



# Unlocking open-source data for emerging tech: introducing the Emerging Technology Observatory

November 18, 2022  
Zach Arnold | Emily Weinstein



# Agenda

---

- About CSET and ETO
- Analyzing S&T emergence at scale: the Map of Science
  - How the Map works
  - Use cases
- Building a broader picture in key domains: the Supply Chain Explorer
  - How the Explorer works
  - Use cases
- Wrap-up/Q&A

# About CSET and ETO

- The [Center for Security and Emerging Technology](#) studies the security implications of emerging technologies, including AI, advanced computing, and biotechnology.
  - Nonpartisan and data-driven
  - “Compete” Line of Research: Analyze the state of technological innovation and competitiveness in the United States and their role in national power.
- CSET’s newly launched [Emerging Technology Observatory](#) builds data resources to inform critical decisions on emerging tech issues.
  - [eto.tech](#) launched in October with three tools: [Map of Science](#), [Supply Chain Explorer: Advanced Chips](#), [Country Activity Tracker](#)
    - Future tools will cover AI capabilities, Chinese tech ecosystem, open-source software, other supply chains, STEM talent flows, ...
  - Free to access; nonpartisan, nonprofit, 100% philanthropically funded

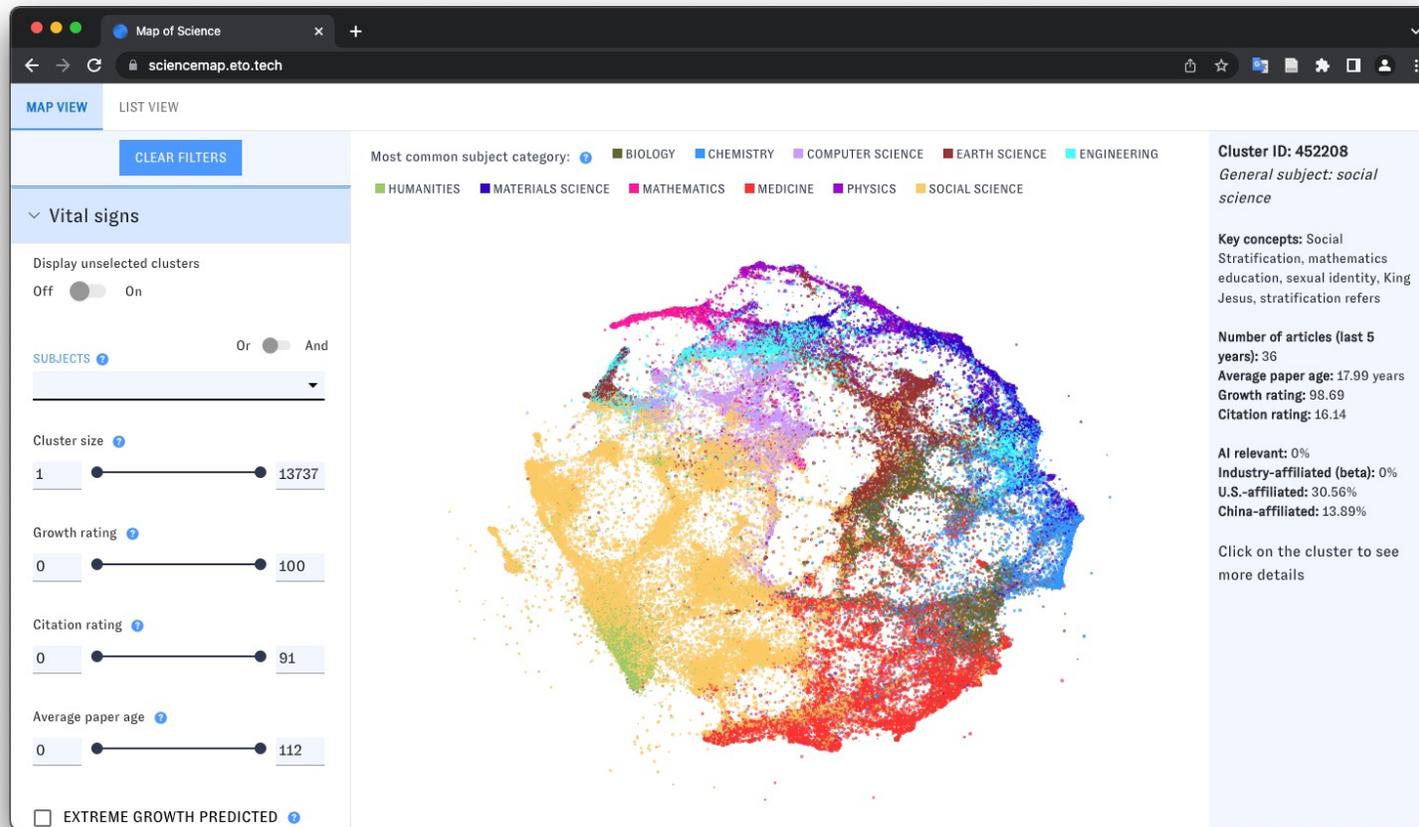
# Analyzing S&T emergence at scale: the Map of Science

- The [Map of Science](#) is ETO's tool for exploring global research across topics, sources, and languages
- Detailed metadata on **130 million academic articles**; citation-based **clustering** to provide a legible structure
- UI lets users quickly filter, browse, and drill down on areas of interest
- Core uses: detect and understand emerging topics; provide “entry points” for further analysis and action
  - Less suited for: finding predetermined “needles,” tracking nonpublic R&D, longitudinal analysis



2D visualization of the Map clusters ([fastest-growing 10% highlighted](#); clusters with more intercluster citation links are closer together)

