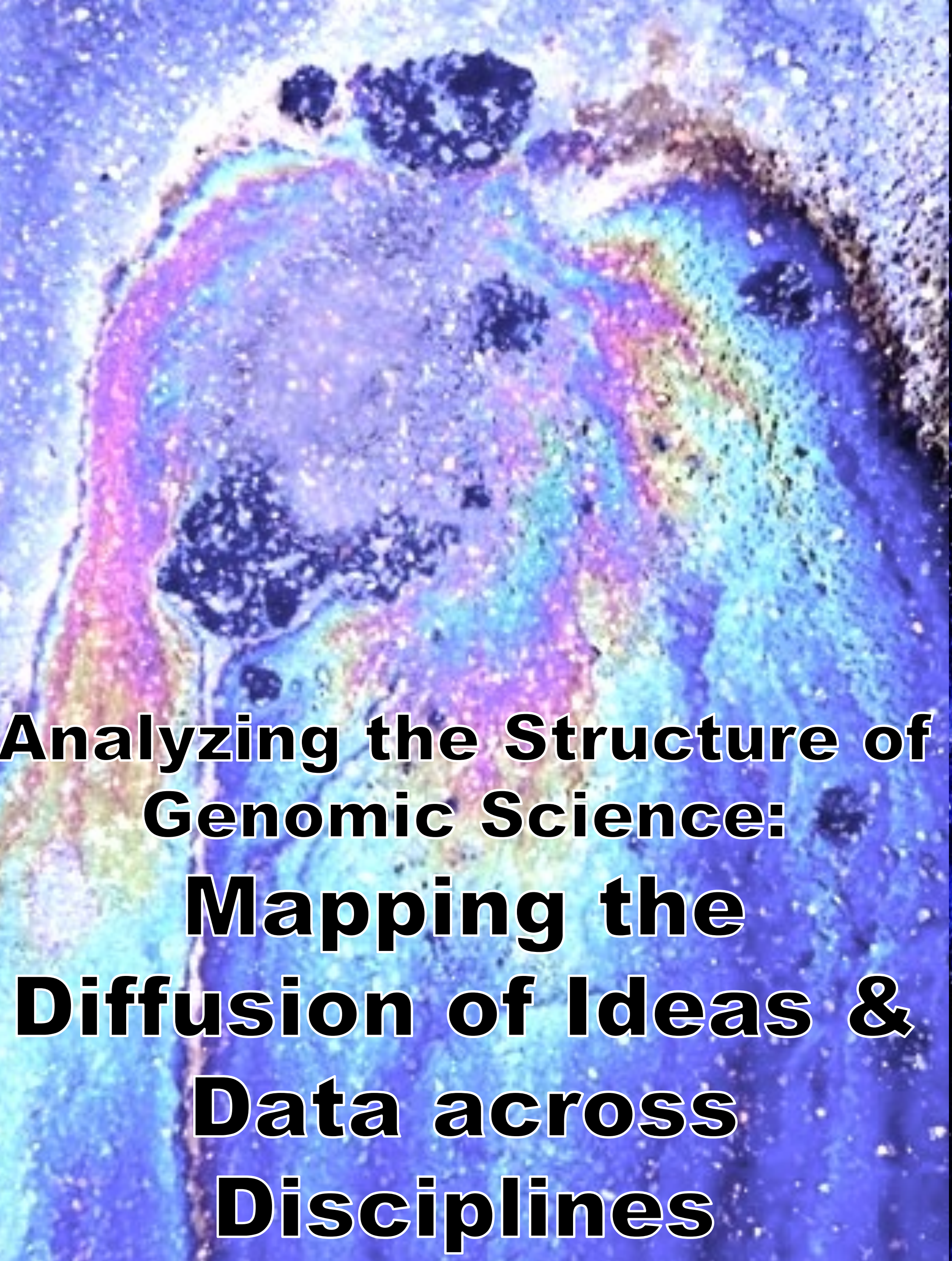
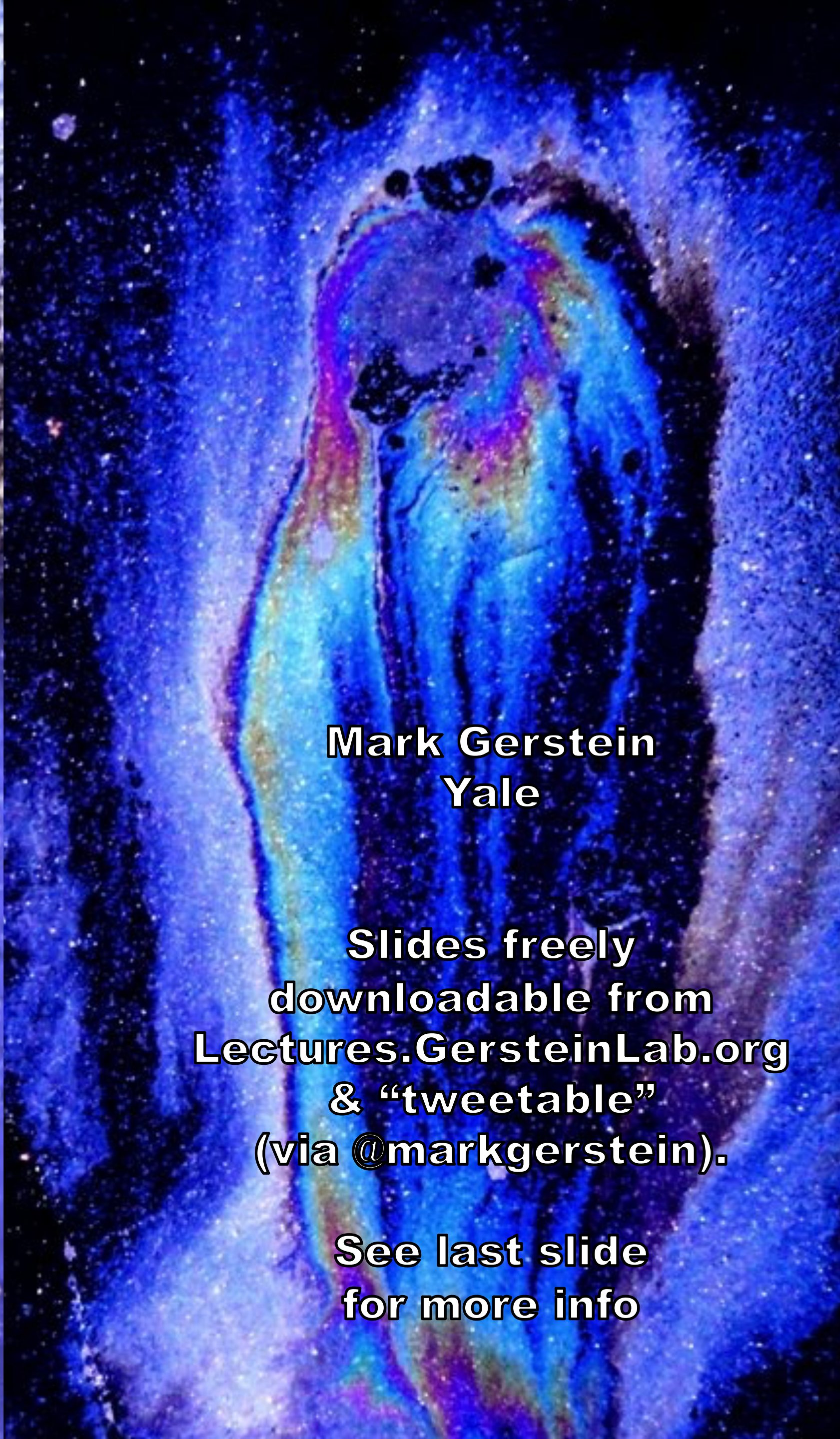


# Analyzing the Structure of Genomic Science: Mapping the Diffusion of Ideas & Data across Disciplines

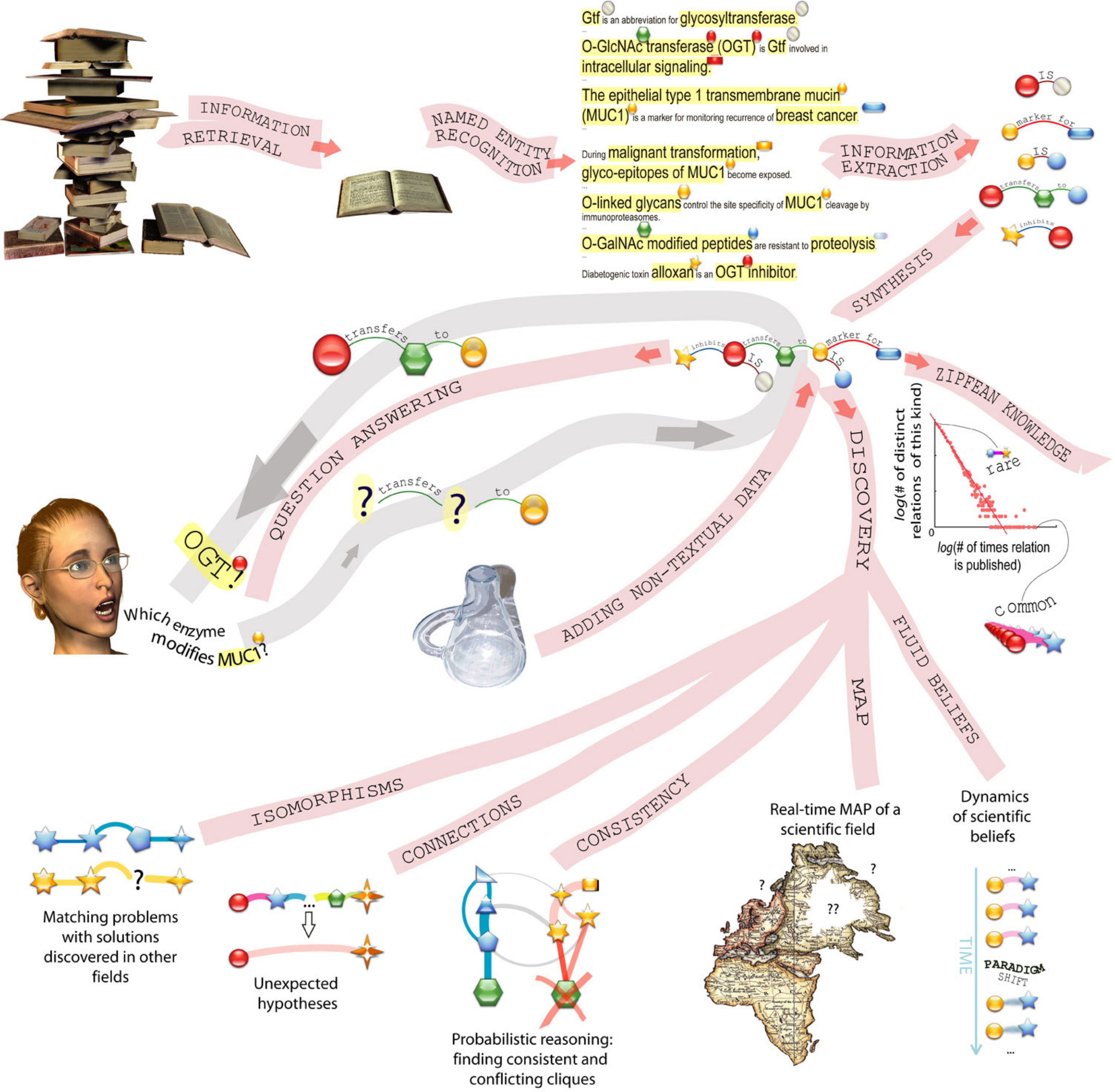


Mark Gerstein
Yale

Slides freely downloadable from Lectures.GersteinLab.org & “tweetable” (via @markgerstein).

See last slide for more info

# Vision for Mining Large-scale Structured Literature



Gtf is an abbreviation for glycosyltransferase

O-GlcNAc transferase (OGT) is Gtf involved in intracellular signaling.

The epithelial type 1 transmembrane mucin (MUC1) is a marker for monitoring recurrence of breast cancer.

During malignant transformation, glyco-epitopes of MUC1 become exposed.

O-linked glycans control the site specificity of MUC1 cleavage by immunoproteasomes.

O-GalNAc modified peptides are resistant to proteolysis.

Diabetogenic toxin alloxan is an OGT inhibitor.

Which enzyme modifies MUC1?

Matching problems with solutions discovered in other fields

Unexpected hypotheses

Probabilistic reasoning: finding consistent and conflicting cliques

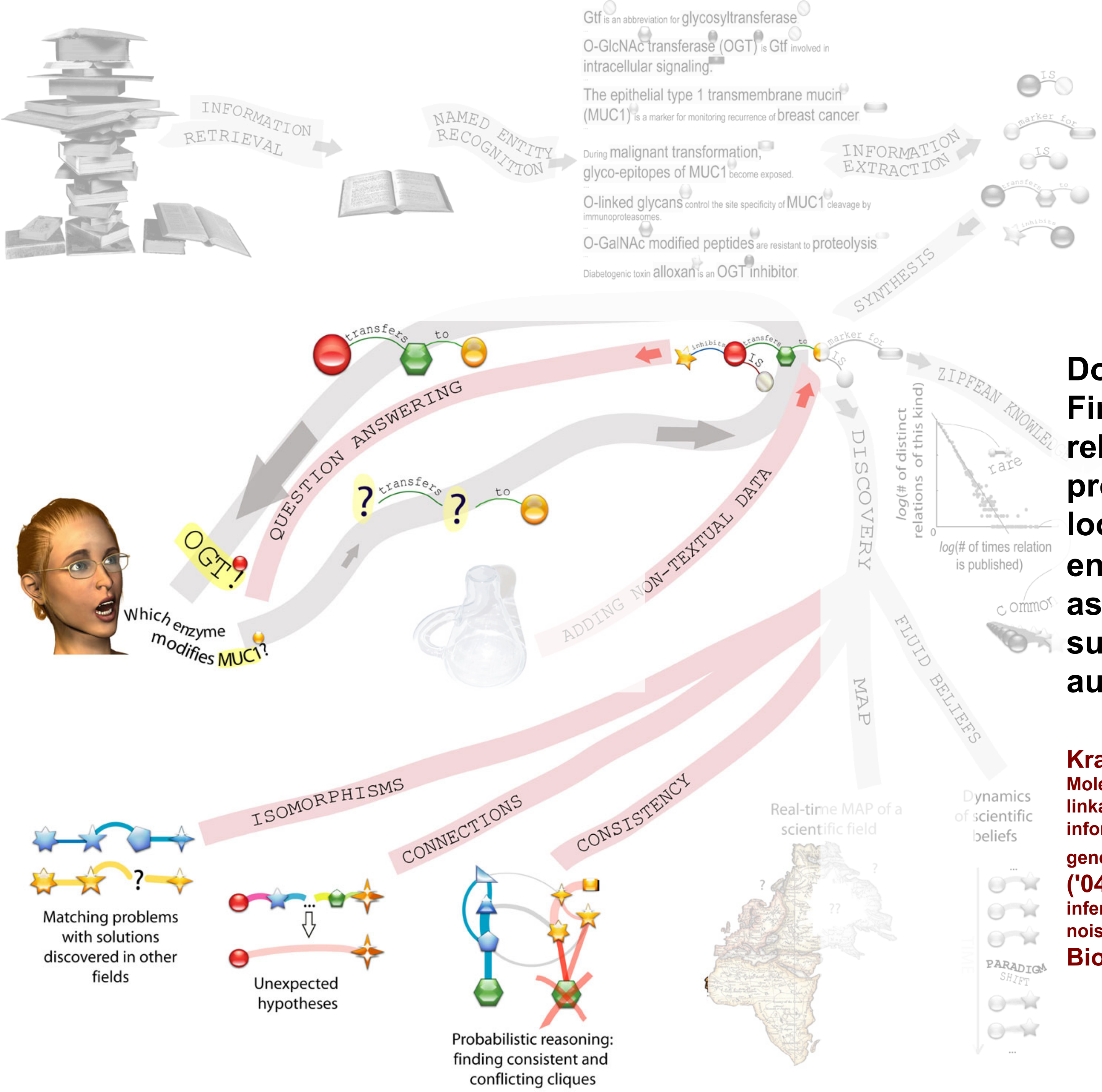
Real-time MAP of a scientific field

Dynamics of scientific beliefs

Harnessing the “Data Exhaust” from large-scale efforts

[Rzhetsky et al, Cell ('08), PLOS CB ('09); Bourne et al. PLOS CB '08]

# Vision for Mining Large-scale Structured Literature

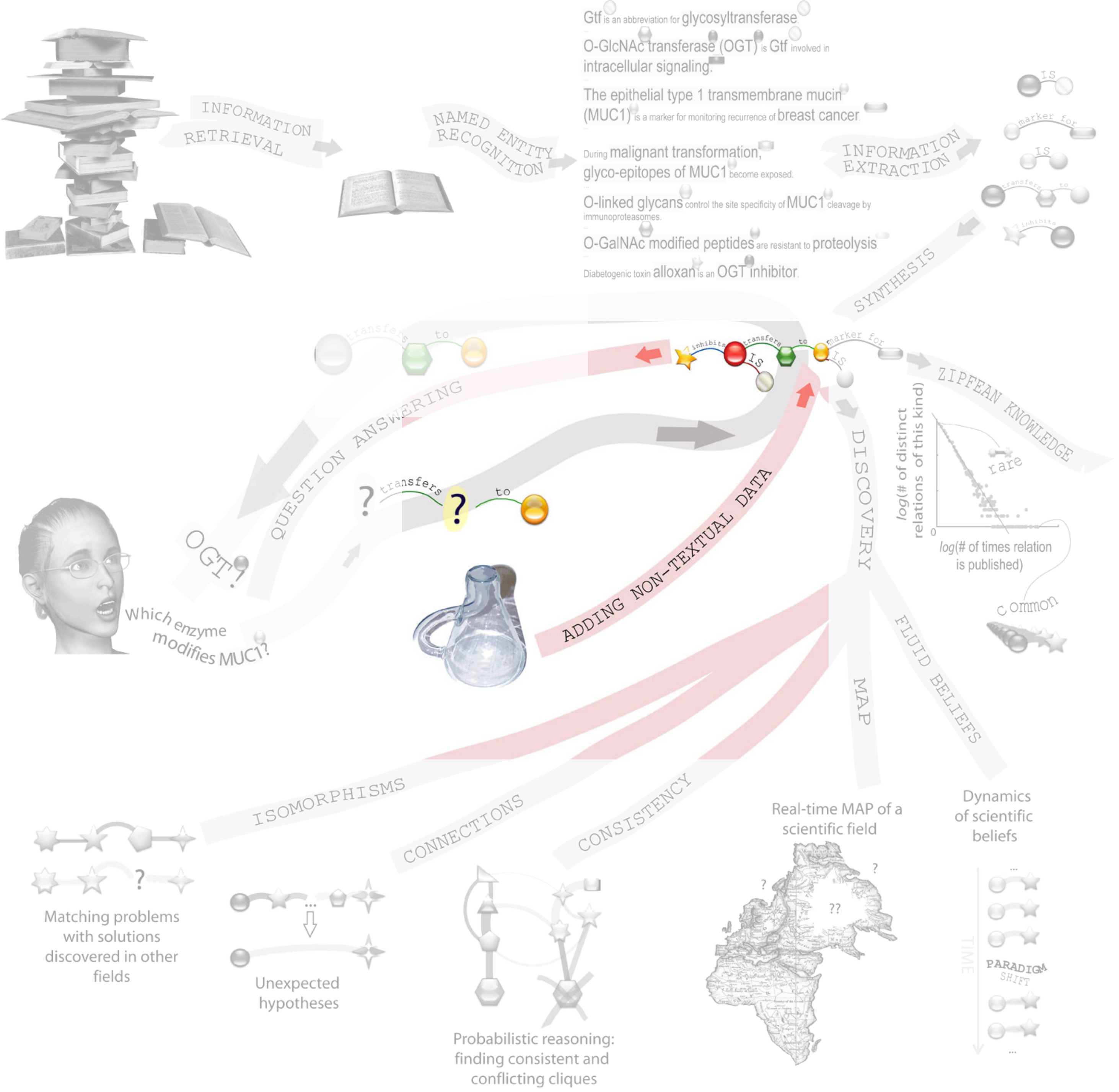

INFORMATION RETRIEVAL
NAMED ENTITY RECOGNITION
Gtf is an abbreviation for glycosyltransferase
O-GlcNAc transferase (OGT) is Gtf involved in intracellular signaling.
The epithelial type 1 transmembrane mucin (MUC1) is a marker for monitoring recurrence of breast cancer
During malignant transformation, glyco-epitopes of MUC1 become exposed.
O-linked glycans control the site specificity of MUC1 cleavage by immunoproteasomes.
O-GalNAc modified peptides are resistant to proteolysis
Diabetogenic toxin alloxan is an OGT inhibitor
INFORMATION EXTRACTION
IS
marker for
IS
transfers to
inhibits
SYNTHESIS
transfers to
inhibits transfers to
marker for
IS
ZIPFEAN KNOWLED
log(# of distinct relations of this kind)
0
rare
log(# of times relation is published)
common
QUESTION ANSWERING
? transfers ? to
OGT?
Which enzyme modifies MUC1?
ADDING NON-TEXTUAL DATA
DISCOVERY
MAP
FLUID BELIEFS
ISOMORPHISMS
CONNECTIONS
CONSISTENCY
?
Matching problems with solutions discovered in other fields
...
Unexpected hypotheses
Probabilistic reasoning: finding consistent and conflicting cliques
Real-time MAP of a scientific field
?
??
?
Dynamics of scientific beliefs
...
TIME
PARADIGM SHIFT
...

**Doing better science: Finding new protein relationships (e.g. protein interactions), looking for inconsistencies in arguments, assembling consensus definitions automatically**

**Krauthammer et al. Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. PNAS ('04); Iossifov et al. Probabilistic inference of molecular networks from noisy data sources. Bioinformatics ('04)**

[Rzhetsky et al, Cell ('08), PLOS CB ('09); Bourne et al. PLOS CB '08]

# Vision for Mining Large-scale Structured Literature



Gtf is an abbreviation for glycosyltransferase

O-GlcNAc transferase (OGT) is Gtf involved in intracellular signaling.

The epithelial type 1 transmembrane mucin (MUC1) is a marker for monitoring recurrence of breast cancer.

During malignant transformation, glyco-epitopes of MUC1 become exposed.

O-linked glycans control the site specificity of MUC1 cleavage by immunoproteasomes.

O-GalNAc modified peptides are resistant to proteolysis.

Diabetogenic toxin alloxan is an OGT inhibitor.

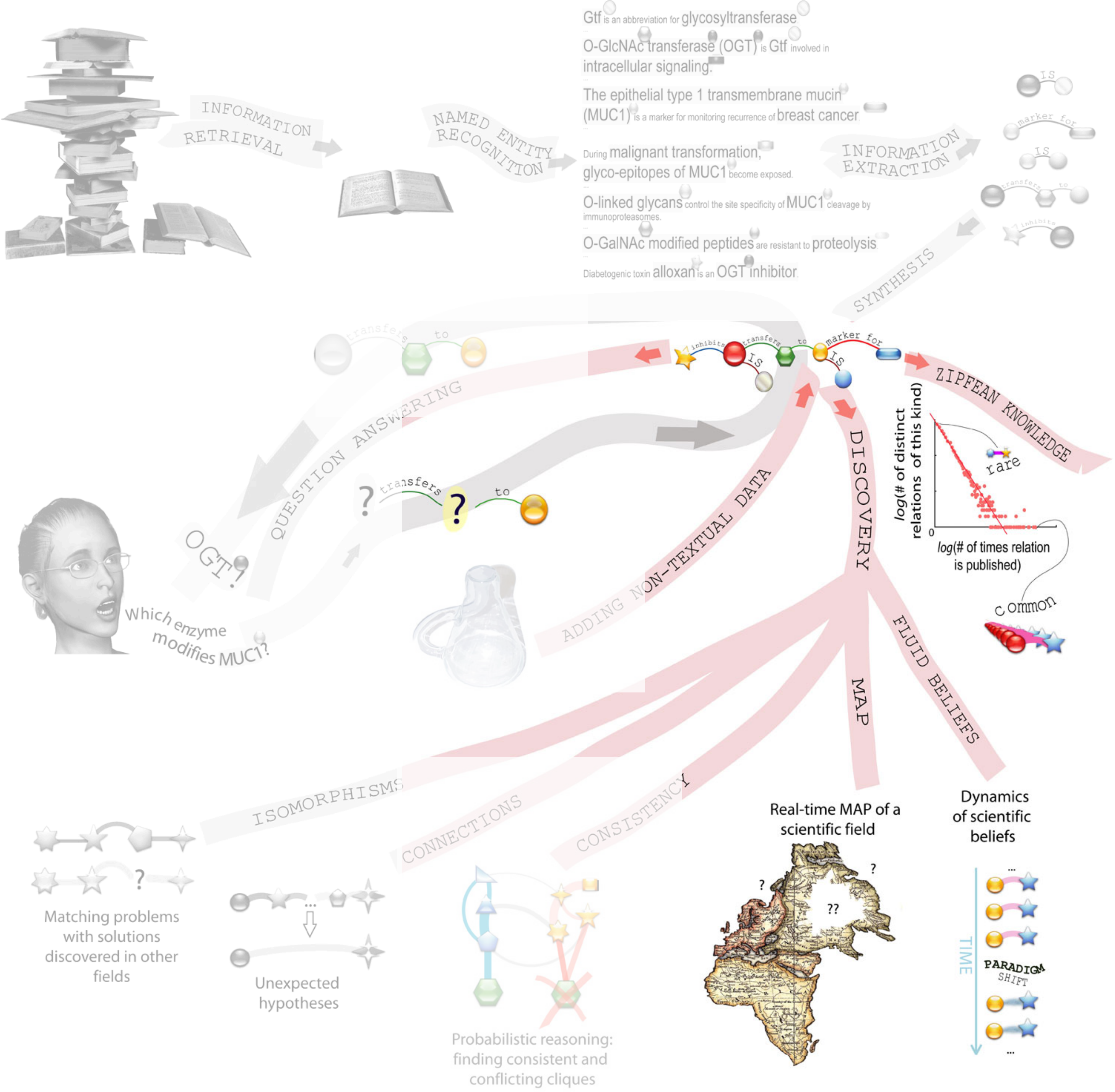
Which enzyme modifies MUC1?

**Making it understand-able (through “mashup”)**

**SciVee, podcasts**

[Rzhetsky et al, Cell ('08), PLOS CB ('09); Bourne et al. PLOS CB '08]

# Vision for Mining Large-scale Structured Literature



Gtf is an abbreviation for glycosyltransferase

O-GlcNAc transferase (OGT) is Gtf involved in intracellular signaling.

