Big Data Strategy

Aija Leiponen, Cornell University, aija.leiponen@cornell.edu

Plan for the session

- Introduction – how did we get to this point? What are the key technological and market trends?
- Data governance – what are the economic properties of data? What are the implications for commercialization?
- Data platforms – what is the platform revolution about? How will data platforms emerge?
- Summary and conclusion – what do we know, what do we not know?

Primarily from papers co-authored with Pantelis Koutroumpis and Llewellyn Thomas

- "Invention Machines: How Control Instruments and Information Technologies Drove Global Technological Progress Over a Century of Invention"
  - Presented at the NBER Summer Institute July 2016
- "Economic Characteristics of Data Goods"
  - Working paper available
- "The (Unfulfilled) Promise of Data Marketplaces"
  - Under review in Information Systems Research special issue on digital platforms
- Ongoing empirical work on data contracts with Joy Wu

Introduction: Economics of information and IT

Aija Leiponen
Cornell University

Characteristics of ICT

- Digital

  → information that can be reduced to bits
- General Purpose Technologies

  → pervasive, ubiquitous use
- Network technologies

  → network effects
- Require skills – "IT literacy"

  → skill-biased technical change (premium for skills in the labor market)

Designing a Digital Future

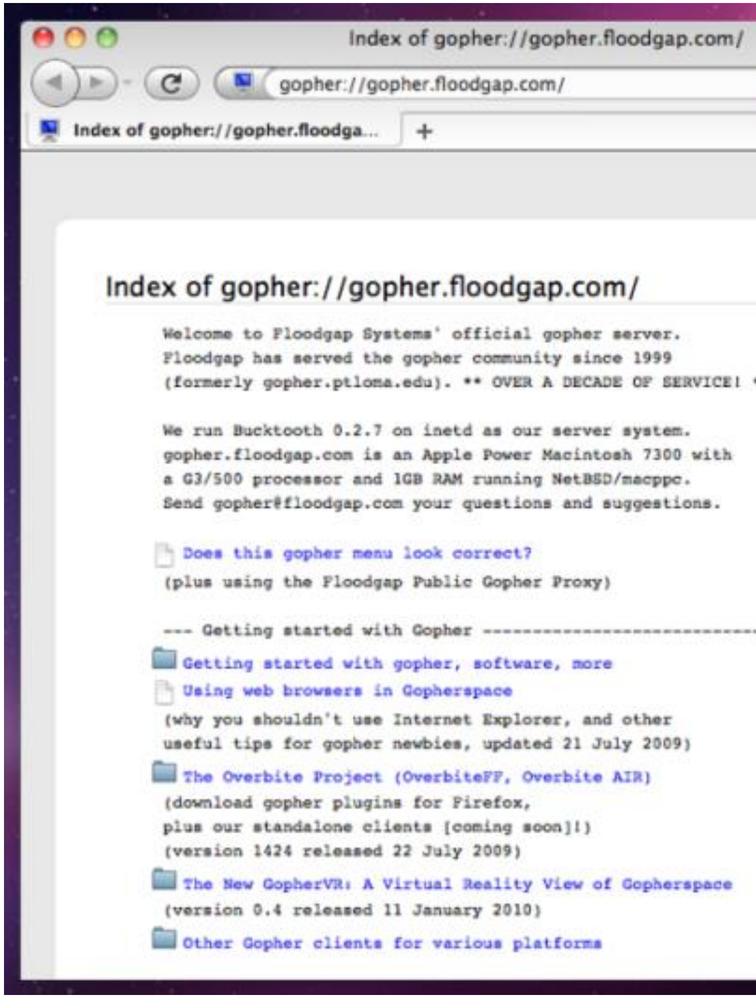
"Advances in ICT

- …are a key driver of economic competitiveness
- … are crucial to achieving our priorities in energy and transportation, education and life-long learning, healthcare, and national security
- … accelerate the pace of discovery in nearly all other fields
- …are essential to achieving the goals of open government"

US President's Council of Advisors on Science and Technology, 2010

Web 1.0: Information revolution

- Rise of the personal computer (PC) 1970s
- FTP, telnet, gopher – menu based
- World Wide Web (Berners-Lee @ CERN 1990). NCSA Mosaic → "static" webpages
- Battle of the browsers – Mozilla → Netscape → Firefox vs. Internet Explorer, now Google Chrome
- Search engines – Archie (1990), Excite (Stanford undergrads 1993), Yahoo Directory (fav websites of Jerry Yang 1994), Lycos (Carnegie-Mellon 1994), Overture (paid search 1998)
- **Google**
  - Founded in 1998; IPO 2004
  - crawl → index → ranking: number and quality of links to a site


Index of gopher://gopher.floodgap.com/
gopher://gopher.floodgap.com/
Index of gopher://gopher.floodga...
+

Index of gopher://gopher.floodgap.com/

Welcome to Floodgap Systems' official gopher server.
Floodgap has served the gopher community since 1999
(formerly gopher.ptloma.edu). ** OVER A DECADE OF SERVICE!

We run Bucktooth 0.2.7 on inetd as our server system.
gopher.floodgap.com is an Apple Power Macintosh 7300 with
a G3/500 processor and 1GB RAM running NetBSD/macppc.
Send gopher@floodgap.com your questions and suggestions.

Does this gopher menu look correct?
(plus using the Floodgap Public Gopher Proxy)

--- Getting started with Gopher ---------------------------

Getting started with gopher, software, more

Using web browsers in Gopherspace
(why you shouldn't use Internet Explorer, and other
useful tips for gopher newbies, updated 21 July 2009)

The Overbite Project (OverbiteFF, Overbite AIR)
(download gopher plugins for Firefox,
plus our standalone clients [coming soon]!)
(version 1424 released 22 July 2009)

The New GopherVR: A Virtual Reality View of Gopherspace
(version 0.4 released 11 January 2010)

