



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

---

FACULTAD DE CIENCIAS

ANÁLISIS DE LA CIENCIOMETRÍA EN BIOCIENCIAS

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

MATEMÁTICAS APLICADAS

PRESENTA:

MINERVA MARÍA ROMERO PÉREZ

DIRECTORA DE TESIS:

DRA. LAYLA MICHÁN AGUIRRE



Ciudad Universitaria, Ciudad de México 2024



# Índice general

<b>1</b>	<b>Introducción</b>	<b>1</b>
§1.1	Lenguaje natural y lenguaje máquina . . . . .	1
§1.2	Los metadatos . . . . .	2
§1.3	Las bases de datos bibliográficas . . . . .	3
§1.4	PubMed . . . . .	4
§1.5	Las APIs . . . . .	4
§1.6	Revisiones de literatura . . . . .	6
§1.7	La cienciometría . . . . .	6
§1.8	Leyes e indicadores cienciométricos . . . . .	6
§1.8.1	Ley de Lotka . . . . .	6
§1.8.2	Ley de branford . . . . .	7
§1.8.3	Ley de Zipf . . . . .	7
§1.9	Cienciometría para ciencias de la vida y de la salud . . . . .	7
§1.10	Análisis de redes . . . . .	7
§1.11	Objetivos . . . . .	7
<b>2</b>	<b>Material y método</b>	<b>9</b>
§2.1	Planeación . . . . .	9
§2.2	Recuperación de la literatura en bases de datos . . . . .	9
§2.3	Definición de criterios de inclusión, de exclusión y las variables . . . . .	10
§2.3.1	Metadatos bibliográficos . . . . .	11
§2.4	Extracción de los datos . . . . .	13
§2.5	Procesamiento de la información de la literatura . . . . .	13
§2.6	Análisis . . . . .	13

§2.7 Visualización de datos . . . . .	14
§2.8 Integración y contextualización de la información . . . . .	14
<b>3 Resultados</b>	<b>15</b>
<b>4 Discusión</b>	<b>17</b>

# 1 Introducción

La información es una fuente importante para hacer investigación se (Hull et al., 2008)  
Las biociencias son ... Los análisis métricos de las publicaciones científicas ha crecido  
junto con la tecnología informática e informacional en la Web...

Entre métodos más usados para realizar los análisis están: la minería de textos, la  
ciencia de datos, los análisis de redes, los análisis métricos y los grandes datos...

instituciones e investigadores que utilicen la cienciometría...

Actualmente, existen instituciones como el Centre for Science and Technology Studies  
(CWTS) traducido al español como Centro de Estudios de Ciencia y Tecnología, en el  
cual con la bibliometría y la cienciometría estudian la investigación científica ([https://  
www.cwts.nl/about-cwts](https://www.cwts.nl/about-cwts)) ... Silvio Peroni... Ludo Wltman... Daniel Torres Salinas ...  
La ciencia abierta es un ...

## 1.1. Lenguaje natural y lenguaje máquina

El lenguaje natural es la forma en la que las personas hablamos y vemos la información,  
esta es la que entendemos y es como visualizamos comúnmente las cosas en la web. Mientras  
que el lenguaje máquina es estructurado, interoperable y no es comprensible a simple vista,  
es leído por computadoras y es como funcionan los programas y la web.

## 1.2. Los metadatos

Los metadatos son comúnmente conocidos como los datos sobre los datos. Estos constituyen Información estructurada que describe, explica, localiza y hace que sea más fácil de recuperar, utilizar o manejar la información (NISO,XX). Estos pueden ser estructurales, descriptivos o administrativos. Los estructurales son ... (Gartner, 2016)(Pomerantz, 2015)

El uso de metadatos adecuados permiten una correcta estructuración y descripción de los datos, puesto que detallan las características de la información que se genera, como la fecha de creación, la licencia de uso, el autor, el tipo de objeto digital (texto, imagen, sonido, video) o la dirección electrónica en la que se encuentra alojado.

Los metadatos bibliográficos son útiles para el fácil procesamiento de la información, ayudan al estudio de las publicaciones científicas, con ellos se pueden generar análisis para ver patrones y responder preguntas científicas sobre las investigaciones. Este tipo de metadatos se originan a partir de la información proporcionada por los autores, revistas y bases de datos bibliográficas.

