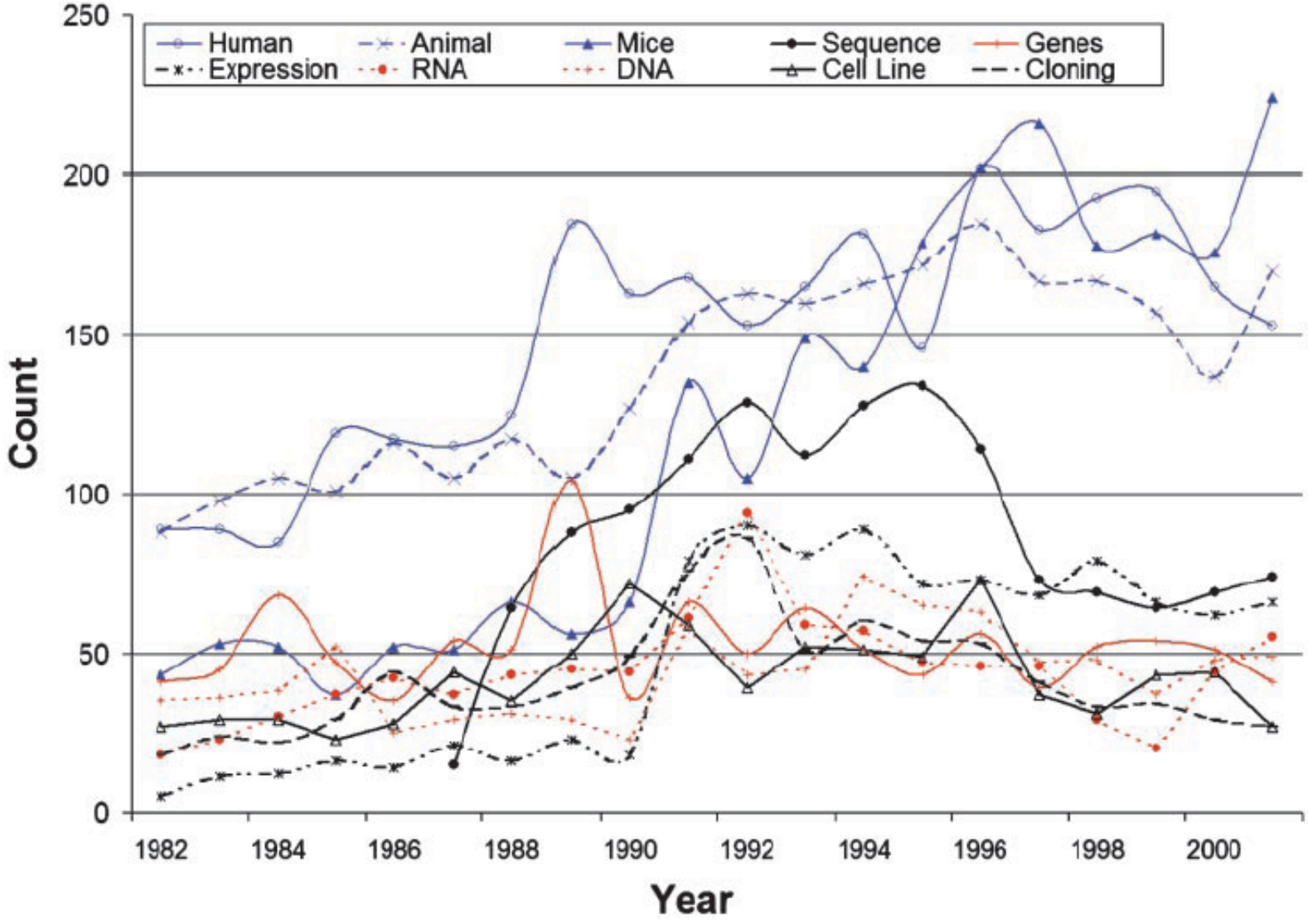
TOPIC DYNAMICS: AN ALTERNATIVE MODEL OF 'BURSTS' IN STREAMS OF TOPICS

Dan He and D. Stott Parker
Department of Computer Science, UCLA
SIGKDD 2010
Contact: danhe@cs.ucla.edu

BACKGROUND

- Monitor occurrences of topics in streams.
- Burst: period of "elevated" occurrence of events.
- Very useful to identify when a topic becomes "hot" -- emergent research areas, change of focus.
- Critical for resource allocation or grant investment from funding agencies, research labs, institutions and corporations.
- Challenge: identify bursts for PubMed/MedLine indexed biomedical literature abstracts
- PubMed: A database containing 20 million biomedical publication citations

HOW TO DEFINE "ELEVATED" OCCURRENCES OF TOPICS?



EXISTING BURST MODELS

- There are several well-known bursts models:
  - Kleinberg's model(Kleinberg et al. 2009, Kleinberg 2002): bursts are defined relative to an infinite state automaton in which each state represents a message arrival rate.
  - Shasha's model(Zhu and Shasha 2003): bursts are defined on wavelet-like hierarchies of fixed-length time intervals. Bursts occur in those intervals in which the event frequency of occurrence exceeds a given threshold.

LIMITATIONS OF EXISTING BURST MODELS

- The terms in PubMed are arranged in a hierarchy.
- Existing burst models assume term independence. However, in a hierarchy, the terms are often correlated.
- The occurrence of a child term might contribute to the occurrence of a parent term.
- Occurrences of the terms might want to be accumulated in the hierarchy.

LIMITATIONS OF EXISTING BURST MODELS

- The definition of burst “strength” or “intensity” is vague.
- In Kleinberg’s model, burst strength is defined as the state of the infinite automaton.
- Bursts in the same state have the same strength.
- The number of states is usually small.
- Burst strength is discrete.

LIMITATIONS OF EXISTING BURST MODELS

- With PubMed, the underlying process is not clearly "memoryless", an assumption in Kleinberg's model.
- Biomedical publications are released in batches. The time interval between the batches are fixed.
- The Kleinberg and Shasha models are computationally expensive.
- And more…..

TOPIC DYNAMICS MODEL

- Use kinetics concepts from physics – Mass and Position – to derive velocity, momentum, acceleration and force.
- View bursts as intervals of increasing "momentum".

USEFUL MEASURES OF A TOPIC

- **Position:** x(t)
  - a linearly-ordered measure or intensity of a topic at time t. Examples:
    - number of articles containing the topic
    - number of pages including the topic
    - number of accesses or downloads
- **Mass:** m(t)
  - aggregate weight or importance of a topic at time t. Examples:
    - number of article citations
    - journal impact factors
    - journal relevance measures.

