be able to use measures of central tendency and dispersion to tell a 1 dimensional story about data. 2. To be able to use frequency and the shape of a distribution to tell a 1.x dimensional story. 3. To be able to use percentile to tell a story about a point in a distribution. 4. To be able to read these stories from a graphical representation.

Knols: dimensionality of data; mean, median, mode, standard deviation, range, interquartile range; frequency; histogram; cumulative frequency; percentile

2.2. Lesson 9: Scatter plots and time series: Telling stories about association and trends

Purpose: in this lesson we expand the idea of a data story to two variables - in two ways: XY data and XT data. This allows us to introduce the ideas of association and correlation and trends. As a parting shot in the lesson we introduce dependent and independent variables and plant the question of correlation/causation and hidden intervening variables (perhaps not in so many words).

Lesson Goal(s): 1. To be able to tell an association story by looking at a scatterplot (recognizing more and less and positive and negative correlation and no correlation). 2. To be able to read and tell the trend story in time-series data. 3. To be able to complicate association and trend stories by distinguishing dependent and independent variables and raising the question of third variables.

Knols: Scatterplots; Time-series data; Association; Correlation (positive, negative, none); Trend analysis; Dependent variables; Independent variables; Hidden intervening (third) variables; Correlation vs. causation

2.3. Lesson 10: Telling Stories about Bins: Contingency Tables and A/B Tests

Purpose: This lesson continues the multivariate theme of the previous class. We draw on our previous work with histograms as "binned" data and show how we can bin multidimensional data in a contingency table. We follow up on having learned about IV and DV with the discipline of "percentage down and compare across." The example we use is A/B testing and so we get to introduce that idea along the way.

Lesson Goal(s): 1. To be able to extend the idea of histograms and bins to 2 dimensions. 2. Percentage down, compare across. 3. To raise the question of when is a difference a difference (to be revisited in next unit)

Knols: Multivariate Analysis; Histograms; Binned Data; Contingency Tables; Independent and Dependent Variables; Percentage Down, Compare Across; A/B Testing

2.4. Lesson 11: Telling Stories About Missing Variables

This lesson extends the contingency table discussion in LP10 to the multivariate case. We look at simple cases where partialling out reveals a relationship when none had been there before (cf. Simpson's paradox). We practice telling the story of invisible effects and searching for covariates (without using the term, perhaps) that would allow an effect to become visible.

Lesson Goal(s): 1. To understand why treatment/control/causal stories are hard to tell. 2. To be able to do a simple table partialing. 3. To extend this idea conceptually to selection bias and controlling for other variables, statistical matching. 4. To be able to explain the logic of Simpson's paradox.

Knols: Contingency Tables; Multivariate Analysis; Invisible Effects; Covariates; Treatment/Control/Causal Stories; Table Partialing; Selection Bias; Controlling for Variables; Statistical Matching; Simpson's Paradox; Confounding Variables

2.5. Lesson 12: Telling Stories about Uncertainty

In this lesson we introduce measurement uncertainty and error but not in a computational sense. Rather, how to talk about (and how to understand talk about) error and uncertainty. Focus is on the reporting uncertainty in polling.

Lesson Goal(s): 1. Conceptual understanding of error bars and confidence intervals. 2. Ability to tell the correct story about results with errors and uncertainty.

Knols: Uncertainty; Measurement Uncertainty; Error; Confidence Intervals; Margin of Error

2.6. Lesson 13: Synthesis

Lesson Goal(s): 1. Review the Unit

3. Basic Analytic Tools for Decision Making

A data story is only as convincing as the analysis behind it is correct. In this unit we learn several analytical tools and how to characterize the uncertainty in the answers they provide. When we acquire some mastery of data analytical tools we will be prepared for the next unit in which we examine the stuff itself: what is data? where does it come from? how can we make sure it is what it is?

3.1. Lesson 14: Quantifying Uncertainty: Samples and Reality

In this lesson we learn how to quantify the error and uncertainty we talked about in LP 12. We introduce the normal distribution and errors as normally distributed in natural processes and how this allows us to compute confidence intervals.

Lesson Goal(s): 1.Understanding all data as a sample of reality and analysis as building a model of reality. 2. Understanding error bars and confidence intervals in principle. 3. Telling a story about error and uncertainty with understanding of the math behind it.

Knols: Error; Uncertainty; Normal Distribution; Confidence Intervals; Sampling; Model; Statistical Significance; Standard Error; Sampling Error

3.2. Lesson 15: Descriptive analytics

This lesson reprises the material from LP3 at a higher level of sophistication. We look at the same or similar data as was used to generate illustrations and examples there and reproduce the values and charts.

Lesson Goal(s): 1. To be able to explain and demonstrate how to generate some of the descriptive statistics and graphics. 2. To draw on earlier units to tell a correct and compelling story based on carrying out some of these analyses.

Knols: Descriptive Analytics; Descriptive Statistics; Data Visualization; Charts and Graphs; Data Storytelling; Measures of Central Tendency; Measures of Dispersion; Frequency Distribution; Histograms; Scatterplots; Boxplots; Percentiles

3.3. Lesson 16: A First Look at Regression

This lesson introduces students conceptually to regression. We proceed from XY and XT visualizations that we saw previously and talk about how we might model the process that produced the data we see. Using Y=MX+B from high school we introduce the change in Y per unit change in X interpretation and then extend this to logistic regression and multiple regression, again, conceptually.

Lesson Goal(s): 1.Understand linear, logistic, and multiple regression in principle. 2. Able to articulate the change in Y per unit change in X interpretation in connection with examples.

