# 开放科学浪潮下，图书馆员的新角色

Calvin Wu| 吴非

EBSCO 销售经理

EBSCO Information Services

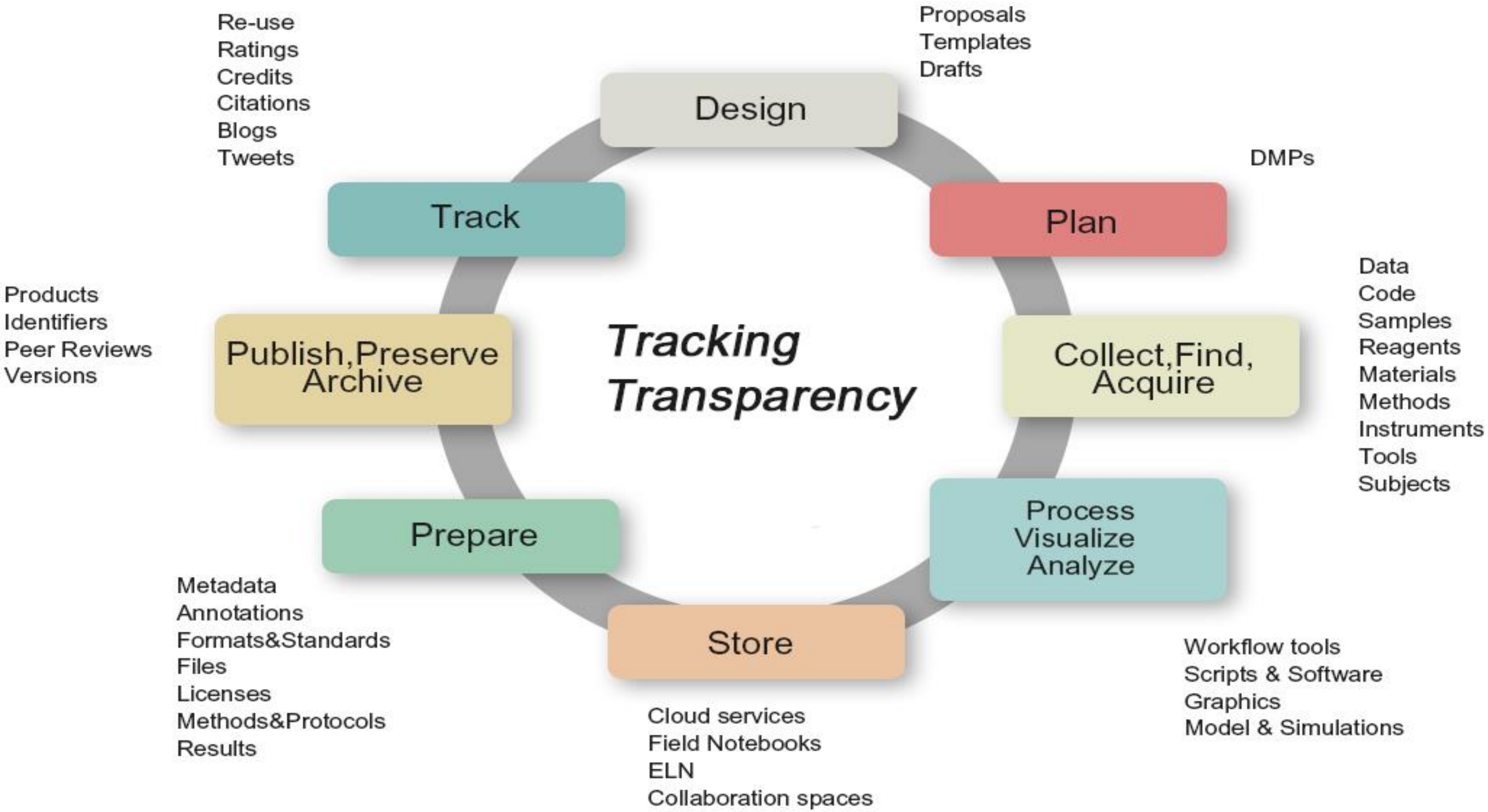
01

# 开放科学的过去与现在

? 一项研究的生命周期

! 开放阶段的前移

# 一项研究的生命周期也许是这样：



Tracking Transparency

Design

Proposals
Templates
Drafts

Plan

DMPs

Collect,Find, Acquire

Data
Code
Samples
Reagents
Materials
Methods
Instruments
Tools
Subjects

Process Visualize Analyze

Workflow tools
Scripts & Software
Graphics
Model & Simulations

Store

Cloud services
Field Notebooks
ELN
Collaboration spaces

Prepare

Metadata
Annotations
Formats&Standards
Files
Licenses
Methods&Protocols
Results

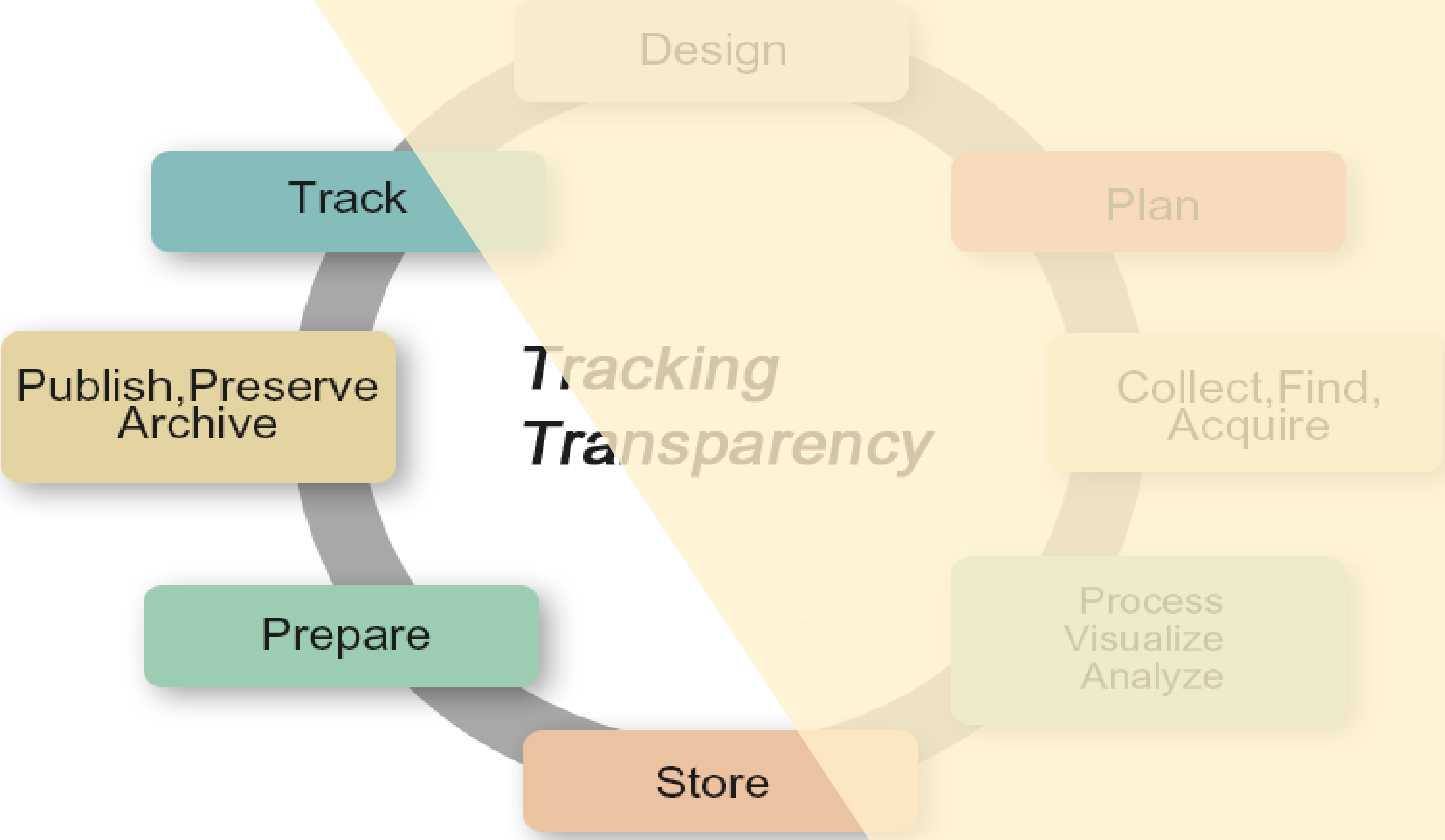
Publish,Preserve Archive

