

# Open Research Data, Urgent Problems, and Incremental Solutions

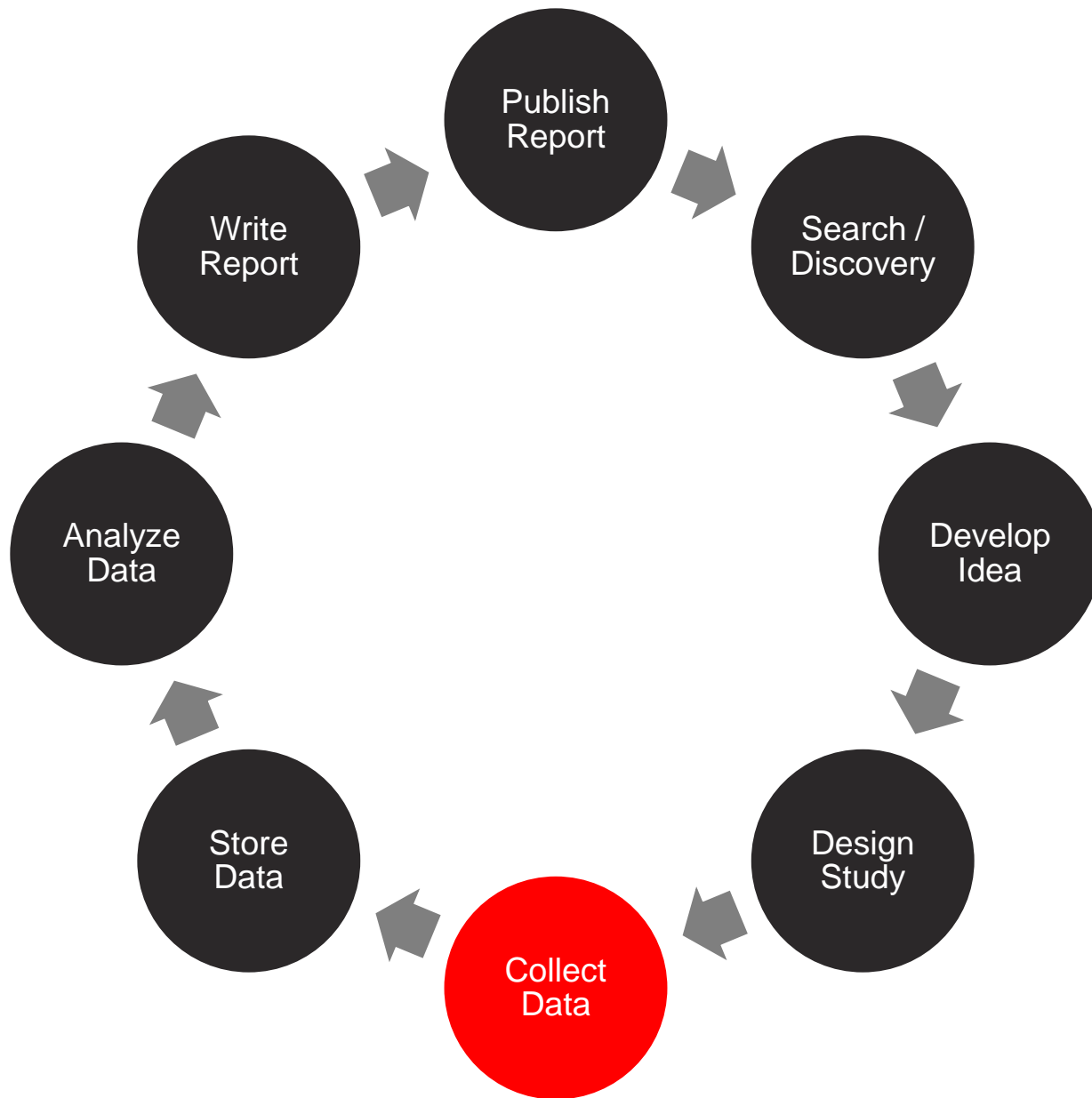
Jeffrey Spies

221B | SHARE | UVA

 @jeffspies

**Step One**

**Broaden the definition of  
data.**



# Research Lifecycle



Context is important.