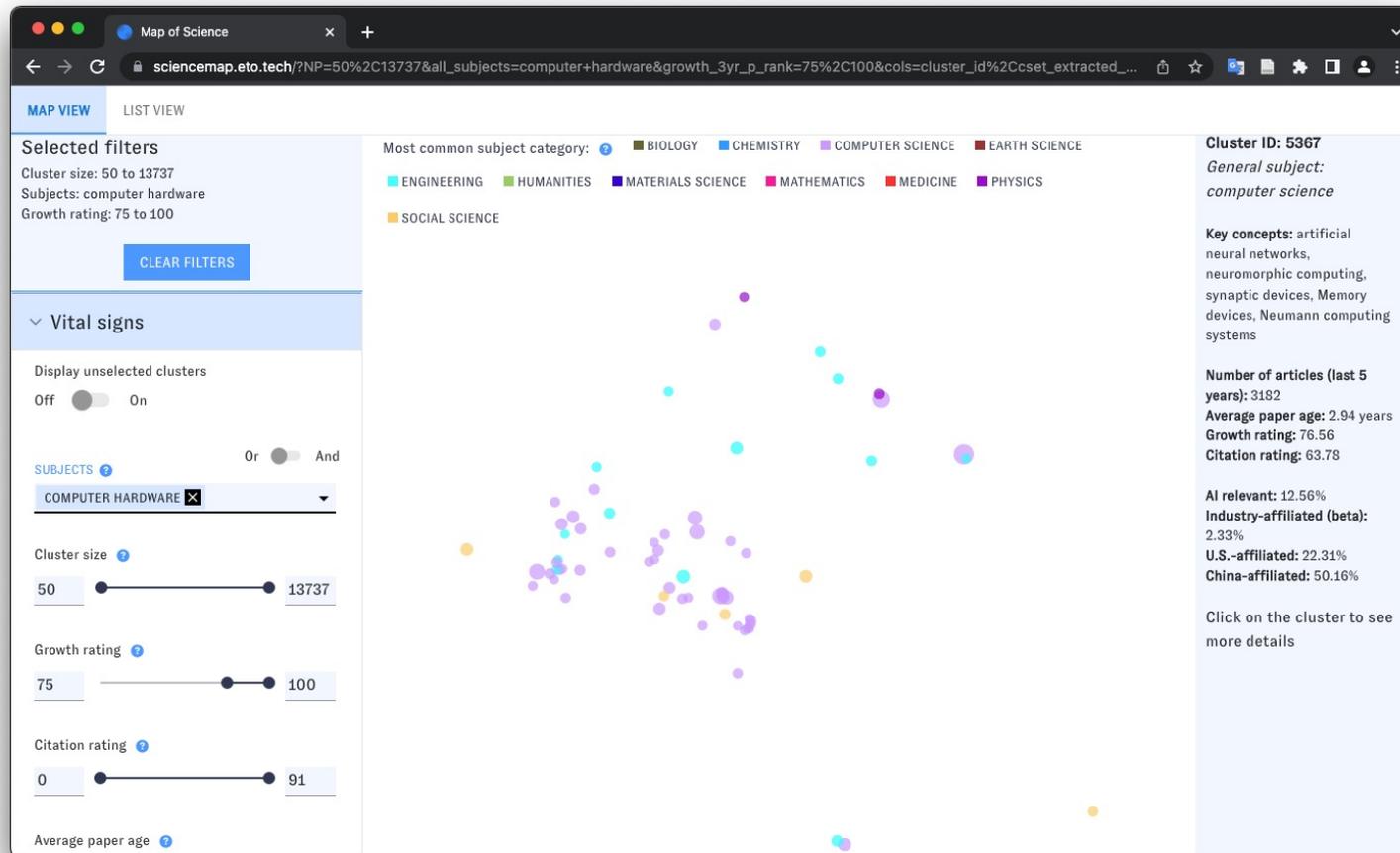
# The Map in practice



“What are the next big trends in computing hardware?”

For more information: [ETO documentation](#), [Rahkovsky et al. 2021](#), [Dunham et al. 2020](#), [Klavans et al. 2020](#)

# The Map in practice



- [Start with clusters:](#)
  - with a lot of computer hardware research (model-tagged)
  - With high growth
  - With a decent number of recent articles
- **From 110k+ clusters to ~70**

For more information: [ETO documentation](#), [Rahkovsky et al. 2021](#), [Dunham et al. 2020](#), [Klavans et al. 2020](#)

# The Map in practice

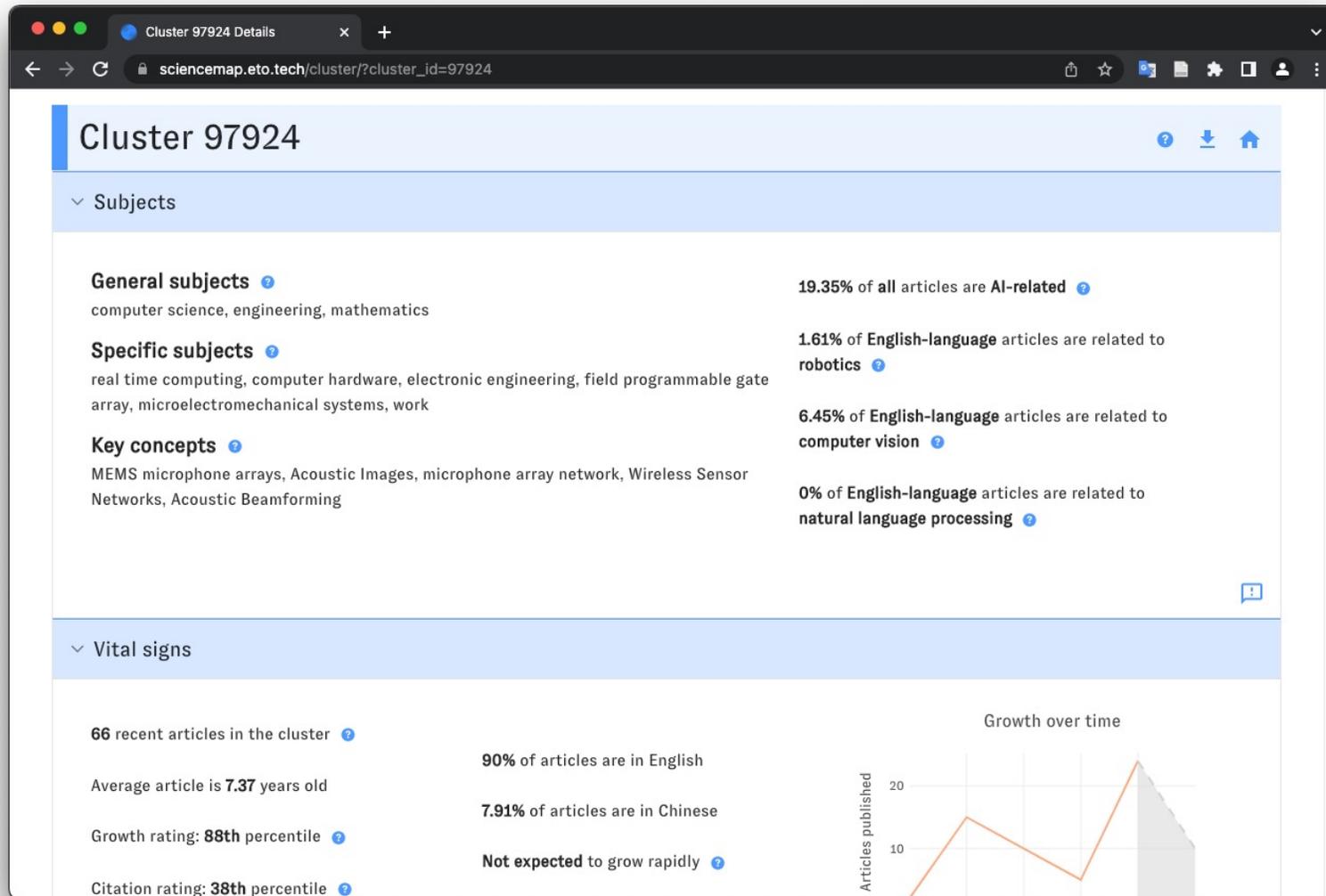
Top clusters		+ ADD/REMOVE COLUMNS		
Cluster ID	Most common subject category	CSET phrases	Cluster size ?	Growth rating ? ↓
<a href="#">109172</a>	computer science	Power Grid Operation, Environment Monitoring System, Monitoring System Based, power grid, Grid Operation Situation	84	97.80
<a href="#">109530</a>	computer science	Robot Operating System, mobile robot, Hector SLAM, SLAM algorithms, robot navigation	75	97.23
<a href="#">72278</a>	physics	semiconductor optical amplifier, optical frequency encoded, logic gates, optical NAND gates, Reflective Semiconductor Optical	75	93.82
<a href="#">104245</a>	computer science	Medical Things, Health Monitoring Systems, structural health monitoring, IoT, compressed ECG signals	61	93.61
<a href="#">116016</a>	social science	Health Monitoring System, health vitals, Patient Health Monitoring, Smart Health Care, IoT	61	93.39
<a href="#">91826</a>	computer science	Elevator Button Recognition, autonomous elevator button, mobile robots, button recognition system, Button Operation	57	92.30
<a href="#">100477</a>	computer science	Home Security System, Motion Detection System, Maintenance Monitoring System, Detection System Based, smart surveillance system	72	91.04

Use the Map's list view to browse and sort

Key concepts (model-derived) typically give a rough sense of “what the cluster is about”

For more information: [ETO documentation](#), [Rahkovsky et al. 2021](#), [Dunham et al. 2020](#), [Klavans et al. 2020](#)

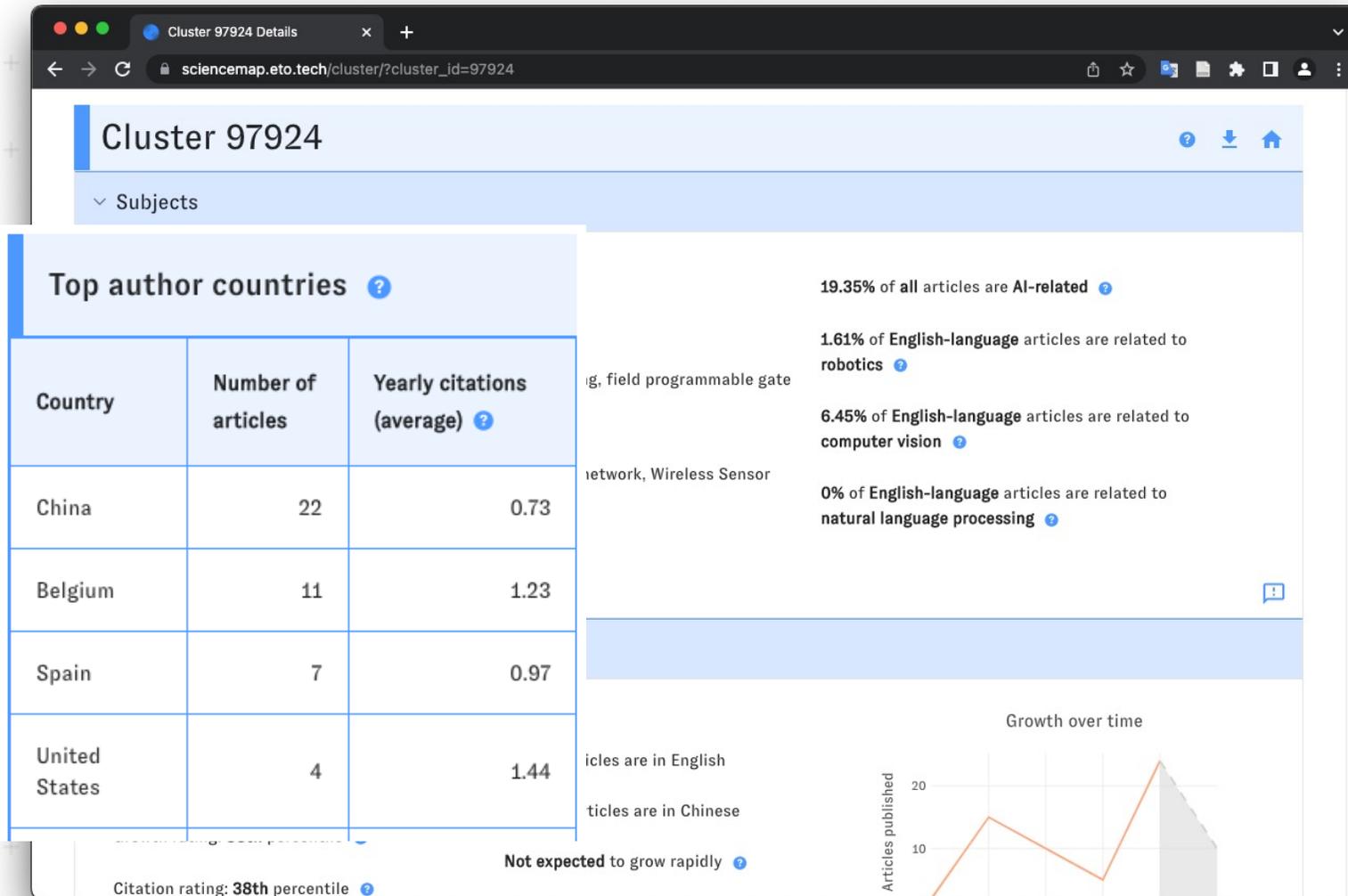
# The Map in practice



Switch to detail view to understand any particular cluster in depth

For more information: [ETO documentation](#), [Rahkovsky et al. 2021](#), [Dunham et al. 2020](#), [Klavans et al. 2020](#)

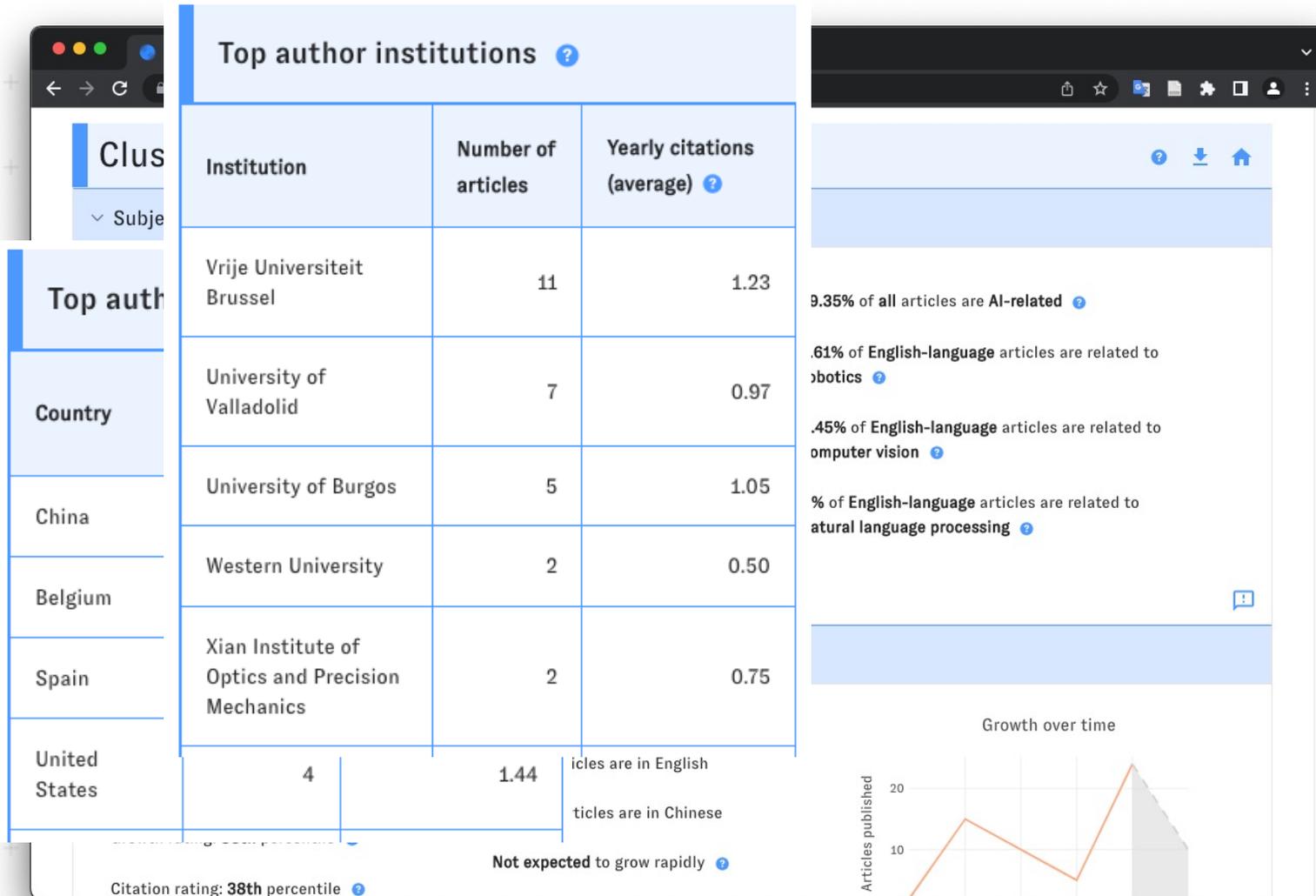
# The Map in practice



Switch to detail view to understand any particular cluster in depth

For more information: [ETO documentation](#), [Rahkovsky et al. 2021](#), [Dunham et al. 2020](#), [Klavans et al. 2020](#)

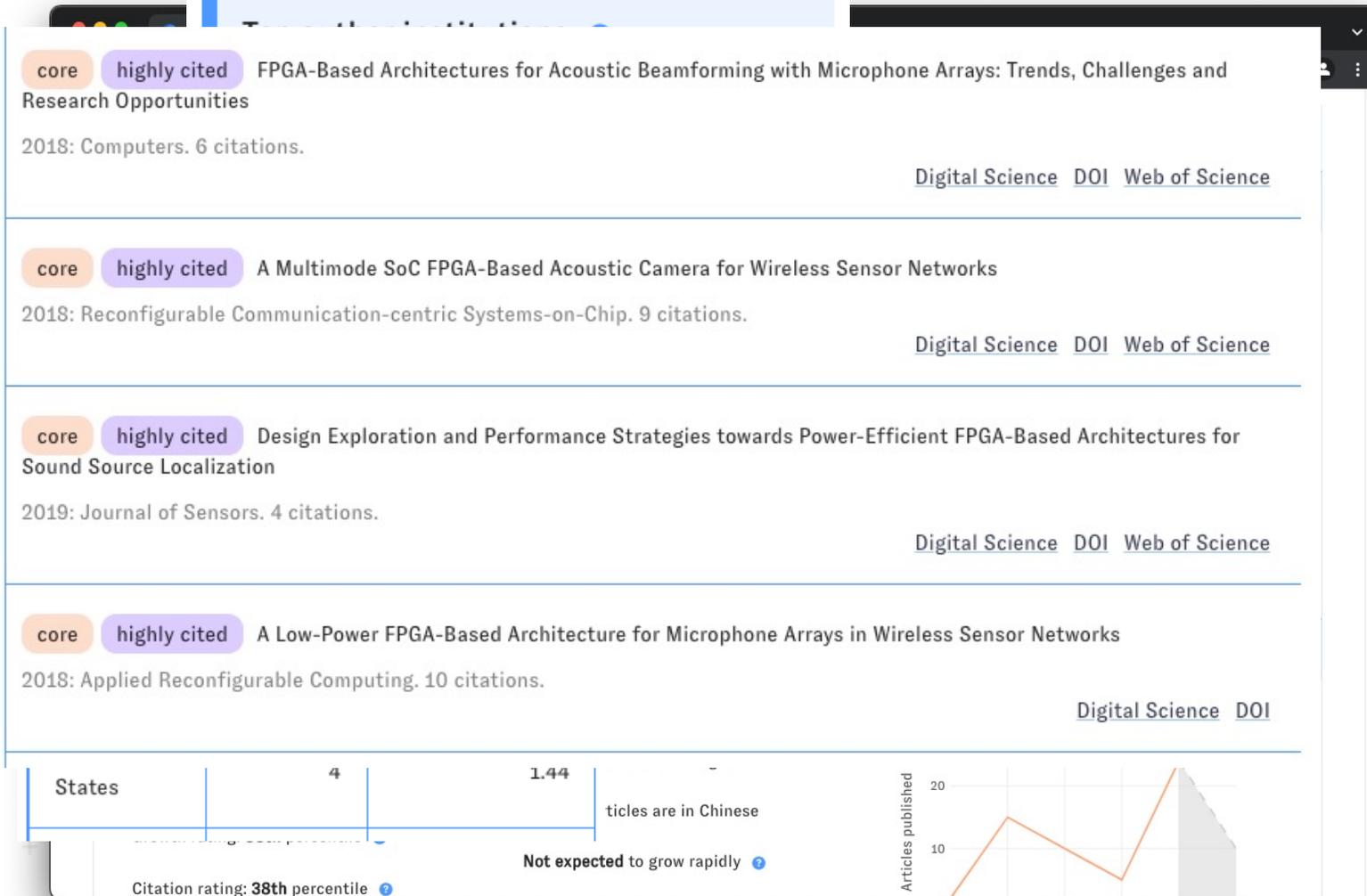
# The Map in practice



Switch to detail view to understand any particular cluster in depth

For more information: [ETO documentation](#), [Rahkovsky et al. 2021](#), [Dunham et al. 2020](#), [Klavans et al. 2020](#)

# The Map in practice



Switch to detail view to understand any particular cluster in depth

For more information: [ETO documentation](#), [Rahkovsky et al. 2021](#), [Dunham et al. 2020](#), [Klavans et al. 2020](#)

# The Map in practice

The screenshot displays a research map interface. On the left, a list of papers is shown, each with a 'core' and 'highly cited' badge. The papers are:

- FPGA-Based Architectures for Acoustic Beamforming with Microphone Arrays: Trends, Challenges and Research Opportunities (2018: Computers. 6 citations.)
- A Multimode SoC FPGA-Based Acoustic Cam (2018: Reconfigurable Communication-centric Systems-on-Chip. 9 citations.)
- Design Exploration and Performance Strategies for Sound Source Localization (2019: Journal of Sensors. 4 citations.)
- A Low-Power FPGA-Based Architecture for Microphone Arrays (2018: Applied Reconfigurable Computing. 10 citations.)

At the bottom left, a 'States' bar shows 4 states and 1.44 articles are in Chinese. A citation rating of 38th percentile is shown. A line graph at the bottom right shows 'Articles published' over time, with a peak around 2018.

On the right, a table titled 'Top patent titles' lists patent titles and their number of within-cluster citations:

Patent title	Number of within-cluster citations
Systems and methods for managing the processing of information acquired by sensors within an environment	1
Two-dimensional microphone array with improved directivity	1
Array microphone system and method of assembling array microphone system	1

Switch to detail view to understand any particular cluster in depth

For more information: [ETO documentation](#), [Rahkovsky et al. 2021](#), [Dunham et al. 2020](#), [Klavans et al. 2020](#)

# Using the Map of Science

- Understanding key S&T players and trends in China (and elsewhere)
- Comparing/contrasting key S&T focus areas between the U.S. and China
  - [Concentrations of AI-Related Topics in Research: Robotics](#)
    - Robotics-related RC in Engineering:
      - Japan dominates research for this RC, followed by China and the United States, respectively.
      - Over 11 percent of papers written in Chinese
- Tracking technologies that are “emerging” to inform policy
  - [Terrorism, AI, and Social Media Research Clusters](#)

# Building a broader picture: the Supply Chain Explorer

- ETO's [Supply Chain Explorer](#) is an interactive, high-level visualization of the supply chain for advanced computer chips
  - Builds on several years of CSET research and data acquisition
- Primarily for non-specialists: an accessible orientation and reference
- Core uses:
  - Visually explore the chip supply chain as a series of stages and processes, each involving different tools, materials, and providers
  - Assess countries' and companies' role in the supply chain
  - Identify “chokepoints” and other structural features

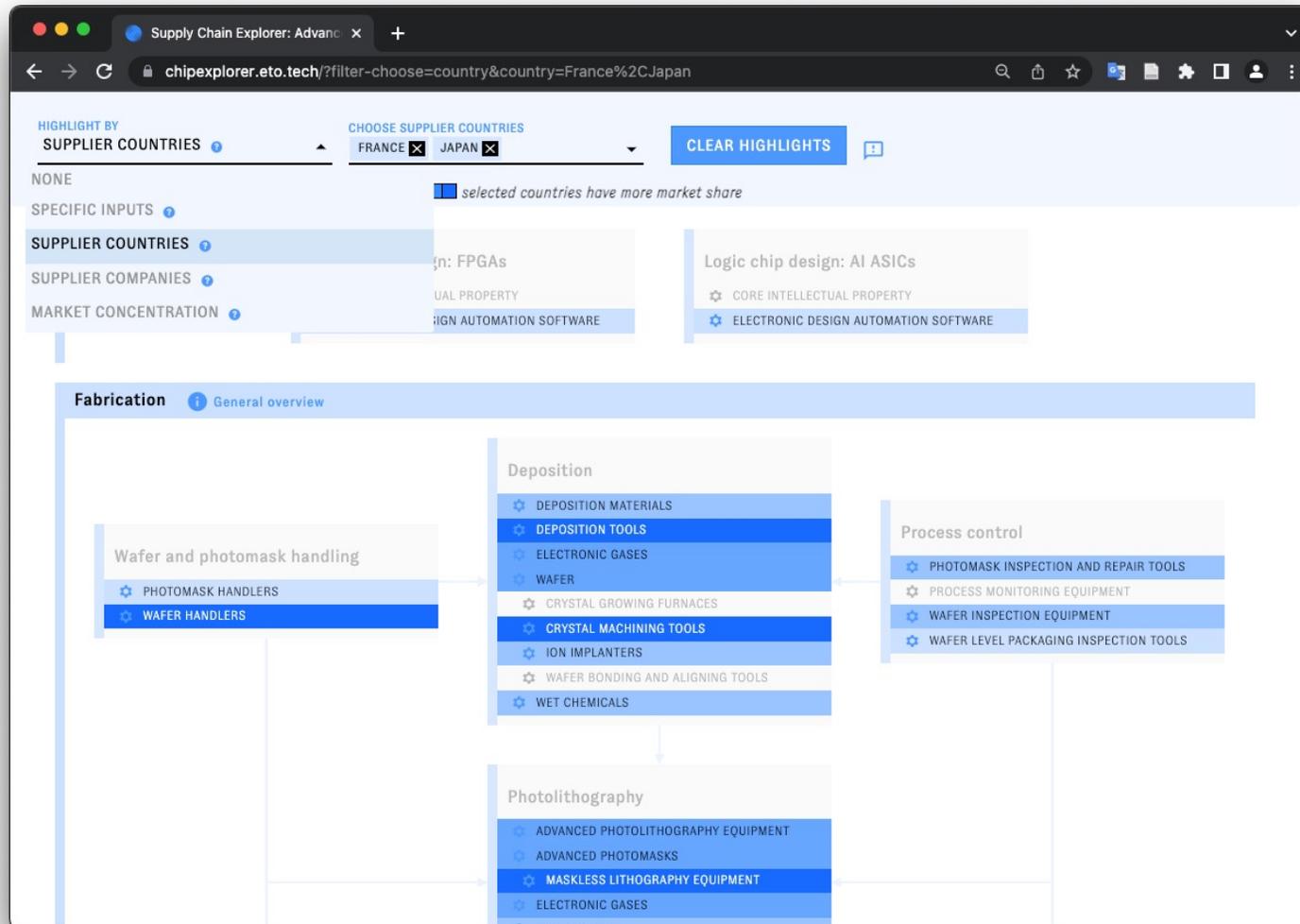
# How the Explorer works

The screenshot shows the 'Supply Chain Explorer' web application. The browser address bar displays 'chipexplorer.eto.tech/?parentNode=N35&selectedNode=N9'. The interface includes a 'HIGHLIGHT BY NONE' dropdown and a 'CLEAR HIGHLIGHTS' button. A central diagram shows three interconnected boxes: 'Wafer and photomask handling' (containing photomask and wafer handlers), 'Process control' (containing inspection and monitoring equipment), and a central list of tools including deposition materials, electronic gases, wafers, crystal growing furnaces, crystal machining tools, ion implanters, wafer bonding and aligning tools, and wet chemicals. Below this, a detailed view for 'Crystal machining tools' is shown, featuring a list of categories (Deposition, Crystal Machining Tools, etc.), a photograph of a technician in a cleanroom, and descriptive text about the process. The text states: 'Crystal growing furnaces and machining tools are necessary to produce all wafers-thin, disc-shaped materials used to produce chips. The furnace forms a cylindrical ingot of silicon from polycrystalline raw silicon; machining equipment then cuts the ingot into wafers used for chip fabrication. These tools have relatively low value and complexity relative to other semiconductor manufacturing equipment.' It also includes 'Global market size: \$110 million (2019)' and 'Supplier countries'.

Structures CSET's research on advanced logic chips into a streamlined, high-level dataset of processes, inputs, and providers

For more information: [ETO documentation](#), [Explorer dataset](#), [Khan et al. 2021](#), [Khan et al. 2021](#)

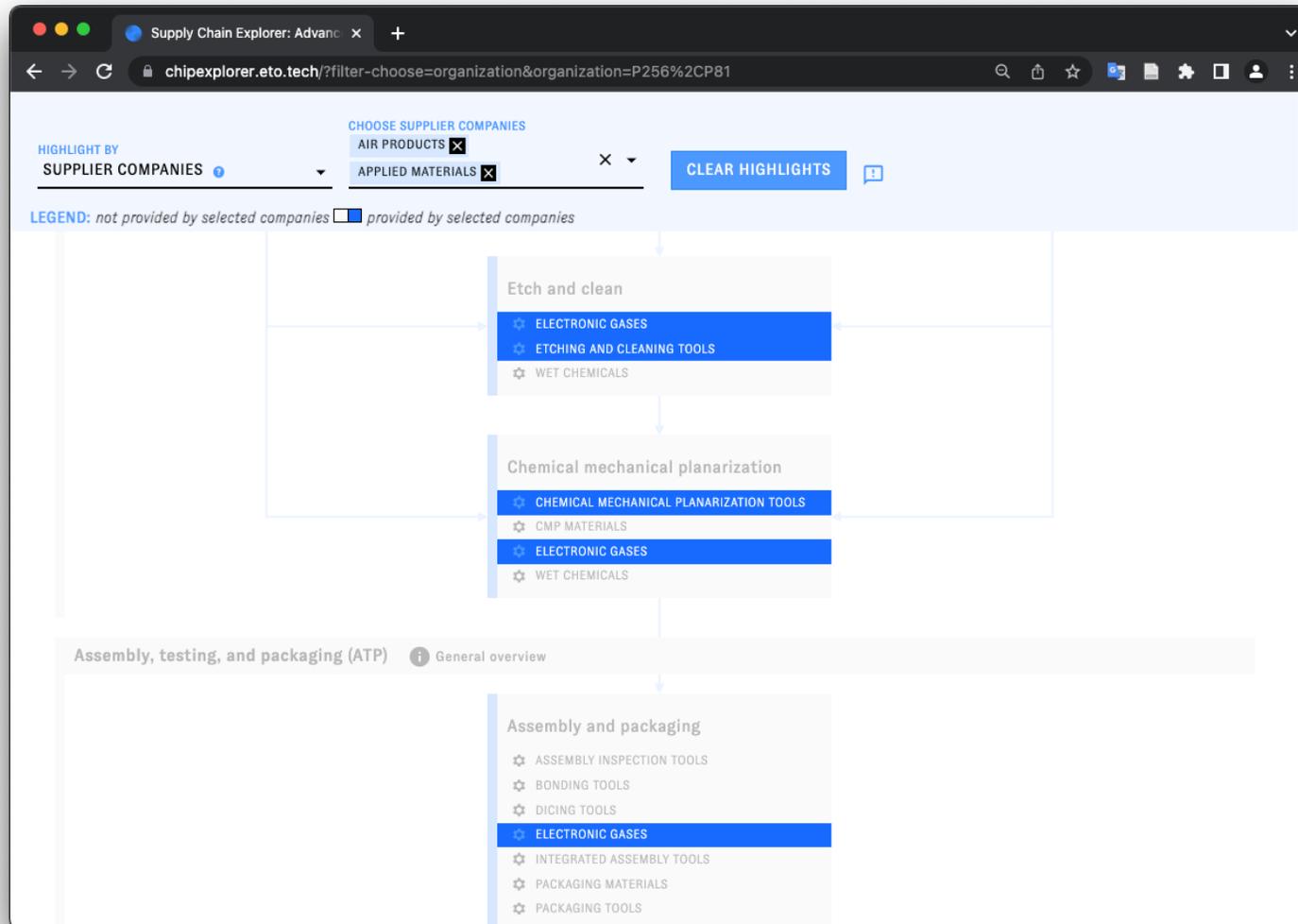
# How the Explorer works



Country, company, and chokepoint filters help users assess likely areas of sensitivity/tech transfer activity going forward

For more information: [ETO documentation](#), [Explorer dataset](#), [Khan et al. 2021](#), [Khan et al. 2021](#)

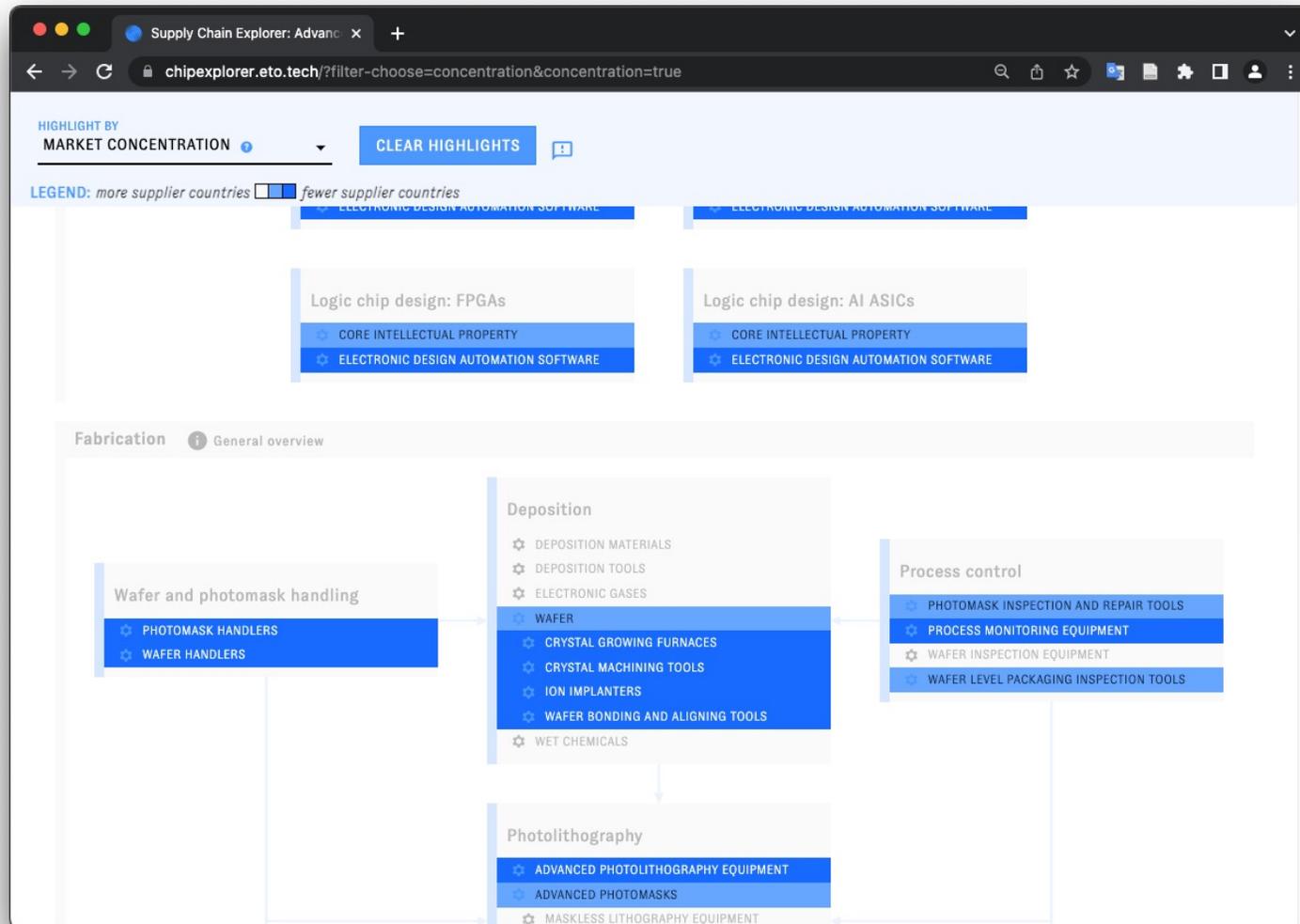
# How the Explorer works



Country, company, and chokepoint filters help users assess likely areas of sensitivity/tech transfer activity going forward

For more information: [ETO documentation](#), [Explorer dataset](#), [Khan et al. 2021](#), [Khan et al. 2021](#)

# How the Explorer works



Country, company, and chokepoint filters help users assess likely areas of sensitivity/tech transfer activity going forward

For more information: [ETO documentation](#), [Explorer dataset](#), [Khan et al. 2021](#), [Khan et al. 2021](#)

# Using the Explorer

- Example: Using this tool to make sense of U.S. export control policies
  - Attempting to map new controls onto the chip supply chain and pre-existing controls to inform Indicators and Warnings (early-stage) project
  - Building on previous CSET research:
    - Khan, “[U.S. Semiconductor Exports to China: Current Policies and Trends](#)” (2020)
    - Khan, “[Securing Semiconductor Supply Chains](#)” (2021)

# Keep in touch

---

- [Support is available](#) for ETO resources
- Requests and feedback are always welcome through ETO's [website](#) or [cset\\_eto@georgetown.edu](mailto:cset_eto@georgetown.edu)
- Follow [ETO's blog](#) for news, demos and insights
- Reach out to CSET's analysis team: [cset@georgetown.edu](mailto:cset@georgetown.edu)
- [Sign up](#) on CSET's website to receive our latest research, biweekly newsletter, and event invitations
  - To request briefings, contact Danny Hague ([danny.hague@georgetown.edu](mailto:danny.hague@georgetown.edu))



**Emily Weinstein**  
Research Fellow  
[emily.weinstein@georgetown.edu](mailto:emily.weinstein@georgetown.edu)



**Zach Arnold**  
Analytic Lead, ETO  
[zachary.arnold@georgetown.edu](mailto:zachary.arnold@georgetown.edu)

# Backup slides

# About CSET

## **Purpose**

- Study the security implications of emerging tech, including AI, advanced computing, and biotech
- Deliver nonpartisan, data-driven analyses to decisionmakers
- Prepare the next generation of policymakers, analysts, and diplomats to wrestle with future technology dilemmas

## **Structure**

- Over 50 dedicated staff examining emerging technology and security issues
- Dedicated analytic, data science, and translation teams

## **Funding**

- Supports independent research based on CSET-identified priorities
- >\$100M from philanthropic grants
- Not accepting government or foreign funding

# CSET lines of research

- **Applications:** Applications of emerging tech with emphasis on national security-relevant missions
- **Assessment (new):** AI/ML standards, testing, safety; and accidents, harm, and vulnerabilities
- **CyberAI:** AI/ML support to cyber ops; failure modes of AI/ML; competition in cyber + AI/ML
- **Workforce:** Domestic and foreign AI/ML talent pipelines and trends; workforce development, education and management
- **Compete:** Global tech competition; R&D and innovation ecosystems; tech alliances and diplomacy; research security
- **Supply Chains:** Analyze supply chains and chokepoints to maintain U.S. and allied tech leadership
- **Peer Watch:** Analyze emerging tech development of strategic competitors; indicators and warnings
- **Regions:** Country and regional emerging tech capabilities; regional alliances and diplomacy
- **Bio-Risk (new):** Risky biotech research discovery; ethical asymmetries; workforce and infrastructure



# Compete LOR

Analyze the state of technological innovation and competitiveness in the United States and their role in national power.

