

# Iterative Automated Named Entity Recognition (NER) to Bootstrap Biomedical Lexicon Generation

Swarmfest 2007

July 13, 2007 Chicago, IL

Gary An, MD (1) and Steven Lytinen, PhD (2)

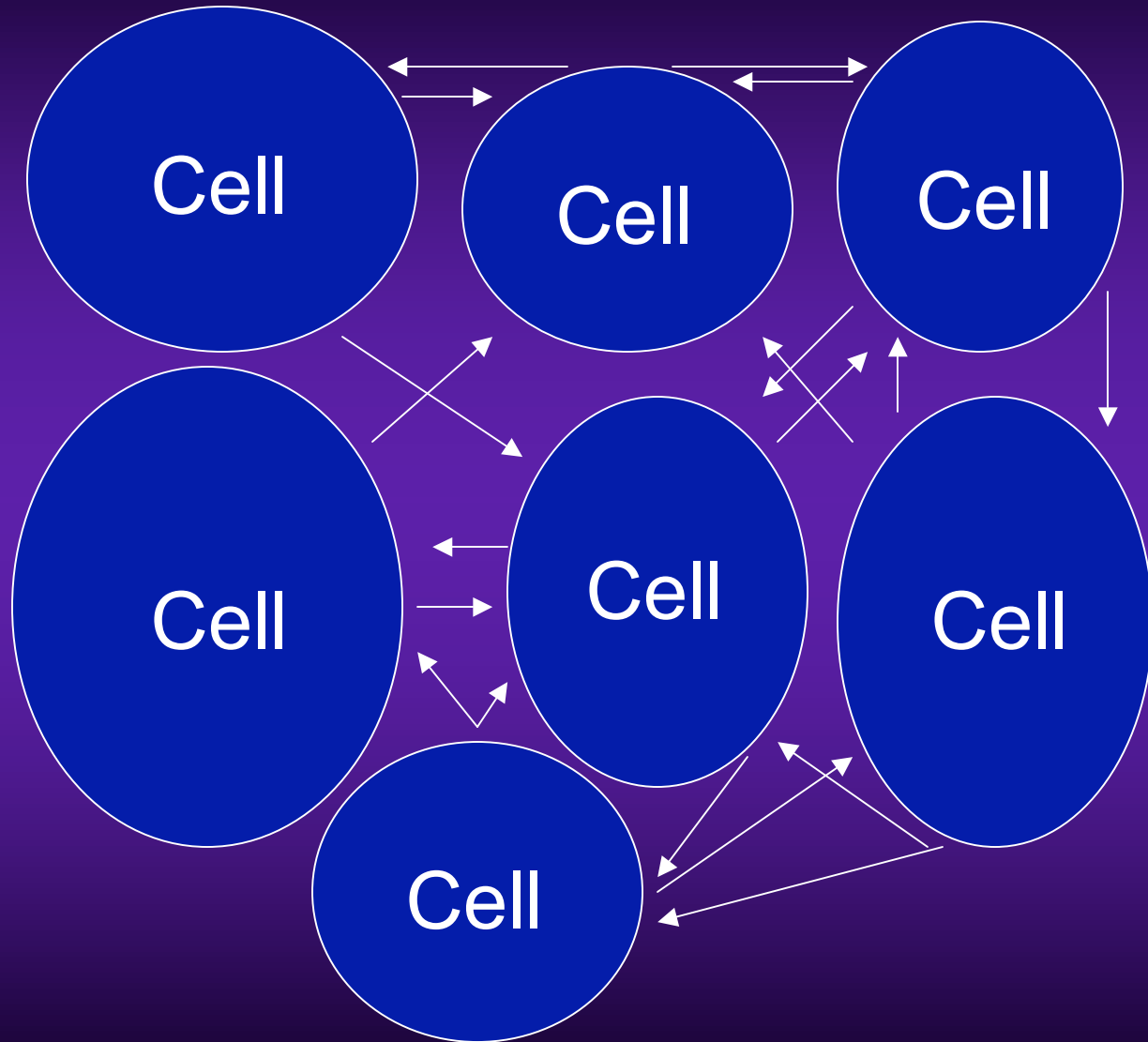
1. Department of Surgery, Northwestern University
2. School of Computer Science, Telecommunications  
and Information Systems, DePaul University