Cada revista científica tiene una fecha de publicación de sus números, lo que genera los metadatos de fecha de publicación, las bases de datos generan los metadatos de cuando reciben y agregan las publicaciones, entonces se van generando distintos metadatos con fechas, para realizar una revisión de literatura y tomar el conjunto de publicaciones se tienen que tener en consideración estas, ya que hay que tener muy bien definido el intervalo de tiempo del que se van a obtener los textos.

La diferencia entre la lectura natural y la lectura máquina de los metadatos se muestra en la figura 1, en la cual se puede ver que el lenguaje natural es como encontramos comúnmente la información de la Web, mientras que el lenguaje máquina muy pocos lo conocen, pero es la forma en la que se maneja la información para su análisis, manejo y procesamiento.

Formato RIS son las siglas de la empresa que lo desarrollo Research Information Systems, en español Sistemas de información de investigación, es un formato de etiquetado en el cual se separan los metadatos, se agrega la etiqueta espacio, guion, espacio y el metadato

asociado a la etiqueta. Se utiliza para el intercambio de información de citas. Por ejemplo, los gestores bibliográficos aceptan este formato para manejar las citas, y las bases de datos bibliográficas manejan este para tener los metadatos y poder exportarlos.

### 1.3. Las bases de datos bibliográficas

Las bases de datos son fuentes en las cuales permiten procesar, guardar, ordenar, mantener y presentar datos. Son sistematizadas, presentan la información en tablas con un orden establecido, por lo que se tiene mucha información en poco espacio. Las bases de datos permiten realizar búsquedas a partir de distintos criterios, procesar datos de forma cuantitativa y cualitativa, interrelacionar los resultados utilizando distintas variables y actualizar la información fácil y rápidamente.

Las bases de datos se extienden por módulos y se pueden diseñar, almacenar, manejar y analizar en una computadora, son compatibles con otras bases de datos, en estas se pueden hacer consultas ('queries') para realizar cálculos, las bases de datos relacionales son las más utilizadas, presentan los datos en dos o más tablas (constituidas por registros o renglones y columnas) que están relacionadas entre sí por medio de la repetición de un atributo. Las bases de datos pueden ser estáticas en donde la información no varía en el tiempo o dinámicas las cuales se actualiza, cambia o elimina la información después de algún tiempo.

Las bases de datos biológicas son una herramienta indispensable para el quehacer de los biólogos, estas permiten...

Las bases de datos bibliográficas se encargan de manejar información estructurada, sobre los artículos, libros, capítulos o memorias. Las bases de datos bibliográficas manejan los metadatos dados por los autores y revistas, algunas agrega metadatos como identificadores, ontologías y tesauros.

Existen muchas bases de datos bibliográficas, desde especializadas en áreas como la medicina, que solo contienen artículos o solo libros y hay multidisciplinarias y que contienen cualquier tipo de publicación como artículos, preprints, libros, patentes, etc. Entonces es importante saber qué información contienen y qué información necesitamos para elegir la

o las más relevantes para trabajar. Entre las características más importantes a considerar están: criterios de inclusión de documentos, contenido (tipo de registros), tipo de acceso, la cobertura (representación) tipológica, temática, geográfica y temporal de la colección. Dentro de las bases de datos biológicas se encuentran las bases de datos bibliográficas sobre biociencia, estas resultan útiles para generar nuevo conocimiento. Entonces algo fundamental en la Web son los datos ligados, estos nos permiten tener la información conectada desde diferentes enfoques ...

Indexar es agregar a un catálogo cualquier tipo de información, por ejemplo revistas, artículos, términos estructurados, etc. Entonces se tiene que las bases de datos bibliográficas indexan revistas de acuerdo a ciertos criterios que implementa la base de datos.

## 1.4. PubMed

Una de las bases de datos bibliográficas más importantes para la medicina es PubMed, una de sus características más importantes es que es de acceso abierto, es amigable y tienen una gran cantidad de herramientas, estas características y más pueden consultarse en ... (artículo herramientas de pubmed texto, <https://zenodo.org/records/10659207>). Para este trabajo es de interés el tesauro MeSH asociado a PubMed, este es un vocabulario controlado ...

## 1.5. Las APIs

Una API o Interfaz de Programación de Aplicaciones (por sus siglas en inglés, Application Programming Interface), es un conjunto de reglas y protocolos que permiten que diferentes aplicaciones o sistemas informáticos se comuniquen entre sí. Estas definen los métodos y datos que las aplicaciones pueden utilizar para interactuar unas con otras de manera estandarizada y predecible.

Las APIs son esenciales para la integración y la interoperabilidad entre sistemas de software, y se utilizan en una variedad de contextos, incluyendo aplicaciones web, aplicaciones móviles, sistemas operativos, bases de datos, servicios en la nube y más. Algunos



de los conceptos clave relacionados con las APIs son:

- Estándares y protocolos: para garantizar la consistencia y la compatibilidad.
- Interfaz de usuario vs. interfaz de programación: en donde la interfaz de usuario permite que las personas interactúen con la aplicación, mientras que la API o interfaz de programación permite que los programas y sistemas interactúen entre sí sin necesidad de intervención humana
- Datos estructurados: las APIs suelen transmitir datos en un formato estructurado, esto puede ser en formato JASON (JavaScript Object Notation) o XML (eXtensible Markup Language) y facilitan la interpretación y el procesamiento de la información por parte de las aplicaciones
- Seguridad: las APIs pueden requerir autenticación y autorización para garantizar que solo las aplicaciones autorizadas puedan acceder a sus servicios. Esto es fundamental para proteger la privacidad y la seguridad de los datos.
- Documentación: Las buenas APIs suelen estar acompañadas de documentación detallada que describe cómo utilizarlas, qué métodos están disponibles y qué datos se pueden esperar en las respuestas.

Las APIs son utilizadas en una amplia gama de aplicaciones, desde la integración de servicios de terceros en aplicaciones móviles hasta la automatización de tareas en sistemas empresariales. Permiten a los desarrolladores aprovechar las funcionalidades de otros sistemas de manera eficiente, lo que acelera el desarrollo de software y la creación de aplicaciones más poderosas y versátiles.

En este trabajo nos interesa explorar las APIs de bases de datos bibliográficas, ya que estas son las que nos van a permitir extraer metadatos de las publicaciones de una manera eficiente y rápida para poder realizar análisis de la información

Adicionalmente, PubMed se tiene disponible la API (Application Programming Interface) que sea la cual define los métodos y datos que las aplicaciones pueden utilizar para interactuar con la base de datos unas con otras de manera estandarizada y predecible, para

lo que permite para exportar registros y para procesarlos por métodos informáticos que implican programación y conocimiento en el funcionamiento de las aplicaciones. Además, hay una buena documentación disponible con los detalles técnicos, lo que promueve que incluso los usuarios puedan crear aplicaciones.

## 1.6. Revisiones de literatura

Una revisión de literatura es ...

Existen distintos tipos de revisiones de literatura, en capítulo biorevisiones de literatura, X se pueden ver algunas con sus ejemplos y..

## 1.7. La cienciometría

La cienciometría puede considerarse un tipo de revisión de literatura ( capítulo biorevisiones de literatura,X)

Para realizar análisis cienciométricos existen aplicaciones web y software para computadora desde los cuales se puede procesar los datos, algunas de estas herramientas son amigables pero, algunas no ... (XX)

La ciencimetría abierta es ..., las ventajas de esta son (XX)

## 1.8. Leyes e indicadores cienciométricos

Cuando se hace un análisis bibliométrico se utilizan las tres leyes de la bibliometría y los indicadores cienciométricos, las cuales ...

### 1.8.1. Ley de Lotka

La proporción de autores que han publicado  $n$  artículos dentro de un campo científico es inversamente proporcional al cuadrado de  $n$ . De lo cual obtenemos lo siguiente (Urbizagástegui Alvarado, 1999):

$$A(n) = \frac{C_1}{n^2}$$

De donde:  $A(n)$  : es la función que indica la fracción de autores que han publicado  $n$  artículos  $C_1$  : es el número de publicaciones por cada autor  $n$  : el número de publicaciones dentro de un campo científico

### 1.8.2. Ley de branford

### 1.8.3. Ley de Zipf

## 1.9. Cienciometría para ciencias de la vida y de la salud

La cienciometría es

### 1.10. Análisis de redes

Para entender la definición de red es necesario dar una introducción a la teoría de gráficas. Una gráfica es un conjunto  $(X, A)$  donde  $X$  es el conjunto no vacío de elementos llamados nodos o vértices, y  $A$  es el conjunto de pares no ordenados de vértices o nodos que llamados aristas o vértices.