The (meta)data across the research lifecycle is the data necessary for effective **assessment and metascience, meta-analyses, reproducibility, replicability, reuse, discovery, and innovation.**

This can be thought of as a **knowledge graph**, but one not only composed of semantic relationships extracted from published text.

Semantic relationships extracted  
from published text +  
relationships between **people,**  
**places, data, methods, software,**  
**preprints, peer-review, etc.**



## Step Two

The situations with respect to incentives is bleak.

Incentives for individual success are focused on **getting it published, not getting it right.**

**Table 5 Comparison of Average Amount of Cash Awards\* for a Paper Published in Selected Journals (2008-2016)**

	2008	2009	2010	2011	2012	2013	2014	2015	2016
<i>Nature, Science</i>	\$26,212	\$26,006	\$25,781	\$25,365	\$33,990	\$36,658	\$38,908	\$43,783	\$43,783
<i>PNAS</i>	\$3,156	\$3,025	\$3,353	\$3,443	\$3,664	\$3,619	\$3,751	\$3,513	\$3,513
<i>PLOS One</i>	\$1,096	\$1,086	\$1,035	\$994	\$991	\$915	\$941	\$984	\$984
<i>MIS Quarterly</i>	\$2,613	\$2,570	\$2,553	\$2,654	\$2,876	\$2,861	\$2,992	\$2,938	\$2,938
<i>JASIST</i>	\$1,737	\$1,758	\$1,741	\$1,887	\$2,066	\$2,303	\$2,435	\$2,488	\$2,488
<i>Journal of Documentation</i>	\$1,082	\$1,087	\$1,042	\$1,111	\$1,167	\$1,265	\$1,329	\$1,408	\$1,408
<i>Library Hi Tech</i>	\$781	\$775	\$726	\$741	\$740	\$768	\$795	\$783	\$783
<i>LIBRI</i>	\$650	\$644	\$577	\$560	\$538	\$509	\$517	\$484	\$484

\* All the amounts are full amount (in USD) awarded to the first author

Incentives to publish lead  
to bias to produce what is  
publishable.

Rates of retraction and p-value errors are positively correlated with impact factor.

Statistical power is  
negatively correlated with  
impact factor.

# Why Most Published Research Findings Are False

Corollary 1: The **smaller the studies** conducted in a scientific field, the less likely the research findings are to be true.

Corollary 2: The **smaller the effect sizes** in a scientific field, the less likely the research findings are to be true.



Therefore, likelihood of truth is negatively correlated with impact factor.

**Step Three**

There is urgency.

**67**

TRIALS CHECKED

**9**

TRIALS WERE PERFECT

**354**

OUTCOMES NOT  
REPORTED

**357**

NEW OUTCOMES  
SILENTLY ADDED

On average, each trial reported just 58.2% of its specified outcomes. And on average, each trial silently added 5.3 new outcomes.

**58**

LETTERS SENT

**18**

LETTERS PUBLISHED

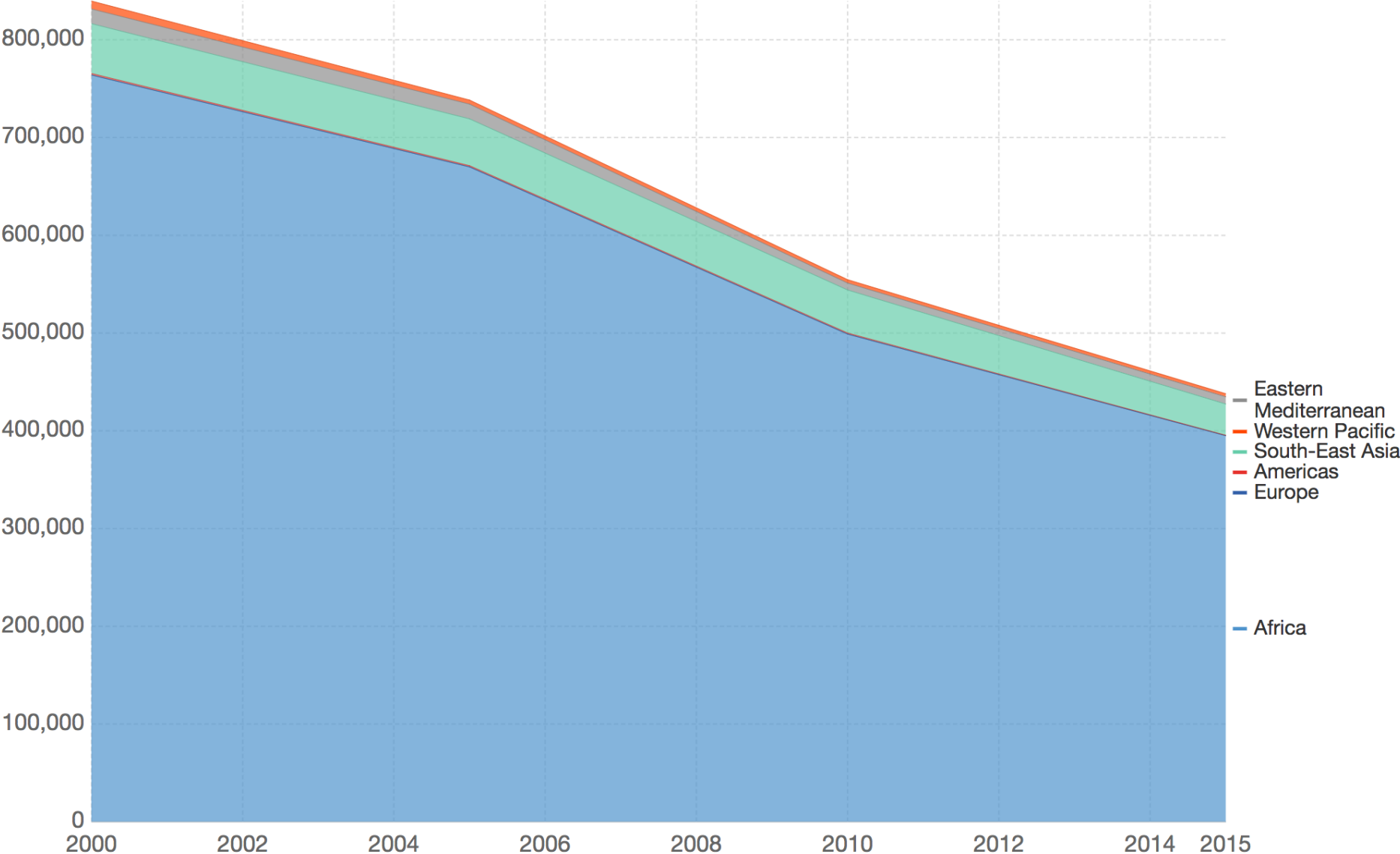
**8**

LETTERS  
UNPUBLISHED AFTER 4  
WEEKS

**32**

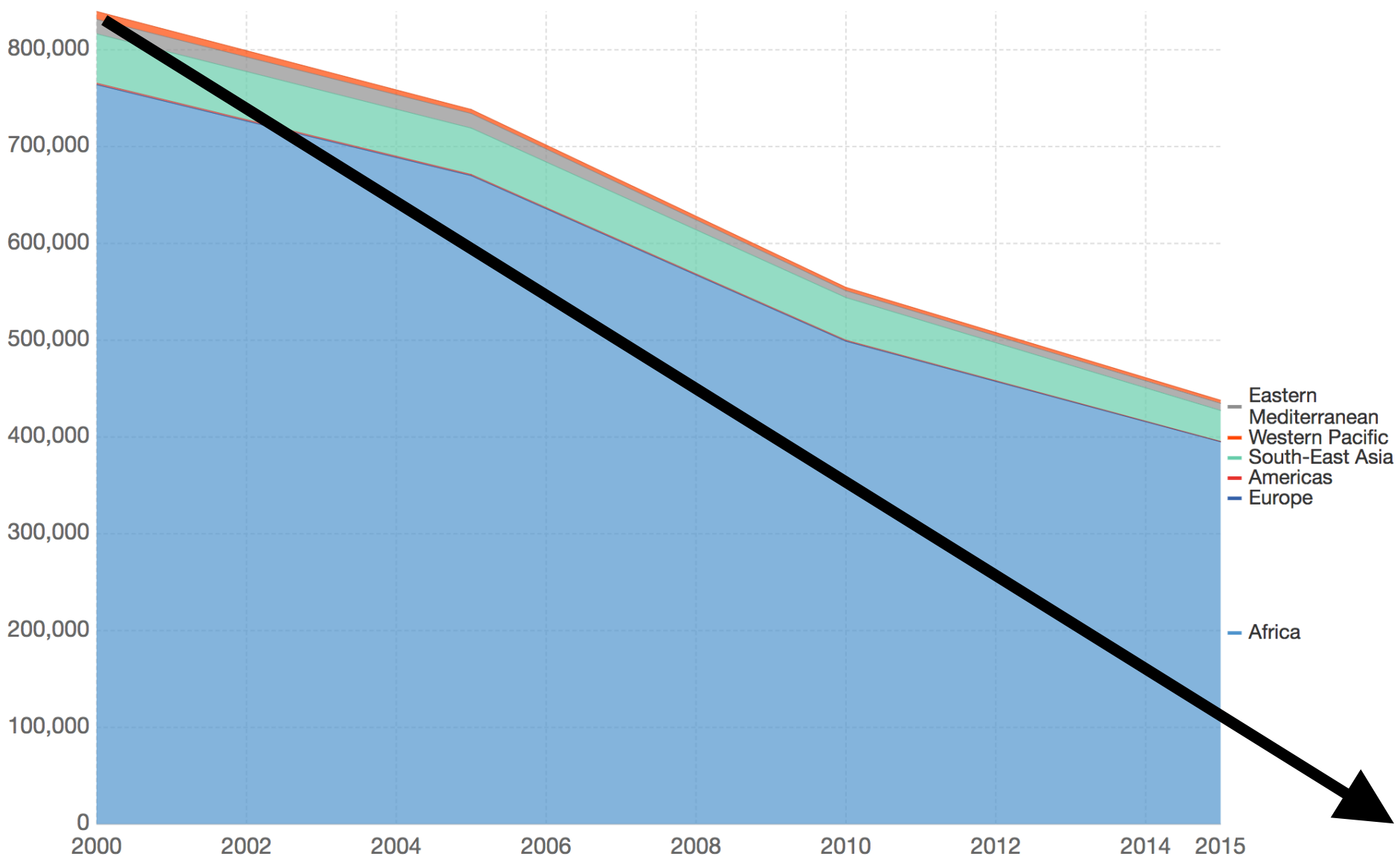
LETTERS REJECTED BY  
EDITOR

# Global malaria deaths by world region



Source: Deaths by World Region (WHO)

# Global malaria deaths by world region



Source: Deaths by World Region (WHO)



There are maybe 10 other people in the world that can understand my work much less reproduce or extend it.

If I'm wrong

And 1 more person can  
extend my work, I have  
increased the effort  
dedicated to that  
problem by 10%.

