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

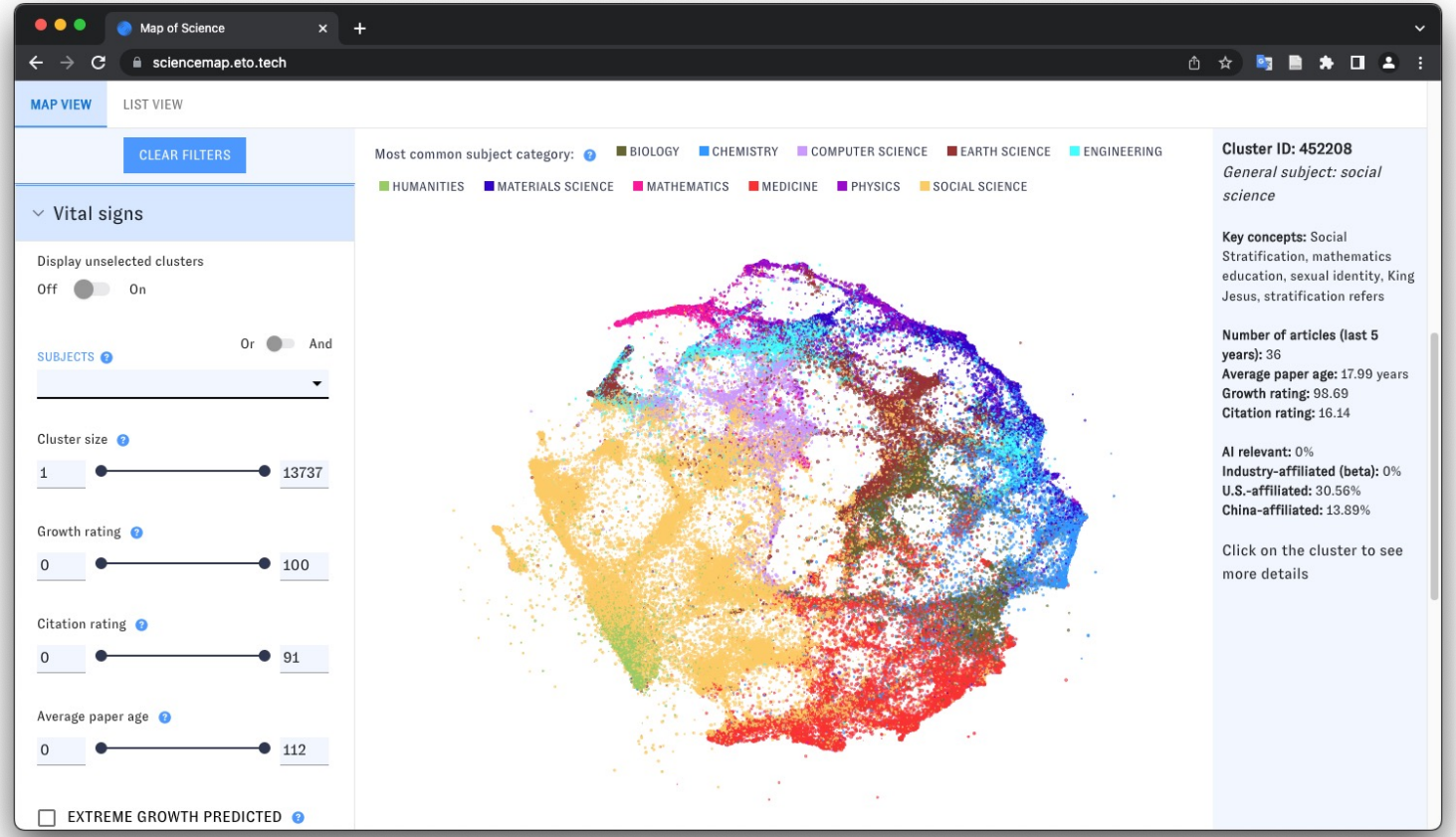
Analyzing S&T emergence at scale: the Map of Science