The epithelial type 1 transmembrane mucin (MUC1) is a marker for monitoring recurrence of breast cancer

During malignant transformation, glyco-epitopes of MUC1 become exposed.

O-linked glycans control the site specificity of MUC1 cleavage by immunoproteasomes.

O-GalNAc modified peptides are resistant to proteolysis

Diabetogenic toxin alloxan is an OGT inhibitor

Matching problems with solutions discovered in other fields

Unexpected hypotheses

Probabilistic reasoning: finding consistent and conflicting cliques

Real-time MAP of a scientific field

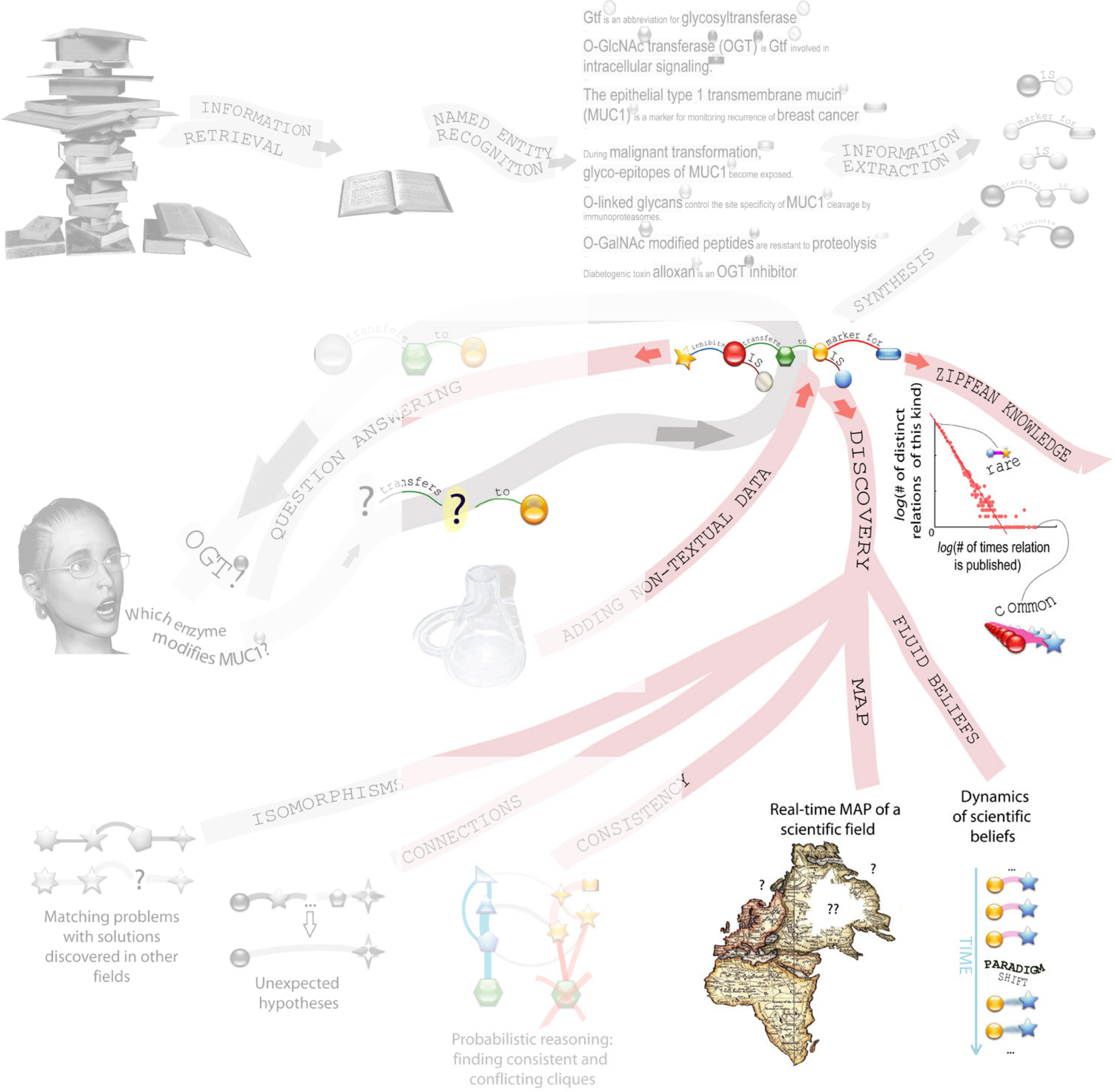
Dynamics of scientific beliefs

**Mapping Science**

**+**

**Studying its Dynamics & Evolution**

[Rzhetsky et al, Cell ('08), PLOS CB ('09); Bourne et al. PLOS CB '08]



Gtf is an abbreviation for glycosyltransferase

O-GlcNAc transferase (OGT) is Gtf involved in intracellular signaling.

The epithelial type 1 transmembrane mucin (MUC1) is a marker for monitoring recurrence of breast cancer

During malignant transformation, glyco-epitopes of MUC1 become exposed.

O-linked glycans control the site specificity of MUC1 cleavage by immunoproteasomes.

O-GalNAc modified peptides are resistant to proteolysis

Diabetogenic toxin alloxan is an OGT inhibitor

Which enzyme modifies MUC1?

Matching problems with solutions discovered in other fields

Unexpected hypotheses

Probabilistic reasoning: finding consistent and conflicting cliques

Real-time MAP of a scientific field

Dynamics of scientific beliefs

# Vision for Mining Large-scale Structured Literature

- Revealing patterns of collaboration
- Understanding basis of terms & nomenclature
- Tracking the evolution of ideas
- Models for the evolution of science;
- Helping set policy & research directions

[Rzhetsky et al, Cell ('08), PLOS CB ('09); Bourne et al. PLOS CB '08]

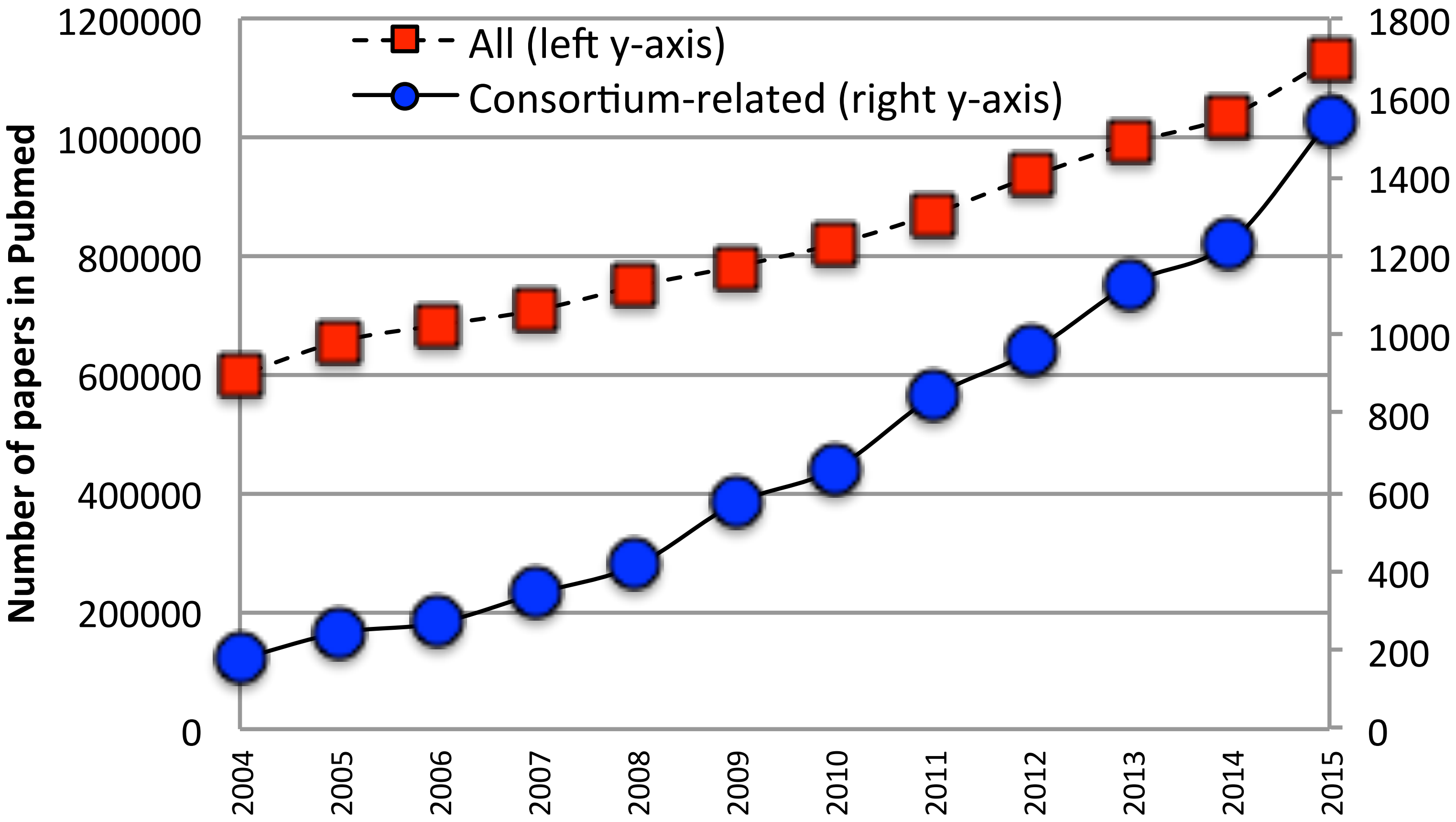
## Analyzing the Structure of Genomic Science: Mapping the diffusion of ideas & data across disciplines

- Intro: Using the Data Exhaust from Consortia
- PubNet Tool for analyzing co-publication patterns
  - Different patterns for Str. Genomics centers
- Analysis of Evolution of the ENCODE Consortium
  - Differences in “modularity” for members & users
  - Key role for brokers
- Other Examples of Idea & Data Diffusion
  - RNAi & SRA bases
- Quant. Model of Information Diffusion
  - Based on PLOS ALM
  - Random multiplicative process
  - Moderated by parm. w/ 2 time regimes

# Analyzing the Structure of Genomic Science: Mapping the diffusion of ideas & data across disciplines

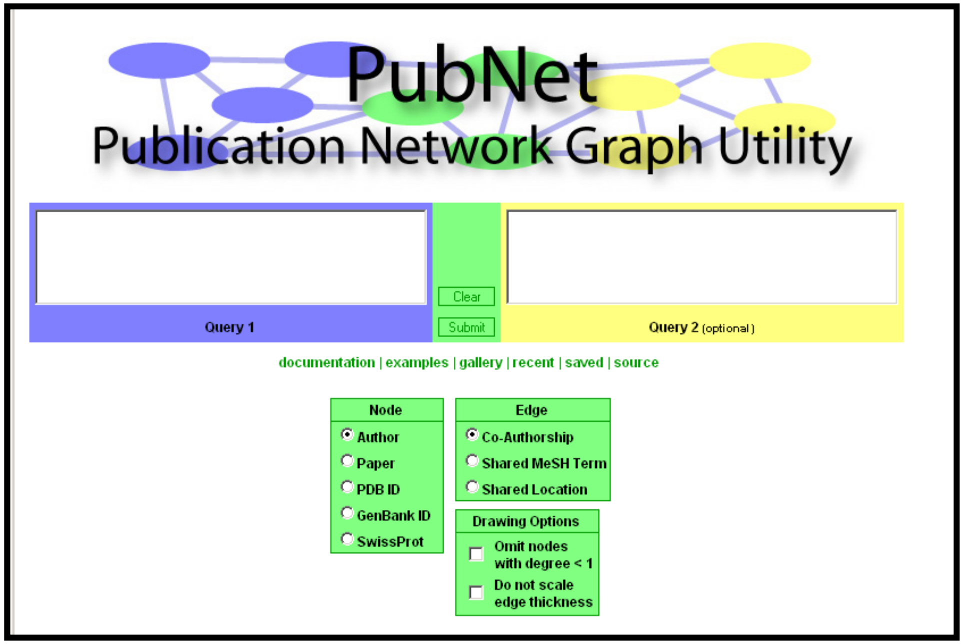
- Intro: Using the Data Exhaust from Consortia
- PubNet Tool for analyzing co-publication patterns
  - Different patterns for Str. Genomics centers
- Analysis of Evolution of the ENCODE Consortium
  - Differences in "modularity" for members & users
  - Key role for brokers
- Other Examples of Idea & Data Diffusion
  - RNAi & SRA bases
- Quant. Model of Information Diffusion
  - Based on PLOS ALM
  - Random multiplicative process
  - Moderated by parm. w/ 2 time regimes

