Economic Measurement of AI: An Agenda and Preliminary Progress

By Kristina McElheran, University of Toronto

[Preliminary and incomplete]

I. Introduction

A group of technologies commonly referred to as "artificial intelligence" have been developing and diffusing rapidly over the past few years. [EXAMPLES HERE]. It is widely believed that these technologies have the potential to dramatically affect economic and social outcomes such as productivity, innovation, inequality, and the nature of work (Agarwal et al. 2017, OTHER CITATIONS HERE]. Given the potential – and some would argue, the threat[1] – of these technologies, it is crucial that the research community be proactive in considering the possible implications of these new technological developments.

To that end, the NBER held the first conference on the Economics of Artificial Intelligence in 2017 to convene a group of leading economists to set a research agenda for researchers working in the area of AI. The direction for the conference was:

> "The context is this: imagine back to 1995 when the internet was about to begin transforming industries. What would have happened to economic research into that revolution had the leading economists gathered to scope out a research agenda at that time? Today, we are facing the same opportunity with regard to AI. This time around we are convening a group of 30 leading economists to scope out the research agenda for the next 20 years into the economics of AI."

A range of topics were introduced, from X to Y, AND OTHER DETAILS OF THE CONFERENCE HERE. The emphasis of the conference was largely theoretical, with the exception of Erik/Chad/Dan's paper, which provided some preliminary statistics… [MORE HERE]

This is understandable, as we have no systematic data on the use of AI in the economy at the moment. The technology is still relatively new, and the "hype" far outstrips the actual practice at the moment (GOOD CITATION HERE?). However, a notable gap in the agenda-setting process was a systematic discussion of how the profession should think about and pursue empirical measurement. In the published volume that followed the conference, Raj and Seamans (2018) outlined some considerations for the measurement of robotics in the economy, some of which apply here as well. These and other key considerations – particularly as they concern AI, and not just automation, in general -- warrant early and serious attention.

Why is measurement of AI so important to worry about at this early stage? Recent work at the frontier of economics, management, and technology suggests that we already risk being late to the party when it comes to measuring the diffusion and impact of AI. Research on the economic impact of the internet was severely influenced by data constraints, and many important dimensions of that technological sea change remain poorly understood to this day. Recent work has advanced our

[1] Musk and Hawking stuff….

measurement – and, hence, our understanding – of technologies that emerged from the Internet Age such as "big data" analytics and cloud computing, however we also learned a great deal about how far we have to go.

This paper outlines key issues in the measurement of new technologies at early stages of their development. It presents some of the first empirical evidence we have on technologies that will serve as critical "test cases" for measuring and understanding AI in the broader economy: predictive analytics and cloud computing. Finally, it presents key insights from these recent efforts that can be applied to future data collection efforts.

This paper proceeds as follows. First, I briefly motivate why we need empirical evidence and the importance of early measurement of new technologies. Some examples of challenges, sucesses, and lessons learned from studying the internet foreshadow core concerns for AI. Next, I present early findings on two key technologies: cloud computing and predictive analytics. Cloud computing is essential for gauging the diffusion and application of AI in the economy due to the important infrastructure it currently provides – particularly for young firms -- for advances in machine learning and artificial intelligence. Predictive analytics is a broader class of IT-based practices that largely relies on older techniques (such as multiple regression) but will quickly incorporate artificial intelligence. Understanding how firms adopt, deploy, and answer surveys about these technologies will have important implications for how we measure artificial intelligence as it goes "mainstream" in the economy. Finally, I outline some key take-aways from these efforts and considerations for future data collection efforts. Finally, I preview upcoming efforts that promise to greatly advance our understanding of the diffusion and impacts of artificial intelligence in the economy.

Throughout, I lean heavily on my own research and that of my co-authors, advisors and close colleagues. This is not to say that there have not been other, even more meaningful contributions concerning the economics of technology and technological change. However, many of the key issues of interest for measurement never make it into the published versions of our papers (which merits consideration as an issue all of its own), and so I have less visibility to the key tradeoffs and data quality issues other projects have faced. Also, to give credit where credit is due would culminate in a truly lengthy review article on the state of the field – a worthy undertaking, but beyond the scope of this paper.

II. Why measurement matters

a. *Second Machine Age Reasons*: technology is now digital, exponential, interdependent, etc. diffusing rapidly… "A difference in degree has become a difference in kind". We cannot just apply the things we thought we learned about prior technologies.
    i. Give an example? For instance, we thought IT was largely a large-firm game, but the cloud is changing this.
b. Discussion of GPT's and how much had to change in the economy for them to have impact. It was not clear, ex ante, what had to change or how. It might be reasonable to expect a similar level of uncertainty and an ongoing need to test our assumptions and the magnitudes involved.
c. Dramatic differences in the rate of change across technologies and how hard it is to predict, from a theory perspective. DO WE HAVE EVIDENCE ON THIS?

Finally, and very simply, facts matter. And facts seem to matter more when the hype level is very high, because the marginal fact may be quite valuable. I personally know one young person who decided not to pursue a radiology career last year because machine learning algorithms "already can read xrays better than humans can" (citation). Except it turns out that they cannot (Citation). And, while perhaps they are on track to be better than humans on a time scale that matters for people making education and career decisions today, that is a different conversation –and one that would surely benefit from more data.

Firms invest in response to the "hype" cycle as well.[FORRESTER HYPE CYCLE HERE?] While there are historical examples of first-mover advantages that receive a great deal of attention (Examples here?), there are many reasons for firms to wait until technology is better developed and complementary resources – either within or outside the firm – are better developed.[2] A mistaken understanding of how mature a new technology is will certainly not improve the efficiency of firm investments.

To focus the discussion, this paper is going to restrict attention to **firm** adoption and exploitation of technology. While consumer adoption of technologies is an important and worthy topic of discussion, it is already receiving considerable attention in the Information Systems, Economics of Digitization, and related fields. This is in no small part due to the availability of new digital data on consumer behavior. Of potential concern is how much of the profession's resources are being directed to these particular "lampposts"[3] not because these are the questions of greatest economic import, but because the costs of trying to research firm behavior without readily available data are so high.

*Why **early** measurement matters*

Aside from the very useful notion of an "early warning" system and the value of having up-to-date statistics, there is a very basic econometric justification for tackling data collection and measurement early. Where possible, sound empirical research relies on techniques to address unobserved heterogeneity in firms that could influence the adoption or returns to adoption. Typically, we employ fixed-effect estimators that rely on changes over time within a given firm or establishment to identify the effect of interest. To do this, we need a reliable "pre-treatment" to identify early effects; barring this, we need sufficient change over time to trace out the correct magnitudes. But if a technology diffuses quickly, we may lack enough within-firm variation to strip out unobserved firm characteristics. Throughout this essay, I will describe specific examples of too little, too late or near misses in data collection that help motivate why sooner will be better when it comes to sound measurement of AI.

Why industry context matters

Note Raj and Seamans work here and make clear that, while there are important links between robotics and AI – in particular, the potential for respondent confusion! – there are important differences as well. **In particular, a key difference between robotics and un-embodied AI such as machine learning is the sectors in which they will matter.** This has implications for Census' collection efforts, as the Manufacturing surveys have historically been the most innovative – but that is not where phenomena of

[2] For instance, Bai, Jin, McElheran, and Williams (2018) find that later adopters of Enterprise Resource Planning technologies did better than earlier adopters. OTHER EXAMPLES HERE?

[3] Lamppost quote here…

concern such as automation of white collar work is likely to take place first. For instance, much more attention should be paid to services.

III. What we learned from studying The Commercial Internet

The example of the commercial internet is instructive for thinking about data collection methods and measurement issues in the early years of new technology diffusion. The network technology we know as The Internet became available to firms in 1994 and diffused so quickly and widely that, by 1995, most researchers consider availability to be nearly ubiquitous [DETAILS, CITATIONS HERE]. Use of the internet varied among firms in important ways. For instance, …EXAMPLES HERE]

The primary windows into this phenomenon in the U.S. have been administrative data collected by the U.S. government and proprietary data sets collected for a variety of non-statistical purposes (discussed below).

Method 1: Administrative Data

We learned many interesting things from official statistics collected by the U.S. Census Bureau. Unfortunately, there was a considerable lag between the introduction of the Internet and official data collection. Administrative data collection on the use of the internet did not start until the US Census Bureau conducted the Computer Network Use Supplement to the 1999 Annual Survey of Manufactures. This rich survey revealed important variation in how firms actually applied Internet-based technologies (McElheran 2015, Atrostic and Nguyen YEARS), but was only done in one year and only in manufacturing. Few researchers outside the Census used this data to study the impacts of internet use.[4] One exception is Forman and McElheran (2018), which examines the extent to which firms shifted transactions from inside the firm to market-based exchange in the wake of the commercial internet. It relies extensively on the CNUS to disentangle internal versus external coordination over the internet to disentangle the effects.

