

Communication and data curation

# V1. Communication issues in data curation

As we have seen: communication is central to almost every area of data curation . . .

And is itself an area of data curation

Scientific communication of data and the results of data analysis  
is an essential part of science, technology, and scholarship,  
with effects in two directions

Of course more data, and more data analysis, is a good thing,  
but the increase is causing a crisis in scientific communication

# Communication and the data curation

As we have seen *communication* is central to almost every area of data curation

*[And it itself an independent area]*

# Areas of curatorial activities

<b>Collection:</b>	Support the collection and acquisition of data
<b>Organization:</b>	Employ an appropriate data model and use appropriate standards
<b>Storage:</b>	Support reliable and effective storage
<b>Preservation:</b>	Ensure that data will be understandable and useable in the future
<b>Discoverability:</b>	Support the ability to search for and locate relevant data
<b>Access:</b>	Support the ability to retrieve and distribute data
<b>Workflow:</b>	Support the ability to systematize data workflows
<b>Identification:</b>	Support the ability to identify, authenticate, and validate data
<b>Integration:</b>	Support integration of data from different sources using different data models
<b>Reformatting:</b>	Support reformatting for use by different tools or to match new format standards
<b>Reproducibility:</b>	Support ability to reproduce results, ensuring scientific validity
<b>Sharing:</b>	Support sharing data between researchers, teams, and institutions.
<b>Communication:</b>	Support representation, publishing, and visualizations that provide insight
<b>Provenance:</b>	Support identifying what inputs and calculations are responsible for data values
<b>Modification:</b>	Support management of corrections and updates
<b>Compliance:</b>	Ensure compliance to legal, regulatory, and local policy requirements
<b>Security:</b>	Ensure that data is secure from tampering or inappropriate access and distribution

# Our definition of data science (again)

Data science is concerned with all aspects of  
the **creation, management, analysis**, and **communication** of data  
focusing in particular on  
the application of *computational methods* to *digital data*

The data science objective: *extracting useful knowledge from data*

# Methods of curatorial action

There are many methods and techniques employed to achieve the objectives just listed, but five categories stand out as particularly important:

## **Analysis**

To determine needs, and develop relevant data models and *metadata*, and reformat, correct, or update data.

## **Documentation**

To record essential information (typically via *metadata*)

## **System design and implementation**

To support all data curatorial activities

To support the generation and use of data documentation and processing documentation

## **Policy**

To specify objectives, procedures, practices, and formats.

## **Process**

To ensure success and efficiency by managing the development of appropriate organizational units and roles, providing training, advocating for change, and managing curatorial activities.

# What we will take up next

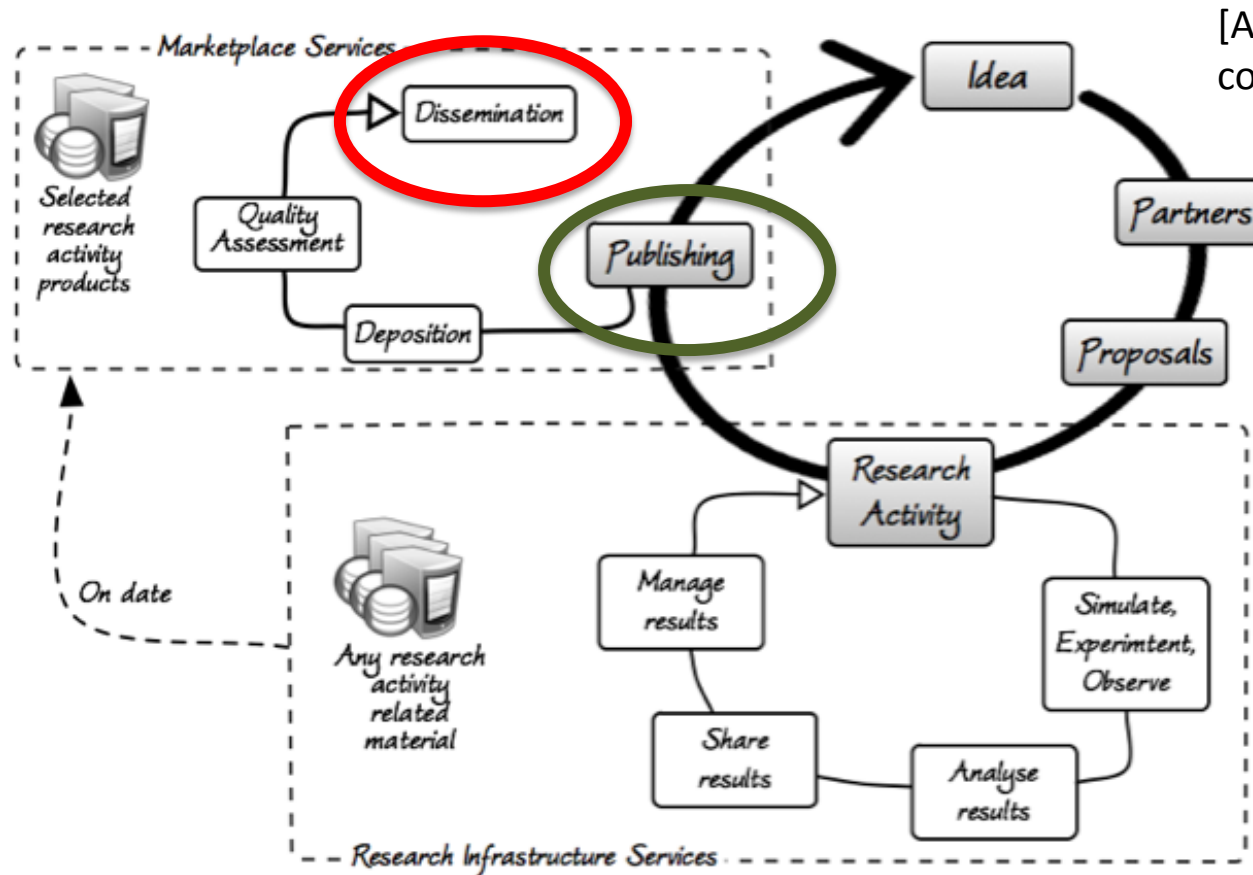
Ok, communication is relevant to many aspects of data curation

But we have a particular focus this week:

*Scientific and technical communication*

particularly journal literature, where scientific results are reported

# Scientific communication is how data gets noticed



[After all, if the results of analyzing data are not communicated, then what's the point of it all?]

Scientific and technical communication is a critical part of the data lifecycle, with effects flowing both ways:

- from the research process,
- and back into the research process.



. . . the crisis

*But scientific and technical publishing is in crisis*

as we'll see in the next video

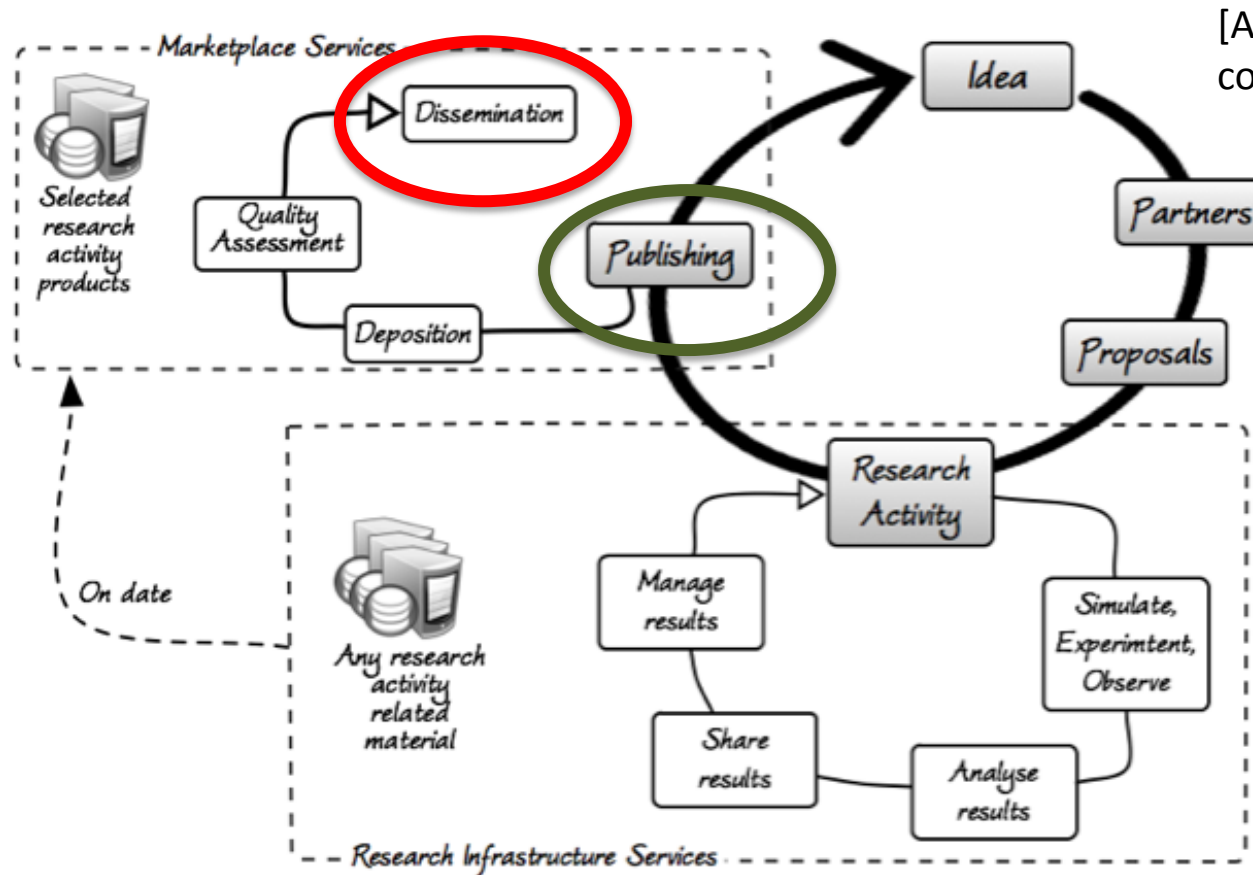
# V2. The crisis in data-driven scientific communication

A story about Lisa

The scientific literature explosion

Solutions that won't (completely) work

# Scientific communication is how data gets noticed



[After all, if the results of analyzing data are not communicated, then what's the point of it all?]

Scientific and technical communication is a critical part of the data lifecycle, with effects flowing both ways:

- from the research process,
- and back into the research process.

. . . the crisis

*But scientific and technical publishing is in crisis*

a problem caused by data  
and that can be addressed with data

as we'll see in the next video

# Introducing Lisa (DOB: January 1, 2000)

Today she is 17, just starting college

All her life she has been using the Web, Google, FB, smart phones...

*Now let's look ahead just 8 years, to **2026***

Lisa has just finished her doctoral coursework in molecular biology,  
and she is about to start her research.

*She walks up to the science reference desk...*

# Lisa at the science reference desk in 2026

*Why is she there?*

- Does she need to know some fact?
- Does she need to find a resource?
- Does she need to know how to use a resource?

She begins:

“I’ m studying the role of the P53 in Huntington’ s disease... ”\*

The reference librarian interrupts...

“So you’ d like to find some articles to read on P53?”

[ *they both laugh* ]

**Why do they laugh?**

[\*]Inspired by John Wilbanks, various presentations 2005-2007;  
and comments by MacKenzie Smith, Associate Director, MIT Libraries

*Because the reference librarian is making a joke.*

In 2026  
*no one is “looking for articles to read”*  
(at least not in science)

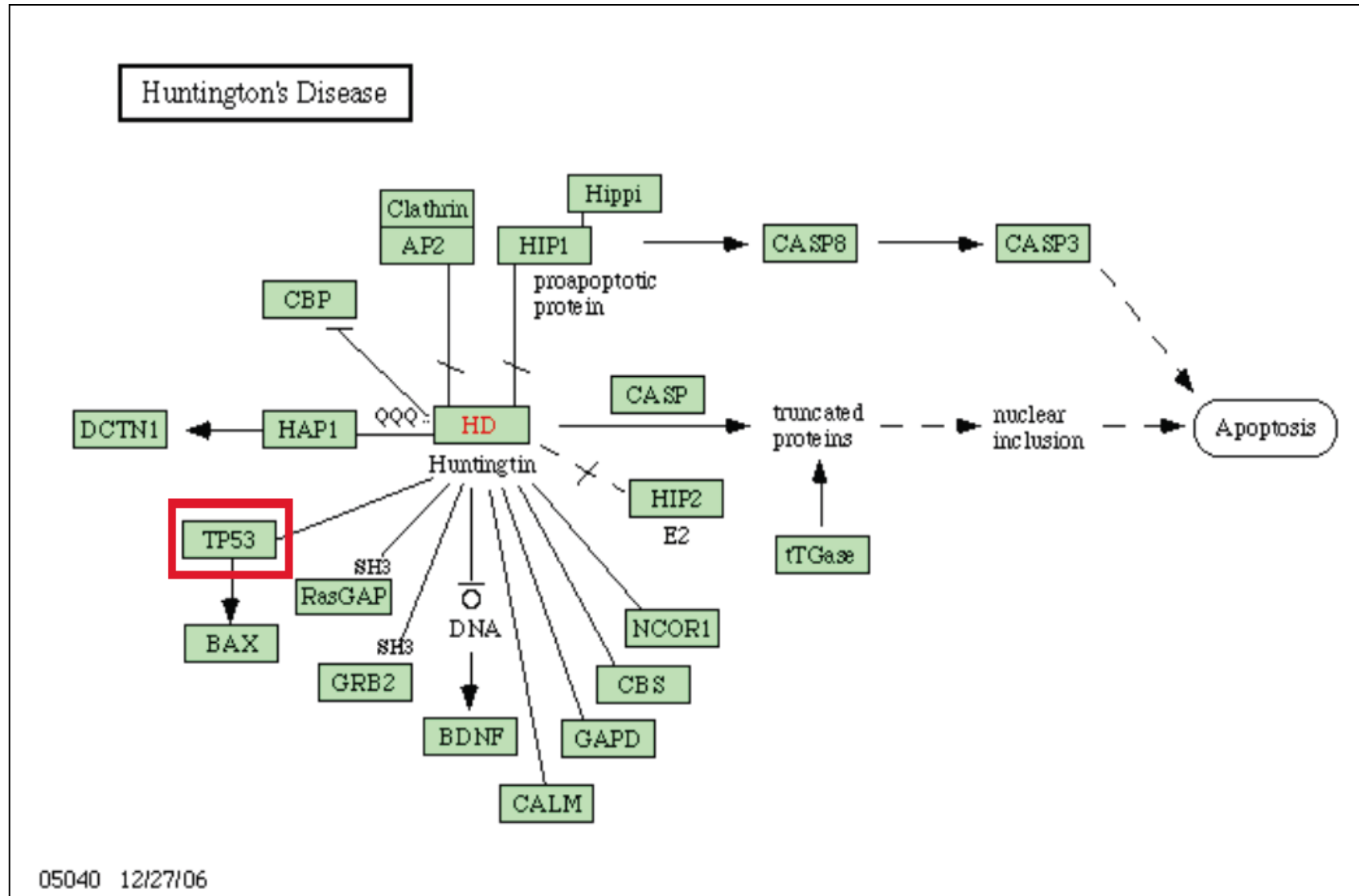
in 2026 engaging with the scientific literature  
will (*finally*) be like

*“flying a jet plane through information space”\**

and not at all like *finding and reading* articles

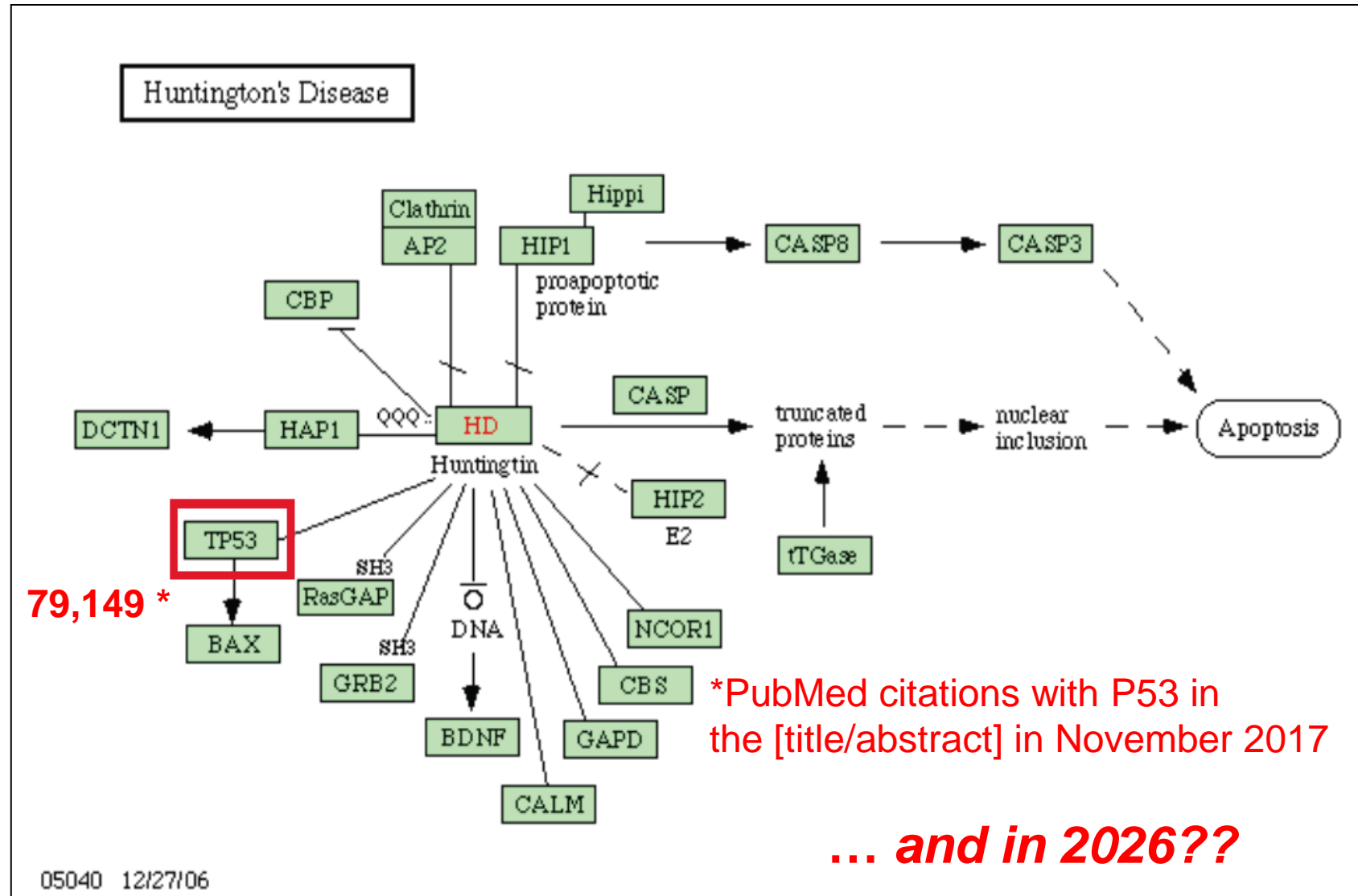
\*attributed to Alan Kay, mid-1980s,

# Lisa's problem: P53





# Lisa's problem



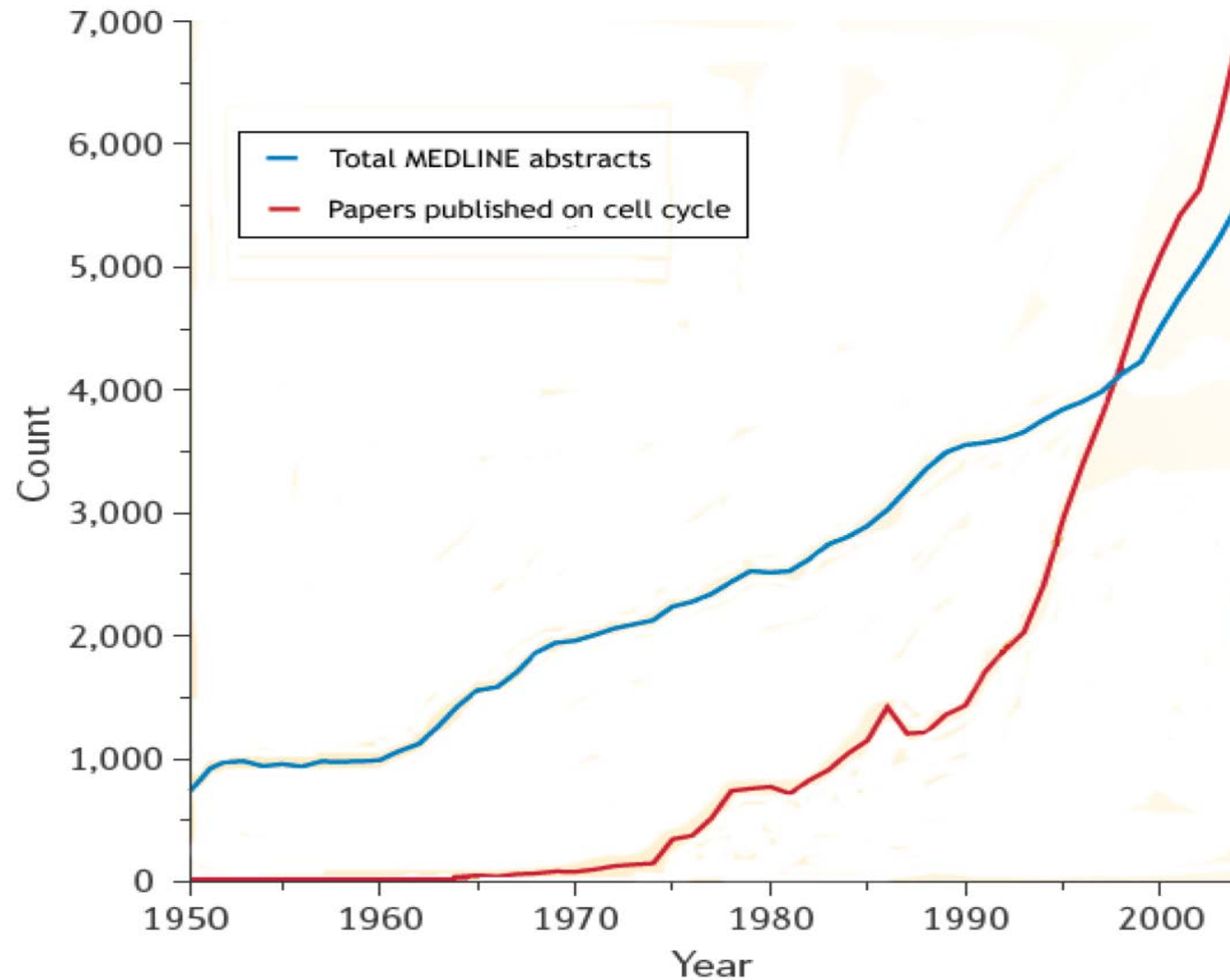
# A tipping point has been reached

“Nowadays ... sets of relevant papers [are] identified that surpass human capability for reading, interpretation, and synthesis.”

— Barend Mons  
“Which gene did you mean?”

This is the problem that contemporary data generation and tools has created.

# *Are you kidding me???*



[Axis is  $\times 10^{-2}$  for total Medline abstracts]

Adapted from Jensen, Saric, & Bork; *Nature* (2006).

# Responses to the problem

**One response:** *text mining* instead of *reading*

- » information extraction
  - » “undiscovered public knowledge”  
and hypothesis generation
- (Swanson and Smalheiser)

**Another response:** *tools for strategic reading*

# V3. The solution to the data crisis is . . . *more* data

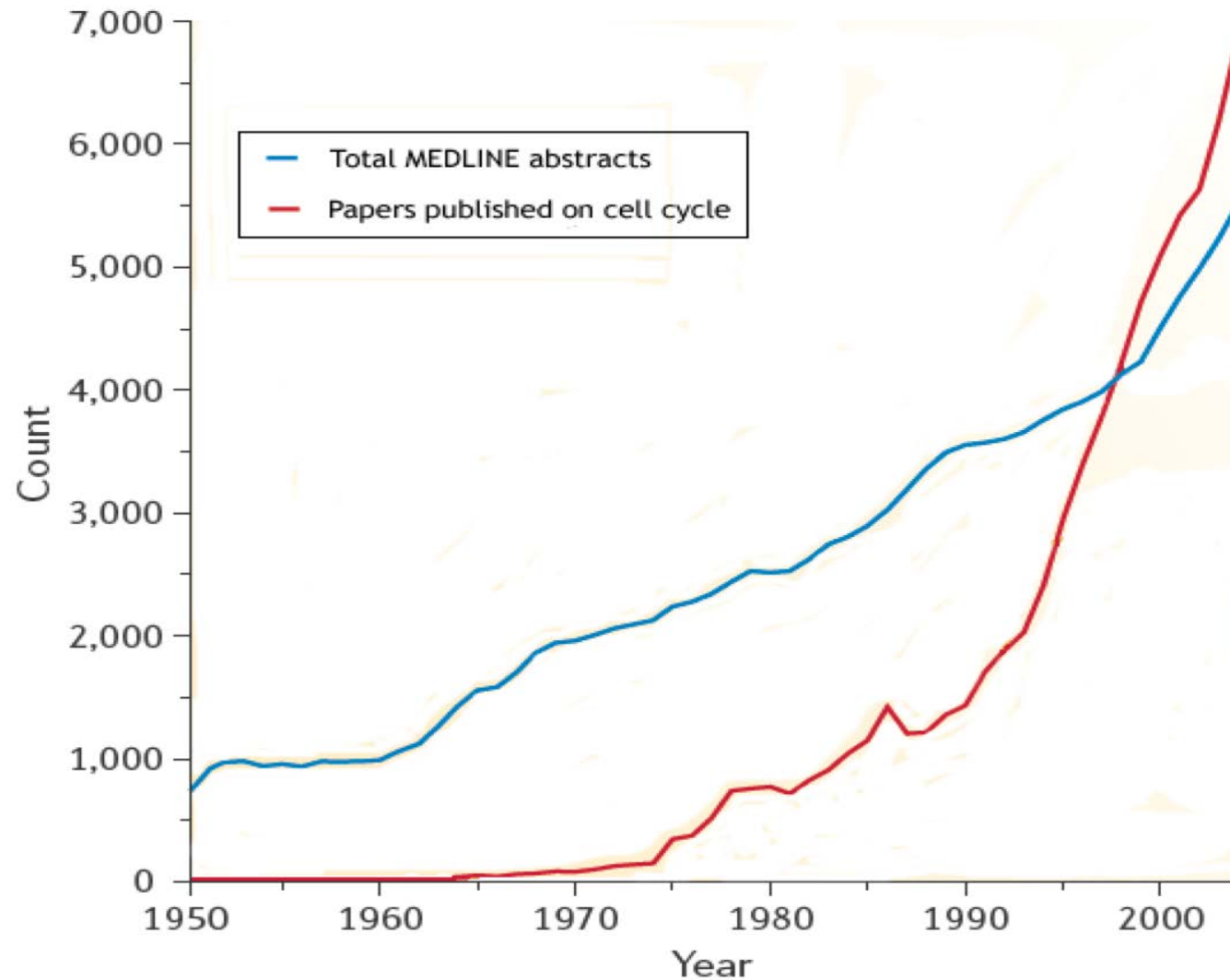
Recall the problem: more relevant articles than any specialist could possibly read

Background: How *do* scientists and other professionals read these days?

How could we help them read better?

Examples, obstacles, the future

# *Are you kidding me???*



[Axis is  $\times 10^{-2}$  for total Medline abstracts]

Adapted from Jensen, Saric, & Bork; *Nature* (2006).

# Some big numbers

- 25,000 publishers
- 1.5 million scholarly articles a year
- 600,000 abstracts added to Medline each year

So, how do scientists and other professionals engage the literature



# User behavior during (scientific) search

in, e.g., PubMed, SCOPUS, Google Scholar, Web of Science, ADS, etc.

The *search trance*.

Researchers engage with the literature as if playing a video game

They rapidly, almost subconsciously...

- develop queries likely to find known items, or retrieve subject or topic results
- track references backward and citations forward,
- make rapid relevance judgments: assessments of impact, and quality
- locate and compare key terms, equations, definitions, protocols, findings

This is almost sub-cognitive, kinaesthetic, even trance-like,

sessions are often considered successful

— *even though no article to read was found and read!!*

Their goal appears to be *not* finding an article to read. . .

but avoiding reading

Now, this is nothing new...

- indexing and citation analysis help decide whether articles are relevant...  
*without reading them.*
- abstracts and literature reviews help us take advantage of articles ...  
*without reading them.*
- the articles we do read, in their analyses and summaries help us take advantage of other articles  
... *without reading them.*
- friends, colleagues, and, best of all, graduate students, help us take advantage of articles ...  
*without reading them.*

## Paper based strategic reading

... engineers describe a common pattern for utilizing document components by zooming ... and filtering information ...

[they] first read the abstract, then skim section headings.

Next ... lists, summary statements, definitions, and illustrations.

... they disaggregate and re-aggregate article components for use in their own work ... perhaps by using a marker to highlight ... perhaps by creating a mental register...

— Bruce Schatz et al.  
“Federated Search of Scientific Literature”  
*IEEE Computer*, 1999.

## Online strategic reading

*[informant]* ...

I used the sections of the papers for the equations....

I even wouldn't read all the other parts of the article ...

I look for specific surface tensions, experimental measurements ...

I sometimes need to look specifically at other methods and theories.

— Ann Bishop  
“Document Structure and Digital Libraries:  
How Researchers Mobilize Information in Journal Articles”  
*Information Processing and Management*, 1999.

# Longstanding behaviors, sure: but newly urgent

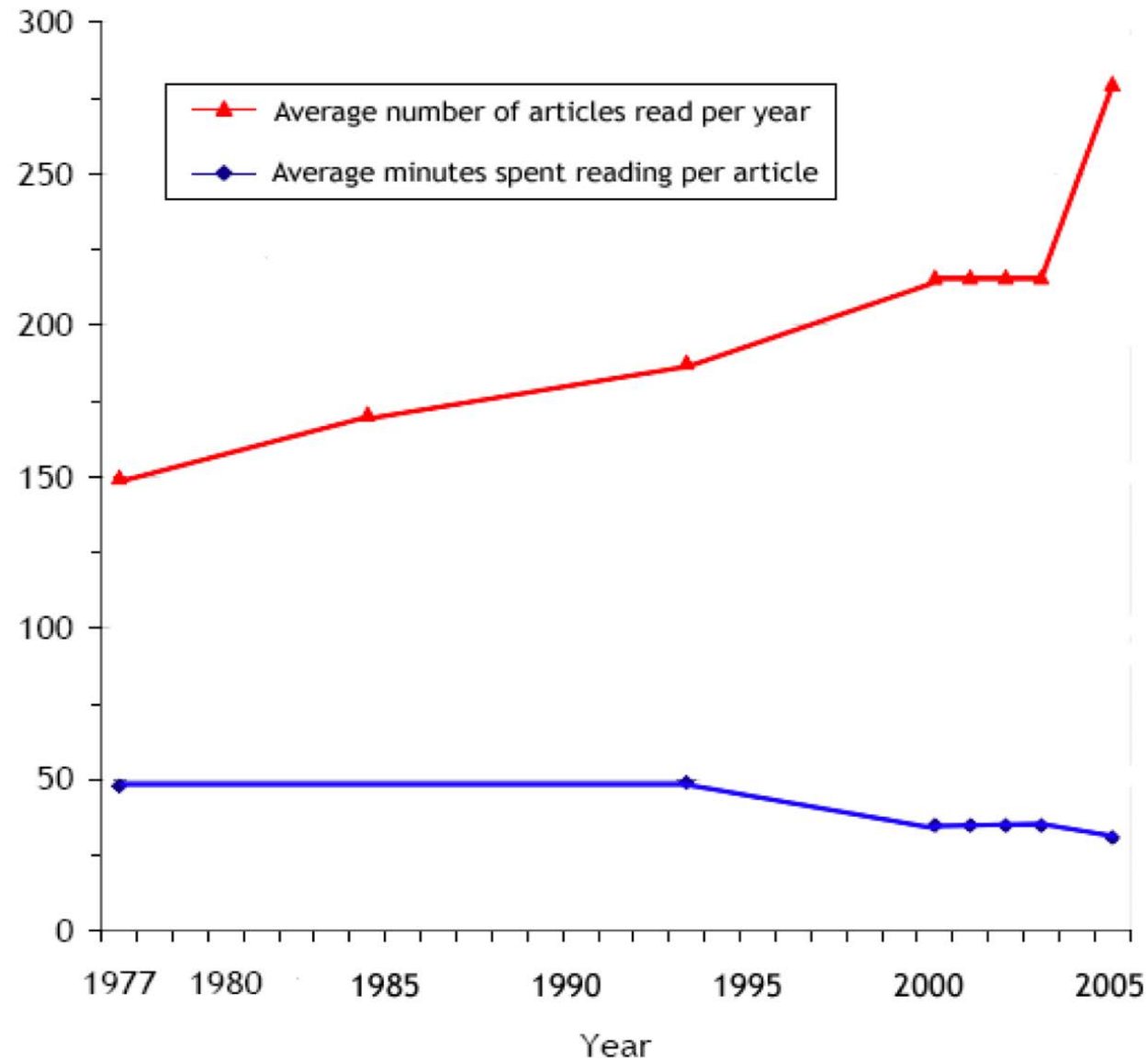
Key studies from C. Tenopir and D. King (2000-2007) show ...

- time *searching* and *browsing* has been rising rapidly from 1984 to 2000.
- until the mid-1990s the number of articles “read” was more or less steady.
- but since then the number “read” (*not “browsed”*) has been climbing
- and so *reading time per article is dropping*, in some fields fairly fast (< 24 minutes!)

In addition studies from C. Palmer et al (2000-2007) show ...

- researchers use sophisticated techniques to mobilize information according to varied research needs.
- these information needs vary with discipline, research life cycle (local and global), and research strategies, as well as with varying affordances of current technology
- the pace of evolution and innovation in these techniques is increasing, and increasingly driven by researchers themselves and not information specialists or publishers

# *Faster, faster, faster, more more more*



Tenopir et al. 1977-2005

# So what to do??

Human reading of natural language prose is uniquely valuable, providing nuance, clarity, and insight.

So reading cannot be totally replaced by text mining

So let's provide tools that support *strategic reading*

For example

- computationally available data items accessible with discipline-specific tools  
(chemical formulae, proteins, equations, etc.)
  - advanced navigation and viewing  
optimized for domain-specific and objective-specific browsing and analysis,
  - typed hypertext linking with links as first class objects,
  - data-driven interactive diagrams and graphics
  - computable equations,
  - supportive ontological inferencing
  - thoroughgoing interoperability with other tools
- ... and so on, and on, and on.

the grand old dream of radical new functionality as envisioned by  
Paul Otlet, Vannevar Bush, Douglas Engelbart, Ted Nelson et al.

## More of what scientists want. . .

The *datument* is a hypermedia document accessible to robots and humans

... for transmitting “complete” information including content and behaviour.

... the machine is ... semantically aware of the document content [through] domain-specific XML components...

We argue that a cultural change in our approach to information is needed.

P. Murray-Rust and H.S. Rzepa[\*]

“The Next Big Thing: From Hypermedia to Datuments,”

*Journal of Digital Information*, 5:1 2004

[\*]Chemistry faculty at Cambridge University and  
Imperial College London, respectively.

Imagine what could be achieved if articles, rather than consisting entirely of free-form natural languages, contained explicit assertions about biological knowledge in unambiguous machine readable form ... some progress is being made...

Mathew Cockerill, Editorial,  
*BMC Bioinformatics*, 6:140 2005



# In a nutshell

As scientific ontologies are integrated into the publishing workflow  
many enhancements to scientific communication will become possible

...support for text mining, information extraction, and literature-based discovery.

And one is not so obvious,

... support for the long-standing practice of *strategic reading*

“Strategic Reading and the Future of Scientific Publishing”  
Allen H. Renear, Carole L. Palmer,  
*Science*, August 16 2009.

# Necessary data standards are now, finally, in place

## **Character** encoding interoperability

Unicode/UTF-xx

[Adoption: nearly total]

## **Data structure serialization** interoperability

XML, JSON

[Adoption: nearly total]

## **Syntactic** interoperability

i.e. RDF(S), OWL

[Adoption: underway]

## **Semantic** interoperability

RDF/OWL ontologies; linked data.

[Adoption: substantial]

## **Document markup meta-languages**

XML

[Adoption: nearly total]

## **Document markup languages**

e.g, NLM/DTD, XHTML, TEI, DocBook, DITA

[Adoption: widely adopted]

## **Metaphysical** interoperability

“upper” ontologies

[Adoption: (hard to say)]

## **Domain ontologies and terminologies**

hundreds

[Adoption: steady improvements]

Some examples . . .

## Search Interface

**Keywords ?**

depression

☐ Exact match ☐ Case sensitive

**Categories ?**

List >

response to stimulus (GO)

Select category 2 from list above

Select category 3 from list above

Select category 4 from list above

Advanced Search Options : [on](#) | [off](#) Location (abstract)

## Categories/Ontology

Term	Variants
response to stimulus	Response to stimulus, response to stimuluses, Response to stimulus
physiological response to stimulus	Physiological response to stimulus, physiological response to stimuli
response to external stimulus	Response to external stimulus, response to external stimuluses, Res
response to environmental stimulus	Response to environmental stimulus, response to environmental sti
response to stress	Response to stress, response to stresses, Response to stresses
response to endogenous stimulus	Response to endogenous stimulus, response to endogenous stimulu

# Textpresso for Neuroscience

perception of stimulus	Perception of stimulus, perception of stimuluses, Perception of stim
stimulus detection	Stimulus detection, stimulus detections, Stimulus detections
stimulus sensing	Stimulus sensing, stimulus sensings, Stimulus sensings

5 matches found in 3 documents. Search time: 0.085 seconds.

Global links/files: [all results in endnote](#) [all results in print version](#) [all results in xml](#)

Score: 4.00

**Title:** Soluble oligomers of beta amyloid ( 1-42 ) inhibit long-term potentiation but not long-term depression in rat dentate gyrus .

**Authors:** Wang HW Pasternak JF Kuo H Ristic H Lambert MP Chromy B Viola KL Klein WL Stine WB Krafft GA Trommer BL

**Journal:** Brain Res

**Year:** 2002

☐ Bibliographic Information

☐ Abstract

☐ Matching Sentences

**Sen. 18:** Longterm potentiation ( LTP ) and long-term depression ( LTD ) are complementary cellular models of learning and memory that constitute an attractive means of detecting per 134 H . -W Wang et al / Brain Research 924 ( 2002 ) 133 140 turbations of synaptic functioning in the absence of overt neuronal death . [Field: body, subscore: 2.00]

**Sen. 3:** We therefore examined the effects of soluble oligomers of Ab potentiation ( LTP ) and long-term depression ( LTD ) , two cellular models of memory , in the dentate gyrus of rat hippocampal slices . [Field: body, subscore: 1.00]

**Sen. 3:** We therefore examined the effects of soluble oligomers of A beta ( 1-42 ) on long-term potentiation ( LTP ) and long-term depression ( LTD ) , two cellular models of memory , in the dentate gyrus of rat hippocampal slices . [Field: abstract, subscore: 1.00]

Supplemental links/files: [reference in endnote](#) [reference in xml](#) [online text](#) [related articles](#) [Pubmed citation](#)

## Results Set

Hoffmann, R; Valencia, A (Jul 2004).  
"A gene network for navigating the literature."  
*Nature Genetics*. **36** (7): 664.

**iHOP**  
Information Hyperlinked  
Over Proteins

Search for a gene synonym or accession number...

in

Search  
Interface

Symbol	Name	Synonyms
<b>SNF1</b>	AMP-activated serine/threonine protein kinase found in a complex containing Snf4p and members of the Sip1p/Sip2p/Gal83p family; required for transcription of glucose-repressed genes,...	Carbon catabolite derepressing protein kinase, CAT1, CCR1, D8035.20, GLC2, HAF3, PAS14, YDR477W
WikiGenes	<a href="#">edit this page</a> <b>new</b>	
UniProt	<a href="#">P06782</a>	
IntAct	<a href="#">P06782</a>	
PDB Structure	<a href="#">2FH9, 2QLV</a>	
NCBI Gene	<a href="#">852088</a>	
NCBI RefSeq	<a href="#">NF_010765</a>	
NCBI UniGene	<a href="#">852088</a>	
<a href="#">Homologues of SNF1 ...</a>		
<a href="#">Definitions for SNF1 ...</a>		
<a href="#">Most recent information for SNF1 ...</a>		
<a href="#">Enhanced PubMed/Google query ...</a>		

Results Set

Sentences in this view contain interactions of SNF1 - Interaction Information is available whenever you see this symbol - [Read more.](#)

For a summary overview of the information in this page [click here.](#) **new**

We show that **SNF4** binds to the **SNF1** regulatory domain in low **glucose**, whereas in high **glucose** the regulatory domain binds to the kinase domain of **SNF1** itself. [1996] **new**

We first show that the fraction of cellular **Snf4** protein that is **complexed** with **Snf1** is reduced in a sip1 delta sip2 delta gal83 delta triple mutant. [1997]

This **gene activation** depended on the previously identified derepression genes **CAT1** (**SNF1** ) (encoding a protein kinase) and **CAT3** (**SNF4** ) (probably encoding a subunit of Cat1p [**Snf1** p]). [1995]

The **SNF4** -beta-galactosidase protein **coimmunoprecipitated** with the **SNF1** protein kinase, thus providing evidence for the physical association of the two proteins. [1989]

Increased **SNF1** **gene dosage** partially compensates for a mutation in **SNF4** , and the **SNF4** **function** is required for maximal **SNF1** protein kinase

**MeSH-Term** - [Click for options...](#)

We have here addressed the role of the **Snf4** **SNF1** protein kinase in response to **glucose** availability in **Saccharomyces**

Regulation of **Snf1** kinase. Activation requires **Snf1** subunit. [2001]

Yeast **Snf1** is a prototype of activating kinase subunits of eukaryotic **Snf1** (AMPK-related protein kinases) controlling **glucose** and stress

**Hyperlink** - This term has been predicted to be a gene:  
**SNF4** - "Protein kinase activator found in a complex containing Snf1p and members of..." (Saccharomyces cerevisiae)

**Evidence from large-scale screens for interactions between SNF1 and...**  
- **SNF4** (from: TAP & HMS & IntAct)

Muller HM, Kenny EE, Sternberg PW  
"Textpresso: an ontology-based  
information retrieval and extraction  
system for biological literature"  
*PLoS Biol.* 2004 Nov;2(11)..