# Increase in Consortium Science



## Examples Illuminating Current State of Affairs:

# Using Network Representations to Make Maps of Science -- Studying the Publication Patterns of Genomics Consortia



PubNet

Publication Network Graph Utility

Clear

Submit

Query 1

Query 2 (optional)

documentation | examples | gallery | recent | saved | source

Node

● Author

○ Paper

○ PDB ID

○ GenBank ID

○ SwissProt

Edge

● Co-Authorship

○ Shared MeSH Term

○ Shared Location

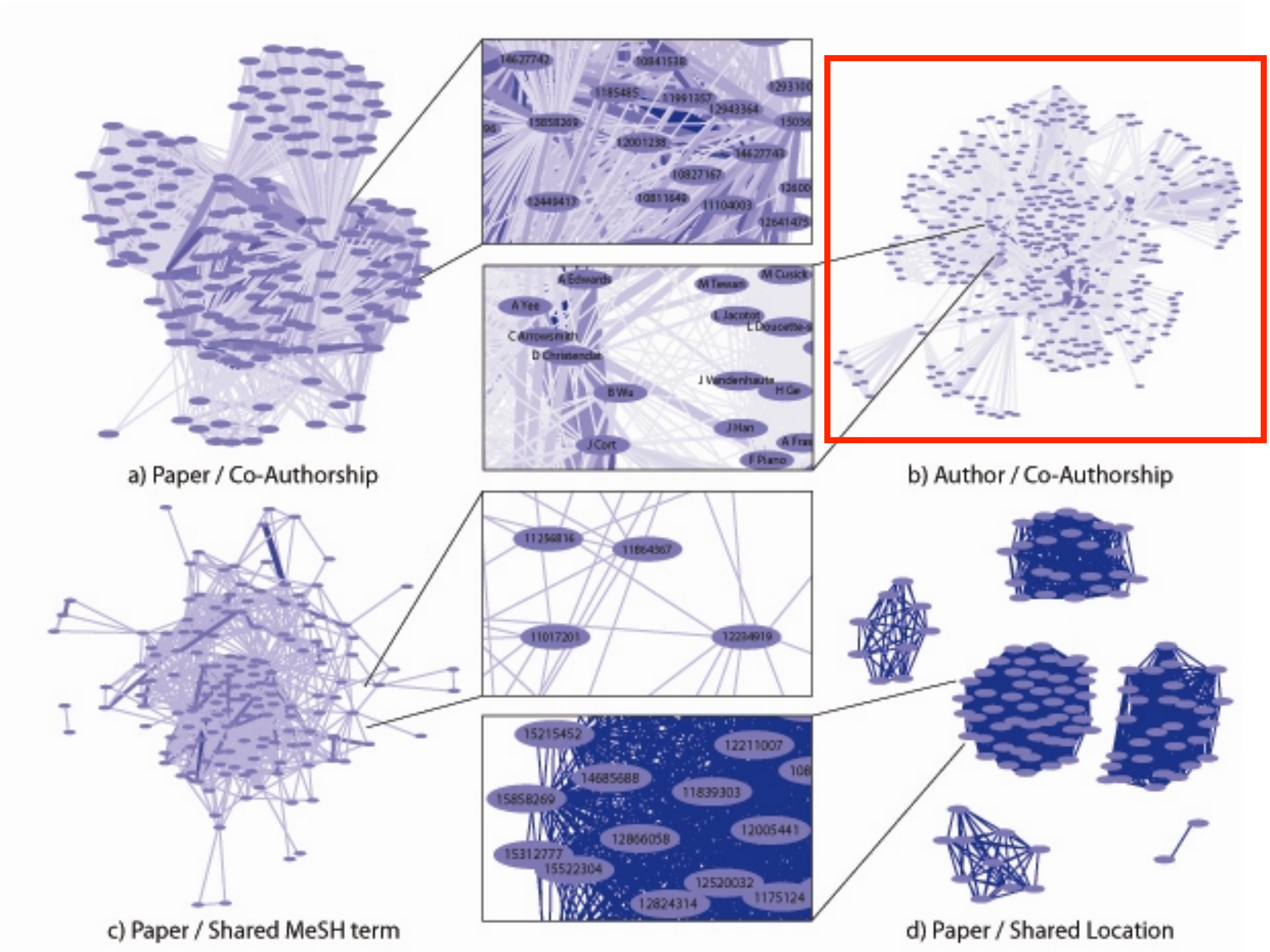
Drawing Options

□ Omit nodes with degree < 1

□ Do not scale edge thickness

[Douglas et al. GenomeBiol. ('05), pubnet.gersteinlab.org]

# Different Representations of the Publication Network of a Structural Genomics Center (NESG)



a) Paper / Co-Authorship

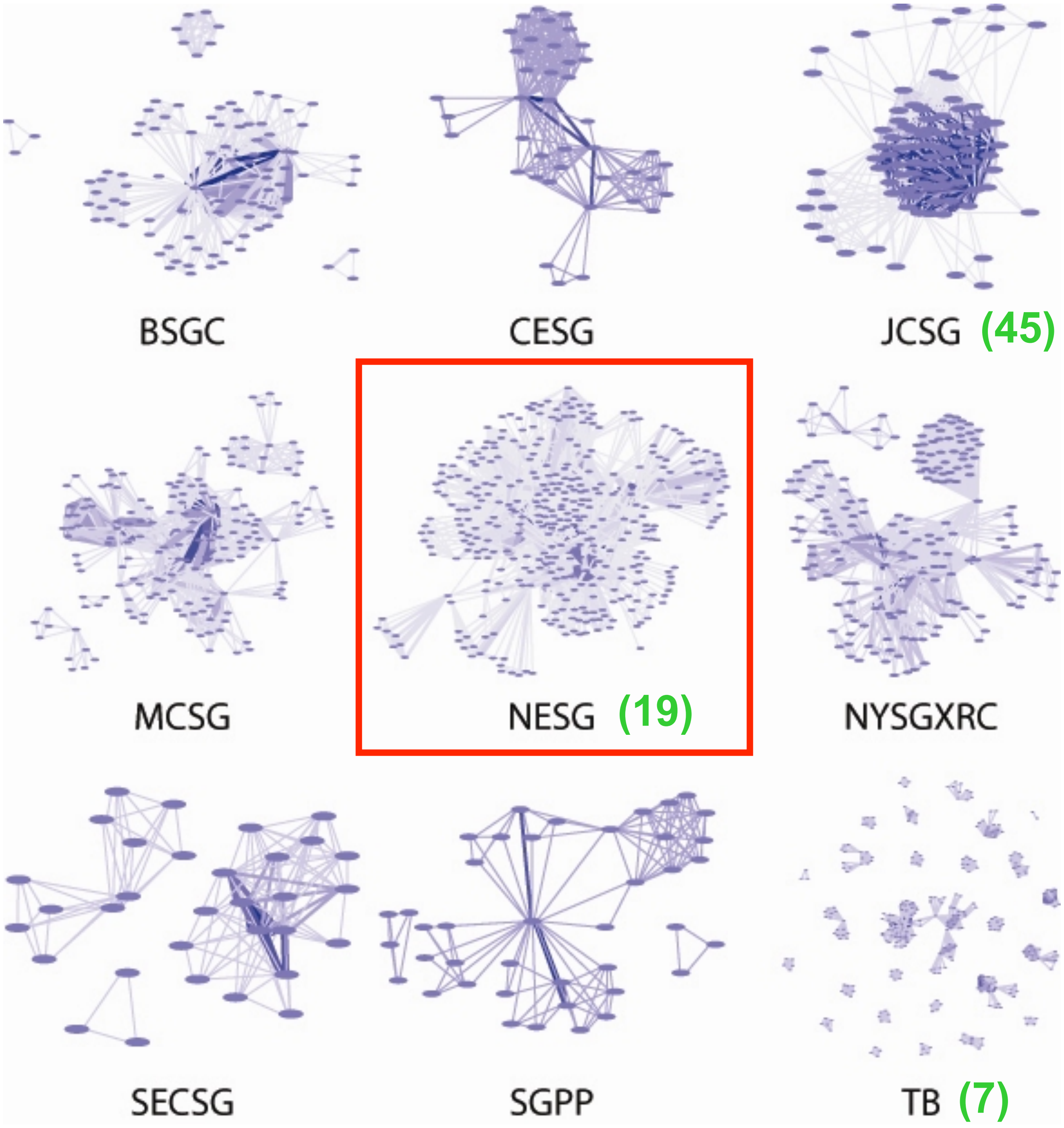
b) Author / Co-Authorship

c) Paper / Shared MeSH term

d) Paper / Shared Location

[Douglas et al. GenomeBiol. ('05), pubnet.gersteinlab.org]

# Co-authorship Networks comparing the 9 NIH Structural Genomics Centers

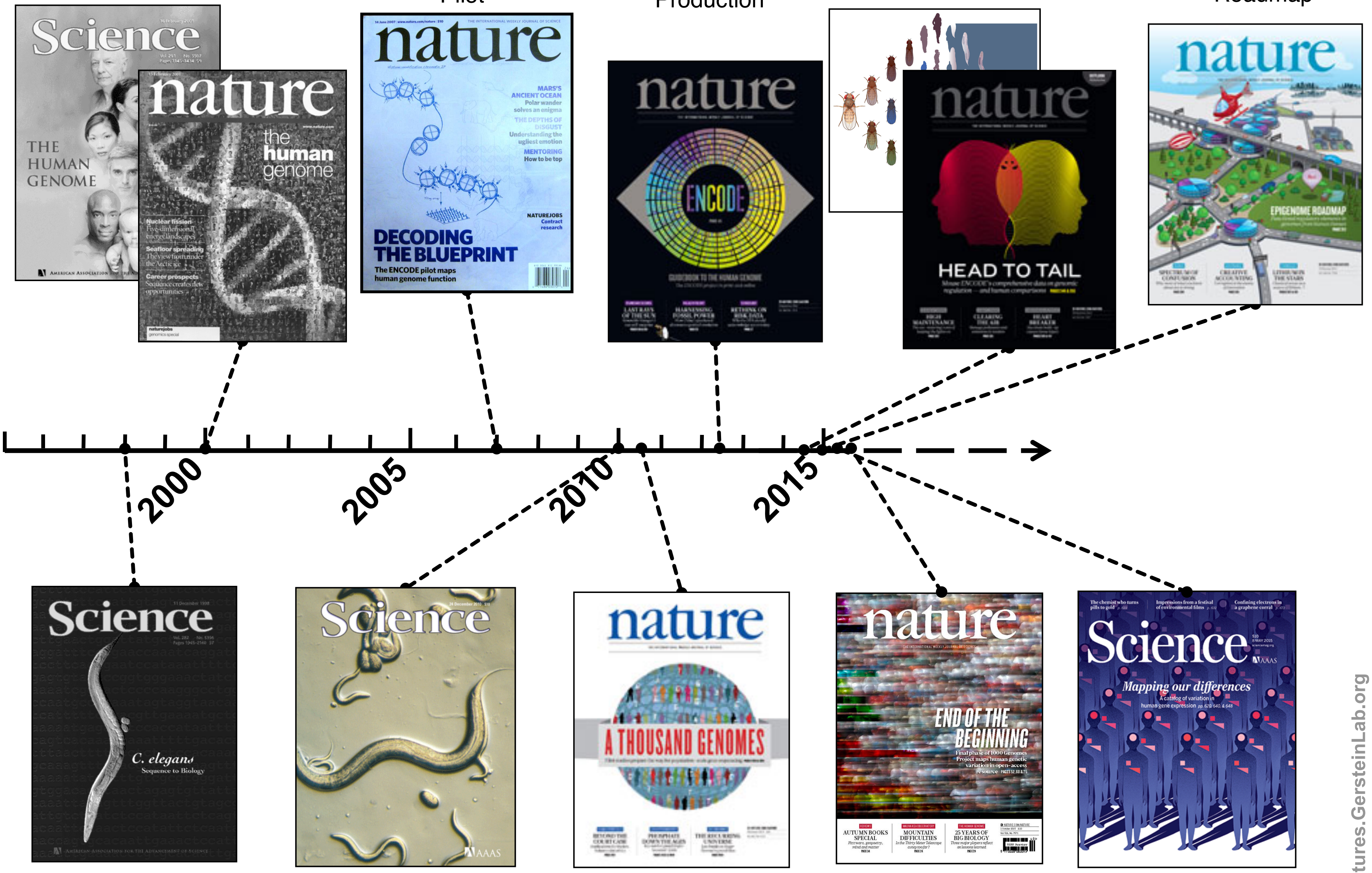


**Average Degree**

[Douglas et al.
GenomeBiol. ('05),
pubnet.gersteinlab.org]

# Analyzing the Structure of Genomic Science: Mapping the diffusion of ideas & data across disciplines

- Intro: Using the Data Exhaust from Consortia
- PubNet Tool for analyzing co-publication patterns
  - Different patterns for Str. Genomics centers
- Analysis of Evolution of the ENCODE Consortium
  - Differences in "modularity" for members & users
  - Key role for brokers
- Other Examples of Idea & Data Diffusion
  - RNAi & SRA bases
- Quant. Model of Information Diffusion
  - Based on PLOS ALM
  - Random multiplicative process
  - Moderated by parm. w/ 2 time regimes