Products
Identifiers
Peer Reviews
Versions

Track

Re-use
Ratings
Credits
Citations
Blogs
Tweets

# 一项研究的生命周期也许是这样：



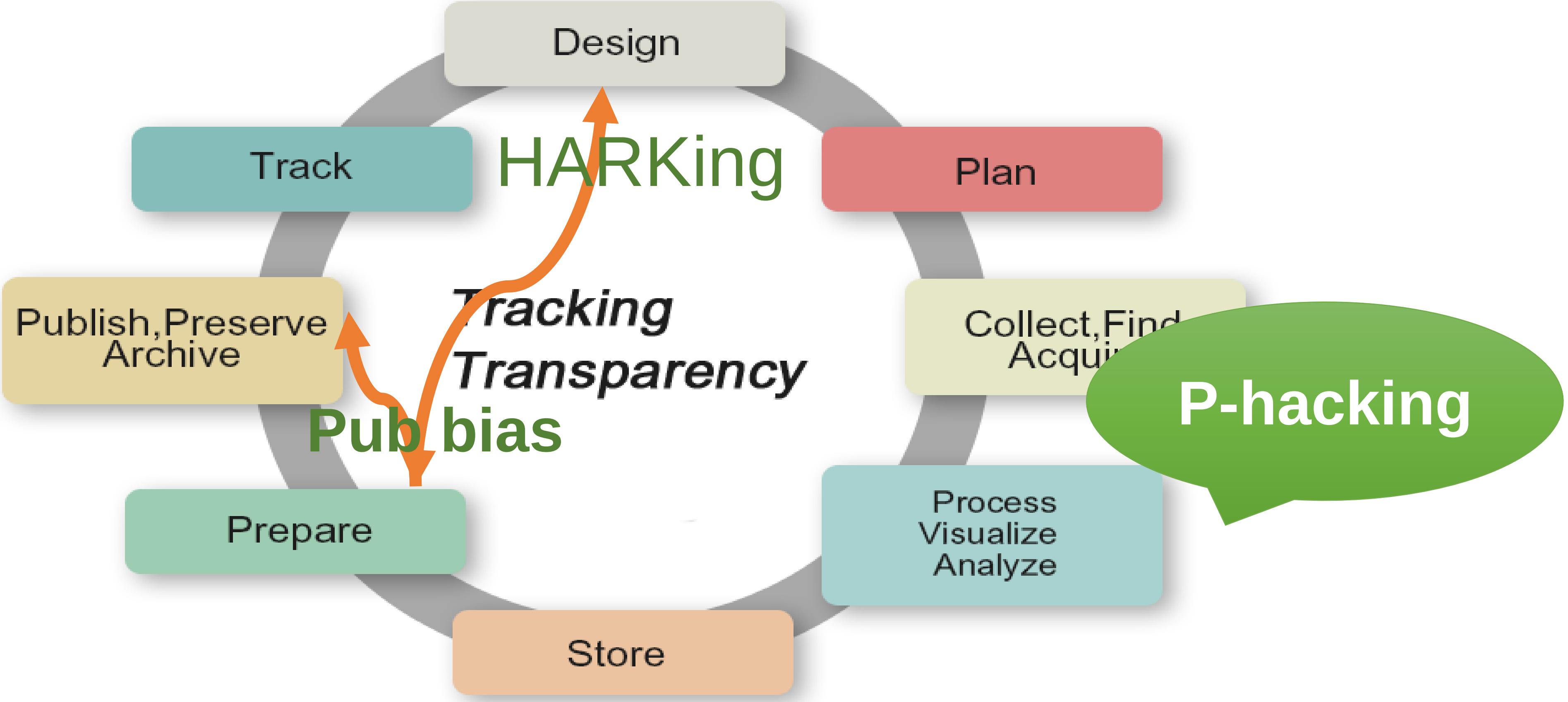
# 开放科学的背景：顶级期刊纷纷响应，2010s成为热议话题的”R WORDS”：

Repeatability(可重复性)

Replicability(可复制性)

Reproducibility(可再现性)

# 如果只分享出版以后的产物，难以避免R-WORDS的危机以及研究不透明的威胁：



今天的开放科学-
期待研究生命周期的每个阶段都会开放，
即便是失败的结果

02

# 走向开放科学的未来

- 开放科学新兴的3D模型
- No raw data, No science, 没有原始数据，就没有科学
- 未来图书馆员可能的新角色

*Fig. 1: 2D-model of Open Science (Based on the Continuum of Openness in Lyon, 2009).*

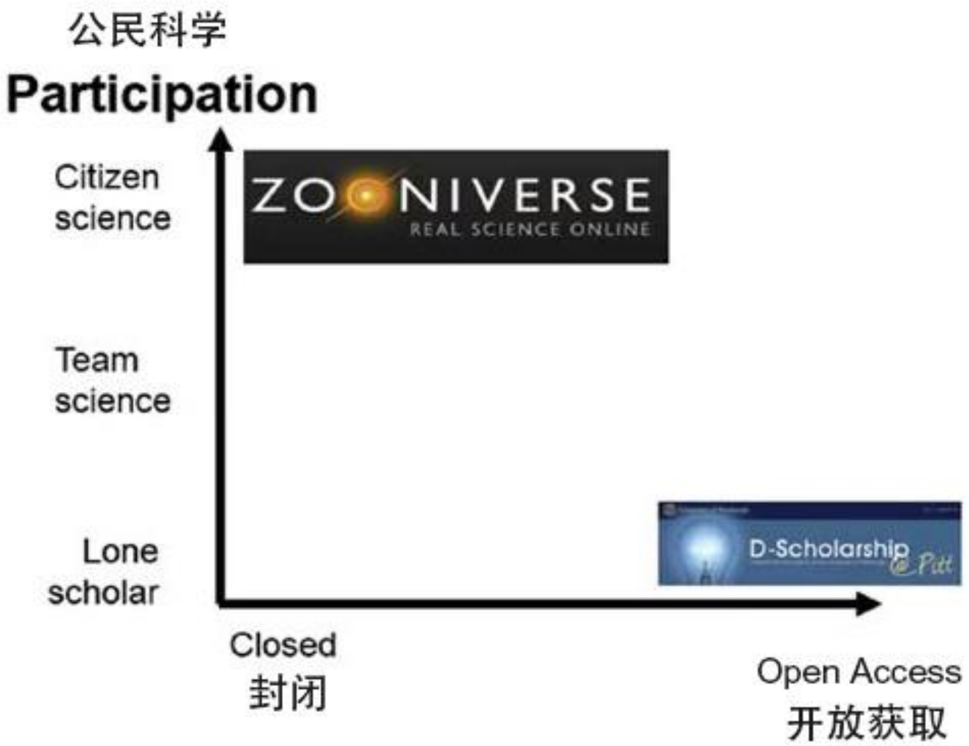
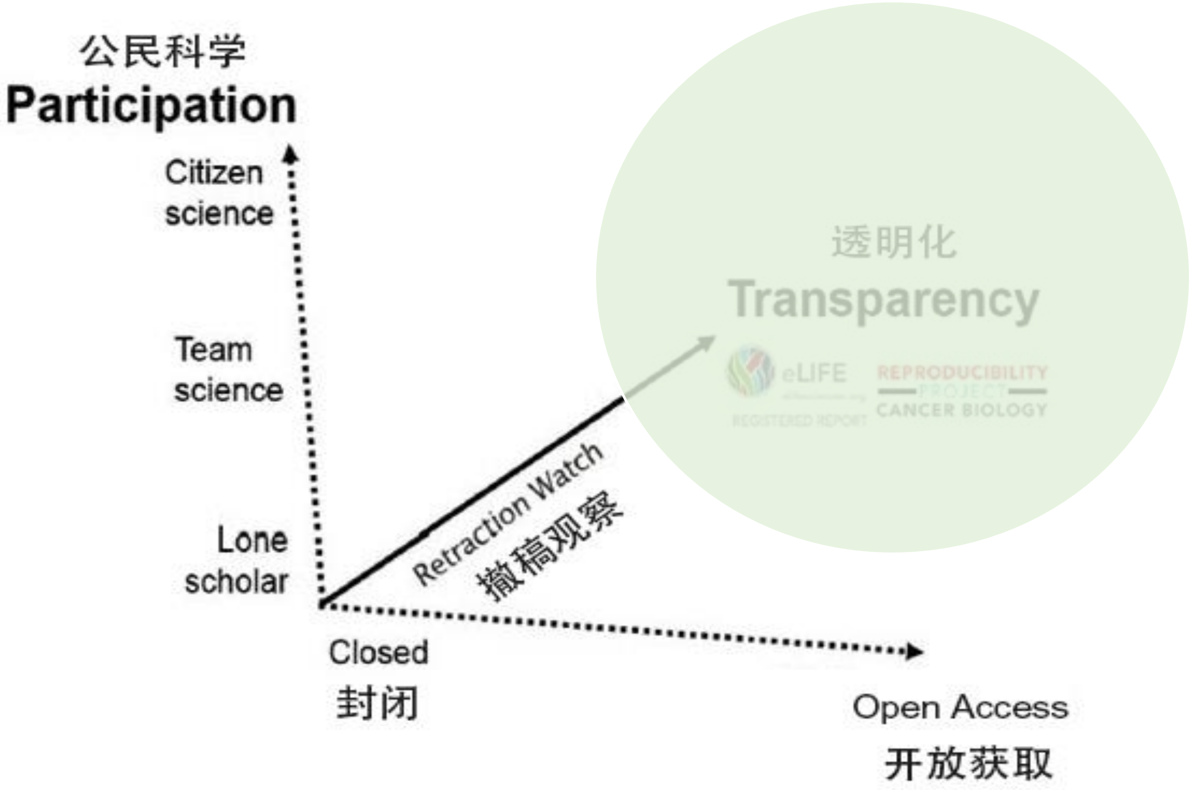
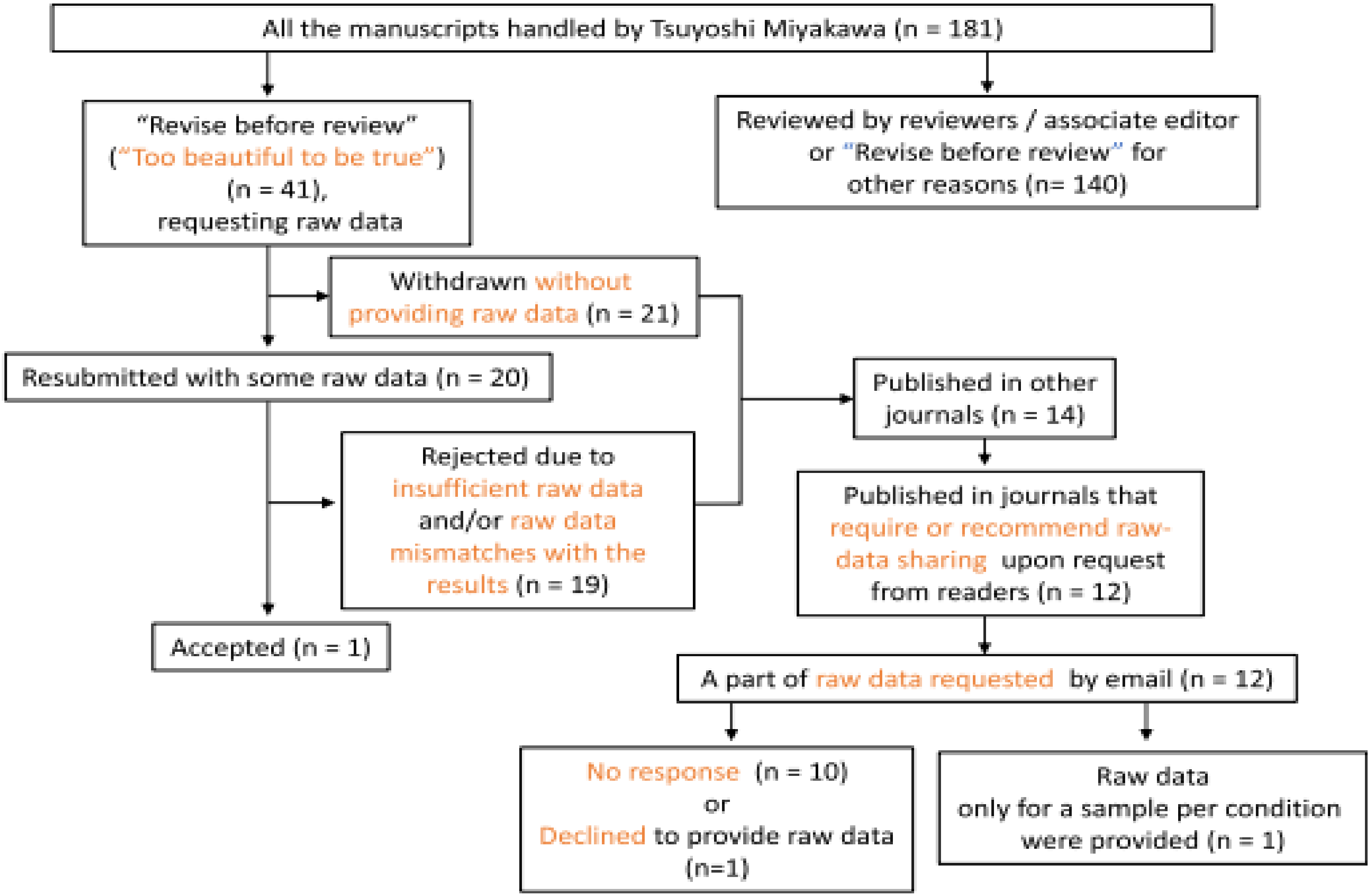


*Fig. 2: 3D-Model of Open Science.*



# 开放科学新兴的3D模型

*Liber Quarterly Volume 25 Issue 4 2016*



All the manuscripts handled by Tsuyoshi Miyakawa (n = 181)

"Revise before review" ("Too beautiful to be true") (n = 41), requesting raw data

Reviewed by reviewers / associate editor or "Revise before review" for other reasons (n= 140)

Withdrawn without providing raw data (n = 21)

Resubmitted with some raw data (n = 20)

Published in other journals (n = 14)

Rejected due to insufficient raw data and/or raw data mismatches with the results (n = 19)

Published in journals that require or recommend raw-data sharing upon request from readers (n = 12)

Accepted (n = 1)

A part of raw data requested by email (n = 12)

No response (n = 10) or Declined to provide raw data (n=1)

Raw data only for a sample per condition were provided (n = 1)

**Fig. 1** Flowchart of the manuscripts handled by Tsuyoshi Miyakawa in *Molecular Brain* from December 2017 to September 2019

Tsuyoshi Miyakawa

# 原始数据的缺乏或数据捏造是造成不可重复性的另一个可能原因

让原始数据能够公开获取不仅对于数据挖掘、再使用非常重要，而且对于去证实研究的结果是否基于实际的数据，也同样重要

- Tsuyoshi Miyakawa

Miyakawa, T. No raw data, no science: another possible source of the reproducibility crisis. *Mol Brain* **13,** 24 (2020). https://doi.org/10.1186/s13041-020-0552-2

# 国务院办公厅关于印发科学数据管理办法的通知

「政府预算资金资助形成的科学数据应当按照开放为常态、不开放为例外的原则，由主管部门组织编制科学数据资源目录，有关目录和数据应及时接入国家数据共享交换平台，面向社会和相关部门开放共享。 」

http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm

# 图书馆员可能的新角色-数据透明化代理人，即Data Librarian@开放科学有关透明化的基本术语

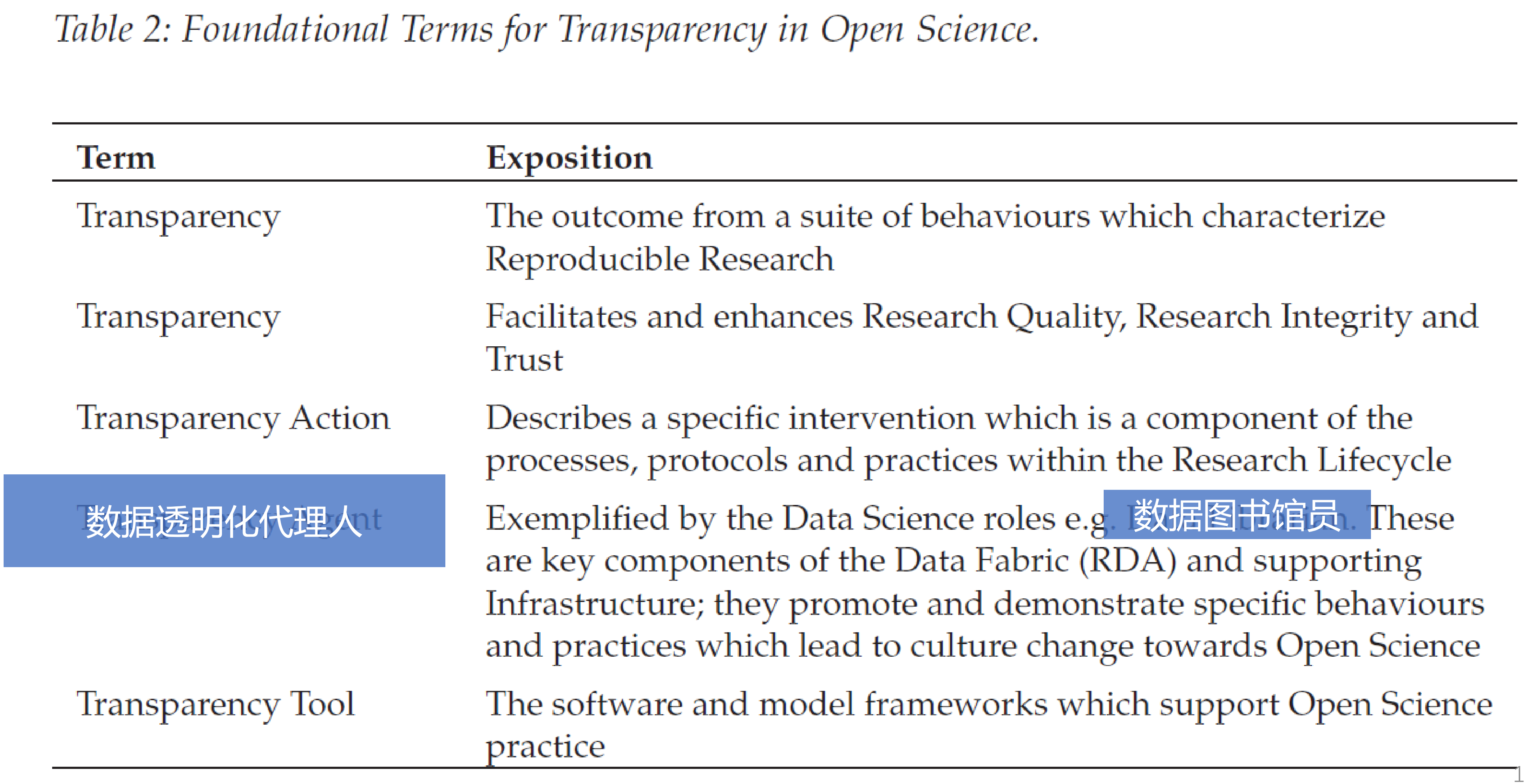
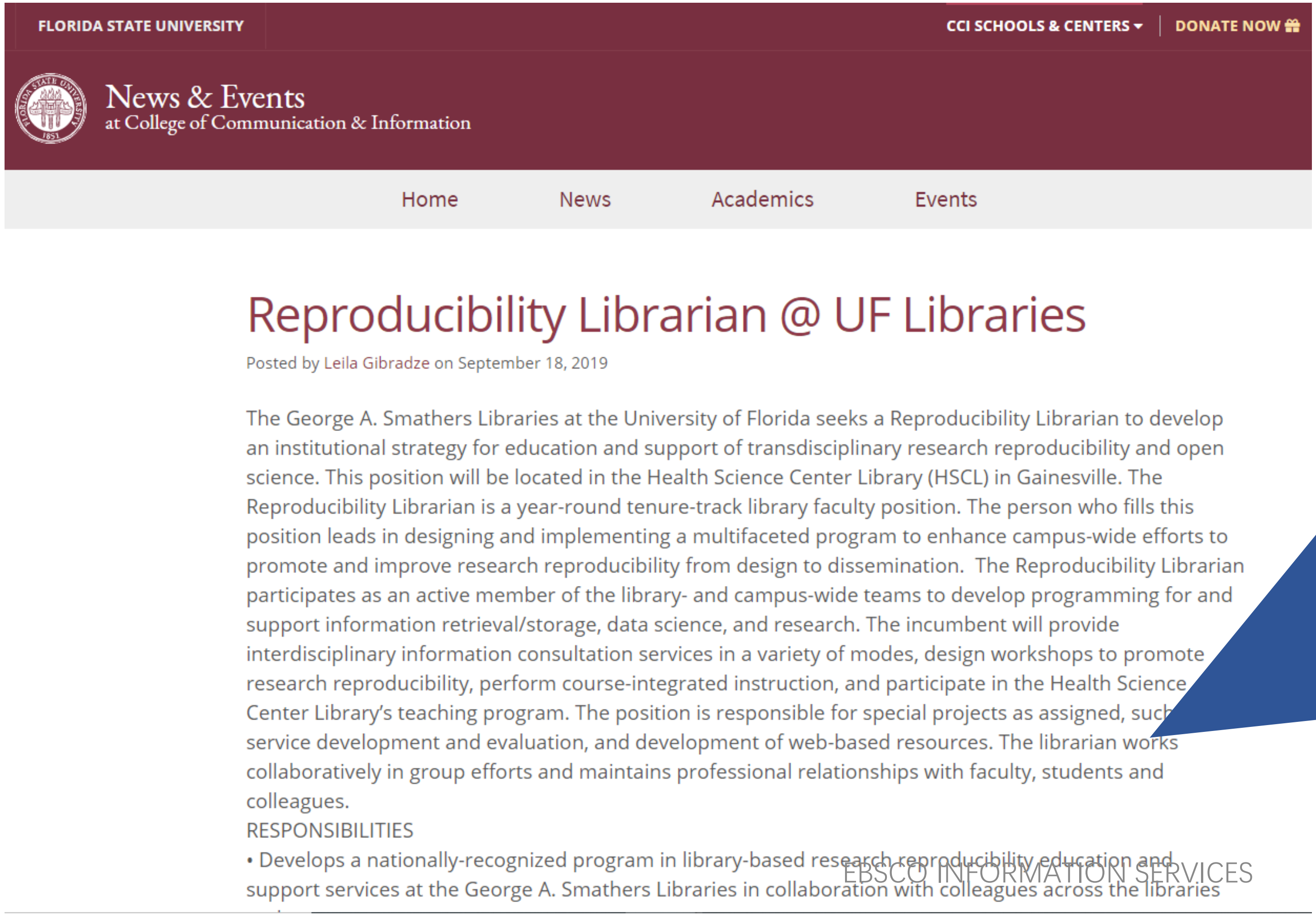


*Table 2: Foundational Terms for Transparency in Open Science.*

| Term | Exposition |
|---|---|
| Transparency | The outcome from a suite of behaviours which characterize Reproducible Research |
| Transparency | Facilitates and enhances Research Quality, Research Integrity and Trust |
| Transparency Action | Describes a specific intervention which is a component of the processes, protocols and practices within the Research Lifecycle |
| 数据透明化代理人 | Exemplified by the Data Science roles e.g. 数据图书馆员. These are key components of the Data Fabric (RDA) and supporting Infrastructure; they promote and demonstrate specific behaviours and practices which lead to culture change towards Open Science |
| Transparency Tool | The software and model frameworks which support Open Science practice |

# 美国学术图书馆出现新型的图书馆职位-研究再现图书馆员

## 美国佛罗里达大学图书馆2019年9月招聘一位“研究再现”图书馆员



### Reproducibility Librarian @ UF Libraries

Posted by Leila Gibradze on September 18, 2019

The George A. Smathers Libraries at the University of Florida seeks a Reproducibility Librarian to develop an institutional strategy for education and support of transdisciplinary research reproducibility and open science. This position will be located in the Health Science Center Library (HSCL) in Gainesville. The Reproducibility Librarian is a year-round tenure-track library faculty position. The person who fills this position leads in designing and implementing a multifaceted program to enhance campus-wide efforts to promote and improve research reproducibility from design to dissemination.  The Reproducibility Librarian participates as an active member of the library- and campus-wide teams to develop programming for and support information retrieval/storage, data science, and research. The incumbent will provide interdisciplinary information consultation services in a variety of modes, design workshops to promote research reproducibility, perform course-integrated instruction, and participate in the Health Science Center Library's teaching program. The position is responsible for special projects as assigned, such service development and evaluation, and development of web-based resources. The librarian works collaboratively in group efforts and maintains professional relationships with faculty, students and colleagues.

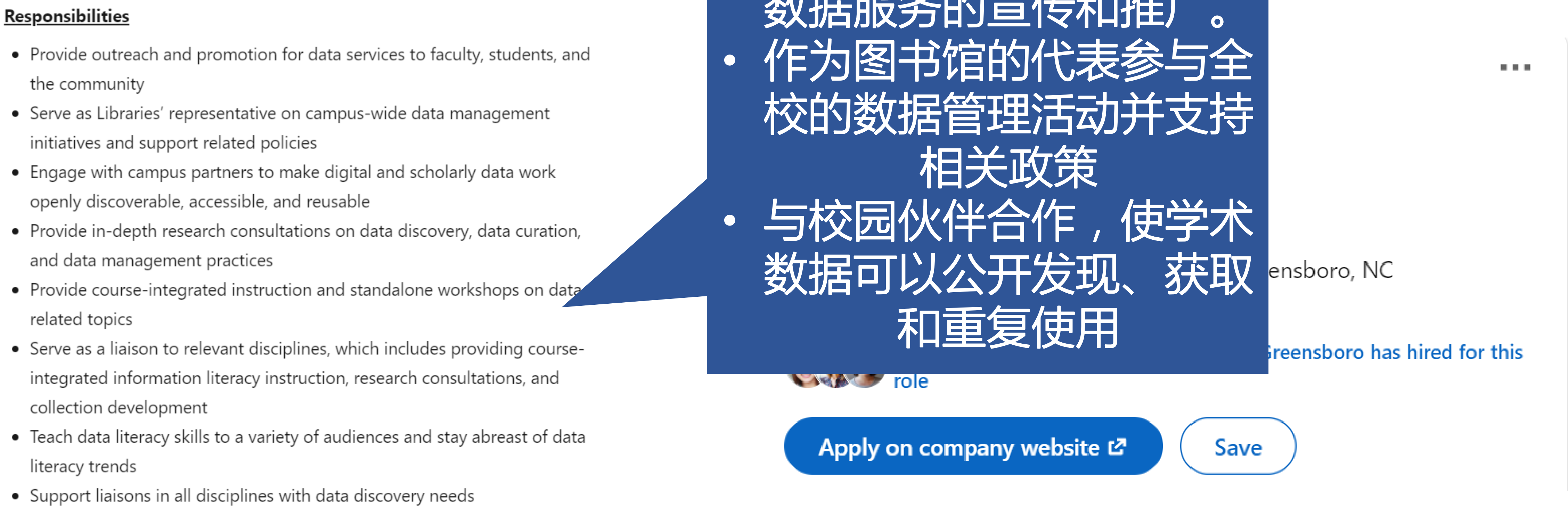
RESPONSIBILITIES

• Develops a nationally-recognized program in library-based research reproducibility education and support services at the George A. Smathers Libraries in collaboration with colleagues across the libraries

- 在George A. Smathers图书馆与整个图书馆和校园的同事合作，开发一个全国公认的以图书馆为基础的研究可重复性教育和支持服务项目。
- 为佛罗里达大学的学生、教职员工提供研究可重复性和开放科学领域的专业知识和咨询服务。
- 设计并提供关于提高研究可重复性和进行开放科学的技术指导方案，包括实验和计算分析的记录和保存

# 新型图书馆职位-数据服务图书馆员@北卡罗来纳大学格林斯博罗分校2021年2月

- 向教师、学生和社区提供数据服务的宣传和推广。
- 作为图书馆的代表参与全校的数据管理活动并支持相关政策
- 与校园伙伴合作，使学术数据可以公开发现、获取和重复使用



