Towards A Conceptual Model of Computational Sustainability

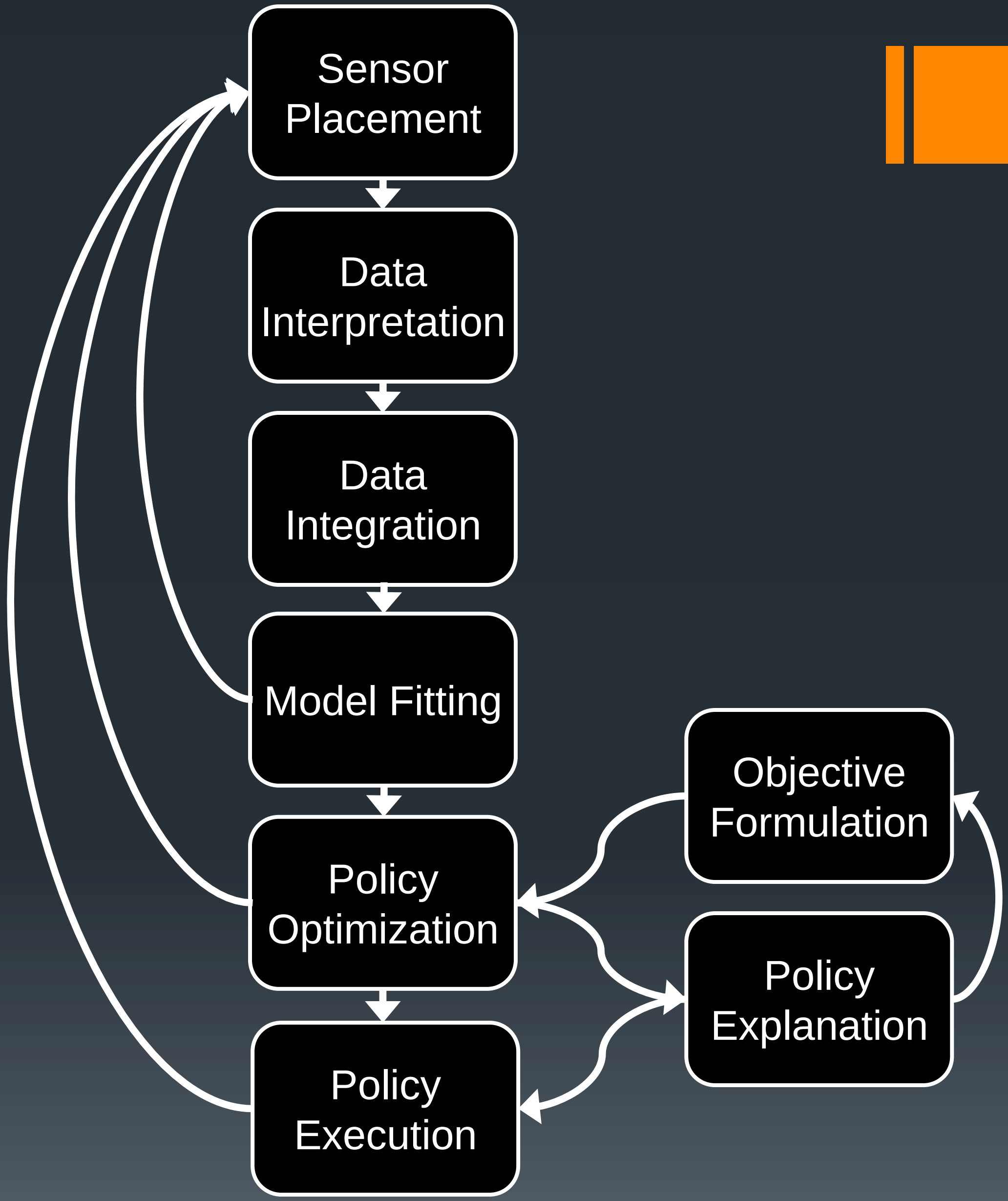
Tom Dietterich

Oregon State University

Computational Sustainability

- The study of computational methods that can contribute to the sustainable management of the earth’s ecosystems
  - biological
  - social
  - economic

- Data → Models → Policies



Sensor Placement

Data Interpretation

Data Integration

Model Fitting

Policy Optimization

Policy Execution

Objective Formulation

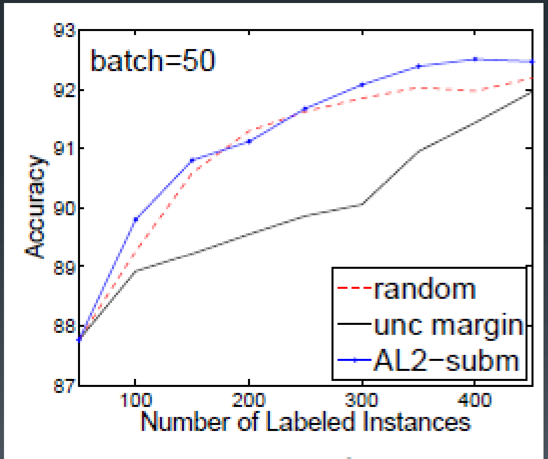
Policy Explanation

Example Research Efforts

Sensor Placement

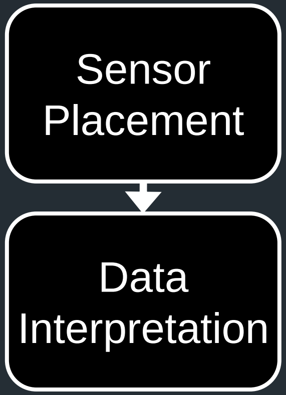
- Objectives
  - detection probability
  - improving model accuracy
  - improving causal understanding
  - improving policy effectiveness

- Active Learning for eBird (Damoulas & Dilkina)
- Others?



(f) eBird

Data Interpretation



Sensor Placement

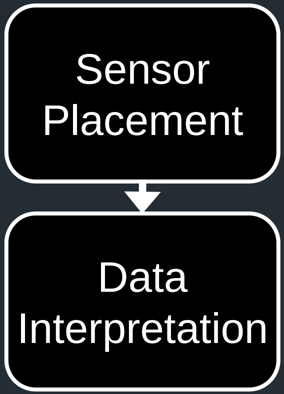
Data Interpretation

- Insect identification for population counting (Dietterich, Todorovic, Lin, et al.)
  - Freshwater macro-invertebrates
  - Rice pests
  - Raw data: images
  - Interpreted data: Count by species
- Understanding tree swallow roosts from Doppler radar (Sheldon, et al)
  - Raw data: Doppler radar images
  - Interpreted data: Location and approx. size of swallow roosts over whole US
- Estimating Bird Migration from Doppler Radar (BirdCast project)
- Sensor Network Data Cleaning (Dereszynski & Dietterich)

image: Qing Yao

Rice Pest Project

- Working with Dr. Qing Yao from Zhejiang Sci-Tech University
- Challenge: Classifying overlapping specimens


