

# Why Field Methods is the most important course in the HLT curriculum

Kevin Cohen

Center for Computational Pharmacology  
University of Colorado School of Medicine

# Suzanna Lewis

- Fruitfly geneticist
- 5 kids
- Latte + 3 shots

# Suzanna Lewis

It is the middle of the night (2:38 to be precise), I am away from friends and family, It has been this way for over 2 years, I can't sleep because of all the work there is yet to do, and there is no end in sight. So **when do the magic little elves appear out of nowhere and get everything done?**

p.s. I am serious.

Suzanna Lewis

pray for elves

*D. melanogaster* gene Pray For Elves, abbreviated as PFE, is reported here. It has also been known in FlyBase as CG15151. Similar sequences have been identified in *Caenorhabditis elegans*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* and *Saccharomyces cerevisiae*.

(FlyBase report FBal0138651)

*D. melanogaster* gene **Pray For Elves**, abbreviated as **PFE**, is reported here. It has also been known in FlyBase as **CG15151**. Similar sequences have been identified in *Caenorhabditis elegans*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* and *Saccharomyces cerevisiae*.

(FlyBase report FBal0138651)

# Center for Computational Pharmacology

- "Wet lab" scientists
  - molecular biologists
  - pharmacologists
  - geneticists
- Computer scientists
  - artificial intelligence
  - mathematicians
  - linguists

- Central problem in biology today:

"drinking from a firehose"

# Center for Computational Pharmacology

Build a knowledge-base  
for interpretation of  
high-throughput assays

# NLP Projects

- Text data mining
  - information extraction
  - entity identification
  - information retrieval

# Information retrieval

- "Classic" information retrieval problem:
  - X bazillion documents (web pages)
  - Y of them are relevant,  $Y \lll X$
  - find them

"Needle in a haystack"

# Information retrieval

- Molecular biology problem:
  - There are hundreds of relevant papers, and I can find them easily
  - Please sort this pile...

Haystack is made of needles....

- Cluster by areas of interest to subject matter experts
- For a linguist:
  - Phonetics
  - Phonology
  - Morphology
  - Syntax

- Cluster by areas of interest to subject matter experts
- For molecular biology of substance abuse:
  - Cell cycle
  - Behavior
  - Drug interactions
  - Pathways

# Information retrieval

- Why I care about information retrieval even though I'm not interested in it: false positives for entity identification

# Entity identification

- Classic entity identification problem:
  - list of classes
    - person
    - organization
    - location
    - date
    - monetary amount
  - unstructured text
- Find the classes in the text

# Entity identification

John Denver will be playing at the  
Denver Buffalo Company in Denver at  
10 p.m. on Friday.

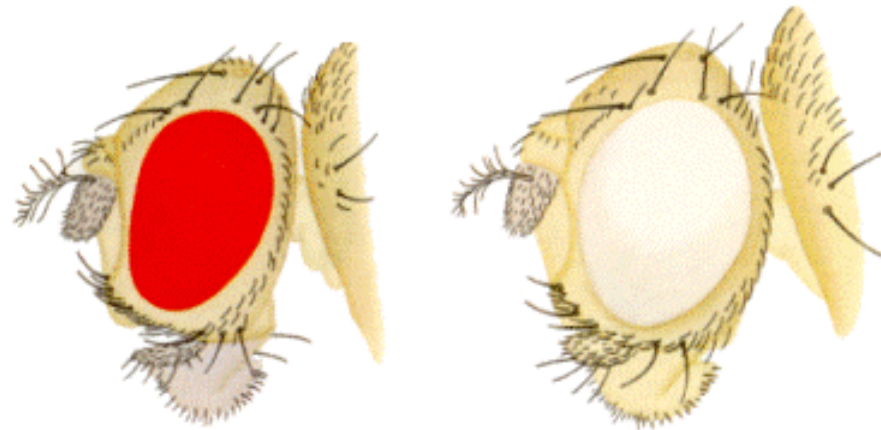
# Entity identification

<PERSON>John Denver</PERSON>  
will be playing at the  
<ORG>Denver Buffalo Company</ORG>  
in <LOCATION>Denver</LOCATION>  
at <TIME>10 p.m. on Friday</TIME>.