- 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

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

- 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

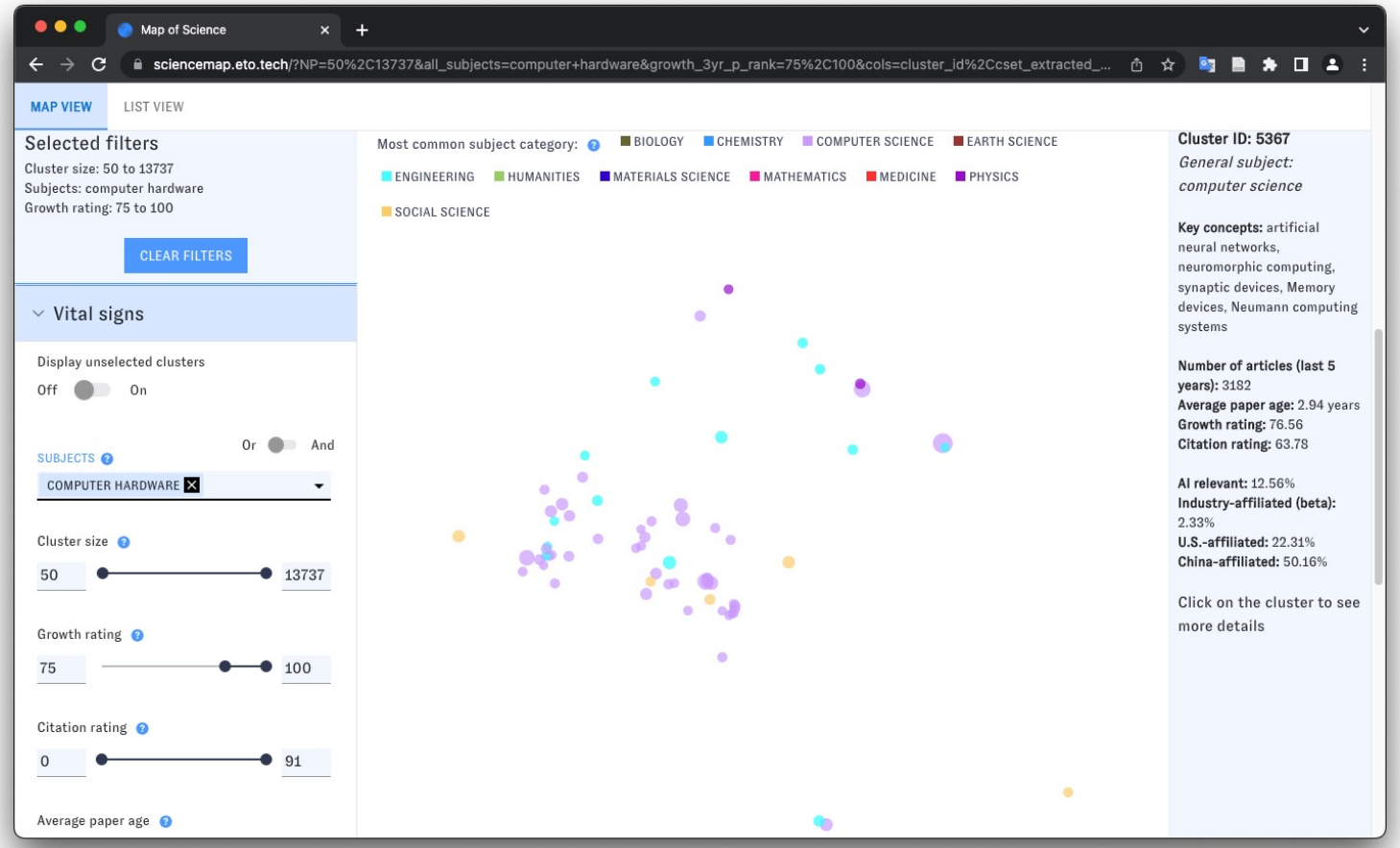
The Map in practice



"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

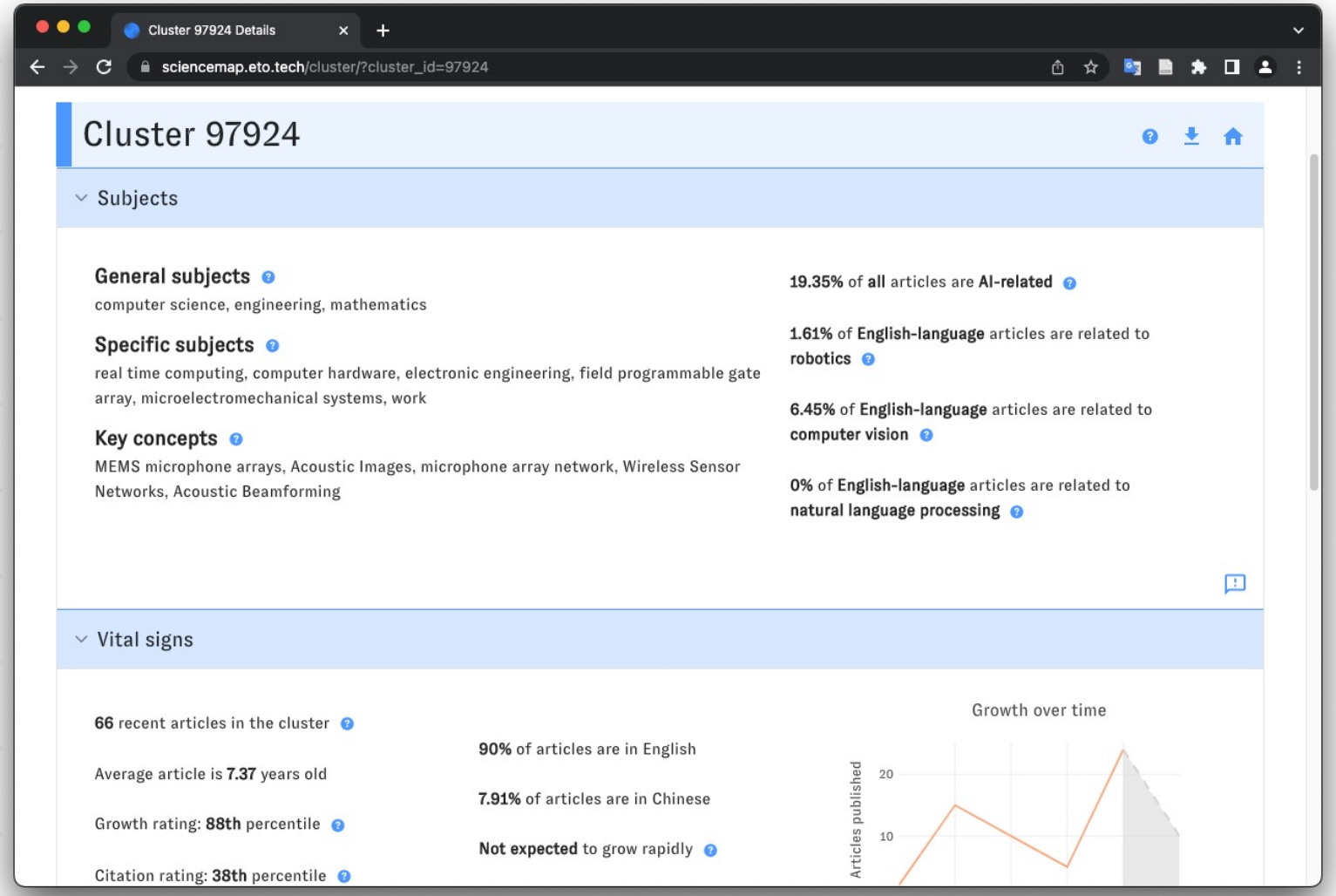
Cluster ID	Most common subject category	CSET phrases	Cluster size	Growth rating
109172	computer science	Power Grid Operation, Environment Monitoring System, Monitoring System Based, power grid, Grid Operation Situation	84	97.80
109530	computer science	Robot Operating System, mobile robot, Hector SLAM, SLAM algorithms, robot navigation	75	97.23
72278	physics	semiconductor optical amplifier, optical frequency encoded, logic gates, optical NAND gates, Reflective Semiconductor Optical	75	93.82
104245	computer science	Medical Things, Health Monitoring Systems, structural health monitoring, IoT, compressed ECG signals	61	93.61
116016	social science	Health Monitoring System, health vitals, Patient Health Monitoring, Smart Health Care, IoT	61	93.39
91826	computer science	Elevator Button Recognition, autonomous elevator button, mobile robots, button recognition system, Button Operation	57	92.30
100477	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

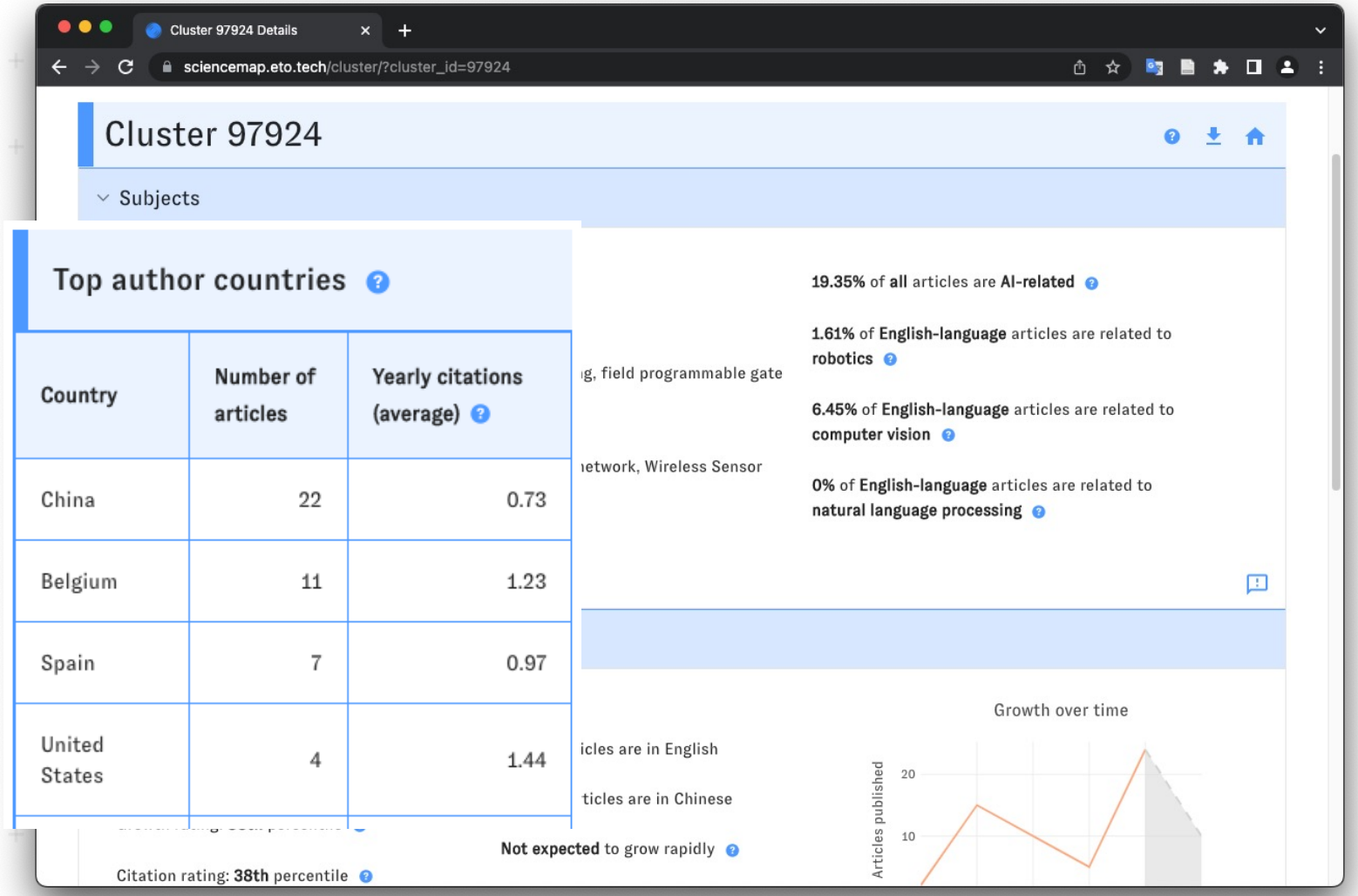
The Map in practice



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]()

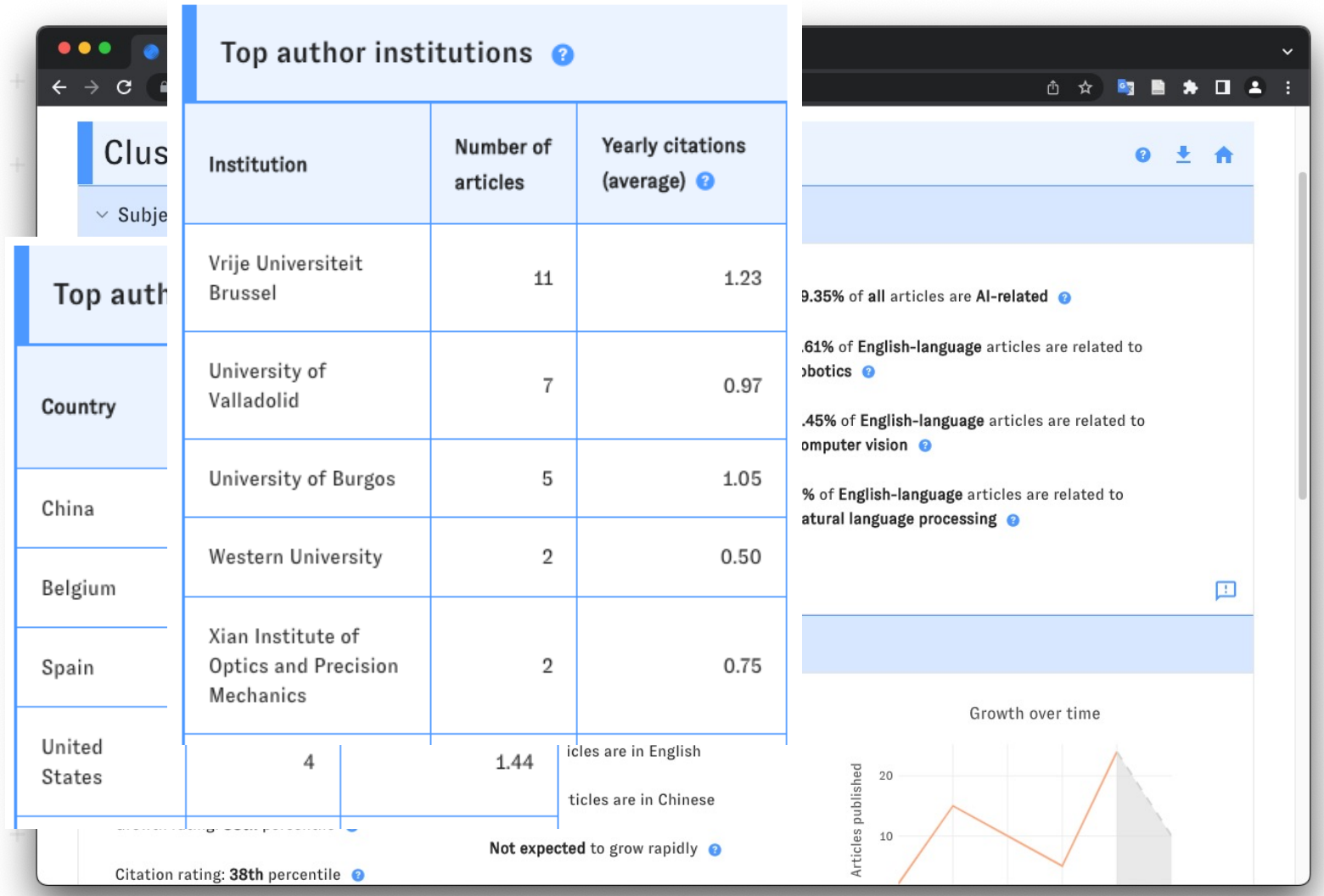
The Map in practice



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

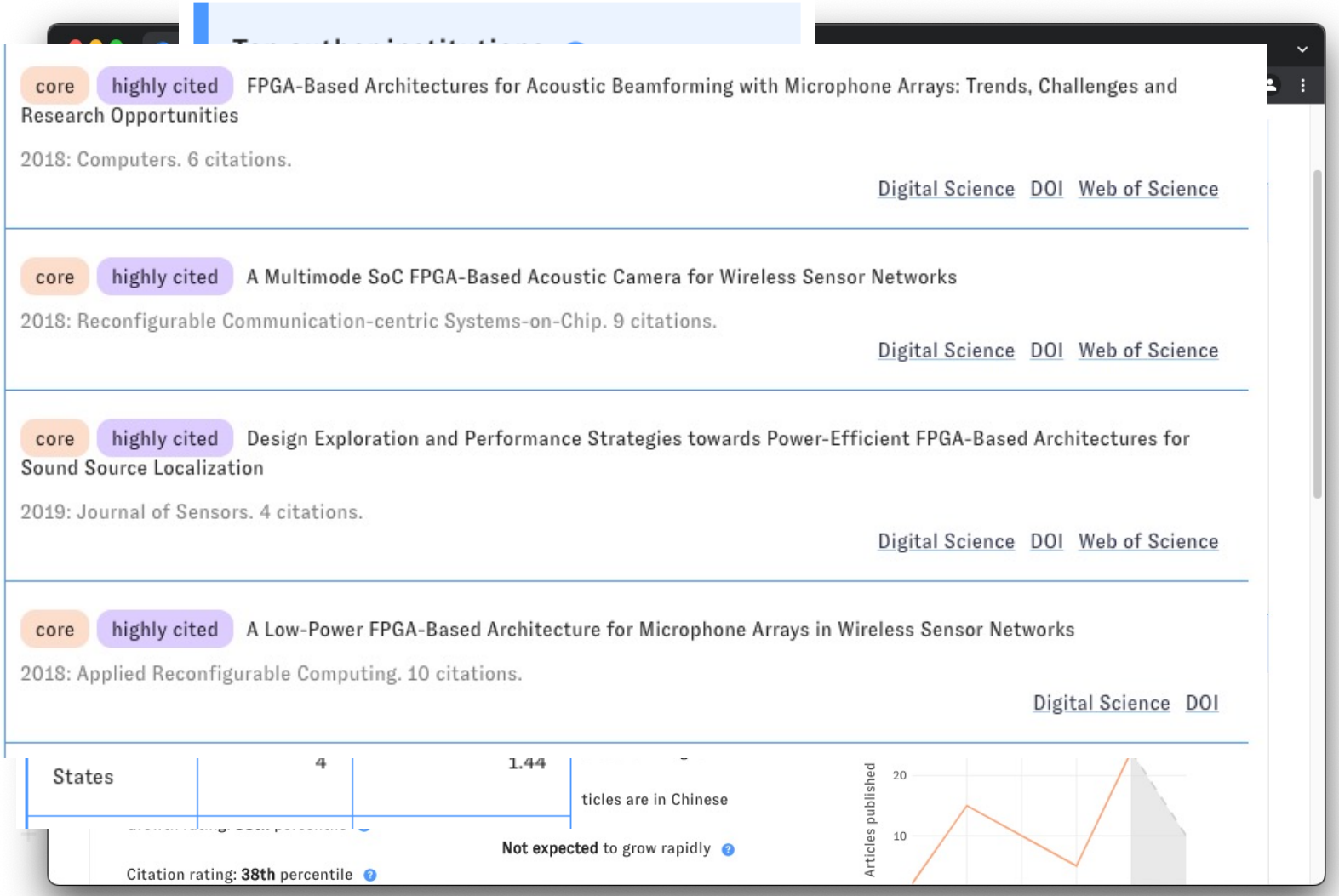
The Map in practice



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]()

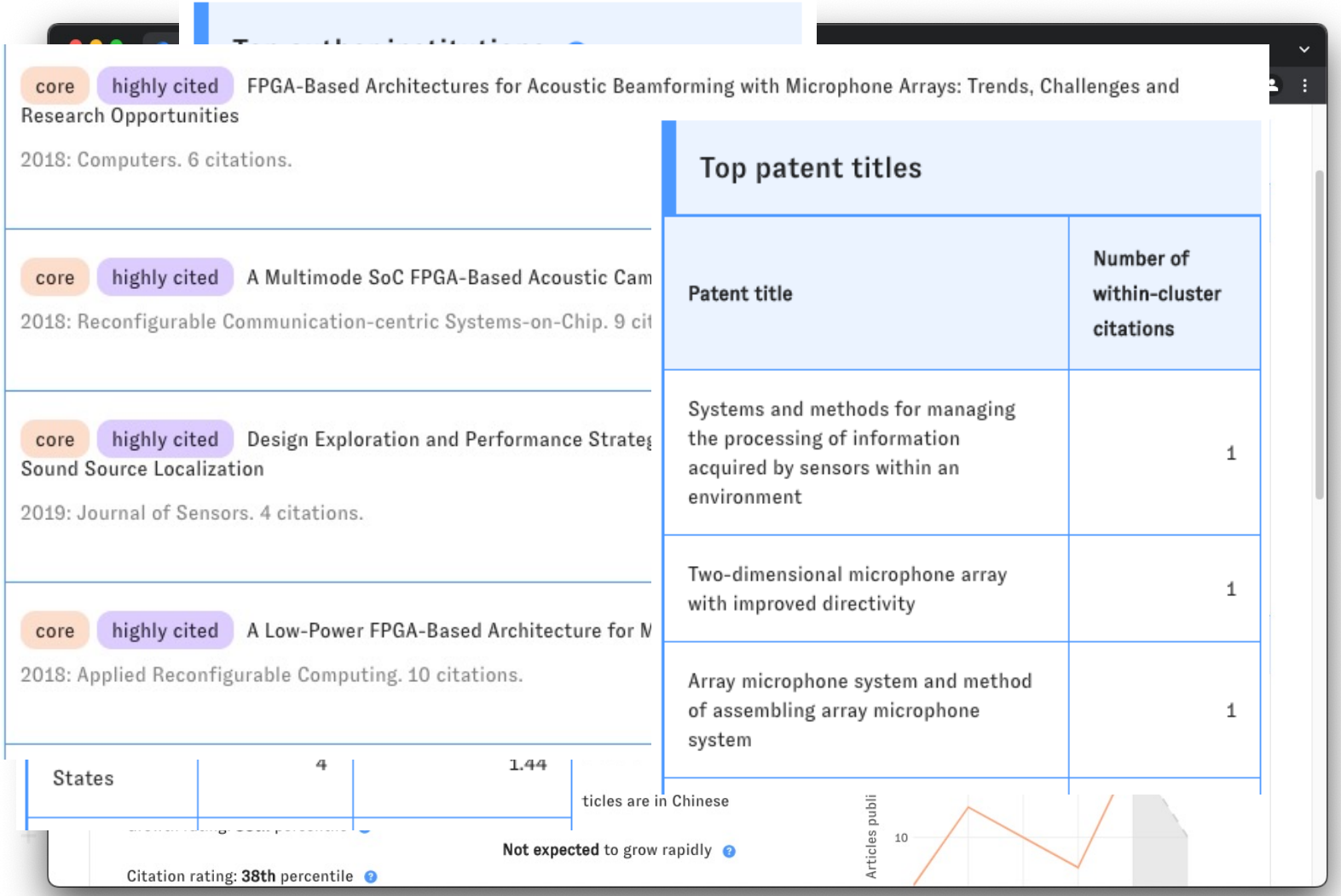
The Map in practice



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



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

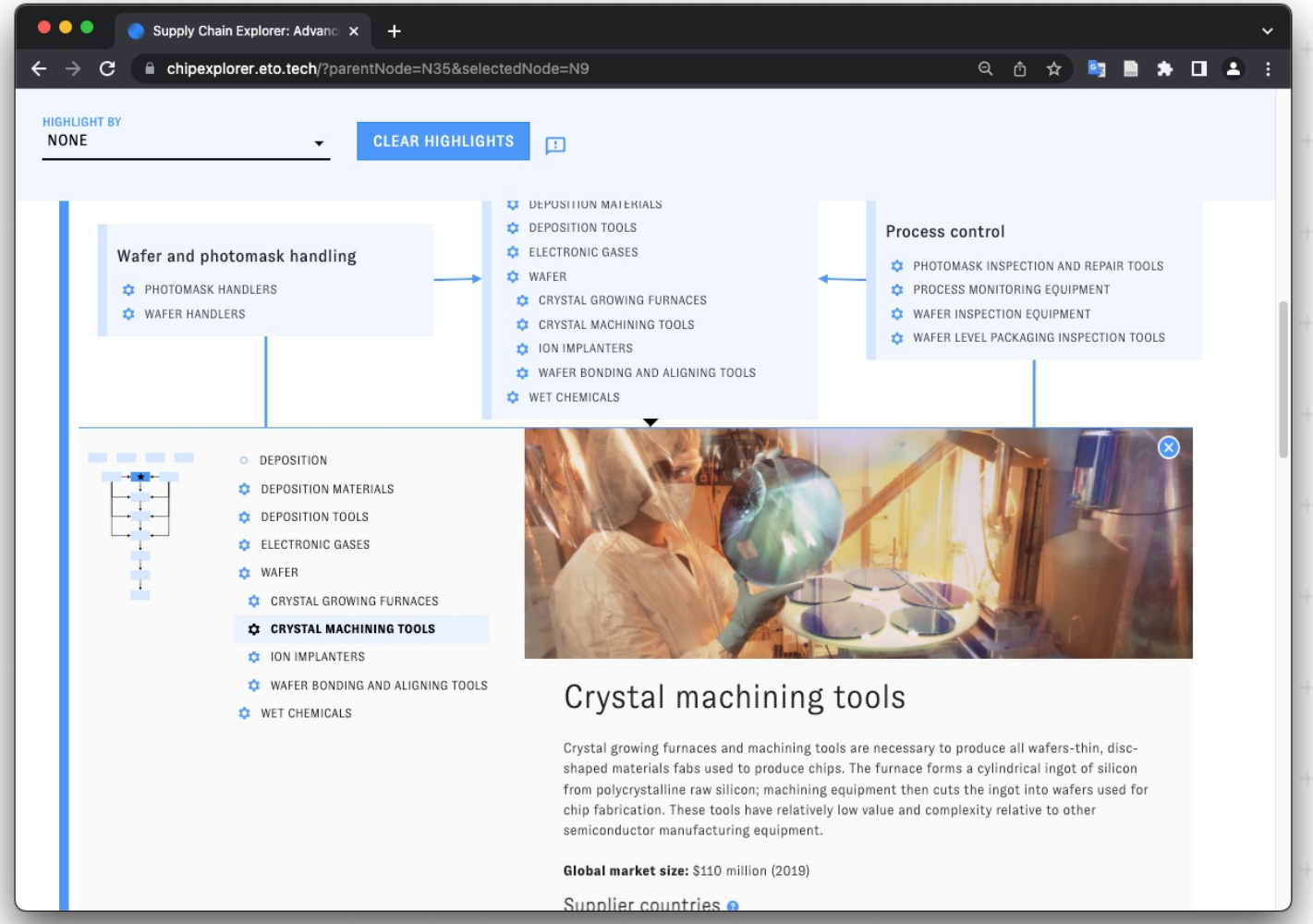
Using the Map of Science

- 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

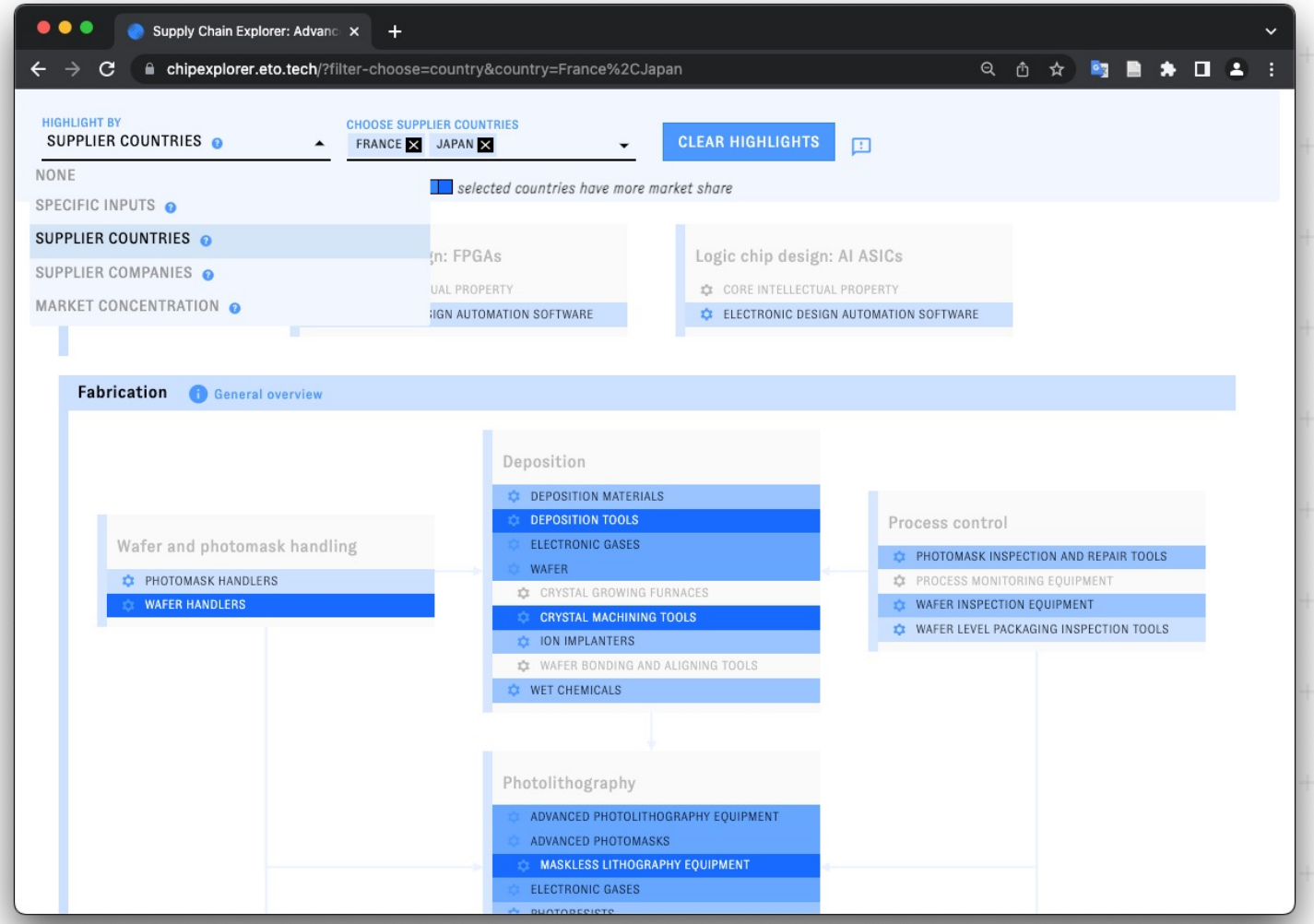
How the Explorer works



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

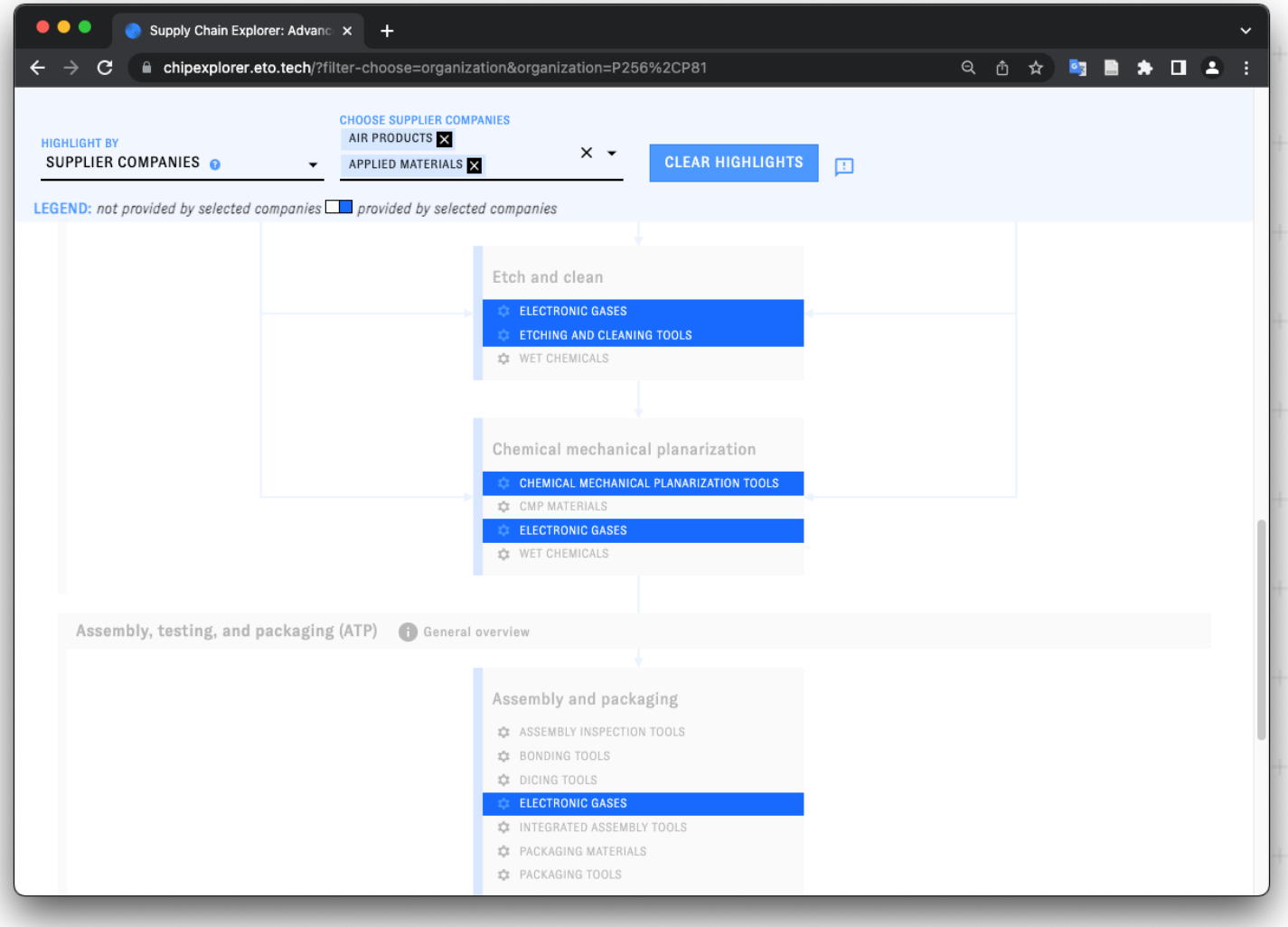
How the Explorer works



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

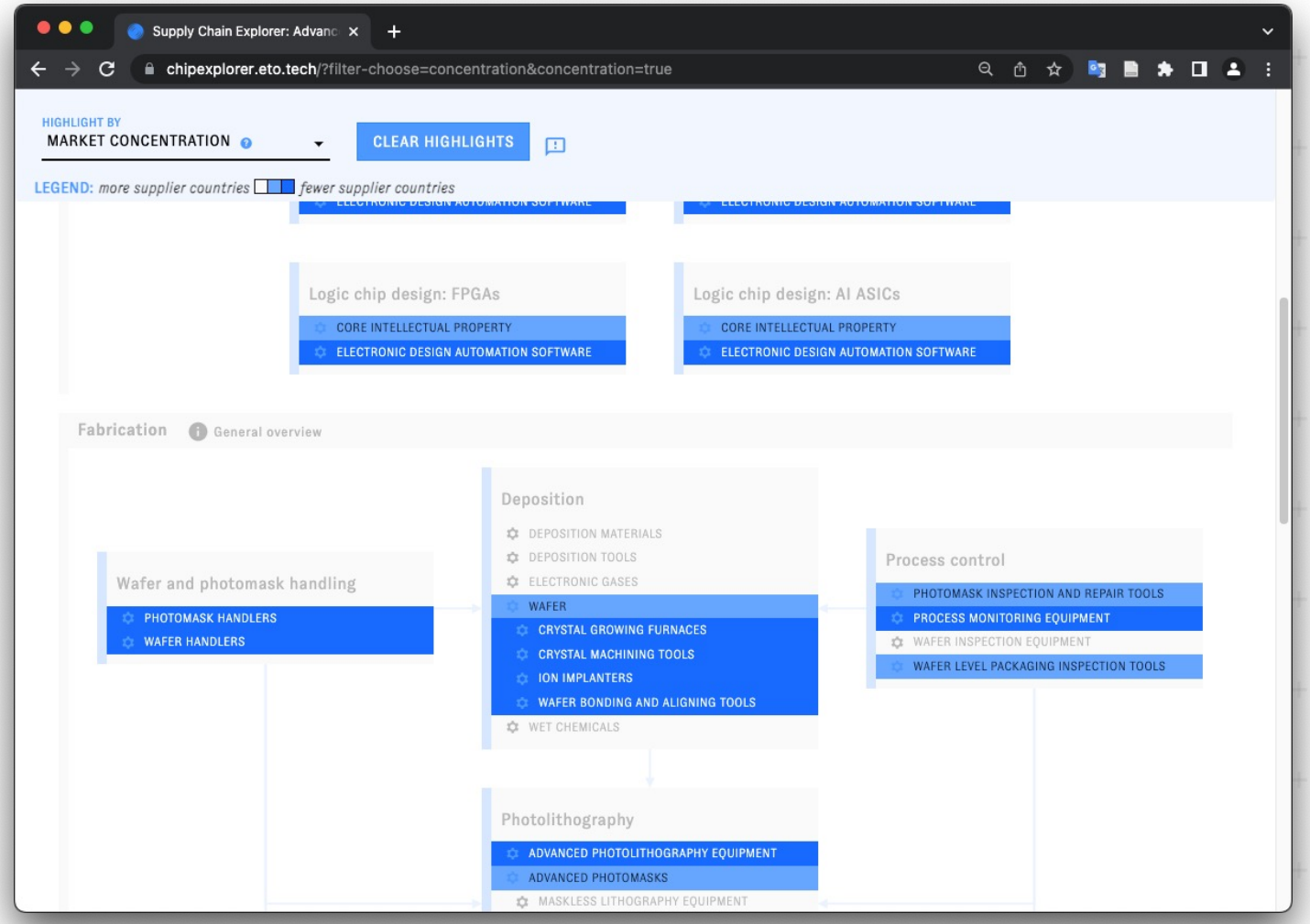
How the Explorer works



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
- Follow ETO's blog for news, demos and insights
- Reach out to CSET's analysis team: 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)

Emily Weinstein
Research Fellow
emily.weinstein@georgetown.edu

Zach Arnold
Analytic Lead, ETO
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