Sensor Placement

Data Interpretation

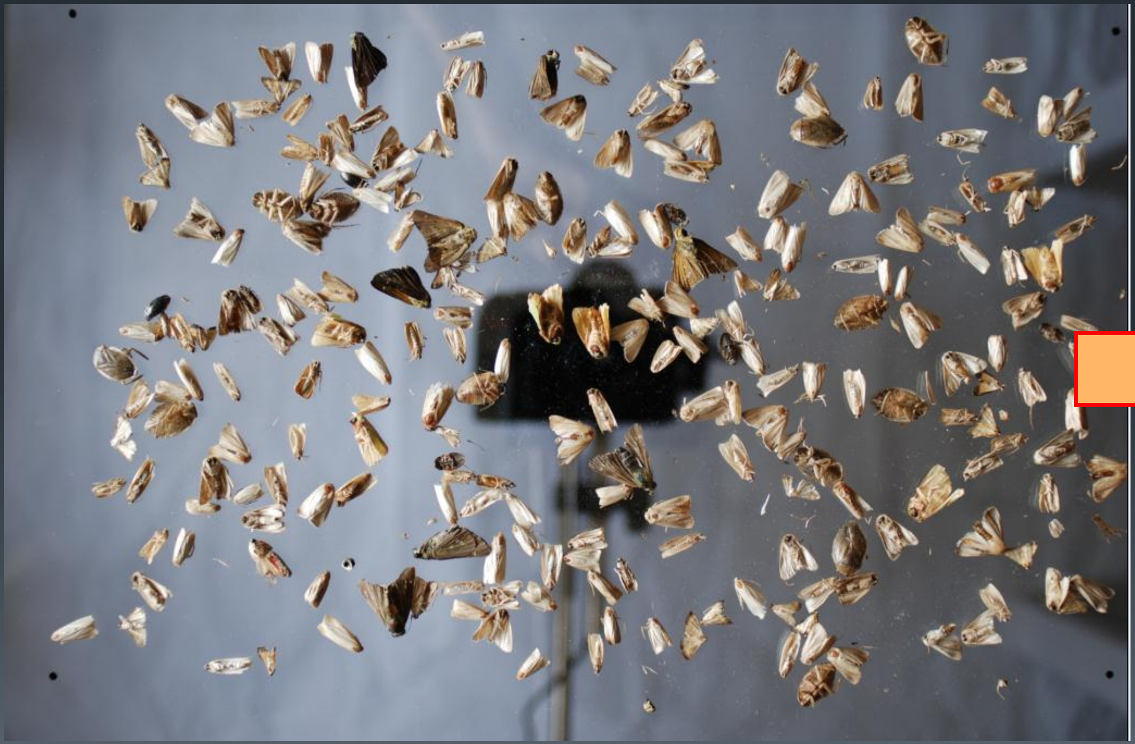
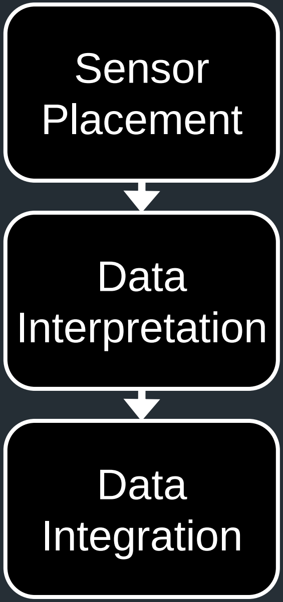


image: Qing Yao

Species	Count
Nilaparvata lugens(	12
Sogatella furcifera	8
Laodelphax striatellus	0
Cnaphalocrocis medinalis	0
Chilo suppressalis	45
Sesamia inferens	18

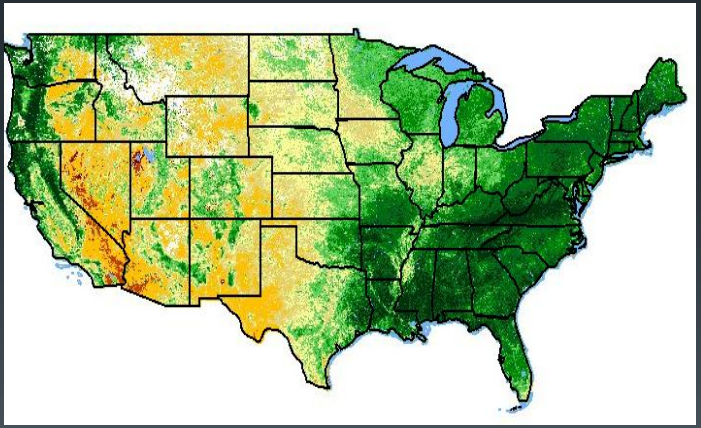
Data Integration

- eBird Reference Data Set + BirdCast
  - Landsat (30m; monthly)
    - land cover type
  - MODIS (500m; daily/weekly)
    - land cover type
    - “greening” index
  - Census (every 10 years)
    - human population density
    - housing density and occupation
  - Interpolated weather data
    - rain, snow, solar radiation, wind speed & direction, humidity
  - Integrated weather data (daily)
    - warming degree days
  - Digital elevation model (rarely changes)
    - elevation, slope, aspect


