projectMOSAIC
UNC CHARLOTTE

Twitter Text Analytics

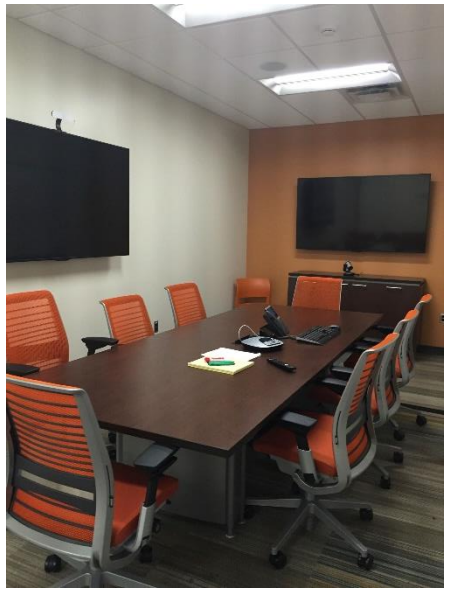
Project Mosaic Workshop

Date 9/27 – 9/29/2016

Ryan Wesslen, Computing & Informatics

rwesslen@uncc.edu

Project Mosaic



- Project Mosaic: What do we do?
  - Build research methods capability in social sciences
  - Facilitate research across social science disciplines
  - Promote social science research
- Project Mosaic Services
  - Social sciences research incubator
    - Facilitate connections
    - Bring people together to exchange ideas and pursue external funding
    - Information sharing on research funding opportunities
  - Consulting
    - Free to UNC Charlotte faculty, staff and graduate students
  - Workshops
    - Open to entire campus community
    - Provides cutting-edge tools for research and a forum for researchers to network within campus

projectMOSAIC UNC CHARLOTTE

Workshop Agenda

- Day 1: Why Study Twitter?
  - Case Study to find Panther Tweets
- Day 2: Intro to Text Analysis with quanteda
  - Presidential Candidate Tweets
- Day 3: Topic Modeling with topicmodels
  - Charlotte Twitter dataset

projectMOSAIC
UNC CHARLOTTE

Workshop materials

All workshop materials can be found here:

www.github.com/wesslen/fall-2016-pm-twitter-text

projectMOSAIC
UNC CHARLOTTE

Day 1: Intro to Twitter Text

projectMOSAIC
UNC CHARLOTTE

Why research Twitter and social media?

1. Measurement
2. Micro vs Macro
3. Simultaneity Problem

Ways to analyze Twitter and Social Media

1. Text
2. Social Network
3. Geospatial

Our focus for our workshops focuses on Text.

projectMOSAIC
UNC CHARLOTTE

Presidential Campaign Case Study

Sneak preview for case study on Day 2

projectMOSAIC
UNC CHARLOTTE

Can Rhetoric explain voter choice?

Rhetoric

- Candidate Speeches
- Candidate Twitter mentions

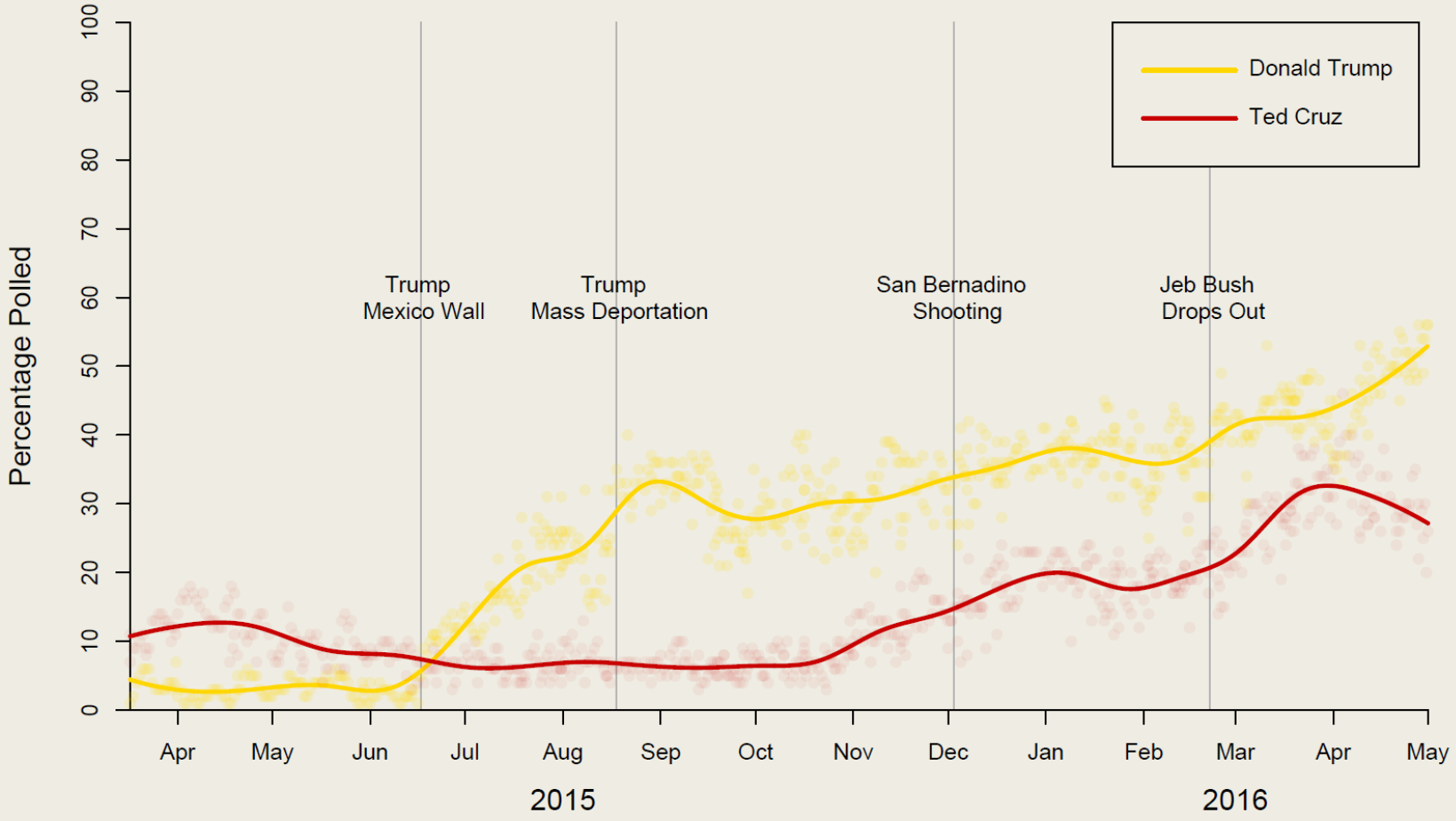
Candidates: Clinton, Cruz, Sanders and Trump