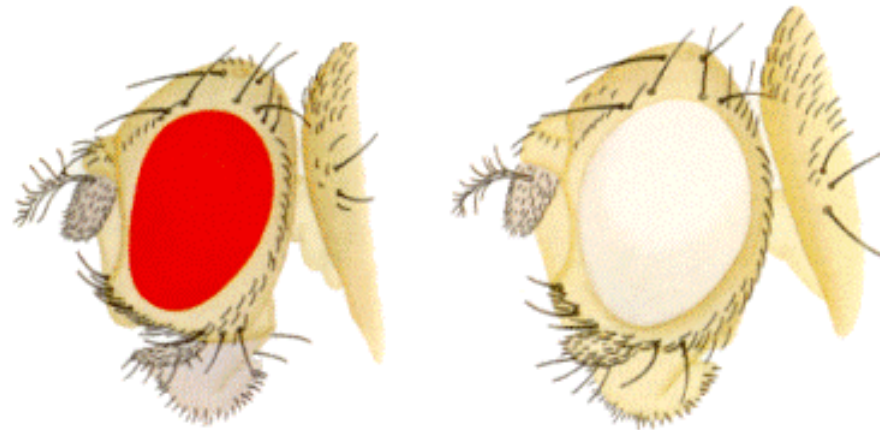
# Entity identification

- Molecular biology entity identification problem:
  - small list of classes
  - much harder
    - Usual case-related cues don't help
    - More variability of content
    - Huge lexical ambiguity problem
    - Common English
  - as posed, not useful

white

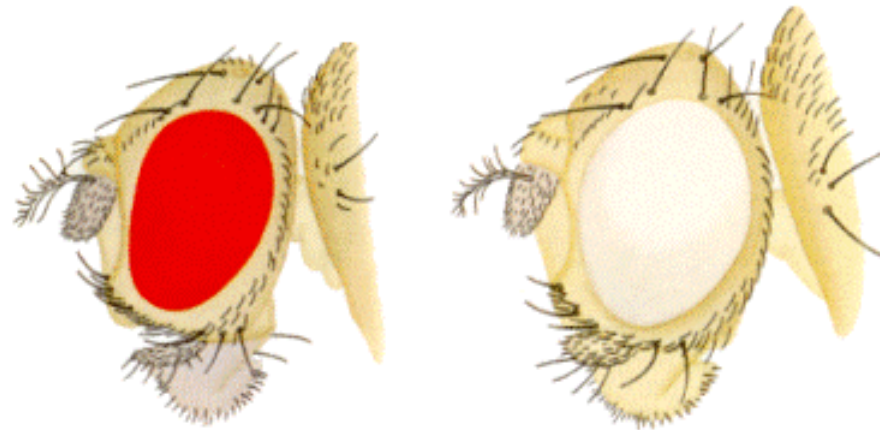


# white



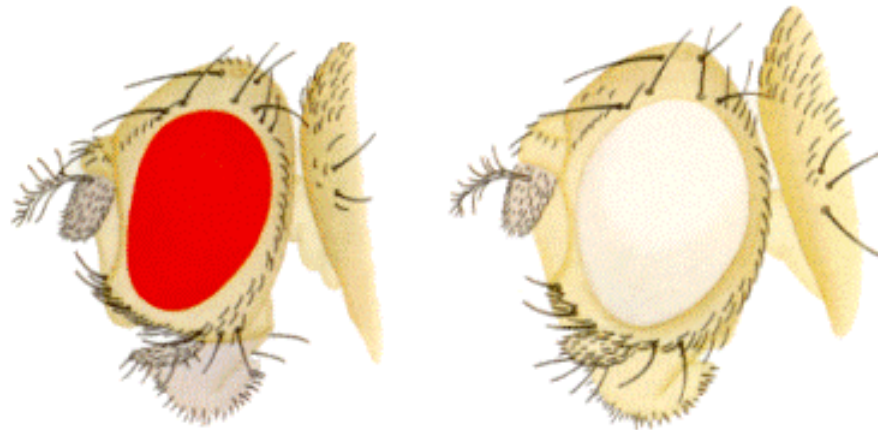
"wild-type" (not mutated)