Sensor Placement

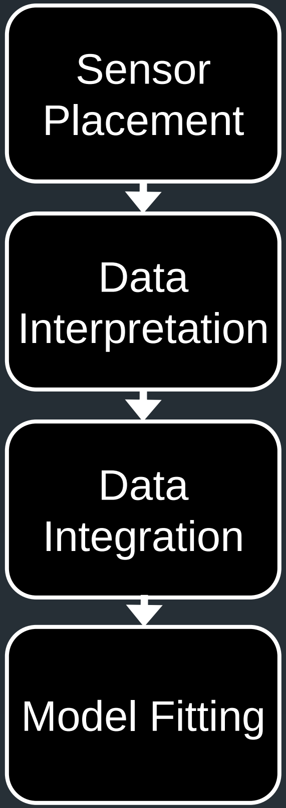
Data Interpretation

Data Integration

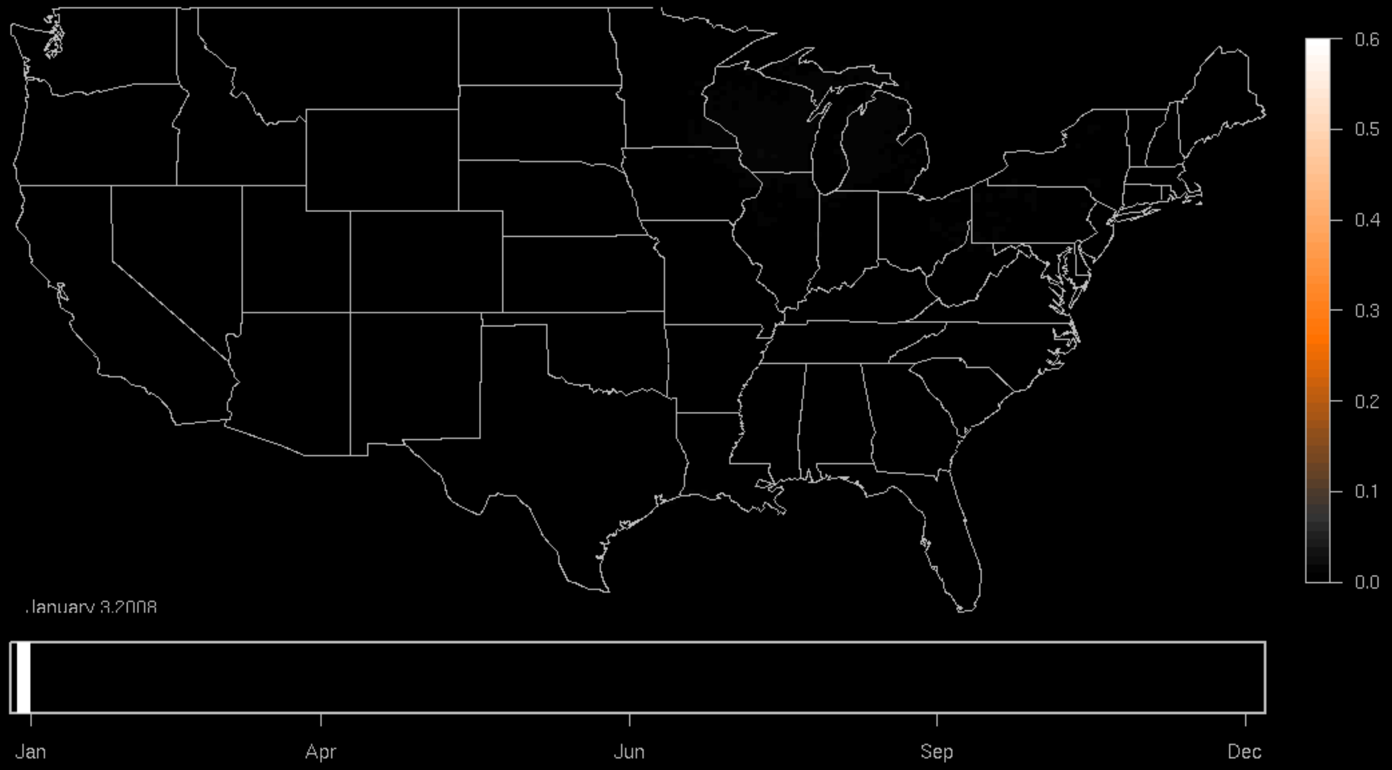

Landsat NDVI: http://ivm.cr.usgs.gov/viewer/

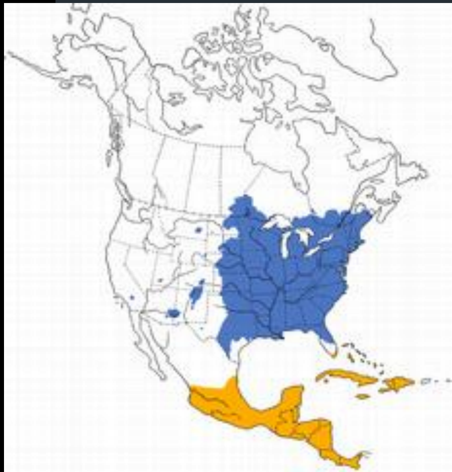
Model Building and Model Integration

- STEM: bird species distribution models (Fink, et al.)
- OD-BRT: Occupancy Models parameterized via boosted regression trees (Hutchinson, et al.)
- ODE: Occupancy, Detection & Expertise (Wong, et al.)
- Discovering plant communities from field observational data (Lettkeman & Dietterich)
- Multiple-Species SDMs (Wong, Dietterich, et al.)
- Moth Emergence Model (Sheldon, Dietterich, et al.)
- Aral Sea Fisheries (Conrad, et al.)
- Bird Migration Model: Collective Graphical Model (Sheldon)
- Oregon Centennial Fire Model (Montgomery, et al.)



Example Fitted Model: STEM Model of Bird Species Distribution



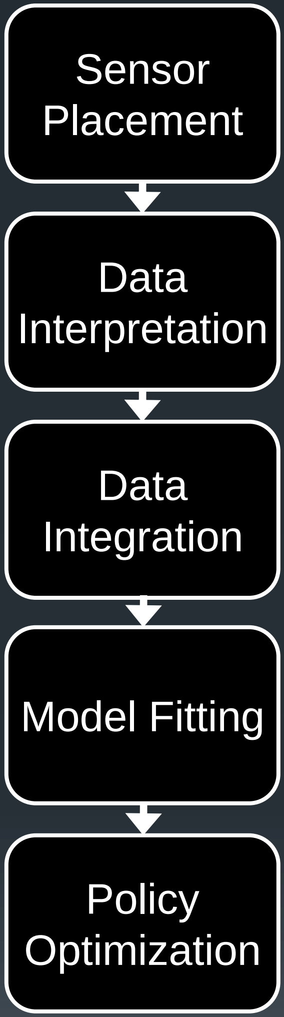


Indigo Bunting

slide courtesy of Daniel Fink

Policy Optimization

- Halibut Fisheries (Ermon, Conrad et al)
- Wildfire Management :
  - LetBurn vs. Suppress (Montgomery, Houtman, et al.)
  - Spencer & Shmoys
- Invasive Species Management
  - Tamarisk: (Albers, Hall, Taleghan, Dietterich)
  - Spencer & Shmoys
- Red Cockaded Woodpecker (Sheldon, Finseth, et al.)
- Johne's Diease (Toese, et al.)
- +++



