Current & Future NLP Research

A Few Random Remarks

Computational Linguistics

- We can study anything about language ...

- 1. Formalize some insights
- 2. Study the formalism mathematically
- 3. Develop & implement algorithms
- 4. Test on real data

Reprise from Lecture 1: What's hard about this story?

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there.

- These ambiguities now look familiar
- You now know how to solve some (e.g., conditional log-linear models):
  - PP attachment
  - Coreference resolution (which NP does "it" refer to?)
  - Word sense disambiguation
    - Hardest part: How many senses? What are they?
- Others still seem beyond the state of the art (except in limited settings):
  - Anything that requires much semantics or reasoning
    - Quantifier scope
    - Reasoning about John's beliefs and actions
    - "Deep" meaning of words and relations

examples mostly from Terry Winograd in the 1970's, via Doug Lenat

Deep NLP Requires World Knowledge

- The pen is in the box.
  The box is in the pen.

- The police watched the demonstrators because they feared violence.
  The police watched the demonstrators because they advocated violence.

- Mary and Sue are sisters.
  Mary and Sue are mothers.

- Every American has a mother.
  Every American has a president.

- John saw his brother skiing on TV. The fool
  ... didn't have a coat on!
  ... didn't recognize him!

- George Burns: My aunt is in the hospital.
  I went to see her today, and took her flowers.
  Gracie Allen: George, that's terrible!

Big Questions of CL

- What formalisms can encode various kinds of linguistic knowledge?
  - Discrete knowledge: what is possible?
  - Continuous knowledge: what is likely?
  - What kind of p(...) to use (e.g., a PCFG)?
  - What is the prior over the structure (set of rules) and parameters (rule weights)?
  - How to combine different kinds of knowledge, including world knowledge?
- How can we compute efficiently within these formalisms?
  - Or find approximations that work pretty well?
  - Problem 1: Prediction in a given model. Problem 2: Learning the model.
- How should we learn within a given formalism?
  - Hard with unsupervised, semi-supervised, heterogeneous data ...
  - Maximize p(data | θ) · p_{prior}(theta)?
  - Pick θ to directly minimize error rate of our predictions?
  - Online methods? (adapt θ gradually in response to data, then forget)
  - Don't pick a single θ at all, but consider all values even at test time?
  - Learn just the feature weights θ, or also which features to have?
  - What if the formalism is wrong, so no θ works well?

Some of the Active Research

- Syntax:
  - Non-local features for scoring parses; discriminative models
  - Efficient approximate parsing (e.g., coarse to fine)
  - Unsupervised or partially supervised learning (learn a theory more detailed than one's Treebank)
  - Other formalisms besides CFG (dependency grammar, CCG, ...)
  - Using syntax in applied NLP tasks
- Machine translation:
  - Best-funded area of NLP, right now
  - Models and algorithms
  - How to incorporate syntactic structure?
  - "Low-resource" and morphologically complex languages?

Some of the Active Research

- Semantic tasks (how would you reduce these to prediction problems?)
  - Sentiment analysis
  - Summarization
  - Information extraction, slot-filling
  - Discourse analysis
  - Textual entailment
- Speech:
  - Better language modeling (predict next word) – syntax, semantics
  - Better models of acoustics, pronunciation
    - fewer speaker-specific parameters
      - to enable rapid adaptation to new speakers
    - more robust recognition
      - emotional speech, informal conversation, meetings
      - juvenile/elderly voices, bad audio, background noise
    - Some techniques to solve these:
      - non-local features
      - physiologically informed models
      - dimensionality reduction

Some of the Active Research

- All of these areas have learning problems attached.
- We're really interested in **unsupervised** learning.
- How to learn FSTs and their probabilities?
- How to learn CFGs? Deep structure?
- How to learn *good* word classes?
- How to learn translation models?

Semantics Still Tough

- **"The perilously underestimated appeal of Ross Perot has been quietly going up this time."**
  - Underestimated by whom?
  - Perilous to whom, according to whom?
  - "Quiet" = unnoticed; by whom?
  - "Appeal of Perot" ⇐ "Perot appeals ..."
    - a court decision?
    - to someone/something? (actively or passively?)
  - "The" appeal
  - "Go up" as idiom; and refers to amount of subject
  - "This time" : meaning? implied contrast?

Deploying NLP

- Speech recognition and IR have finally gone commercial.
- And there is a **ton** of text and speech on the Internet, cellphones, etc.
- But not much NLP is out in the real world.
- What killer apps should we be working toward?