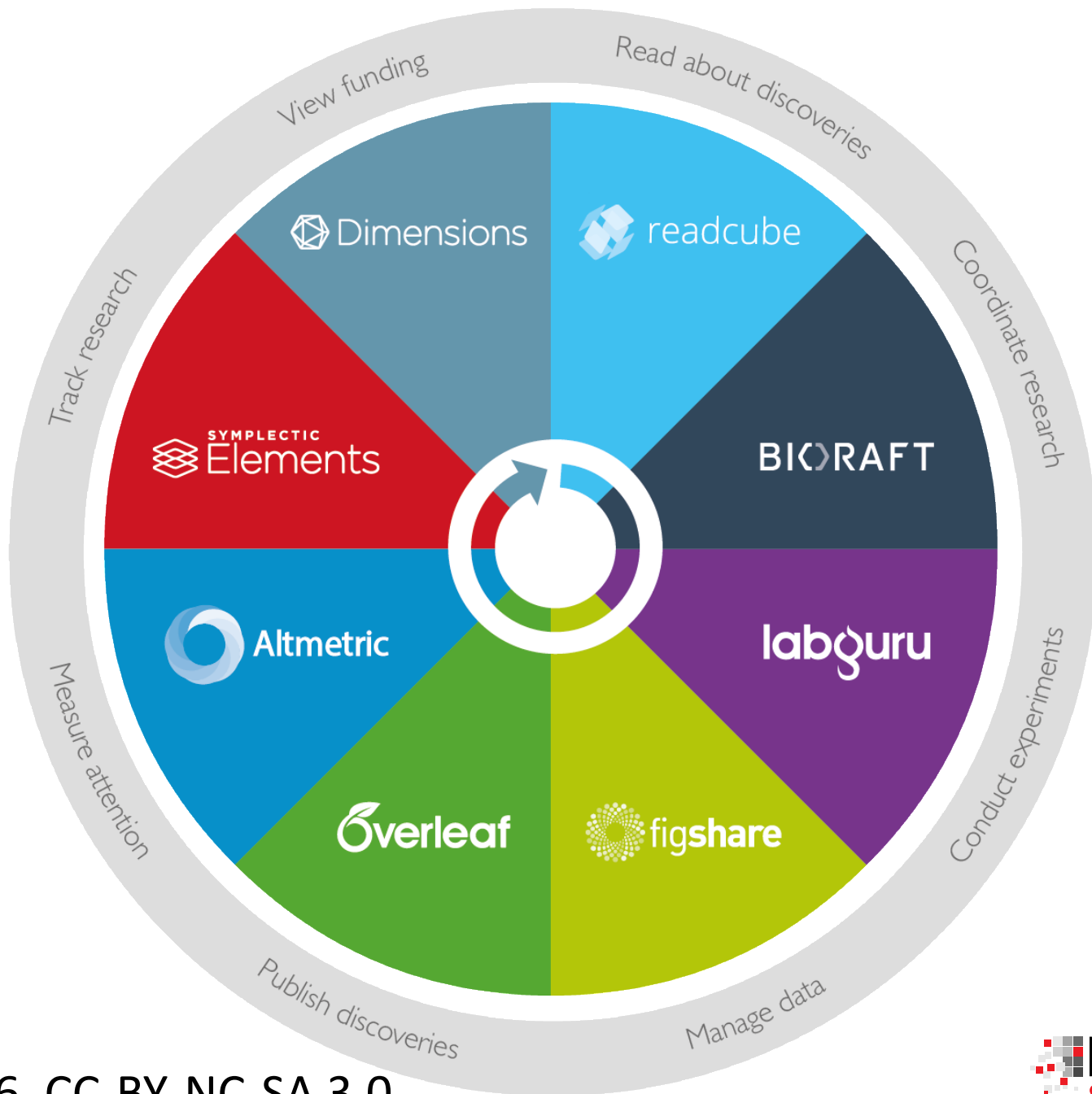
...1 more person out of  
the 99.98% of the world  
without access.



## **Step Four**

**Vendor lock-in is the dominant business model in scholarship.**

If research data goes the way of text, fewer people will be able to afford it, and it will stifle scholarly and commercial innovation.



## Step Five

Given urgency and pace of lock-in based businesses, we don't have time to wait for culture change.

## Step Six

There are stakeholders that can immediately move incentives: funders and policy makers.

## Step Seven

Researchers need a partner; they lack incentives and time.

There is a ready partner  
with the expertise to help  
if given the chance: the  
research library.

They have process  
expertise.



The data they processed  
in the past was the book;  
it is now digital streams of  
information.

They are value-oriented.

Generally speaking, they lack respect/recognition from researchers.

## **Step Eight**

**Start with resourced  
solutions and nudges;  
increment.**

1. List the URL to your department's promotion and tenure policy.

2. Does that policy mention the value of data sharing?

3. List the librarian/archivist with whom you will partner on data stewardship and sharing.

## **Step Nine**

**Support the development  
of a free, open knowledge  
graph.**



## Step Ten

Support the development of open infrastructure that welcomes both vendors and scholars, but enforces values via a practical integration of distributed and decentralized technology.

jeff@221b.io