**Responsibilities**

- Provide outreach and promotion for data services to faculty, students, and the community
- Serve as Libraries' representative on campus-wide data management initiatives and support related policies
- Engage with campus partners to make digital and scholarly data work openly discoverable, accessible, and reusable
- Provide in-depth research consultations on data discovery, data curation, and data management practices
- Provide course-integrated instruction and standalone workshops on data related topics
- Serve as a liaison to relevant disciplines, which includes providing course-integrated information literacy instruction, research consultations, and collection development
- Teach data literacy skills to a variety of audiences and stay abreast of data literacy trends
- Support liaisons in all disciplines with data discovery needs

03

# 支持开放科学的工具

# Code Ocean ：

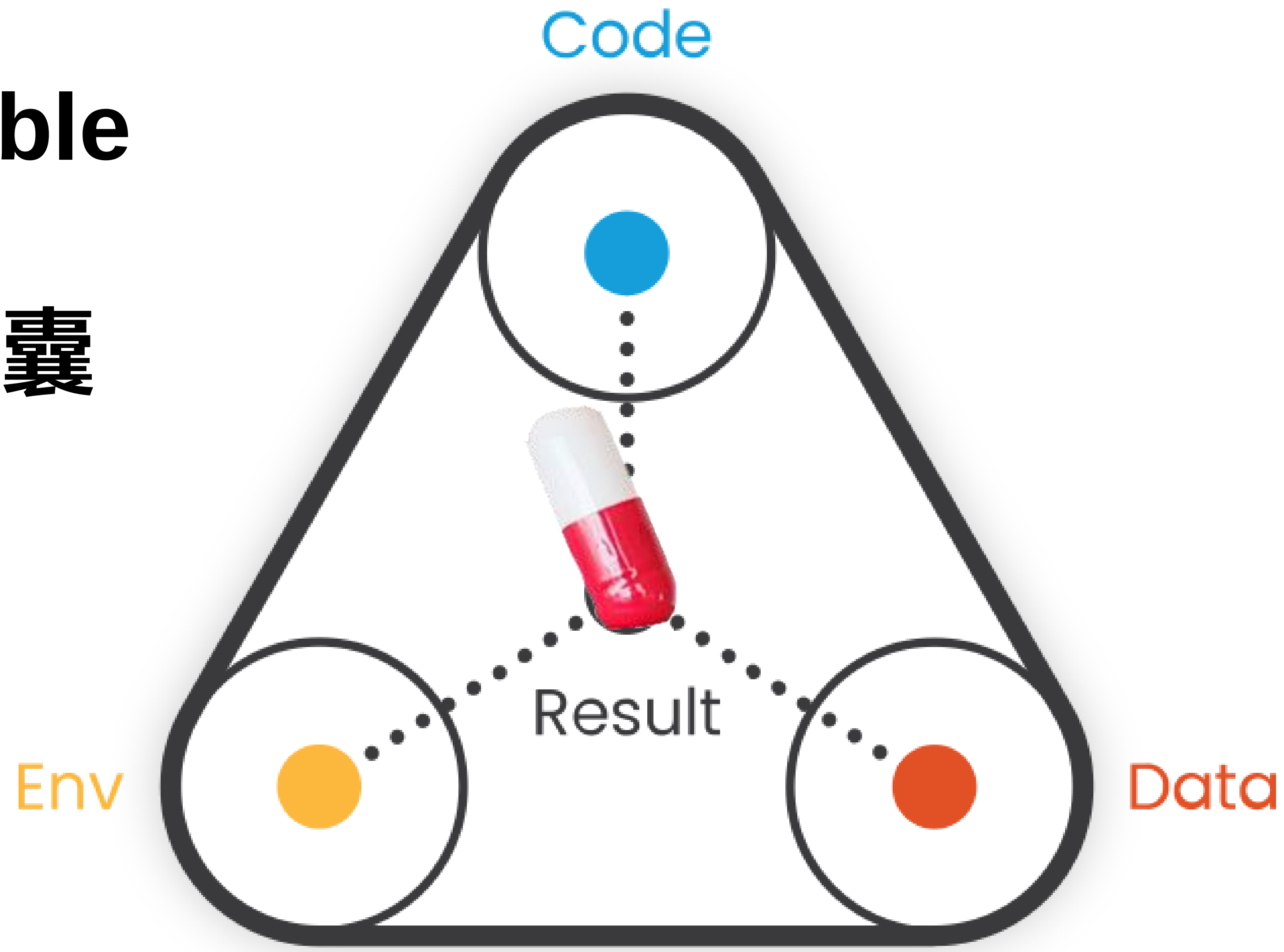
A centralized platform for the creation, sharing, publication, preservation and reuse of executable code and data.

Code Ocean是一个创建、共享、发布、保存和重复使用可执行代码和数据的集中平台。

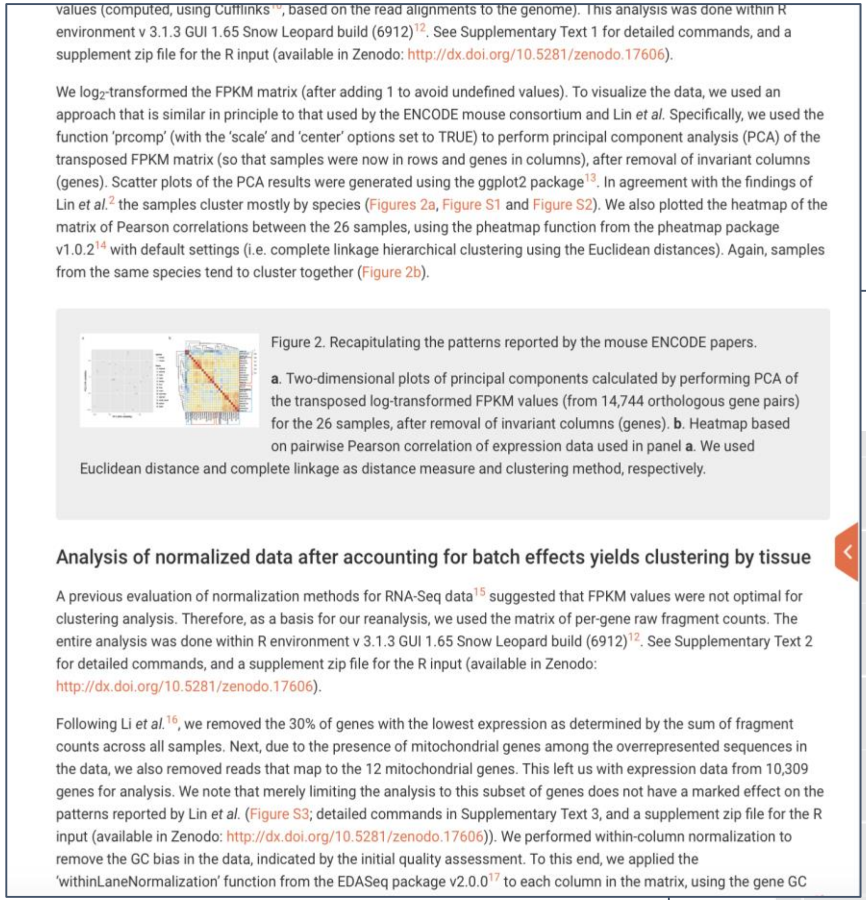
# Code Ocean ：

Code Ocean 平台可以将研究产出做成一个标准、安全和可执行的研究包，称为「Capsule胶囊」

# Reproducible Capsule
# 可再现的胶囊



# Code Ocean ：



values (computed, using Cufflinks[10], based on the read alignments to the genome). This analysis was done within R environment v 3.1.3 GUI 1.65 Snow Leopard build (6912)[12]. See Supplementary Text 1 for detailed commands, and a supplement zip file for the R input (available in Zenodo: http://dx.doi.org/10.5281/zenodo.17606).

We $\log_2$-transformed the FPKM matrix (after adding 1 to avoid undefined values). To visualize the data, we used an approach that is similar in principle to that used by the ENCODE mouse consortium and Lin *et al.* Specifically, we used the function 'prcomp' (with the 'scale' and 'center' options set to TRUE) to perform principal component analysis (PCA) of the transposed FPKM matrix (so that samples were now in rows and genes in columns), after removal of invariant columns (genes). Scatter plots of the PCA results were generated using the ggplot2 package[13]. In agreement with the findings of Lin *et al.*[2] the samples cluster mostly by species (Figures 2a, Figure S1 and Figure S2). We also plotted the heatmap of the matrix of Pearson correlations between the 26 samples, using the pheatmap function from the pheatmap package v1.0.2[14] with default settings (i.e. complete linkage hierarchical clustering using the Euclidean distances). Again, samples from the same species tend to cluster together (Figure 2b).

Figure 2. Recapitulating the patterns reported by the mouse ENCODE papers.