- Resources (see Linguistic Data Consortium, LREC conference)
  - Treebanks (parsed corpora)
  - Other corpora, sometimes annotated
    - CORPORA mailing list
    - Mechanical Turk, annotation games
  - WordNet; morphologies; maybe a few grammars
  - Research tools:
    - Published systems (write to the authors & ask for the code!)
    - Toolkits: finite-state, machine learning, machine translation, info extraction
    - Dyna – a new programming language being built at JHU
    - Annotation tools
    - Emerging standards like VoiceXML
- Still out of the reach of J. Random Programmer

Deploying NLP

- Sneaking NLP in through the back door:
  - Add features to existing interfaces
    - "Click to translate"
    - Spell correction of queries
    - Allow multiple types of queries (phone number lookup, etc.)
    - IR should return document **clusters** and **summaries**
    - From IR to QA (question answering)
    - Machines gradually replace humans @ phone/email helpdesks
  - Back-end processing
    - Information extraction and normalization to build databases: CD Now, New York Times, ...
    - Assemble good text from boilerplate
  - Hand-held devices
    - Translator
    - Personal conversation recorder, with topical search

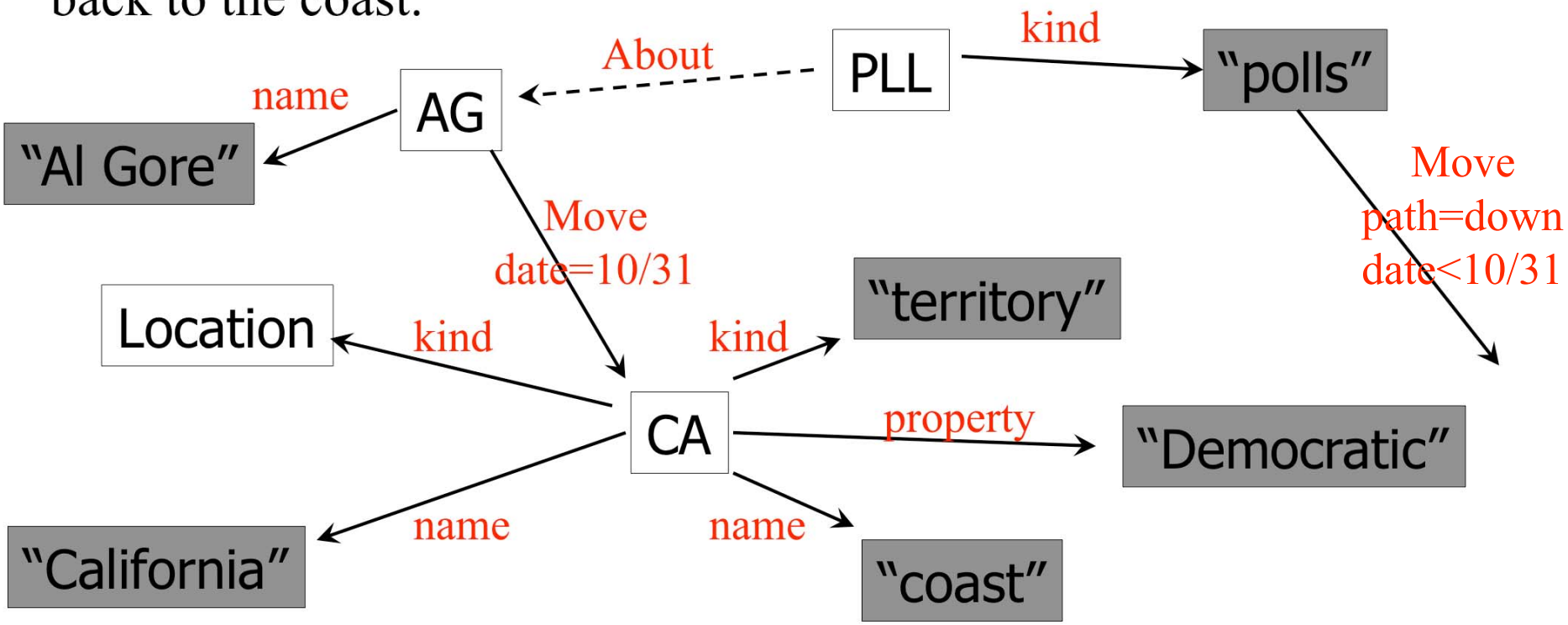
IE for the masses?

"In most presidential elections, Al Gore's detour to California today would be a sure sign of a campaign in trouble. California is solid Democratic territory, but a slip in the polls sent Gore rushing back to the coast."