white



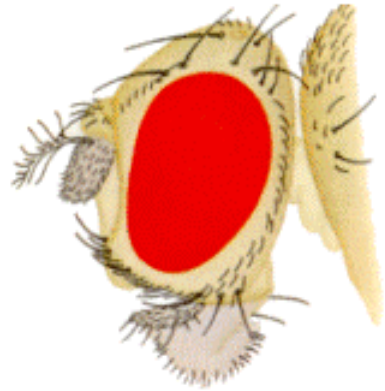
"mutant"

white



white

# Case is meaningful

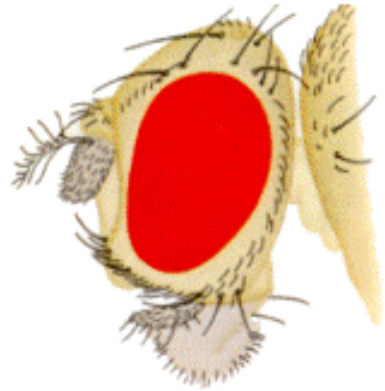


White



white

# Case is meaningful



White

Symbol: W



white

Symbol: w

# Case is meaningful

Misshapen (Msn) has been proposed to shut down *Drosophila* photoreceptor (R cell) growth cone motility in response to targeting signals linked by the SH2/SH3 adaptor protein Dock.

# Case is meaningful

**Misshapen** (Msn) has been proposed to shut down *Drosophila* photoreceptor (R cell) growth cone motility in response to targeting signals linked by the SH2/SH3 adaptor protein **Dock**. (Ruan et al. 2002)

...even sentence-initially.

**sunday driver** (*syd*) was identified in a screen for novel axonal transport mutants in *Drosophila*. **Syd** is a ~137kDa protein that is broadly conserved in evolution with homologous proteins identified in *C. elegans*, mouse and human. (Bowman 2000)

# Case is meaningful

**Misshapen** (Msn) has been proposed to shut down **Drosophila photoreceptor** (R cell) growth cone motility in response to targeting signals linked by the SH2/SH3 adaptor protein **Dock**. Here, we show that Bifocal (Bif), a putative cytoskeletal regulator, is a component of the Msn pathway for regulating R cell growth targeting. bif displays strong genetic interaction with msn.

Surely you could determine  
on a document-by-document  
basis...

**Misshapen** (Msn) has been proposed to shut down **Drosophila photoreceptor** (R cell) growth cone motility in response to targeting signals linked by the SH2/SH3 adaptor protein **Dock**. Here, we show that **Bifocal** (Bif), a putative cytoskeletal regulator, is a component of the **Msn** pathway for regulating R cell growth targeting. **bif** displays strong genetic interaction with **msn**.

Surely you could determine  
on a document-by-document  
basis...

Axonal traffic jams with a **sunday driver**:  
Identification of a broadly conserved  
transmembrane protein required for  
axonal transport in *Drosophila*.  
(Bowman 2000)

# Evolution

- What it looks like
- What it acts like
- Metaphor
- ...

# Looks like...

- white
- swiss cheese
- clown
- daschund
- dreadlocks

# Acts like...

- ether a go-go
- lush
- agnostic
- amontillado

# Metaphor/metonymy

- lot
- maggie
- scott of the antarctic
- always early -> british rail
- asp -> cleopatra
- tudor -> vasa -> gustavus
- nanos -> smaug

# whimsy

- chablis, merlot, zinfandel, retsina, moonshine (16 zebrafish genes)
- milkah, murashka, zolotistyuy, zloday (32 Drosophila genes)

- fuculokinase
- GABA
- Heat shock protein 60
- calmodulin
- dHAND
- suppressor of p53

- cheap date
- lush
- ken and barbie
- ring
- to
- the
- there
- a

# Worst gene names

- sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A

# Worst gene names

- sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A

# Worst gene names

- sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A
- SEMA5A

# Worst gene names

- sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A
- SEMA5A
- Tyrosine kinase with immunoglobulin and epidermal growth factor homology domains
- tie

So: what to do?

# EI as POS tagging

- Retinoic acid downmodulates erythroid differentiation and *GATA1* expression in purified adult-progenitor culture.  
(Labbaye et al. 1994)

- Retinoic/**JJ** acid/**NN**  
downmodulates/**VBZ** erythroid/**JJ**  
differentiation/**NN** and/**CC**  
*GATA1*/**NN** expression/**NN** in/**IN**  
purified/**VBN** adult-progenitor/**JJ**  
culture/**NN** ./ . (Labbaye et al. 1994)

- Retinoic/**JJ** acid/**NN**  
downmodulates/**VBZ** erythroid/**JJ**  
differentiation/**NN** and/**CC**  
*GATA1*/**GENE** expression/**NN** in/**IN**  
purified/**VBN** adult-progenitor/**JJ**  
culture/**NN** ./ . (Labbaye et al. 1994)



# Entity identification

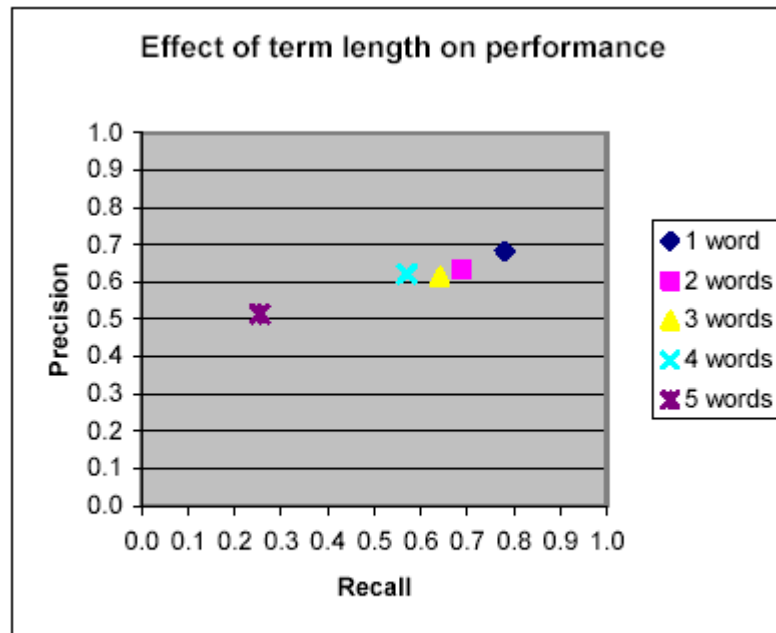


Figure 2: Effect of term length on performance.

# Location vs. identification

- Rat somatotropin
- Somatotropin
- Growth hormone

- Somatotropin is upregulated by X
- Transcription of rat somatotropin is blocked by Y
- Growth hormone is expressed by cells of type Z

- Somatotropin
  - Upregulated by: X
- Rat somatotropin
  - Transcription blocked by: Y
- Growth hormone
  - Expressed by cell: Z

- Somatotropin
  - Upregulated by: X
- Rat somatotropin
  - Transcription blocked by: Y
- Growth hormone
  - Expressed by cell: Z

- Somatotropin
  - Upregulated by: X
  - Transcription blocked by: Y
  - Expressed by cell: Z

- HSP-60
- HSP60
- Hsp-60
- Hsp 60
- Hsp (rat) 60

- Krauthammer et al. (2000): heuristics might be useful

Contrast vs. Variability

- Paradigmatic
  - Case
  - Hyphenation
  - Parenthesization
  - Vowel sequences
- Syntagmatic
  - Left vs. right edge
  - Character vs. word