USEFUL MEASURES OF A TOPIC

- **mass:** $m(t)$
- **position:** $x(t)$
- **velocity:** $v(t) = dx(t)/dt \simeq \Delta x(t)/\Delta t$
- **momentum:** mass $\cdot$ velocity $\simeq m(t) \cdot v(t)$
- **acceleration:** $dv(t)/dt$
- **force:** mass $\cdot$ acceleration $\simeq m(t) \cdot dv(t)/dt$

TOPIC DYNAMICS MODEL

- With the physical notions just defined, we can model a burst in terms of increasing momentum – derivative of momentum is positive.
- More specifically, assume that mass is constant, and that a burst is a time interval over which acceleration (= force) is positive.
  - **momentum:** mass $\cdot$ velocity $\approx m(t) \cdot v(t)$
  - **acceleration:** $dv(t)/dt$
  - **force:** mass $\cdot$ acceleration $\approx m(t) \cdot dv(t)/dt$
  - If weighted by mass, positive force.
  - If not weighted by mass, positive acceleration.

STOCK MARKET ANALOGY

- The trends for the terms in biomedical publication can be visualized using charts that are highly reminiscent of the stock market.
- **Position** x(t): measures "value" like stock price.
- **Mass** m(t): measures "importance" like trading volume.
- In our model, we consider m(t) = 1 for all terms, and x(t) is the frequency of a MeSH topic term in all publications for year t.

TREND ANALYSIS INDICATOR

- EMA (Exponential Moving Average): For a variable $x = x(t)$ (Position) which corresponds to discrete time series $x = \{x_t \mid t = 0,1,2...,\}$, the n-day EMA is defined as:

$$EMA(n)[x]_t = \alpha x_t + (1-\alpha)EMA(n-1)[x]_{t-1}$$
$$= \sum_{k=0}^{n} \alpha(1-\alpha)^k x_{t-k} = EMA(n)$$

$$\alpha = 2/(n+1)$$

TREND ANALYSIS INDICATOR

- MACD (Moving Average Convergence/ Divergence): MACD for a variable $x(t)$ (Position) is defined as the difference of its n_1 -day and n_2-day moving averages:

$$MACD(n_1, n_2) = EMA(n_1) - EMA(n_2)$$

- This difference is an estimate of $\Delta x / \Delta t$ and hence an estimate of velocity.
- In technical stock analysis, a common choice is $n_1 = 12, \quad n_2 = 26$

TREND ANALYSIS INDICATOR

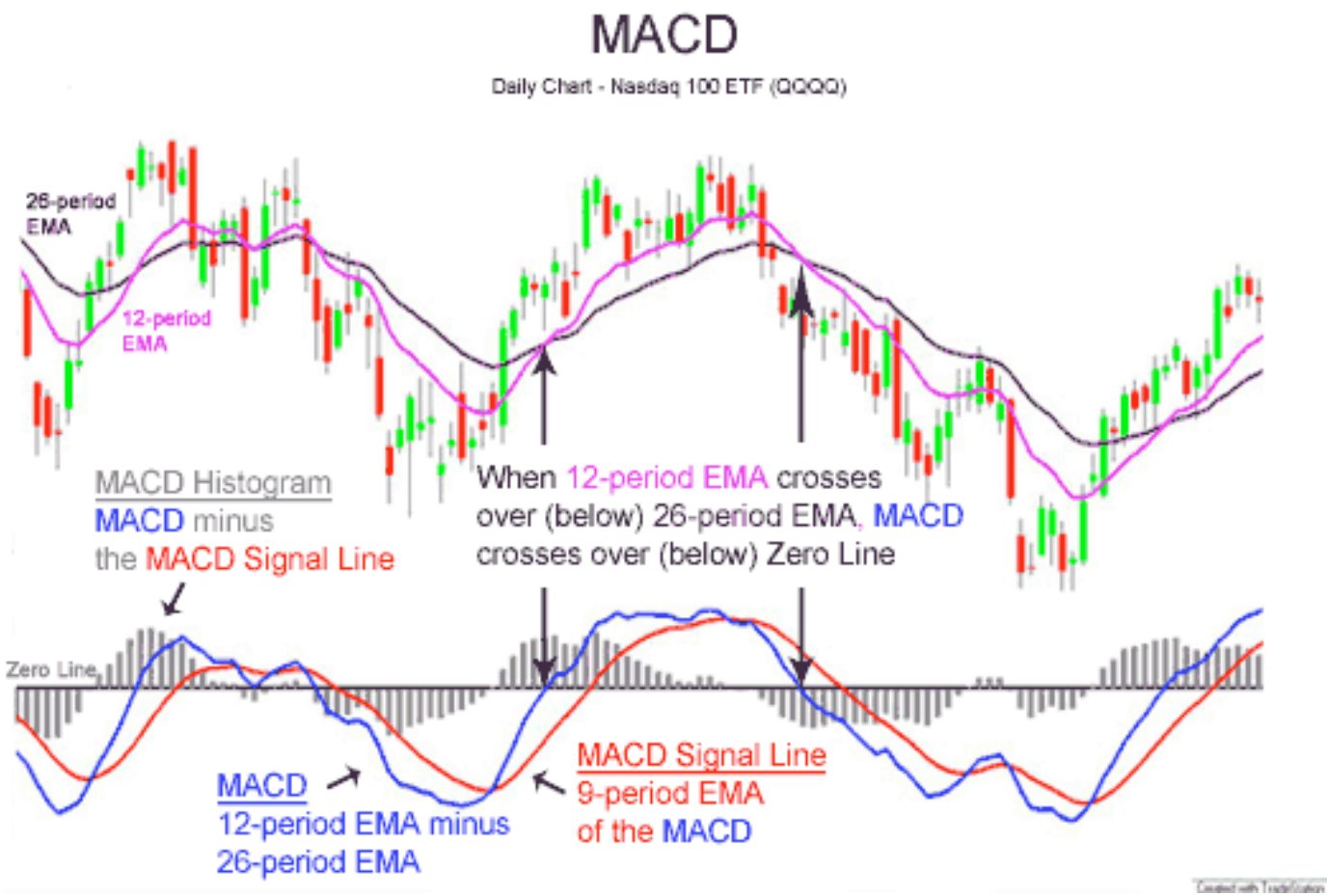
- MACD histogram: MACD histogram is an estimate of the derivative of the MACD:

$$signal(n_1, n_2, n_3) = EMA(n_3)[MACD(n_1, n_2)]$$

$$histogram(n_1, n_2, n_3) = MACD(n_1, n_2) - signal(n_1, n_2, n_3)$$

- MACD histogram is the difference between MACD and its moving average, and thus is an estimate of the "second derivative" of $x(t)$, or the acceleration.
- In technical stock analysis, a common choice is (12, 26, 9)

MACD EXAMPLE with parameters (12, 26, 9)



STOCK MARKET ANALOGY

- We define bursts as intervals with positive acceleration.
- MACD histogram value is an estimate of Acceleration.
- ***Burst period: time interval over which the MACD histogram value is positive.***
- **Starting time of burst:** starting time of the positive MACD histogram interval.
- **Ending time of burst:** ending time of the positive MACD histogram interval.
- **Strength of burst:** the MACD histogram value.

PUBMED TOPICS

- MeSH (Medical Subject Headings): A hierarchy of topics integrated with PubMed/MEDLINE.

```
C01.539                 Infection
C01.539.778                 Sexually Transmitted Diseases
C01.539.778.281                 Sexually Transmitted Diseases, Bacterial
C01.539.778.281.201                 Chancroid
C01.539.778.281.301                 Chlamydia Infections
C01.539.778.281.401                 Gonorrhea
C01.539.778.281.451                 Granuloma Inguinale
C01.539.778.281.859                 Syphilis
```

- All citations in PubMed/MEDLINE are labeled (annotated) with MeSH topics (terms) by experts.

MESH HIERARCHY

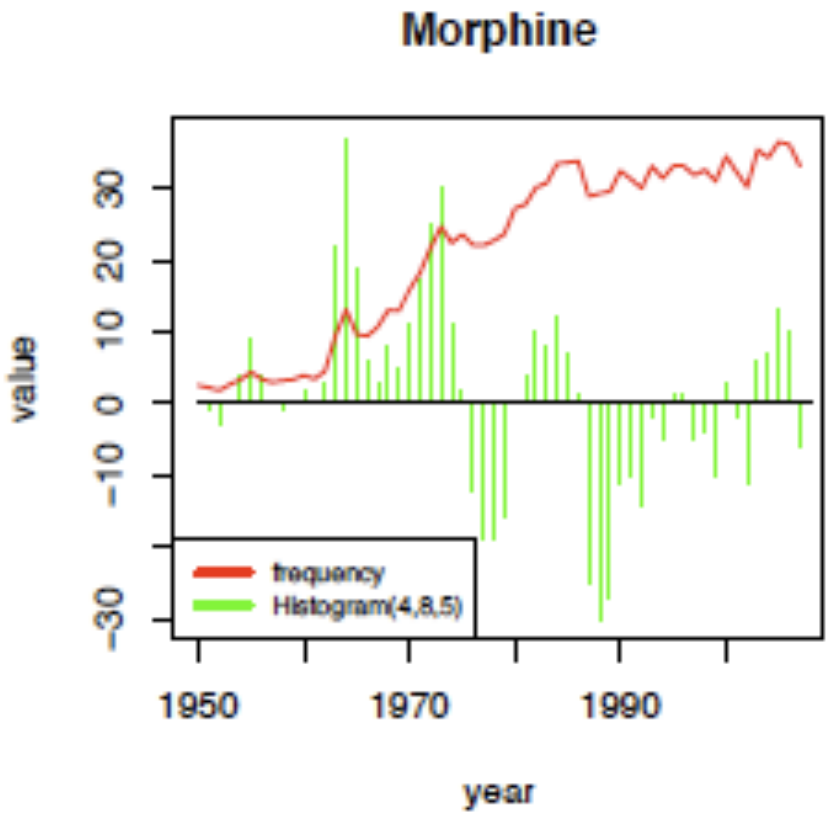
- Approximately 50,000 terms arranged in a hierarchy.

hierarchy depth	number of MeSH terms
1	109
2	1495
3	6527
4	12446
5	12702
6	7960
7	4304
8	1820
9	800
10	242
11	38

Setup

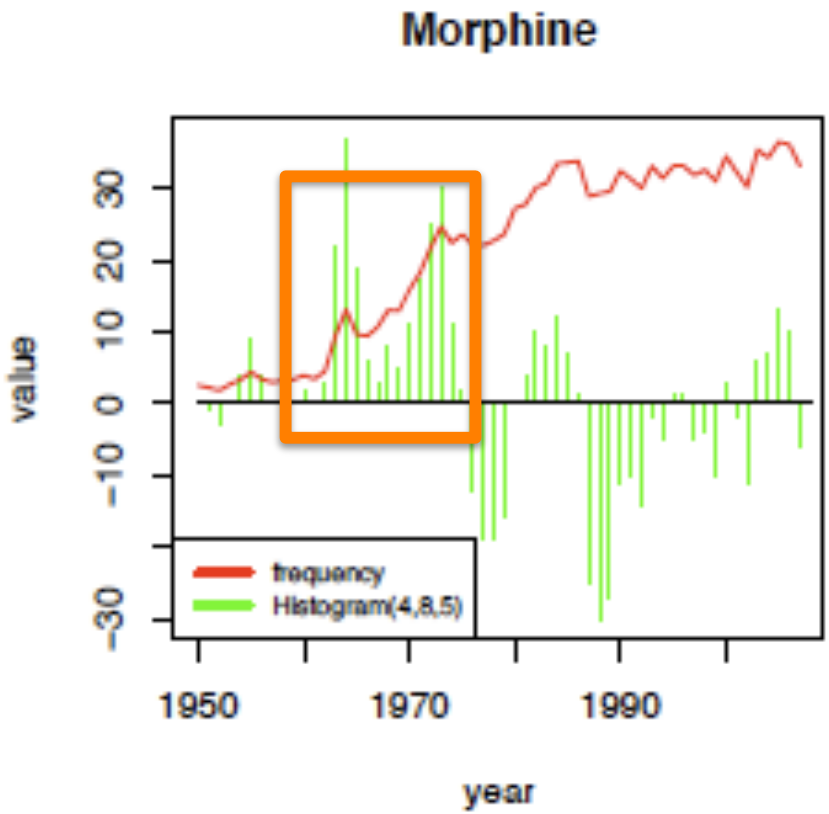
- Aggregate term frequency counts over annual time window from 20 million biomedical publication citations.
- Analysis is on annual numbers of publications.
- Aggregate the frequency of a term to all its ancestors in the MeSH hierarchy.
- Apply our model on all MeSH terms.
- We use MACD histogram parameter (4, 8, 5) in this analysis—all frequency values are annual.