Voter Choice

- Polls
- Money fundraised
- Voter Awareness (Google Trends, Twitter mentions)

- How do people update their vote preferences for a candidate?
- Consider theory of “enlightened preferences” (Gelman and King 1993)

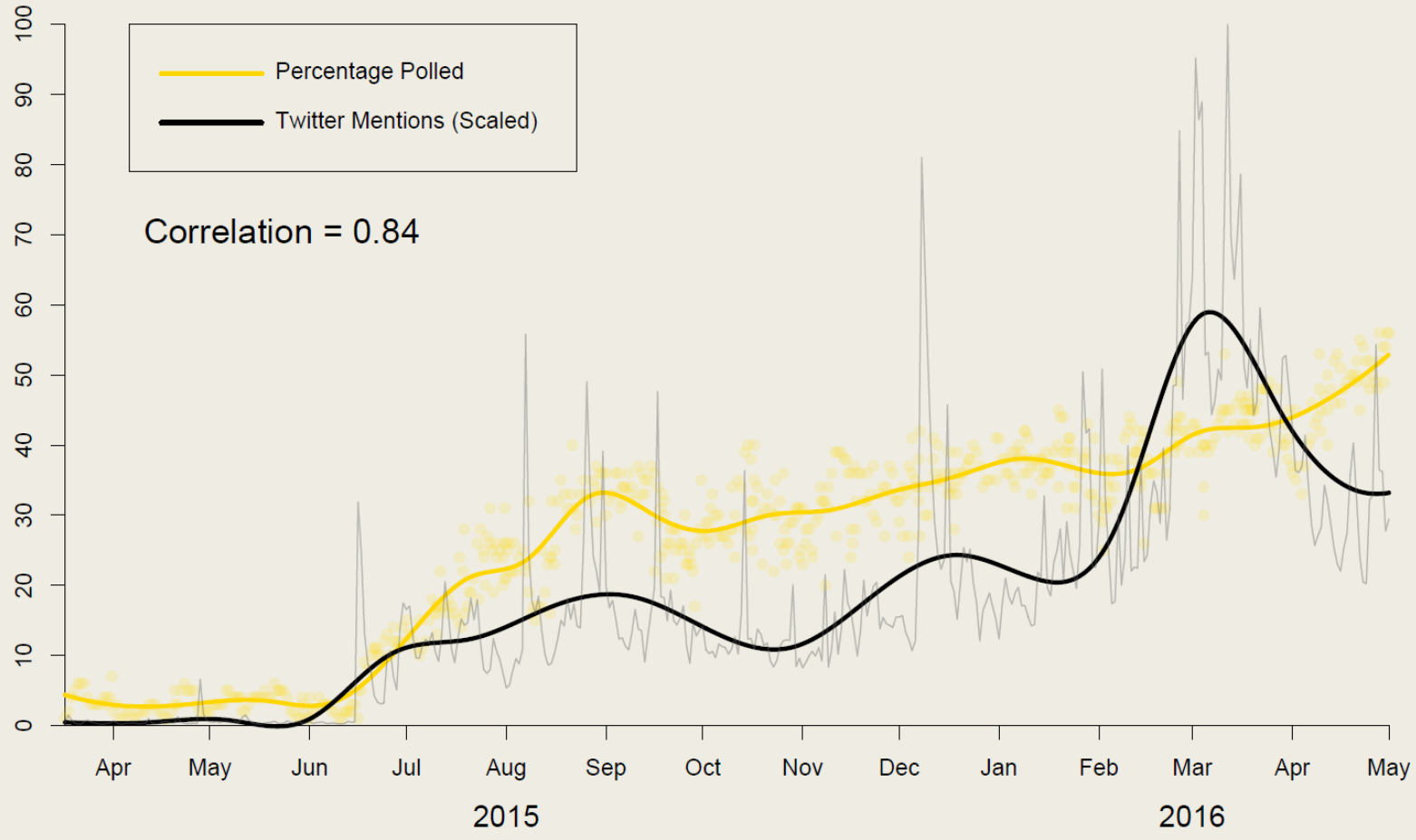
Donald Trump vs Ted Cruz polls



Data Source: http://fivethirtyeight.com/

projectMOSAIC
UNC CHARLOTTE

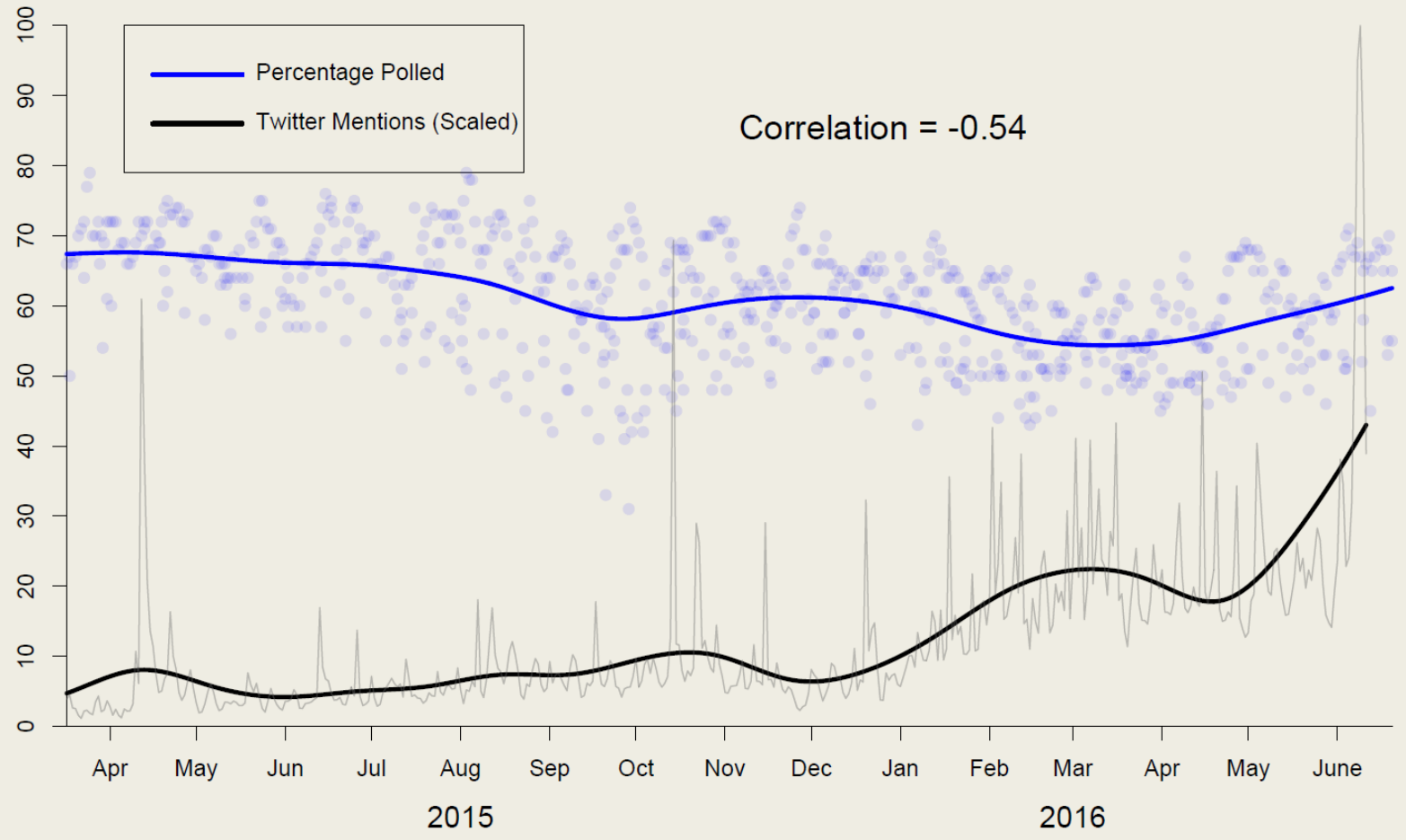
Trump polls vs Twitter mentions



Twitter Data Source: Gnip Full Archive Search

projectMOSAIC
UNC CHARLOTTE

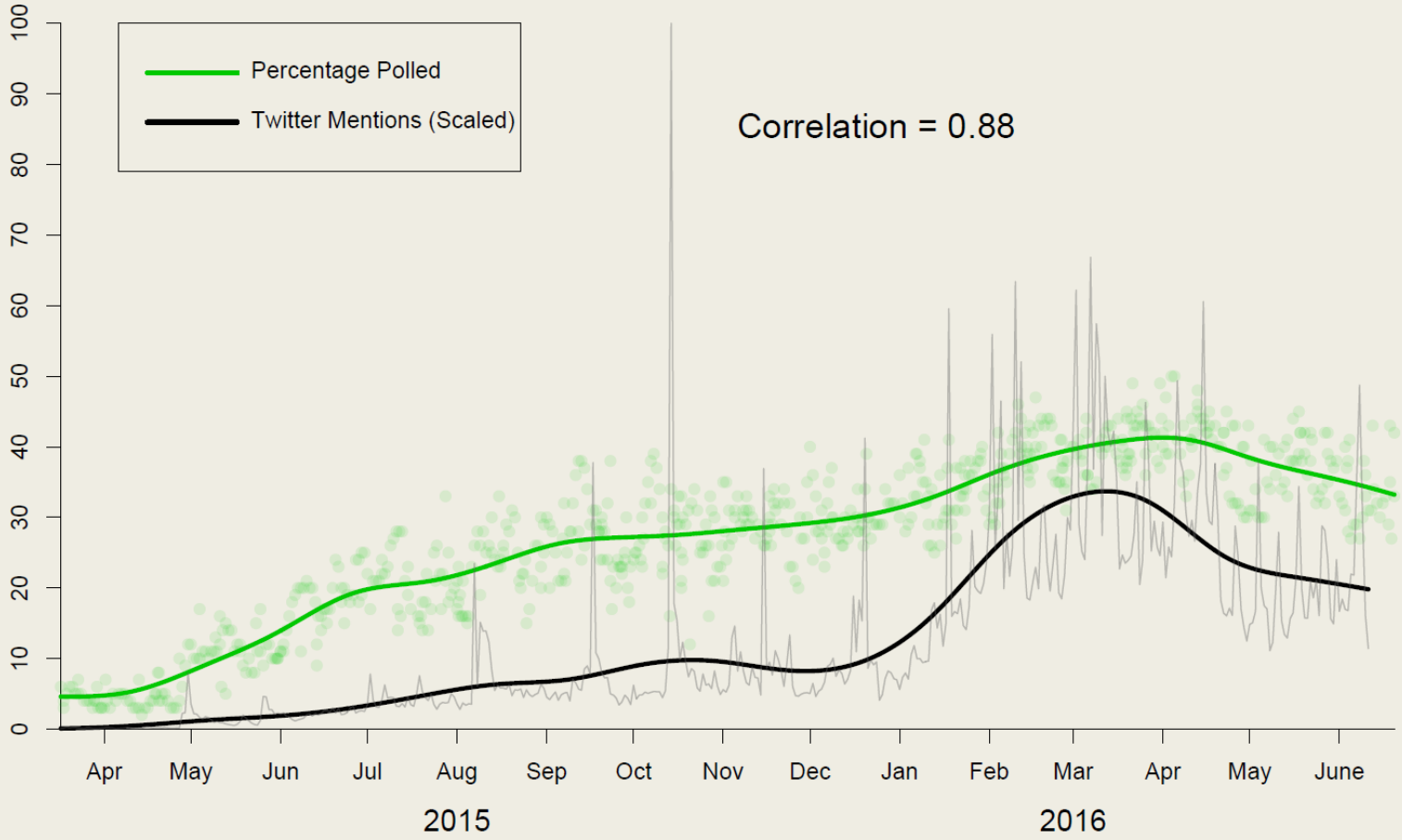
Hillary polls vs Twitter mentions



Twitter Data Source: Gnip Full Archive Search

projectMOSAIC
UNC CHARLOTTE

Bernie polls vs Twitter mentions

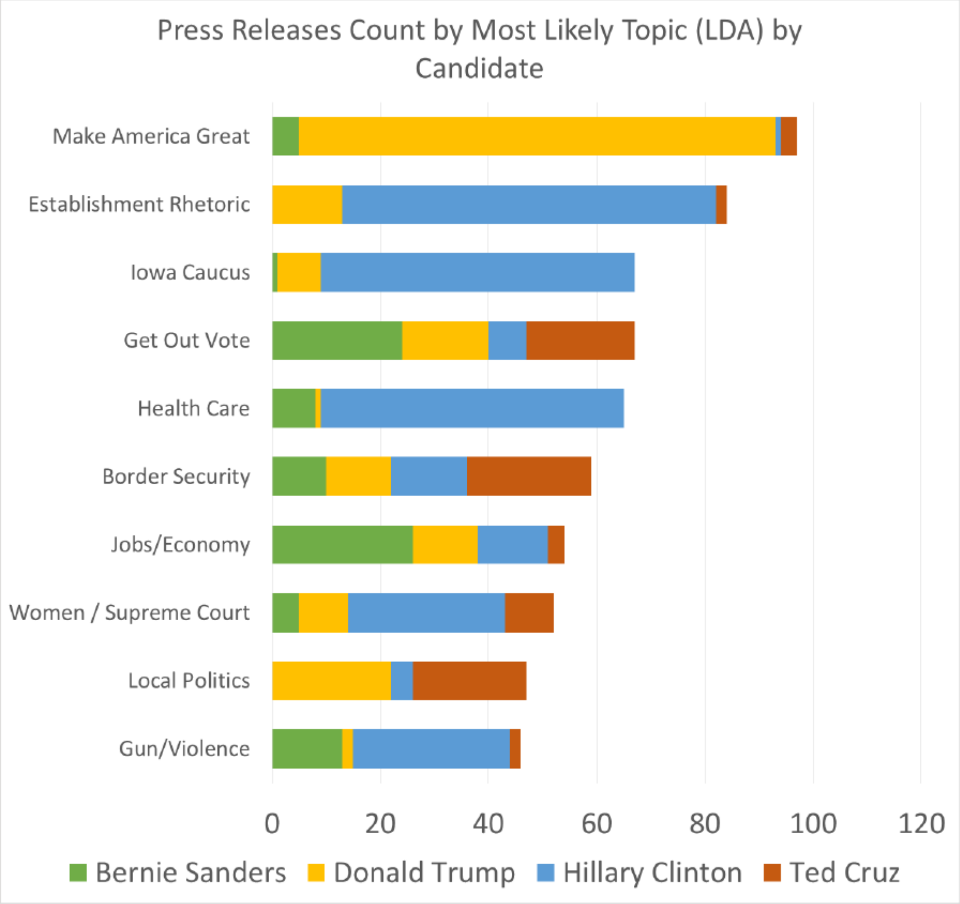


Twitter Data Source: Gnip Full Archive Search

Is Hillary the insider candidate?

Press Release Comparison Cloud

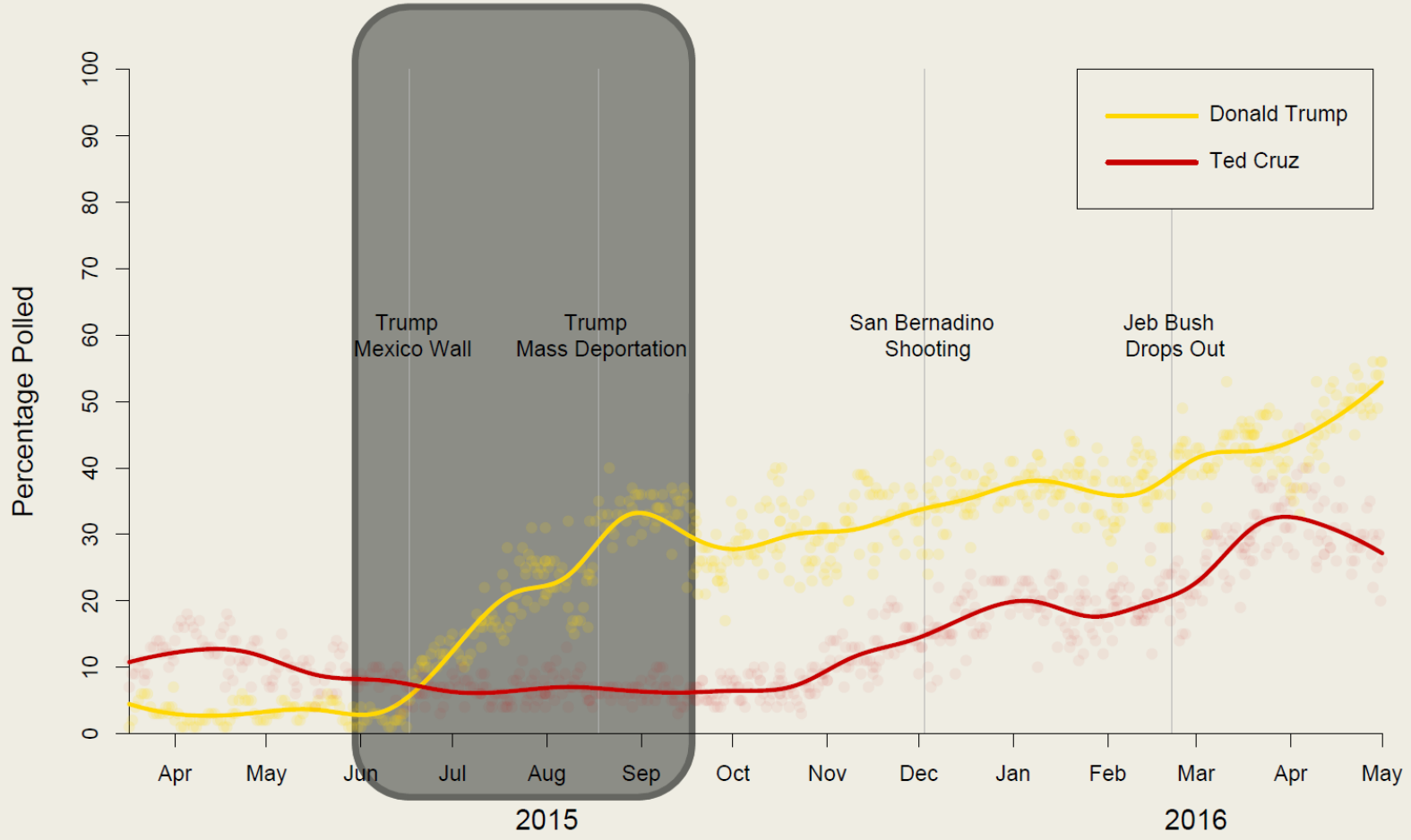
Press Release Topics by Candidate for 10 Topic LDA



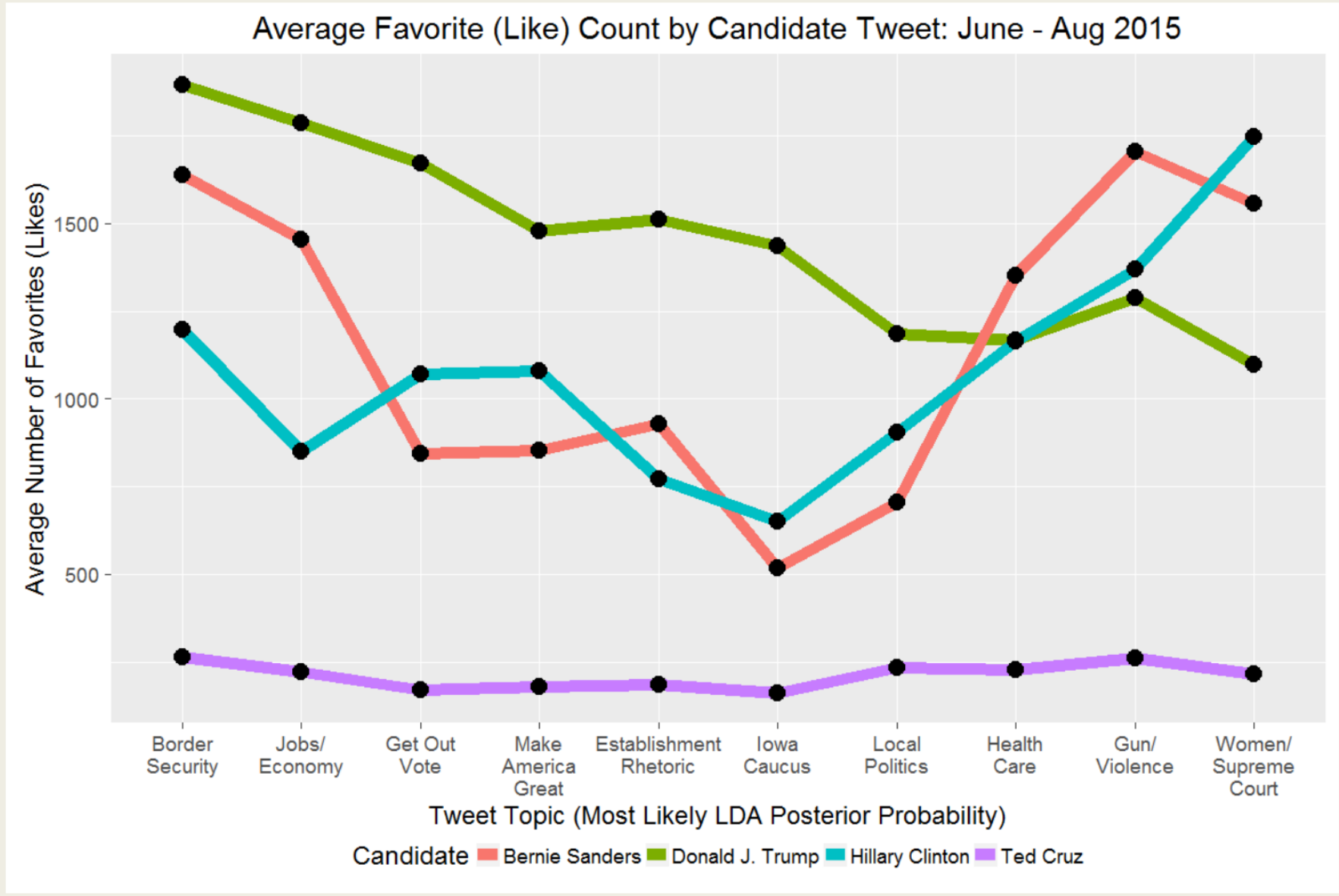
Data source: USCB Presidential Website

See Appendix for topic words, counts

Donald Trump vs Ted Cruz polls



Data Source: http://fivethirtyeight.com/


Average Favorite (Like) Count by Candidate Tweet: June - Aug 2015
Average Number of Favorites (Likes)
1500
1000
500
Border Security
Jobs/ Economy
Get Out Vote
Make America Great
Establishment Rhetoric
Iowa Caucus
Local Politics
Health Care
Gun/ Violence
Women/ Supreme Court
Tweet Topic (Most Likely LDA Posterior Probability)
Candidate
Bernie Sanders
Donald J. Trump
Hillary Clinton
Ted Cruz

Case Study: Day 1

- We will examine a 20% sample of about 250k Geolocated Tweets from the Charlotte area for three months (Dec 2015 – Feb 2016).

- The exercise can be found on the github site:

https://github.com/wesslen/fall-2016-pm-twitter-text

projectMOSAIC
UNC CHARLOTTE

Day 2: Quanteda Package

Intro to text analysis in R

projectMOSAIC
UNC CHARLOTTE

Text Analysis with the Quanteda Package

We will review the quanteda package documentation found here:

- https://cran.r-project.org/web/packages/quanteda/vignettes/quickstart.html

Case Study: Day 2

- We will examine Tweets from the four major Presidential Candidates (Clinton, Trump, Sanders, Cruz) from June 2015 – June 2016.