Knols: Regression Analysis; Data Modeling; Linear Regression; Interpretation of Coefficients; Logistic Regression; Multiple Regression; Dependent and Independent Variables

3.4. Lesson 17: Regression and Diagnostic analytics

In this lesson we take the core of what we just learned about regression and apply it to several diagnostic decision problems (where we are trying to ascertain what our data can tell us about what happened).

Lesson Goal(s): 1. Understand the basic logic of asking a causal question. 2. Understanding the deep challenge of making a valid causal inference. 3. Being able to walk through and explain a valid causal inference.

Knols: Diagnostic Analytics; Causal Questions; Causal Inference; Validity of Causal Inference; Linear Regression; Logistic Regression; Multiple Regression; Dependent and Independent Variables; Confounding Variables; Correlation vs Causation; Residual Analysis; Goodness of Fit; Assumptions of Regression Models

3.5. Lesson 18: Regression and Predictive analytics

We continue to build our regression capacity in this lesson. The logic of this lesson is to see the same underlying model applied to three different situations. We might start with an insurance underwriting scenario and then move on to consumer behavior and then human resources or education admissions.

Lesson Goal(s): 1. To be able to more rigorously than before distinguish predictive analytics. 2. To be able to follow a guided code or workbook scenario to reproduce a predictive analysis.

Knols: Regression Analysis; Predictive Analytics; Machine Learning; Training and Testing Data; Model Validation; Overfitting and Underfitting; Feature Selection; Forecasting; Prediction Error Metrics e.g., Mean Squared Error, Mean Absolute Error

3.6. Lesson 19: Regression and Prescriptive analytics

In this lesson we extend our analytical reach to the decision question of "what should we do?" This takes us into the realm of optimization. We learn about the varieties of things we might maximize or minimize and the generic notion of an objective function. Focal example for the class is recommender systems. In the prep work we want students to have a chance to experience the challenge of finding global optima vs the relative ease of local optima as a way of contextualizing the big data challenge in optimization.

Lesson Goal(s): 1. To be able to explain the concept of optimization and give examples. 2. To distinguish between global and local optimization. 3. To be able to explain what a recommender system does and how this is a decision problem.

Knols: Regression; Prescriptive Analytics; Decision Optimization; Objective Function; Maximization and Minimization; Global and Local Optima; Recommender Systems; Constraint Satisfaction; Linear Programming

3.7. Lesson 20: Synthesis

Lesson Goal(s): 1. Review the Unit

4. Gathering the Right Data in the Right Way

The decision depends on the story depends on the analysis depends on the data. In this unit we learn about different types of data, properties of data, methods of measurement, sources of data, methods of data collection and generation, data cleaning, and feature engineering.

4.1. Lesson 21: Data sources and data collection methods

This lesson makes a quick nod to the underlying issues of measurement (accuracy, validity, precision, reliability) and type of measurement (nominal, ordinal, interval, and ratio) and to conventional data collection methods (survey research) but the focus in on "data exhaust" as "unobtrusive" measurement and methods of mass measurement (like buttons, etc.). The goal of the session is to impress upon the student that data collection is a thing and is a hard thing but also that the recognition of potential data sources being collected incidentally is also a thing. We telescope a later session by putting a pin in the idea of unwanted surveillance as an aspect of data collection.

Lesson Goal(s): 1. To be able to explicitly identify data collection all around us. 2. Be able to explain the four qualities of measurement, especially validity. 3) To be able to explain why raw data produced by various methods might not be ready for analysis.

Knols: KNOLS: Measurement; Accuracy; Validity; Precision; Reliability; Nominal Data; Ordinal Data; Interval Data; Ratio Data; Data Exhaust; Unobtrusive Measurement; Data Collection Ethics; Surveillance; Web Scraping; APIs; IoT Data

4.2. Lesson 22: Data quality and cleaning; structured vs unstructured data; feature engineering.

This lesson picks up on the last point of the previous lesson: raw data is rarely ready for analysis. Using an example we learn several tools and techniques for converting raw data into analyzable form.

Lesson Goal(s): 1. Able to describe basic data headaches such as date formats, spacing, special characters and how to handle them. 2. Able to follow guided workbooks and construct a usable dataset from raw data (one that will be recognized as used earlier in the course).

Knols: KNOLS: Data Cleaning; Structured Data; Unstructured Data; Feature Engineering; Missing Data Imputation; Outlier Detection; Data Wrangling Tools (e.g., Python pandas); Regular Expressions

4.3. Lesson 23: Data storage, data privacy and ethics

In this lesson we introduce the ideas of data governance, privacy, and ethics. The overarching purpose of the lesson is to inform students about the existence of this dimension of data informed decision making. They are asked to complexify the "information wants to be free" hypothesis and to reckon with the fact that tech companies and other users of big data want clear guidelines and regulation.

Lesson Goal(s): 1. To be able to conceptually distinguish between governance, privacy, and ethics. 2. To be able to articulate more than one perspective on data privacy. 3. To be able to articulate reasons that good governance and regulation can benefit individual firms and an entire industry.

Knols: Data Governance; Data Privacy; Data Ethics; Data Storage; Data Regulation; Privacy Laws (e.g., GDPR, CCPA); Data Breach; Informed Consent; Anonymization and Pseudonymization; Data Stewardship; Responsible AI; Privacy-preserving Techniques (e.g., Differential Privacy).

4.4. Lesson 24: Assignment Presentations

Students have the opportunity to practice and demonstrate their progress on the learning objectives of the course.