VALIDATION OF TOPIC DYNAMICS MODEL



Morphine: A drug to relieve severe or agonizing pain and suffering

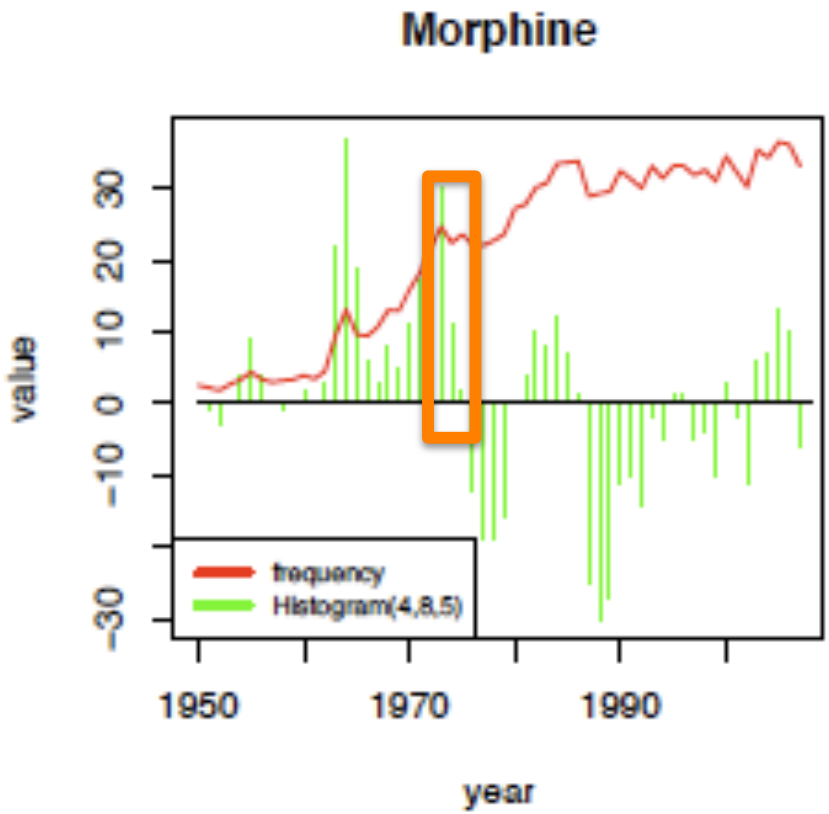
VALIDATION OF TOPIC DYNAMICS MODEL



Morphine: A drug to relieve severe or agonizing pain and suffering

The Vietnam War started in 1963 and ended in 1975

VALIDATION OF TOPIC DYNAMICS MODEL

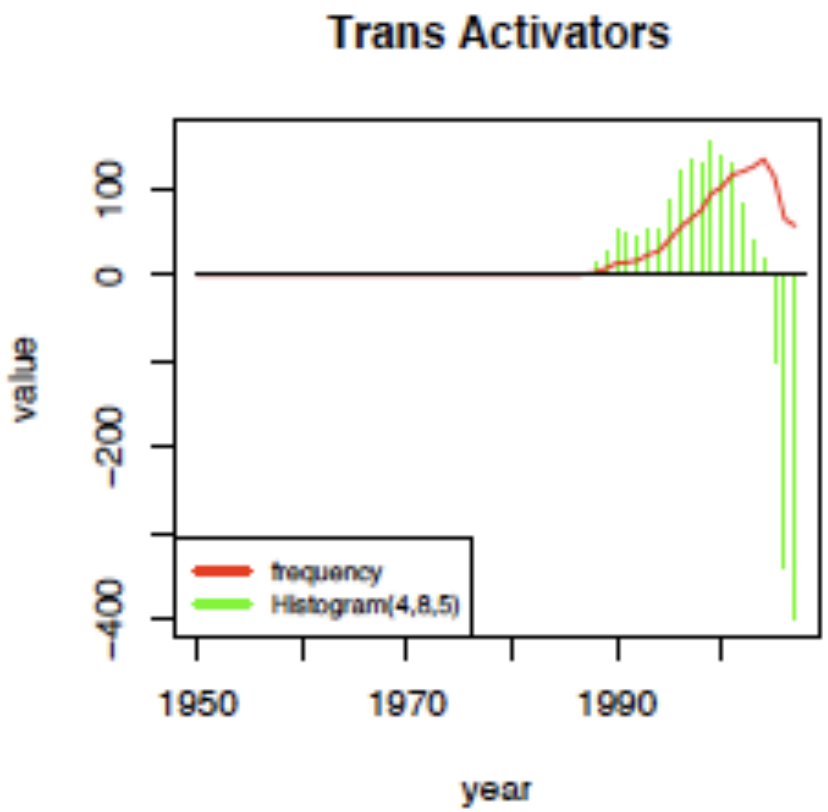


Morphine: A drug to relieve severe or agonizing pain and suffering

The Vietnam War started in 1963 and ended in 1975

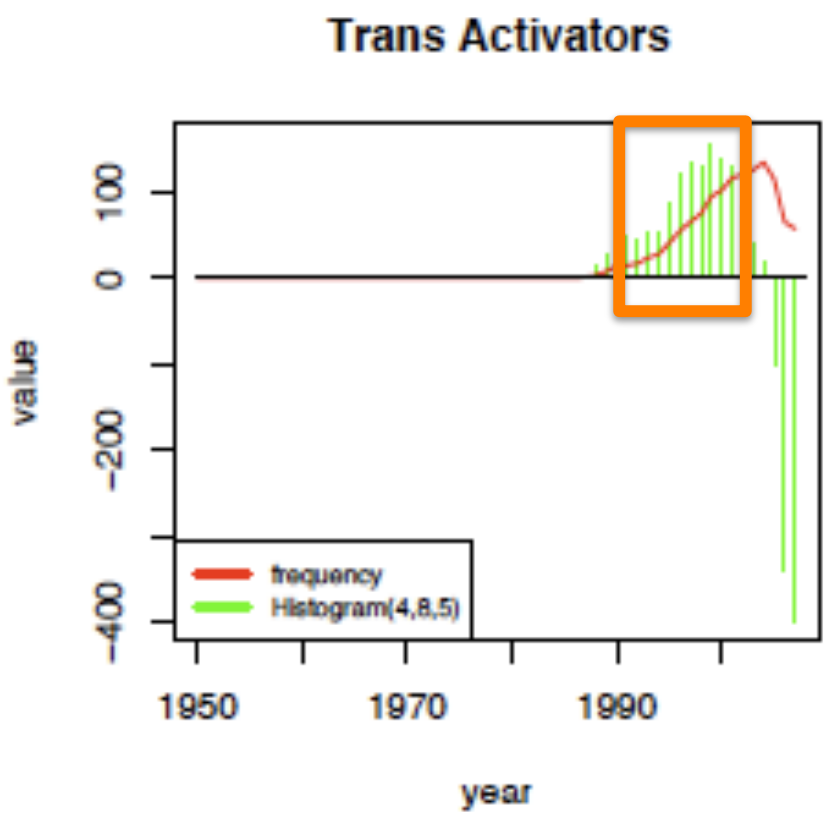
Towards the end of the war, publications on morphine slows, resulting a drop of burst strength