HumanGenome Project

ENCODE Pilot

ENCODE Production

ComparativeENCODE

Epigenome Roadmap

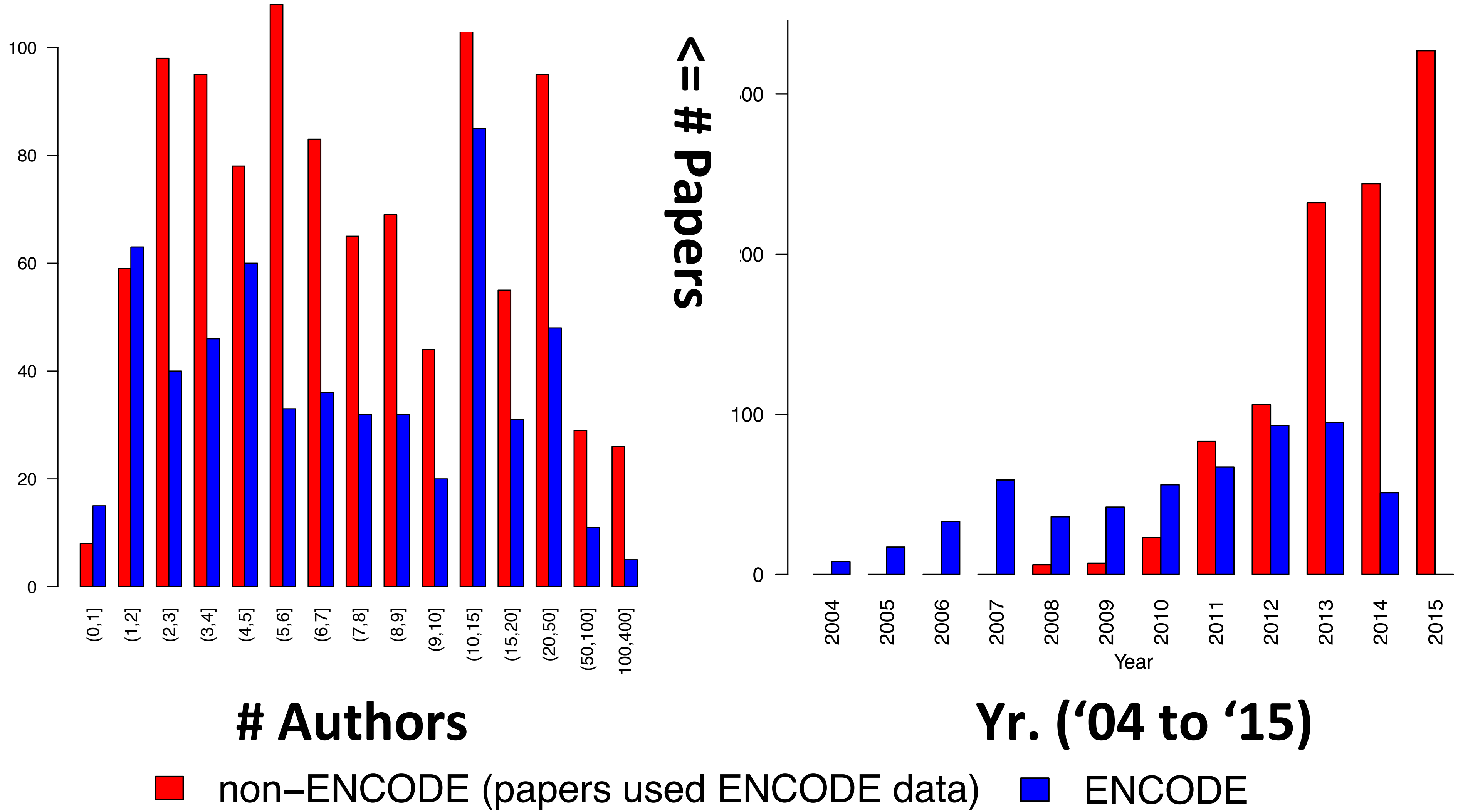
Worm Genome

modENCODE

1000 Genomes Pilot

1000 Genomes Phase 3

GTEx



With help of NHGRI, identified:

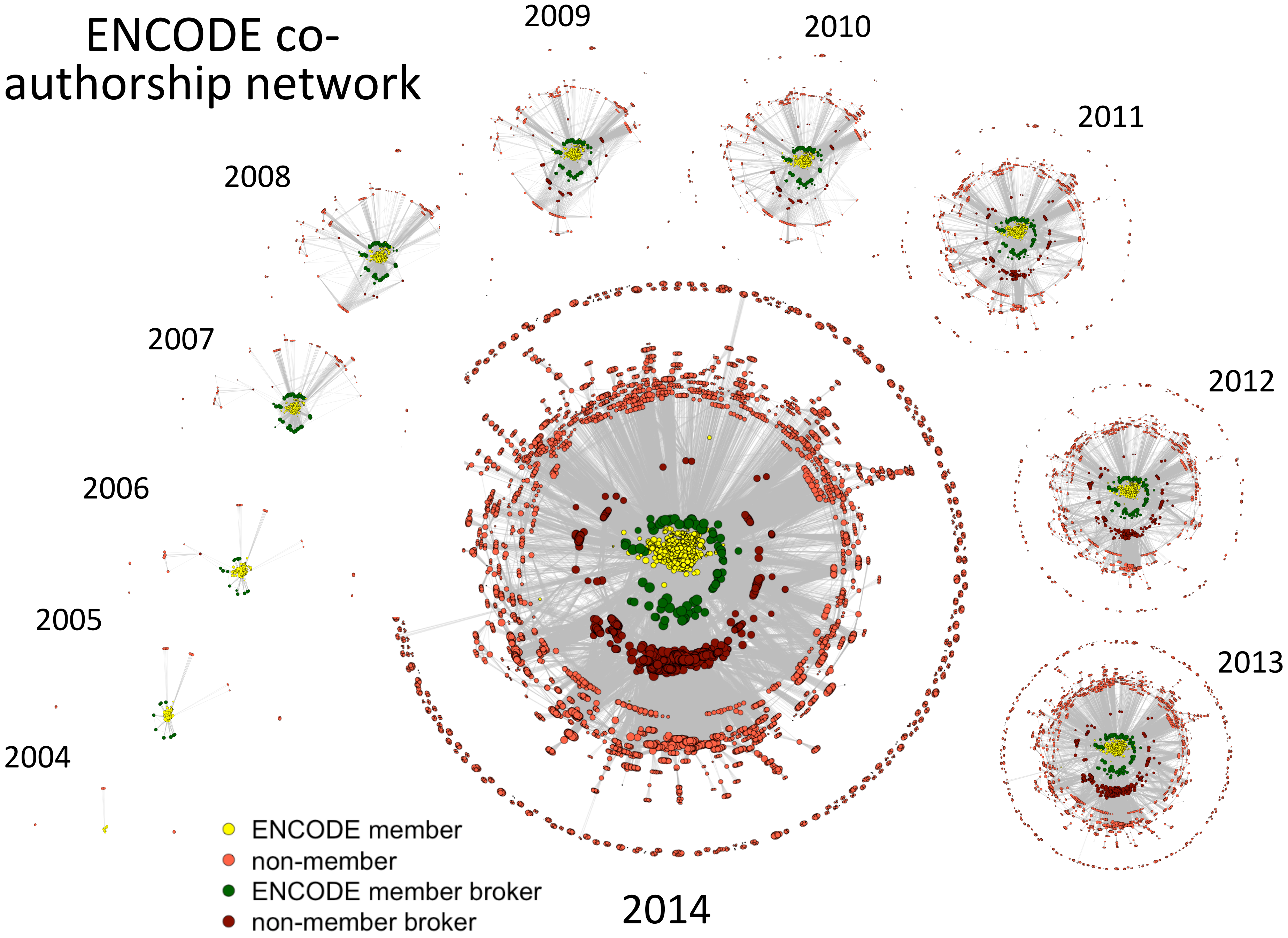
1,786 ENCODE members & 8,263 non-members

from 558 consortium papers supported by ENCODE funding &

702 community papers that used ENCODE data but were not supported by ENCODE funding

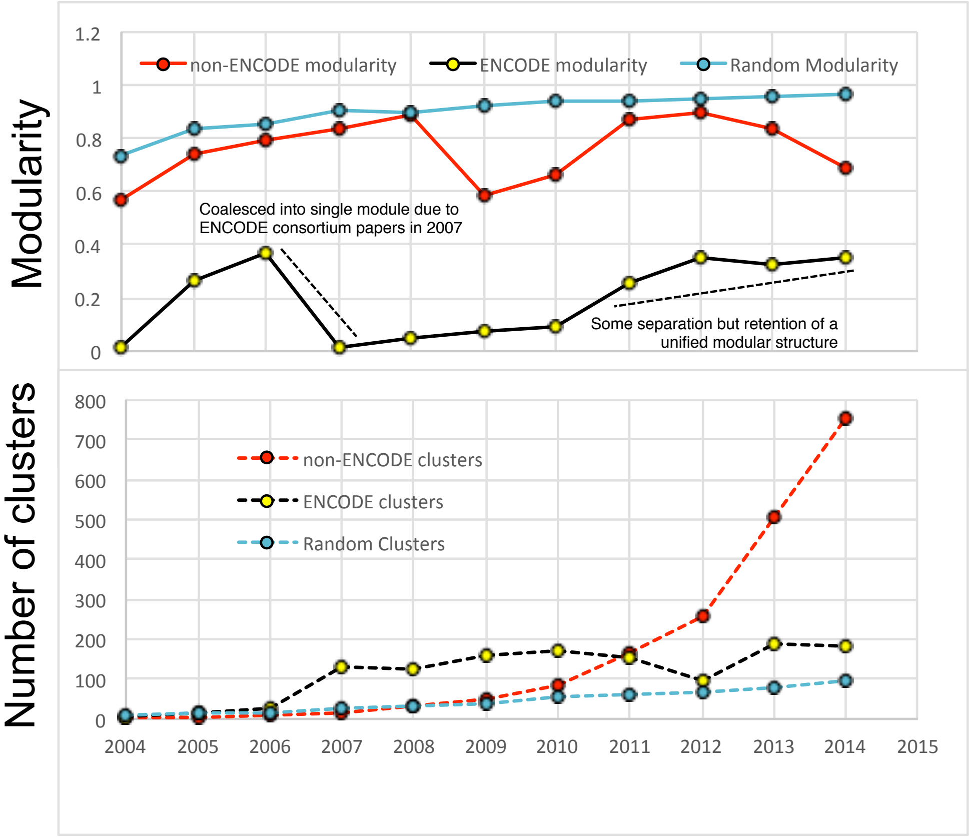
[Wang et al., TIG ('16)]

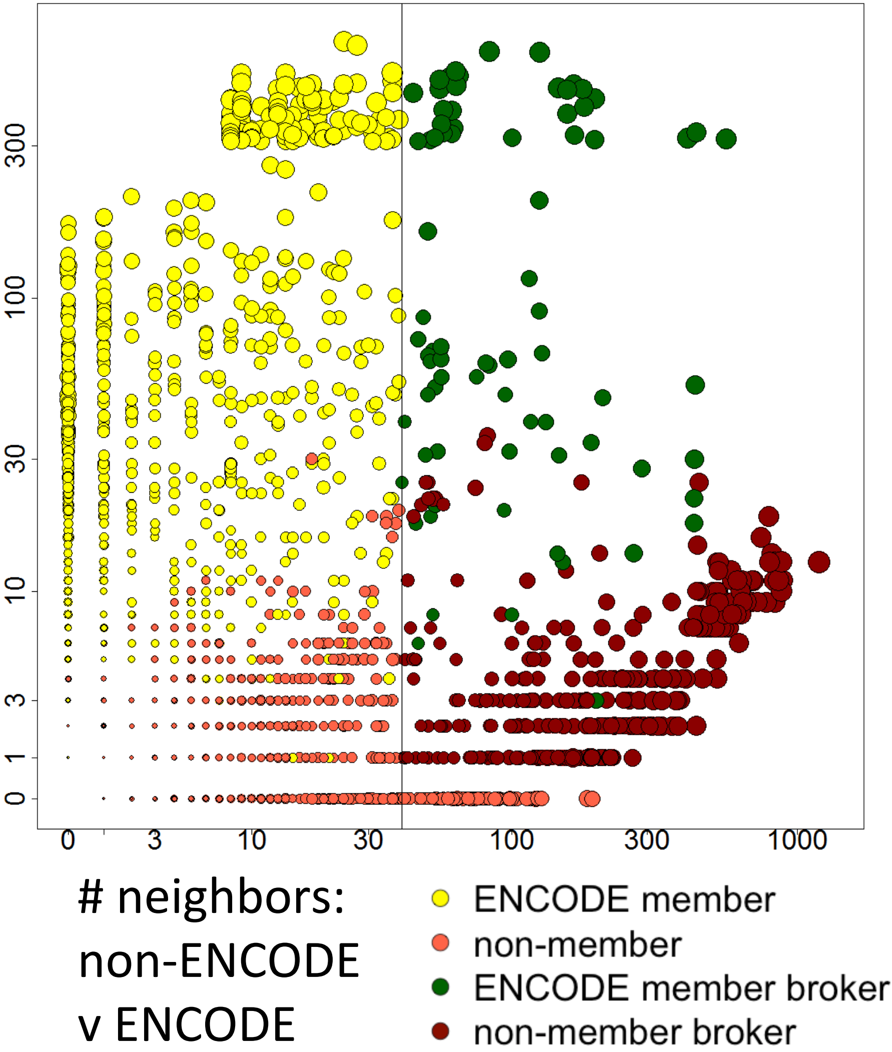
# ENCODE co-authorship network



[Wang et al., TIG ('16)]