- The exercise can be found on the github site:

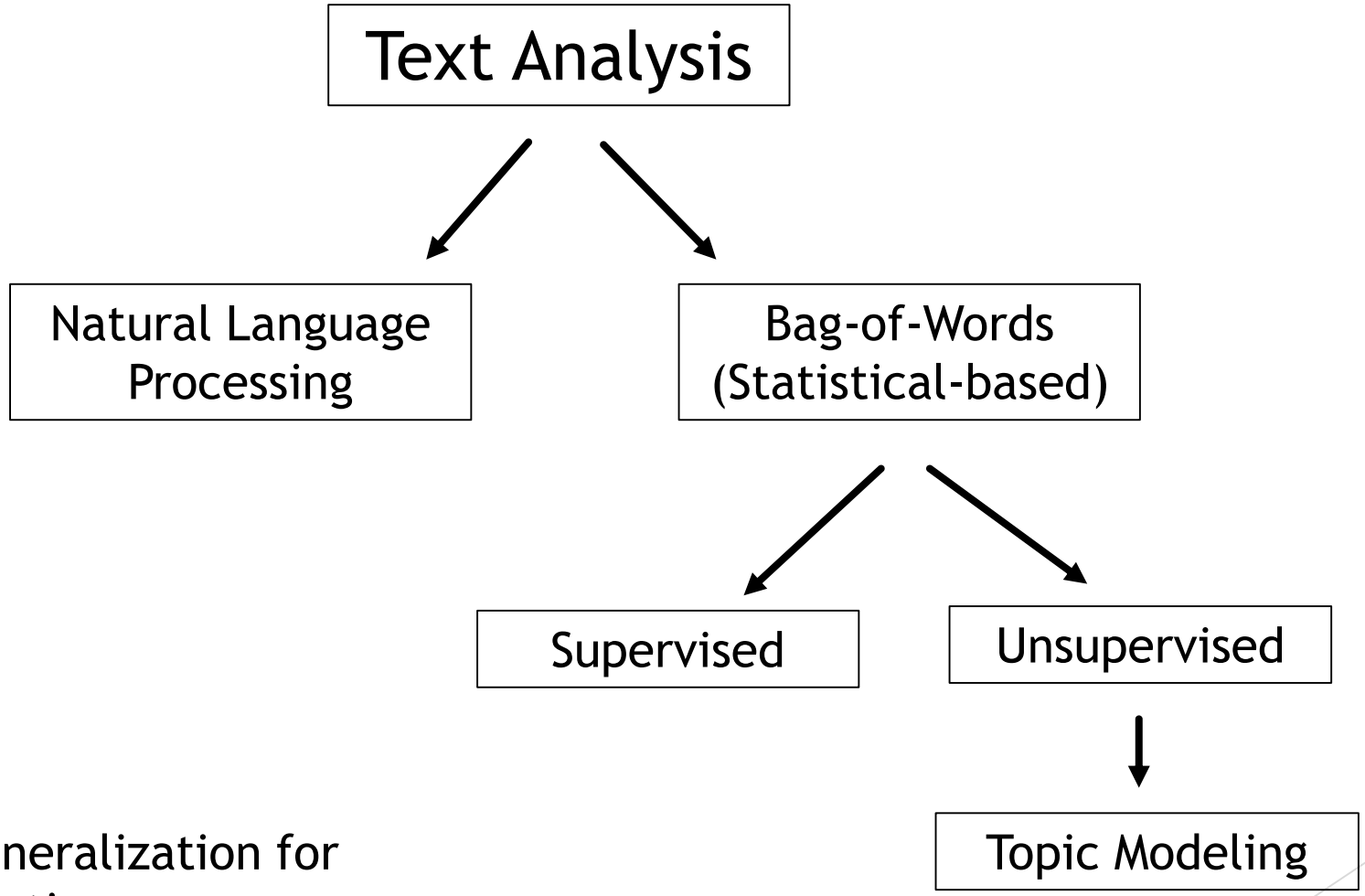
https://github.com/wesslen/fall-2016-pm-twitter-text

