# Integrating Syntactic and Semantic Annotation of Biomedical Text

Seth Kulick, Mark Liberman, Martha Palmer and Andrew Schein

The University of Pennsylvania

Support from: NSF ITR-EIA-0205448

# Contributors

## The University of Pennsylvania

- Ann Bies, Susan Davidson, Hubert Jin, Aravind Joshi, Seth Kulick, Jeremy Lacivita, Mark Liberman, Mark Mandel, Mitch Marcus, Marty McCormick, Tom Morton, Martha Palmer, Eric Pancoast, Fernando Pereira, Andrew Schein, Val Tannen, Lyle Ungar, Peng Wang

## eGenome (Children's Hospital of Philadelphia)

- Yang Jin, Peter White, Scott Winters

## GlaxoSmithKline

- Jim Butler, Paula Matuszek, Robin McEntire

## Other

- Robert Gaizauskas, Jun-ichi Tsujii, Bonnie Webber

# Goal

- Information Extraction from the biomedical literature, particularly Medline
  - Enzyme Inhibition Relations

    Expression of CYP3A11 and PXR was suppressed by inactivation of HNF4alpha

    customer: GlaxoSmithKline
  - Mutation/Malignancy Relations

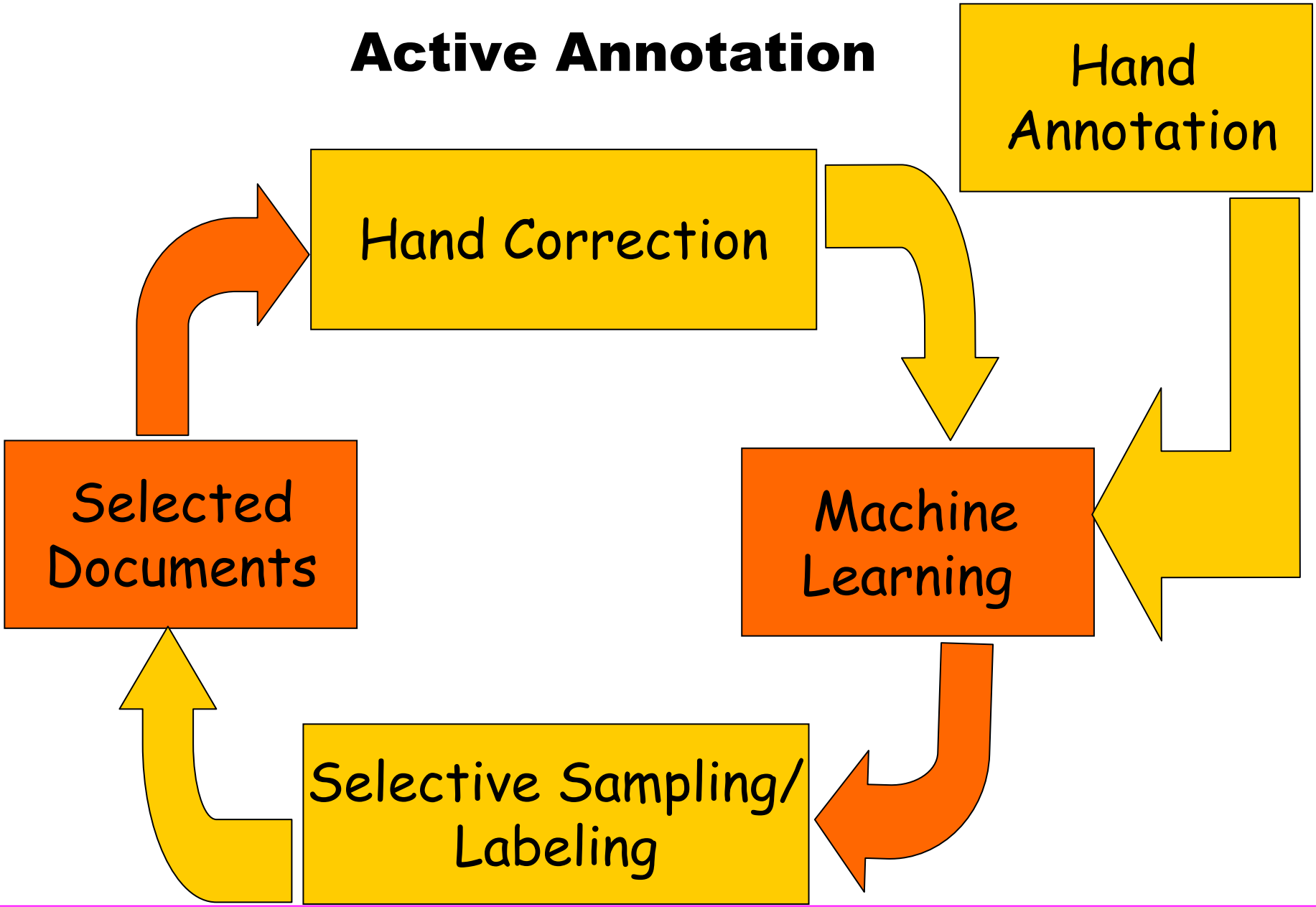
    Ki-ras mutations were detected in 17.2% of the adenomas.