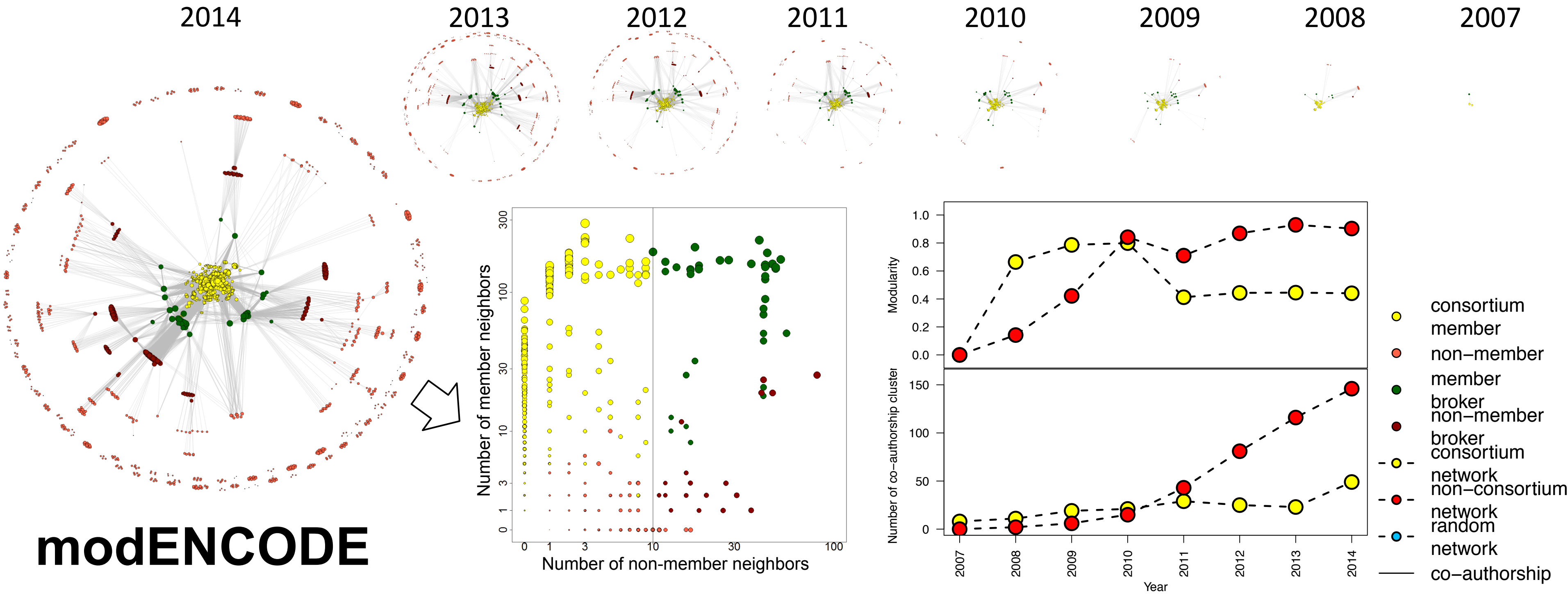
# Network statistics highlight change in modularity with consortium rollouts (L) & importance of broker role (R)





[Wang et al., TIG ('16)]

# Similar Findings in terms of modularity & broker scientists in the modENCODE consortium as for ENCODE



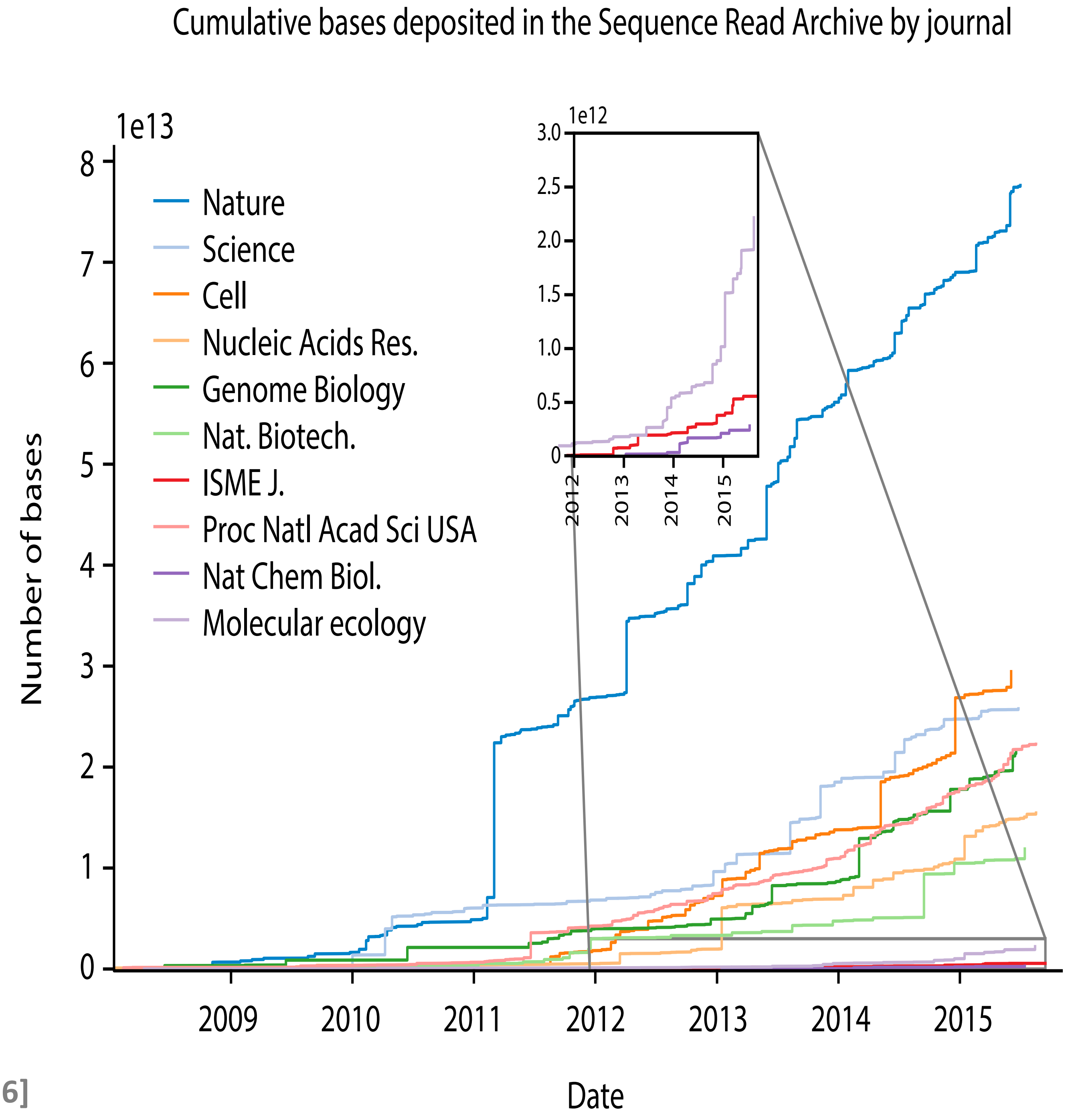
modENCODE

[Wang et al., TIG ('16)]

# Analyzing the Structure of Genomic Science: Mapping the diffusion of ideas & data across disciplines

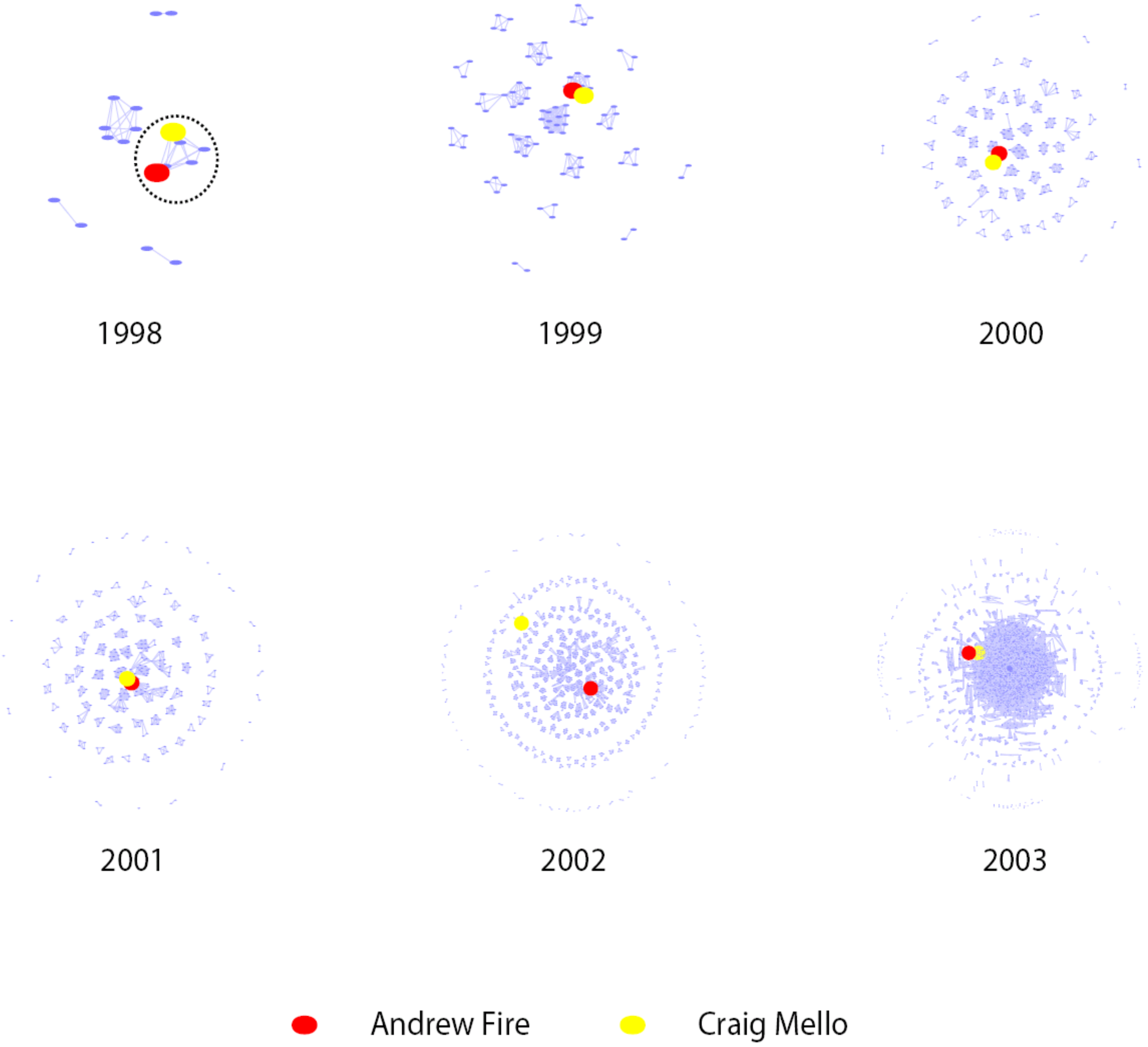
- Intro: Using the Data Exhaust from Consortia
- PubNet Tool for analyzing co-publication patterns
  - Different patterns for Str. Genomics centers
- Analysis of Evolution of the ENCODE Consortium
  - Differences in “modularity” for members & users
  - Key role for brokers
- Other Examples of Idea & Data Diffusion
  - RNAi & SRA bases
- Quant. Model of Information Diffusion
  - Based on PLOS ALM
  - Random multiplicative process
  - Moderated by parm. w/ 2 time regimes

# Diffusion of a data type (sequenced bases) measured by occurrence in specialty journals



[Muir et al. GenomeBiol. '16]

# RNAi: Birth of a Field in the Literature Culminating in the 2006 Nobel



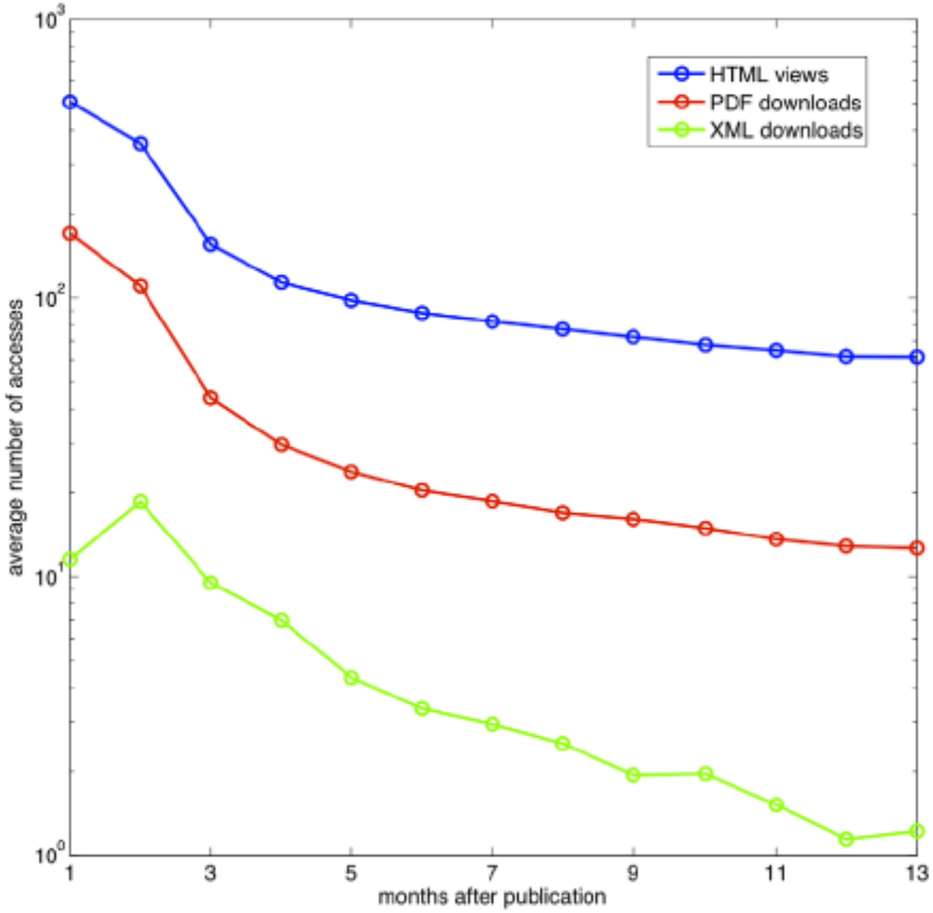
Source:
Gerstein & Douglas. PLoS Comp. Bio. 3:e80 (2007)
PubNet.GersteinLab.org