NAME	AG	"Al Gore"	
NAME	CA	"California"	
NAME	CO	"coast"	
MOVE	AG	CA	TIME=Oct. 31
MOVE	AG	CO	TIME=Oct. 31
KIND	CA	Location	
KIND	CA	"territory"	
PROPRTY	CA	"Democratic"	
KIND	PLL	"polls"	
MOVE	PLL	?	PATH=down, TIME<Oct. 31
ABOUT	PLL	AG	

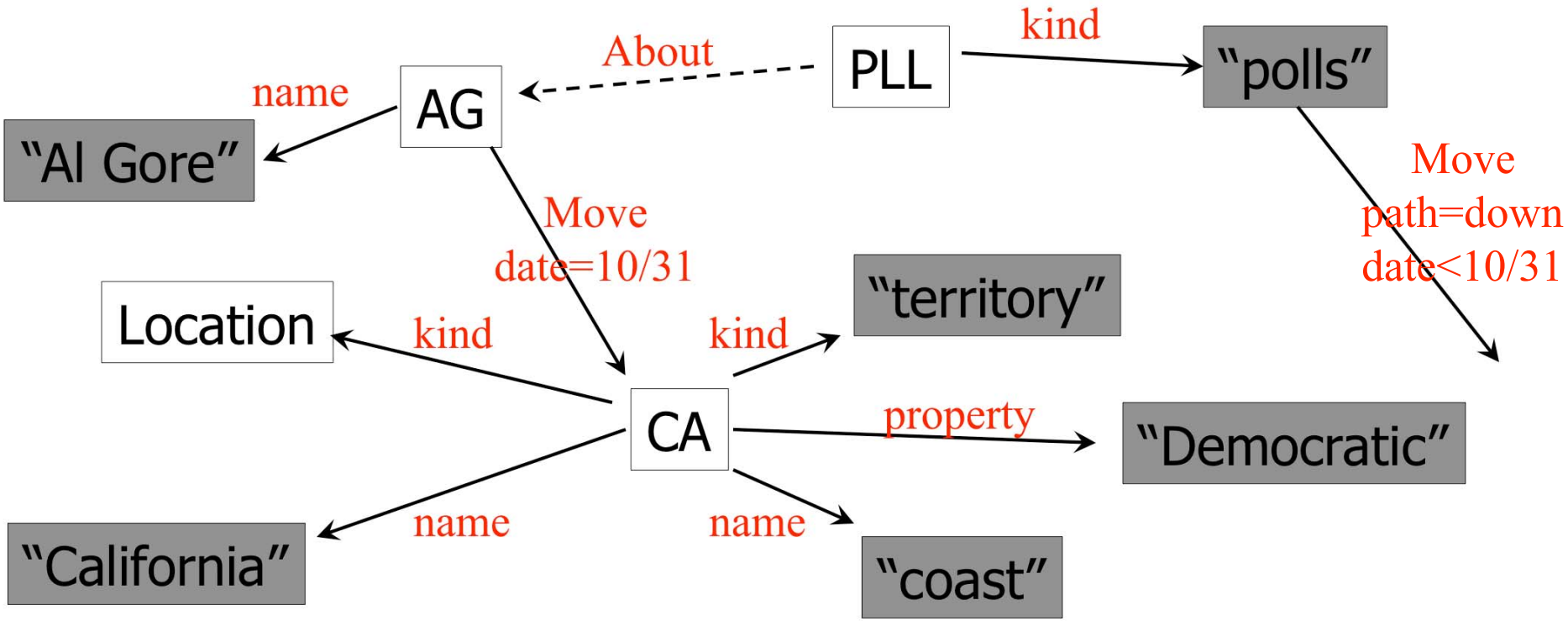
IE for the masses?

"In most presidential elections, Al Gore's detour to California today would be a sure sign of a campaign in trouble. California is solid Democratic territory, but a slip in the polls sent Gore rushing back to the coast."



IE for the masses?

- "Where did Al Gore go?"
- "What are some Democratic locations?"
- "How have different polls moved in October?"



IE for the masses?

- Allow queries over meanings, not sentences
- Big semantic network extracted from the web
- Simple entities and relationships among them
- Not complete, but linked to original text
- Allow inexact queries
  - Learn generalizations from a few tagged examples
- Redundant; collapse for browsability or space

Dialogue Systems

- Games
- Command-and-control applications
- “Practical dialogue” (computer as assistant)
- The Turing Test

Turing Test

Q: Please write me a sonnet on the subject of the Forth Bridge.

A [either a human or a computer]: Count me out on this one. I never could write poetry.

Q: Add 34957 to 70764.

A: (Pause about 30 seconds and then give an answer) 105621.

Q: Do you play chess?

A: Yes.

Q: I have my K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?

A: (After a pause of 15 seconds) R-R8 mate.

Turing Test

Q: In the first line of your sonnet which reads “Shall I compare thee to a summer’s day,” would not “a spring day” do as well or better?