Sensor Placement

Data Interpretation

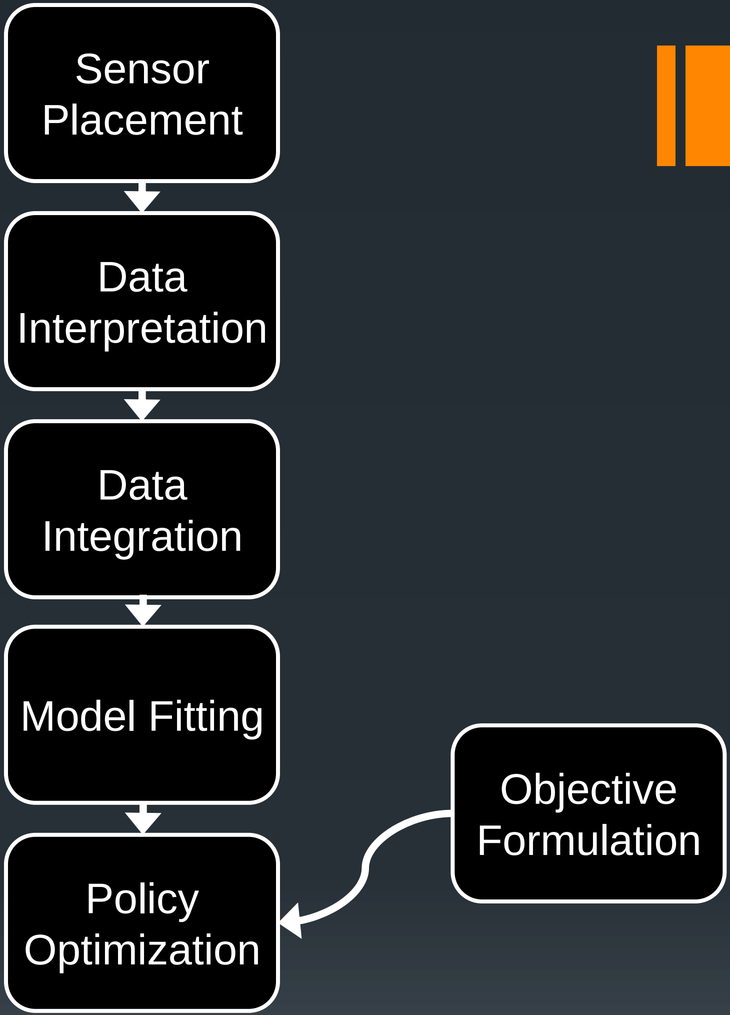
Data Integration

Model Fitting

Policy Optimization

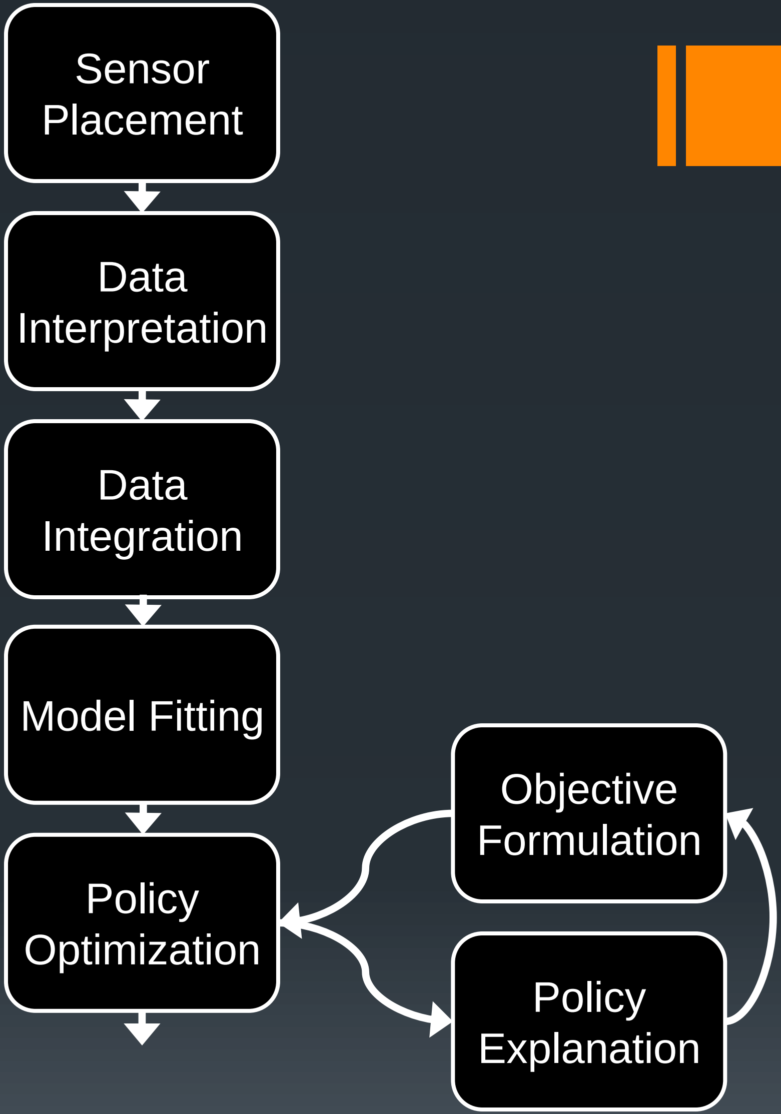
Objective Formulation

- Any?



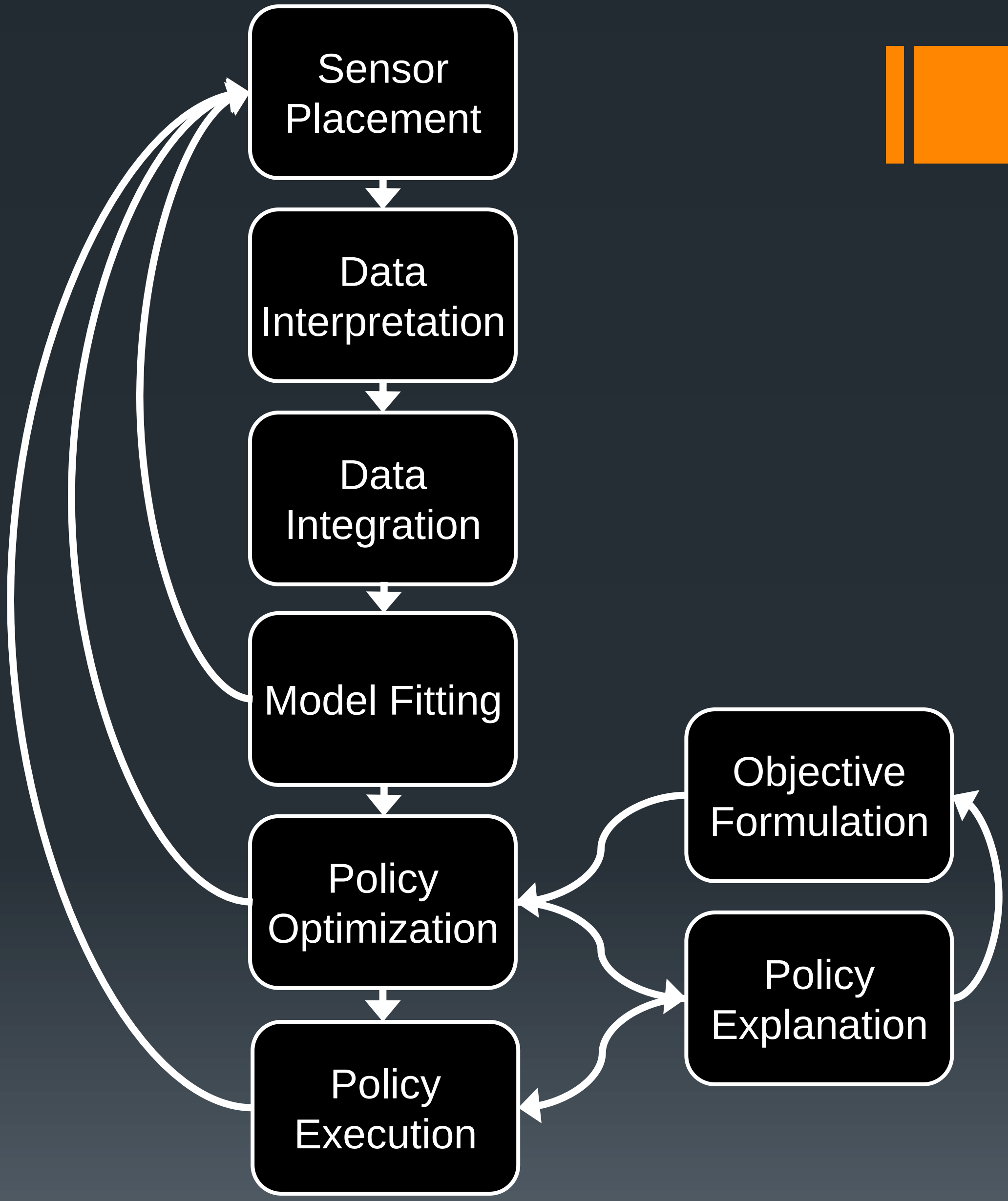
Policy Explanation

- Wildfire Management (McGregor)
- Invasive Species (Taleghan)
- Others?



Policy Execution

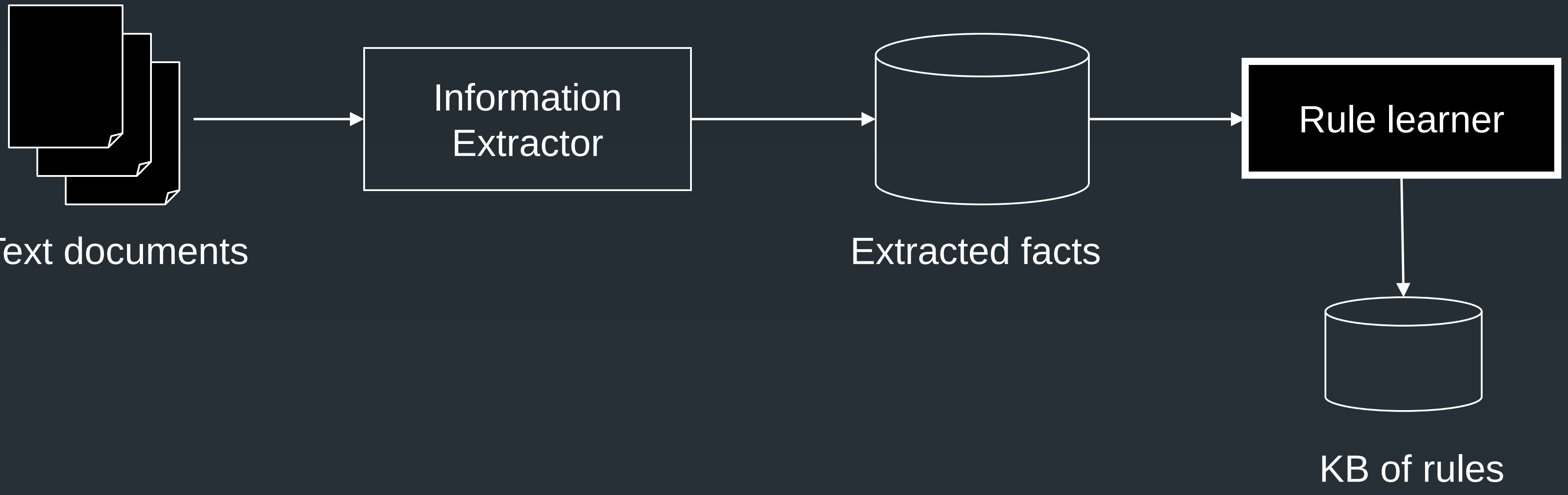
- RCW?
- Fisheries?
- Invasives?



Learning Rules from Incomplete Examples via a Probabilistic Mention Model

Mohammad Shahed Sorower, Janardhan Rao Doppa, Thomas G. Dietterich

Motivation and Goal



- **Goal:**
  - Induce general rules by reading about concrete facts
    - *Ex: gameWinner(G,T1) :- teaminGame(G,T1), teaminGame(G,T2), gameLoser(G,T2)*

Challenges in Learning Rules from Natural Text

- Extracted ground facts are highly incomplete
  - Only a very small part of the “whole truth” is mentioned in a document
  - Even less is successfully extracted by NLP methods

- Incompleteness is not “missing at random”
  - Speaker seeks to achieve communication goals concisely
  - Mention “newsworthy” or “surprising” facts
  - Let the reader fill in the rest by applying background knowledge

Example

"Given the commanding lead of Kansas city on the road, the Denver Broncos' 14-10 victory surprised many"

- TeamInGame(g1,KansasCity)
- TeamInGame(g1, DenverBroncos)
- GameWinner(g1, DenverBroncos)
- GameTeamScore(g1, DenverBroncos,14)
- GameTeamScore(g1, KansasCity, 10)
- AwayTeam(g1, KansasCity)

Does not mention

- GameLoser(g1,KansasCity)
- HomeTeam(g1,DenverBroncos)

Hard to learn rules such as

- Winner ⬄ not Loser
- HomeTeam ⬄ not AwayTeam
- Winner is team that scores the most points

Example 2:

“Ahmed Said Khadr, an Egyptian-born Canadian, was killed last October in Pakistan”

- BornIn(Khadr, Egypt)
- CitizenOf(Khadr, Canada)
- KillingEvent(e1)
- Location(e1, Pakistan)
- Victim(e1, Khadr)

How can we learn the rule

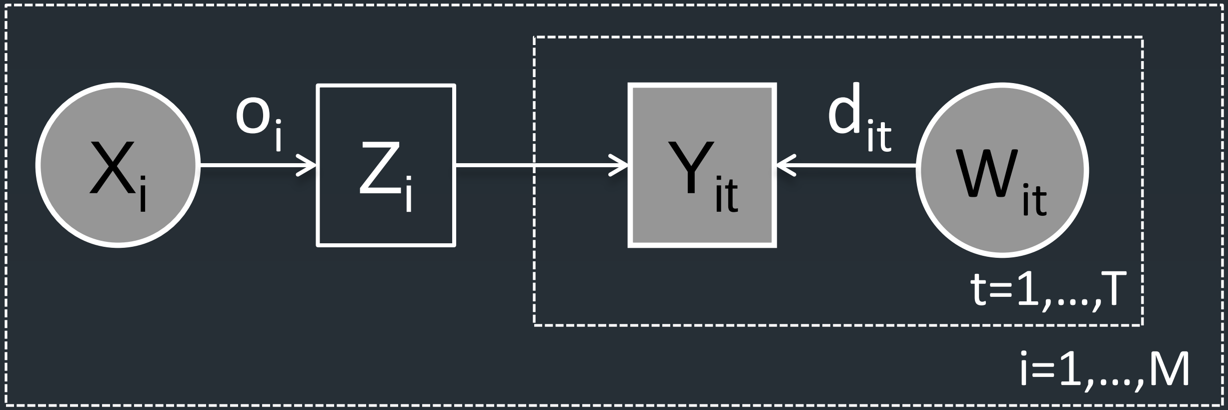
CitizenOf(P,C) :- BornIn(P,C) ???

Most articles only mention both CitizenShip and BirthPlace when they are *not* equal

- Pilot Study corpus: 23 BirthPlace mentions of which 14 violate the rule

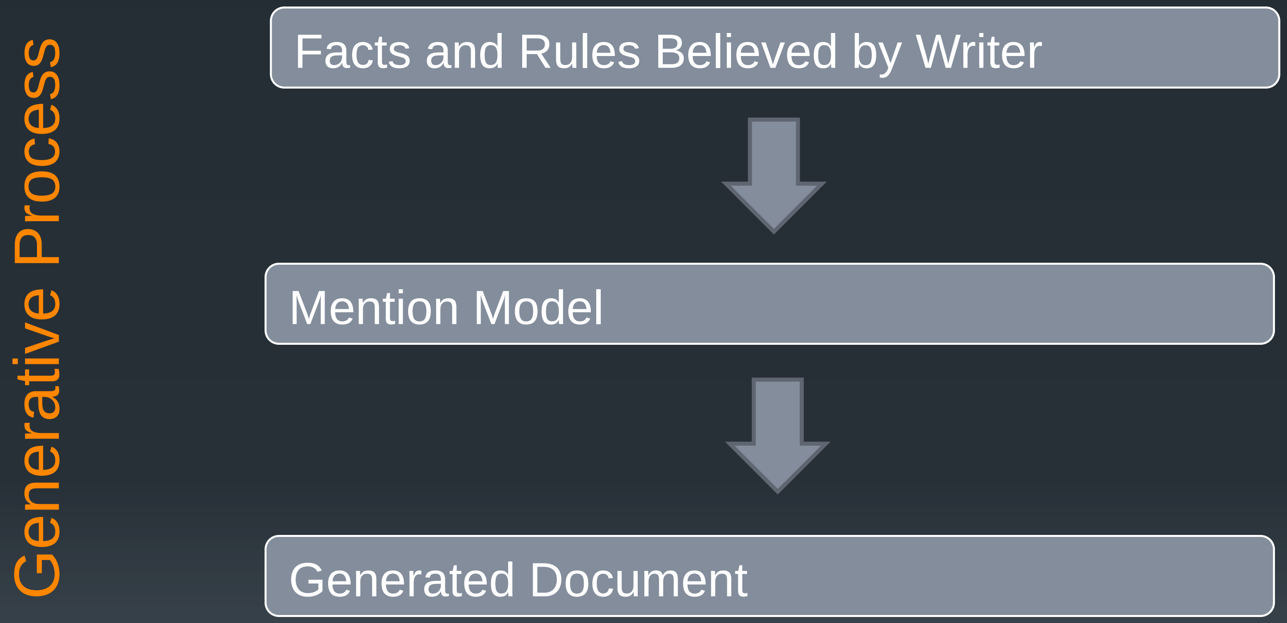
Occupancy-Detection Model

MacKenzie, et al, 2006



Key Idea: Explicit model of the observation process

Idea: Learn an explicit model of the observation process = "Mention Model"