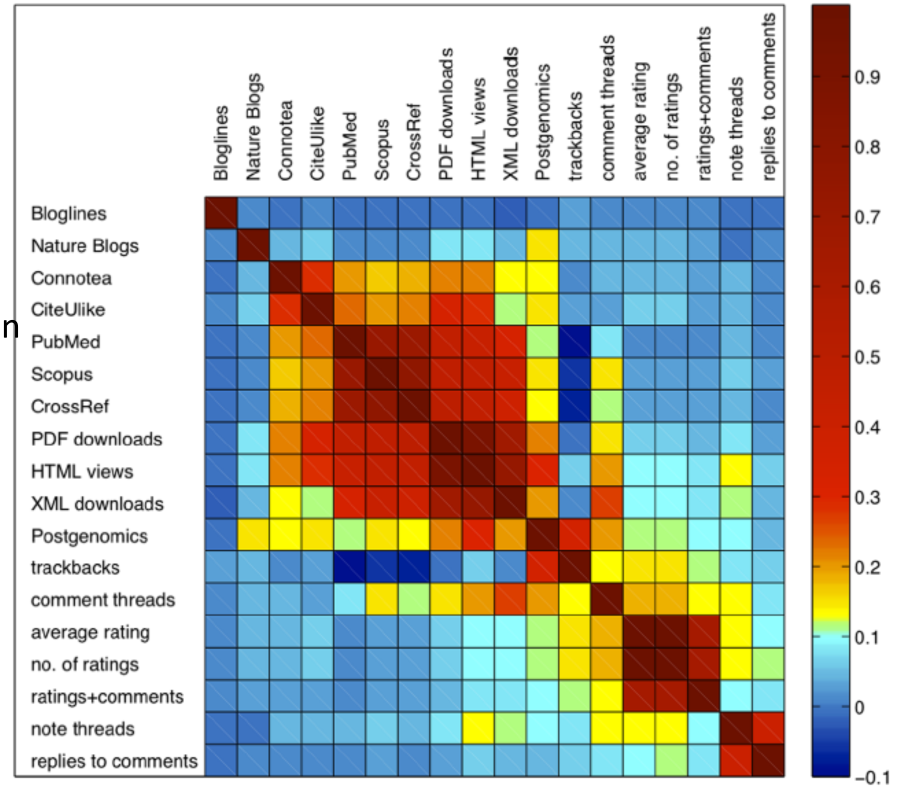
## Analyzing the Structure of Genomic Science: Mapping the diffusion of ideas & data across disciplines

- Intro: Using the Data Exhaust from Consortia
- PubNet Tool for analyzing co-publication patterns
  - Different patterns for Str. Genomics centers
- Analysis of Evolution of the ENCODE Consortium
  - Differences in "modularity" for members & users
  - Key role for brokers
- Other Examples of Idea & Data Diffusion
  - RNAi & SRA bases
- Quant. Model of Information Diffusion
  - Based on PLOS ALM
  - Random multiplicative process
  - Moderated by parm. w/ 2 time regimes

# Spread of information as a diffusion process

The knowledge of a scientific publication is a piece of information





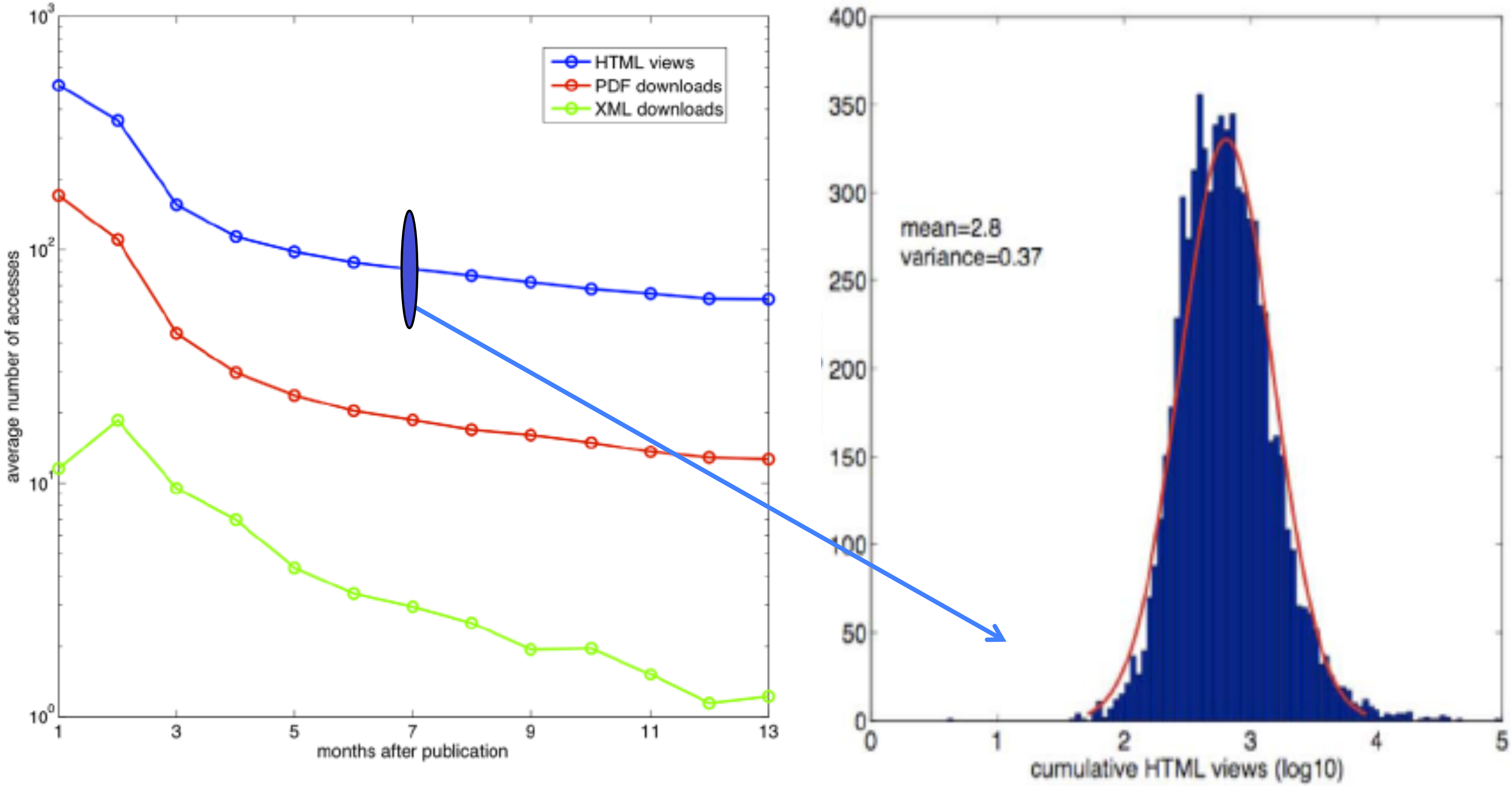
Based on PLOS ALM data for ~7000 papers

the access of different articles follows a log-normal distribution

Yan and Gerstein PLoS One 2011.

# Spread of information as a diffusion process

The knowledge of a scientific publication is a piece of information

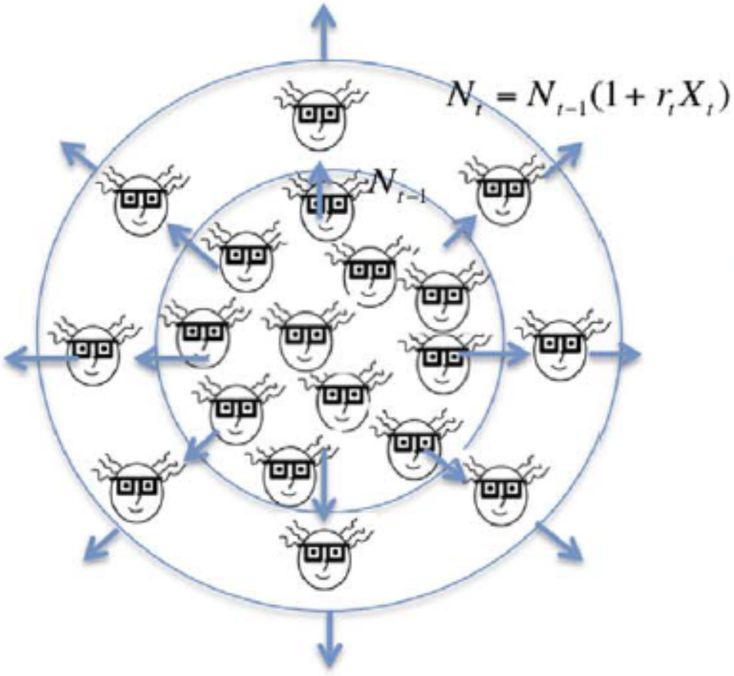


the access of different articles follows a log-normal distribution

Yan and Gerstein PLoS One 2011.

# Modeling Information diffusion

share the information with friends?



log-normal distribution suggests a simple model,
random multiplicative process for a given paper p:

$$N_t(p) = N_{t-1}(p)\,(\,1 + r_t\,X_t(p)\,)$$

$N_t$=cumulative number of accesses for a given paper up to time t

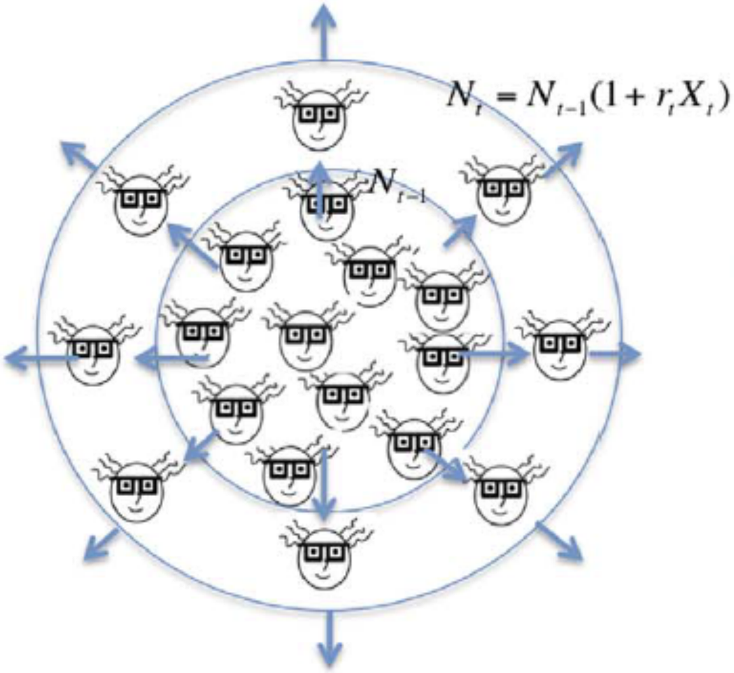
$X_t$ = iid random variable whose mean represents avg. fraction of scientists willing to "spread" the paper at t

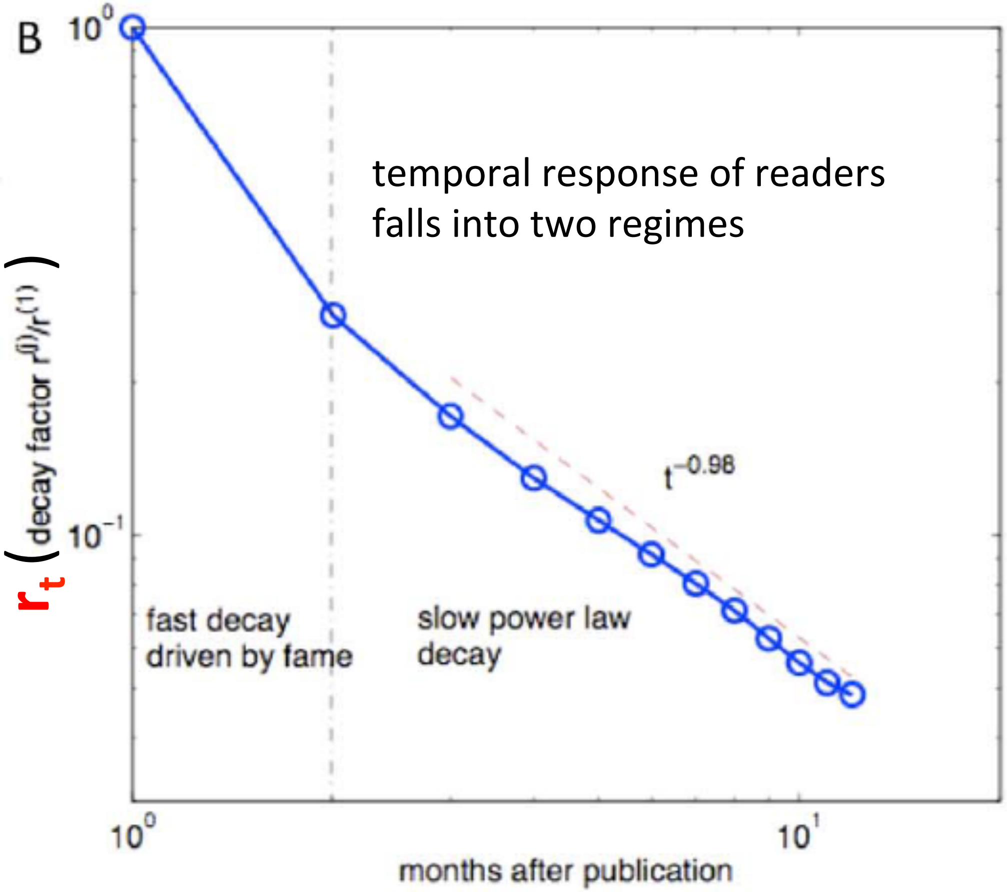
$r_t$ = moderating parameter on how the "spreading" changes over time

Yan and Gerstein PLoS One 2011.