A: It wouldn’t scan.

Q: How about “a winter’s day”? That would scan all right.

A: Yes, but nobody wants to be compared to a winter’s day.

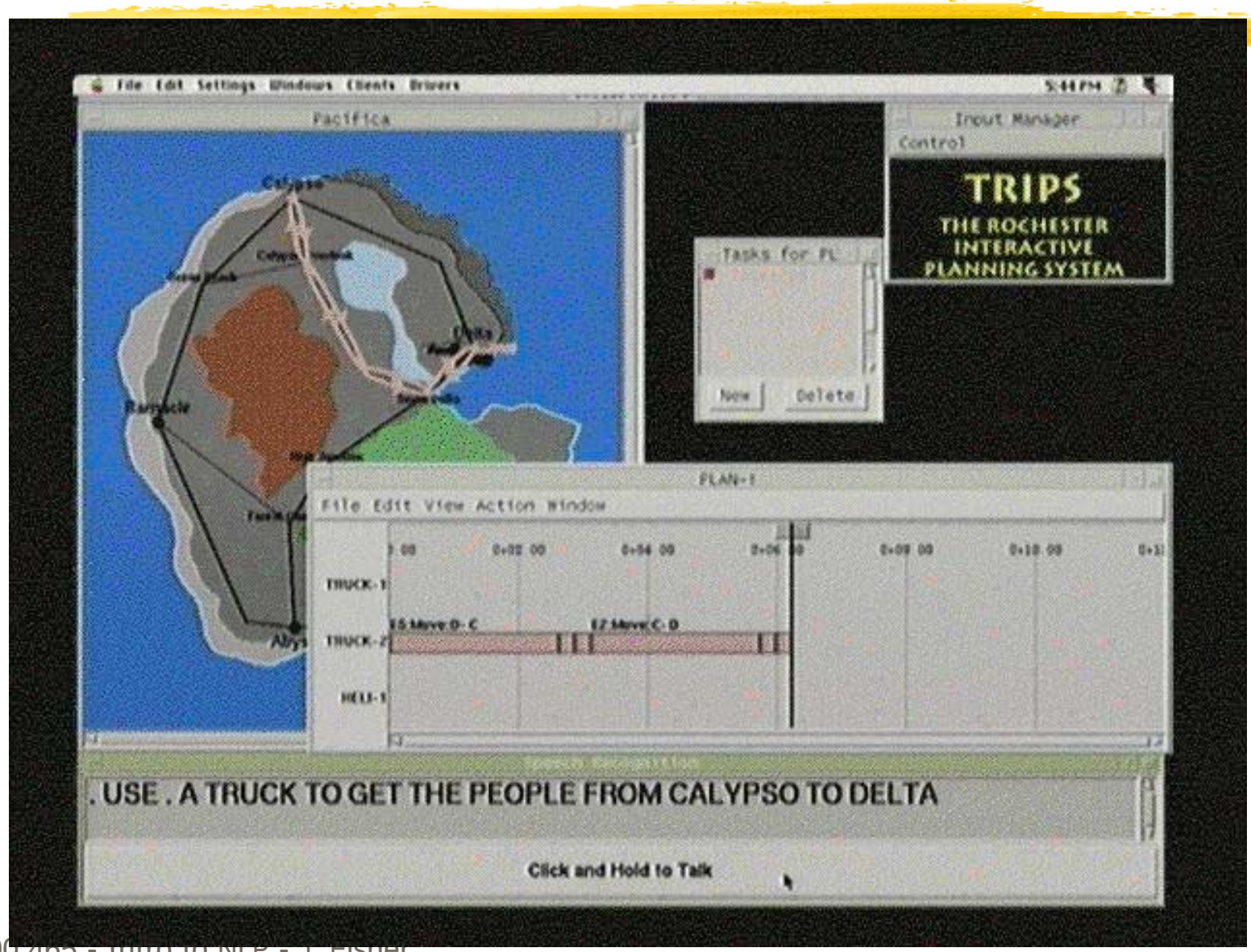
Q: Would you say Mr. Pickwick reminded you of Christmas?

A: In a way.