Generative Process

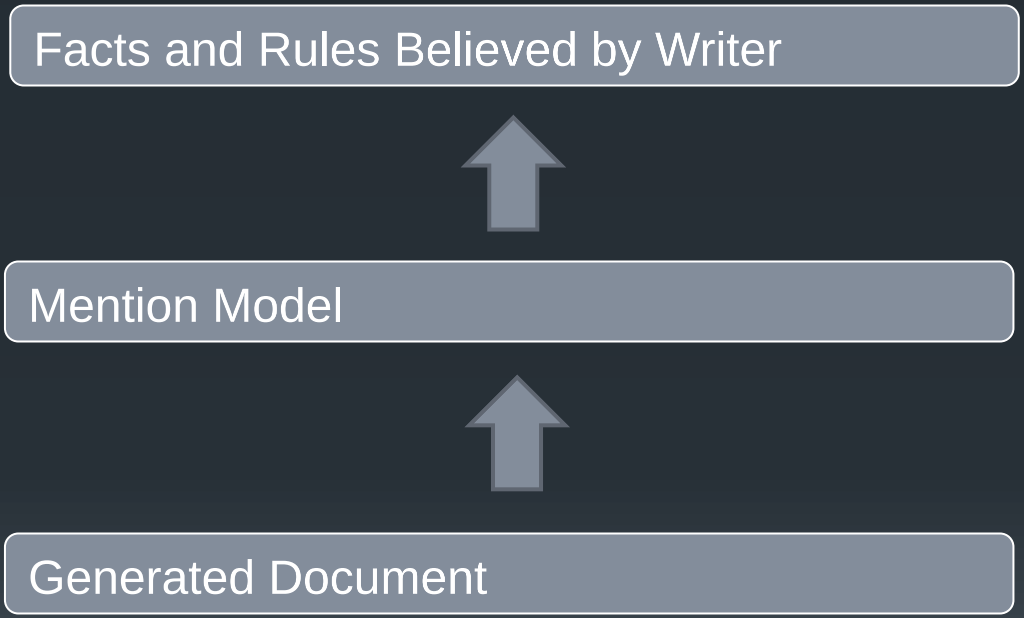
Facts and Rules Believed by Writer

Mention Model

Generated Document

Learn Rules by Probabilistic Inversion of the Mention Model

Learning Process



Facts and Rules Believed by Writer

Mention Model

Generated Document

Mention Model: Grice’s Maxims of Cooperative Conversation

- Be Truthful
  - Do not say things you believe are false
  - Do not omit things that would lead the hearer to believe falsehoods [Added]
- Quantity of Information
  - Say as much as is necessary
  - Do not say more than is necessary
- Be Relevant
- Be Clear

Formalization

- Reader believes K, is told F, and will infer G:

  $K, Mention(F) \vdash_{reader} G$

- Mention true facts:

  $F \Rightarrow Mention(F)$ [with some probability]

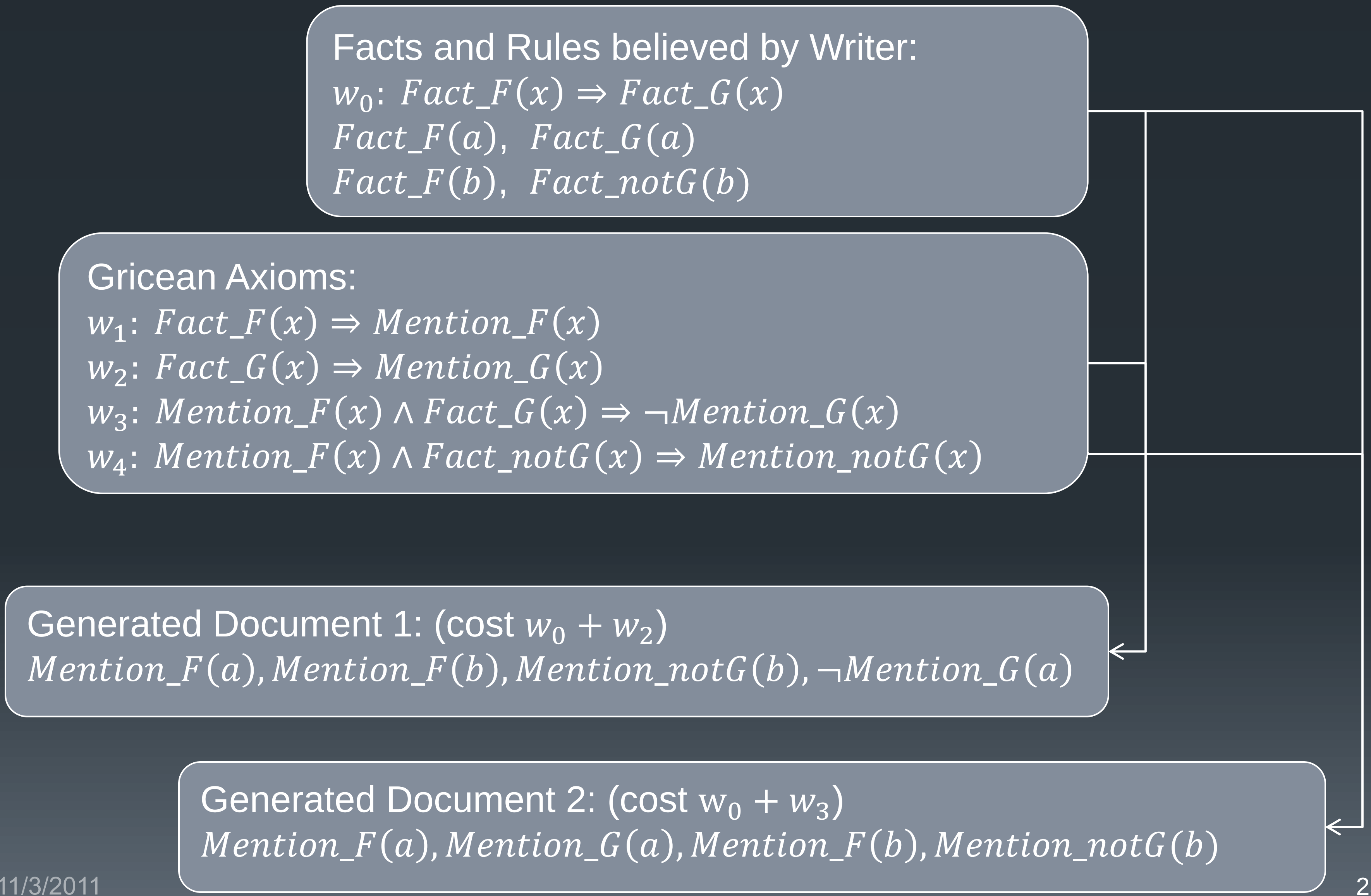
- Don't mention facts that can be inferred:

  $Mention(F) \wedge G \wedge (K, Mention(F) \vdash_{reader} G) \Rightarrow \neg Mention(G)$

- Mention facts needed to prevent incorrect inferences

  $Mention(F) \wedge \neg G \wedge H \wedge (K, Mention(F) \vdash_{reader} G) \wedge (K, Mention(F)$
  $\wedge Mention(H) \vdash_{reader} \neg G) \Rightarrow Mention(H)$