## Examines

- Investments and incentives to strengthen the innovation ecosystem
- Export controls and sanctions policy
- Trade rules and antitrust regulation

**Line of Research Lead:** Emily Weinstein, [esw54@georgetown.edu](mailto:esw54@georgetown.edu)



# Motivation for the ETO

- Useful, accessible ET data is hard to find, harder to develop
  - Open-source data is essential for understanding the contemporary emerging tech landscape - but making it useful takes sustained investment and specialized resources.
- CSET has unusual capabilities that position us to fill this gap
  - Identifying, acquiring, enriching, integrating, validating, documenting, distributing, and maintaining data
- We can build shared infrastructure that empowers others and prevents duplication of effort
- We welcome requests and feedback as we expand ETO's (free!) platform and toolkit in the coming months

# ETO works in progress

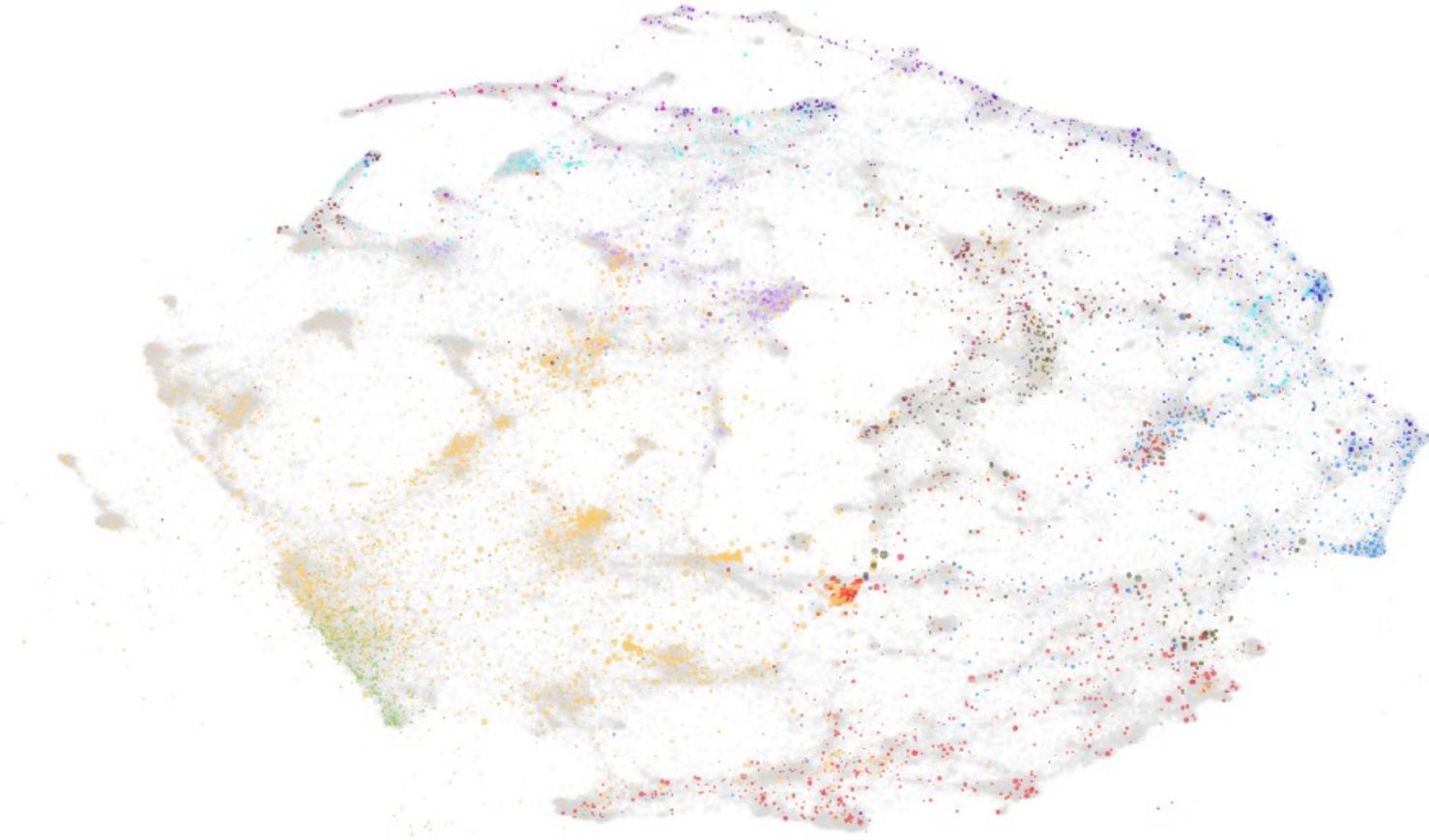
---

- Map of Science/S&T landscape analysis: topic-specific dashboards, tech maturity metrics
- Supply Chain Explorer: new topics (energy storage, rare earths, pharma, ?), semiconductor updates
- Open-source software ecosystem analysis
- Mapping research institutions
- Tools for China tech ecosystem research
- Tracking global STEM talent and leading-edge AI systems

# The thinking behind the Map of Science

- The global research literature is accessible in theory, but it's hard to draw insight in practice
  - Overwhelming scale
  - Sources scattered or locked away
  - Usual means of parsing (keywords, top sources, SMEs, etc.) are fragile and costly
- Help users identify emerging and critical areas in S&T by:
  - Assembling an unparalleled unified corpus
  - Using structural features of the corpus (citation links) to identify areas of interest
  - Using other corpus metadata to interpret results and inform action
- Aspiring to accessible, scalable, replicable insight
  - Complementing other more subjective/domain-specific processes
  - Pinpoint fast-growing subfields, understand context around active topics, ...

# How the Map works



2D visualization of the Map clusters ([fastest-growing 10% highlighted](#); clusters with more intercluster citation links are closer together)

- + Group 130m articles into ~110k clusters based *solely* on citation patterns
- +
  - o Maximizing modularity of citation
  - o Correlates (broadly) with intuitively interesting characteristics - topic, language, etc.

For more information: [ETO documentation](#), [Rahkovsky et al. 2021](#)

# How the Map works

Top clusters <span style="float: right;">+ ADD/REMOVE COLUMNS</span>					
Cluster ID	Most common subject category	CSET phrases	Cluster size <span>+</span>	Growth rating <span>+</span> ↓	Patent impact rating <span>+</span>
<a href="#">109172</a>	computer science	Power Grid Operation, Environment Monitoring System, Monitoring System Based, power grid, Grid Operation Situation	84	97.80	0.00
<a href="#">109530</a>	computer science	Robot Operating System, mobile robot, Hector SLAM, SLAM algorithms, robot navigation	75	97.23	50.16
<a href="#">72278</a>	physics	semiconductor optical amplifier, optical frequency encoded, logic gates, optical NAND gates, Reflective Semiconductor Optical	75	93.82	57.42
<a href="#">104245</a>	computer science	Medical Things, Health Monitoring Systems, structural health monitoring, IoT, compressed ECG signals	61	93.61	0.00
<a href="#">116016</a>	social science	Health Monitoring System, health vitals, Patient Health Monitoring. Smart Health Care. IoT	61	93.39	50.16

Map “list view” showing key concepts and user-selected metadata fields for a set of fast-growing clusters

Create cluster-level metadata based on constituent articles

- Simple aggregation of metadata from underlying datasets
- Model-based characterization - topic, language, key concepts, projected growth
- Mapping to CSET patent data (Dimensions/1790)

For more information: [ETO documentation](#), [Dunham et al. 2020](#), [Klavans et al. 2020](#), [Shen et al. 2018](#)

# How the Map works

▼ Vital signs

Or  And

SUBJECTS ⓘ

GENERAL: MATERIALS SCIENCE ✕

Cluster size ⓘ

100  13737

Growth rating ⓘ

0  100

Citation rating ⓘ

50  91

Average paper age ⓘ

0  5

EXTREME GROWTH PREDICTED ⓘ

> Countries and languages

▼ Vital signs

664 recent articles in the cluster ⓘ

Average article is 3.96 years old

Growth rating: 87th percentile ⓘ

Citation rating: 63rd percentile ⓘ

81.6% of articles are in English

12.96% of articles are in Chinese

Expected to grow rapidly ⓘ

Growth over time

Year	Articles published
2017	~10
2018	~60
2019	~90
2020	~130
2021	~190
2022	~170

Data is incomplete after 2021.