    customer: eGenome
- Annotate 1-10K abstracts for each domain

# Approach to Information Extraction

- Phase 1:
  - Develop definitions and ontologies
  - Annotate data according to definitions
- Phase 2: Train corpus-based algorithms exploiting various annotation:
  - Parsing
  - Predicate-argument analysis
  - Reference resolution
- Phase 3: "Active Annotation"

# Active Annotation



# Challenge: Diversity in Expression

1. "Activation of the C-Ki-ras genes by point mutations in codons 12 or 13..."
2. "Point mutations in codons 12 and 13 activated C-Ki-ras"
3. "Point mutations in codons 12 and 13 were activators of C-Ki-ras gene"

Want to populate a factbank with:

activation(C-Ki-ras, point mutation in codon 12)

activation(C-Ki-ras, point mutation in codon 13)

# Approaches to Handling Diversity

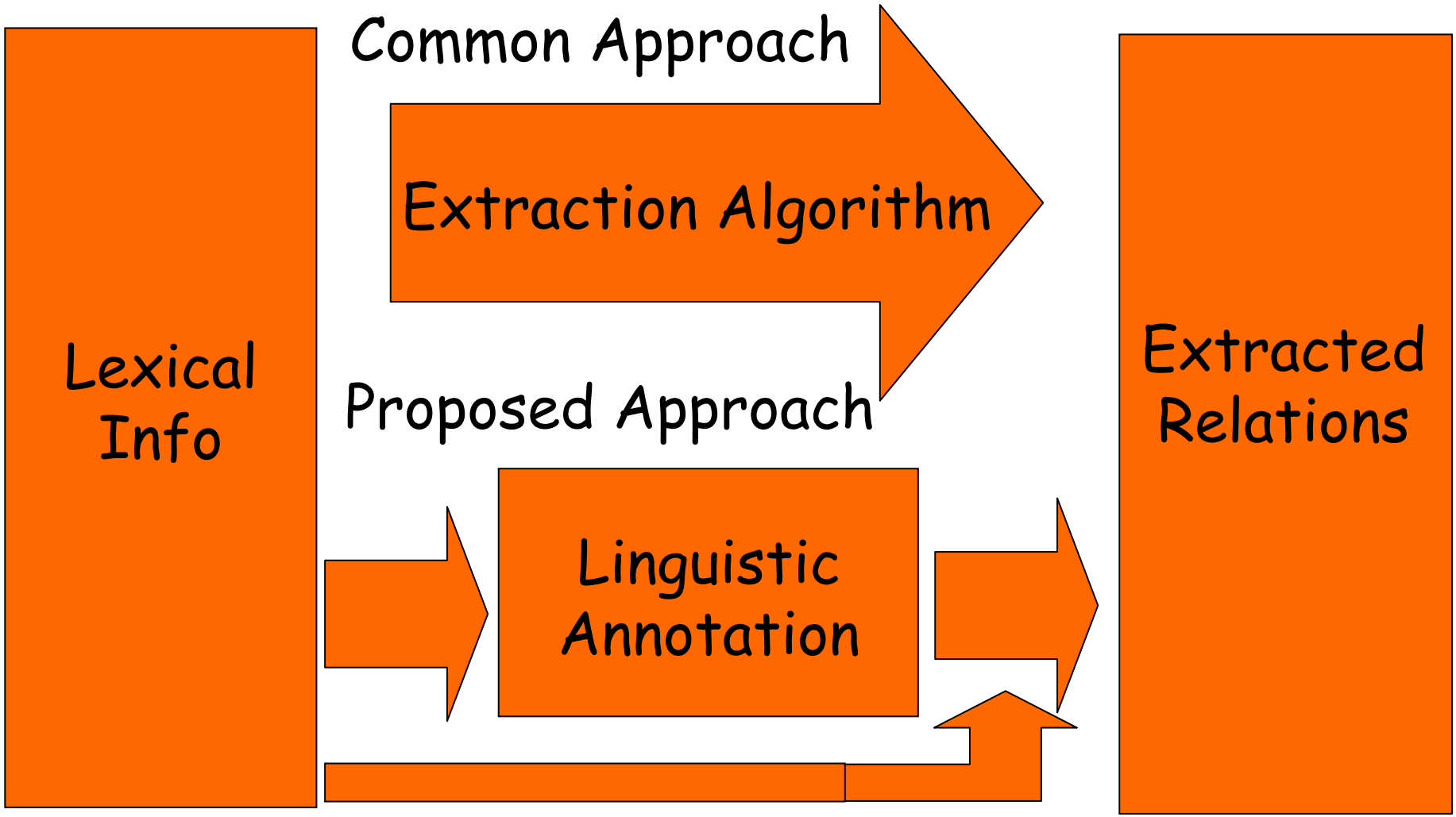
- Current Approach is to either:
  - Hand build extraction patterns to cover all variant expressions

    **or**

  - Annotate **lots** of data to get examples of variant expressions (for machine learning)
- Proposed Approach:

  Linguistic analysis of the sentences

# Information Extraction Approaches



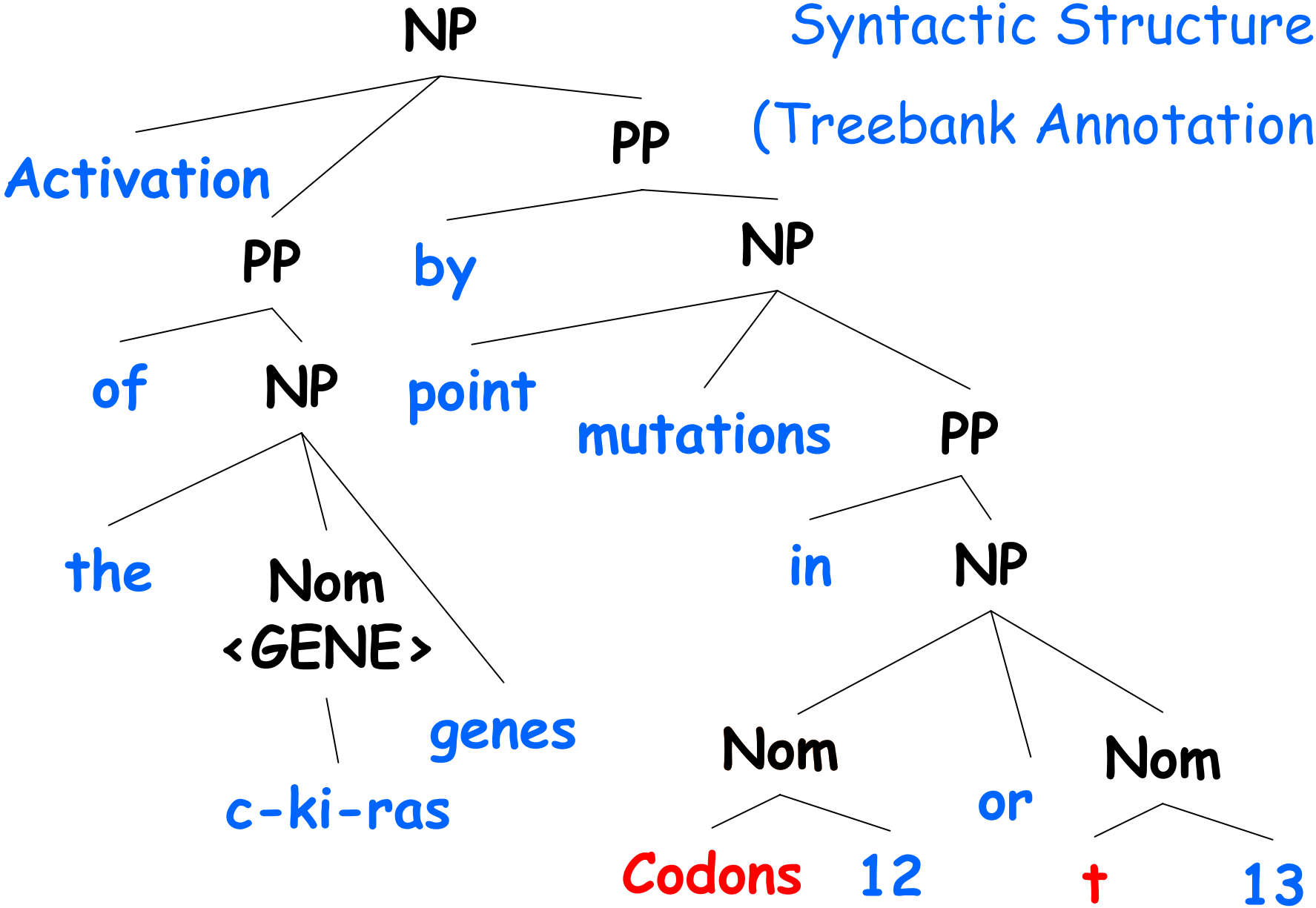
# Our Annotation Effort

**Together for the first time…**

Annotations include:

- Treebank (Syntax)
- Probank (predicate-argument structure)
- Entities (genes, malignancies)
- Reference and Coreference
- Factbanking (end goal)

# Syntactic Structure (Treebank Annotation)



# More Examples of Coordination

- "the ortho and meta positions"
  - (= the ortho positions and meta positions)
- "PLC and cytochrome P450 arachidonate epoxygenase activity"
  - (= PLC arachidonate epoxygenase activity and cytochrome P450 arachidonate…)
- "enhanced CYP2C9 expression and 11,12 EET production"
  - (= enhanced CYP2C9 expression and enhanced 11,12 EET production)

# Predicate-Argument Annotation: Propbank

- "Point mutations in codons 12 and 13 were activators of C-K-ras genes"
- "Activation of the C-K-ras genes by point mutations in Codons 12 or 13..."
- Predicate-Argument Structure (Propbank):
  - REL: activation
    activatee: c-ki-ras genes
    activator: point mutations in codons 12 or 13
  - REL: mutations
    type: point
    position(s): Codons 12 or 13

# Why Combine Treebank and Propbank?

- Treebank indicates constituents
  - subject, verb, direct object, etc.
- Propbank indicates roles of constituents
  - "agent," "theme," "quantification", etc.
  - inhibitor, inhibitee, inhibition rate
- Prior work combines Treebank/Propbank for financial text IE:

  (Surdeneau et al., 2003, Gildea and Palmer, 2002)

# Entity Annotation

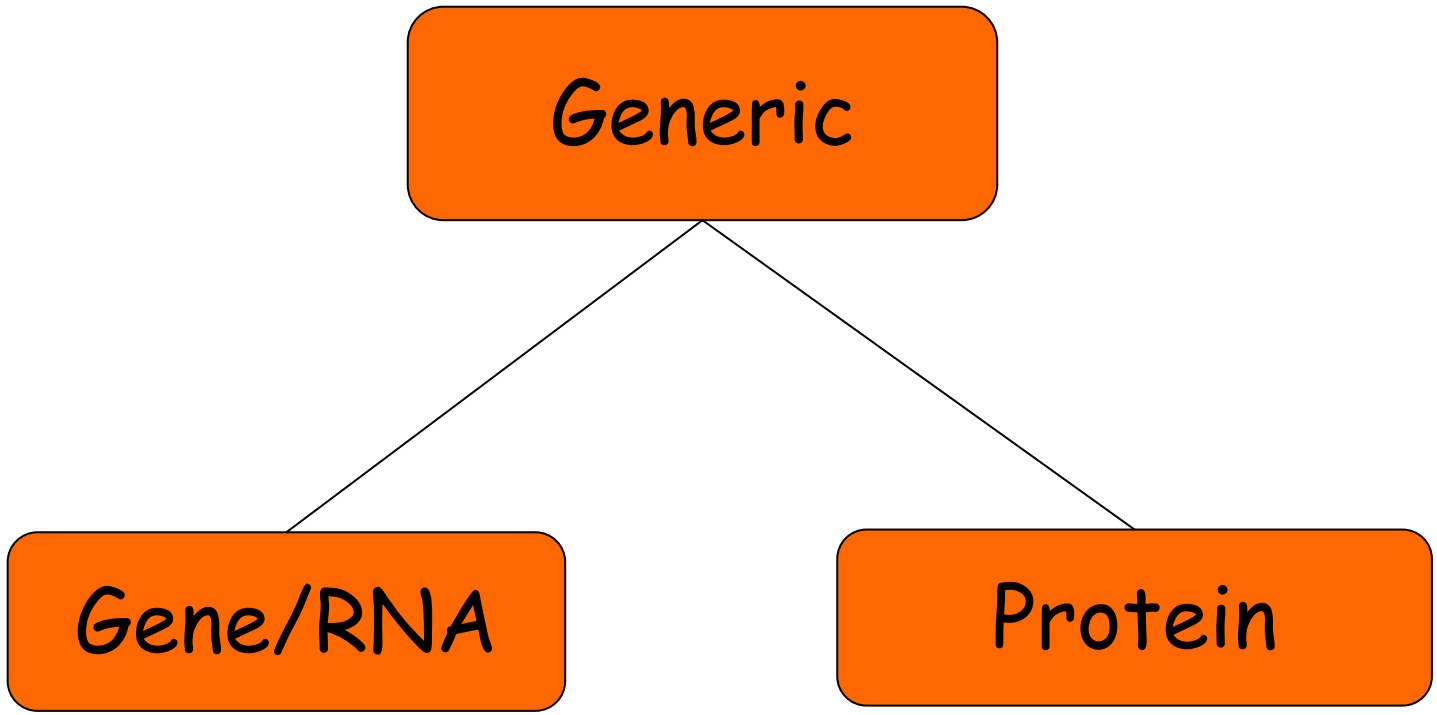
- Entities we annotate include:

  "gene", "protein", "substance", "malignancy"

- Metonymy Issues:
  - is a reference a gene or a protein?
  - We use subtypes, following ACE conference convention
  - Gene is broken in to three categories:

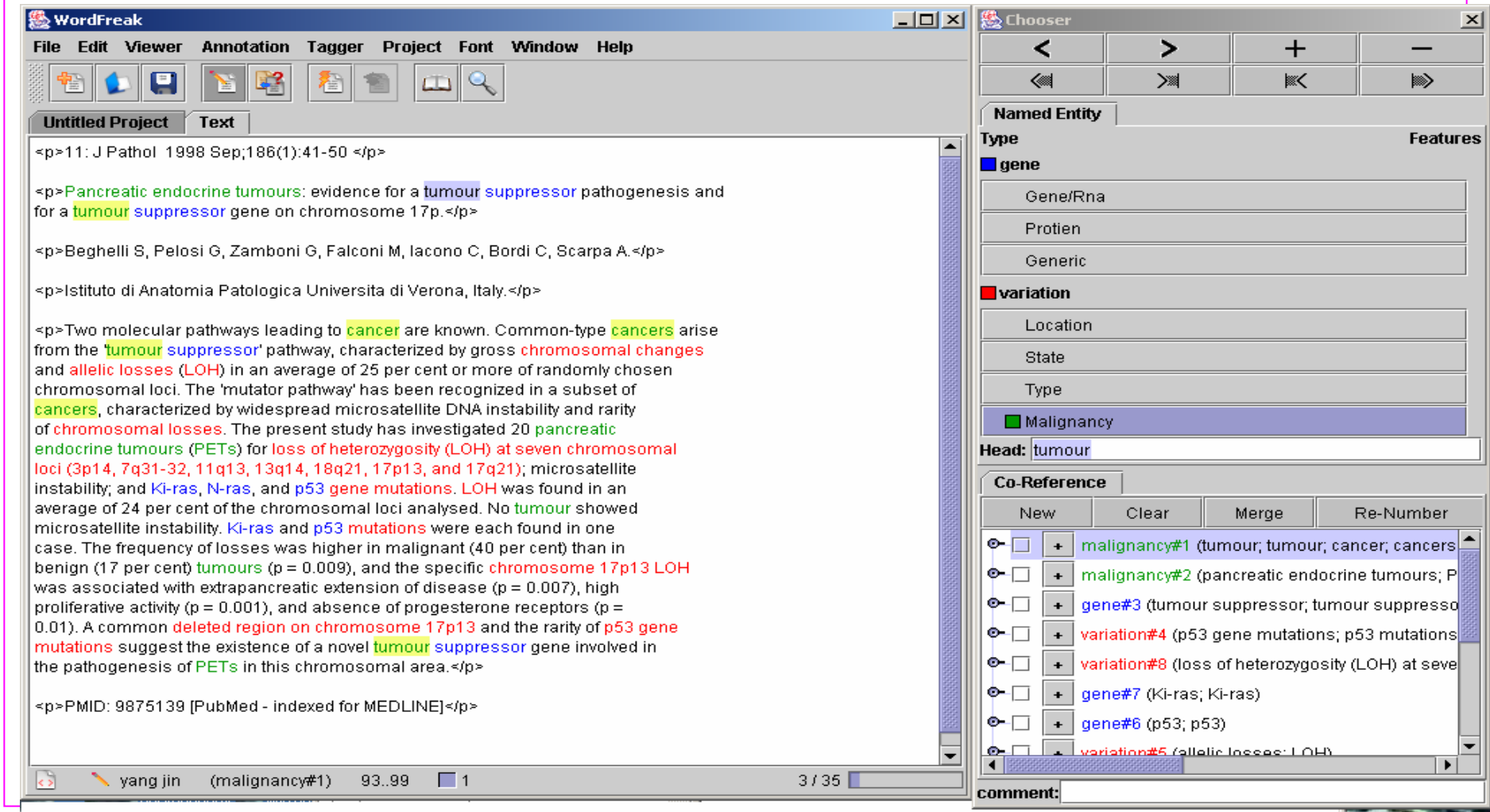
    "Generic," "Gene/RNA" and "Protein"

# The Gene Entity



# WordFreak Annotation Tool

Morton, Lacivita, Pancoast: www.annotation.org



# Reference and Co-reference Annotation

- Co-reference is an equivalence relation
- subtypes prevent nonsense in a co-ref graph

Example of reference types:

"K-Ras is a member of the Ras family of Oncogenes. The protein form is actively expressed in…"

class-membership(K-Ras, Ras family)

anaphor(K-Ras_protein, protein form)

# Current Activities

- In Progress:
  - Entity Annotation of "Gene," "Chemical," "Malignancy," "genetic variation," etc.
  - POS annotation
  - Training Treebank Syntactic Annotators
- Starting Up:
  - Start coreference annotation
  - Build our first entity tagging models

# Some Projected Milestone Dates

- January 2004 -

  Entity tagging and coreference on oncology domain complete. We publish:

  - annotation guidelines
  - data
  - baseline statistical taggers

- May 2004 -

  First draft syntactic analysis of oncology domain

(1-10K Medline abstracts)

# Some Annotation Projects and Related Research

- GENIA Project and U Tokyo Work:
  http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA
- Pasta system and Sheffield Work:
  http://nlp.shef.ac.uk/research/areas/bio.html
- GENIES system and Columbia/CUNY Work
- Modeling Linguistic Phenomenon:
  - Ray/Craven, IJCAI-2001
  - Pustejovsky et al. 2003

# The End.

# Some Examples Follow

# Reference and Co-reference

- Our reference subtypes are:
  - Acronyms (definitions and linkages)
  - Anaphor (such as pronouns)
  - Classes versus their members
  - "Is-a" relation,

    i.e. "{CYP450}, {an enzyme} found in…"
  - Standardized database reference

# Complex Coordination Example

Inhibition of CB -52 and -101 metabolism

Note coordination of "CB" and also "metabolism"!

The sentence above can be represented as:

Inhibition of CB-52 metabolism and CB-101 metabolism)