# Basic Science Paradigm

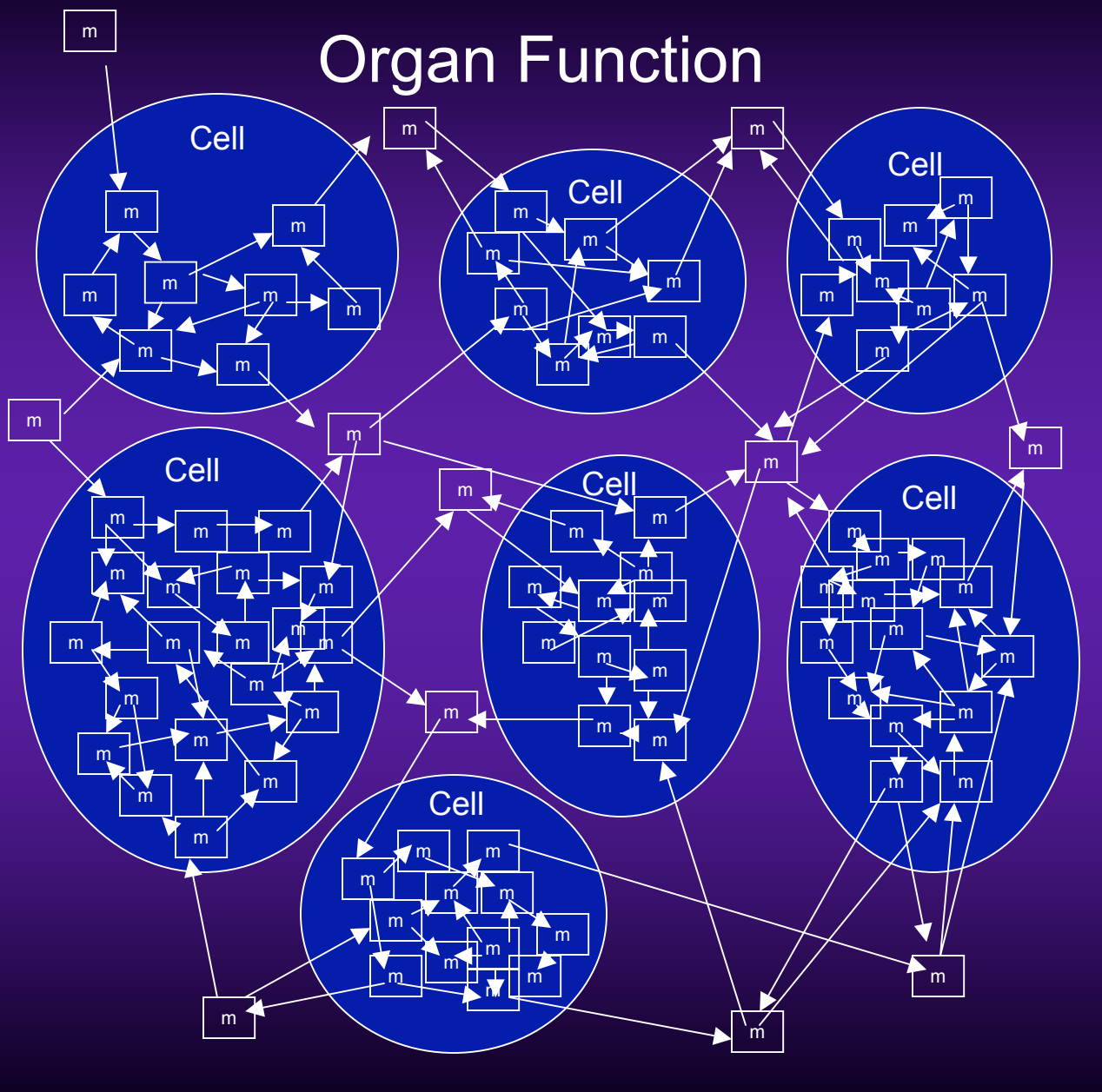
- Examines a system via reduction and isolation of its components
- “Good” experiment  $\Rightarrow$  solves for one variable  $\Rightarrow$  linear analysis
- Reconstructs system behavior by summing the results of the linear experiments

# Organ Function

# Organ Function



# Organ Function



# The Challenge: Integrating and Representing Community Knowledge

- Vast Advances in scope and scale of Biomedical Research
- “Information Overload”
  - eg. Pubmed Search for “NF-kB” > 22,000 citations
- How does it fit together?
- What is does the Community as a whole actually know?

# The Response: Utilization of Technology to Facilitate Knowledge Concatenation

- Computerized Text Analysis and Information Extraction
  - Natural Language Processing (NLP)
    - Make Biomedical Corpus more efficiently accessible
    - Extract Relationships/Hypotheses from Biomedical Corpus
    - Facilitate Formal Knowledge Representation (KR)
    - Develop “Knowledge Ecologies”

# Aspects of Information Extraction (IE)

- Two Different Hierarchies
  - Functional
    - Term Identification
    - Relationship Identification
    - Relationship Characterization
    - Knowledge/Relationship Discovery
  - Process
    - Named Entity Recognition (NER)
    - Ontology-Based IE
    - Ontology Driven IE

# This Talk...

- Two Different Hierarchies
  - Functional
    - Term Identification
    - Relationship Identification
    - Relationship Characterization
    - Knowledge/Relationship Discovery
  - Process
    - Named Entity Recognition (NER)
    - Ontology-Based IE
    - Ontology Driven IE