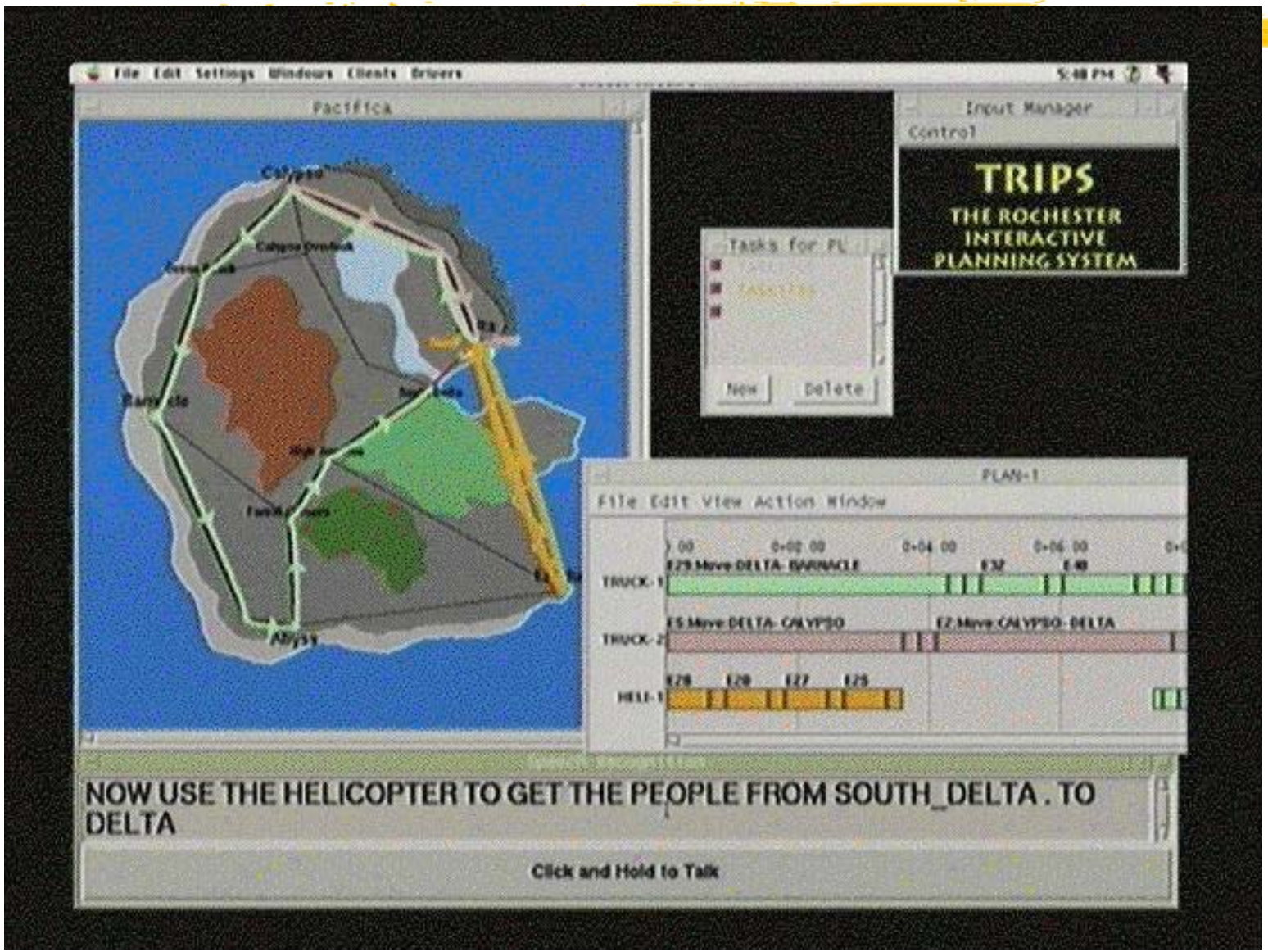
Q: Yet Christmas is a winter’s day, and I do not think Mr. Pickwick would mind the comparison.

A: I don’t think you’re serious. By a winter’s day one means a typical winter’s day, rather than a special one like Christmas.