# Modeling Information diffusion

## Refine values of $r_t$ for different times





temporal response of readers falls into two regimes

Yan and Gerstein PLoS One 2011.

# Analyzing the Structure of Genomic Science: Mapping the diffusion of ideas & data across disciplines

- Intro: Using the Data Exhaust from Consortia
- PubNet Tool for analyzing co-publication patterns
  - Different patterns for Str. Genomics centers
- Analysis of Evolution of the ENCODE Consortium
  - Differences in "modularity" for members & users
  - Key role for brokers
- Other Examples of Idea & Data Diffusion
  - RNAi & SRA bases
- Quant. Model of Information Diffusion
  - Based on PLOS ALM
  - Random multiplicative process
  - Moderated by parm. w/ 2 time regimes

## Analyzing the Structure of Genomic Science: Mapping the diffusion of ideas & data across disciplines

- Intro: Using the Data Exhaust from Consortia
- PubNet Tool for analyzing co-publication patterns
  - Different patterns for Str. Genomics centers
- Analysis of Evolution of the ENCODE Consortium
  - Differences in “modularity” for members & users
  - Key role for brokers
- Other Examples of Idea & Data Diffusion
  - RNAi & SRA bases
- Quant. Model of Information Diffusion
  - Based on PLOS ALM
  - Random multiplicative process
  - Moderated by parm. w/ 2 time regimes

# Acknowledgments

"Encode authors"

**D Wang**,

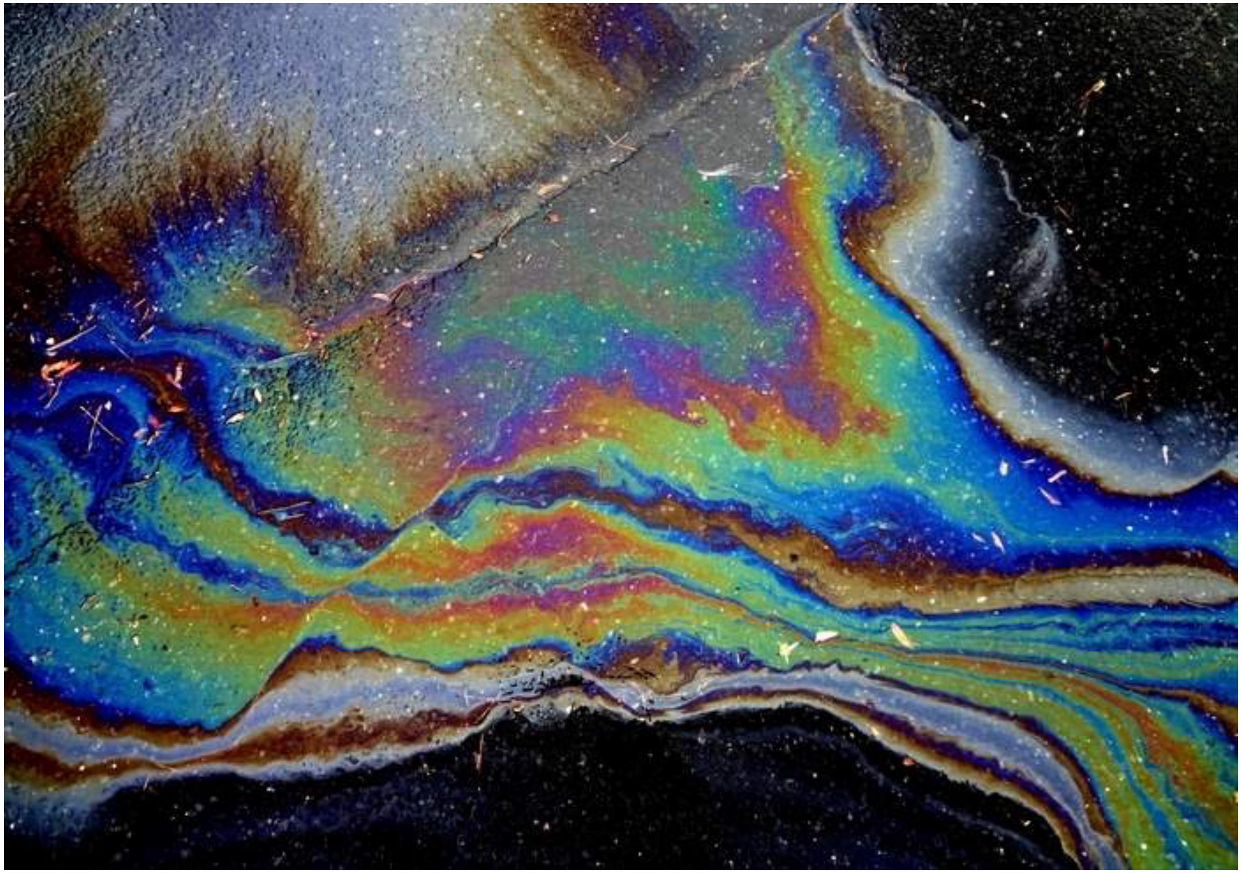
KK Yan,

J Rozowsky,

E Pan

"Info Flow"

**KK Yan**



Cost of Seq. & SRA bases

P Muir, S Li, S Lou, D Wang, DJ Spakowicz, L Salichos, J Zhang, GM Weinstock, F Isaacs, J Rozowsky

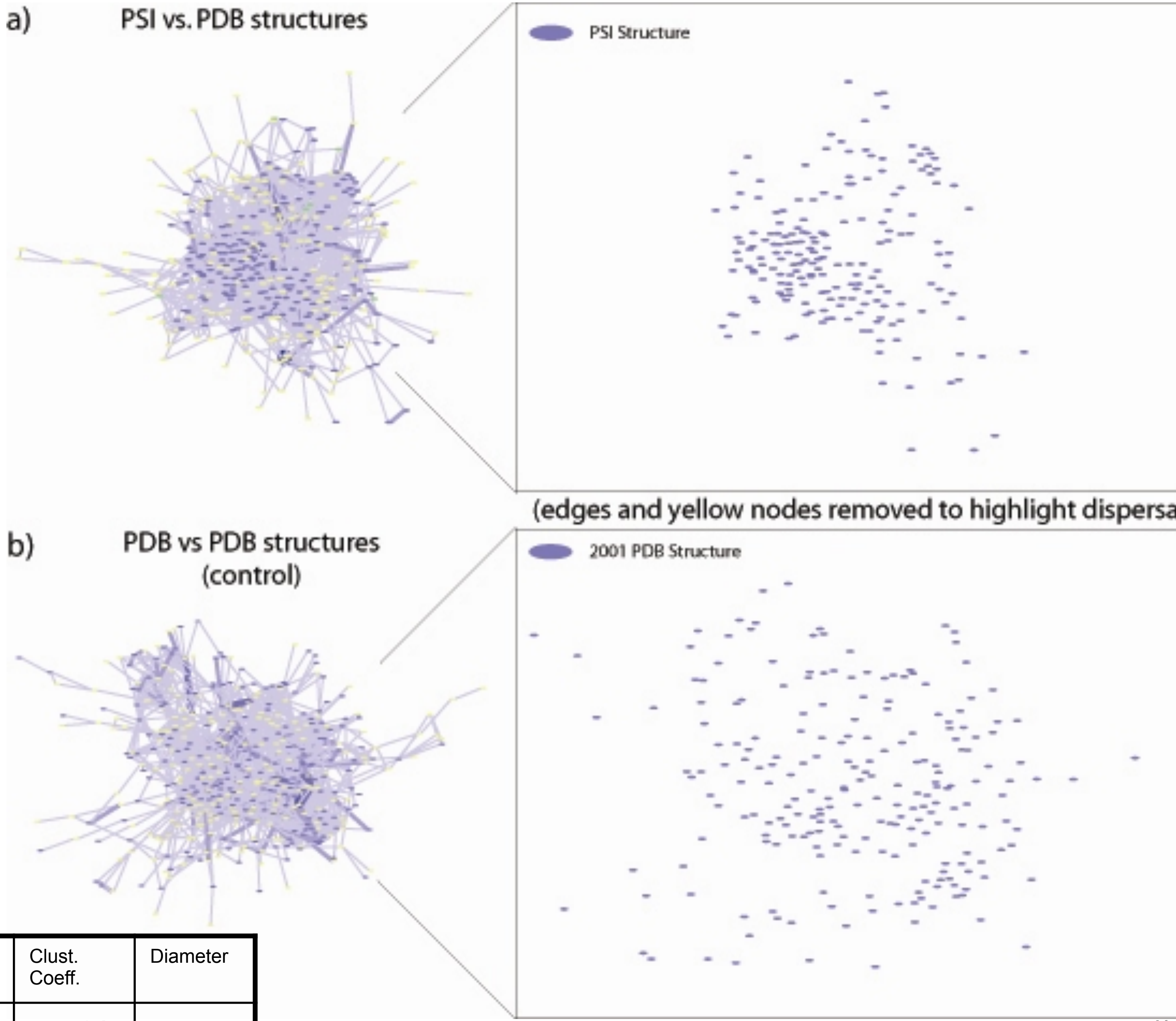
PubNet.gersteinlab.org & RNAi -- SM Douglas

Vision for "Text Mining" -- A Rzhetsky, M Seringhaus

Hiring Postdocs. See gersteinlab.org/jobs

# Extra

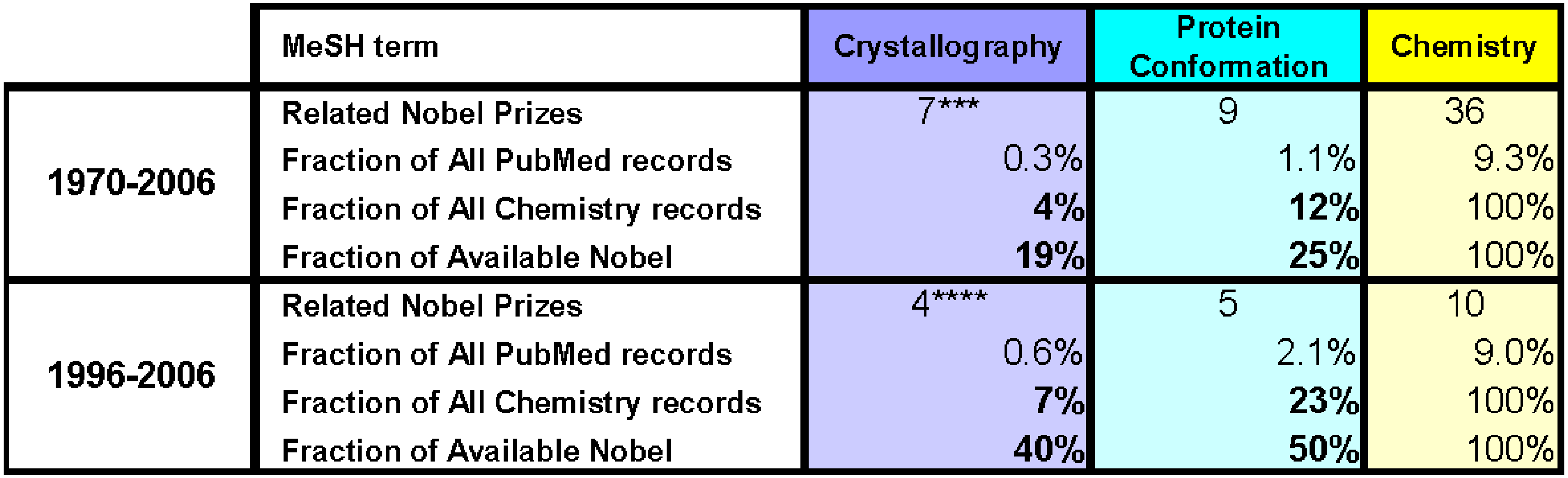
# Clustering structures determined by struc. genomics consortia according to functional similarity: Is there a functional bias in consortia structures?



| | Avg. Degree | Avg. Path | Clust. Coeff. | Diameter |
|---|---|---|---|---|
| PSI | 24 | 2.6 | 37% | 7 |
| PDB | 6 | 3.9 | 31% | 9 |

[Douglas et al. GenomeBiol. ('05), pubnet.gersteinlab.org]

# Over-representation of crystallography among the Nobel Prizes, highlighted by the 2006 Nobels



| | MeSH term | Crystallography | Protein Conformation | Chemistry |
|---|---|---|---|---|
| **1970-2006** | Related Nobel Prizes | 7*** | 9 | 36 |
| | Fraction of All PubMed records | 0.3% | 1.1% | 9.3% |
| | Fraction of All Chemistry records | **4%** | **12%** | 100% |
| | Fraction of Available Nobel | **19%** | **25%** | 100% |
| **1996-2006** | Related Nobel Prizes | 4**** | 5 | 10 |
| | Fraction of All PubMed records | 0.6% | 2.1% | 9.0% |
| | Fraction of All Chemistry records | **7%** | **23%** | 100% |
| | Fraction of Available Nobel | **40%** | **50%** | 100% |

[Seringhaus & Gerstein, Science (2007)

# Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2015.
  - Please read permissions statement at www.**gersteinlab.org/misc/permissions.html** .
  - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
  - Paper references in the talk were mostly from Papers.GersteinLab.org.

- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info .
  - In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt