Appendix 5A

DIMENSIONS OF DATA CHALLENGES

Timeliness and frequency of data collection. These will vary with context. For example, during the COVID-19 pandemic, new small and medium-sized firms were emerging daily in response to mask and respirator shortages. High-frequency data collection was important to determine whether US production capacity for these items could meet demand. In contrast, daily data collection for semiconductor manufacturing is unlikely to be of similar value since it takes 3 years and $10 billion to build a greenfield semiconductor facility, and 6 months to a year to redesign a chip to be produced in a different fabrication facility.

Accuracy and completeness of data collected. Is the sample complete or reflective of the total population? Can the data be trusted? For example, during the pandemic companies reported on business-to-business sites whether they sold masks and in some cases listed themselves as manufacturers although their manufacturing was not domestic. And patents may represent a limited sample of inventions (research shows that only 42% of new-to-market products are patented). Finally, data can be deceptive, depending on how they are presented. For example, news and policy outlets have reported an exponential rise in China's patenting, but researchers have shown that this rise is due to coinvention by Chinese employees at multinational firms with non-Chinese inventors as leads on the invention. The two reports may lead to different conclusions and policy actions, and by themselves are likely too little to effectively inform clear policy action.

Granularity. Data often lack the granularity necessary to answer important questions. For example, trade data don't provide product-level information or volumes, making it difficult to map critical supply chains. Paper or patent keywords or human-selected classifications may not capture the evolution of an emerging technology for which the most relevant terms may be evolving.

Privacy protections. Depending on the type of data, it may be important to maintain individual or firm privacy. In some cases (e.g., the US Census Bureau or Bureau of Labor Statistics), individual or firm privacy is maintained by the government; in others, it's by a nongovernmental third party such as an FFRDC or trusted university partner. For example, private firms prefer that supply chain data and data on composite materials used in emerging technology standards not be held by government.

Ease of access and the cost of collecting, storing, and validating data. Technology assessment is particularly challenging when data do not exist or are not in an easily accessible form. Researchers may need to collect data themselves (e.g., workers' knowledge and skills on a particular machine in a selected fabrication location), data may not be easily accessible if they are maintained by private entities (e.g., private firms' manufacturing or supplier data, or Amazon's supply chain data), or data may be geographically or institutionally distributed (e.g., efforts to understand national access to critical products may require collecting data from multiple private firms). Sometimes data are available but costly or slow to access, such as Census data. Data may be differentially available to different individuals, depending on their status, affiliation, cultural norms, or other factors. A final dimension of data collection, storage, and validation is the cost—in time or money—of those activities. For example, the collection and accurate interpretation of undocumented firm- or community-level data may require being on-site for weeks at a firm or in a community. Data validation may require contacting individual firms to confirm publicly posted information or crowd-sourcing information from locals. Storage costs depend on the data size.

INTERSECTION OF DATA TYPES AND DIMENSIONS

Appendix table 5A-1 shows the range of data used by different Network projects to answer different types of questions.

The technology's stage of the S-curve to some extent determines what data are available and their prevalence (in the case of measures of the economy or societal effects). Data used to assess technologies at an early stage of the S-curve generally focus on discovery and invention activity; data on inputs to the innovation process may include human capital, investment, and grants, often combined with bibliometric data, which are predominant at these stages (and generally publicly available). For technologies further up the S-curve, data related to commercialization and diffusion may include patent licenses and company products, but can be more challenging to collect because they are often proprietary. Data needs on the outputs of innovative activity include standard bibliometrics (patents and papers), products and services developed, measures of technology adoption, and productivity and labor market effects of technology use. As adoption increases, data on the development of supply chains, interactions between the technology and economic and social systems, and dependencies around the technology become critical for assessment. Common types of data here include product designs and attributes; user preferences; prices, production process inputs, tasks, and organization; labor requirements including skills, wages, and hours worked; and production, consumption, and trade quantities. Also needed are qualitative data (e.g., from ethnography, interviews, and surveys) from scientific and technical experts; organizations including firms, governments, and nonprofits; their (potential) customers, communities, and the public on technology commercialization and adoption bottlenecks; processes by which outcomes are achieved; on-the-ground realities of new technologies; organizational behavior; and implementation of rules and legislation.