Una gráfica puede ser dirigida o no, le llamamos dirigida a las que cumplen que

Una red es una gráfica dirigida en la cual se le asocia una función  $f$  a los arcos, nodos o ambos (Hernández, 2020). En general, el término red se refiere a problemas reales como redes metabólicas, redes de cooperación en artículos o redes de documentos web (Barabási & Pósfai, 2016).

### 1.11. Objetivos

El objetivo de mi investigación es analizar la bibliografía en Biociencias para entender cómo se ha usado la cienciometría, y otros métodos métricos relacionados, como la bibliometría, la infometría y la altmetría, en la literatura sobre el COVID-19 en los tres enfoques que

son la enfermedad, la pandemia y el virus, en los cuales se ha contribuido en nuevo conocimiento. Además, quiero reunir las herramientas informáticas, los procesos y los protocolos informacionales e informáticos implementados en los análisis métricos para seleccionar los casos más relevantes y la tecnología de vanguardia, como, los grandes datos, el análisis de redes, la semántica, las ontologías, los datos ligados, algoritmos e inteligencia artificial. Este proyecto es multi, intra, trans e interdisciplinario, es original, inédito e integrativo, porque generalmente los análisis bibliométricos estudian un tema científico específico, pero en este caso tengo como objetivo principal estudiar los artículos cientiométricos publicados en biociencias, esto es, haré cientiometría de la cientiometría en ciencias de la vida y de la salud para sistematizarla y sintetizarla con los siguientes propósitos:

- Entender las diferentes formas en las que se ha utilizado la cientiometría para el COVID-19.
- Explorar y explicar las ventajas de hacer cientiometría abierta.
- Reunir y explorar las herramientas digitales amigables y de código disponibles para realizar análisis de literatura y los métodos informáticos más útiles e innovadoras para PubMed.
- Descubrir la mejor manera de integrar, representar y contextualizar la información por medio de las visualizaciones.
- Reunir todos los análisis realizados para México.
- Realizar analisis de redes para ver las tendencias y patrones de los datos.
- Probar y estudiar las leyes de la blibliometría para analizar artículos.

## 2 Material y método

Esta investigación *in silico* centrada en el uso de recursos informáticos, basada en la literatura y sustentada en teorías matemáticas, constó de ocho etapas.

Durante estas etapas utilicé la base de dato bibliográfica PubMed, el lenguaje de programación R con las paqueterías, bibliometrix y ggplot2. Para programar utilicé Rstudio en una computadora Asus VivoBook con sistema operativo de 64 bits, procesador AMD Ryzen 7 4700U y RAM de 16.0 GB.

### 2.1. Planeación

Para recuperar la literatura de una forma eficiente, use las APIs (Application Programming Interface) de las fuentes, para poder usaras, exploré su funcionamiento en distintos recursos como textos, plataformas, videos, etc.

Además, entender el mundo de los análisis de literatura y las bases de datos bibliográficas, revise artículos, publicaciones, entrevistas y vídeos de expertos. El análisis científico es un proceso largo, generalmente realizada por equipos multidisciplinarios, que involucra una gran cantidad de datos, y que ha tenido un impulso a partir de la era de la digitalización, ya que resulta más sencillo tener acceso a las bases de datos de las cuales se extrae la información.

### 2.2. Recuperación de la literatura en bases de datos

Para esta etapa con ayuda del esquema de la Figura 2 se organizan los elementos para hacer una recuperación de literatura, (ri biocolores 2023)

Las biociencias son un campo de estudio muy amplio y por ende intentar buscar literatura de todo este sería una tarea titánica, es por esto que realice una primera búsqueda, esta fue genaral para ver todas las publicaciones sobre cienciometría en PubMed en esta búsqueda utilice términos generales para mostrar y visualizar la gran cantidad de publicaciones que existen. La consulta utilizada fue evolucionando hasta obtener una consulta interesante.

**Consulta 0:** scientometric\* OR bibliometr\* OR altmetr\* OR cybermetr\* OR infometr\*

Ya que se obtuvo un conjunto muy grnade de datos elegí el tema de COVID-19, ya que me parece interesante para encontrar las nuevas tecnologías y métodos innovadores, en estos tres años hay una gran cantidad de análisis métricos que se realizaron desde distinto enfoques como lo son la enfermedad, la pandemia y el virus, a pesar de que también hay publicaciones que fueron publicadas en el periodo de la pandemia estas no las consideré, puesto que no hablan propiamente del COVID-19.

Para este tema obtuve distintas consultas interesantes.

**Consulta 1:** (Scientometr\* OR Bibliometr\* OR Altmotr\* OR Cybermetr\* OR Infometr\* OR entitymetr\*)AND (covid-19 OR SARS-CoV-2)

**Consulta 2:** (("Bibliometrics"[Mesh] OR entitymetr\* OR scientometr\* OR Altmotr\* OR Cybermetr\* OR Infometr\* OR "metrics") AND (COVID-19"[Mesh] OR "SARS-CoV-2"[Mesh])) NOT "Journal Impact Factor"[Mesh]

## 2.3. Definición de criterios de inclusión, de exclusión y las variables

Los artículos que incluí para la etapa dos tenían que cumplir con los siguientes criterios:

- Tener el identificador DOI.
- Tener el identificador PMID.
- Ser sobre el virus, la enfermedad o la pandemia del COVID-19 o SARS-CoV-2.

- Artículos publicados entre el 2020 y 2023.
- No ser publicaciones que fueran publicadas en el periodo de la pandemia del COVID-19
- Realiza un análisis cuantitativo explícitamente.
- No ser de factor de impacto, es decir, que evalúen a instituciones o autores.
- Variables

### 2.3.1. Metadatos bibliográficos

Los metadatos asociados a las publicaciones tienen un ciclo por el que se le va asociando los metadatos, primero la revista le pide al autor ingresar datos como título, resumen, palabras clave y autores, segundo la revista al publicar genera metadatos como el DOI y otros identificadores, tercero las bases de datos cuando indexan los documentos generan metadatos como identificadores, cuarto los usuarios pueden generar nuevos metadatos obteniéndolos desde el documento para poder realizar análisis más robustos, en la tabla 2.1 se muestra en la columna de origen de los metadatos la etapa en la cual se originó cada metadato extraído de los documentos.

Los metadatos son información que se evalúa en un análisis de la literatura, es por esto que se usan como variables Durante el procedimiento registré todas las variables factibles para ser analizadas (Tabla 2.1) y después seleccioné las variables que usé para las siguientes etapas. Para este análisis yo utilicé los metadatos que genera la revista, la base de datos y los que generé que se encuentran en la tabla 2.1 con \*

Variables
Autores
Adscripción de los autores
Título
Identificadores (DOI, PMID, PMCID)
Términos MeSH
Anotaciones SciLite
Acceso
Resumen
Fecha de publicación
Tipo de publicación
Revista
Palabras clave
Idioma
Citas
Fuente/Base de datos
Número de registros analizados
Objetivo del estudio
Tipo de análisis métrico
Fecha de inicio del estudio
Fecha de fin del estudio
Intervalo de tiempo del estudio
Consulta
Software
API's utilizadas
Tipo de estadísticas
Tipo de tecnologías usadas como IA, análisis de textos, bibliometría, análisis de redes, etc.
Otras bases de datos asociadas utilizadas durante el estudio
Herramientas digitales utilizadas

Tabla 2.1: *Tabla con las variables que se pueden evaluar de estas, solo seleccione las que se encuentran en color rosa.*



Las variables que serán usadas se eligieron para poder ver las leyes de la bibliometría y para hacer análisis de redes.

- Los autores son las personas que contribuyeron para la realización de la publicación, cada investigador tiene un área de especialidad.
- Las revistas son publicaciones periódicas que tienen un enfoque y alcance específico, en las que se publican los textos revisados por pares que cumplen con los requisitos de la revista.
- Las palabras clave de un texto son uno o más términos en los cuales se describe la investigación. Estas son metadatos básicos de cualquier texto y son generados por los autores al publicar en una revista.
- Los términos MeSH son términos estandarizados del tesauro MeSH, estos se utilizan para indexar, catalogar y buscar información biomédica y relacionada con la salud. Estos términos en PubMed se pueden ver y descargar como metadatos asociados a los textos (“Medical Subject Headings”, 2023).

## 2.4. Extracción de los datos

El ciclo de los metadatos

Para obtener los datos necesarios para el análisis de manera eficiente y rápida se extrajeron los metadatos mediante la API de PubMed,

## 2.5. Procesamiento de la información de la literatura

la

## 2.6. Análisis

Dos grandes formas en las cuales se pueden procesar metadatos biométricamente son desde aplicaciones en línea amigables, las cuales muchas veces están asociadas a bases de datos

y desde software en los que se necesita conocimiento en programación y resultan no ser amigables.

Ejemplos

Para realizar el análisis utilicé Bibliometrix una paquetería de R (Aria & Cuccurullo, 2017)

## 2.7. Visualización de datos

de

## 2.8. Integración y contextualización de la información

Durante la

Para el registro de los artículos y notas, utilice [hypothes.is](https://hypothes.is/). Una herramienta digital de gran utilidad que nos parece muy valiosa para la investigación digital.

Para mantenerme actualizada de nuevas publicaciones de los distintos temas en biociencias, genere los RSS

### 3 Resultados

De la consulta 0 obtuve 18,825 resultados, de los cuales podemos ver la cantidad de publicaciones que hay al año y se muestran en la Figura 3.1 en la cual se puede ver que a partir del 2020 aumentaron drásticamente las publicaciones sobre cienciometría, bibliometría, y los análisis de la literatura.

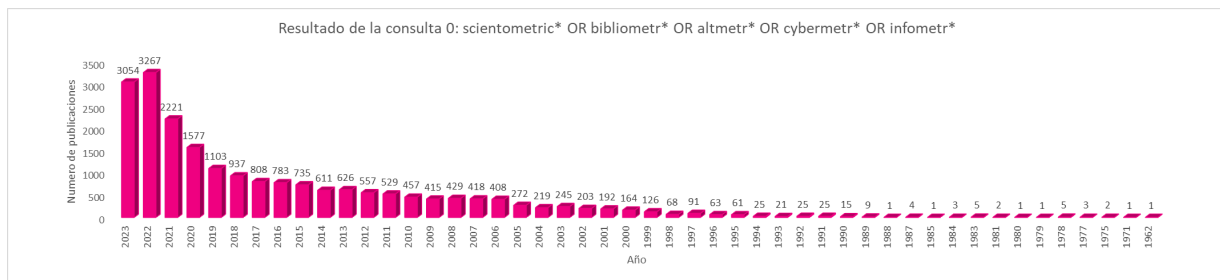


Figura 3.1: *Gráfica de la cantidad de publicaciones por año de la consulta 0*

De la consulta 1 obtuve 1,011 resultados de los cuales...

De la consulta 2 obtuve 1,784 resultados de los cuales...



## 4 Discusión

De esta manera se tienen los desdoblamientos hiperfinos, los cuales se muestran en la



# Bibliografía

- Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), 959-975. <https://doi.org/10.1016/j.joi.2017.08.007>
- Barabási, A.-L., & Pósfai, M. (2016). *Network science*. Cambridge university press. <http://networksciencebook.com/>
- Gartner, R. (2016). *Metadata*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-40893-4>
- Hernández, A., María del Carmen. (2020). *Introducción a la teoría de redes*. papirhos, IM-UNAM. <http://www.libros.unam.mx/introduccion-a-la-teoria-de-redes-9786073038980-libro.html>
- Hull, D., Pettifer, S. R., & Kell, D. B. (2008). Defrosting the Digital Library: Bibliographic Tools for the Next Generation Web (J. McEntyre, Ed.). *PLoS Computational Biology*, 4(10), e1000204. <https://doi.org/10.1371/journal.pcbi.1000204>
- Medical Subject Headings [Publisher: U.S. National Library of Medicine]. (2023, junio). Consultado el 30 de noviembre de 2023, desde <https://www.nlm.nih.gov/mesh/meshhome.html>
- Pomerantz, J. (2015). *Metadata*. The MIT Press.

