Storytelling with data

01 | Analytics Communication — Scopes, Audiences, Challenges

Scott Spencer | Columbia University

introductions

Meeting your professor

Education

Doctor of Jurisprudence
Honors in research and writing
Focus — analysis

Master of Science
Sports Management
Focus — data science analytics
Won, SABR analytics competition

Bachelor of Science
Chemical Engineering
Focus — numerical methods, statistical process control

Scott Spencer

Columbia University

Faculty, Lecturer, Alumnus

Consultant, Data Scientist

Professional sports
Example — Major League Baseball research and development for player performance & manager decision-making

Data for good
Example — Bayesian, generative modeling effects of climate change on perceived expectations of property values

Innovation
Example — whether invented attributes of an edible oil previously existed or was made or sold by competitor

Teaching and Research

Developing generative models
Building Bayesian, generative models to enable decision-making in complex fields such as sports performance.

Communicating uncertainty
Writing monograph on quantitative persuasion amid uncertainty. Developing R packages to tie human perception to graphical representation of data.

Contributing open-source software
Contribute to interfaces to Stan, a probabilistic programming language.

hy-flex

Class meets Wednesdays 8:10-10PM
Zoom and 329 Pupin Laboratories

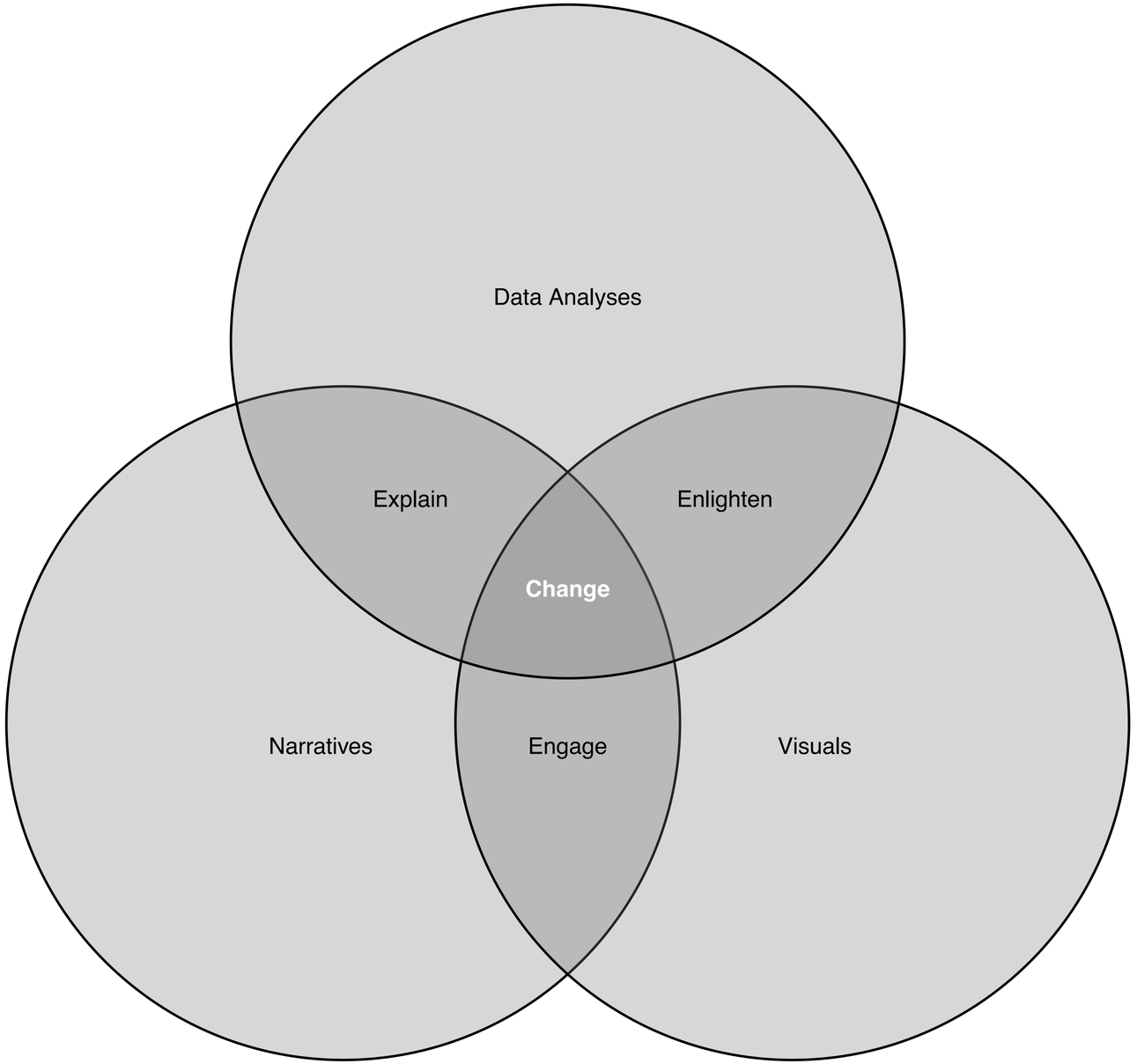
office hours

Professor Spencer
Click to schedule appointment

Associate Scherling
Click to schedule appointment

Who are your fellow students and future colleagues? **Let's say hello!**

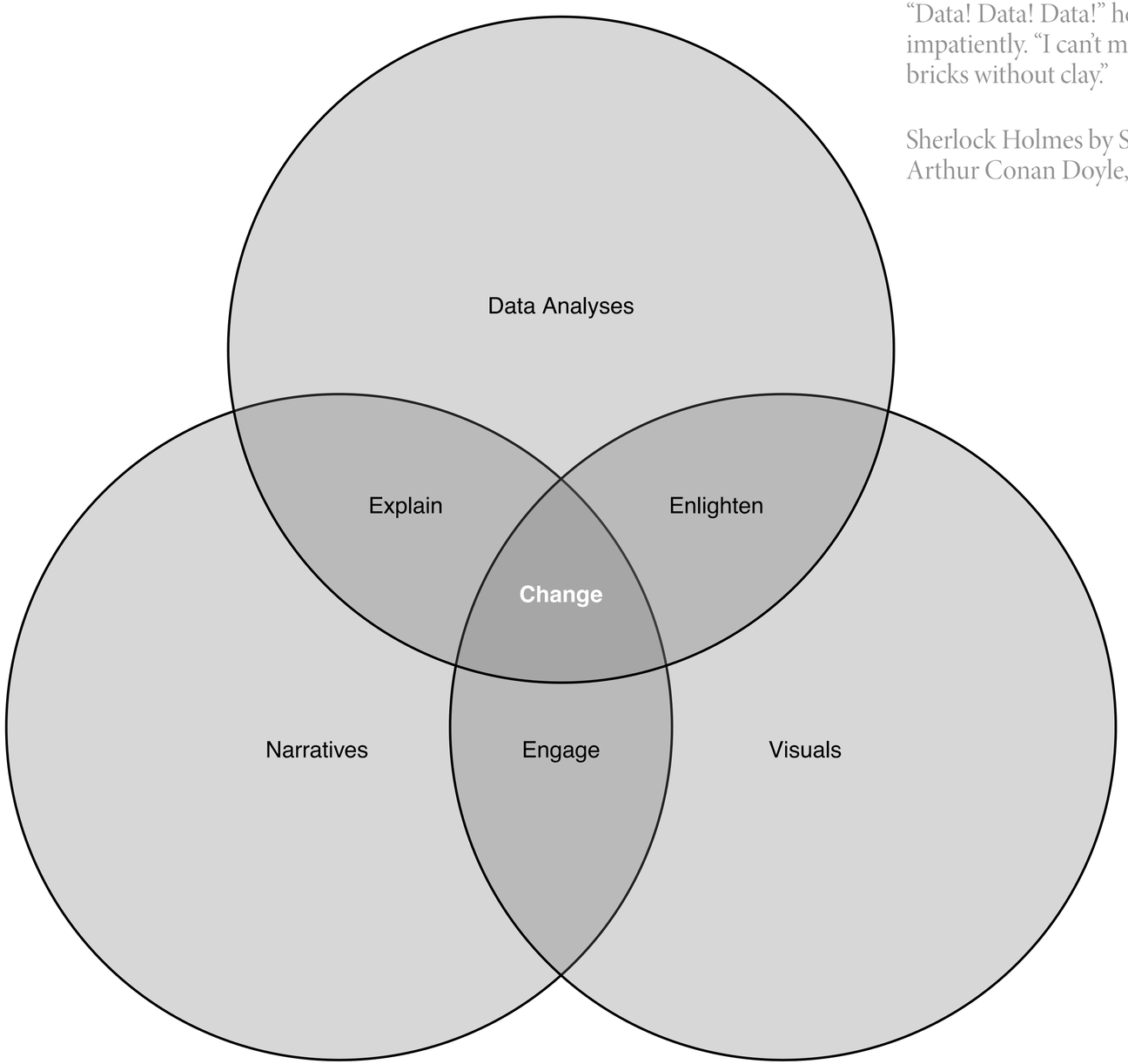
course overview


Data Analyses
Explain
Enlighten
Change
Narratives
Engage
Visuals

course overview | components

"Data! Data! Data!" he cried impatiently. "I can't make bricks without clay."

Sherlock Holmes by Sir Arthur Conan Doyle, *author*



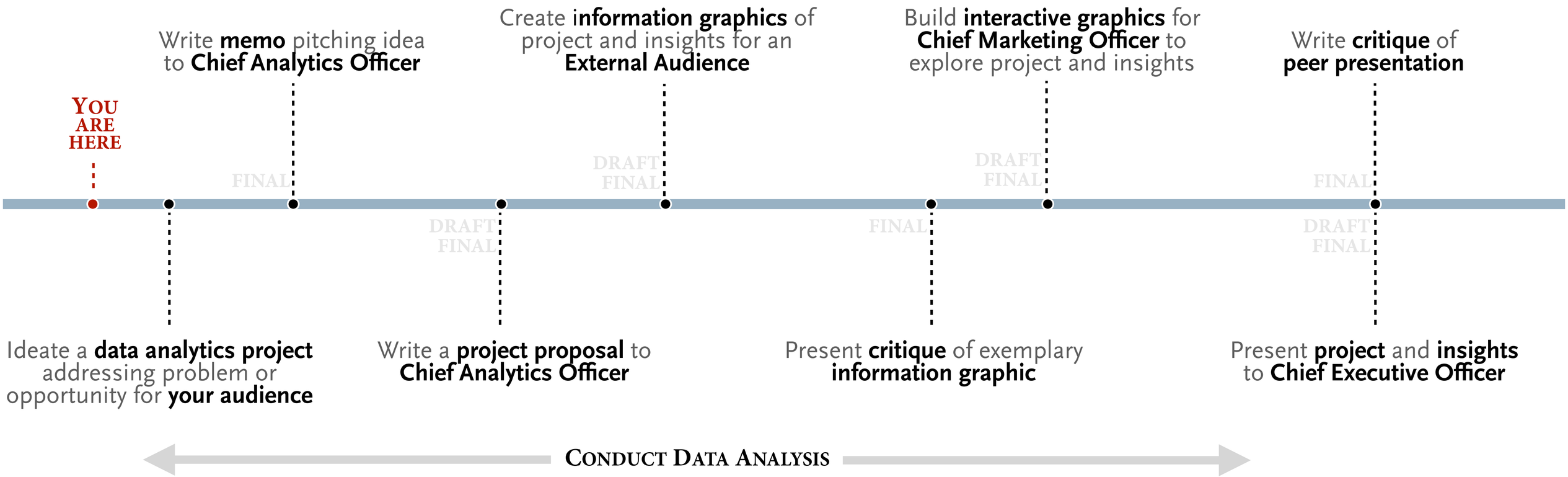
No one ever made a decision because of a number. They need a story.

Daniel Kahneman, *psychologist, behavioral economist, and author*

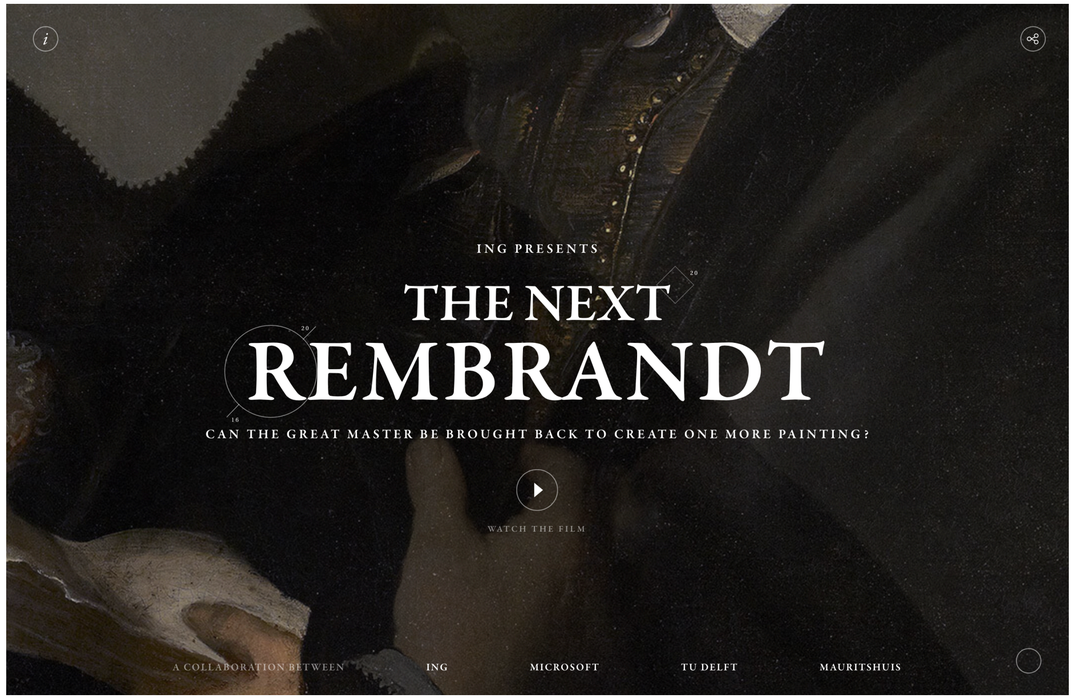
The greatest value of a picture is when it forces us to notice what we never expected to see.

John W Tukey, *mathematician*

course overview | main course deliverables



data analytics project scope



What were the **data** the analysts worked with?

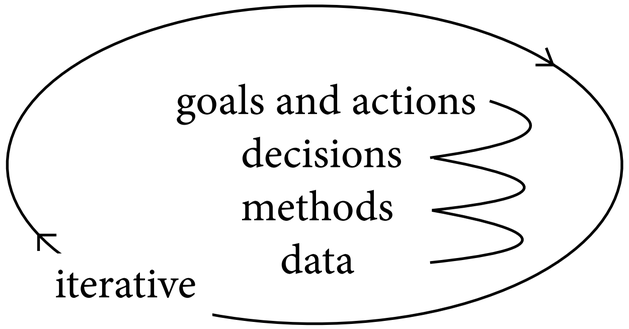
How specific were their explanations of **project scope** and **methods**?

Who may have been their **audience**?

Assuming the audience, **appropriate detail**?

Do you feel this is a **story**? Why or why not?

process



initial questions

What **problem** is to be solved?

Is the problem **important**?

Could an answer have **impact**?

Do **data** have a role in **solving the problem**?

Are the right **data available**?

In what **contexts** are the data generated?

Is the **organization** ready to **tackle the problem** and **take actions from insights**?

Rebalancing is one of the **biggest challenges** of any bike share system, especially in … New York where residents don't all work a traditional 9-5 schedule, and … people work in a variety of other neighborhoods.

— Simmons, Dani. Citi Bike spokeswoman.

Identifying events and user behavior

What **events** may be correlated with or cause empty or full bike docking stations?

What potential **user behaviors** or **preferences** may lead to these events?

From what **analogous** things could we draw **comparisons** to provide **context**?

Measurements of events and behaviors

How may these events and behaviors have been **measured and recorded**?

What **data are available**? Where? What form?

May these data be **sufficient to find insights** through analysis, useful for decisions and goals?

Examples of publicly available data sources

Data are recorded of each **bike** unlocked and docked, along with remaining **dock** capacities at the locations, dates, and times of each event: https://www.citibikenyc.com/system-data

Taxi pickup and drop-off locations and times: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Subway lines entrance/exit locations: https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49

Historical **weather**: https://www.weather.gov/documentation/services-web-api

Traffic data and more: http://www.nyc.gov/html/dot/html/about/datafeeds.shtml#realtime

analytics project scope | a few (of many, many) starting points for finding — and get help finding — data

Columbia University Library Research Data Services

Research Data Services is jointly supported by the Libraries and CUIT, providing support and consulting for research data needs at Columbia University. Our **expert staff are available to help** with many aspects of the research data lifecycle including **research**, **data management**, **finding data**, recommendations for **cleaning** and **understanding** data, **mapping** and **visualizing** your data.

https://library.columbia.edu/services/research-data-services.html

Columbia Library Clio database search

Real-time and historical SEC EDGAR filings, scanned images of company annual reports and foreign exchange filings.

https://clio.columbia.edu/databases?q=research+reports

Social media: Ravindran, Sharan Kumar, and Vikram Garg. *Mastering Social Media Mining with R*. Packt Publishing, 2015. Print. Clio: https://clio.columbia.edu/catalog/14225862

Web: Munzert, Simon et al. *Automated Data Collection with R*. Wiley, 2015. Print. Clio: https://clio.columbia.edu/catalog/11269563

R's base installation, and many R **packages** contain built-in datasets. The command `data()` gives you the base R datasets, and including the installed package name, say, `data(package="rethinking")` lists the datasets in the package.

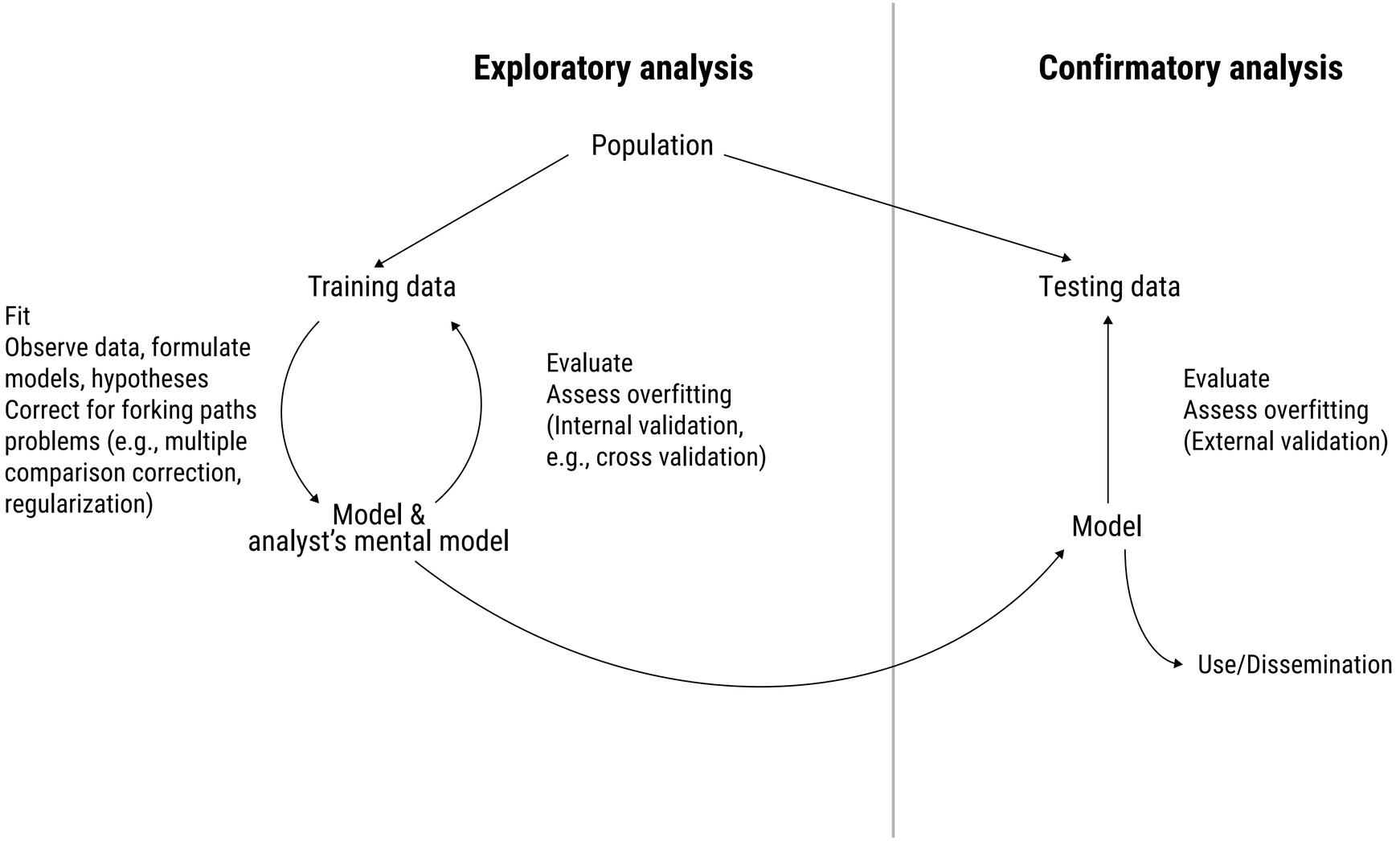
The **General Social Survey** includes more than 40 years of personal-interview survey questions on social characteristics and attitudes in the United States. http://gss.norc.org

Kaggle is an online community of data scientists owned by Google who publish data sets, over 14,000 now, for public use. https://www.kaggle.com/datasets

NYC OpenData provides public access to numerous data sets gathered from NYC agencies. https://opendata.cityofnewyork.us/data/

Data.gov is a USA federal collection of datasets. https://www.data.gov

Google Dataset Search is just like a regular Google search, but focused on datasets. https://toolbox.google.com/datasetsearch


Exploratory analysis
Confirmatory analysis
Population
Training data
Testing data
Fit
Observe data, formulate models, hypotheses
Correct for forking paths problems (e.g., multiple comparison correction, regularization)
Evaluate
Assess overfitting
(Internal validation, e.g., cross validation)
Model &
analyst's mental model
Evaluate
Assess overfitting
(External validation)
Model
Use/Dissemination

common components communicated

I. Title
II. Abstract
III. Project description
    A. Results from prior agency support
    B. Problem statement and significance
    C. Introduction and background
        - Relevant literature review
        - Preliminary data
        - Conceptual, empirical, or theoretical model
        - Justification of approach or novel methods
    D. Research plan
        - Overview of research design
        - Objectives or specific aims, hypotheses, and methods
        - Analysis and expected results
        - Timetable
    E. Broader impacts
IV. References cited
V. Budget and budget justification

I. Title
II. Abstract
III. Project description
    A. Results from prior agency support
    B. Problem statement and significance
    C. Introduction and background
        - Relevant literature review
        - Preliminary data
        - Conceptual, empirical, or theoretical model
        - Justification of approach or novel methods
    D. Research plan
        - Overview of research design
        - Objectives or specific aims, hypotheses, and methods
        - Analysis and expected results
        - Timetable
    E. Broader impacts
IV. References cited
V. Budget and budget justification

Title | accurately represents the *content* and *scope* of the proposal.

I. Title
II. Abstract
III. Project description
    A. Results from prior agency support
    B. Problem statement and significance
    C. Introduction and background
        - Relevant literature review
        - Preliminary data
        - Conceptual, empirical, or theoretical model
        - Justification of approach or novel methods
    D. Research plan
        - Overview of research design
        - Objectives or specific aims, hypotheses, and methods
        - Analysis and expected results
        - Timetable
    E. Broader impacts
IV. References cited
V. Budget and budget justification

Abstract | frames the goals and scope of the study, briefly describes the methods, and presents the hypotheses and expected results or outputs.

Sets up proper expectations, so be careful to avoid misleading readers into thinking that the proposal addresses anything other than the actual research topic.

Try for no more than two short paragraphs.

I. Title
II. Abstract
III. Project description
    A. Results from prior agency support
    B. Problem statement and significance
    C. Introduction and background
        - Relevant literature review
        - Preliminary data
        - Conceptual, empirical, or theoretical model
        - Justification of approach or novel methods
    D. Research plan
        - Overview of research design
        - Objectives or specific aims, hypotheses, and methods
        - Analysis and expected results
        - Timetable
    E. Broader impacts
IV. References cited
V. Budget and budget justification

Significance | begins with the big picture and then funnels the reader through the hypotheses to the goals or specific aims of the research.

I. Title
II. Abstract
III. Project description
    A. Results from prior agency support
    B. Problem statement and significance
    C. Introduction and background
        - Relevant literature review
        - Preliminary data
        - Conceptual, empirical, or theoretical model
        - Justification of approach or novel methods
    D. Research plan
        - Overview of research design
        - Objectives or specific aims, hypotheses, and methods
        - Analysis and expected results
        - Timetable
    E. Broader impacts
IV. References cited
V. Budget and budget justification

Literature review | sets the stage for the proposal by discussing the most widely accepted or influential papers on the research.

The key here is to provide context and be able to show where the work would extend what has been done or how it fills a gap or resolves uncertainty, etc.

I. Title
II. Abstract
III. Project description
    A. Results from prior agency support
    B. Problem statement and significance
    C. Introduction and background
        - Relevant literature review
        - Preliminary data
        - Conceptual, empirical, or theoretical model
        - Justification of approach or novel methods
    D. Research plan
        - Overview of research design
        - Objectives or specific aims, hypotheses, and methods
        - Analysis and expected results
        - Timetable
    E. Broader impacts
IV. References cited
V. Budget and budget justification

Preliminary data | can help establish credibility, likely success, or novelty of the proposal.

But avoid overstating the implications of the data or suggesting you've already solved the problem.

I. Title
II. Abstract
III. Project description
  A. Results from prior agency support
  B. Problem statement and significance
  C. Introduction and background
    - Relevant literature review
    - Preliminary data
    - Conceptual, empirical, or theoretical model
    - Justification of approach or novel methods
  D. Research plan
    - Overview of research design
    - Objectives or specific aims, hypotheses, and methods
    - Analysis and expected results
    - Timetable
  E. Broader impacts
IV. References cited
V. Budget and budget justification

Research plan | The goal is to keep the reader focused on the overall significance, objectives, specific aims, and hypotheses while providing important methodological, technological, and analytical details.

Contains the details of the implementation, analysis, and inferences of the study.

Convince the reader that the project can be accomplished.

I. Title
II. Abstract
III. Project description
  A. Results from prior agency support
  B. Problem statement and significance
  C. Introduction and background
    - Relevant literature review
    - Preliminary data
    - Conceptual, empirical, or theoretical model
    - Justification of approach or novel methods
  D. Research plan
    - Overview of research design
    - Objectives or specific aims, hypotheses, and methods
    - Analysis and expected results
    - Timetable
  E. Broader impacts
IV. References cited
V. Budget and budget justification

Objectives, hypotheses, aims, methods |

Objectives refer to broad, scientifically far-reaching aspects of a study, while *hypotheses* refer to a more specific set of testable conjectures. Specific *aims* focus on a particular question or hypothesis and the *methods* needed and outputs expected to fulfill the aims.

Of note, these points will typically have already been briefly introduced earlier, *e.g.*, in the abstract. Bring in more detail here.

I. Title
II. Abstract
III. Project description
    A. Results from prior agency support
    B. Problem statement and significance
    C. Introduction and background
        - Relevant literature review
        - Preliminary data
        - Conceptual, empirical, or theoretical model
        - Justification of approach or novel methods
    D. Research plan
        - Overview of research design
        - Objectives or specific aims, hypotheses, and methods
        - Analysis and expected results
        - Timetable
    E. Broader impacts
IV. References cited
V. Budget and budget justification

Analysis and expected results | If early data are available, show how you will analyze them to reach your objectives or test your hypotheses.

If such data are unavailable, consider culling data from the literature to show how you expect the results to turn out and to show how you will analyze your data when they are available.