Method 2: Proprietary Data

Much of what we know about the early days of the internet relies heavily on proprietary data, with important pros and cons that matter for thinking about the measurement of AI. A key resource in this stream of research has been the Harte Hanks Intelligence Infocorp (variously referred to as HH or IC) and various reincarnations thereof. This data set DESCRIBE HERE. And has been used extensively in research on IT diffusion and Productivity (HUGE LIST OF CITATIONS HERE).

The purpose of this data set is to provide market research to vendors of ICT goods and services. It has the following virtues…. In particular, it has the virtue of being establishment-level, large, and representative of the non-small-business economy (it primarily surveys establishments that have 100 or more employees). Some tradeoffs….It has been expensive to acquire, as wel. In recent years, they have ceased their direct-calling approaches and have turned to a proprietary algorithm for predicting technology use at firms, limiting its ongoing usefulness for economic research.

i. Forman, Goldfarb, Greenstein work

[4] Census employees BK Atrostic and Sang Nguyen show…. Patricia Angle (YEAR) studies e-commerce.

Other proprietary data sets briefly described here…????

Key Learnings from Internet Diffusion and Productivity Research

1. **Pre-treatment data is important.** Getting a clean and representative "before" year is incredibly useful for identification. The rapid diffusion of the Internet helped with that – because the demarcation of "before" and "after" is very crisp (see also Athey and Stern E911 on benefits of this). However, capturing this well relies on panel data, and, in particular, panel data that keeps a reasonably large and stable sampling frame. The Census annual surveys rotate their sample frame every five years, so getting good coverage of a "before year" can be difficult (particularly if adoption rates are low and samples become small). For instance, McElheran and Forman (2018) had to rely on the complete Census in 1992 to get a large-enough "pre-treatment" sample. Many proprietary data sets are collected without any intention of supporting panel data analysis. For instance, the linkages between establishments within firms and across years is extremely incomplete and error-prone in the Harte Hanks data, requiring extensive data cleaning efforts before analyzing it.[5]

2. **Within-firm variation is an issue.** While a lot of our theorizing and evidence concern firms as the unit of analysis, there is important within-firm heterogeneity in the use and impacts of technology. Forman et al. (YEAR) find that not all units of a firm have the complementary human resources that they need to deploy internet technology, but that they can leverage internal resources elsewhere in the firm. McElheran (2014) shows that decision rights over IT purchasing vary tremendously among units of multi-establishment firms, which has important implications for where and how new technology is applied (and how well-suited it is to its local operating environment).

3. **Supply chains matter**. A key distinction in the heyday of the commercial internet was between business-to-business ("B2B") and business-to-consumer ("B2C") applications. No official statistics were collected on this distinction, though documentation on E-Stats suggested that the e-commerce statistics be interpreted as B2B. However, McElheran (2015) demonstrates important differences in the use of the internet according to whether plants are in primarily B2B or B2C industries. Also, Harte Hanks data on B2B vs. B2C shows that there is considerable B2C activity in the manufacturing sector, a fact corroborated by the much earlier Survey of Manufacturing Technologies (which asked about details concerning a plant's client base).

   A critical shortcoming of most current data sets is the difficulty of seeing the buying and supplying relationships on which adopting firms depend. Some headway has been made by looking at 10-K filings, in which firms list their "material customers." However, it is only available for public companies and only shows one direction of the supply chain relationship.[6]

IV. Big Data, Data-Driven Decision Making and Analytics

While it seems like ancient history already, the hot thing in technology before AI was "big data" and predictive analytics. The promise of big data and advanced analytics of that data was to take

[5] We wrote up a lot of notes on this, can I post this somewhere and reference it? Include it in an Appendix?
[6]We have preliminary evidence that this matters for evaluating the impact of technologies. Bai et al. (2018) find that the benefits of ERP adoption spill over to the suppliers of firms that adopt ERP systems.

advantage of a new ability to measure things that we have never before been able to measure to achieve new efficiencies and innovation. MORE HERE.

Measurement was a central concern for these technologies as they emerged, too. Advances in understanding the diffusion and economic implications of big data and analytics have come primarily from new techniques to measure these approaches directly via the development and deployment of new surveys and indirectly using public data and machine learning techniques.

Method 3: Survey Data

One tried-and-true approach to learning what firms are doing is to ask them. [Early DDD work here?] In 2010, the Census Bureau worked with economists to collect novel direct data on a range of management practices (Bloom et al. 2016 & 2018). The new Managerial and Organizational Practices Survey went to a sample of over 50,000 establishments in the manufacturing sector whose response was required by law. This is significant, as much of the survey data used in economic and management studies is subject to sampling and response bias that is difficult to correct for and often overlooked.

The first wave of the survey asked about data-related management practices, which gave insights into how firms were making use of their data resources (Brynjolfsson and McElheran 2016 & 2018). The survey was repeated in 2015, and new questions were added about predictive analytics.

- **a.** What we found
  - **i.** Predictive analytics takes up where DDD left off (pending disclosure review – have to be careful how I talk about this until I get approval)
  - **ii.** It is more widespread and diffused faster than anticipated – wish we had caught it earlier!
  - **iii.** **Measurement error is a concern – it is good to combine measures were possible (Same with HH data).**
  - iv. Adoption may not be an absorbing state – it wasn't for DDD (pending disclosure review)
  - v. Easier to compare extensive margins and "extremes" – difficult to deal with continuous shifts.
  - vi. We didn't specify whether predictive analytics means regression analysis or machine learning – so you can have a "technology" that really encompasses different underlying techniques.

Method 4: Using analytics to find big data analytics

- a. Saunders and Tambe on Analytics – 10K scraping
- b. Implication for studying AI: can use machine learning to find machine learning, but only if firms report that they are doing it. Key concern is prototyping and experimentation that firms claim but that really aren't production ready. Cheap talk concerns…

Things we Learned from Studying Big Data & Analytics

1. **Consider that the frontier may move in ways that are hard to capture in your questions.** Annual surveys may not be enough. Plan ahead for a "moving frontier" and new best practices (we struggled a lot with this, have to noodle on how much to reveal!)
2. **Distinction between the "what" and the "how":** Predictive analytics can mean anything to many people. There is a vast difference between analyzing data to make predictions using regression analysis and doing the same things using multivariate regression. But they are called the same things, to the detriment of data collection.

