Panel Discussion:
How do we validate patent metrics derived from semantic analysis?

Josh Lerner
Harvard Business School and NBER

Innovation Information Initiative Technical Working Group Meeting
December 3, 2022
A caveat

• Not an AI expert!
• But have used ML/NLP methodologies to...
  • Predict ultimate patent citations (Lerner-Seru 2022).
  • Identify finance patents (Lerner et al, 2021).
  • To identify disruptive technologies (Kalyani et al. 2022)
The challenge

• Wide variety of choices:
  • Pre-treatment of patent text.
  • NLP methodologies used.
• Huge corpus of patent text.
• Has led to inevitably to an explosion of measures.

• Good news: More options than front page data pioneered by Hall, Jaffe, and Trajtenberg (2001).
• Bad news: Hard to sort through and interpret.
The challenge (2)

• Non-endogeniety of patent text:
  • Historical practices such a “copying claims” to provoke interferences.
  • Gaming of claims around standards (Righi and Simcoe, 2020).
  • Proliferation of software to management assignment and review process
A preliminary checklist (in spirit of Lerner-Seru 2022)

• Tank story.

• Potential questions:
  • Does a human audit validate the sorting?
  • Do the results ultimately make sense?
  • Can the results be validated by using other samples?
  • Can we understand what criteria the ML/NLP models are using: