Science versus Engineering:
Considerations for Computational Communication Research

Frederic R. Hopp

UC Santa Barbara - Media Neuroscience Lab
@medianeuro • medianeuroscience.org

> “This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.”

Anderson, Chief Editor *Wired* (2008)

“This is a world where **massive amounts of data** and applied mathematics replace every other tool that might be brought to bear. **Out with every theory** of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. **Who knows why** people do what they do? The point is they do it, and we can **track and measure** it with **unprecedented fidelity**. With enough data, the **numbers speak for themselves**.”

Anderson, Chief Editor *Wired* (2008)

Big Data > Slow, additive knowledge generation of the scientific method

Data-driven prediction > Theory-driven explanation

Optimized Engineering > Sound, scientific reasoning and understanding

Guiding Considerations for CCR

1) “Understanding the human condition” versus “building better mousetraps”?

2) There and back again: From prediction to explanation?

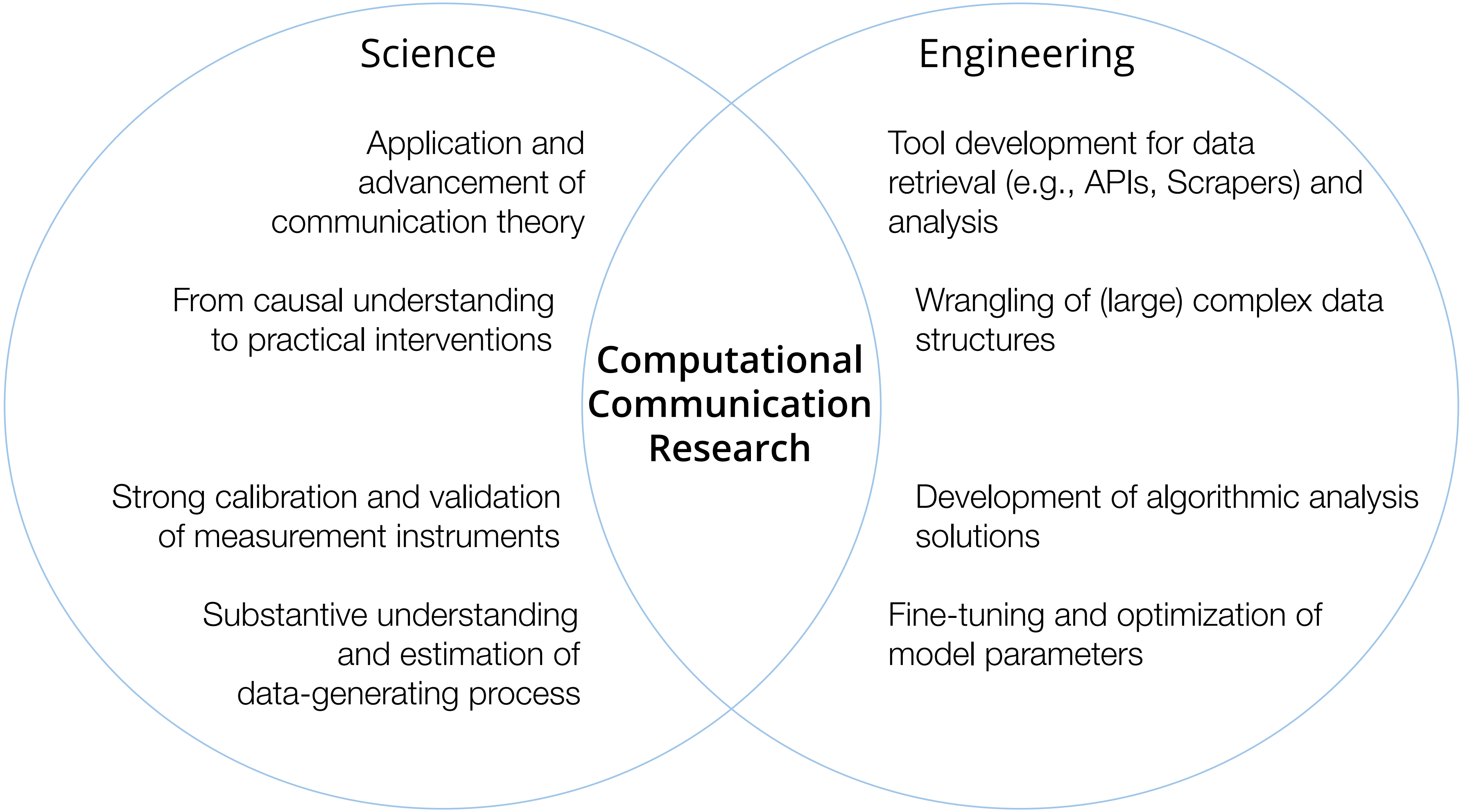
3) Meaningful insights versus Big Data apophenia?

4) Solution-oriented, use-inspired research versus theory agglomeration?

Science versus Engineering (Lin, 2015)

	Science	Engineering
Goal	**Understanding** human behavior and offering **explanations** of social phenomena	Building more effective **computational artifacts** as measured by some **well-defined metric**
Approach	Scientific method Theory-driven	Machine-Learning Data-driven

Computational Communication Research (van Atteveldt & Peng, 2018)



Science

Application and advancement of communication theory

From causal understanding to practical interventions

Strong calibration and validation of measurement instruments

Substantive understanding and estimation of data-generating process

Engineering

Tool development for data retrieval (e.g., APIs, Scrapers) and analysis

Wrangling of (large) complex data structures

Development of algorithmic analysis solutions

Fine-tuning and optimization of model parameters

Computational Communication Research

Common Critiques of CCR

Transparent, parsimonious models versus complex, intricate algorithms

Advancing communication theory versus stacking up the "tool-pile"

Answering substantive research questions versus showcasing machine-learning capabilities

In short: Computational Methods is all Methods and no Science!

⇨ Debate and confusion over two different goals:
Explanation versus **Prediction**

Explanation versus Prediction (Yarkoni & Westfall, 2017)

Explain: Provide an accurate description of a process' causal underpinnings

Data are assumed to arise from a particular data-generating process

Goal: Estimate true parameters of this process

Claims:

1) Improving metrics is neither necessary nor sufficient to make a contribution to knowledge!

2) Engineering creates complex models that can accurately predict outcomes of interest but fail to respect known psychological or neurological constraints!

Predict: Accurately forecast behaviors that have not yet been observed

Data are assumed to be the result of some unknown (possibly unknowable) process

Goal: Find algorithm that results in the same outputs as this process given the same inputs

Claims:

1) Sound scientific reasoning/understanding not necessary to improve engineering/prediction!

2) Focus on explanation yields simple models that appear theoretically elegant but have very limited capacity to predict actual human behavior!

Resolving the Debate - Promises for CCR

Short-term focus on *prediction* can ultimately improve our ability to *explain* the causes of behavior in the long-term.

Examples

- Ships ⇒ Hydrodynamics
- Steam Engines ⇒ Thermodynamics
- Airplanes ⇒ Aerodynamics

- Agent-Based Simulations ⇒ Communication Dynamics?
- Natural Language Understanding ⇒ Narrative Comprehension?
- Finite-state Machines ⇒ News-Event Dynamics?
- Deep neural networks ⇒ ???

“What I cannot create, I do not understand” (Feynman)

“If you cannot measure it, you cannot improve it” (Kelvin)

Resolving the Debate - Promises for CCR

Emphasis on prediction not an opponent of explanation but rather as a complementary goal that can ultimately increase theoretical understanding.

1) Computational modeling ⇒ Deeper understanding of one's data structure and *parameter space* (model complexity)

2) Limiting QRPs:
   a) Minimized *p*-hacking
   b) Increased research efficiency
   c) Evaluation of model performance
   d) (Increased interpretability)

3) Promises of Big Data:
   a) Replicable, reliable science to favor *small effects* (and null-findings!) from *large samples* over large effects from small samples
   b) Inexpensive, fast tests of *risky predictions* over costly, time-consuming, and self-evident hypotheses

Solution-oriented, method-theory synergy (Watts, 2017)

Near-total (too early) focus on *explanation* in Communication Research has produced a plethora of intricate theories with little (or unknown) ability to predict future behaviors with appreciable accuracy.

Incoherency problem

- Historical emphasis on the advancement of theories over the solution of *practical problems*
- Many theories for the same thing and fundamentally *incoherent* when viewed collectively

Use-inspired research

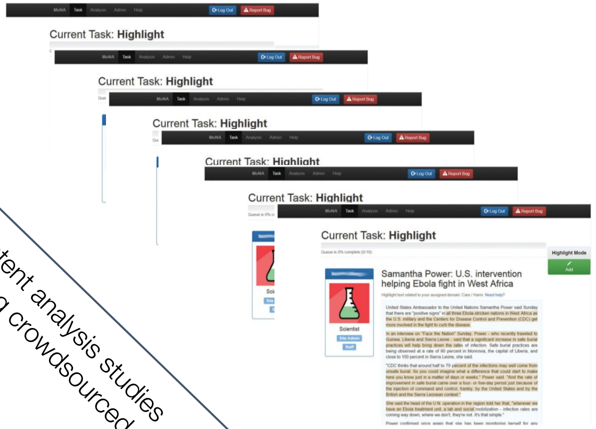
- Replicability over novelty, surprise, or importance
- Advance theory in the service of solving real-world problems