Other Gopher clients for various platforms

Web 2.0: Age of mass collaboration

Social production:
Wikipedia, Flickr,

Open Source Software:
Linux, Apache,
Firefox,…

Creative expression:
YouTube, AcidPlanet,
Kompoz,…

Social networking: Twitter,
Facebook, MySpace,
Orkut, LinkedIn

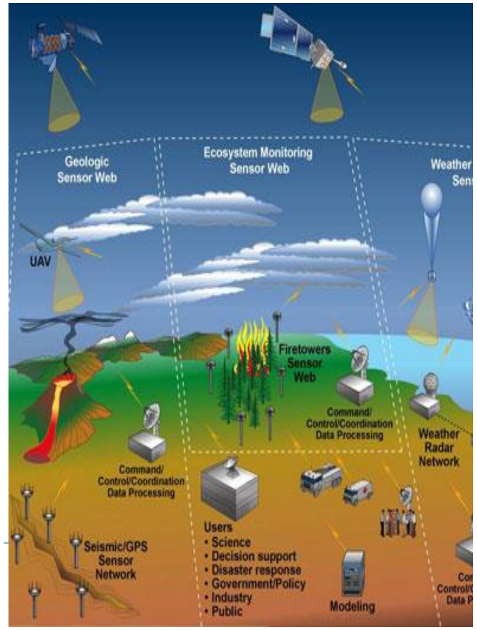
Wisdom of the crowds:
InnoCentive etc.

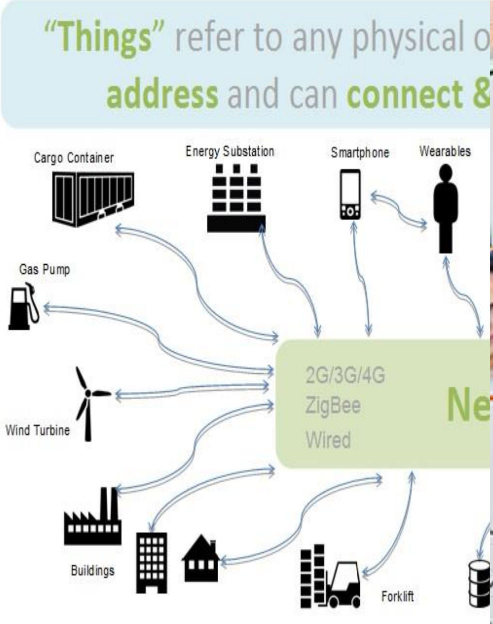
Massively Multiplayer Online Games:
World of Warcraft, Runescape, Everquest…

Virtual worlds (Second Life, Club
Penguin…)

Web 3.0?

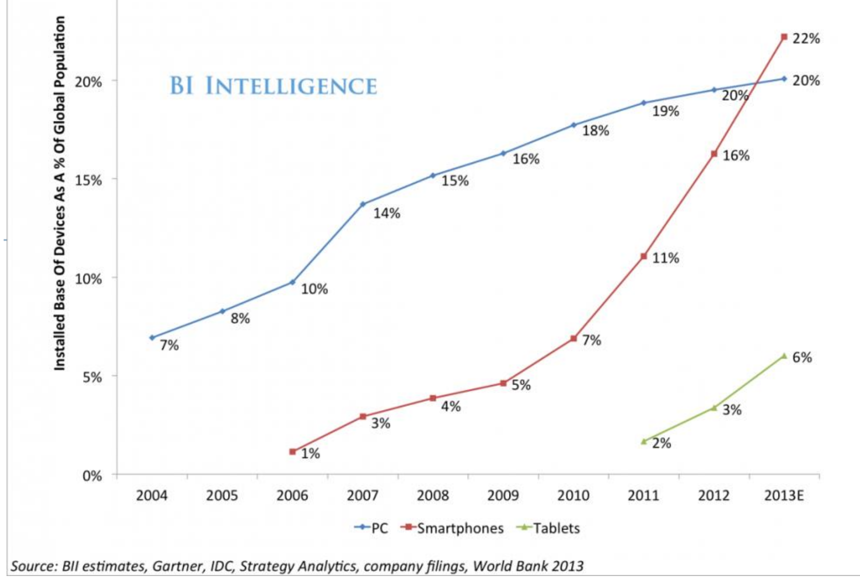
- Capacity to capture and store information growing exponentially
- Sensor networks, social networks, admin data, health records
- Boon for social science… and business innovation?





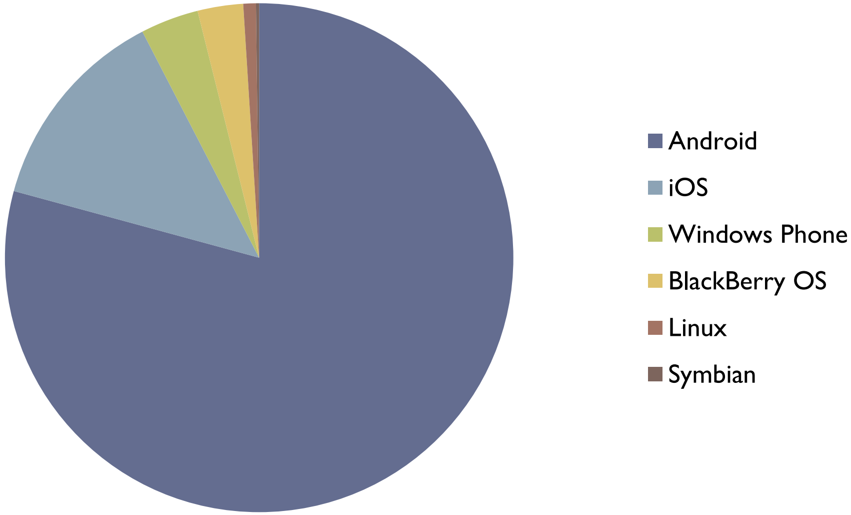
Mobile web

- Smartphone adoption
- App'ification
- Platform competition



Source: BII estimates, Gartner, IDC, Strategy Analytics, company filings, World Bank 2013

2Q13 Market Share

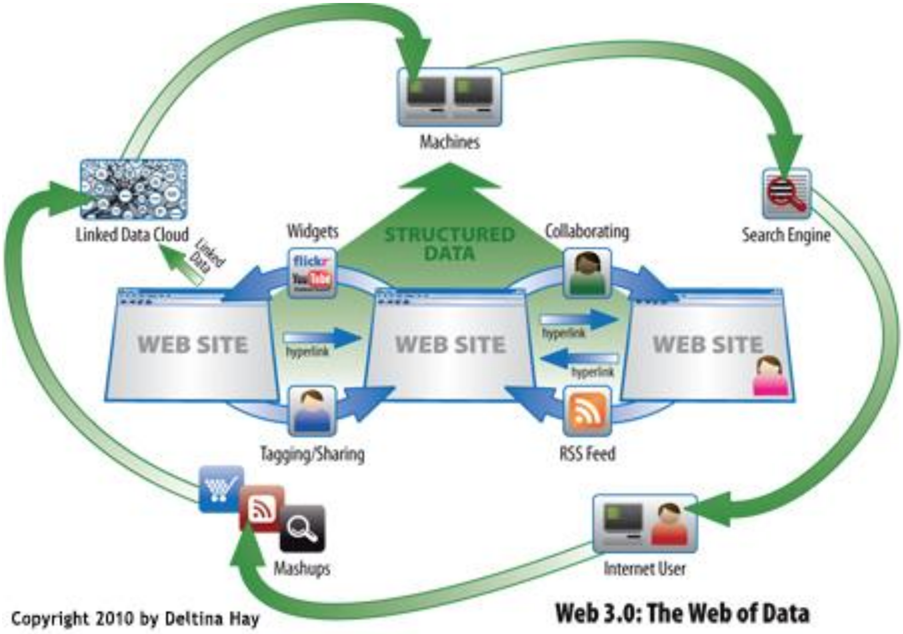


2Q13	Unit Shipments	Market Share	Y-o-Y Change
Android	187.4	79.30%	73.50%
iOS	31.2	13.20%	20.00%
Windows	8.7	3.70%	77.60%
BlackBerry	6.8	2.90%	-11.70%
Linux	1.8	0.80%	-35.70%
Symbian	0.5	0.20%	-92.30%

Semantic Web

- Wikipedia: ”a collaborative movement led by international standards body the World Wide Web Consortium(W3C). The standard promotes common data formats on the World Wide Web. By encouraging the inclusion of semantic content in web pages, the Semantic Web aims at converting the current web, dominated by unstructured and semi-structured documents into a "web of data".

- WWWxBig Data?

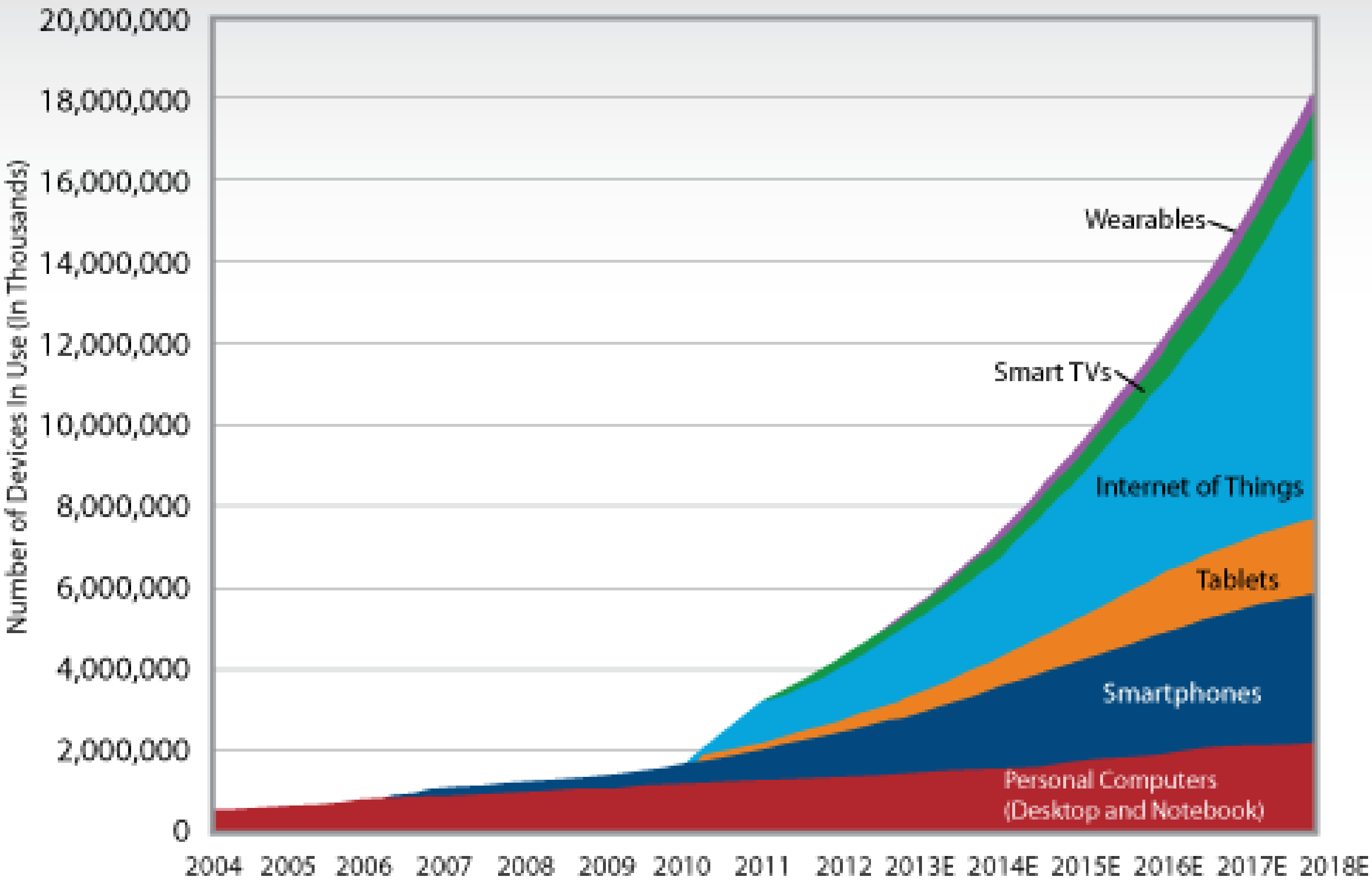


Copyright 2010 by Deltina Hay

Web 3.0: The Web of Data

Everything's Connected...