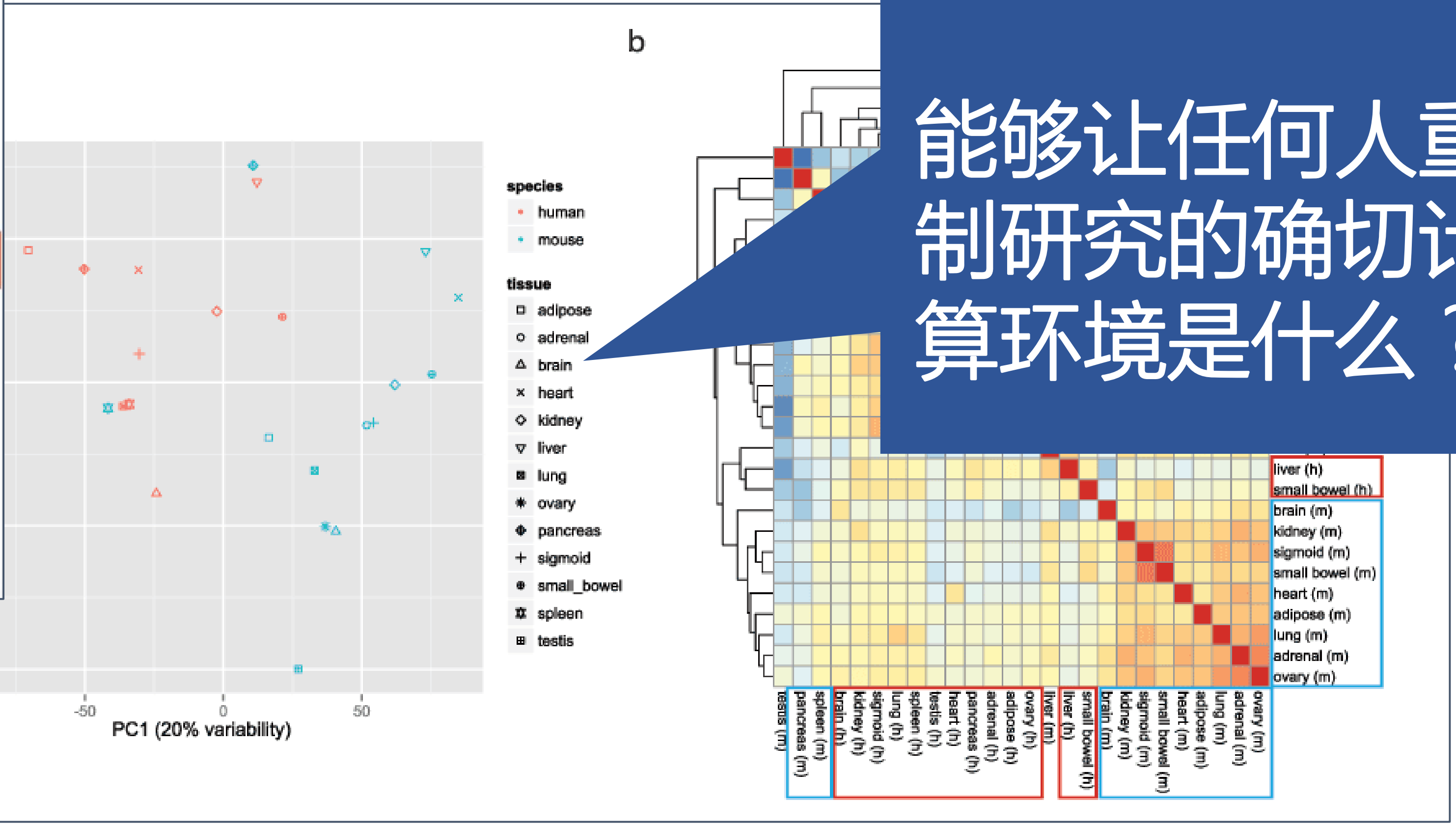
**a**. Two-dimensional plots of principal components calculated by performing PCA of the transposed log-transformed FPKM values (from 14,744 orthologous gene pairs) for the 26 samples, after removal of invariant columns (genes). **b**. Heatmap based on pairwise Pearson correlation of expression data used in panel **a**. We used Euclidean distance and complete linkage as distance measure and clustering method, respectively.

### Analysis of normalized data after accounting for batch effects yields clustering by tissue

A previous evaluation of normalization methods for RNA-Seq data[15] suggested that FPKM values were not optimal for clustering analysis. Therefore, as a basis for our reanalysis, we used the matrix of per-gene raw fragment counts. The entire analysis was done within R environment v 3.1.3 GUI 1.65 Snow Leopard build (6912)[12]. See Supplementary Text 2 for detailed commands, and a supplement zip file for the R input (available in Zenodo: http://dx.doi.org/10.5281/zenodo.17606).

Following Li *et al.*[16], we removed the 30% of genes with the lowest expression as determined by the sum of fragment counts across all samples. Next, due to the presence of mitochondrial genes among the overrepresented sequences in the data, we also removed reads that map to the 12 mitochondrial genes. This left us with expression data from 10,309 genes for analysis. We note that merely limiting the analysis to this subset of genes does not have a marked effect on the patterns reported by Lin *et al.* (Figure S3; detailed commands in Supplementary Text 3, and a supplement zip file for the R input (available in Zenodo: http://dx.doi.org/10.5281/zenodo.17606)). We performed within-column normalization to remove the GC bias in the data, indicated by the initial quality assessment. To this end, we applied the 'withinLaneNormalization' function from the EDASeq package v2.0.0[17] to each column in the matrix, using the gene GC

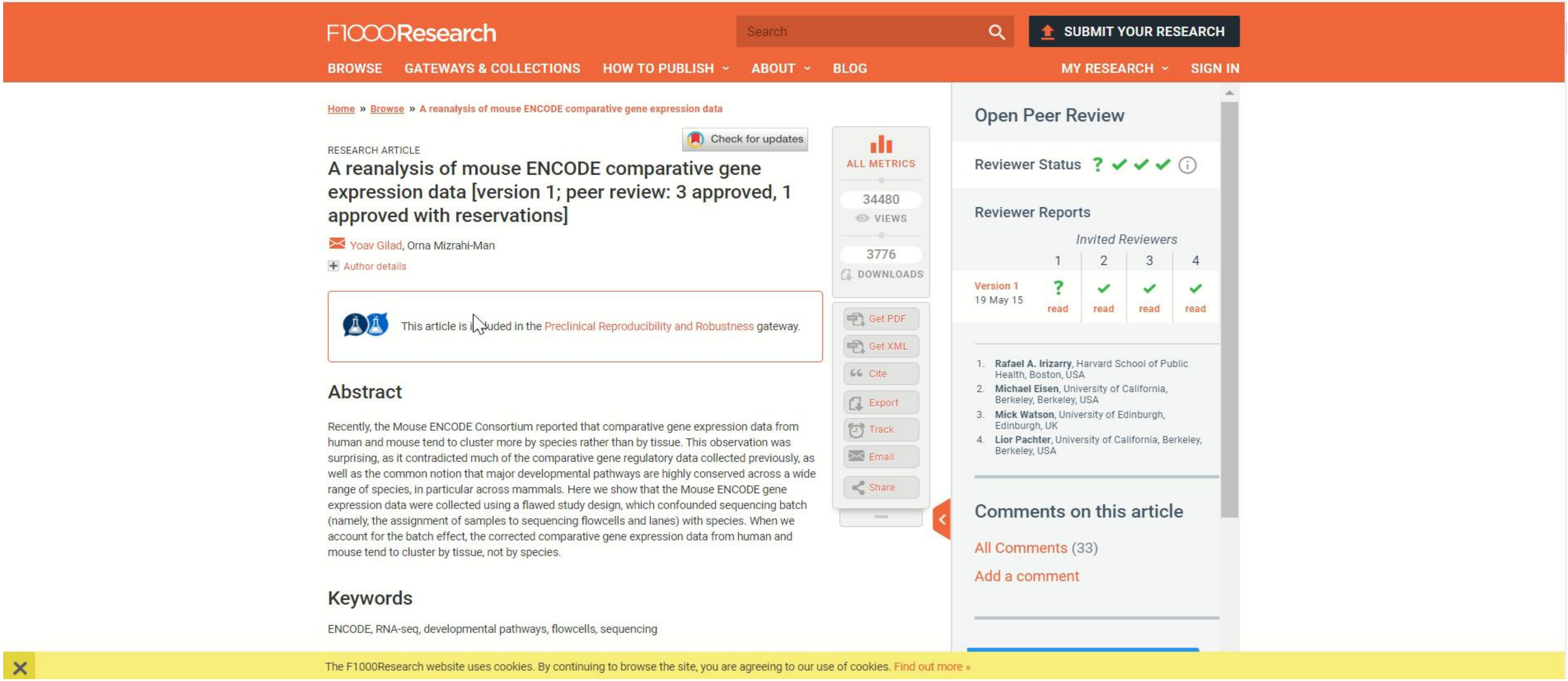
F1000Research



原始程序代码和数据在哪里？

能够让任何人重制研究的确切计算环境是什么？

# Code Ocean ：



Home » Browse » A reanalysis of mouse ENCODE comparative gene expression data

RESEARCH ARTICLE

A reanalysis of mouse ENCODE comparative gene expression data [version 1; peer review: 3 approved, 1 approved with reservations]

Yoav Gilad, Orna Mizrahi-Man

This article is included in the Preclinical Reproducibility and Robustness gateway.

## Abstract

Recently, the Mouse ENCODE Consortium reported that comparative gene expression data from human and mouse tend to cluster more by species rather than by tissue. This observation was surprising, as it contradicted much of the comparative gene regulatory data collected previously, as well as the common notion that major developmental pathways are highly conserved across a wide range of species, in particular across mammals. Here we show that the Mouse ENCODE gene expression data were collected using a flawed study design, which confounded sequencing batch (namely, the assignment of samples to sequencing flowcells and lanes) with species. When we account for the batch effect, the corrected comparative gene expression data from human and mouse tend to cluster by tissue, not by species.

