# Workflow and Provenance

(anything profound, and the cool slides, is from Bertram Ludäscher. Everything else is from Renear

# What is data workflow?

Much of our work with data, especially in scientific applications, consists in *transforming one data set into another*

# Data curation and data workflow

Data curation is concerned with transformations in *two* ways:

managing and documenting transformations involved in data analytics

performing transformation to realize data curation objectives.

(preservation, integration, format conversion, etc.)

# Kinds of data transformations

**Transformations where input and output datasets are identical in propositional content**

transformation to a different data description language (or new version of a language)

transformation to a different serialization (or new version of a serialization)

**Transformations where the input dataset mathematically contains the output dataset**

transformation to a subset matching specific conditions

e.g. simple queries

transformation to a logically or mathematically entailed data of the same kind

e.g., summaries, statistics, visualizations

**Transformations where the input dataset scientifically contains in the output dataset**

transformation to scientifically entailed data of the same kind

here the resulting data set typically contains information different in kind

e.g., a data set about air pressure is transformed to a dataset about altitudes.

# Example Bioinformatics Workflow:

# *Motif-Catcher*

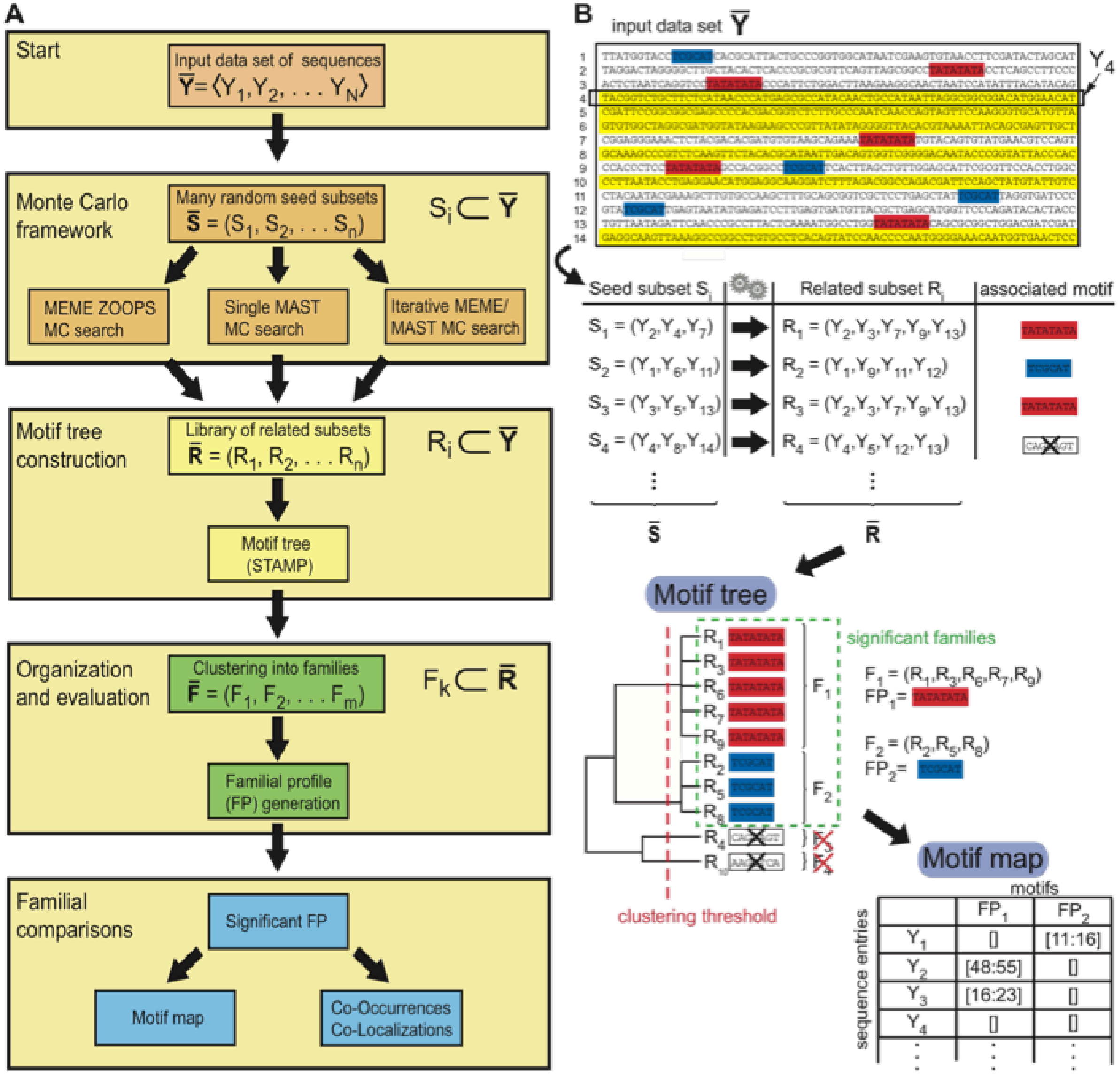


Figure 1: Concept of Monte-Carlo based detection and interpretation of motifs. A) Abstract description of MotifCatcher process. B) Examples illustrating the process with sample data.

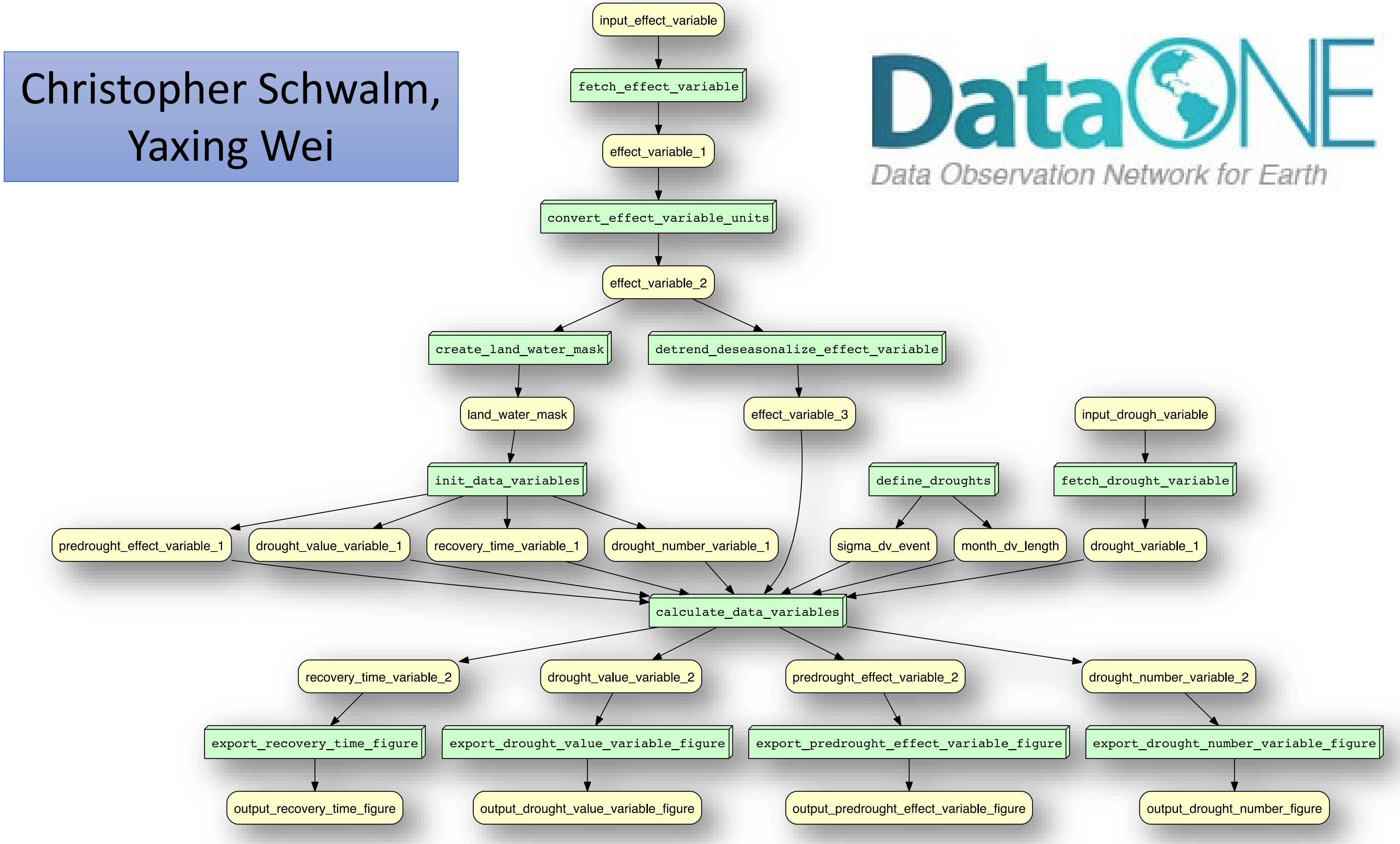
Marc Facciotti *et al.*

UC Davis Genome Center

# Multi-Scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP)

Christopher Schwalm,
Yaxing Wei

DataONE
Data Observation Network for Earth



input_effect_variable

fetch_effect_variable

effect_variable_1

convert_effect_variable_units

effect_variable_2

create_land_water_mask

detrend_deseasonalize_effect_variable

land_water_mask

effect_variable_3

input_drough_variable

init_data_variables

define_droughts

fetch_drought_variable

predrought_effect_variable_1

drought_value_variable_1

recovery_time_variable_1

drought_number_variable_1

sigma_dv_event

month_dv_length

drought_variable_1

calculate_data_variables

recovery_time_variable_2

drought_value_variable_2

predrought_effect_variable_2

drought_number_variable_2

export_recovery_time_figure

export_drought_value_variable_figure

export_predrought_effect_variable_figure

export_drought_number_variable_figure

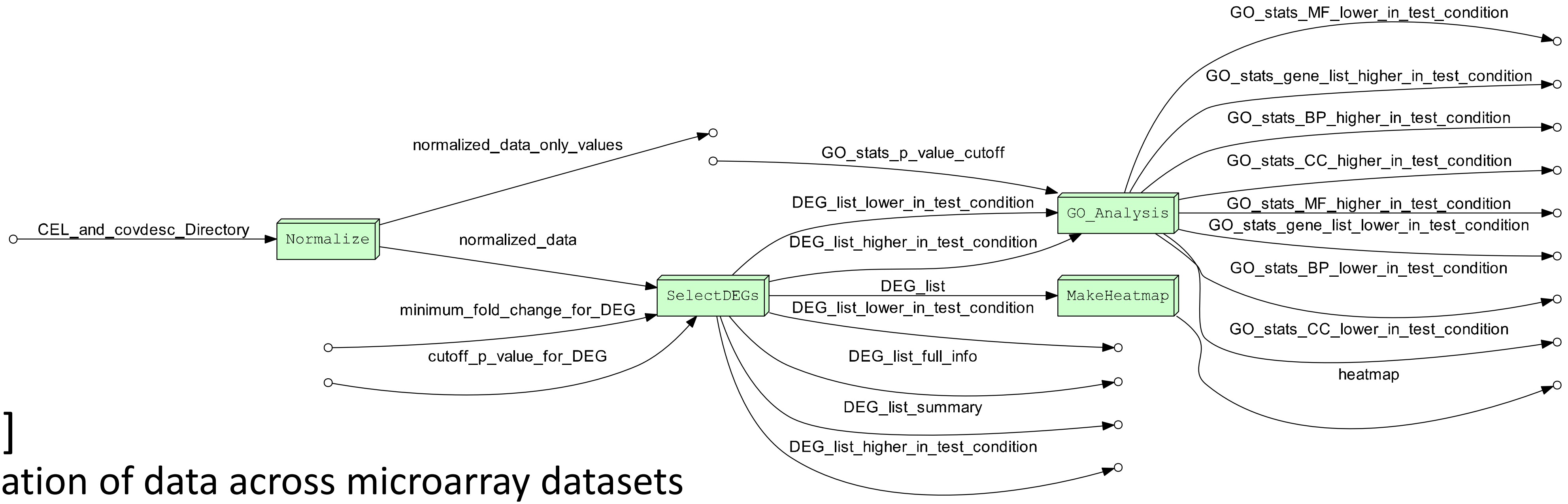
output_recovery_time_figure

output_drought_value_variable_figure

output_predrought_effect_variable_figure

output_drought_number_figure

# Gene Expression Microarray Data Analysis


CEL_and_covdesc_Directory
Normalize
normalized_data_only_values
normalized_data
minimum_fold_change_for_DEG
cutoff_p_value_for_DEG
SelectDEGs
GO_stats_p_value_cutoff
DEG_list_lower_in_test_condition
DEG_list_higher_in_test_condition
DEG_list
DEG_list_lower_in_test_condition
DEG_list_full_info
DEG_list_summary
DEG_list_higher_in_test_condition
GO_Analysis
MakeHeatmap
GO_stats_MF_lower_in_test_condition
GO_stats_gene_list_higher_in_test_condition
GO_stats_BP_higher_in_test_condition
GO_stats_CC_higher_in_test_condition
GO_stats_MF_higher_in_test_condition
GO_stats_gene_list_lower_in_test_condition
GO_stats_BP_lower_in_test_condition
GO_stats_CC_lower_in_test_condition
heatmap

- [**Normalize**]
  - Normalization of data across microarray datasets
- [**SelectDEGs**]
  - Selection of differentially expressed genes between conditions
- [**GO Analysis**]
  - determination of gene ontology statistics for the resulting datasets
- [**MakeHeatmap**]
  - creation of a heatmap of the differentially expressed genes.

Tyler Kolisnik, Mark Bieda

# Why is workflow important

Thoughtfully designed organized workflows support:

Efficiency
Reliability
Modifiability
Reuse
Reproducibility

# Computational provenance

The heart of computational provenance:

*What data was used?*

*What calculations were performed?*

and also: “What in the world exactly happened just now?!”

# Why is provenance important?

Access to provenance information supports

- understanding
- reliability
- reproducibility
- trust
- attribution and credit
- discovery and reuse of data, tools, and algorithms

# Prospective vs Retrospective provenance (Ludaescher)

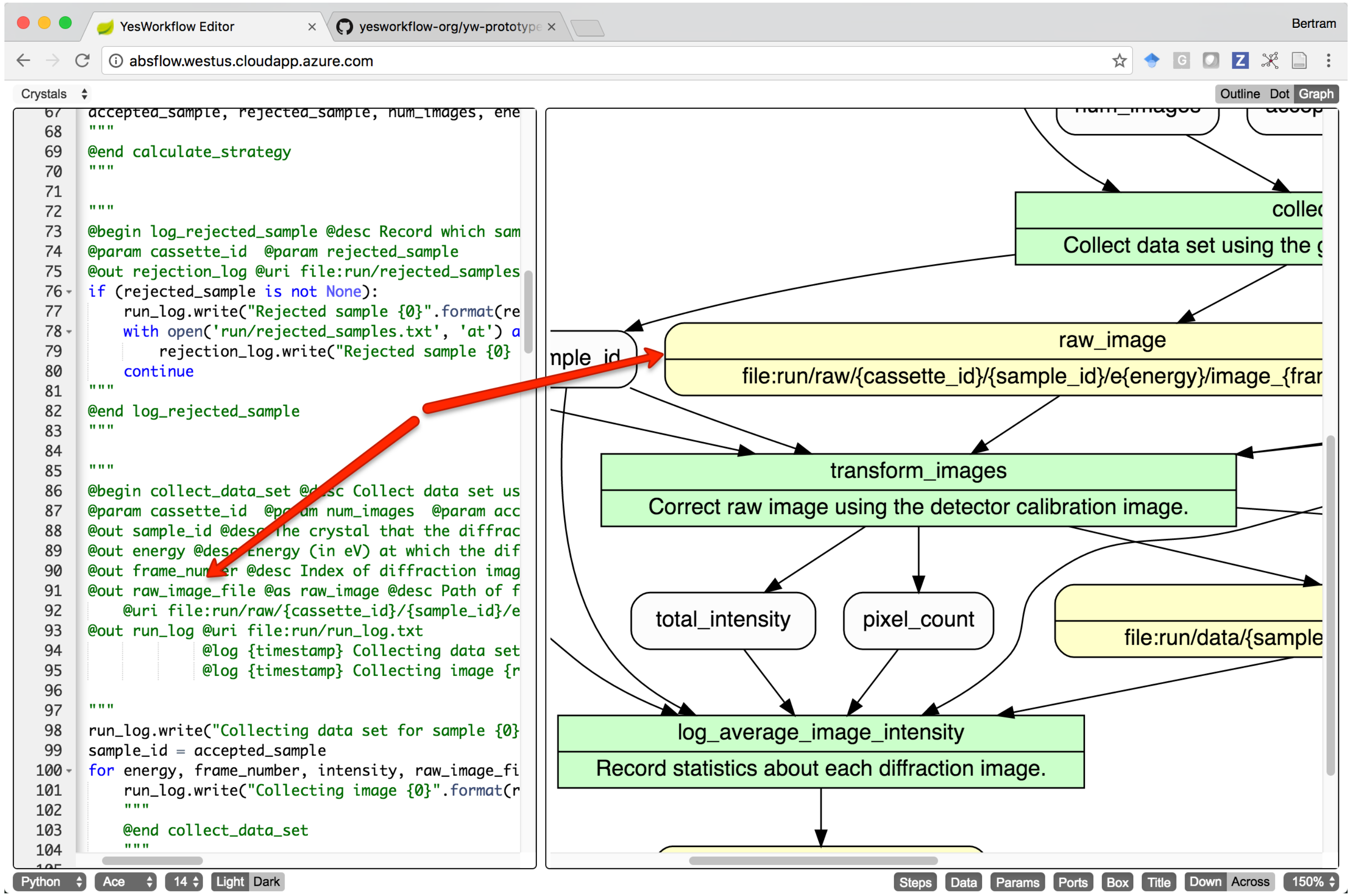
**Prospective**

a specification the workflow scenario

**Retrospective**

generated data on the execution of the workflow scenario

# Scripts are (or can be) workflow!! [try.yesworkflow.org]



YesWorkflow Editor
yesworkflow-org/yw-prototype
Bertram
absflow.westus.cloudapp.azure.com
Crystals
Outline
Dot
Graph
num_images
accep
collec
Collect data set using the g
nple_id
raw_image
file:run/raw/{cassette_id}/{sample_id}/e{energy}/image_{fran
transform_images
Correct raw image using the detector calibration image.
total_intensity
pixel_count
file:run/data/{sample
log_average_image_intensity
Record statistics about each diffraction image.
Python
Ace
14
Light
Dark
Steps
Data
Params
Ports
Box
Title
Down
Across
150%

```
accepted_sample, rejected_sample, num_images, ene
"""
@end calculate_strategy
"""


"""
@begin log_rejected_sample @desc Record which sam
@param cassette_id  @param rejected_sample
@out rejection_log @uri file:run/rejected_samples
if (rejected_sample is not None):
    run_log.write("Rejected sample {0}".format(re
    with open('run/rejected_samples.txt', 'at') a
        rejection_log.write("Rejected sample {0}
    continue
"""
@end log_rejected_sample
"""


"""
@begin collect_data_set @desc Collect data set us
@param cassette_id  @param num_images  @param acc
@out sample_id @desc The crystal that the diffrac
@out energy @desc Energy (in eV) at which the dif
@out frame_number @desc Index of diffraction imag
@out raw_image_file @as raw_image @desc Path of f
    @uri file:run/raw/{cassette_id}/{sample_id}/e
@out run_log @uri file:run/run_log.txt
              @log {timestamp} Collecting data set
              @log {timestamp} Collecting image {r

"""
run_log.write("Collecting data set for sample {0}
sample_id = accepted_sample
for energy, frame_number, intensity, raw_image_fi
    run_log.write("Collecting image {0}".format(r
    """
    @end collect_data_set
    """
```

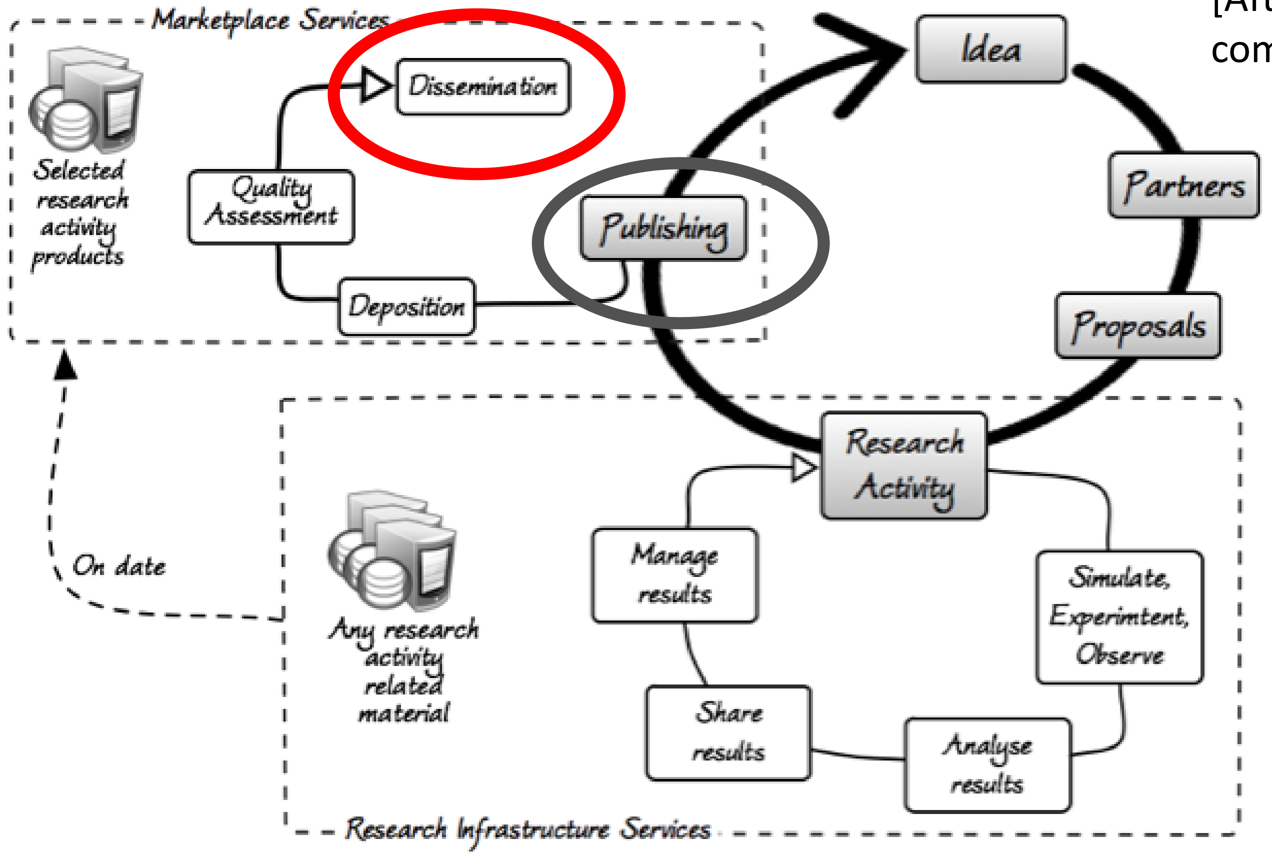
# At least remember this

don’t just sit there typing at the command line,
write a script, and document it

[for crying out loud]

# Communication

# 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.

image from "Science 2.0 Repositories: Time for a Change in Scholarly Communication" Massimiliano Assante et al., *D-Lib Magazine* 2015.

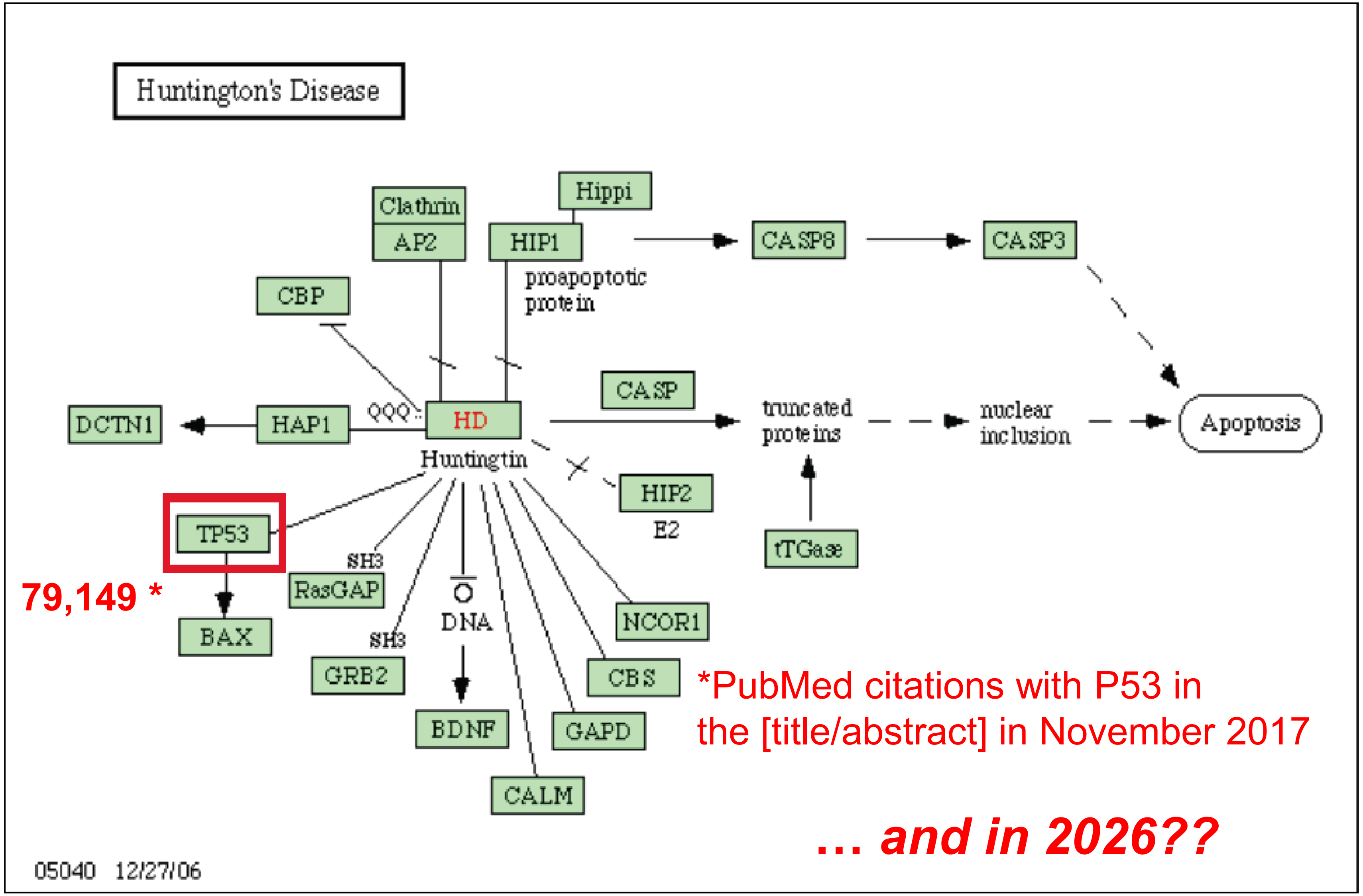
# . . . the crisis

*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

# Lisa’s problem



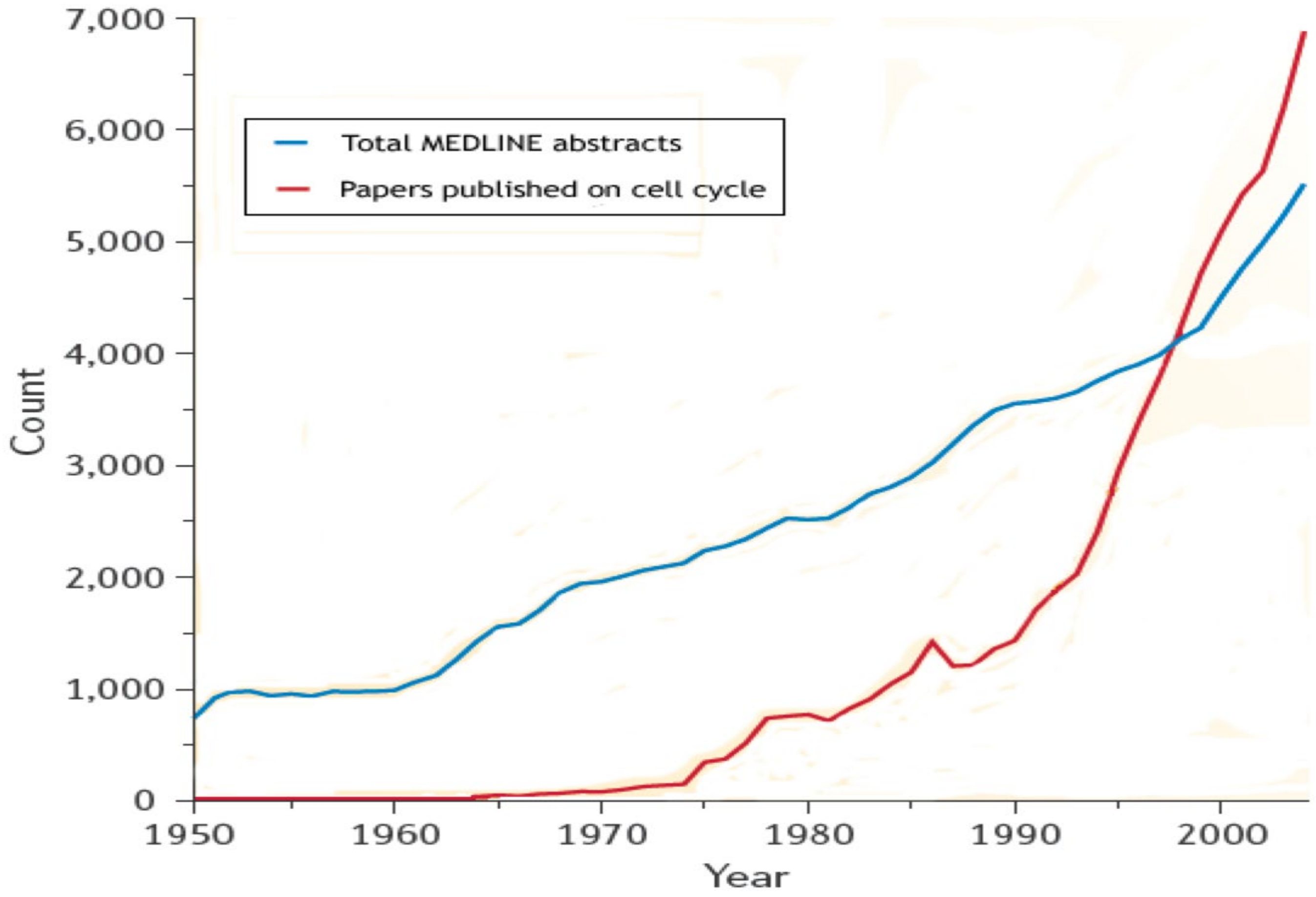
79,149 *

*PubMed citations with P53 in the [title/abstract] in November 2017

***… and in 2026??***

from John Wilbanks, adapting a KEGG diagram

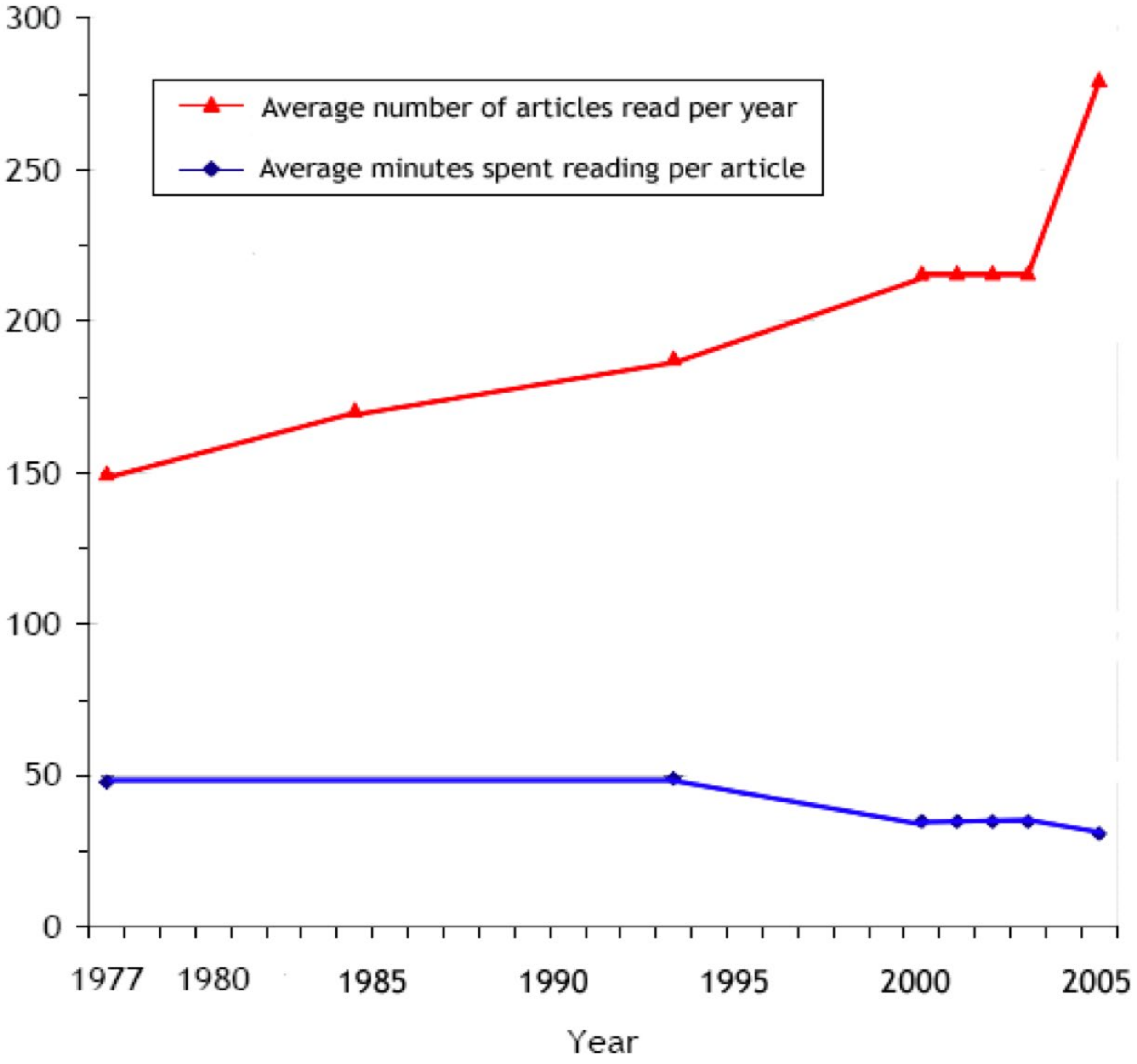
# *Are you kidding me???*



[Axis is x10$^{-2}$ for total Medline abstracts]

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

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



Tenopir et al. 1977-2005

# Responses to the problem

**One response:** *text mining [*instead of L2R/T2B *reading]*

information extraction

"undiscovered public knowledge"

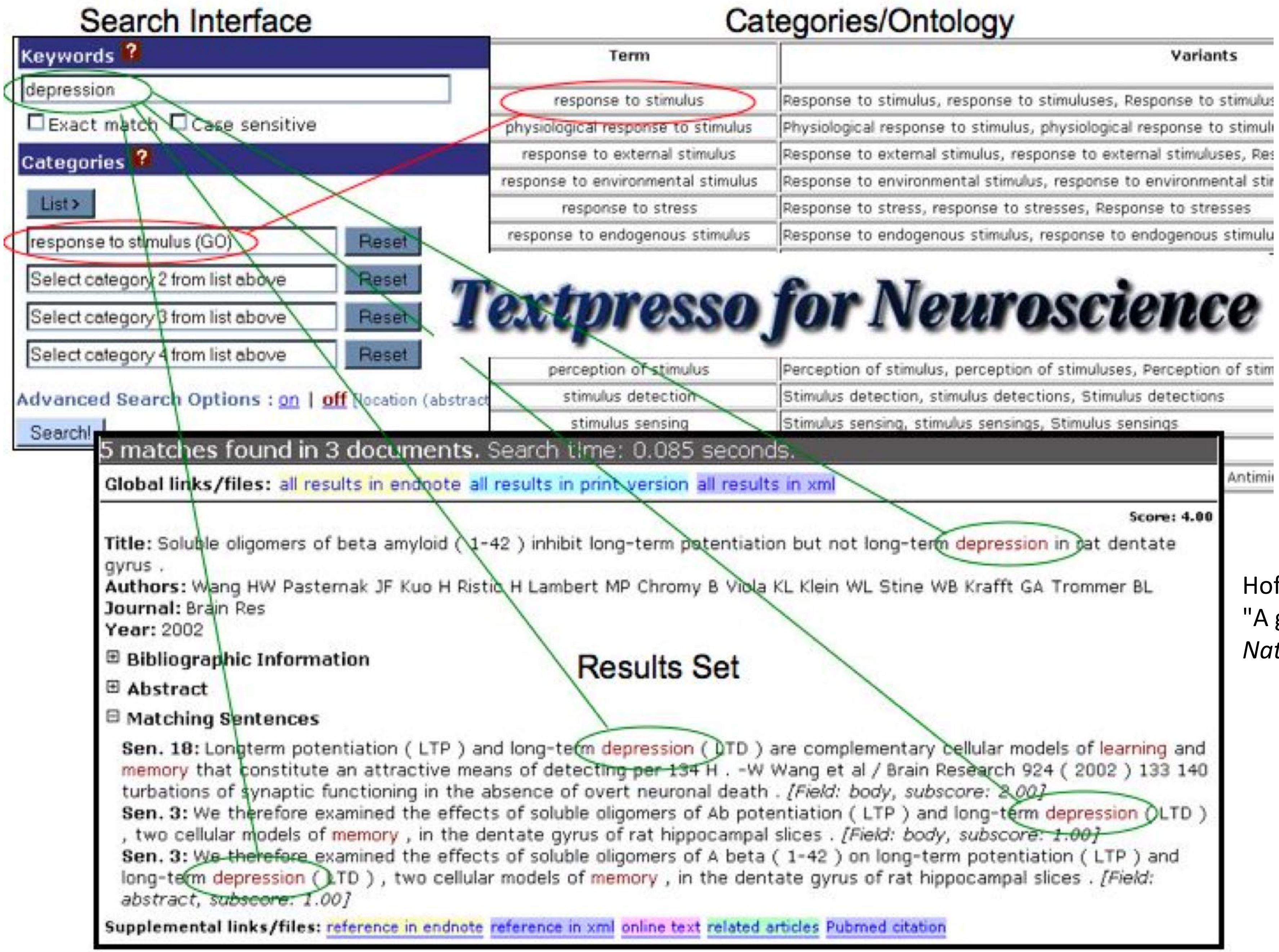
and hypothesis generation

(Swanson and Smalheiser

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

# Necessary data standards are now, finally, in place to support reading tools

| | |
|---|---|
| **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] |



Keywords: depression

□ Exact match □ Case sensitive

Categories: 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

| 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 stimul |
| response to external stimulus | Response to external stimulus, response to external stimuluses, Res |
| response to environmental stimulus | Response to environmental stimulus, response to environmental stir |
| response to stress | Response to stress, response to stresses, Response to stresses |
| response to endogenous stimulus | Response to endogenous stimulus, response to endogenous stimulu |
| 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 |
| | Antimi |

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**

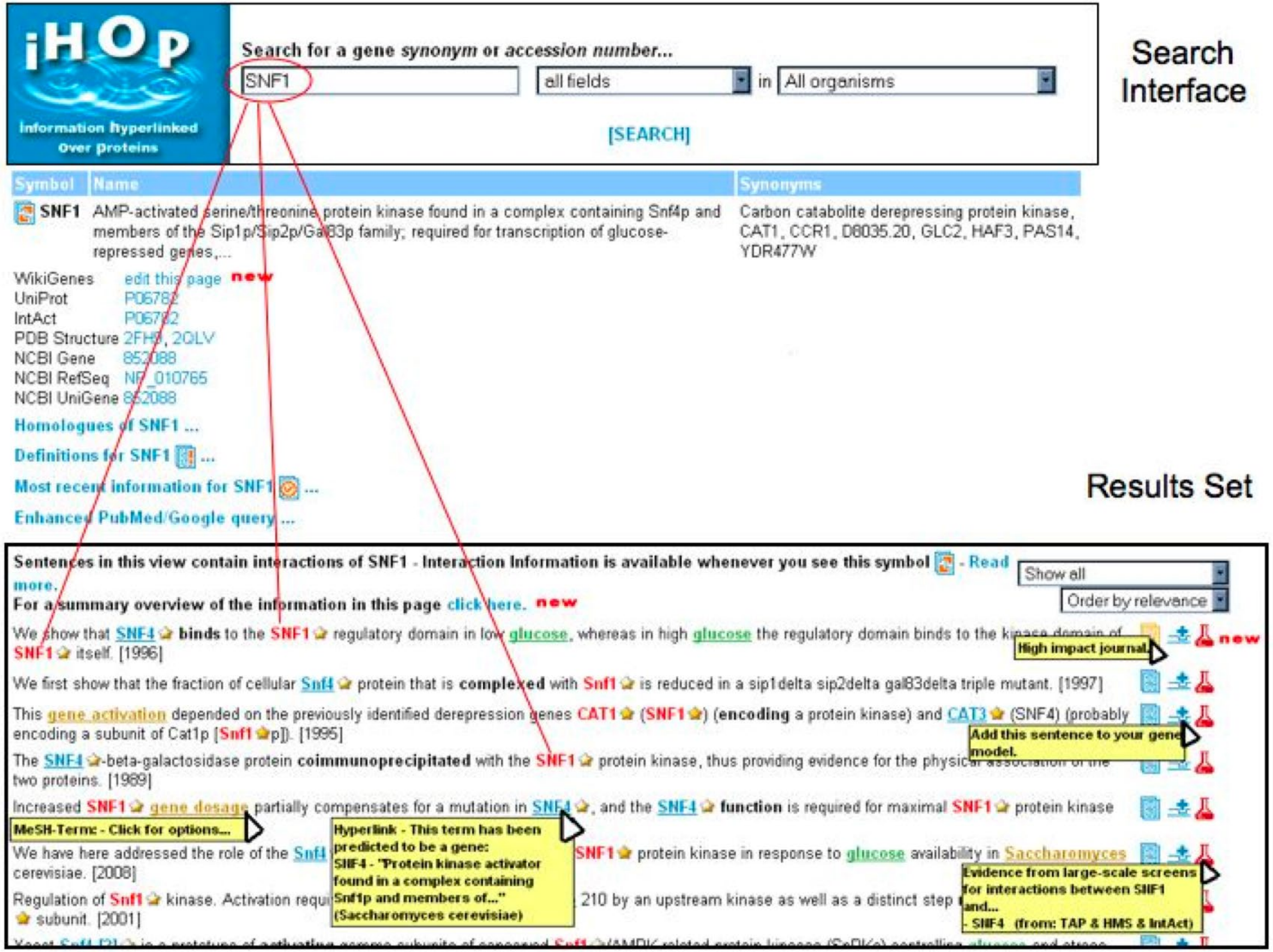
⊞ **Abstract**

⊟ **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

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



Search for a gene *synonym* or *accession number...*

SNF1 | all fields in All organisms

[SEARCH]

| Symbol | Name | Synonyms |
|---|---|---|
| SNF1 | 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 | edit this page new |
| UniProt | P06782 |
| IntAct | P06782 |
| PDB Structure | 2FH9, 2QLV |
| NCBI Gene | 852088 |
| NCBI RefSeq | NP_010765 |
| NCBI UniGene | 852088 |

**Homologues of SNF1 ...**

**Definitions for SNF1 ...**

**Most recent information for SNF1 ...**

**Enhanced PubMed/Google query ...**

**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

Show all
Order by relevance

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

High impact journal

We first show that the fraction of cellular Snf4 protein that is **complexed** with Snf1 is reduced in a sip1delta sip2delta gal83delta 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 [Snf1p]). [1995]

Add this sentence to your gene model.

The SNF4-beta-galactosidase protein **coimmunoprecipitated** with the SNF1 protein kinase, thus providing evidence for the physic 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...

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)

We have here addressed the role of the Snf4 SNF1 protein kinase in response to glucose availability in Saccharomyces cerevisiae. [2008]

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

Regulation of Snf1 kinase. Activation requi 210 by an upstream kinase as well as a distinct step subunit. [2001]

Search Interface

Results Set

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

# Data Mining?

*We wouldn't have to mine the data if we didn't bury in the first place.*

Barend Mons, "Which gene did you mean?" *BMC Bioinfomatics* (2005)

And finally. . .

Automate like you are going to live forever
Document like you are going to  die tomorrow

— *Michael Sperberg-McQueen*

# and for fun. . .

- https://www.youtube.com/watch?v=66oNv_DJuPc