Communication and data curation

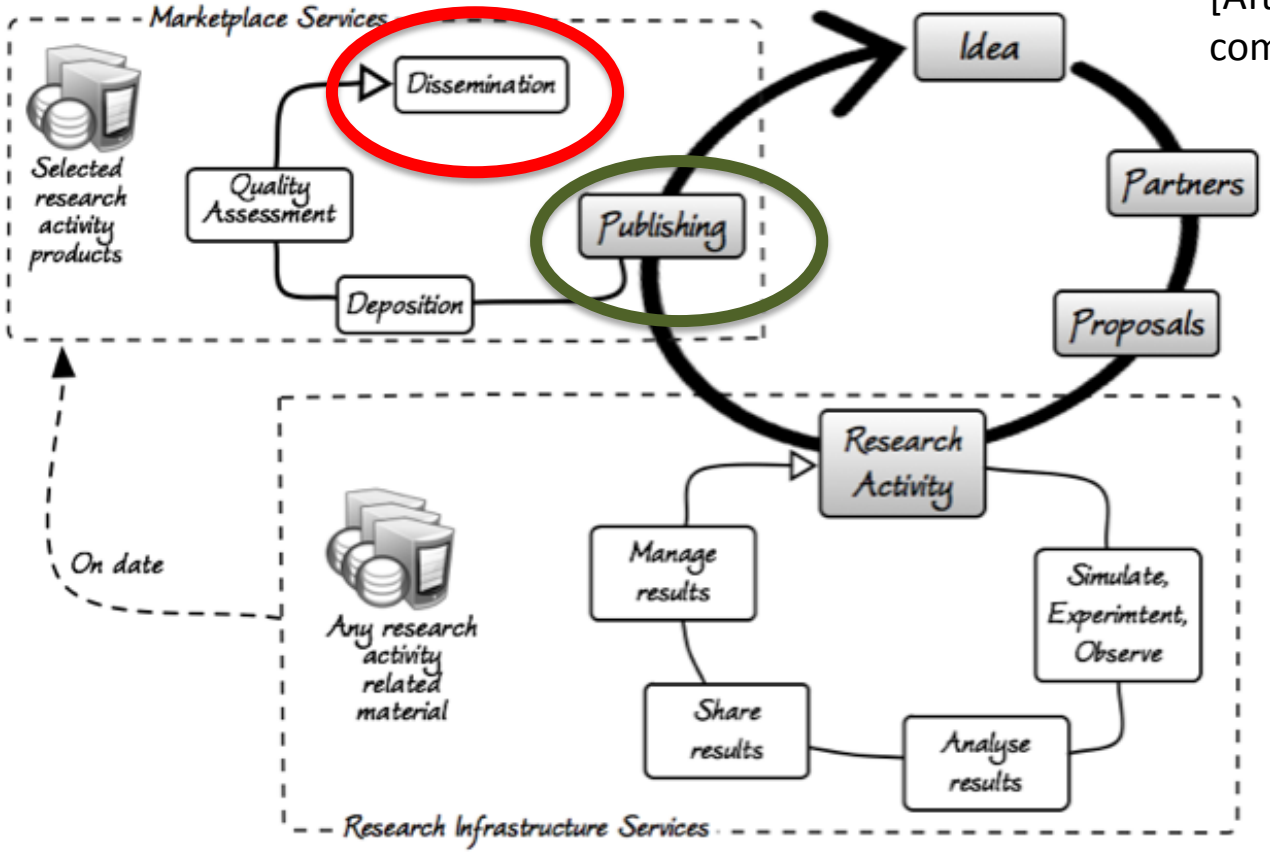
V2. The crisis in data-driven scientific communication

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.

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

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)

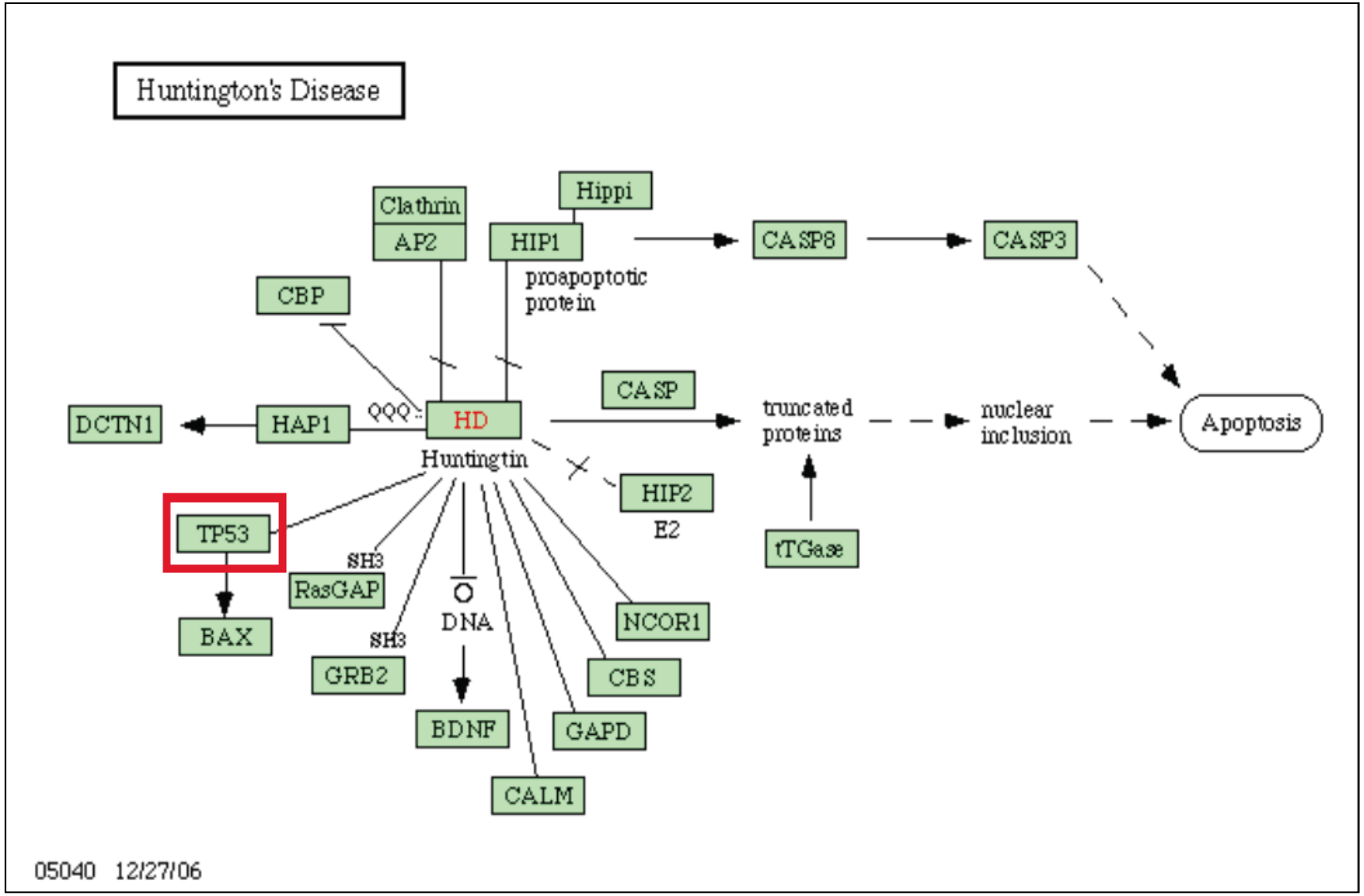
in 2026 engaging with the scientific literature
will (*finally*) be like

*"flying a jet plane through information space"**

and not at all like *finding and reading* articles

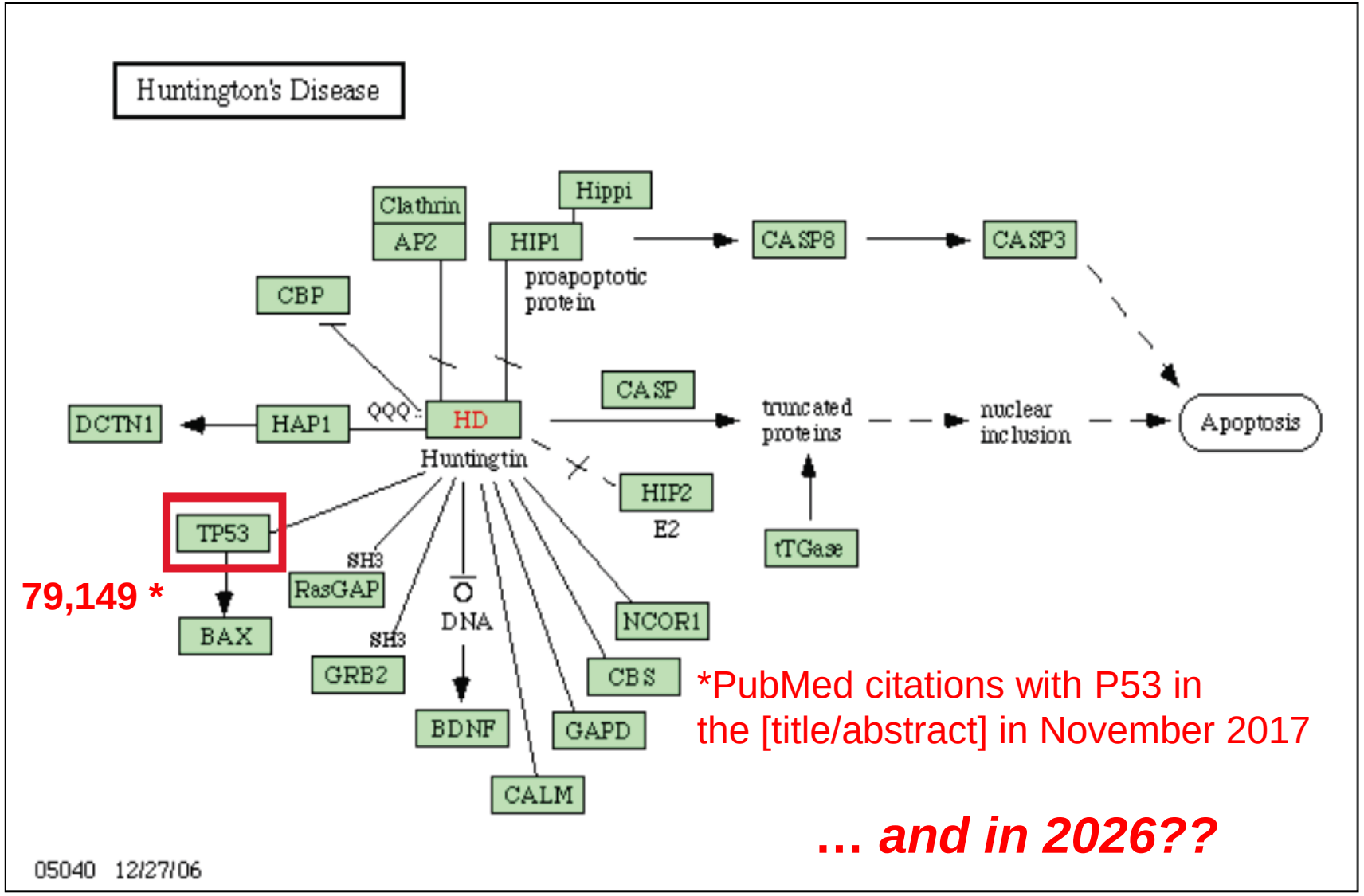
*attributed to Alan Kay, mid-1980s,

Lisa's problem: P53



from John Wilbanks, adapting a KEGG map

Lisa's problem



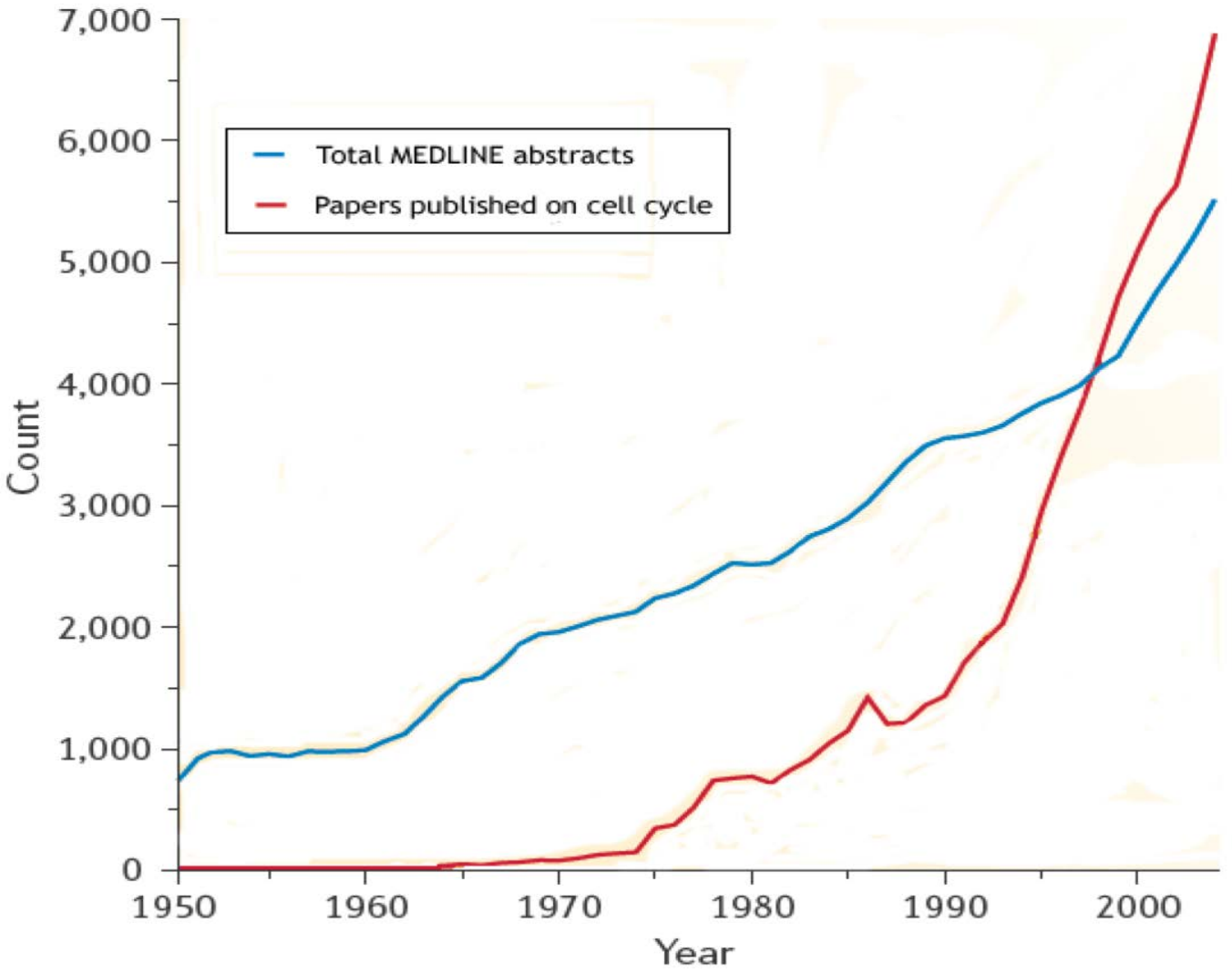
from John Wilbanks, adapting a KEGG diagram

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 x10^{-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*