## Keywords

ENCODE, RNA-seq, developmental pathways, flowcells, sequencing

## Open Peer Review

Reviewer Status ? ✓ ✓ ✓

Reviewer Reports

| | Invited Reviewers | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Version 1<br>19 May 15 | ?<br>read | ✓<br>read | ✓<br>read | ✓<br>read |

1. **Rafael A. Irizarry**, Harvard School of Public Health, Boston, USA
2. **Michael Eisen**, University of California, Berkeley, Berkeley, USA
3. **Mick Watson**, University of Edinburgh, Edinburgh, UK
4. **Lior Pachter**, University of California, Berkeley, Berkeley, USA

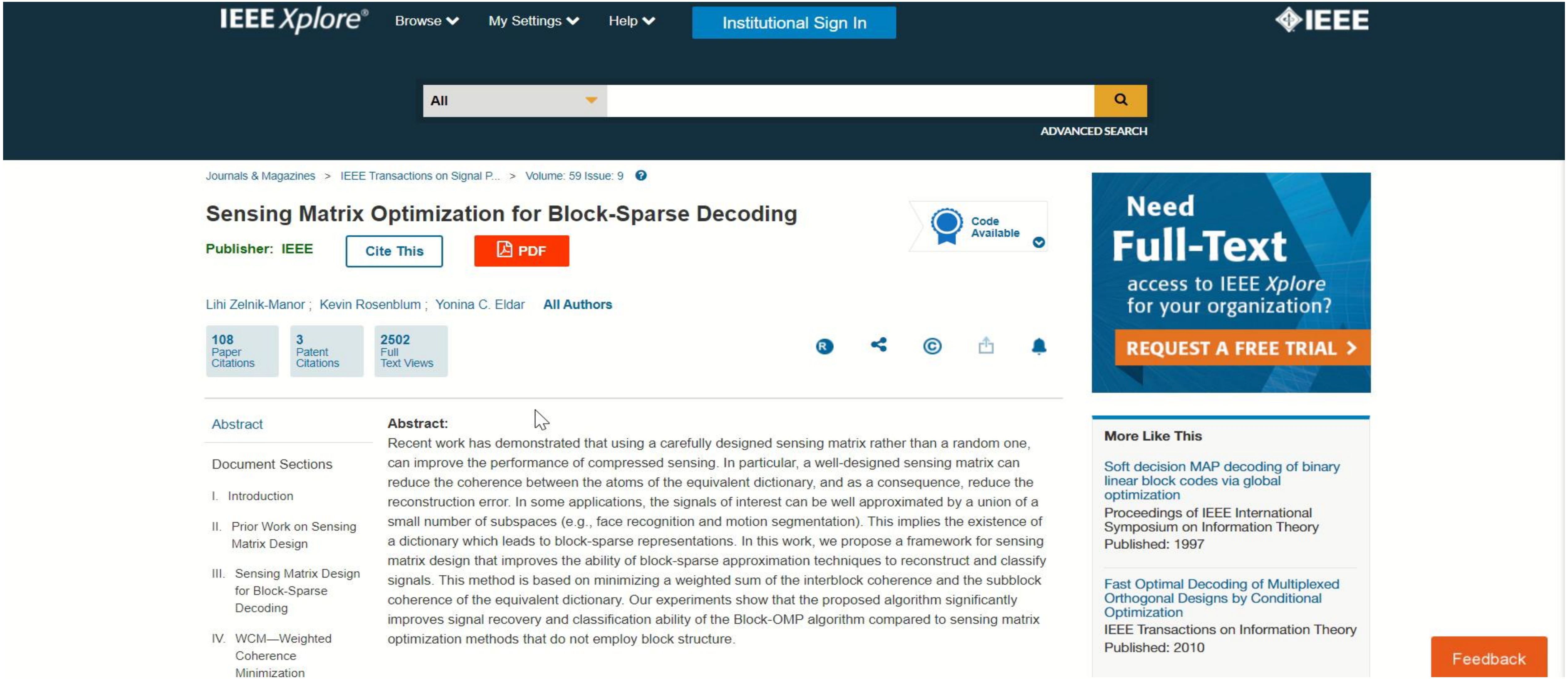
## Comments on this article

All Comments (33)

Add a comment

The F1000Research website uses cookies. By continuing to browse the site, you are agreeing to our use of cookies. Find out more »

# Code Ocean ：



Journals & Magazines > IEEE Transactions on Signal P... > Volume: 59 Issue: 9

## Sensing Matrix Optimization for Block-Sparse Decoding

Publisher: IEEE

Lihi Zelnik-Manor ; Kevin Rosenblum ; Yonina C. Eldar All Authors

Abstract

Document Sections

I. Introduction

II. Prior Work on Sensing Matrix Design

III. Sensing Matrix Design for Block-Sparse Decoding

IV. WCM—Weighted Coherence Minimization

**Abstract:**

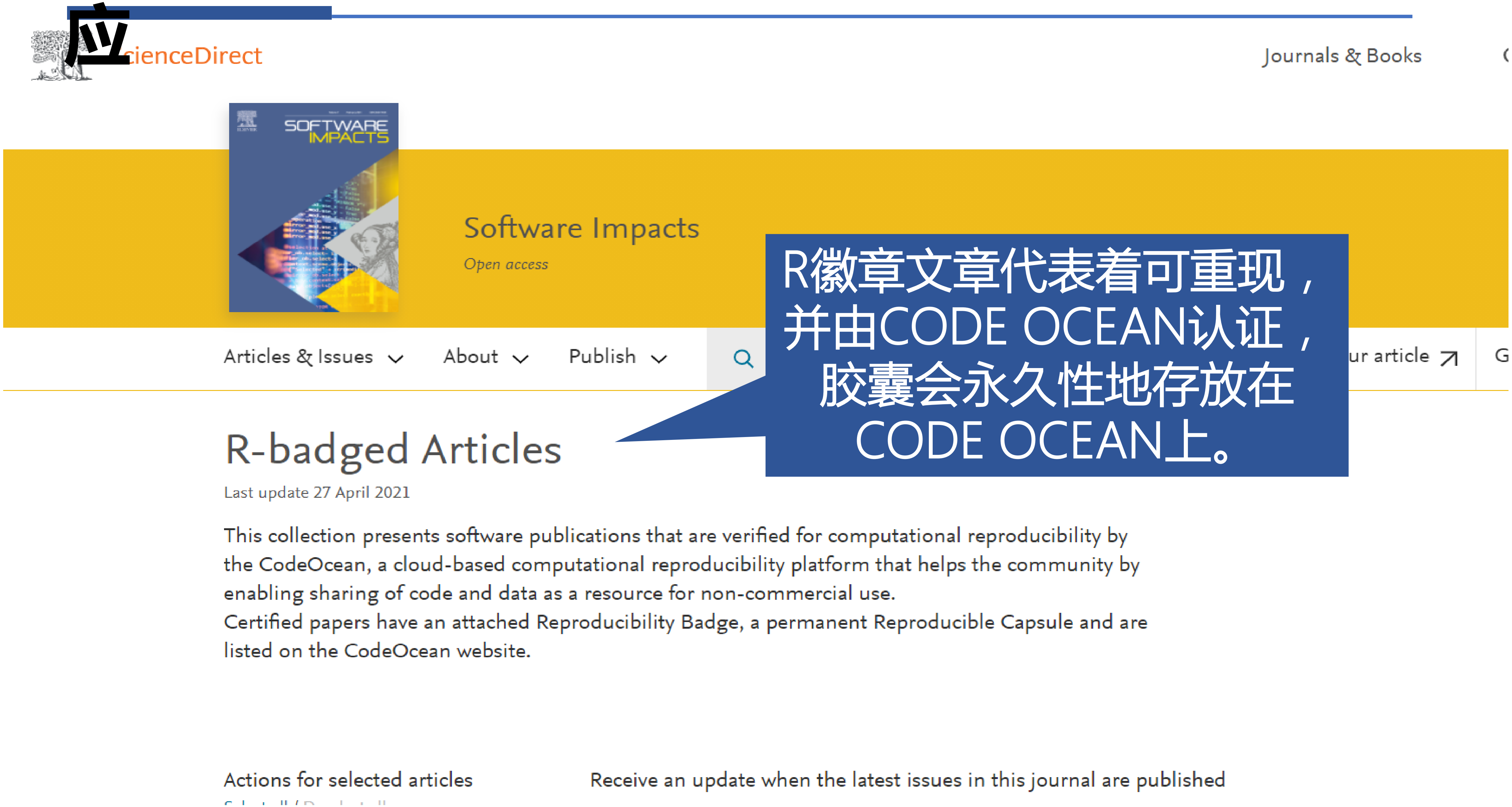
Recent work has demonstrated that using a carefully designed sensing matrix rather than a random one, can improve the performance of compressed sensing. In particular, a well-designed sensing matrix can reduce the coherence between the atoms of the equivalent dictionary, and as a consequence, reduce the reconstruction error. In some applications, the signals of interest can be well approximated by a union of a small number of subspaces (e.g., face recognition and motion segmentation). This implies the existence of a dictionary which leads to block-sparse representations. In this work, we propose a framework for sensing matrix design that improves the ability of block-sparse approximation techniques to reconstruct and classify signals. This method is based on minimizing a weighted sum of the interblock coherence and the subblock coherence of the equivalent dictionary. Our experiments show that the proposed algorithm significantly improves signal recovery and classification ability of the Block-OMP algorithm compared to sensing matrix optimization methods that do not employ block structure.

