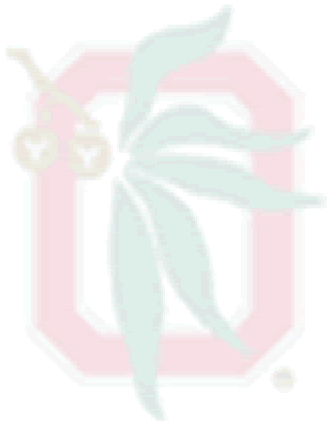


Knowledge Integration in ASR

Eric Fosler-Lussier

The Ohio State University

8 October 2003



Outline

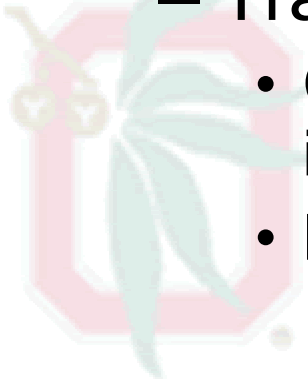
(or, rather, my list of questions)

- What is Knowledge Integration (KI)?
- How has KI influenced ASR to date?
- Where should KI be headed?
 - What types of cues should we be looking for?
 - How should cues be combined?

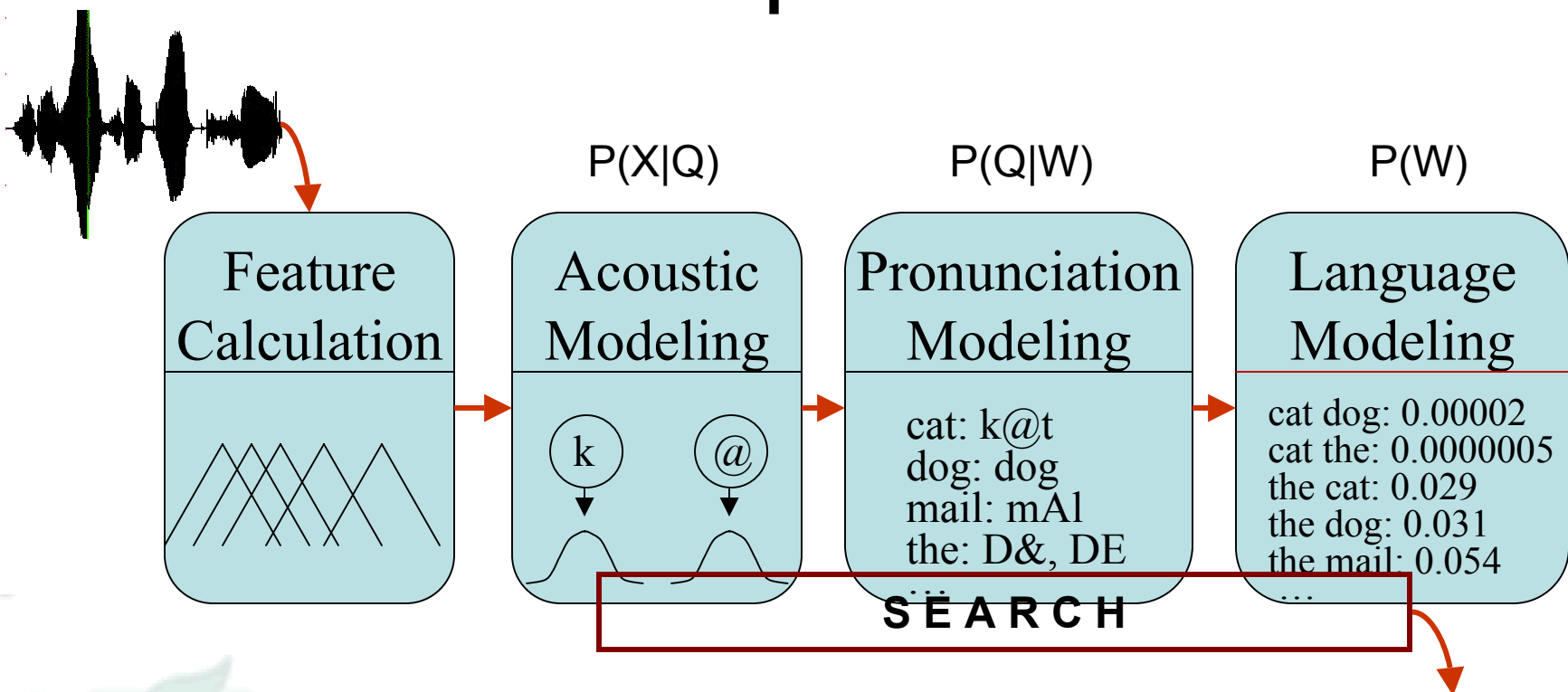


What is Knowledge Integration?

- It means different things to different people
 - Combining multiple hypotheses
 - Bringing linguistic information to bear in ASR
- Working definition:
 - Combining multiple sources of evidence to produce a final (or intermediate) hypothesis
 - Traditional ASR process uses KI
 - Combines acoustic, lexical, and syntactic information
 - But this is only the tip of the iceberg

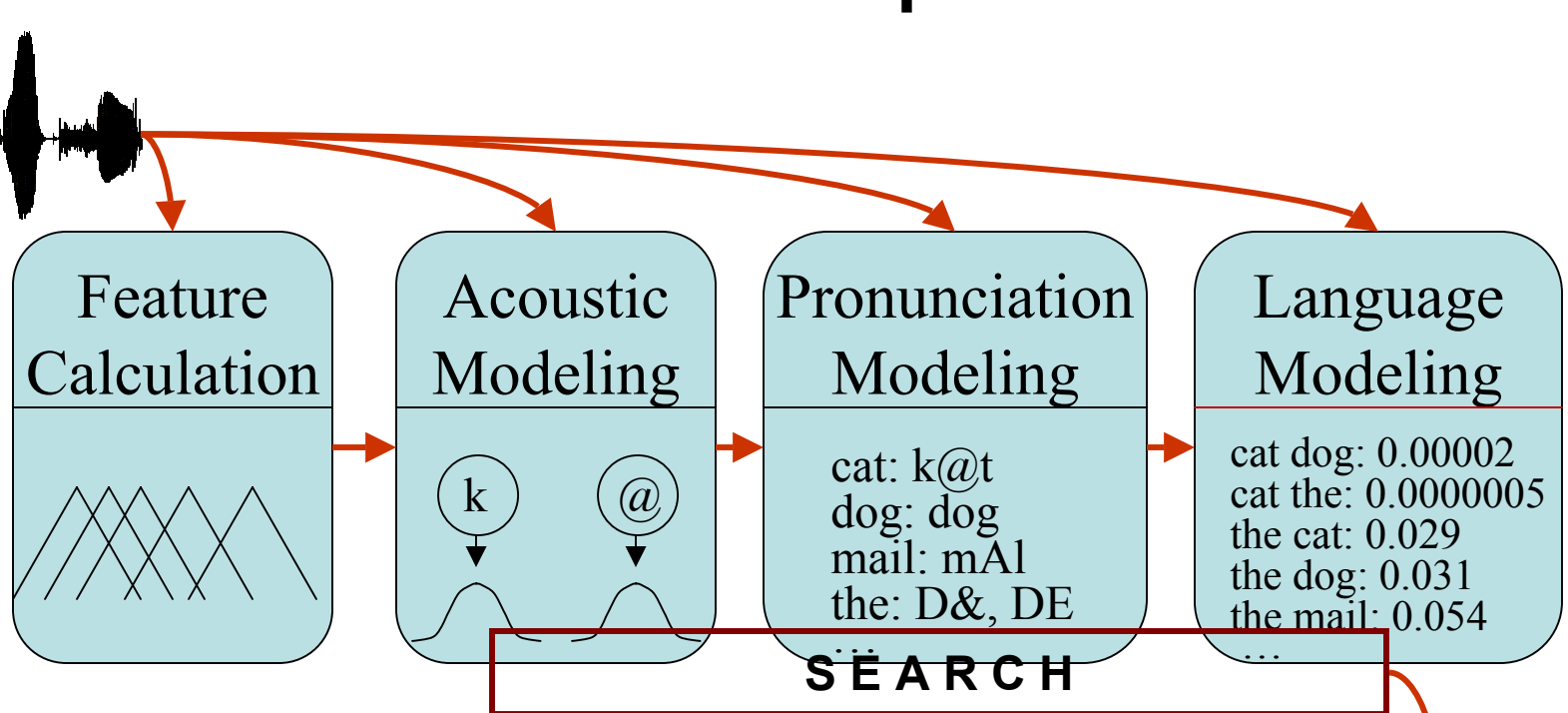


KI examples in ASR



- Acoustic model gives state hypotheses from features
- Search integrates knowledge from acoustic, pronunciation, and language models
- Statistical models have “simple” dependencies

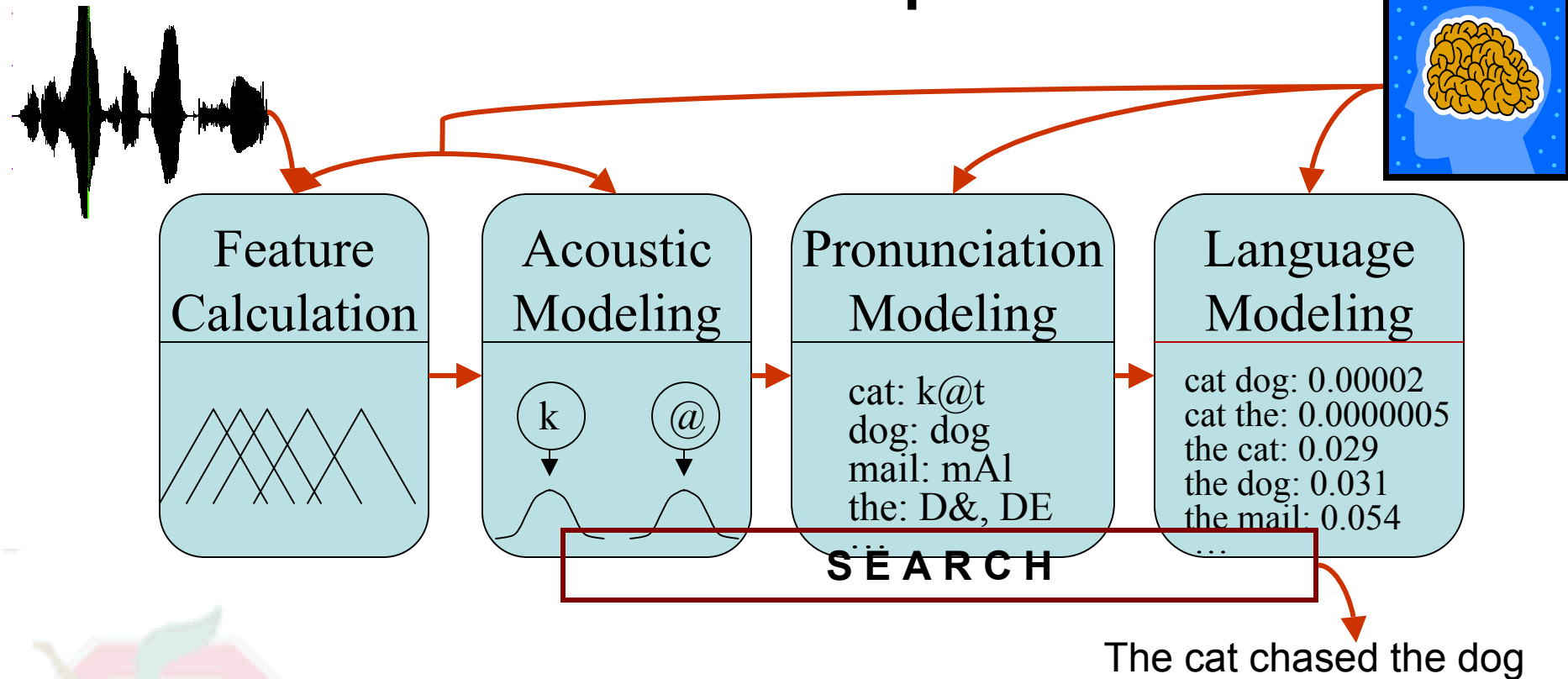
KI: Statistical Dependencies



The cat chased the dog

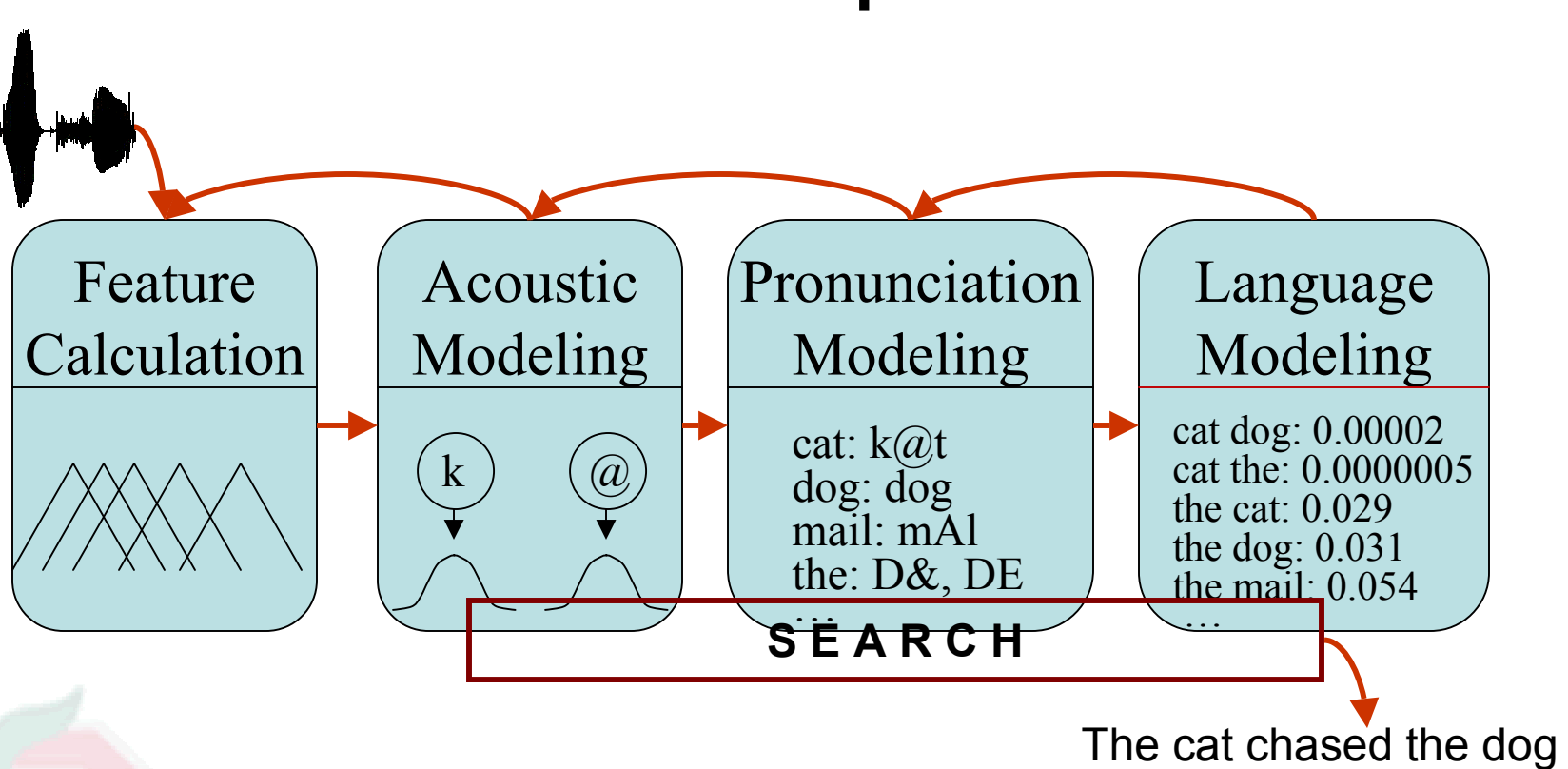
- “Side information” from the speech waveform
 - Speaking rate
 - Prosodic information
 - Syllable boundaries

KI: Statistical Dependencies



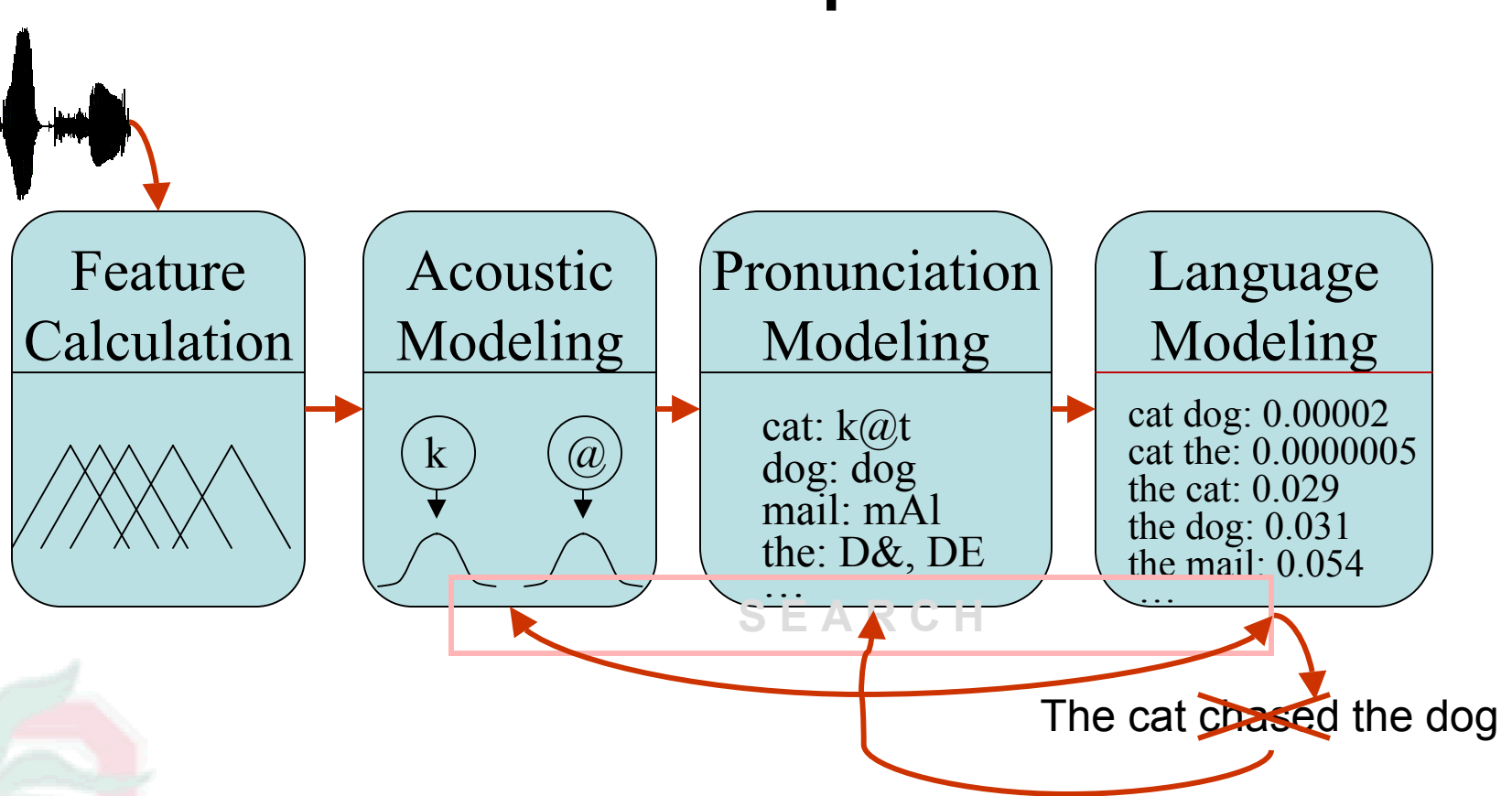
- Information from sources outside “traditional” system
 - Class n-grams, CFG/Collins-style parsers
 - Sentence-level stress
 - Vocal-tract length normalization