TRIPS System



TRIPS System

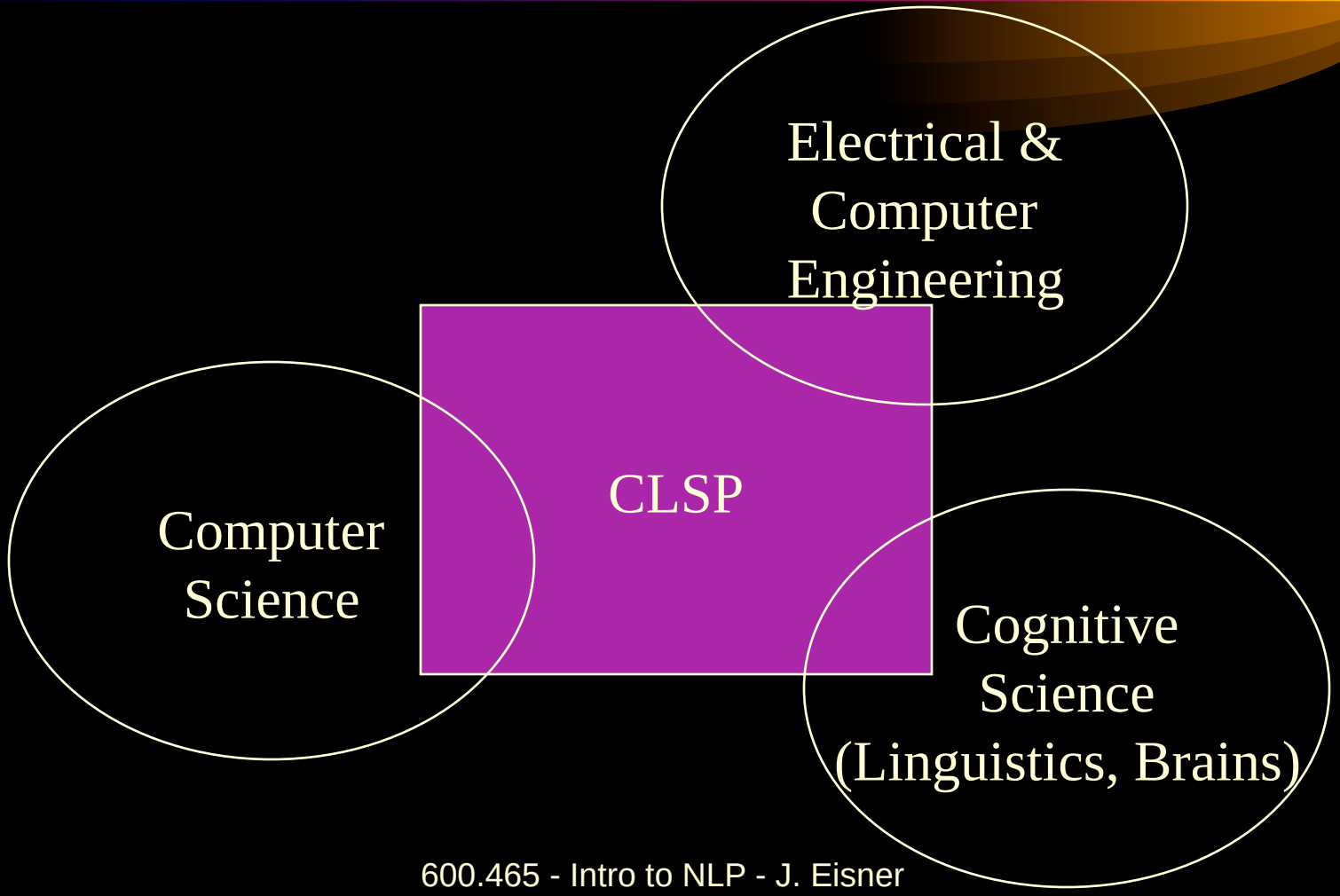


Dialogue Links (click!)

- Turing's article (1950)
- Eliza (the original chatterbot)
  - Weizenbaum's article (1966)
  - Eliza on the web - try it!
- Loebner Prize (1991-2001), with transcripts
  - Shieber: "One aspect of progress in research on NLP is appreciation for its complexity, which led to the dearth of entrants from the artificial intelligence community - the realization that time spent on winning the Loebner prize is not time spent furthering the field."
- TRIPS Demo Movies (1998)

JHU's Center for Language & Speech Processing

(one of the biggest centers for NLP/speech research)



CLSP Vision Statement

- Understand how human language is used to communicate ideas/thoughts/information.
- Develop technology for machine analysis, translation, and transformation of multilingual speech and text.

The form of linguistic knowledge: Mathematical formalisms for writing grammars

Paul Smolensky

Colin Wilson

Kyle Rawlins

+ others

Electrical & Computer Engineering

Computer Science

CLSP

Cognitive Science (Linguistics, Brains)

Recovering meaning in a noisy, ambiguous world: Statistical modeling of speech & language

Fred Jelinek | Sanjeev Khudanpur | Damianos Karakos

Hynek Hermansky | Mounya Elhilali | Andreas Andreou

Electrical & Computer Engineering

CLSP

Cognitive Science (Linguistics, Brains)

Natural Language Processing Lab: All of the above, plus algorithms

David Yarowsky

Jason Eisner

Chris Callison-Burch

Electrical & Computer Engineering

Computer Science

CLSP

Cognitive Science (Linguistics, Brains)

bunch of great students!

