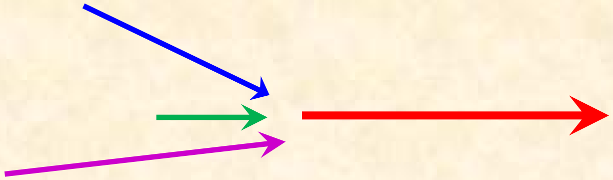
How to Statistically Model Processes?

Statistical discourse analysis



Ming Ming Chiu

University at Buffalo, State University of New York

mingchiu@buffalo.edu

Ask questions via CHAT

Feel free to ask questions at any time.

To reduce your wait time,
Type your questions into the chat.

Types of Research Questions

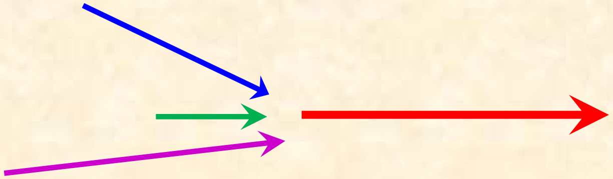
What affects people's actions/processes?

- One student's use of strategies across problems?
- Teachers' sequences of lessons and reflections?
- Classroom conversations?

Choose a research question to explore

How would you address the following issues?

How to Statistically Model Processes?



- Predict whether an action occurs or not
- Smaller unit of analysis
- Analyze time
- Contextual differences
- Complex codes, Missing data, Rare events…

Predict Whether an Action Occurs

- "Is vs. is not" (0 vs. 1) variables

  - Use strategy vs. not
  - Reflect on student motivation vs. not
  - Ask question vs. not

  Use *Logit / Probit*

- Predicting many actions?

  Use *Multivariate Logit / Probit*

Smaller Unit of Analysis

- Unit smaller than individual
  - Strategies of students
  - Reflective notes of teachers
  - Conversation turns of people
- Increase sample size
- Use *Multi-level analysis*

  (aka *Hierarchical Linear Modeling*)

Analyze Time

- Statistically identify critical moments that divide a session into distinct time periods

  - Use *Breakpoint analysis*

- How do sequences of actions/events affect the likelihood of a subsequent event? a, b, c → d?

  - Micro-time context effects

  - Use *Vector Auto-Regression (VAR)* and *Serial correlation test*

- Causal mechanisms A → B → C

  - Use *Multilevel mediation tests* or *Structural Equation Modeling*

Contextual Differences

- Different contexts

  - Micro-time contexts/recent actions

  - Different groups and individuals

  - Different time periods

  - Different settings

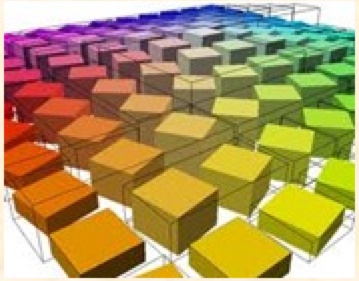
- Test Cross-level interactions via

  Multilevel Slope/Intercept Random Effects

Other Issues

- Model complex categories with

  Multi-dimensional coding



- Estimate missing data with

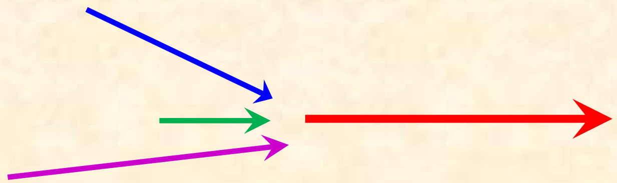
  Markov Chain Monte Carlo

  Multiple Imputation

- Model rare actions/events

  with *Logit bias estimator*

How to Statistically Model Processes?



- Predict whether an action occurs or not
- Smaller unit of analysis
- Analyze time
- Contextual differences
- Complex codes, Missing data, Rare events…

Thank You!

Statistical Discourse Analysis

4 types of Analytic Difficulties

- Time
- Outcomes
- Explanatory variables
- Data set

Statistical Discourse Analysis

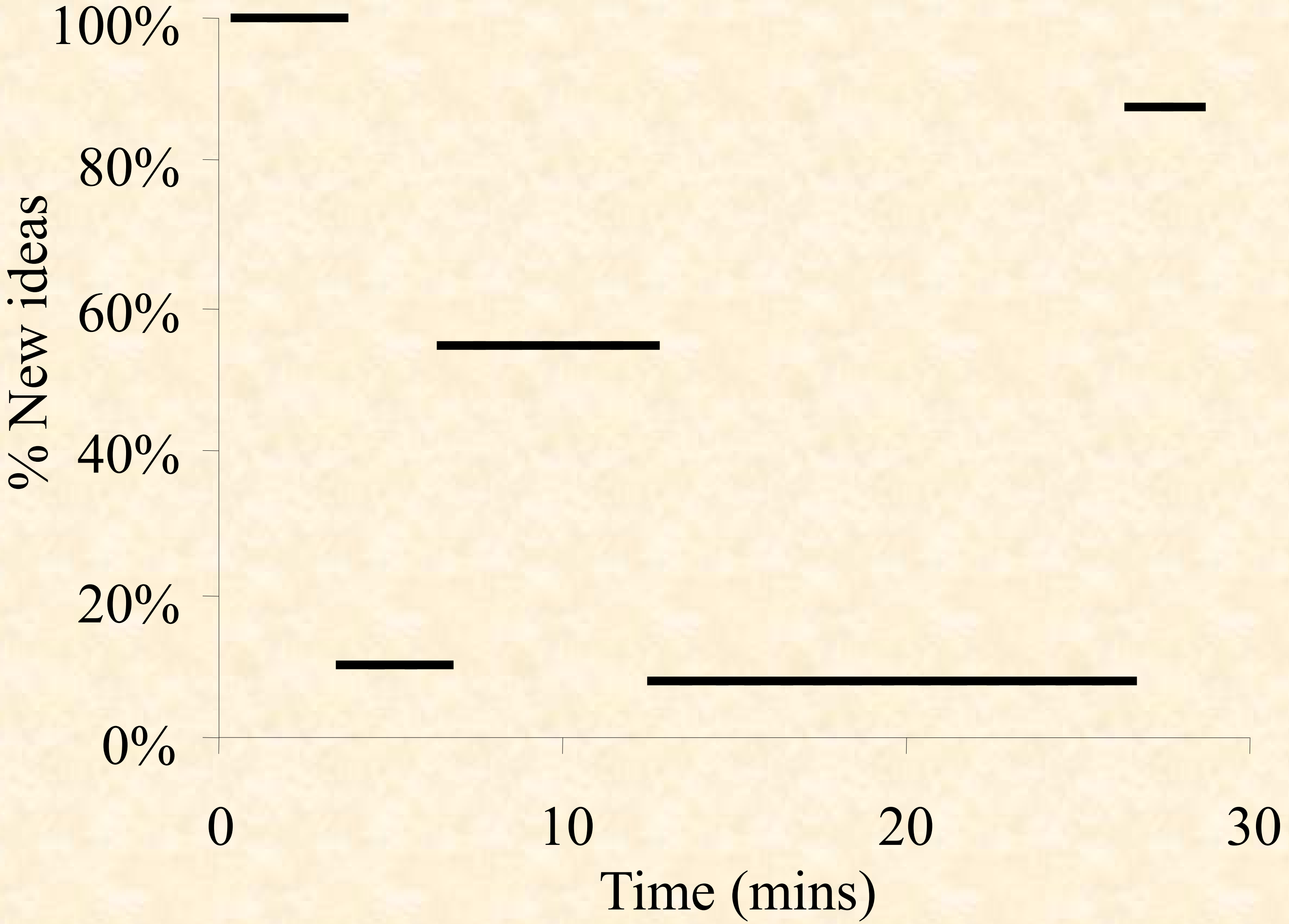
Difficulties regarding Time

- Time periods differ ($T_2 \neq T_4$)
- Serial correlation ($t_8 \rightarrow t_9$)

Strategies

- Breakpoint analysis

Breakpoints in 1 group



Statistical Discourse Analysis

Difficulties regarding Time

- Time periods differ ($T_2 \neq T_4$)
- Serial correlation ($t_8 \rightarrow t_9$)

Strategies

- Breakpoint analysis
- Multilevel analysis (MLn, HLM)
- Test with Q-statistics
- Model with lag outcomes

  e.g. Justify (-1)

Statistical Discourse Analysis

Outcome Difficulties	Strategies
• Discrete outcomes (Yes / No)	• Logit / Probit
• Multiple outcomes (Y_1, Y_2) New idea & Justify	• Multivariate, multilevel analysis

Statistical Discourse Analysis

Explanatory model Difficulties

- People & Groups differ 🚹 ≠ 🚺

- Mediation effects ($X \rightarrow M \rightarrow Y$)

- False positives ($+ + \cancel{\oplus} +$)

- Effect across turns ($X_6 \rightarrow Y_9$)

Effects across several turns

Ben: 10 times 18 is 2 speakers ago = (– 2)

Eva: 28. 1 speaker ago = (– 1)

Jay: Wrong, 180 dollars.

Statistical Discourse Analysis

Explanatory model Difficulties

- People & Groups differ 🚹 ≠ 🚺
- Mediation effects (X→M→Y)
- False positives (+ + $\not\oplus$ +)
- Effect across turns ($X_6 \rightarrow Y_9$)

Strategies

- Multilevel cross-classification
- Multilevel mediation tests
- 2-stage linear step-up method
- Vector Auto-Regression (VAR)

  Lag explanatory variables
  e.g., Disagree (-1), Girl (-1)
  Disagree (-2)

Statistical Discourse Analysis

Data Difficulties

- Missing data (101?001?10)
- Robustness

Strategies

- Markov Chain Monte Carlo multiple imputation
- Separate outcome models
- Use data subsets
- Use original data

Content analysis

Jay: A hundred eighty dollars.

Ben: If we multiply by ten cents, don't we get a hundred and eighty cents?

- Ben
  - Disagrees politely
  - New information
  - Correct
  - Justifies
  - Question

Multi-dimensional Coding

Evaluation of the previous action

- Agree (+), Neutral (Ø), Ignore/New topic (*), Disagree rudely (—), Disagree politely (–)

Knowledge content regarding problem

- New idea (N), Old idea (O), Null-content (**{}**)

Validity

- Correct (√), Wrong (**X**), Null-content (**{}**)

Justification

- Justify (**J**), No justification (**[]**), Null-content (**{}**)

Invitation to participate

- Command (**!**), Question (**?**), Statement (**_.**)

Invitational Form Decision Tree

Minimize Number of Coding Decisions to ↑ inter-coder reliability

- Minimize Depth of decision tree
- Put highly likely actions at the top

Do any of the clauses proscribe an action?

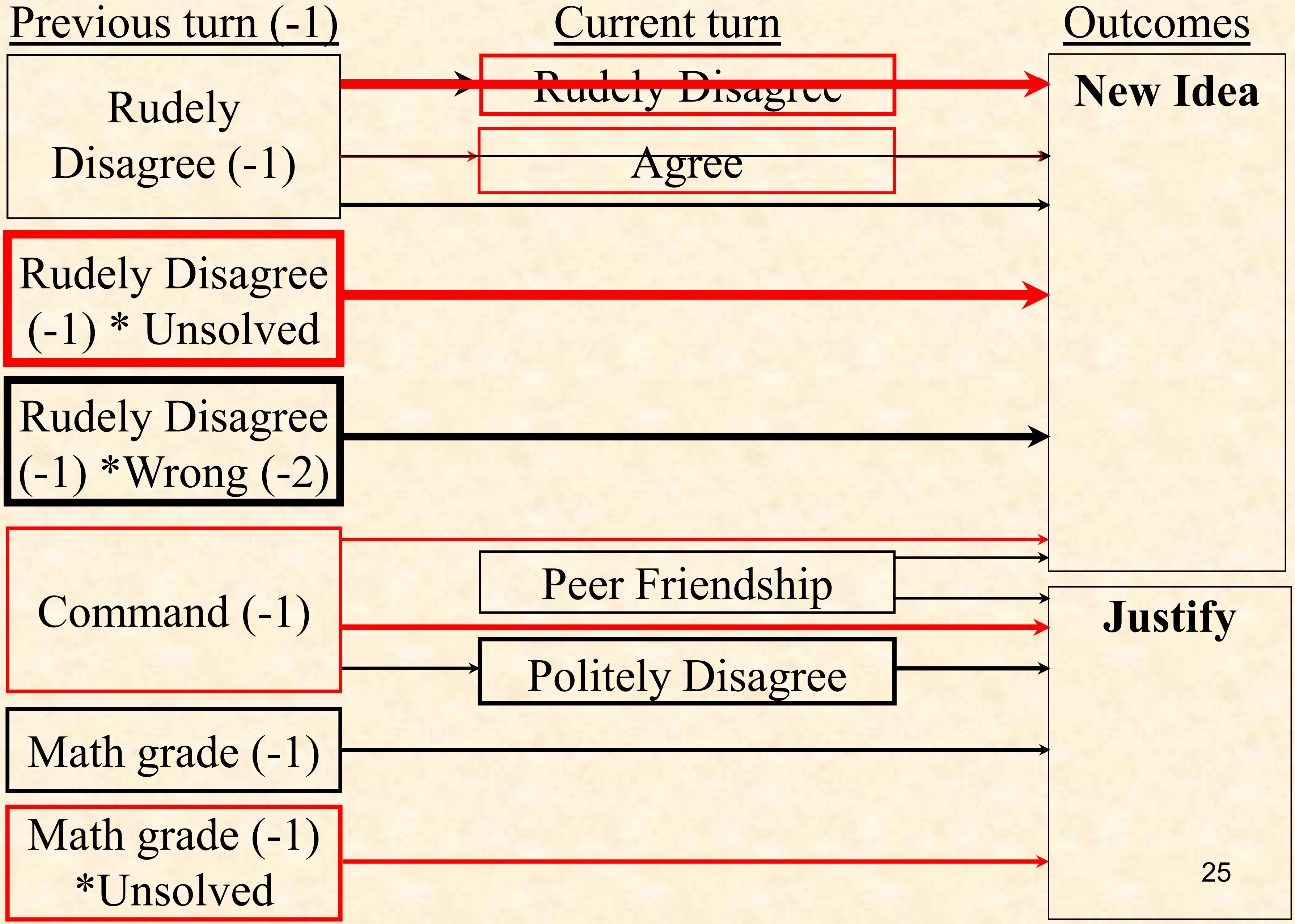
- Yes, code as command (*imperative*)
- No, is the subject the addressee?
  - No, are any of the clauses in the form of a question?
    - No, code as statement (*declarative*)
    - Yes, code as question (*interrogative*)
  - Yes, is the verb a modal?
    - No, should the described action have been performed, but not done?
      - Yes, code as a command
      - No, code as a question
    - Yes, Is it a Wh- question (who, what, where, why, when, how)?
      - Yes, code as an question
      - No, is the action feasible?
        - Yes, code as a command
        - No, code as an question

Based on Labov (2001), Tsui (1992)

Statistical Discourse Analysis

Analytical Difficulty	**Strategy**
• Differences across topics	• Multilevel analysis
• Time periods differ ($T_2 \neq T_4$)	• Breakpoint analysis & Multilevel analysis
• Serial correlation ($t_8 \rightarrow t_9$)	• I^2 index of Q-statistics; Model with lag variables
• Parallel talk (→→⇉⇉)	• Store path: ID prior turn, Vector Auto-Regression
• Discrete outcomes (Yes / No)	• Logit / Probit
• Multiple outcomes (Y_1, Y_2)	• Multivariate outcome models
• Infrequent outcomes (00010)	• Logit bias estimator
• People & Groups differ 🚹 ≠ 🚺	• Multilevel analysis
• Mediation effects (X→M→Y)	• Multilevel mediation tests
• False positives ($+ + \cancel{+} +$)	• 2-stage linear step-up procedure
• Missing data (101?001?10)	• Markov Chain Monte Carlo multiple imputation
• Robustness	• Separate outcome models; Data subsets & unimputed data

Explanatory model: New Idea & Justify



Previous turn (-1)

Current turn

Outcomes

Rudely Disagree (-1)

Rudely Disagree

Agree

New Idea

Rudely Disagree (-1) * Unsolved

Rudely Disagree (-1) *Wrong (-2)

Command (-1)

Peer Friendship

Politely Disagree

Justify

Math grade (-1)

Math grade (-1) *Unsolved

Mathematics

Bayesian Information Criterion

$$-\frac{2\,L}{n} + \left(\frac{k\,\ln(n)}{n}\right)$$

Regression specification

$$\pi_{ijk} = \mathrm{F}(\beta_0 + \mathrm{f}_{0jk} + \mathrm{g}_{00k} + \beta_{00s}\mathbf{S}_{00k} + \beta_{00t}\mathbf{T}_{00k} + \beta_{ujk}\mathbf{U}_{ijk}$$

$$+ \beta_{vjk}\mathbf{V}_{(i-1)jk} + \phi_{vjk}\mathbf{V}_{(i-2)jk} + \gamma_{vjk}\mathbf{V}_{(i-3)jk} + \eta_{vjk}\mathbf{V}_{(i-4)jk})$$