VALIDATION OF TOPIC DYNAMICS MODEL



Trans-Activators: Gene products to regulate the expression of proteins

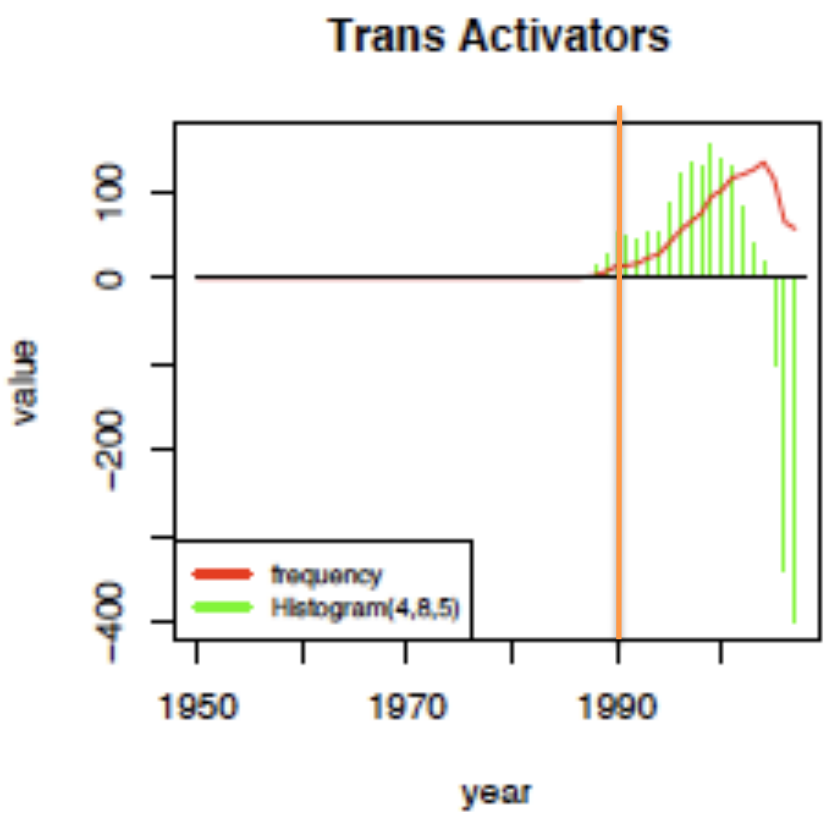
VALIDATION OF TOPIC DYNAMICS MODEL



Trans-Activators: Gene products to regulate the expression of proteins

From the start of the Human Genome Project in 1990 to the cloning of Dolly in 1996, the growth of genetic research generated bursts in DNA-related topics during this period.

VALIDATION OF TOPIC DYNAMICS MODEL

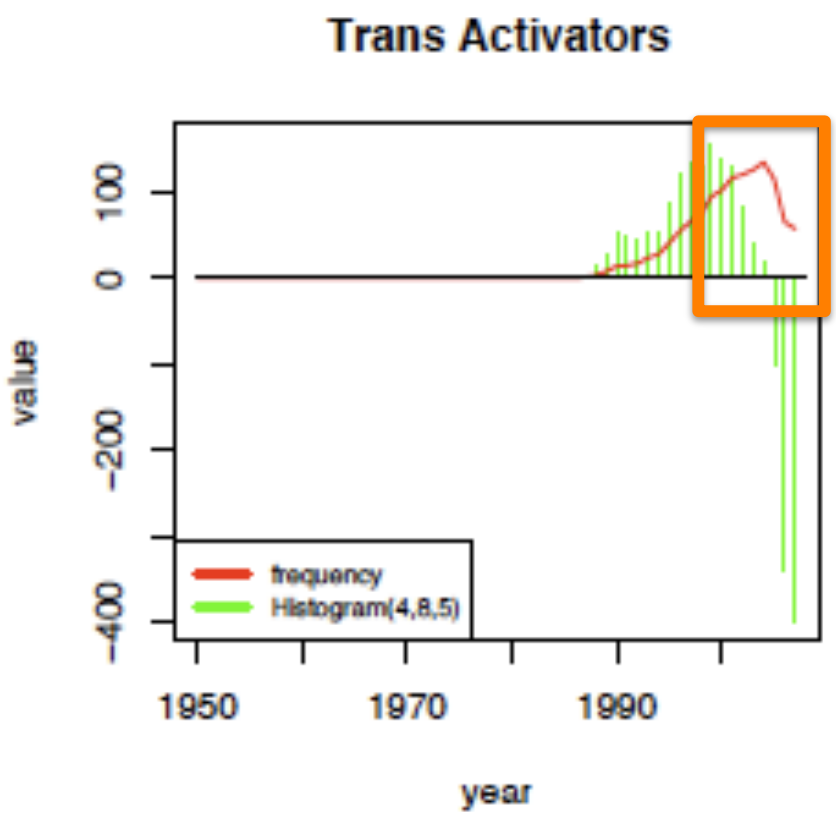


Trans-Activators:

From the start of the Human Genome Project in 1990 to the cloning of Dolly in 1996, the growth of genetic research generated bursts in DNA-related topics during this period.

The burst period starts before 1990.

VALIDATION OF TOPIC DYNAMICS MODEL

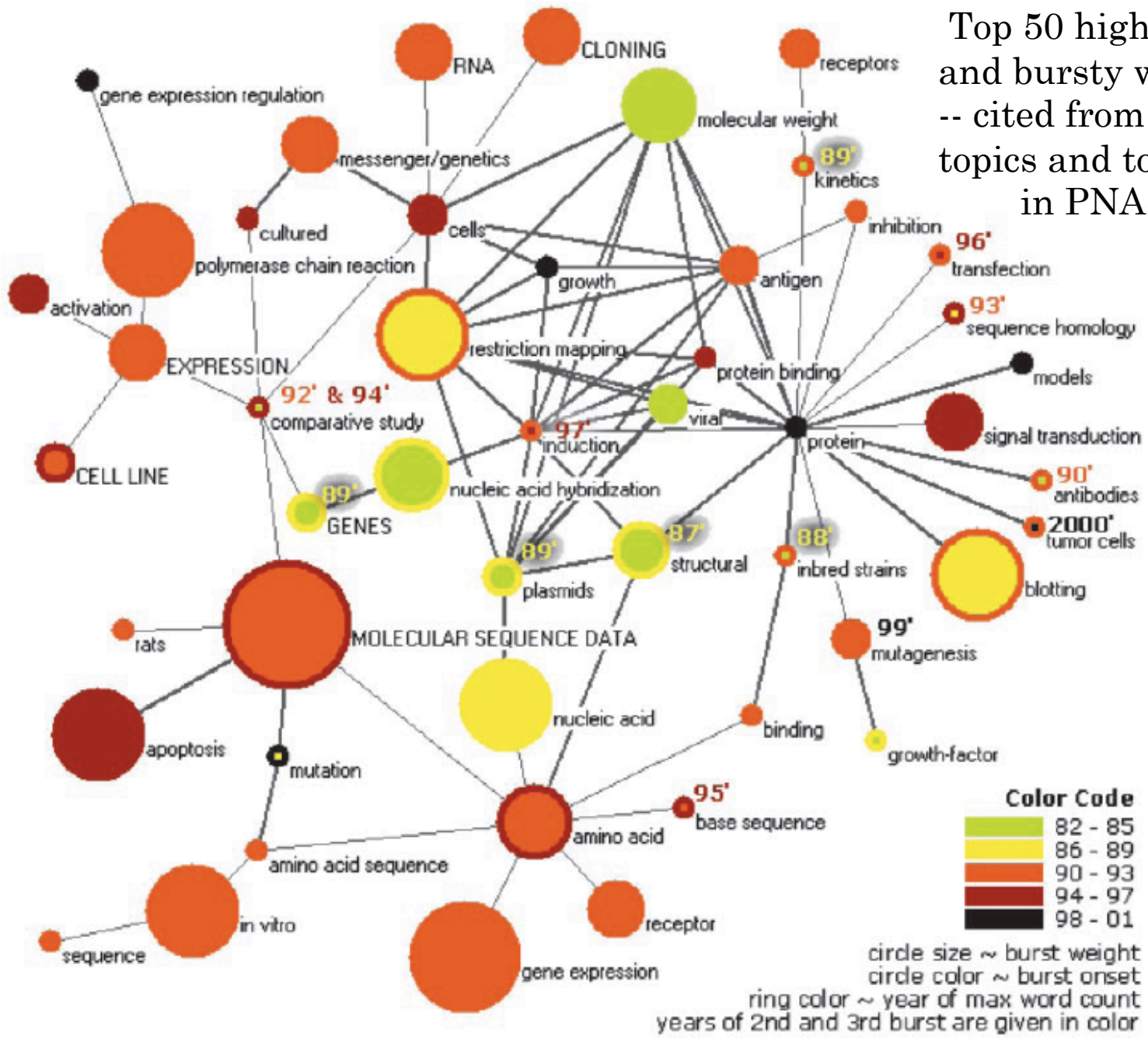


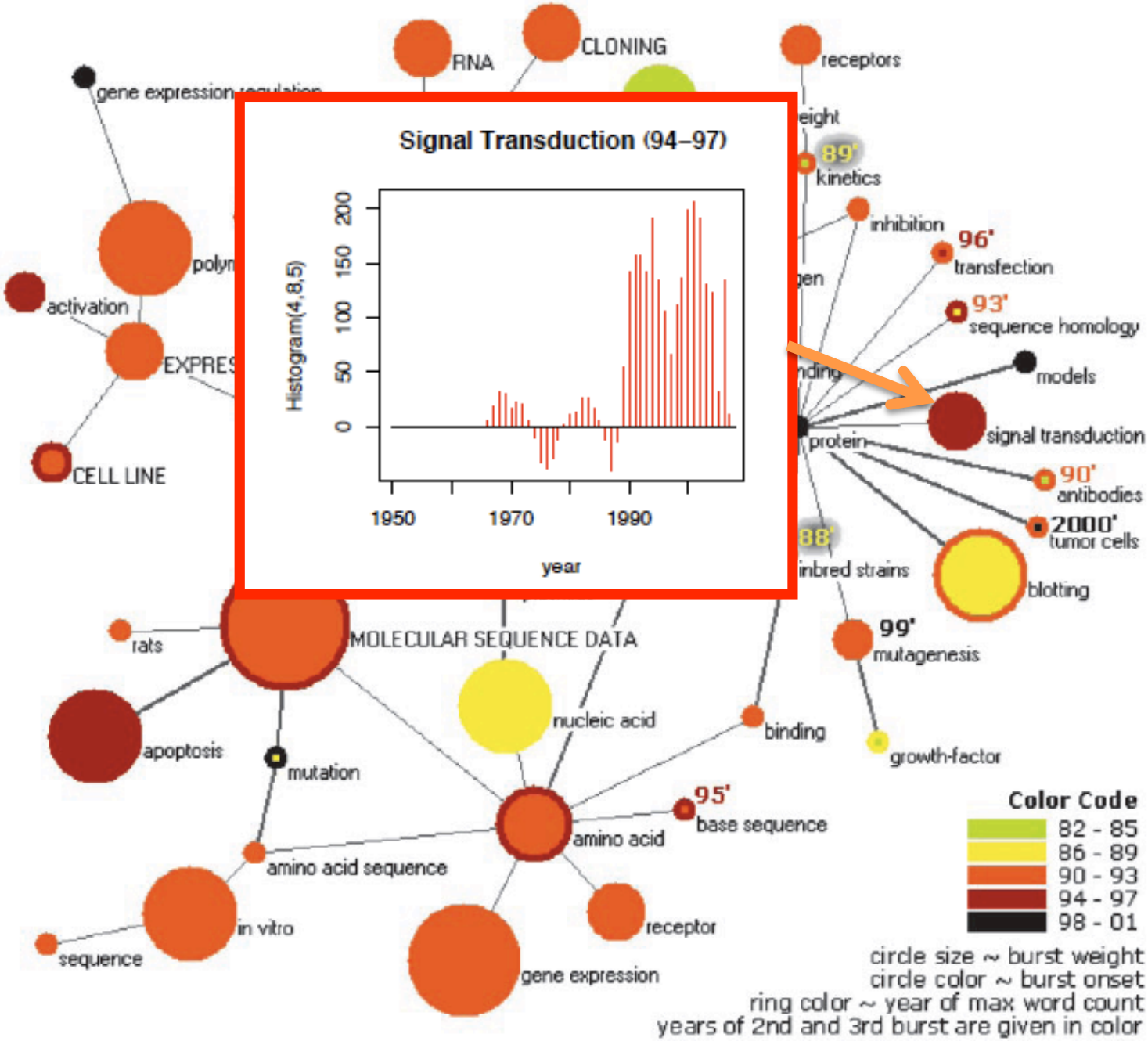
Trans-Activators:

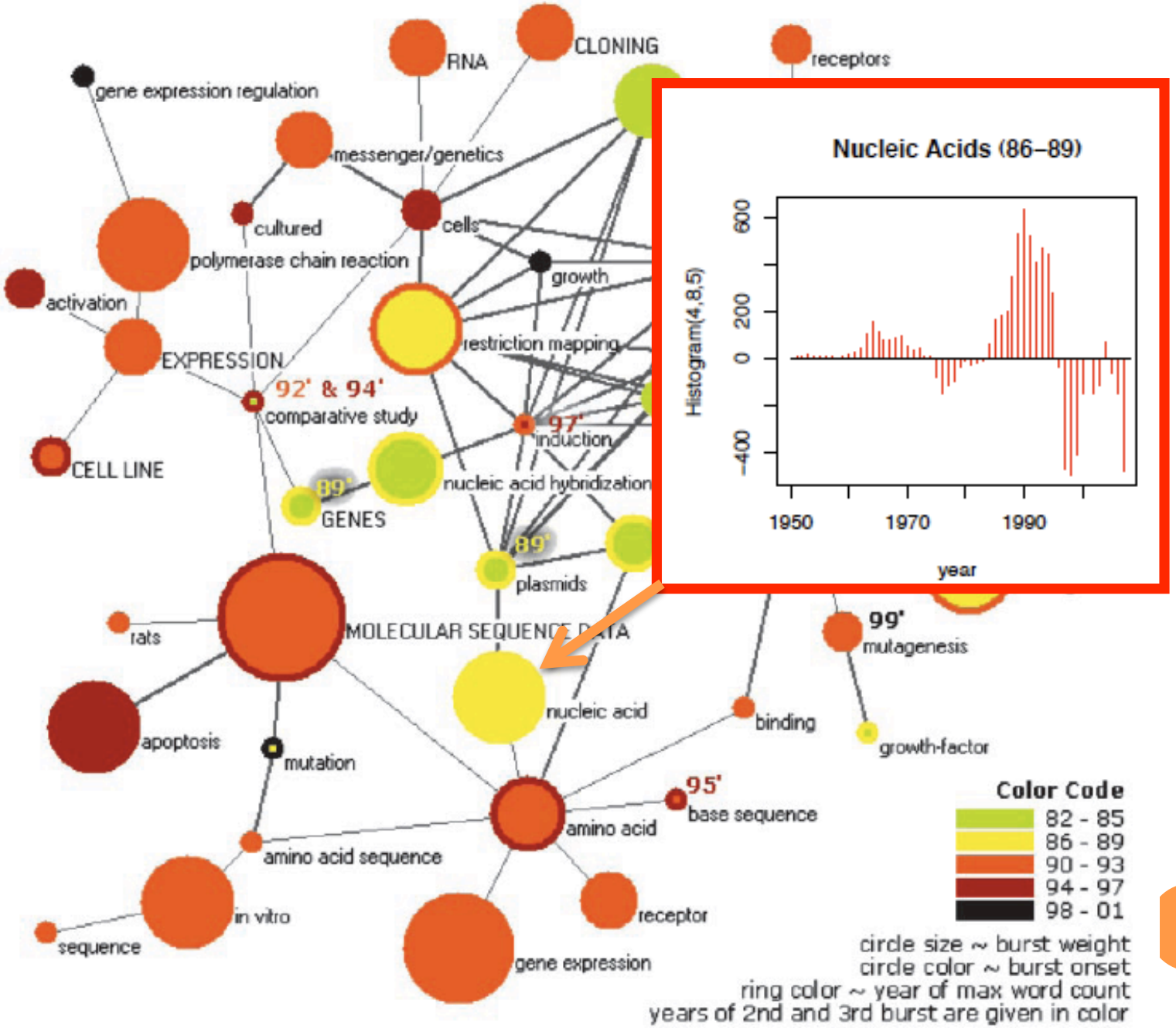
From the start of the Human Genome Project in 1990 to the cloning of Dolly in 1996, the growth of genetic research generated bursts in DNA-related topics during this period.

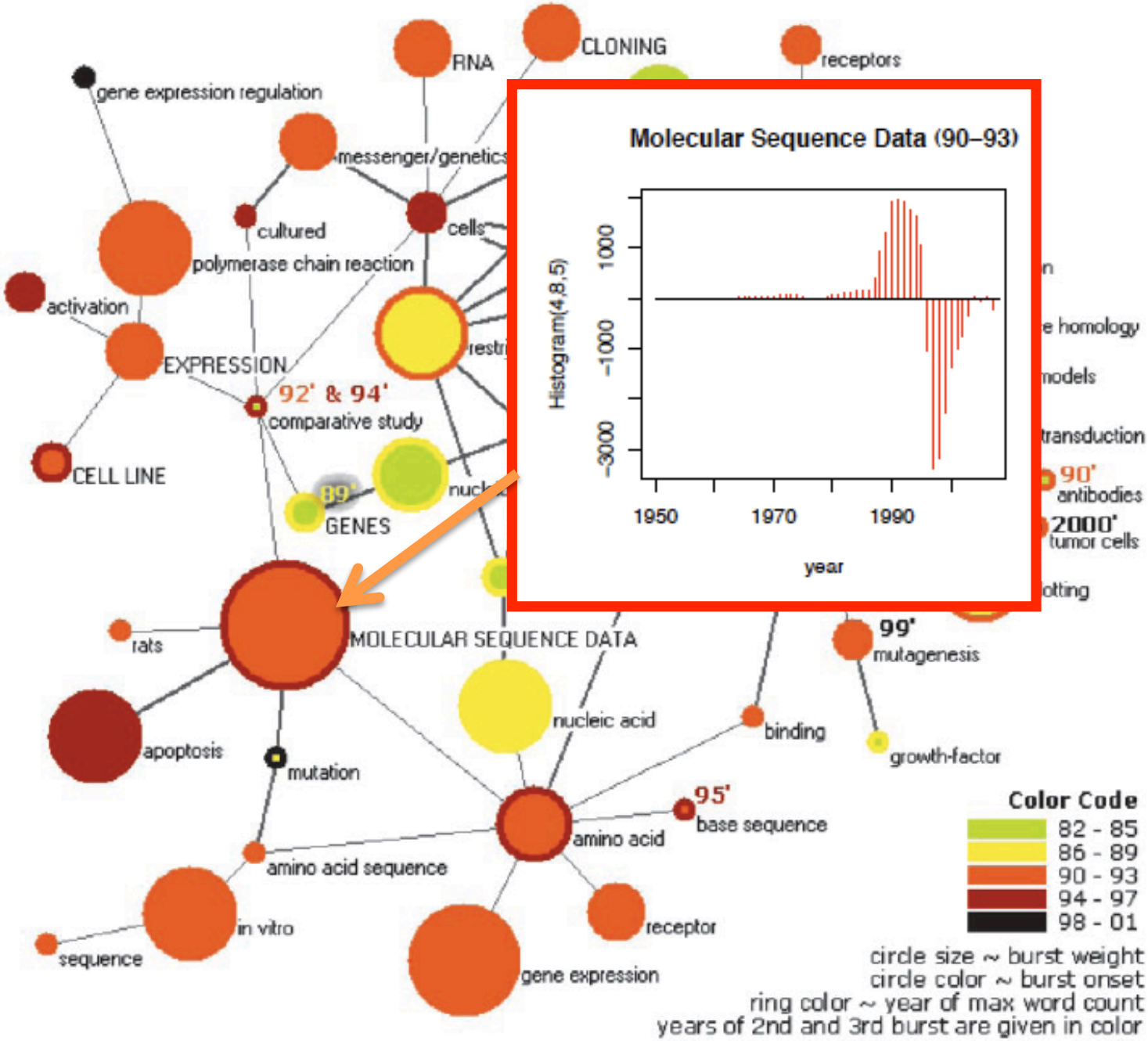
The trend of the histogram clearly precedes the trend of the frequency, suggesting the histogram may be used to forecast oncoming events.

Top 50 highly frequent and bursty words -- cited from "Mapping topics and topic bursts in PNAS"




Signal Transduction (94–97)
Histogram(4,8,5)
0
50
100
150
200
1950
1970
1990
year
RNA
CLONING
receptors
gene expression
89'
kinetics
inhibition
96'
transfection
93'
sequence homology
models
signal transduction
protein
90'
antibodies
2000'
tumor cells
blotting
inbred strains
99'
mutagenesis
growth-factor
binding
activation
CELL LINE
rats
MOLECULAR SEQUENCE DATA
nucleic acid
apoptosis
mutation
amino acid
95'
base sequence
amino acid sequence
in vitro
sequence
receptor
gene expression
Color Code
82 - 85
86 - 89
90 - 93
94 - 97
98 - 01
circle size ~ burst weight
circle color ~ burst onset
ring color ~ year of max word count
years of 2nd and 3rd burst are given in color


RNA
CLONING
receptors
gene expression regulation
messenger/genetics
cultured
cells
polymerase chain reaction
growth
activation
restriction mapping
EXPRESSION
92' & 94'
comparative study
97'
induction
CELL LINE
89'
GENES
nucleic acid hybridization
89'
plasmids
rats
MOLECULAR SEQUENCE DATA
99'
mutagenesis
nucleic acid
binding
apoptosis
mutation
growth-factor
95'
base sequence
amino acid
amino acid sequence
in vitro
receptor
sequence
gene expression
Nucleic Acids (86–89)
Histogram(4,8,5)
600
200
0
−400
1950
1970
1990
year
Color Code
82 - 85
86 - 89
90 - 93
94 - 97
98 - 01
circle size ~ burst weight
circle color ~ burst onset
ring color ~ year of max word count
years of 2nd and 3rd burst are given in color


Molecular Sequence Data (90–93)
Histogram(4,8,5)
1000
−1000
−3000
1950
1970
1990
year
RNA
CLONING
receptors
gene expression regulation
messenger/genetics
cultured
cells
polymerase chain reaction
activation
EXPRESSION
92' & 94'
comparative study
CELL LINE
89'
GENES
MOLECULAR SEQUENCE DATA
rats
apoptosis
mutation
nucleic acid
binding
99'
mutagenesis
growth-factor
95'
base sequence
amino acid
amino acid sequence
in vitro
sequence
receptor
gene expression
e homology
models
transduction
90'
antibodies
2000'
tumor cells
Color Code
82 - 85
86 - 89
90 - 93
94 - 97
98 - 01
circle size ~ burst weight
circle color ~ burst onset
ring color ~ year of max word count
years of 2nd and 3rd burst are given in color

COMPARISON WITH KLEINBERG'S BURST MODEL

- Kleinberg's model focuses on identification of bursts in a stream of mixed topics where it assumes term independence.
- Our Topic Dynamic model focuses on identification of bursts in a stream of single topics. Our model is more appropriate for burst detection in the MeSH hierarchy.

COMPARISON WITH KLEINBERG'S BURST MODEL

- Kleinberg's model can be more reluctant to claim the start of a burst because a cost is associated with the burst state transition.
- Kleinberg's model can be more reluctant to terminate a burst as long as the frequency of the topic is relatively high compared with the frequencies of other topics in the stream.
- Our Topic Dynamics model tends to have shorter burst periods.

word	Bursts (Kleinberg)	Bursts (Our model, S=SIGMOD, B=VLDB)
bases	1975S-1982V	1975V, 1977V, 1978V, 1979V, 1980V, 1982V
object	1990S-1996V	1989S-1992S, 1994S, 2000S
parallel	1989V-1996V	1988S, 1989S-1990V,1991V-1992S, 1993S-1994S, 1995V-1996V
statistical	1981V-1984V	1979V, 1981V-1982V,1987V-1990V
model	1975S-1992V	1975V-1976S, 1978V-1979S, 1981V,1982V, 1983V, 1985S-1985V, 1988S, 1989S, 1990V, 1991V, 1992V
schema	1975V-1980V	1975V-1978V, 1983V
web	1998S-	1995S-1996S, 1997S-1998S, 1999S
approximate	1997V-	1982S-1982V, 1986V-1994S, 1997V-1999S, 2000V
objects	1987V-1992S	1986V-1988S, 1989S-1989V, 1991S, 1995S

Comparison between Kleinberg's burst intervals and our burst intervals on the paper titles from the database conferences SIGMOD and VLDB

EVALUATION ON MACD HISTOGRAM PARAMETERS

word	parameters	Burst Intervals (S = SIGMOD, V= VLDB)
database	(4, 8, 5)	1975V-1976V, 1977V-1979S, 1980S, 1981S …
	(7, 9, 6)	1975V-1981S
truncation	(4, 8, 5)	1990V, 1991V-1992S, 1997V-1998S, 1995V, 2000V
	(7, 9, 6)	1990V-1992S, 1995V, 2000V
parallel	(4, 8, 5)	1991V-1992S, 1993S-1994S, 1995V-1996V, 2001V
	(7, 9, 6)	1991V-1994S, 1995V

Comparison of the burst intervals detected by our model using different MACD histogram parameters on the paper titles from the database conferences SIGMOD and VLDB

MORE ADVANTAGES

- The Topic Dynamics framework may be used to forecast oncoming events.
- The framework takes into account semantic links between topics that are missed by single stream analysis.

CONCLUSIONS

- The topic dynamics model represents an alternative for detecting and analyzing bursts.
- The model works well for tracking topic-bursts of MeSH terms in the biomedical literature.
- Comparison with Kleinberg's burst model
- Advantages over existing burst models:
  - Computationally efficient
  - Can model bursts with two dimensions, not just one
  - Better defined burst strength
  - Incorporate hierarchy of topics
  - May be useful for forecasting oncoming events

THANKS

- Questions?

THANKS

- Questions?

UCLA
Prof. D. Stott Parker

Funding:
NIH, CCB, UCLA scholarship