KI: Statistical Dependencies



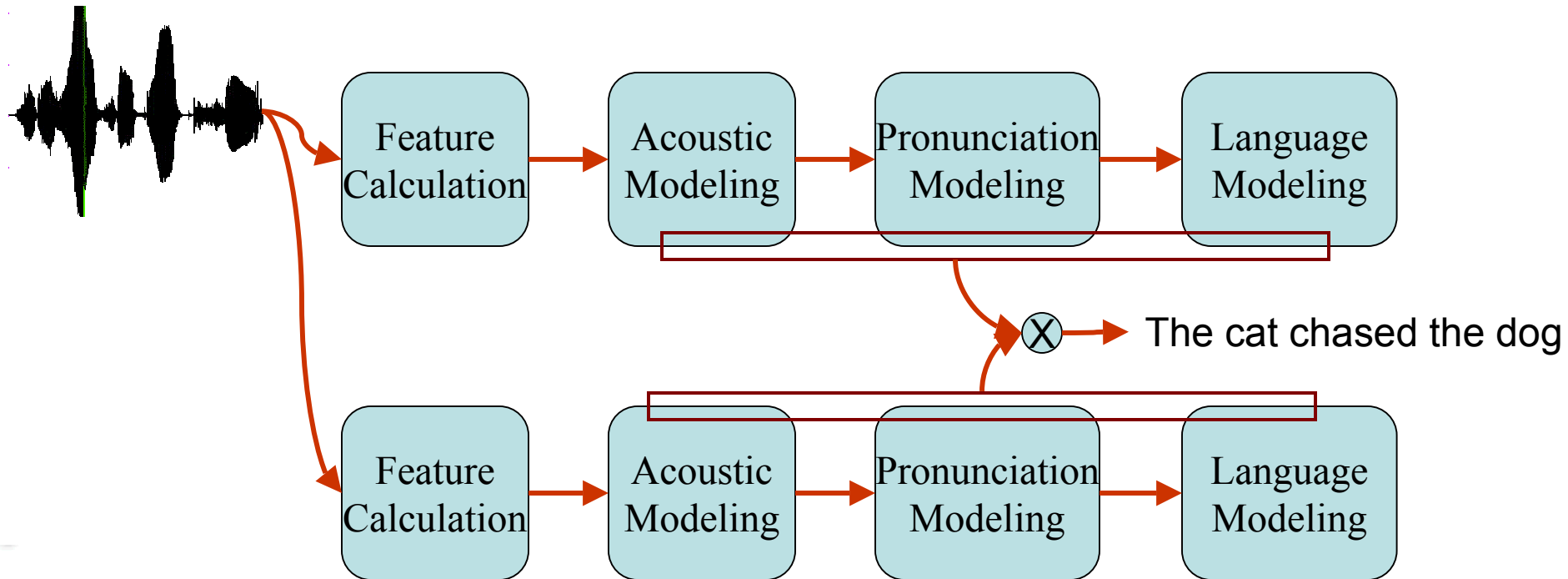
- Information from “internal” knowledge sources
 - Pronunciations w/ multi-words, LM probabilities
 - State-level pronunciation modeling
 - Buried Markov Models

KI: Statistical Dependencies



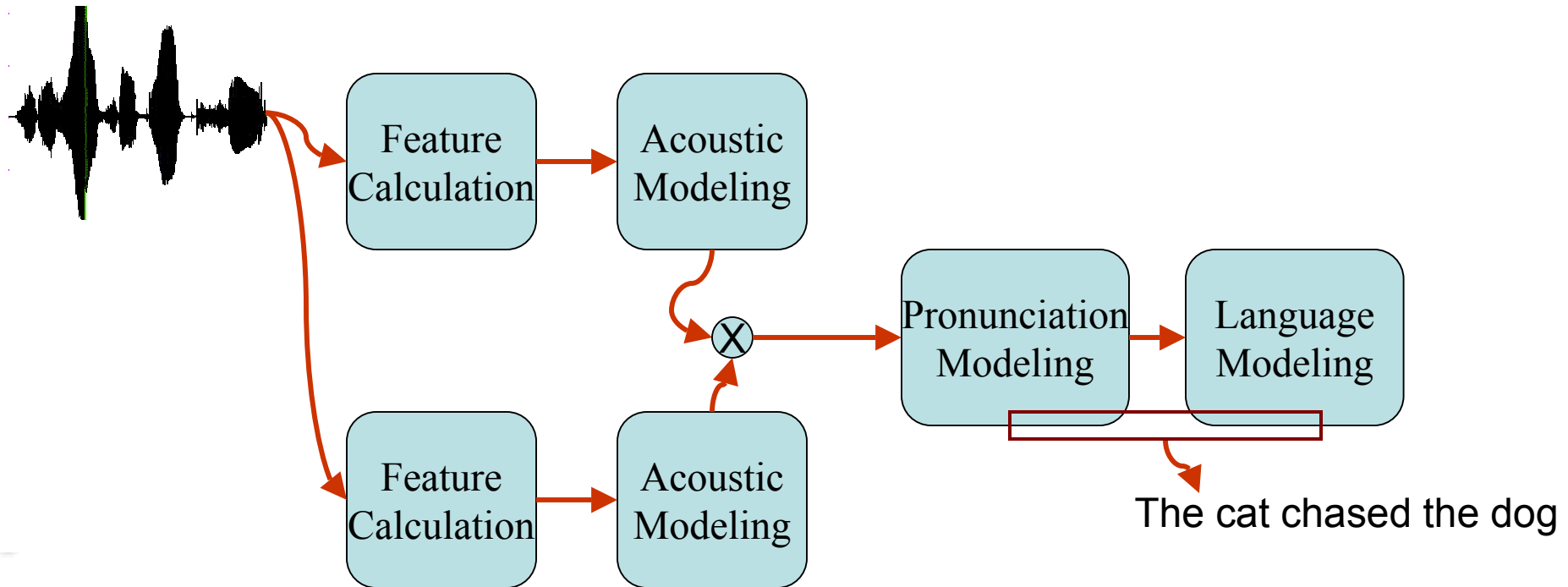
- Information from errors made by system
- Discriminative acoustic, pronunciation, and language modeling