Implementation in Markov Logic



Facts and Rules believed by Writer:

w_0: $Fact\_F(x) \Rightarrow Fact\_G(x)$

$Fact\_F(a),\ \ Fact\_G(a)$

$Fact\_F(b),\ \ Fact\_notG(b)$

Gricean Axioms:

w_1: $Fact\_F(x) \Rightarrow Mention\_F(x)$

w_2: $Fact\_G(x) \Rightarrow Mention\_G(x)$

w_3: $Mention\_F(x) \wedge Fact\_G(x) \Rightarrow \neg Mention\_G(x)$

w_4: $Mention\_F(x) \wedge Fact\_notG(x) \Rightarrow Mention\_notG(x)$

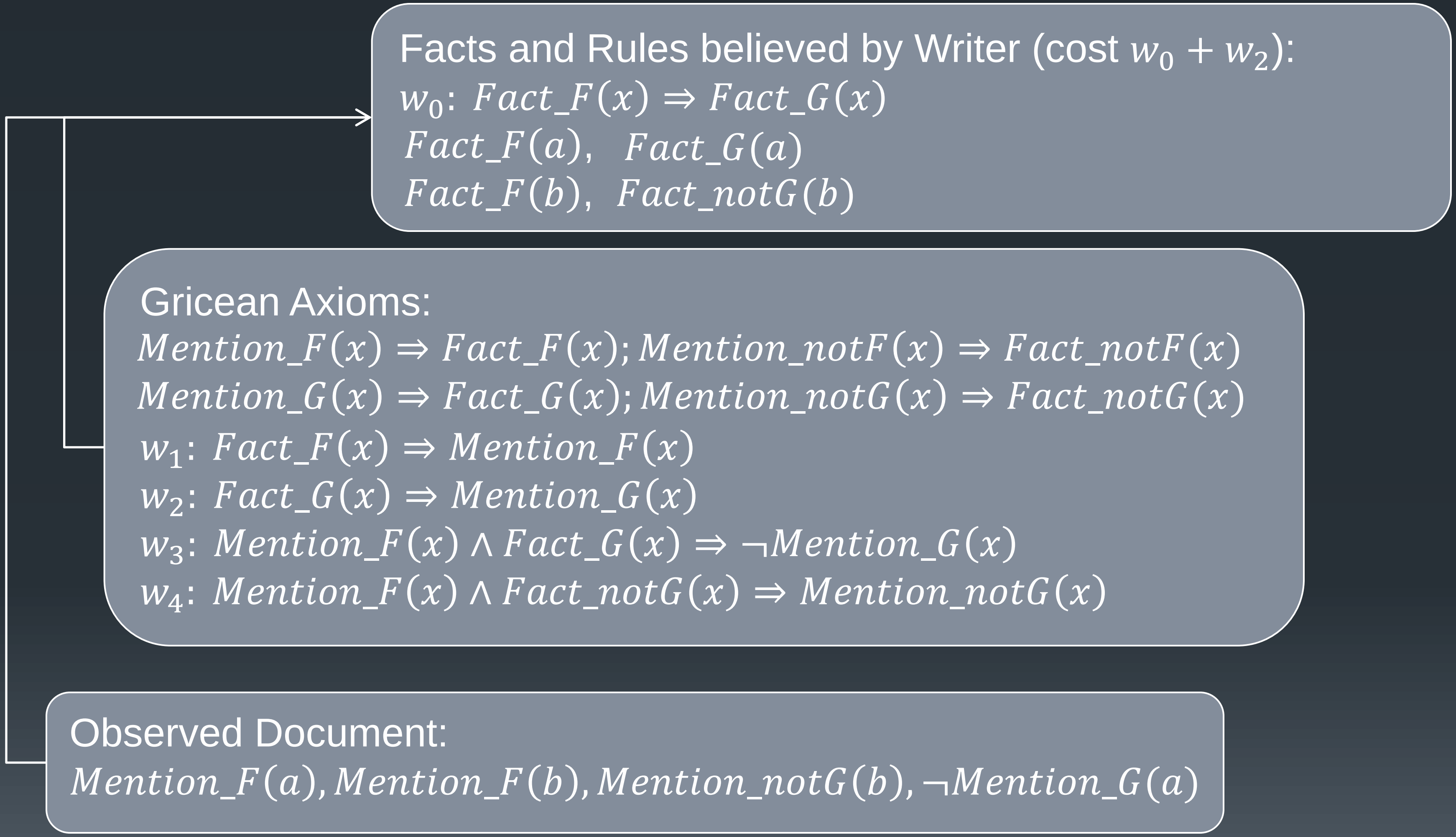
Generated Document 1: (cost $w_0 + w_2$)

$Mention\_F(a), Mention\_F(b), Mention\_notG(b), \neg Mention\_G(a)$

Generated Document 2: (cost $\mathrm{w}_0 + w_3$)

$Mention\_F(a), Mention\_G(a), Mention\_F(b), Mention\_notG(b)$

Inference During Reading



Facts and Rules believed by Writer (cost $w_0 + w_2$):

w_0: $Fact\_F(x) \Rightarrow Fact\_G(x)$

$Fact\_F(a)$, $Fact\_G(a)$

$Fact\_F(b)$, $Fact\_notG(b)$

Gricean Axioms:

$Mention\_F(x) \Rightarrow Fact\_F(x); Mention\_notF(x) \Rightarrow Fact\_notF(x)$

$Mention\_G(x) \Rightarrow Fact\_G(x); Mention\_notG(x) \Rightarrow Fact\_notG(x)$

w_1: $Fact\_F(x) \Rightarrow Mention\_F(x)$

w_2: $Fact\_G(x) \Rightarrow Mention\_G(x)$

w_3: $Mention\_F(x) \wedge Fact\_G(x) \Rightarrow \neg Mention\_G(x)$

w_4: $Mention\_F(x) \wedge Fact\_notG(x) \Rightarrow Mention\_notG(x)$

Observed Document:

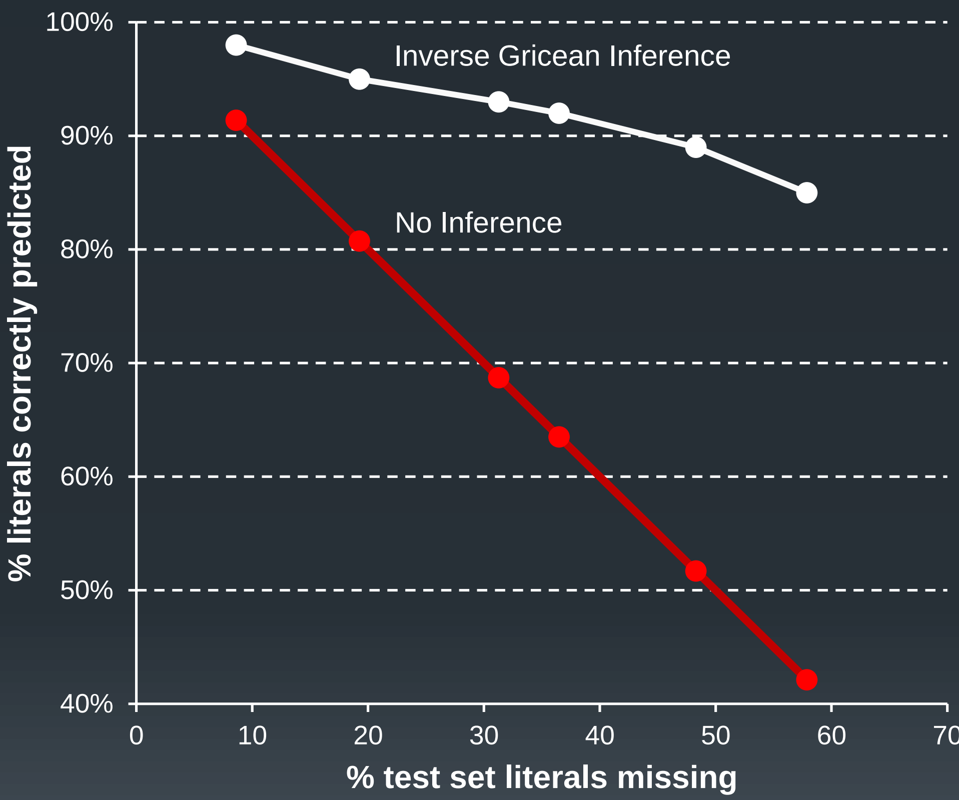
$Mention\_F(a), Mention\_F(b), Mention\_notG(b), \neg Mention\_G(a)$