V. A happy Accident: IT Outsourcing and Cloud Computing

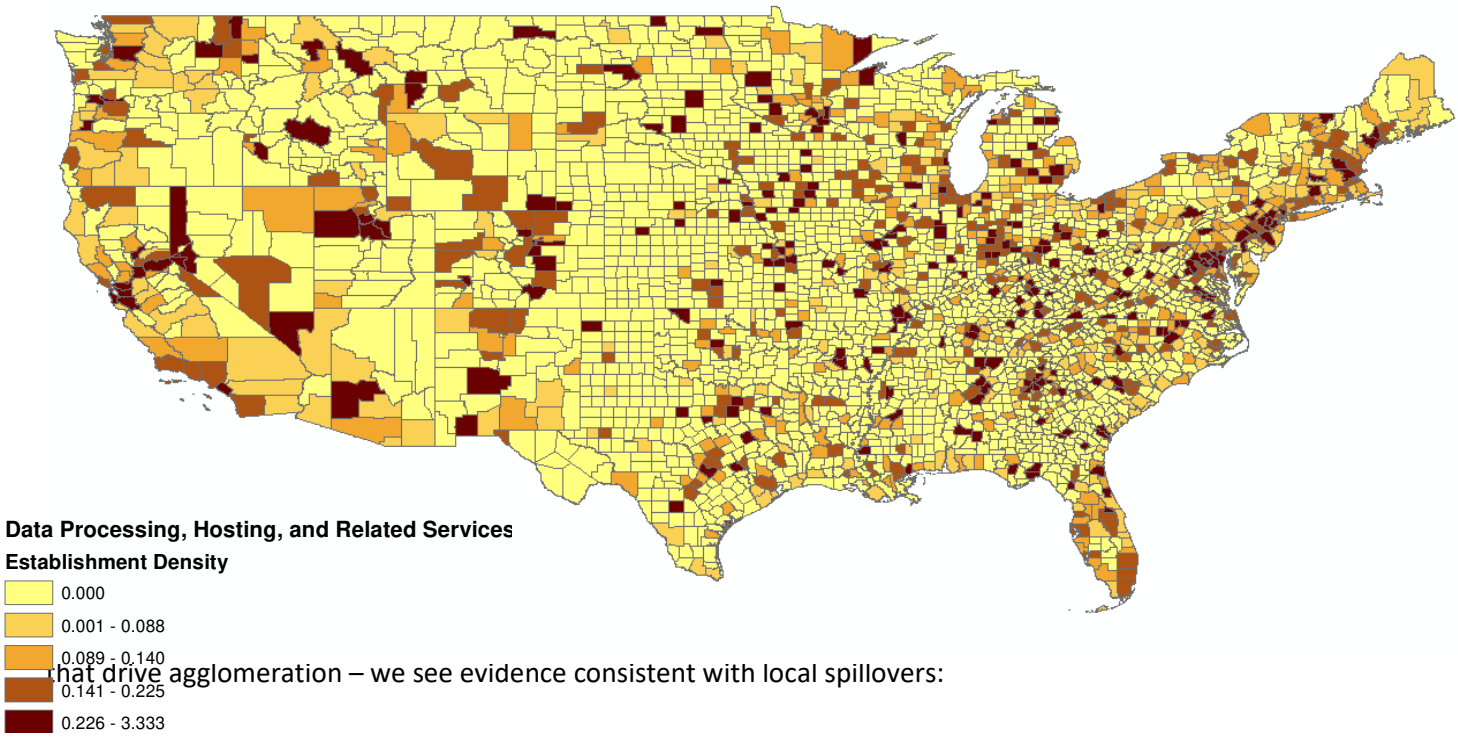
Method 5: Get Lucky

Cloud introduced in 2006, was captured in a measure added to the manufacturing survey in 2006 onward. The only virtue of this approach is that is worked reasonably well. But largely, it is a cautionary tale about getting in front of data collection of new technologies. Many details are missing…

Not in services yet!

- **a.** Census split out different margins of IT spending just as cloud diffused
- **b.** Only did it in Manufacturing
- **c.** What we learned (some of this is in the Cloud paper and some is not)
  - **i.** Diffused fast
  - **ii.** Prices fell fast (Corrado et al. stuff here)
  - **iii.** There is a geographic component:

Figure 1. Data Processing, Hosting, and Related Services



that drive agglomeration – we see evidence consistent with local spillovers:

[how best to show? Another map? Regression output?]

Also differences by age:

Table X. IT Expenditure Breakdown by Type for Young and Older Plants, 2006 – 2014

	(1)	(2)	(3)	(4)
Dependent Variables	**% Expenditure on IT Capital Flows**	**% Expenditure on IT Services**	**% Expenditure on Software**	**% Expenditure on Equipment**
Young	0.009** (0.004)	-0.004 (0.003)	-0.015*** (0.003)	0.010*** (0.003)
Late	-0.054*** (0.002)	0.021*** (0.002)	0.036*** (0.001)	-0.003* (0.002)
Young x Late	-0.006 (0.004)	0.011** (0.003)	0.002 (0.003)	-0.007* (0.004)
Industry Fixed Effects	Y	Y	Y	Y
N	~239,700	~239,700	~239,700	~239,700
R-Squared	0.152	0.086	0.097	0.131

The digitization of the economy in the wake of the Internet provided lots of new data sources. An important one with implications for measuring AI is the approach the Sonny and Lorin took: to find the IT, measure the IT workforce.

Method 6: Measure the Complements

Details of Lorin and Sonny's work here… maybe get some input from Sonny on this?

Could see applying this to AI as long as the skill set is differentiable from general computer science skills.

Key Learnings from Studying Complements:

1. **Complementary human capital matters**.
2. ???

VII. OVERALL, Central Measurement Issues

- **c.** Extensive vs. intensive margin
- **d.** Quality/intensity differences
- **e.** Firm-level vs. establishment level
  - **ii.** How to characterize firms from establishment-level data?
  - **iii.** Broken links in HH (and how we fixed them)
  - **iv.** Where is the locus of decision making? (This varies within firm – McElheran 2014)

v. Where is the boundary of the firm? (Forman and McElheran)
vi. Why it matters (we need to do some work on this)

f. Stocks vs. flows – possibly less important in the age of the Cloud

g. Public vs. Private firms – selected samples in most data sets!

h. Lack of visibility to important covariates

vii. Age (Jin and McElheran 2017)
viii. Organizational Capital (Brynjolfsson, etc.)
1. Recently visible in the MOPS, but still missing things and only for Manufacturing
ix. Traditional Capital – possibly substitutes (Wang – don't we have some findings on this?)

i. Language matters – what is "AI"?

j. Given Rob's article, need to be careful about automation and whether we need to distinguish between robots and AI

.

CONCLUSION: We need to get better at this, but recognize that we cannot run before we can walk. Our current measurement of how firms use technology at every level is poor.

This is a huge list. So, a place for empirical researchers – particularly those of us involved in data collection – to focus is on understanding the most critical places to focus our attention.