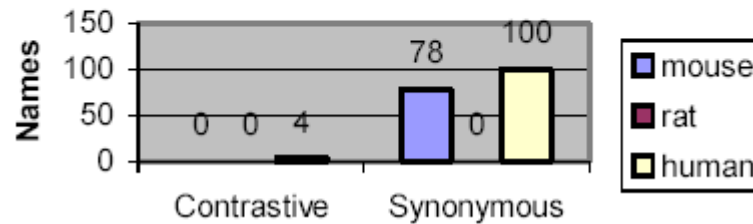
# Entity identification

- Gene Name 1
  - Gene Name 2
  - Gene Name 3
  - Gene Name 4
  - Gene Name 6
  - ...
- Gene 1
  - Gene 1 synonym a
  - Gene 1 synonym b
  - Gene 1 synonym c
  - Gene 2
  - Gene 2 synonym a
  - Gene 2 synonym b
  - Gene 2 synonym c

- Map each name to a reduced form
  - HSP 60 -> HSP X
  - HSP 72 -> HSP X
  - GAB A -> GAB A
- For each reduced form, keep list of full forms that map to it
  - HSP X
    - HSP 60
    - HSP 72

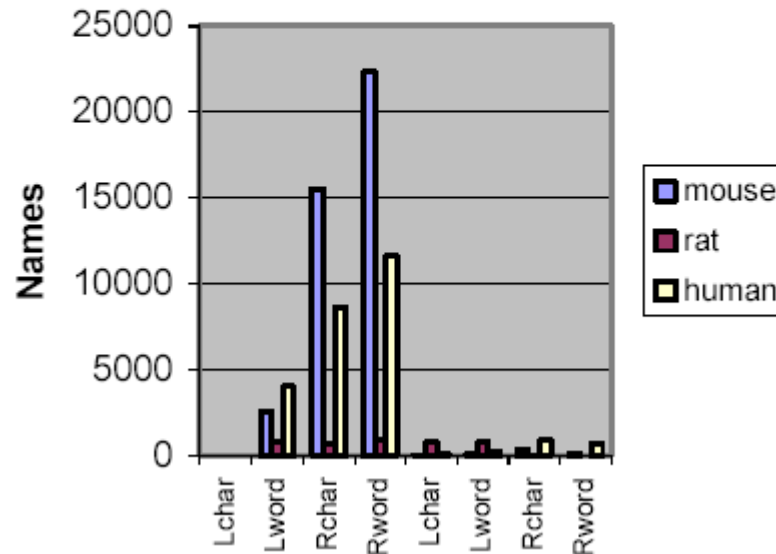
# Entity identification

Graph 1. Hyphenation: contrast and variability



# Entity identification

Graph 4. Edge effects:  
contrastive on left,  
synonymous on right



# Information extraction

- Classic IE problems:
  - terrorist attacks
  - who, how, where, when, how many injured
  - corporate acquisitions
  - corporate succession

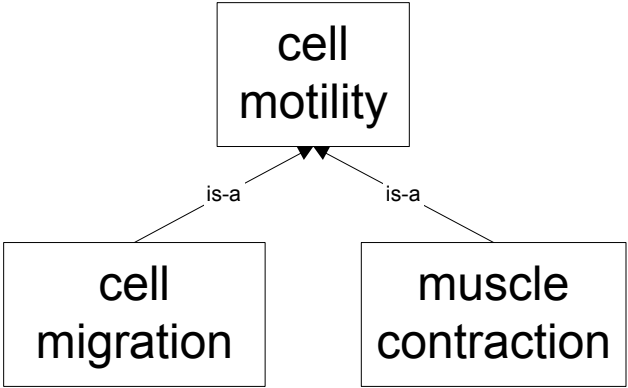
# Information extraction

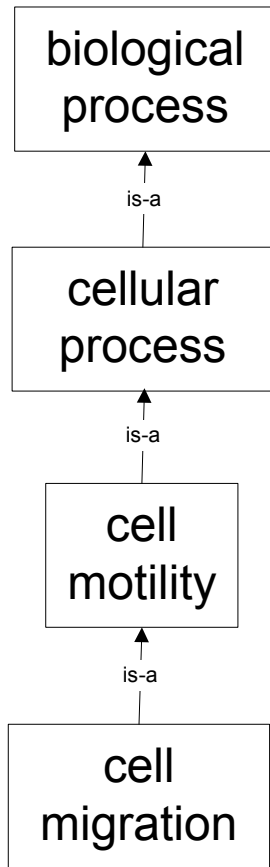
- "Interactions"
- disease association
- subcellular localization
- function

- Interactions

- species
- phenotype
- experimental condition
- cell type
- subcellular localization
- temporal

- Problem: knowledge is expensive
- Solution: publicly available knowledge is now sufficient
- Observation: argument to predicate is often a *Gene Ontology* term
- Problem: variability
- Solution: Ogren et al. 2004





cell  
migration

has-part



regulation of cell  
migration

- membrane
  - inner membrane
    - mitochondrial inner membrane

- membrane
  - inner membrane
    - mitochondrial inner membrane

- Look for strings that correlate with consistent meaning
  - Meaning = relation in the ontology

- 65% of terms
- For a given relation, 47-66% of "derivational" strings are specific to it

- Negative regulation of X
  - Negative regulation of REM sleep
  - 24 new terms
- Limonene X
  - Is-a X
  - Limonene monooxygenase activity

- Enriched conceptual representation
  - Regulation direction
  - Process type
  - Base type
  - Oxygen availability
  - Chirality
  - ...
- Variation in surface forms of terms

These findings suggest that FAK functions in the regulation of cell migration and cell proliferation. (Gilmore and Romer 1996)

Almost:

Regulation of cell migration

Cell proliferation

These findings suggest that FAK functions in the regulation of cell migration and cell proliferation. (Gilmore and Romer 1996)

Right:

Regulation of cell migration

Regulation of cell proliferation

- Regulation of cell proliferation
- Regulation of (TERM and) cell proliferation

What do these have in  
common?

discovery procedure

# How do you know if you're doing a good job or not?

- Precision:  $TP / (TP + FP)$

**bif** displays **strong** genetic interaction with **msn**.

$$TP = 1$$

$$FP = 1$$

$$P = 1/(1+1) = .5$$

# How do you know if you're doing a good job or not?

- Recall:  $TP / (TP + FN)$

**bif** displays strong genetic interaction with **msn**.

$$TP = 1$$

$$FN = 1$$

$$P = 1/(1+1) = .5$$

- *Good:*
  - Community standard
- *Bad:*
  - What do I need to fix??
  - Failure to recognize that this is software
  - Failure to recognize that this is language

- Where did I go wrong? And what am I good at?
  - Software engineering answer: structured testing.

- "Catalogue" of test conditions
  - Zero
  - Non-zero
  - Real
  - Integer
  - Positive
  - Negative
  - Unsigned
  - Smallest number representable
  - One less than ...
  - Largest number representable
  - One more than...

- What are the right test conditions for an entity identification system?
  - Ask a linguist.
- Syntagmatic: environments where gene names can appear
- Paradigmatic: types of gene names
- False positives

- **Environments**

- Sentence-positional

- List position

- Single

- Simple coordination

- Asyndetic coordination

- Typographic

- Parentheses

- Punctuation attached

- POS

- Preceded by adjective

- Preceded by numeral

- Elements of compound NP (X gene, X protein)

- Length
- In vocabulary/OOV
- Name vs. symbol
- Common English
- Orthographic features
- Morphological features
  - Inflectional (N number, genitive; V -ed, -ing)
  - Derivational

- Tagger 1: fails when gene names are only one word long.
- Tagger 2: fails when gene symbols are lower-case initial and sentence-initial.
- Tagger 3: fails on alphanumeric modifiers when:
  - Modifier is numeric
  - Modifier is at rightmost edge
  - Modifies a name, not a symbol

- Tagger 1: fails when gene names are only one word long.
- Tagger 2: fails when gene symbols are lower-case initial and sentence-initial.
- Tagger 3: fails on alphanumeric modifiers when:
  - Modifier is numeric
  - Modifier is at rightmost edge
  - Modifies a name, not a symbol

- George Acquah-Mensah
- Andy Dolbey
- Jens Eberlein
- Philip Ogren

# Two kinds of linguists who should take Field Methods

- Theoreticians
- Non-theoreticians

The most insulting thing  
ever said to me by a linguist

It's a pity that there are  
so many people trying to  
do theoretical work who  
could be doing good  
descriptive work.