KI: Model Combination



- Integrate multiple “final” hypotheses
 - ROVER
 - Word sausages (Mangu et al.)

KI: Model Combination



- Combine multiple “non-final” hypotheses
 - Multi-stream modeling
 - Synchronous phonological feature modeling
 - Boosting
 - Interpolated language models

Summary: Current uses of KI

- Probability conditioning

$$P(A|B) \rightarrow P(A|B, X, Y, Z)$$

- More refined (accurate?) models
- Can complicate overall equation

- Model merging

$$P(A|B) \rightarrow f(P_1(A|B), w_1) + f(P_2(A|B), w_2)$$

- Different views of information are (usually) good
- But sometimes combination methods are not as principled as one would like



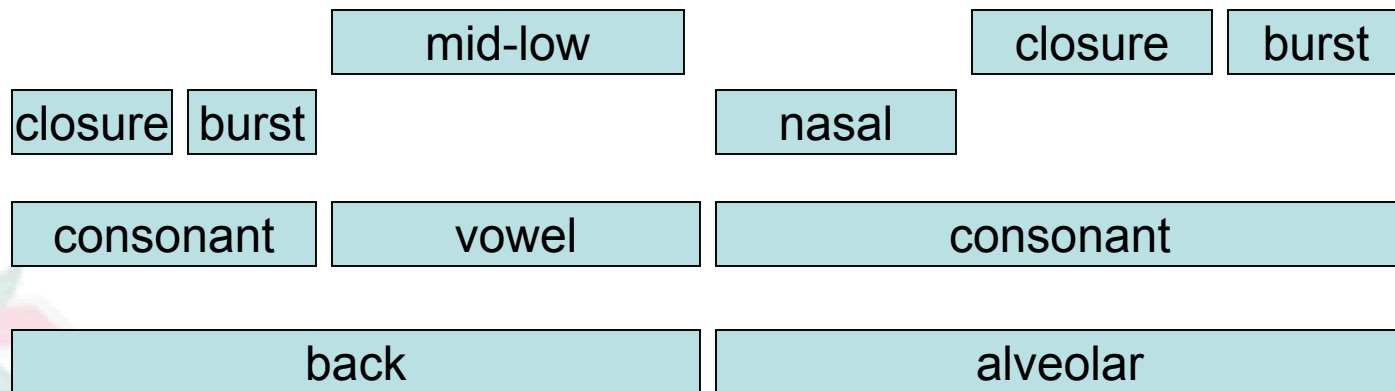
Where should we go from here?

- As a field have investigated many sources of knowledge
 - We learn more about language this way
 - Cf. “More data is better data” school
- To make an impact we need
 - A common framework
 - Easy ways to combine knowledge
 - “Interesting” sources of knowledge

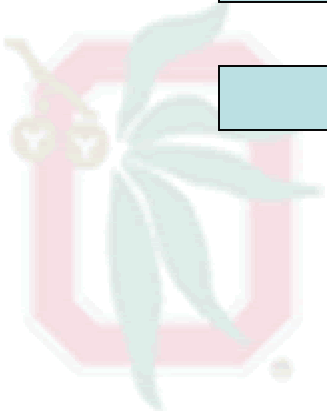


KI in Event-Driven ASR

- Phonological features as events
(from Chin's proposal)

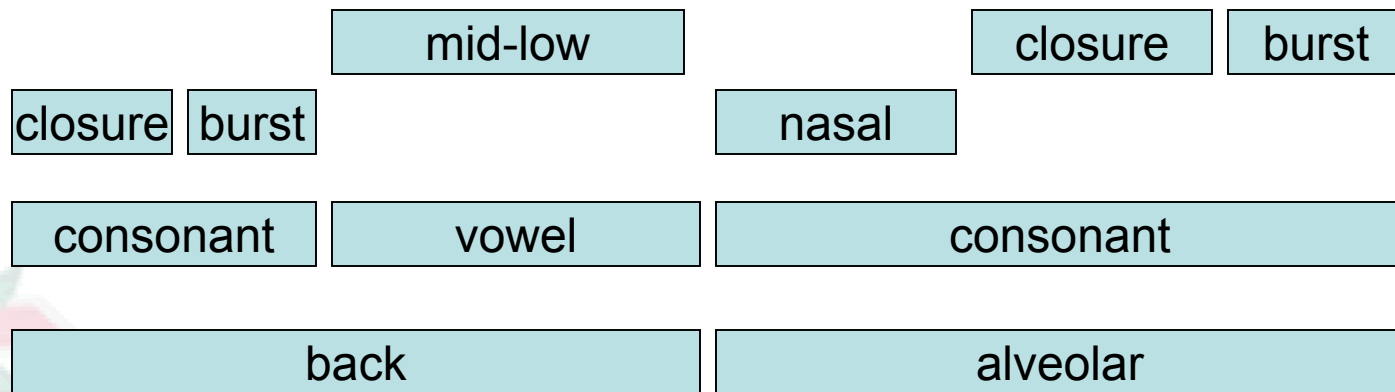


can't



KI in Event-Driven ASR

- Integrating multiple detectors
 - Easy if detectors are of the same type
 - Use both conditioning and model combination