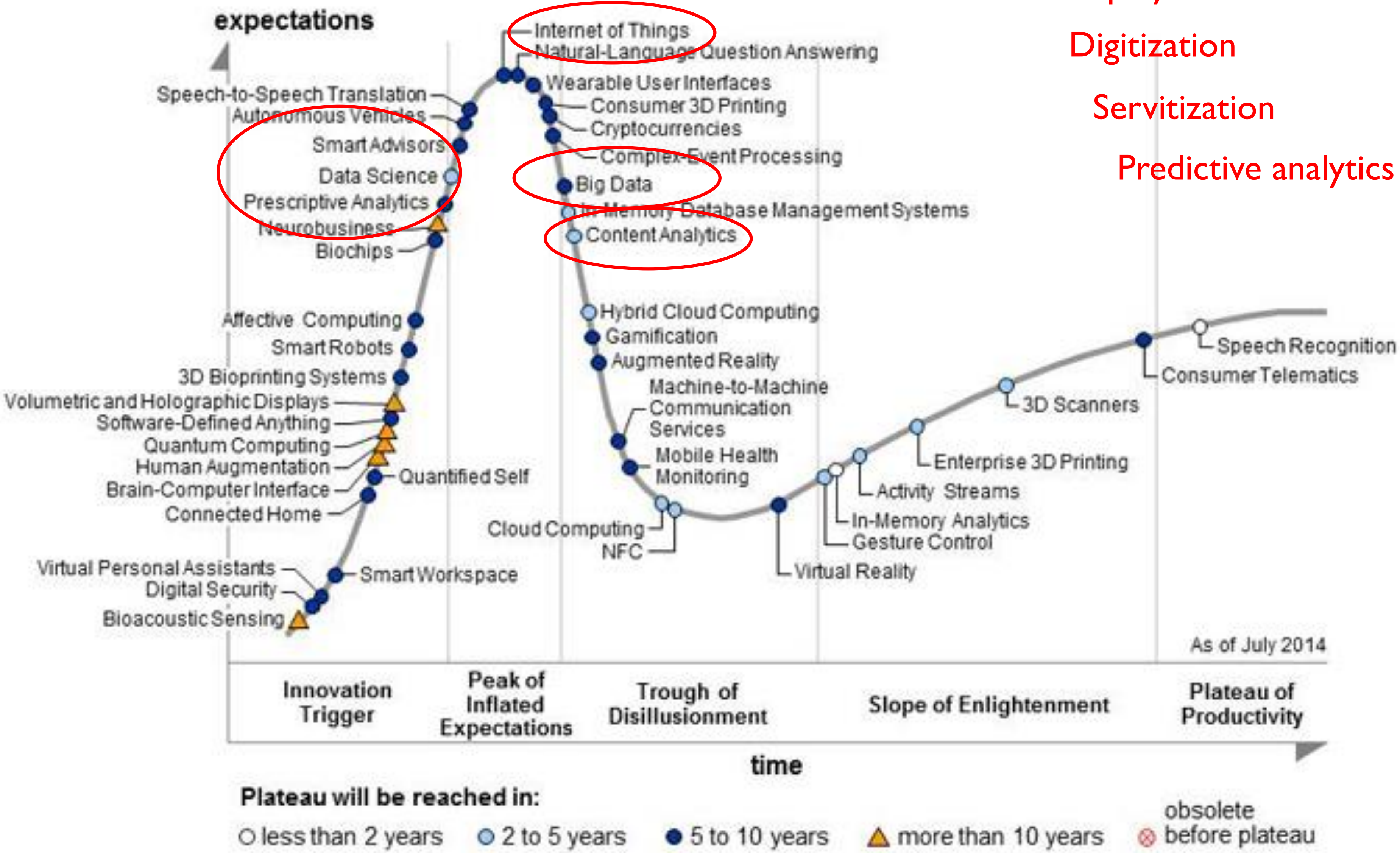
Global Internet Device Installed Base Forecast