Complete a table or diagram, or run statistical tests using the preliminary or "synthesized" data. This can be a good way to show how you would interpret the results of such data.

audiences *we* will consider in this course

our audiences | c-suite executives, general audiences, mixed audiences

Analytics Executives

Lead an organization's data analytics strategy, driving data-related business changes to transform company into a more analytics-driven one.

Chief Executives

Leads management of company; responsible for maximizing company value, high-level decisions on policy and strategy; drives change.

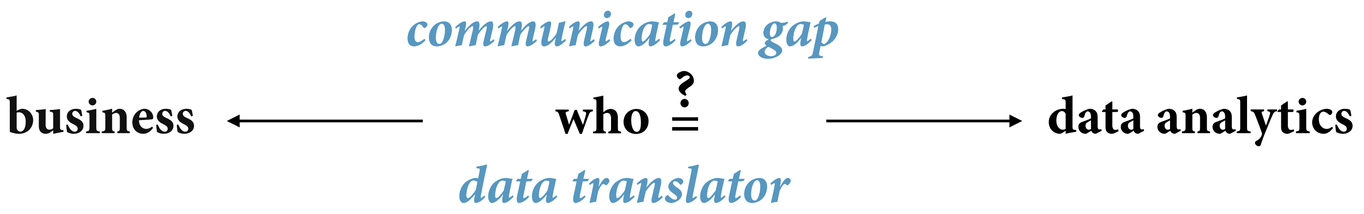
Marketing Executives

Lead responses to changing circumstances; shapes products, sales strategies, and marketing ideas, collaborating across the company.

General and Mixed Audiences

The most challenging audiences to understand and develop persuasive messages.

analytics communication challenges


communication gap
business
who ?=
data analytics
data translator

challenges | bridging the gaps with data translators, qualities needed

project management

data wrangling

data analysis

subject expertise

design

storytelling

resources

References

Spencer, Scott. "Analytics Communication Scopes" and "Audiences and Challenges." In *Data in Wonderland.* 2021. https://ssp3nc3r.github.io/data_in_wonderland.

Berinato, Scott. "*Data Science & the Art of Persuasion*." Harvard Business Review, December 2018, 1–13.

Brady, Chris, Mike Forde, and Simon Chadwick. "*Why Your Company Needs Data Translators*." MIT Sloan Management Review, March 2017, 1–6.

Friedland, Andrew J., Carol L. Folt, and Jennifer L. Mercer. *Writing Successful Science Proposals*. Third edition. New Haven: Yale University Press, 2018.

Friedman, Matthew. "*Citi Bike Racks Continue to Go Empty Just When Upper West Siders Need Them*." News. West Side Rag (blog), August 19, 2017. https://www.westsiderag.com/2017/08/19/citi-bike-racks-continue-to-go-empty-just-when-upper-west-siders-need-them.

Ing, Microsoft, Tu Delft, and Mauritshuis. "*The Next Rembrandt*." The Next Rembrandt, April 2016. https://www.nextrembrandt.com.

Maynard-Atem, Louise, and Ben Ludford. "*The Rise of the Data Translator*." Impact 2020, no. 1 (January 2, 2020): 12–14. https://doi.org/10.1080/2058802X.2020.1735794.

National Science Foundation. *A Guide for Proposal Writing* / National Science Foundation, Directorate for Education and Human Resources, Division of Undergraduate Education. National Science Foundation, 1998.

Oruc, A Yavuz. *Handbook of Scientific Proposal Writing*. CRC Press, 2011.

Oster, Sandra, and Paul Cordo. *Successful Grant Proposals in Science, Technology and Medicine: A Guide to Writing the Narrative*. Cambridge ; New York: Cambridge University Press, 2015.

Saldarriaga, Juan Francisco. "*CitiBike Rebalancing Study.*" Spatial Information Design Lab, Columbia University, 2013. https://c4sr.columbia.edu/projects/citibike-rebalancing-study.

Schimel, Joshua. *Writing Science: How to Write Papers That Get Cited and Proposals That Get Funded*. Oxford ; New York: Oxford University Press, 2012.

Zetlin, Minda. "*What Is a Chief Analytics Officer? The Exec Who Turns Data into Decisions.*" CIO, November 2, 2017.