Rule Learning

- Inputs:
  - Rule templates
  - Extracted mentions from documents
- Outputs:
  - Weighted rules expressed as Markov Logic knowledge base
- Algorithm:
  - Generate all possible rules from the templates
  - Compute # of supporting instances for each rule (on extracted mentions) and keep the top 10 best scoring rules for each head predicate
  - Generate the Gricean rules from these candidate rules
  - Apply the EM algorithm to learn the weights on the fact rules and the Gricean rules

Experiment 1: Synthetic Data

- Synthetic data set
  - NFL games generated from true rules and ground truth
  - Two sets of correlated predicates:
    - GameWinner, GameLoser, GameTeamScore
    - GameHomeTeam, GameAwayTeam
  - Choose one literal from each set and mention it
  - Mention each of the other literals with probability $1 - q$
- Experiment
  - Train on data with 58% of literals missing ($q = 0.97$)
  - Test on data with varying amounts of literals missing



Experiment 2: Real Training; Synthetic Test

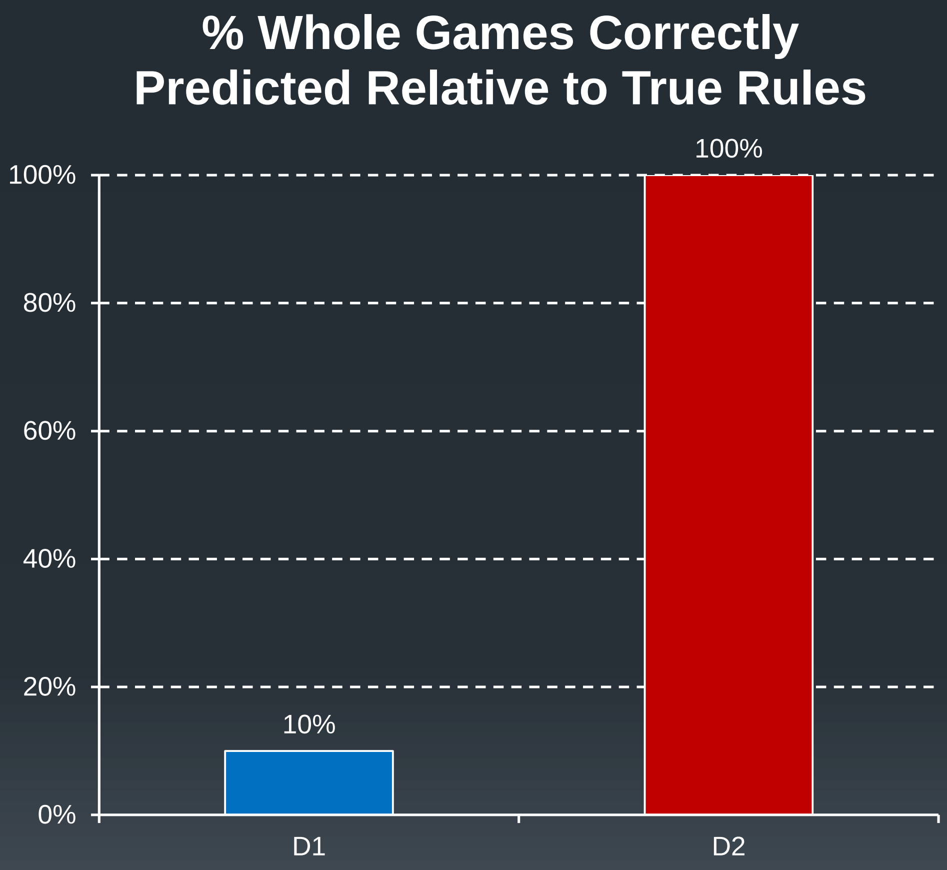
- Data from BBN Extractions 12/16/10
- D1: NFL BBN_training
- D2: NFL BBN_robustness
- Both data sets “repaired” using learned integrity constraints
  - Delete literals in all possible ways to satisfy the integrity constraints
  - Remove duplicates.
  - Data set sizes:
    - D1: 203 records
    - D2: 56 records
- Test set: 100 examples manually created from ground-truth NFL database to cover all missingness scenarios

Home/Away			
Dataset	Both Missing	One Missing	Both Mentioned
D1	85.7%	11.3%	3.0%
D2	17.9%	58.9%	23.2%
Test	20.0%	80.0%	0.0%

Win/Lose			
Dataset	Both Missing	One Missing	Both Mentioned
D1	14.8%	49.2%	36.0%
D2	17.9%	57.1%	25.0%
Test	0.0%	100.0%	0.0%

Experiment 2 Results

- D1: The system was unable to correctly learn the home/away rule (not enough examples where both Home and Away were mentioned)
- D2: The system is able to correctly learn the rules and so it matches the performance of the true rules
- An EM approach applied to D2 only achieves 50%.



Experiment 3

- Birthplace and Citizenship.
- Data: ACE08 Evaluation Corpus
- Citizenship mentioned 583 times
- Birthplace 25 times
- Only 6 articles mention both; 2 of which violate the rule

$$bornIn(X, C) \Rightarrow citizenOf(X, C)$$

	Probability Assigned to the Correct Interpretation	
Configuration	Gricean Method	EM Method
Citizenship missing	1.00	0.969
Birthplace missing	1.00	0.565

Experiment 4

- Somali Hijacking Incidents
- 41 news stories from coordination-maree-noire.eu
- Manual extractions
- 25 stories mention only one fact (ownership or flag)
- 16 mention both, 14 of which violate the rule

$$\mathit{ownershipCountry}(S, C) \Rightarrow \mathit{flagCountry}(S, C)$$

	Probability Assigned to the Correct Interpretation	
Configuration	Gricean Method	EM Method
Ownership missing	1.00	0.459
Flag missing	1.00	0.519

Discussion

- Inverse Gricean Rule Learning is able to learn correct rules from real extractions
  - Extractions should be tuned for high recall
  - Current algorithm relies on observing a decent number of cases where both body and head are correctly extracted
  - Rules are “correct” within Markov Logic, but not necessarily identical with the rules we would write by hand

Concluding Remarks

- We are making exciting contributions in Computational Sustainability
- Some of the ideas we are exploring have application in other parts of computer science

Thank-you

- National Science Foundation Grants 0832804, 0905885, 1125228
- DARPA Contract FA8750-09-C-0179 (BBN Technologies)

- Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF, DARPA, the Air Force Research Laboratory (AFRL), or the US government.