# Named Entity Recognition in the Biomedical Arena

- Very Challenging => More Difficult than in Economics or News Services
- Term Variation = Single Concept Multiple Terms
  - 6-7 Terms for the same thing
  - Two Experts using the same term < 20%
- Term Ambiguity = Single Term Multiple Concepts
  - Domain Specific (eg. Biochem vs Cell Bio)
  - Species Specific (eg. Mouse vs Human)
  - Acronyms: 80% are ambiguous, avg # meanings ~ 15

# Current Approaches to Biomedical NER

- Standardize Nomenclatures => Official Gene Names
  - Difficult to get community to conform
  - 1994 36% Human Genes mentioned by official name
  - 2004 43% Human Genes mentioned by official name
- Build Dictionaries
  - Generally built by hand
  - Very High Initial Investment

# Current Approaches to Biomedical NER

- State of the Art Performance
  - 80% Recall = Proportion of Extraction Info from Total Info
  - 80% Precision = Proportion of Correctly Extracted Info from Total Extracted Info
- These appeared to be capped
- Implications => Constrains the capability of more involved text analysis

# A Different Tack on Biomedical NER: An “ABM-ish” Look

- What is the system/task? => Accumulation and Updating of Biomedical terms
- What are the “agents?” => Biomedical Researchers
- What is the interaction space? => Space of community knowledge
- What defines this space? => Biomedical Corpus and Researcher’s Personal Knowledge

# A Different Tack on Biomedical NER: An “ABM-ish” Look

- Therefore: Look at the Problem in terms of agent behavior within a landscape
  - Local Information
  - Feedback between Agents and Environment
- Smart Agents or Dumb Agents? => How about both?
  - Have pretty “dumb rules,” but allow agents to use their “smartness” => Utilize Domain expertise
  - Distribute the interpretive burden at the point of Algorithmic Limitation
- **Distributed, Evolutionary Process**

# The Community-Level Approach: “Wide and Shallow” NER

- Goal = “Boot-strap” Lexicon Construction
- New Approach: Broad Sweep of range of Texts => not focus on subject of the paper but extract all the information present
- Recall and Precision not for individual texts, but for sets of texts
- Relatively simple tagging rules, but in framework that allows interaction and iteration

# Strategy for Rule-Based String Recognition NER

- Use Biomedical Formalisms to our advantage:
  - String Recognition => Types and combinations of alphanumeric characters
- Not going to worry about classification or ambiguity in rules
- Create Community User Interface for Input and Feedback
  - Rapid/Real Time assessment and modification by users

# The Named Entity Recognition (NER) Program

- Rules-based Java Program
- Identifies Cell and Molecule Names using String Characteristics
- Examples of Rules (over 3 Iterations):
  - Containing Letters and Numbers
  - Containing Hyphens
  - All Upper Case
  - Containing particular suffix, ie “-ase” or “-cyte”
  - Enclosed in Parentheses
- Input => Abstract in Plain Text Format
- Output => Terms and Rule Used to Capture

# Java Program

## Input Abstract

We have previously demonstrated that hepatic matrix metalloproteinase-9 (MMP-9) and gelatinase activity increased significantly after sepsis, and pretreatment with chemically modified tetracycline (CMT-3) inhibited these expressions and improved survivability. Activation of MMP-9 may be associated with TGF- $\beta$ 1 and Caspase-3 signaling pathways. We have been interested in investigating the role of post treatment with CMT-3 on hepatic MMP-9, TGF- $\beta$ 1 and Caspase-3 activity following sepsis. In this study, sepsis was induced in rats by cecal ligation and puncture (CLP) and 2 h later received either CMT-3

Run

## Output Terms

metalloproteinase-9	because of digits
MMP-9	because of parentheses
gelatinase	because of a suffix
sepsis	because of a suffix
CMT-3	because of parentheses
MMP-9	because of all capital letters (can include numbers)
TGF- $\beta$ 1	because of digits
Caspase-3	because of digits

# The Results

- 85 vetted Abstracts from Shock 2005
- Possible Terms (hand curated) = 916
- Number Terms Correctly Extracted = 660
- Number Terms Wrongly Extracted = 341
- **Recall** (Proportion of Correct Extracted Data to Total Desired Data) =  $660/916 = 72\%$
- **Precision** (Proportion of Correct Extracted Data to Total Extracted Data) =  $660 / (341 + 660) = 66\%$
- State-of-the-Art Rule-Based NER =  $\sim 70-80\%$   
Recall and Precision

# The Next Steps: Improve Recall and Precision

1. Incorporate self-generated lexicon database
2. Link to existing Lexical Resources
3. Iterative Development of Rules via Machine Learning
4. User Input via Interactive format
  - Prospectively Input terms as part of abstract submission process
  - Real-time User Editing
  - *Implement for Shock 2008*

# Evolution of Biomedical KR

1. Develop “verb” corpus in similar fashion
2. Semantic Parsing from whole-body text to extract relationships
3. Graphical Representation of Object relationships
4. Dynamically Evolve Ontological Structures... **Next Talk!**
5. Dynamically Instantiate Knowledge => Automated Model Development with Agent based Framework => ABM framework for Biomedical Knowledge Ecology