Human Language Technology Center of Excellence (HLT-CoE)



rocessing

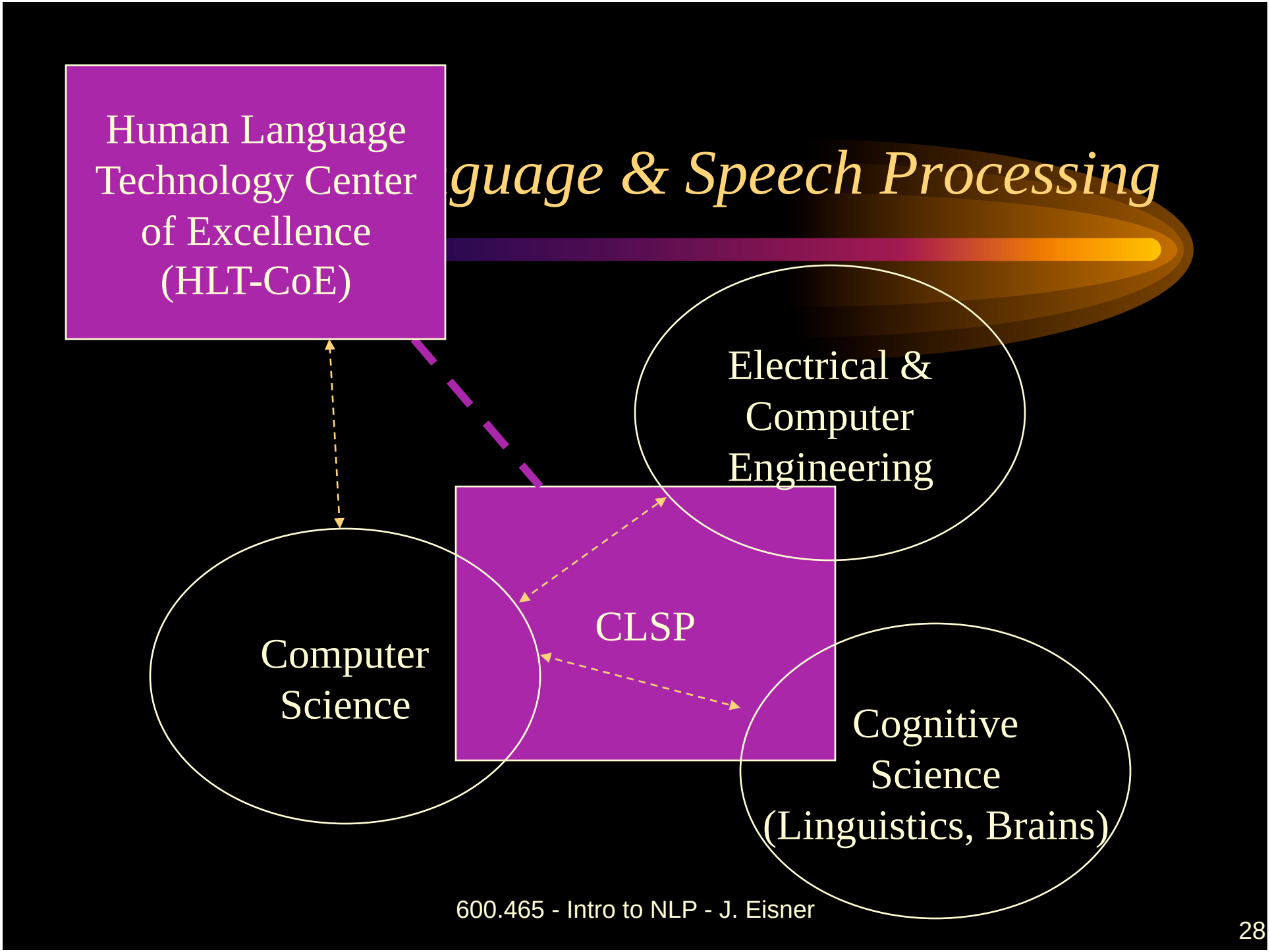
Ken Church | Mark Dredze | Christine Piatko | (+ several others)