**More Like This**

Soft decision MAP decoding of binary linear block codes via global optimization
Proceedings of IEEE International Symposium on Information Theory
Published: 1997

Fast Optimal Decoding of Multiplexed Orthogonal Designs by Conditional Optimization
IEEE Transactions on Information Theory
Published: 2010

# 出版社对于“Reproducibility可重现性”的响应



Software Impacts

*Open access*

R徽章文章代表着可重现，并由CODE OCEAN认证，胶囊会永久性地存放在CODE OCEAN上。

## R-badged Articles

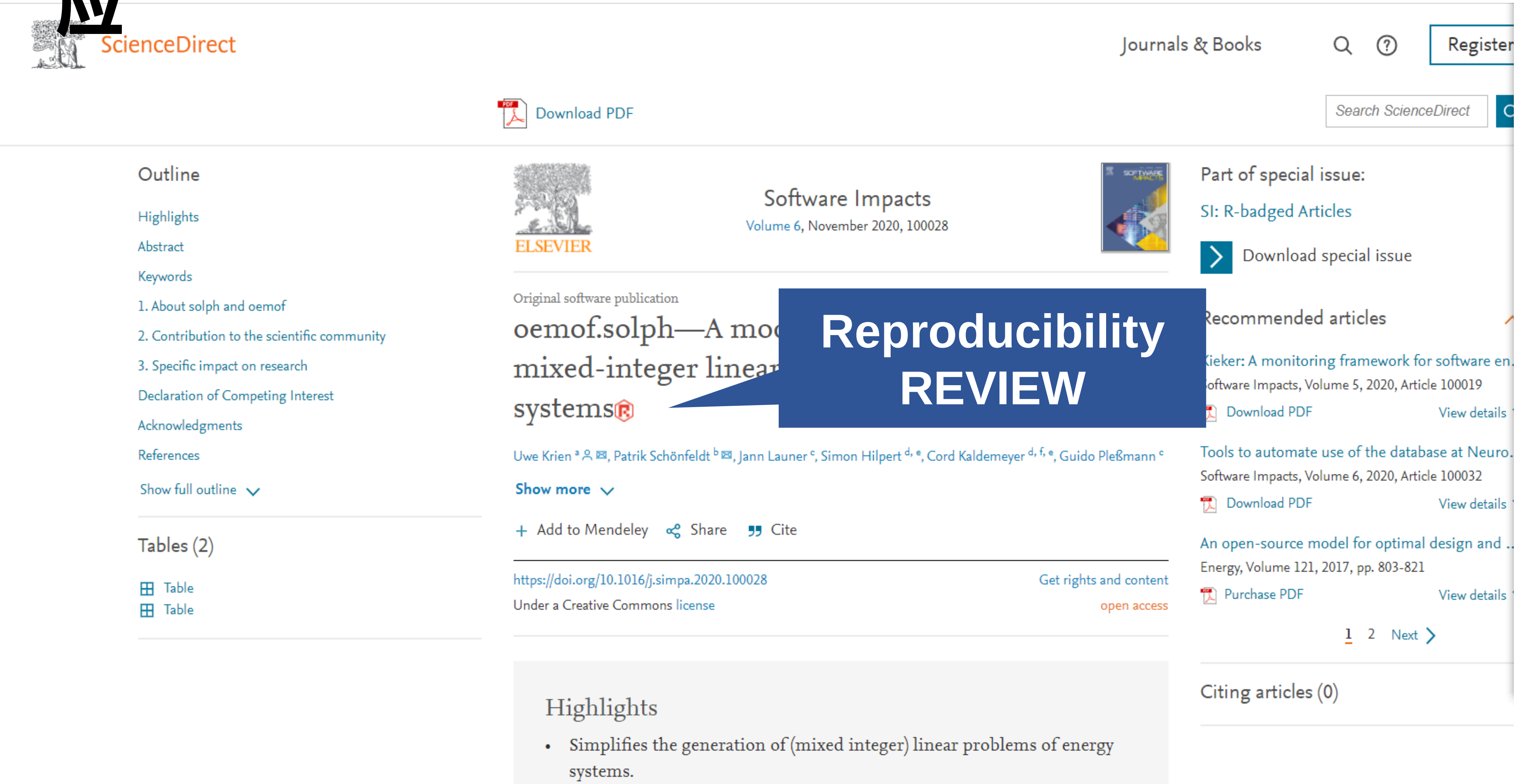
Last update 27 April 2021

This collection presents software publications that are verified for computational reproducibility by the CodeOcean, a cloud-based computational reproducibility platform that helps the community by enabling sharing of code and data as a resource for non-commercial use.
Certified papers have an attached Reproducibility Badge, a permanent Reproducible Capsule and are listed on the CodeOcean website.

Actions for selected articles

Receive an update when the latest issues in this journal are published

# 出版社对于 “Reproducibility可重现性”的响应



ScienceDirect

Journals & Books

Register

Download PDF

Search ScienceDirect

Outline

Highlights

Abstract

Keywords

1. About solph and oemof

2. Contribution to the scientific community

3. Specific impact on research

Declaration of Competing Interest

Acknowledgments

References

Show full outline

Tables (2)

⊞ Table

⊞ Table

ELSEVIER

Software Impacts

Volume 6, November 2020, 100028

Original software publication

oemof.solph—A mo
mixed-integer linea
systems

Reproducibility REVIEW

Uwe Krien [a], Patrik Schönfeldt [b], Jann Launer [c], Simon Hilpert [d, e], Cord Kaldemeyer [d, f, e], Guido Pleßmann [c]

Show more

+ Add to Mendeley  Share  Cite

https://doi.org/10.1016/j.simpa.2020.100028

Get rights and content

Under a Creative Commons license

open access

Highlights

- Simplifies the generation of (mixed integer) linear problems of energy systems.

Part of special issue:

SI: R-badged Articles

Download special issue

Recommended articles

Kieker: A monitoring framework for software en…

Software Impacts, Volume 5, 2020, Article 100019

Download PDF  View details

Tools to automate use of the database at Neuro…

Software Impacts, Volume 6, 2020, Article 100032

Download PDF  View details

An open-source model for optimal design and …

Energy, Volume 121, 2017, pp. 803-821

Purchase PDF  View details

1 2 Next

Citing articles (0)

# Code Ocean ：

CODE OCEAN

Allows anyone to find, create and share code and data, without the need for local deployment

允许任何人查找、创建和共享代码和数据，而无需本地部署

PRINCETON UNIVERSITY

HARVARD UNIVERSITY

PennState

Penn University of Pennsylvania

Columbia University

# protocols.io

protocols.io is a centralized platform for the collaborative creation, management and sharing of research methods.

protocols.io 是协作创造、管理和共享研究方法的集中平台

# 文章的问题：

## 学术文章往往在其研究方法的描述有所欠缺，特别是详尽的步骤、细节或组成的成分

“The hardest part, by far, was figuring out exactly what the original labs actually did. Scientific papers come with methods sections that theoretically ought to provide recipes for doing the same experiments. But often, those recipes are **incomplete**, **missing important steps**, **details**, or **ingredients**. In some cases, the recipes aren’t described at all”

The Atlantic

protocols.io

# Many organizations encourage the use of protocols.io

Journals and publishers recommend protocols.io on manuscript submission

AACR American Association for Cancer Research®

PLOS

GSA® Genetics Society of America

eLIFE

**500+ journals**

Funders require or recommend protocols.io in grant guidelines/policies

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST

Alex's Lemonade Stand

CZ

GORDON AND BETTY MOORE FOUNDATION

NIH

W wellcome

# protocols.io

protocols.io

Allows anyone to find, create and share research methods to support reproducibility and collaborations

允许任何人搜寻、创建和共享研究方法，以支持可重复性与协同合作

Carnegie Mellon University

Stanford University

FRED HUTCH
CURES START HERE

UNIVERSITY OF CALIFORNIA

# 结语：

# 参考文献：

Miyakawa, T. No raw data, no science: another possible source of the reproducibility crisis. *Mol Brain* 13, 24 (2020). https://doi.org/10.1186/s13041-020-0552-2

LYON, L. Transparency: The Emerging Third Dimension of Open Science and Open Data. Liber Quarterly: The Journal of European Research Libraries, [s. l.], v. 25, n. 4, p. 153–171, 2016. DOI 10.18352/lq.10113. Disponível em: http://search.ebscohost.com/login.aspx?direct=true&db=lls&AN=114854749&site=ehost-live&scope=site. Acesso em: 25 abr. 2021.

Benedikt Fasel (1), Jörg Spörri (2,3), Josef Kröll (3), Kamiar Aminian (1) 2017. Functional calibration for trunk and lower limb fixed inertial sensors. protocols.iohttps://dx.doi.org/10.17504/protocols.io.itrcem6

Butler, D., Karpowitz, C., & Pope, J. (2017). Who Gets the Credit? Legislative Responsiveness and Evaluations of Members, Parties, and the US Congress. *Political Science Research and Methods, 5*(2), 351-366. doi:10.1017/psrm.2015.83

L. Zelnik-Manor, K. Rosenblum and Y. C. Eldar, "Sensing Matrix Optimization for Block-Sparse Decoding," in IEEE Transactions on Signal Processing, vol. 59, no. 9, pp. 4300-4312, Sept. 2011, doi: 10.1109/TSP.2011.2159211.

# THANKS

谢谢