> Countries and languages

> Articles and sources

Web-based UI lets users quickly filter, browse, and drill down on clusters of interest

The Map's forte: starting from criteria of emergence, or other similarly broad concepts, and quickly gathering relevant areas of research for further exploration

For more information: [ETO documentation](#)

# The Map in practice

Top clusters <span style="float: right;">➔ + ADD/REMOVE COLUMNS</span>					
Cluster ID	Most common subject category	CSET phrases	Cluster size ?	Growth rating ? ↓	Patent impact rating ? ←
<a href="#">109172</a>	computer science	Power Grid Operation, Environment Monitoring System, Monitoring System Based, power grid, Grid Operation Situation	84	97.80	0.00
<a href="#">109530</a>	computer science	Robot Operating System, mobile robot, Hector SLAM, SLAM algorithms, robot navigation	75	97.23	50.16
<a href="#">72278</a>	physics	semiconductor optical amplifier, optical frequency encoded, logic gates, optical NAND gates, Reflective Semiconductor Optical	75	93.82	57.42
<a href="#">104245</a>	computer science	Medical Things, Health Monitoring Systems, structural health monitoring, IoT, compressed ECG signals	61	93.61	0.00
<a href="#">116016</a>	social science	Health Monitoring System, health vitals, Patient Health Monitoring, Smart Health Care, IoT	61	93.39	50.16
<a href="#">91826</a>	computer science	Elevator Button Recognition, autonomous elevator button, mobile robots, button recognition system, Button Operation	57	92.30	54.37

Add and sort by additional metadata fields to refine the inquiry

- Which of these clusters are helping generate patents?

# Map of Science methodology

Merged scientific research sources:

- Clarivate Web of Science
- Digital Science Dimensions
- Microsoft Academic Graph
- EastView Chinese National Knowledge Infrastructure
- arXiv
- Papers with Code
  
- Create merged corpus - paper disambiguation: string match using paper-level metadata fields (normalized title, normalized abstract, publication year, normalized surnames of authors, DOI, and citations)
- Create citation graph from 1.4 billion direct citation links by maximizing modularity and targeting hundreds of articles per cluster
- Clean up clusters
  - Free articles in clusters with <50 articles, reassign articles to remaining clusters
- Organize the clusters spatially based on linkages (rates of inter-cluster citations)
- Use aggregated paper metadata to characterize clusters
  - subject area, growth, author country affiliations, AI-relevance
- Develop interactive tool for users to explore the clusters and cluster-level information

For more information: [ETO documentation](#), [Dunham et al. 2020](#), [Klavans et al. 2020](#), [Shen et al. 2018](#)

# Extreme growth forecasting in the Map

- Current method (threat score:  $\sim 0.2$ ) uses four indicators:
  - Vitality:  $1/(\text{avg reference age in forecast year})$
  - Stage:  $1/(\text{years since peak year})$
  - Paper Age:  $1/(\text{avg paper age})$
  - Top250: Number of articles in top 250 journals in forecast year
- Experimentation with improved methods ongoing; indicators under evaluation include:
  - Historical growth rate: function of paper, reference, and citation counts over last 10 years
  - Visiting prolific authors: percentage of papers in cluster with authors that are prolific, early career, not in a large lab, not an expert in the research cluster, and increasing their activity

For more information: [ETO documentation](#), [Rahkovsky et al. 2021](#), [Klavans et al. 2020](#)

# Building a broader picture: the Supply Chain Explorer

- ETO's [Supply Chain Explorer](#) is designed to quickly orient non-experts to the essential inputs, players, and relationships involved in producing critical and emerging technologies
  - Complementing more granular, open-ended analytic tools like the Map of Science
- The first version of the Explorer focuses on **advanced computer chips** (<14 nm logic)
  - Builds on several years of CSET research and data acquisition
- High-level, flexible visualization framework designed to be applied to other technologies in the future

# The thinking behind the Explorer

- Chips have rapidly come into focus → many people with little or no background knowledge are now active in the field, often charged with major decisions
- Provide an accessible, rigorous, reasonably comprehensive orientation for newcomers and a handy reference/sharing resource for specialists
- Give government users a *trustable* resource - built by an organization without a financial interest in related policymaking/implementation
- Create a structure and visualization framework applicable to other technologies in the future

# How the Explorer works

The screenshot shows the 'Supply Chain Explorer' web application. The browser address bar displays 'chipexplorer.eto.tech/?parentNode=N35&selectedNode=N9'. The interface features a top navigation bar with 'HIGHLIGHT BY NONE' and a 'CLEAR HIGHLIGHTS' button. The main content area is divided into several sections:

- Wafer and photomask handling:** Includes 'PHOTOMASK HANDLERS' and 'WAFER HANDLERS'.
- Process control:** Includes 'PHOTOMASK INSPECTION AND REPAIR TOOLS', 'PROCESS MONITORING EQUIPMENT', 'WAFER INSPECTION EQUIPMENT', and 'WAFER LEVEL PACKAGING INSPECTION TOOLS'.
- Central list of tools and materials:** Includes 'DEPOSITION MATERIALS', 'DEPOSITION TOOLS', 'ELECTRONIC GASES', 'WAFER', 'CRYSTAL GROWING FURNACES', 'CRYSTAL MACHINING TOOLS', 'ION IMPLANTERS', 'WAFER BONDING AND ALIGNING TOOLS', and 'WET CHEMICALS'.
- Left sidebar:** A tree view showing a hierarchy of categories, with 'CRYSTAL MACHINING TOOLS' selected.
- Bottom section:** A detailed view for 'Crystal machining tools' featuring an image of a technician in a cleanroom, a description of the process, and market data.

**Crystal machining tools**

Crystal growing furnaces and machining tools are necessary to produce all wafers-thin, disc-shaped materials used to produce chips. The furnace forms a cylindrical ingot of silicon from polycrystalline raw silicon; machining equipment then cuts the ingot into wafers used for chip fabrication. These tools have relatively low value and complexity relative to other semiconductor manufacturing equipment.

**Global market size:** \$110 million (2019)

**Supplier countries:** [Link]

Gives users an accessible birds' eye view of the market and global context for areas of particular interest

For more information: [ETO documentation](#), [Explorer dataset](#), [Khan et al. 2021](#), [Khan et al. 2021](#)

# Our resources: Country [AI] Activity Tracker

- Allows country-level comparison of AI research, patenting, investment activity
- Includes metrics for cross-country collaboration and exchange
- Worldwide scope; build your own cohort (or choose from ours)

Showing **DATASET** RESEARCH metrics for **COUNTRIES** QUAD filtered by **FIELD OF STUDY** ALL AI FIELDS

AGGREGATE COUNTRY GROUPS **CLEAR**

India	150,929 #3	29,098 #10	19.28 #211	1,417,741 #9	67.44 #77	539.95 #48
Japan	113,917 #6	32,494 #9	28.52 #207	1,109,540 #11	26.35 #126	58.91 #137
Australia	76,103 #8	52,213 #5	68.61 #113	2,144,174 #6	50.69 #93	166.4 #96

Country co-authorship

COMPARISON FILTER  
TOP 10 COUNTRIES

Each cell lists the number of AI articles (since 2010) with at least one co-author from the country listed in **bold** and at least one co-author from the country listed in the cell. Author countries are inferred from where their institutions are located. Countries listed in **bold** are ordered by the sum of articles co-authored with the countries listed in the cell.

> <b>United States</b>	<b>China (mainland)</b> 62,609	United Kingdom 22,736	Germany 20,024	Australia 19,519	Canada 15,898	Italy 13,467	France 12,086	India 10,371	Japan 9,764	South Korea 8,436
> <b>Australia</b>	<b>United States</b> 19,519	<b>China (mainland)</b> 15,304	United Kingdom 11,812	Mexico 7,454	Germany 3,329	Canada 2,339	France 2,259	Italy 2,000	Singapore 1,926	Switzerland 1,899
> <b>Japan</b>	<b>United States</b> 10,371	<b>China (mainland)</b> 9,764	United Kingdom 8,436	Germany 8,436	France 8,436	Australia 8,436	Switzerland 8,436	Canada 8,436	Taiwan 8,436	Italy 8,436

[cat.eto.tech](https://cat.eto.tech)