projectMOSAIC
UNC CHARLOTTE

Day 3: topicmodels Package

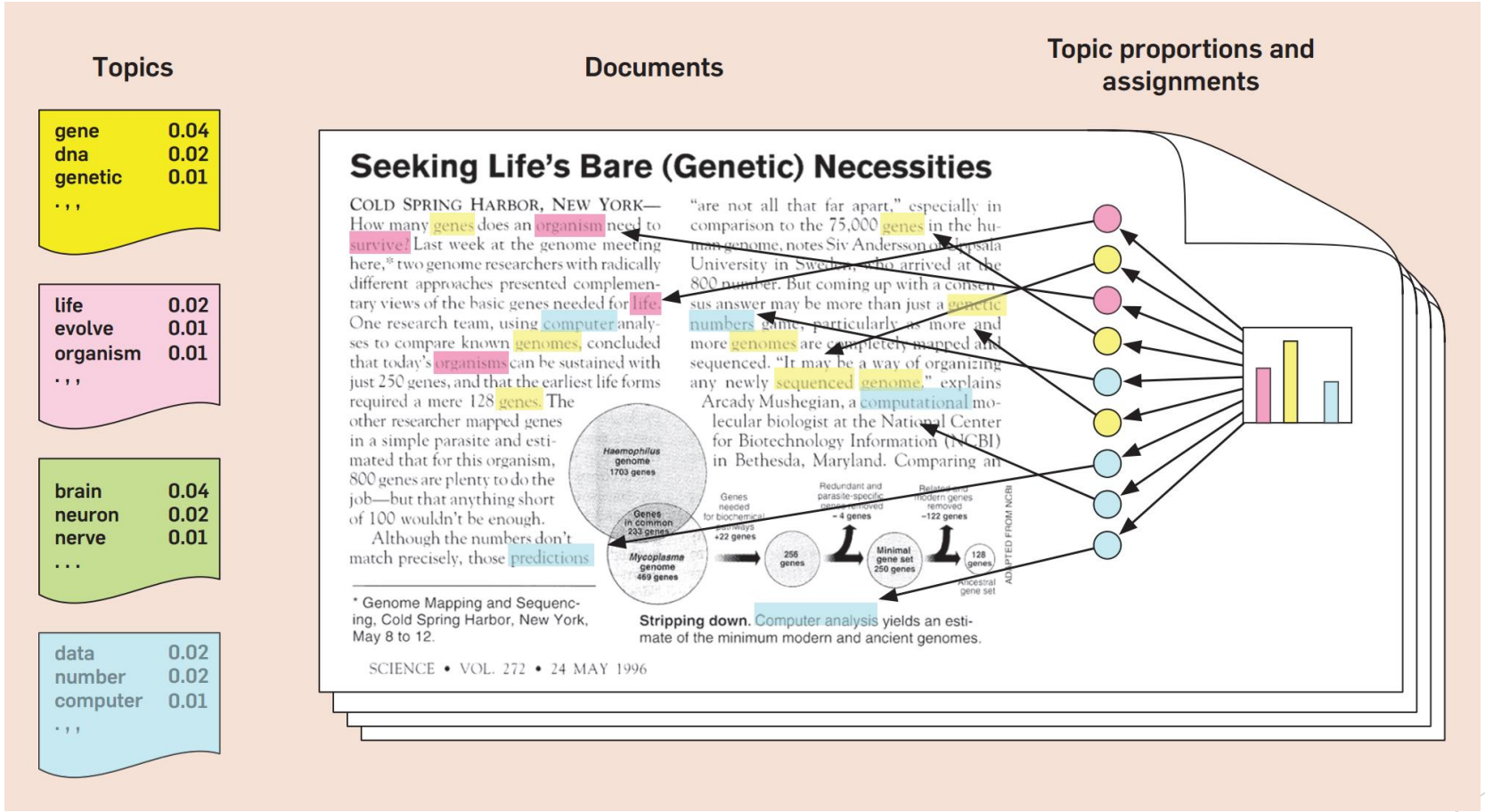
Intro to Topic Modeling (LDA) in R

Topic Modeling



Overgeneralization for illustrative purposes

Latent Dirichlet Allocation (LDA)



David Blei's "Probabilistic Topic Models" (2012)

projectMOSAIC UNC CHARLOTTE

topicmodels package in R

- We'll focus on a hands on introduction.
- For more detailed documentation, see: https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf
- Also, we'll try to run LDAVis package:

Sample: https://gallery.shinyapps.io/LDAelife/

projectMOSAIC UNC CHARLOTTE

Case Study: Day 3

- We will examine the same 20% sample Charlotte Geolocated dataset used in Day 1.

- The exercise can be found on the github site:

https://github.com/wesslen/fall-2016-pm-twitter-text

More About Services

- Research Incubator
  - Affiliates Program
    - Faculty Affiliates are hand picked for their research expertise
    - Our affiliates leverage the core functionality and expertise of Project Mosaic
  - Seed Grants Program
    - Geared towards the formation of new teams of researchers in the social, behavior and economic sciences
    - Aim is to pursue external funding
- Consulting
  - Project Mosaic offers three types of consulting:
    - Software-centric
    - Dissertation/thesis assistance
    - Research collaboration

Make an appointment on our website!

- Workshops
  - Our workshops fulfill a commitment to enhance data literacy and analytical capabilities of UNC Charlotte researchers

Find workshops online on our Events List.

projectMOSAIC UNC CHARLOTTE

Contact Project Mosaic

- Jean-Claude Thill is the director of Project Mosaic. A broadly trained geographer, he is a 'Knight' Distinguished Professor of Public Policy at UNC Charlotte.
- Contact Jean-Claude:
  - Email: Jean-Claude.Thill@uncc.edu
  - Phone: 704-687-5931 ext. 75909

- Leonora is the Administrative Support for Project Mosaic. She manages our not-so-massive paperwork, coordinates meetings and assists with administrative functions.
- Contact Leonora:
  - Email: projectmosaic@uncc.edu
  - Phone: 704-687-5931

Visit our website!
Projectmosaic.uncc.edu

UNC CHARLOTTE
Project Mosaic: A Social Sciences Research Initiative
College of Liberal Arts & Sciences

Additional Resources: Consultants

- Shaoyu Li is the head consultant in the Center of Statistics and Applied Mathematics Consulting Center (CSAMC) and works with Project Mosaic to coordinate consulting requests for statistical and mathematical expertise.
- Contact Shaoyu:
  - Email: shaoyu.li@uncc.edu

- Kailas Venkitasubramanian is a research methodologist and manages the consulting service and the workshop program of Project Mosaic. Kailas is experienced in a variety of applied statistical techniques and works fluently on multiple software platforms.
- Contact Kailas:
  - Email: kvenkita@uncc.edu

projectMOSAIC
UNC CHARLOTTE

Questions?