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

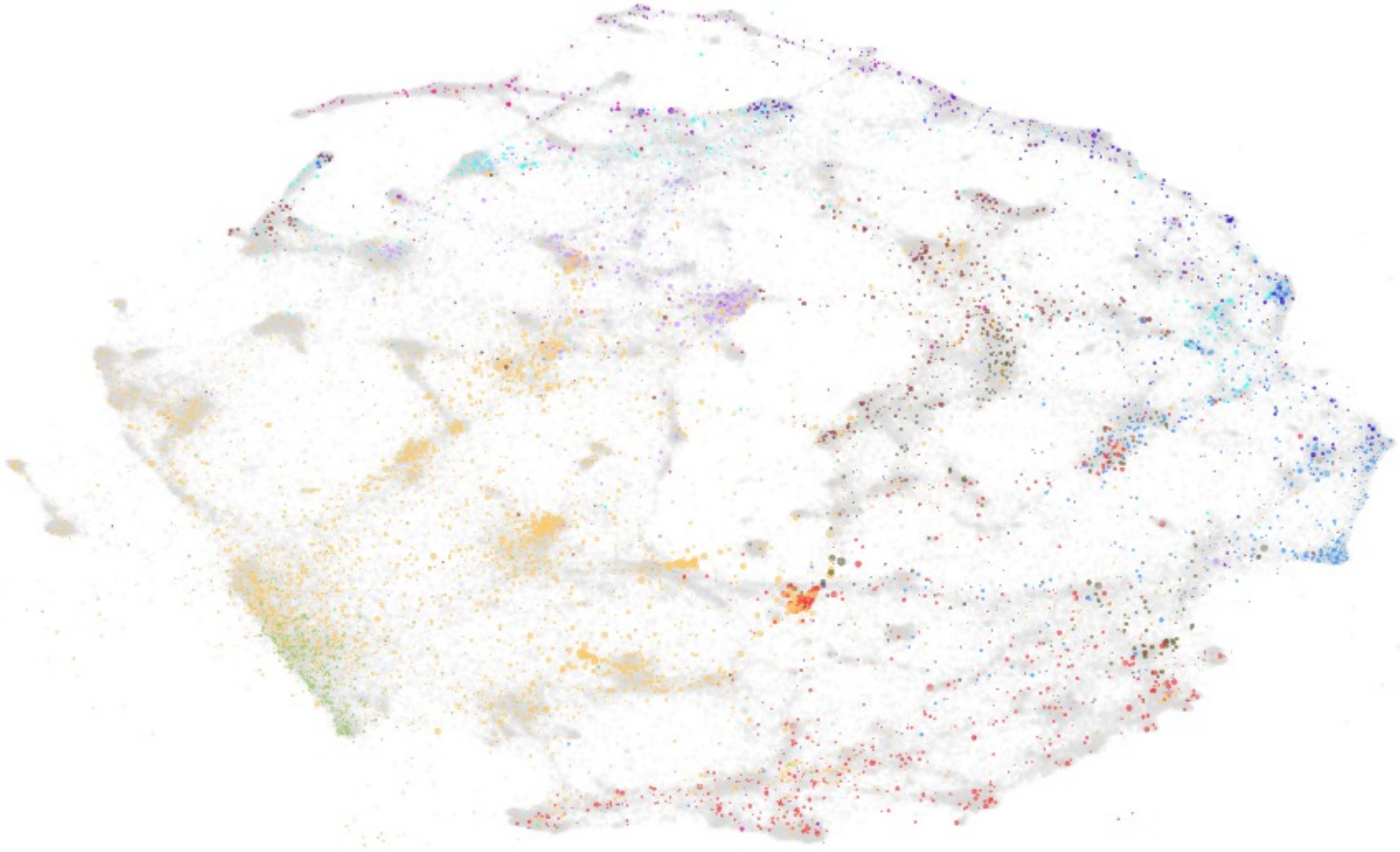
ETO works in progress

- 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

- Maximizing modularity of citation
- Correlates (broadly) with intuitively interesting characteristics - topic, language, etc.

For more information: ETO documentation, Rahkovsky et al. 2021

How the Map works

Top clusters ADD/REMOVE COLUMNS

Cluster ID	Most common subject category	CSET phrases	Cluster size	Growth rating	Patent impact rating
109172	computer science	Power Grid Operation, Environment Monitoring System, Monitoring System Based, power grid, Grid Operation Situation	84	97.80	0.00
109530	computer science	Robot Operating System, mobile robot, Hector SLAM, SLAM algorithms, robot navigation	75	97.23	50.16
72278	physics	semiconductor optical amplifier, optical frequency encoded, logic gates, optical NAND gates, Reflective Semiconductor Optical	75	93.82	57.42
104245	computer science	Medical Things, Health Monitoring Systems, structural health monitoring, IoT, compressed ECG signals	61	93.61	0.00
116016	social science	Health Monitoring System, health vitals, Patient Health Monitoring, Smart Health Care, IoT	61	93.39	50.16

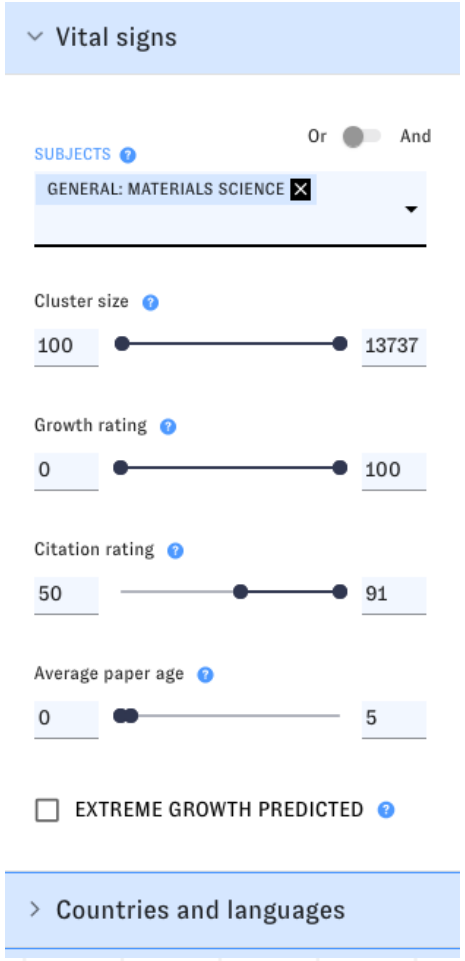
Map "list view" showing key concepts and user-selected metadata fields for a set of fast-growing clusters

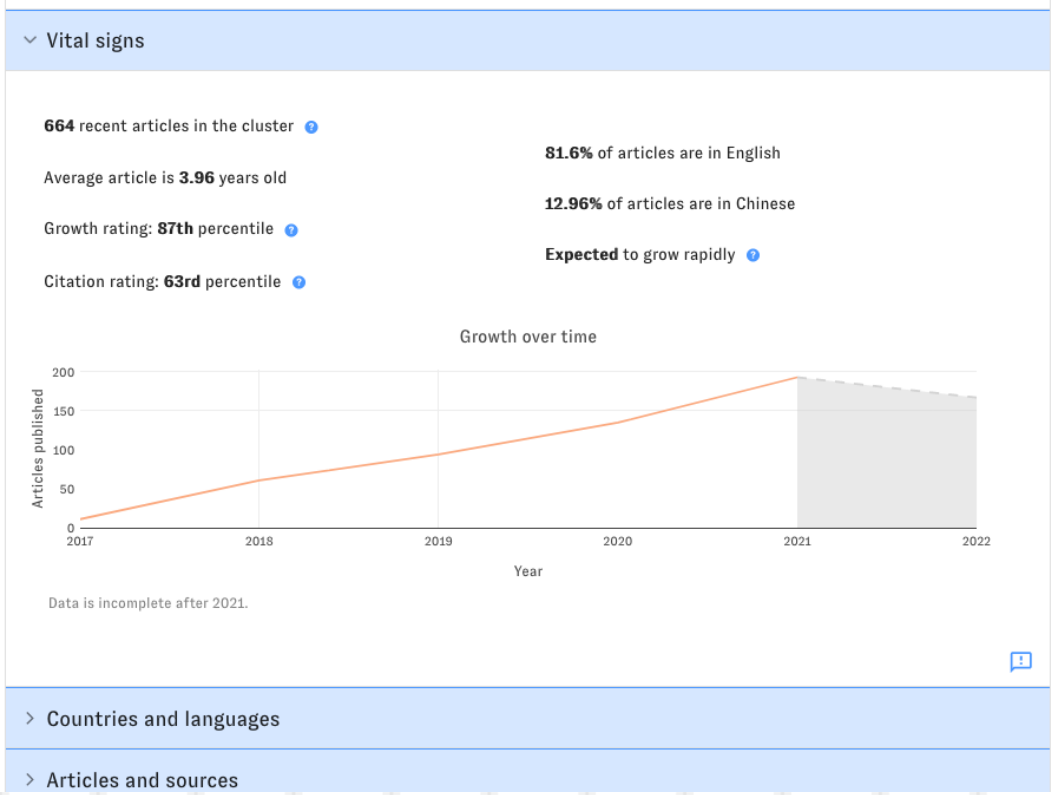
Create cluster-level metadata based on constituent articles

- 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





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

ADD/REMOVE COLUMNS

Cluster ID	Most common subject category	CSET phrases	Cluster size	Growth rating	Patent impact rating
109172	computer science	Power Grid Operation, Environment Monitoring System, Monitoring System Based, power grid, Grid Operation Situation	84	97.80	0.00
109530	computer science	Robot Operating System, mobile robot, Hector SLAM, SLAM algorithms, robot navigation	75	97.23	50.16
72278	physics	semiconductor optical amplifier, optical frequency encoded, logic gates, optical NAND gates, Reflective Semiconductor Optical	75	93.82	57.42
104245	computer science	Medical Things, Health Monitoring Systems, structural health monitoring, IoT, compressed ECG signals	61	93.61	0.00
116016	social science	Health Monitoring System, health vitals, Patient Health Monitoring, Smart Health Care, IoT	61	93.39	50.16
91826	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: ~0.2) uses four indicators:
  - Vitality: 1/(avg reference age in forecast year)
  - Stage: 1/(years since peak year)
  - Paper Age: 1/(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

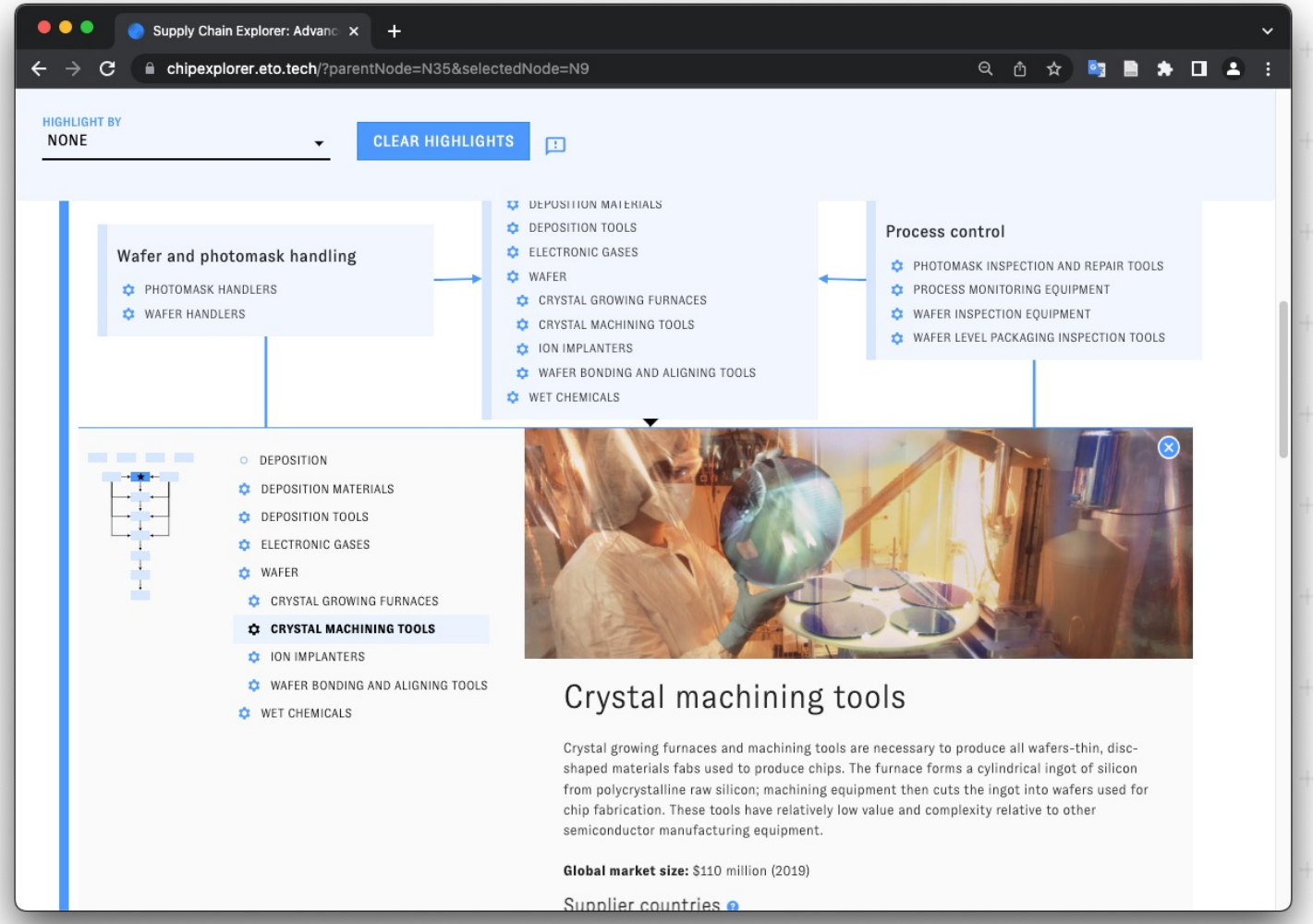
Building a broader picture: the Supply Chain Explorer

- 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

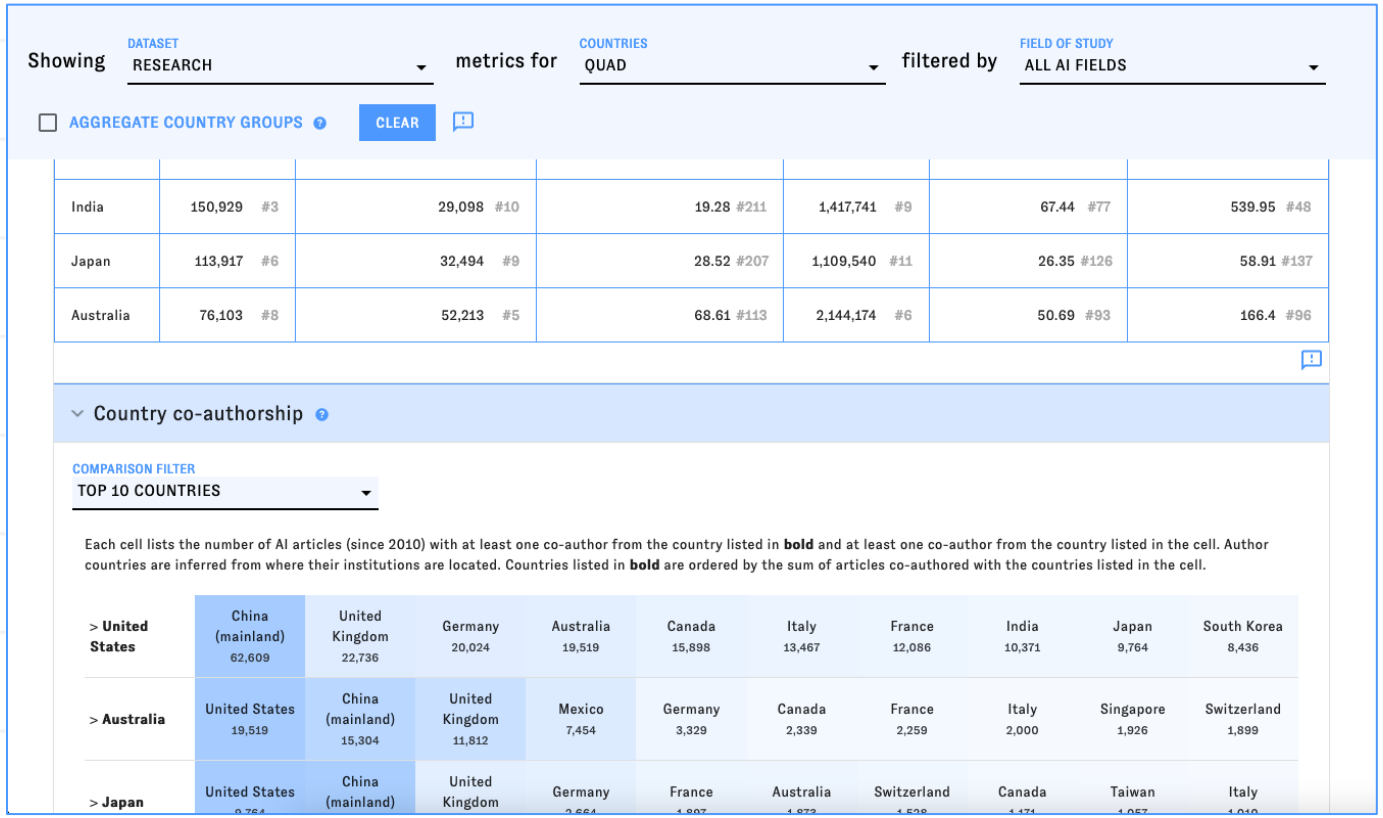


Gives users an accessible birds' eye view of the market and global context for areas of particular interest

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)



cat.eto.tech