Goldilocks problems

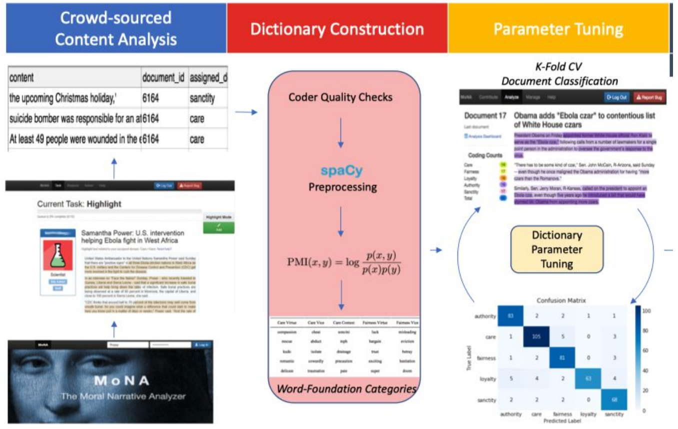
- Research problem that is not too large and complex but sufficiently difficult to justify a genuinely scientific approach
- Modularity ⇒ Address problem in a succession of increasingly ambitious versions

Goldilocks Example

From extracting latent moral information to event forecasting…



6 content analysis studies (including crowdsourced annotations)



Crowd-sourced Content Analysis
Dictionary Construction
Parameter Tuning
Current Task: Highlight
Samantha Power: U.S. intervention helping Ebola fight in West Africa
MoNA
The Moral Narrative Analyzer
Coder Quality Checks
spaCy
Preprocessing
Word-Foundation Categories
K-Fold CV
Document Classification
Document 17
Obama adds "Ebola czar" to contentious list of White House czars
Coding Counts
Dictionary Parameter Tuning
Confusion Matrix
True Label
Predicted Label
authority
care
fairness
loyalty
sanctity

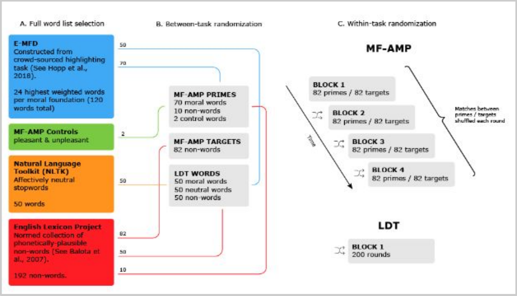
content	document_id	assigned_d
the upcoming Christmas holiday,'	6164	sanctity
suicide bomber was responsible for an at	6164	care
At least 49 people were wounded in the (	6164	care

$$\mathrm{PMI}(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

	authority	care	fairness	loyalty	sanctity
authority	83	2	2	1	1
care	1	105	5	0	3
fairness	1	2	81	0	3
loyalty	5	4	2	63	4
sanctity	2	2	2	0	68

iCoRe - The GDELT Interface for Communication Research
iCoRe
the GDELT interface for communication research

Integration of E-MFD into GDELT for real-time tracking of moral conflict (Hopp et al., 2019)

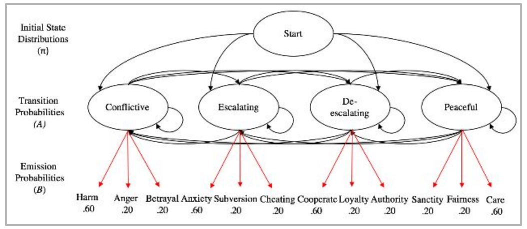


A. Full word list selection
B. Between-task randomization
C. Within-task randomization
E-MFD
Constructed from crowd-sourced highlighting task (See Hopp et al., 2018).
24 highest weighted words per moral foundation (120 words total)
MF-AMP Controls
pleasant & unpleasant
Natural Language Toolkit (NLTK)
Affectively neutral stopwords
50 words
English Lexicon Project
Normed collection of phonetically-plausible non-words (See Balota et al., 2007).
192 non-words.
50
70
2
50
82
50
10
MF-AMP PRIMES
70 moral words
10 non-words
2 control words
MF-AMP TARGETS
82 non-words
LDT WORDS
50 moral words
50 neutral words
50 non-words
MF-AMP
BLOCK 1
82 primes / 82 targets
BLOCK 2
82 primes / 82 targets
BLOCK 3
82 primes / 82 targets
BLOCK 4
82 primes / 82 targets
Time
Matches between primes / targets shuffled each round
LDT
BLOCK 1
200 rounds

Application of E-MFD in behavioral paradigms (Fisher & Hopp, in progress)

Reliable, valid, and manual annotation of morally-relevant content (Weber et al., 2018)

Development of extended Moral Foundations Dictionary for automated analysis of textual corpora (Hopp et al., 2018)



Real-world event prediction based on morally-relevant news frames (Hopp et al.,2019)

Conclusion and Outlook

Case-by-case: Explanation versus Prediction?

- Seeking to identify abstract, generalizable principles ⇒ explanation-focused strategy
- Mimicking the outputs of the true data-generating process when given the same inputs, without care *how* that goal is achieved ⇒ Prediction-focused strategy

Complement accurate predictions with attempt to *understand* the phenomena involved
⇒ Better, more *generalizable solutions*

⇒ Prediction versus explanation not either-or choices, but complementary for opening up new avenues of research and theory

Strong theory

- Clear a priori predictions and sensible explanations of what are otherwise uninterpretable statistical tests

Thank You!

UC Santa Barbara - Media Neuroscience Lab
@medianeuro • medianeuroscience.org