$P(\text{back}|\text{detector1})$

$P(\text{back}|\text{detector2})$

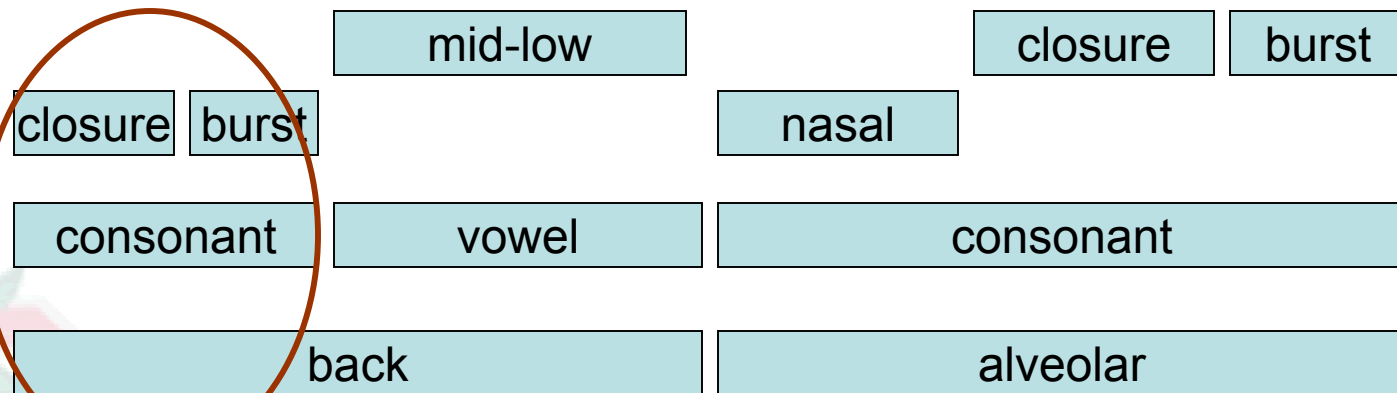
can't

KI in Event-Driven ASR

- Integrating multiple cross-type detectors

- Simplest to use Naïve Bayes assumption

$$P(X|e_1, e_2, e_3) = (P(e_1|X)P(e_2|X)P(e_3|X)P(X))/Z$$

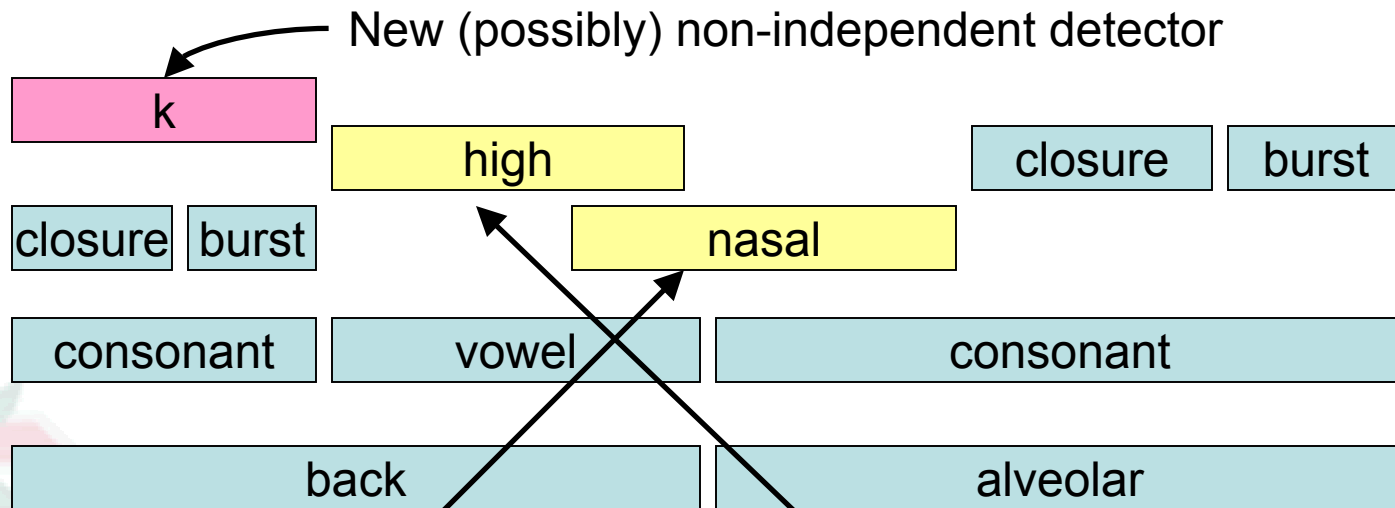


$P(k|features)$

can't

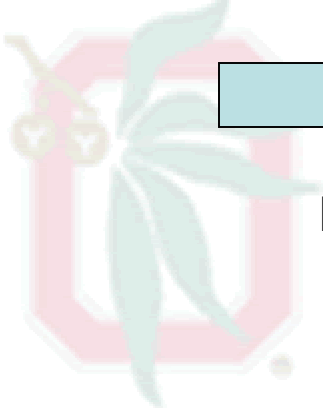
KI in Event-Driven ASR

- Breakdown in Naïve Bayes
 - Detectors aren't always independent



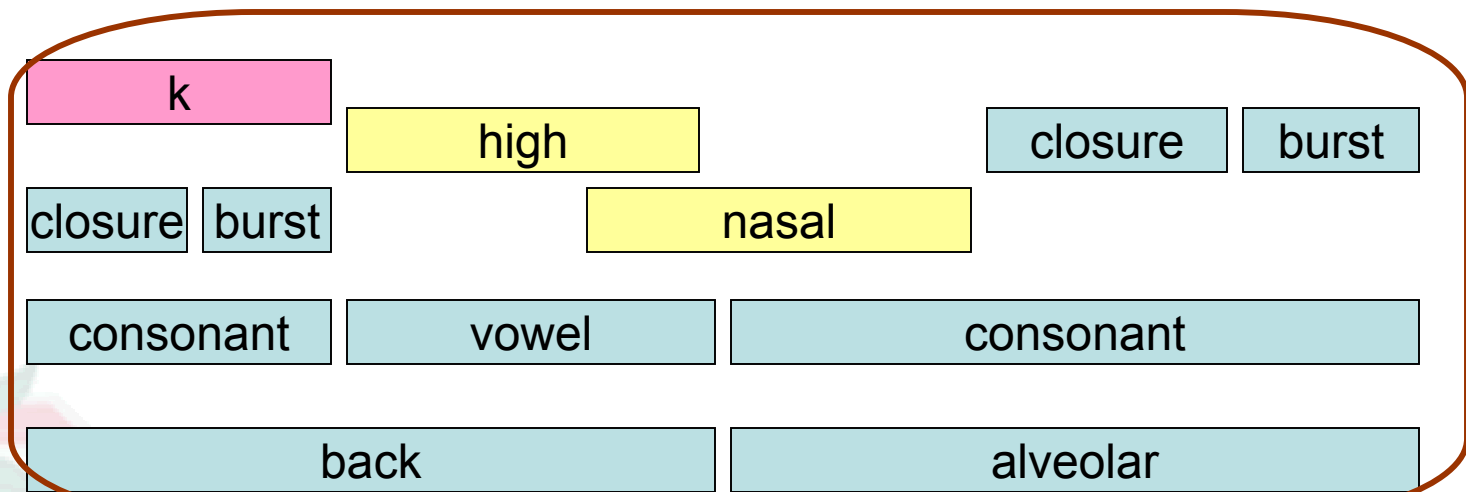
Feature spreading correlated with vowel raising

can't



KI in Event-Driven ASR

- Wanted: Gestalt detector
 - View overall shape of detector streams



P(can't|)

The Challenge of Plug-n-Play

- Shouldn't have to re-learn entire system every time a new detector is added
 - Can't have one global P(can't|all variables)
 - Changes should be localized
 - Implies need for hierarchical structure
- Composition structure should enable combination of radically different forms of information
 - E.g., audio-visual speech recognition



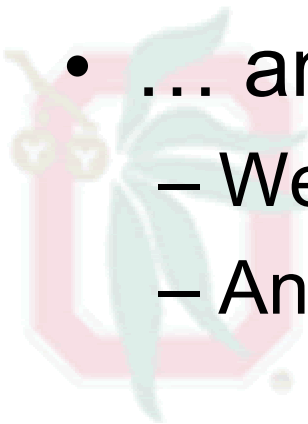
The Challenge of Plug-n-Play

- Perhaps need three types of structures
 - Event integrators
 - Is this a CVC syllable?
 - Problems like feature spreading become local
 - Hypothesis generators
 - I think the word “can’t” is here.
 - Combines evidence from top-level integrators
 - Hypothesis validators
 - Is this hypothesis consistent?
 - Language model, word boundary detection, ...
- Still probably have Naïve Bayes problems



What type of detectors should we be thinking about?

- Phonological features
- Phones
- Syllables? Words? Function Words?
- Syllable/word boundaries
- Prosodic stress
- ... and a whole bunch of other things
 - We've already looked at a number of them
 - And Jim's already made some of these points



Putting it all together

- Huge multi-dimensional graph search
- Should not be strictly “left-to-right”
 - “Islands of certainty”
 - People tend to emphasize the important words
 - ...and we can usually detect them better
 - Work backwards to firm up uncertain segments



Summary

- As a field, we have looked at many influences on our probabilistic models
- Have gained expertise in
 - Probability conditioning
 - Model combination
- Event-driven ASR may provide challenging, but interesting framework for incorporating different ideas