Computer Engineering

CLSP

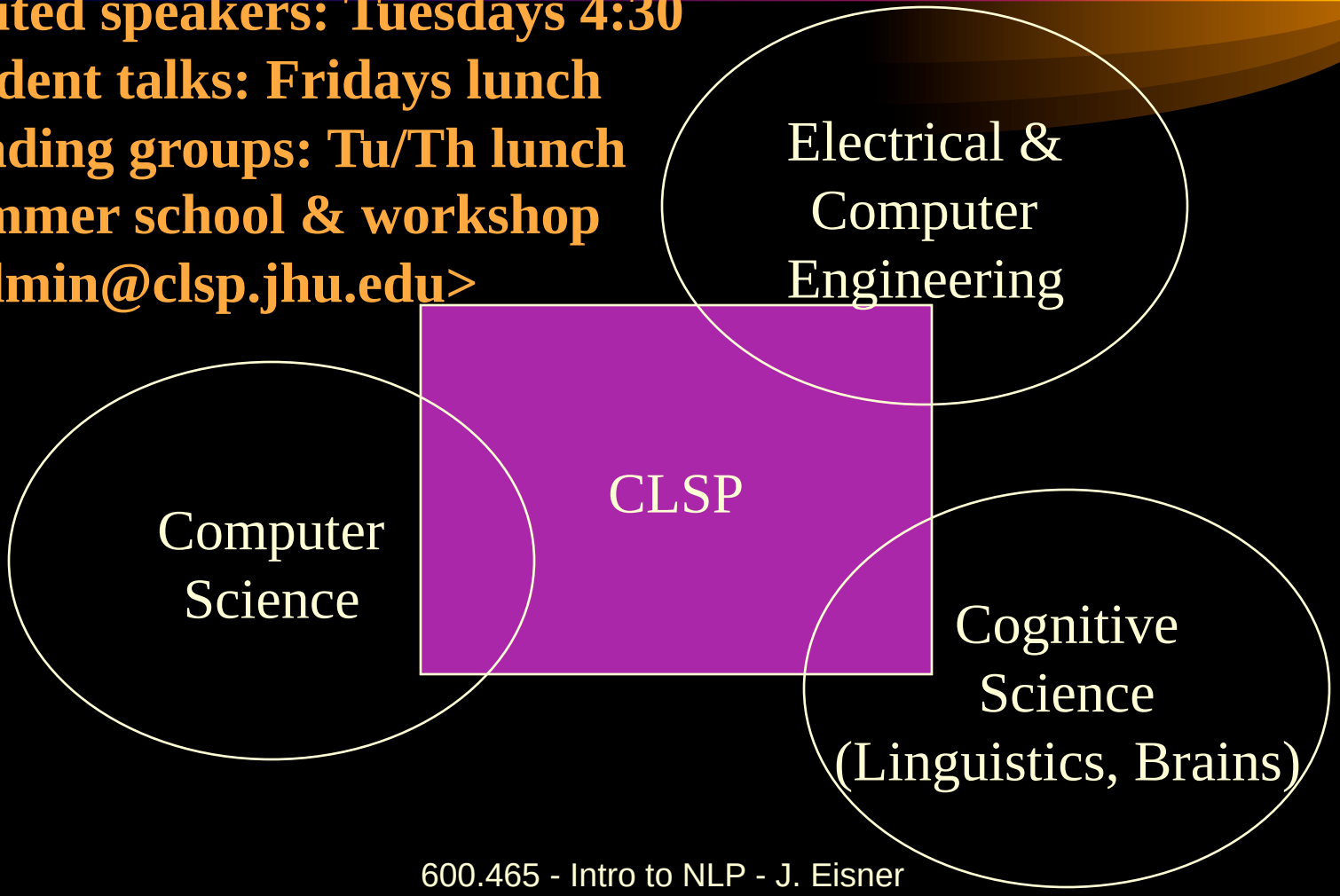
Computer Science

Cognitive Science (Linguistics, Brains)


Human Language Technology Center of Excellence (HLT-CoE)
guage & Speech Processing
Electrical & Computer Engineering
CLSP
Computer Science
Cognitive Science (Linguistics, Brains)

Center for Language & Speech Processing

Invited speakers: Tuesdays 4:30
Student talks: Fridays lunch
Reading groups: Tu/Th lunch
Summer school & workshop
<admin@clsp.jhu.edu>



Why Language?

y_0 ?

Well, at least you can use it to make jokes with …

Why Language?

- Selfish reasons
  - Really interesting data
  - Use both sides of your brain
  - Great problems => lifetime employment?
- $elfish reason$
  - space telescope: “all” cosmological data
  - genome: “all” biological data
  - online text/speech: “all” human thought and culture
    - suddenly PCs can see lots of speech & text – but they can’t help you with it until they understand it!
- Sound fun? 600.465 Natural Language Processing
  - techniques are transferable (comp bio, stocks)

Typical problems & solution

- Map input to output:
  - speech → text
  - text → speech
  - Arabic → English
  - sentence → meaning
  - unedited → edited
  - document → summary
  - document → database record
  - query → relevant documents
  - question → answer
  - email → is it spam?

1. Dream up a model of p(output | input)
2. Fit the model's parameters from whatever data you can get
3. Invent an algorithm to maximize p(output | input) on new inputs

One of two language-learning devices I recently helped build (this is model 1, from 2003)

2005 (fairly fluent)

2004 (pre-babbling)