Source: *Business Insider*

WWW.DAILYRECKONING.COM

Gartner: Emerging Technologies Hype Cycle 2014



Ubiquity

Digitization

Servitization

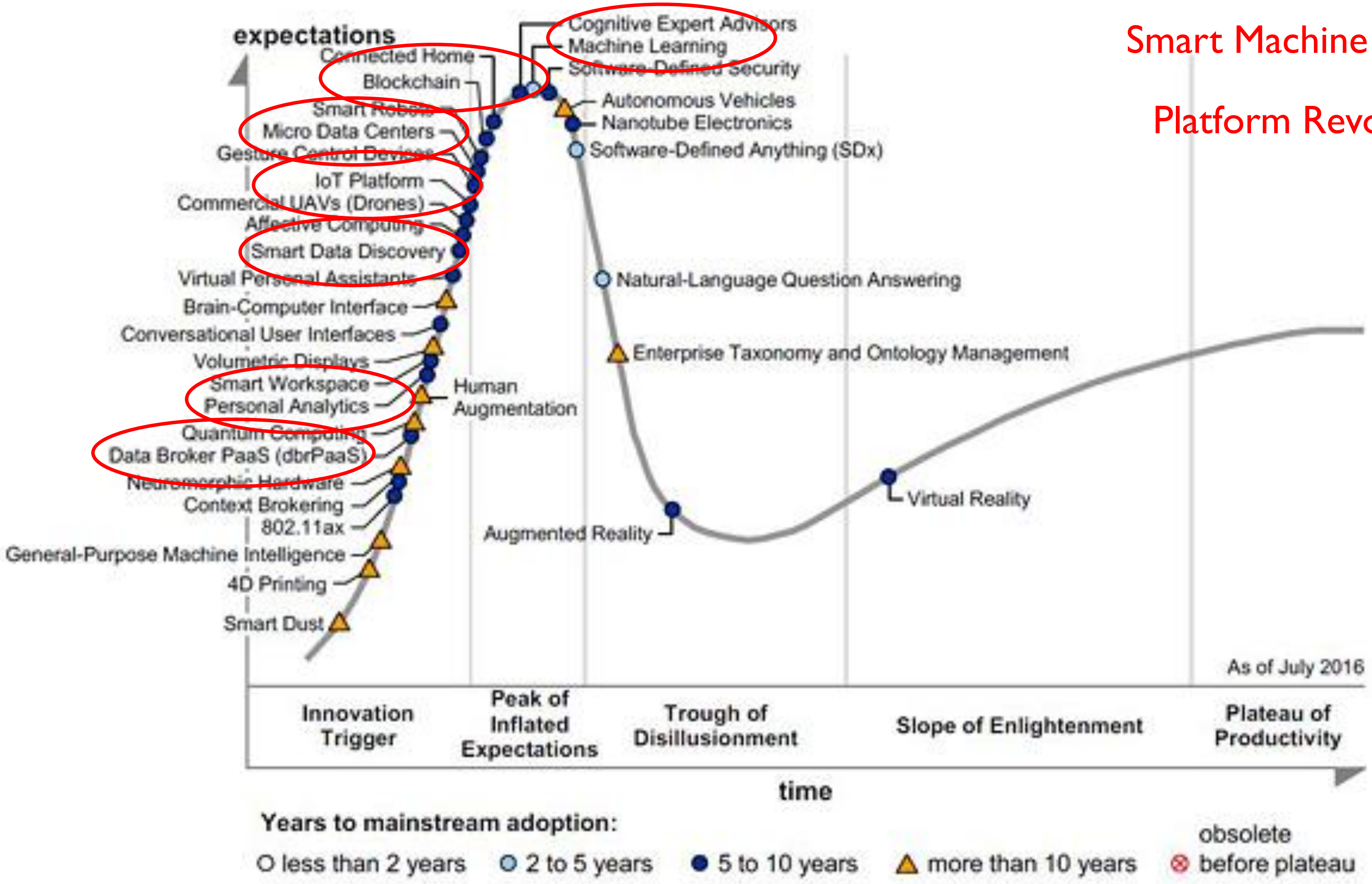
Predictive analytics

Hype cycle 2016

Immersive Experiences

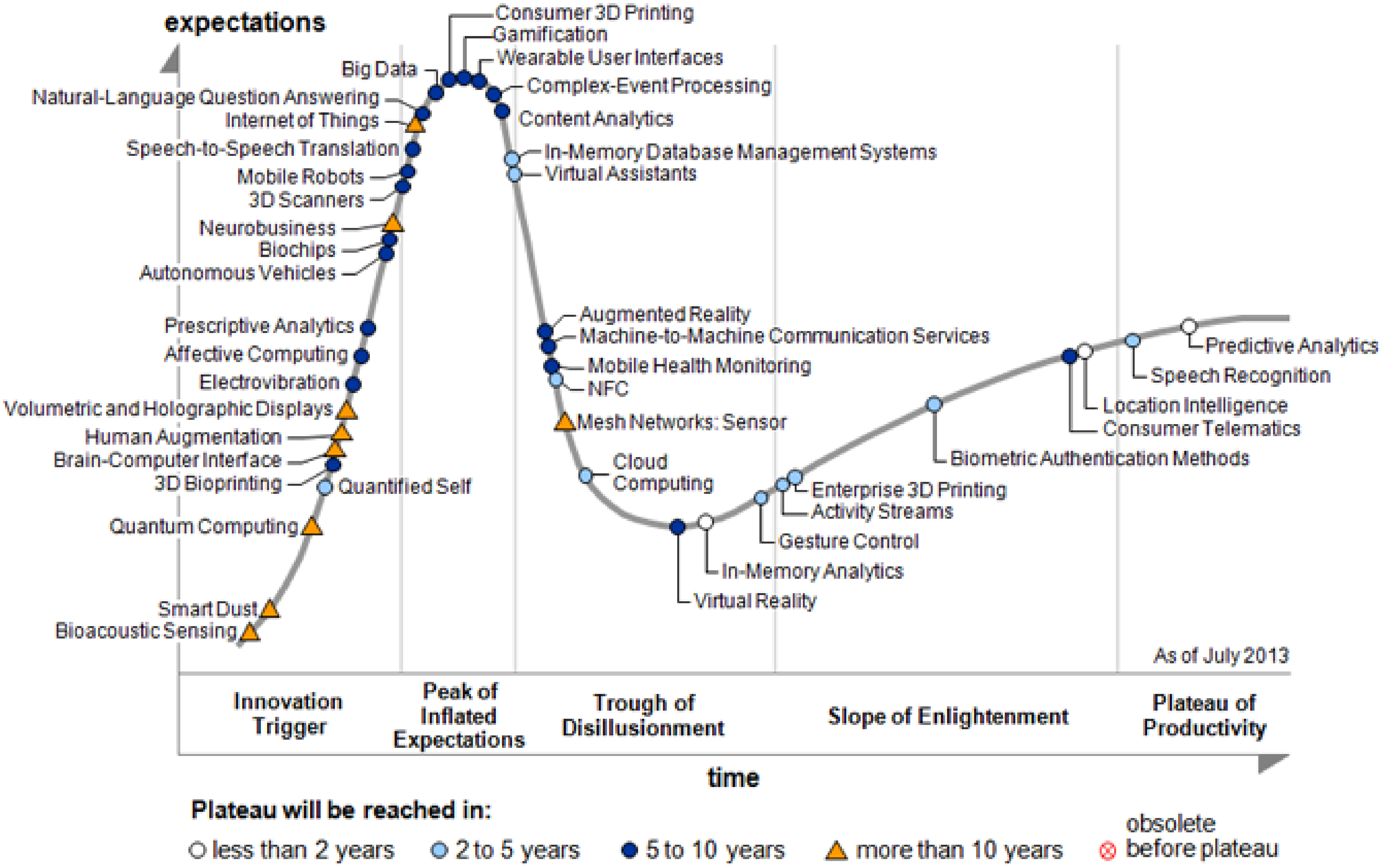
Smart Machine Age

Platform Revolution



Source: Gartner (July 2016)

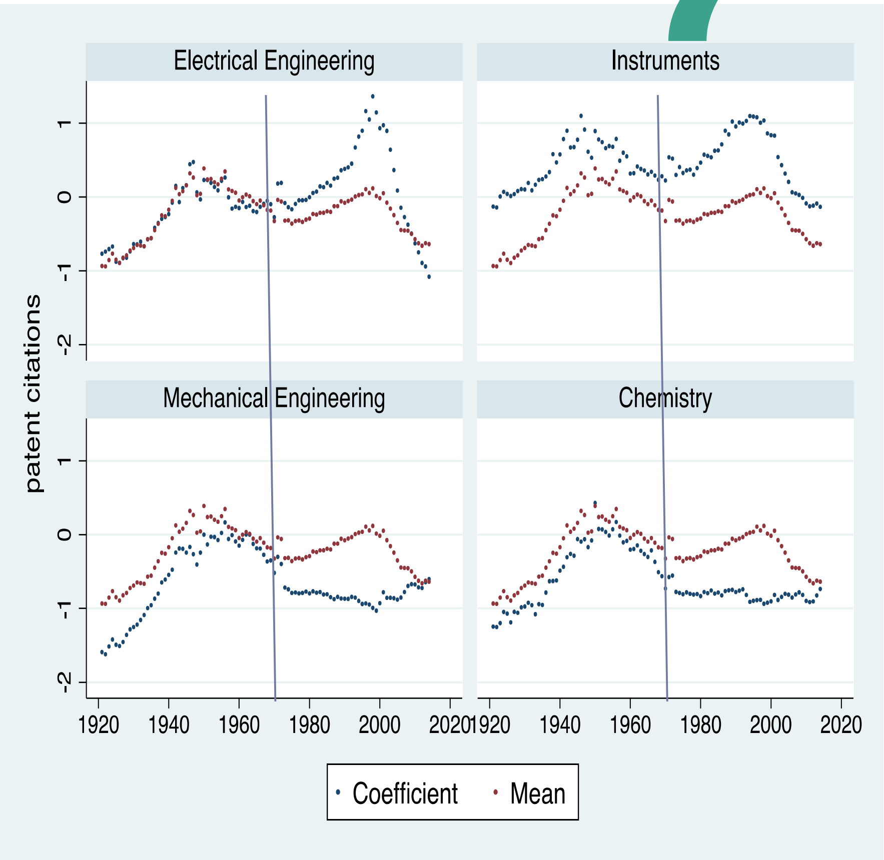
Gartner: Emerging Technologies Hype Cycle 2013

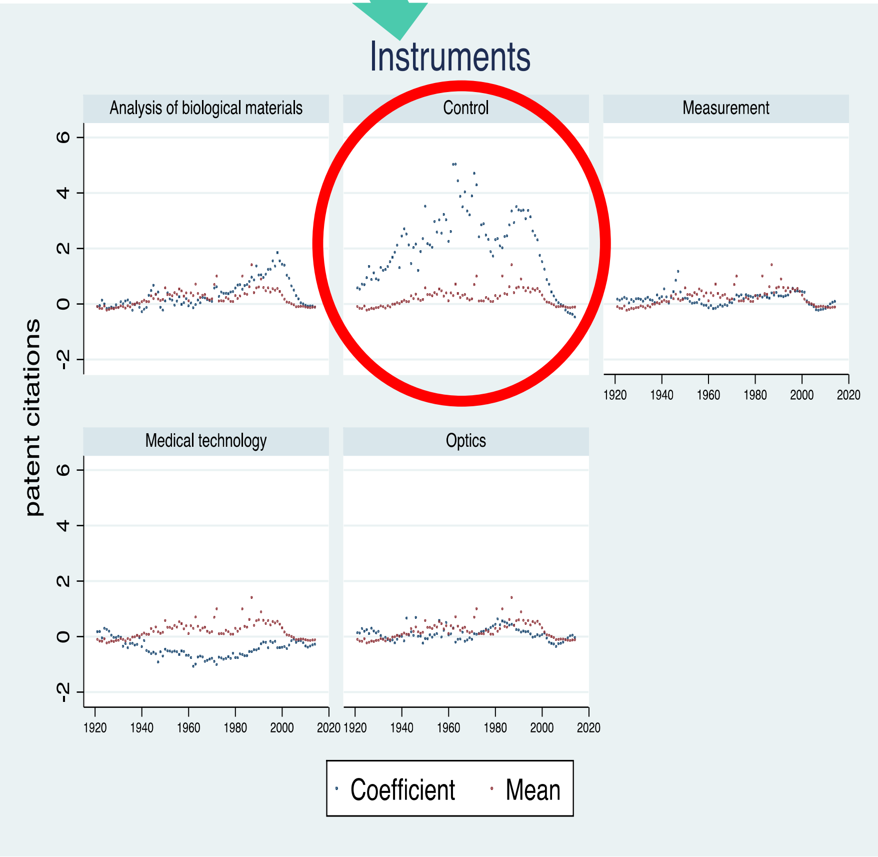


QUIZ: What are the most influential hardware technologies of the 21st century?

- Measured by other technologies *citing* the technology class in patent applications
  - “Prior art citations”
  - Indicate that the new invention is building on the earlier invention; its novelty is delimited by the earlier invention
- Is it chemical engineering, mechanical engineering, electrical engineering, or instruments?
- What kinds of instruments: Measurement, optics, control, medical, or biological?
- Why?

Predicted citations (sector/field X year) (100 years, 160 countries, 90M patents, 160M citations)





Red dots = mean of all sectors

Blue dots = coefficient for sector in question

Count-data model of citations/patent

$$C_i=\beta_{kt}F_{kt}+\gamma_iX_\iota+\varepsilon_i$$

C_i is the sum of all citations received by patent *i*,

F_{kt} is an indicator for patents that belong to field *k* and were published in year *t*.

β_{kt} captures the number of citations at the **field-year** level (34 fields)

controls are included in X_i, the vector of patent characteristics

analysis at the patent level for max degree of flexibility without aggregating inventions at the patent family or extended patent family levels

the size of the dataset exceeds common computing capacities. Most of the analysis took place in the AWS using r3.8xlarge memory optimized instances (244GiB)

Controls

Publication authority and year effects

Family (and extended family) size controls

Total number of inventions granted (annual patent flows)

Total number of citations within each sector each year

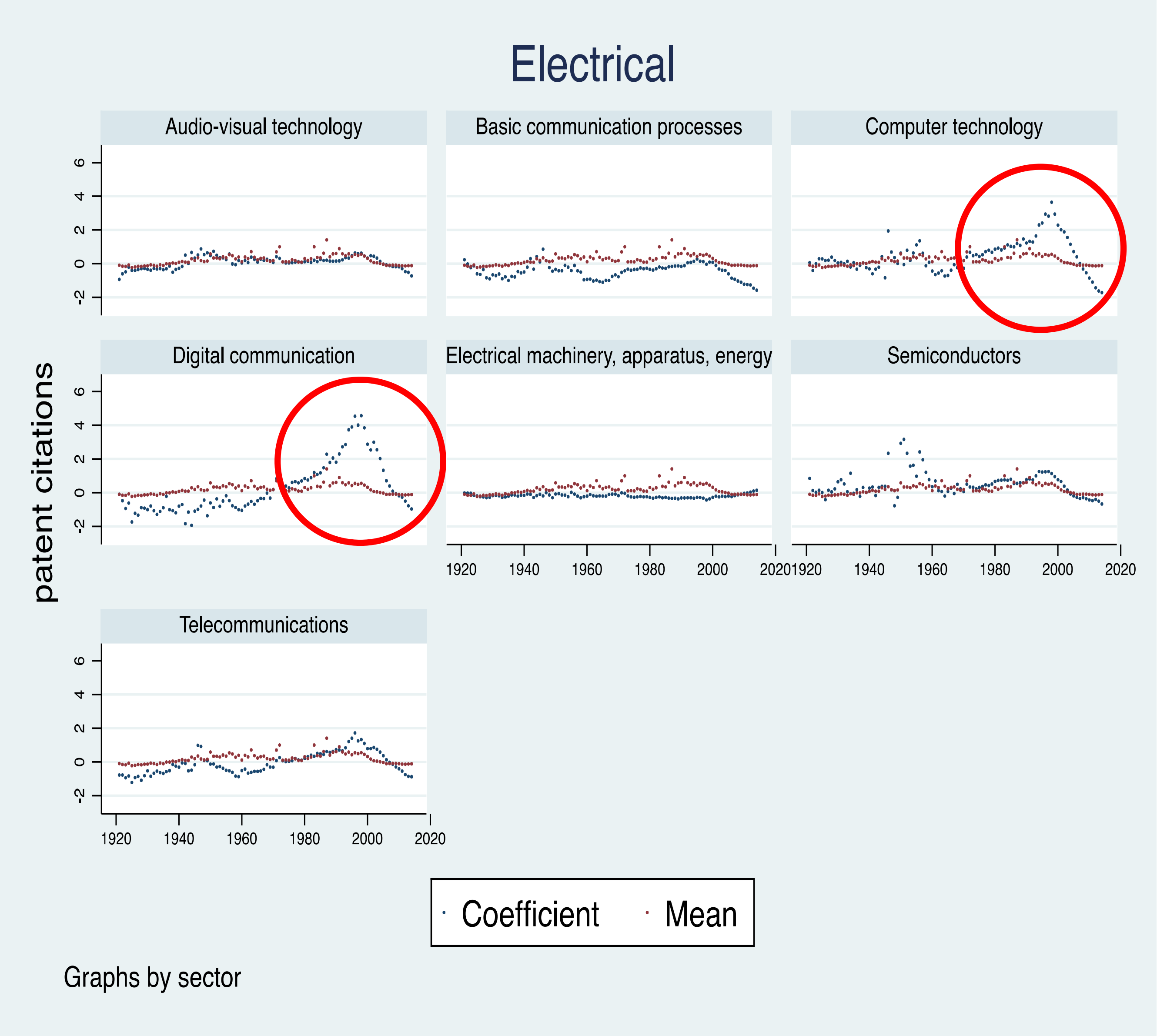
Examiner citations

Number of claims

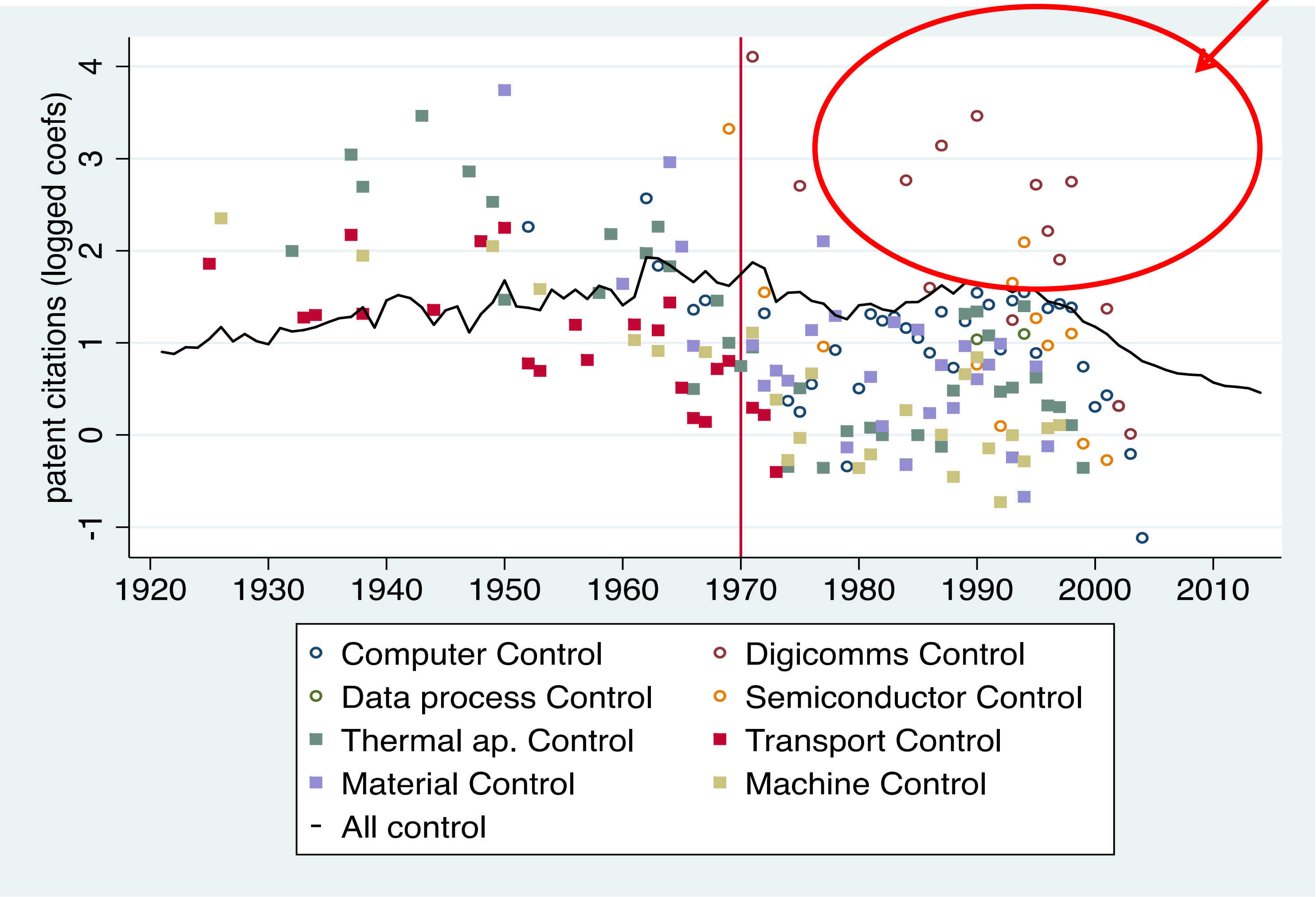
Month of publication

Assuming that ***patents granted by a patent office, at the same time, within the same field, and with the same family size will be treated equally.***

Predicted electrical engineering citations



Predicted control patent citations with co-listed technology fields



Patents listed in both digital comm & control

Why?

Invention machines: *Applicable in many sectors; Facilitate invention in other sectors; A broad and catalytic impact by enabling follow-on invention in many application sectors; Generate massive knowledge spillovers over long periods of time*

- Control instruments
- Digital communication
- Computer technologies

Internet of Things

Instruments enable manipulation of material; computers enable manipulation of information

Automation requires instrumentation

Web 3.0

Control instruments – sensors, indicators, logic devices, actuators

+

Data – social, administrative, industrial, personal

+

Artificial Intelligence – algorithms, machine learning, prescriptive analytics

=

"Second Wave of the Second Machine Age"

(Erik Brynjolfsson/MIT)

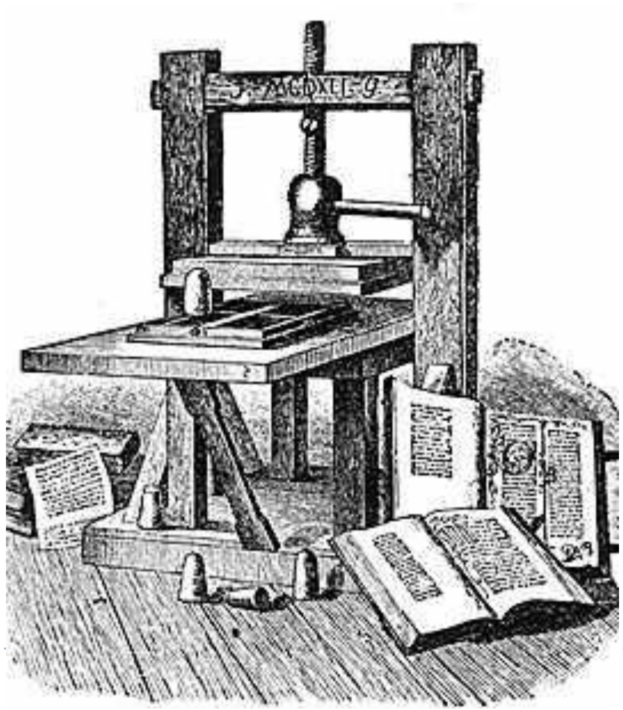
Earlier communication revolutions

- Printing press
- Steam engine
- Telegraph
- Telephone
- Radio
- Television

- Electrification

Printing press

- Johannes Gutenberg (Germany 1452)
- Reading became accessible to common people
- Fiction, entertainment, propaganda
- Mass education
- Network externalities:
  availability of books →
  incentives to *learn to read* →
  demand for books



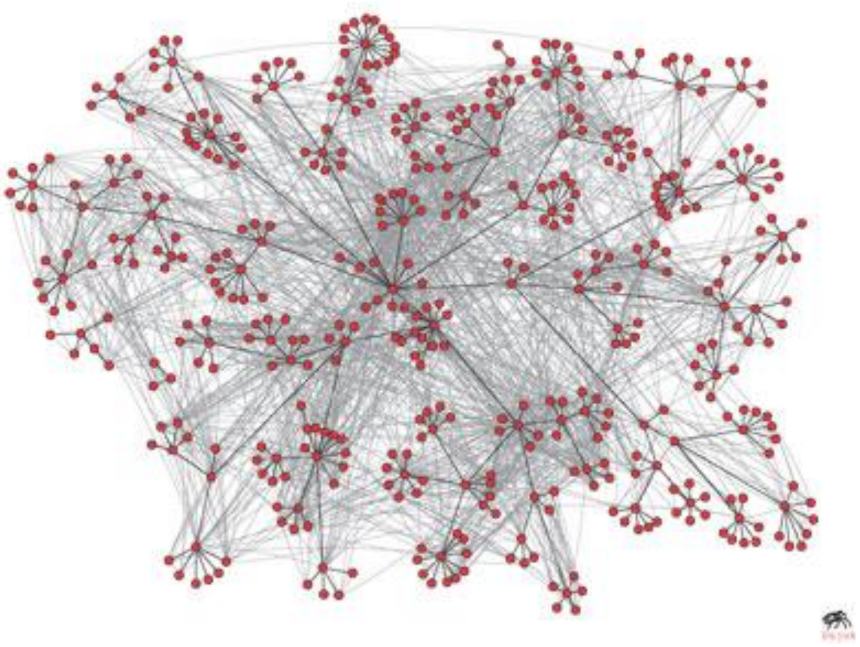
Impact of the printing press

- 30 years later a printing shop in Florence run by nuns charged 3 florins for 1000 copies of Plato's Dialogues, while a scribe would have charged 1 florin for 1 copy
- Availability of paper from China → prices fell, demand increased
- #books produced in 50 years following the invention = #books produced by European scribes in preceding 1000 years!
- Fust was suspected in Paris to be in league with the devil – fear of novelty

➔ How did the printing press change society, lifestyles, economy?

Did the Internet *fundamentally* change the way we…

- learn?
- socialize?
- shop?
- participate?
- engage?
- work?
- share?



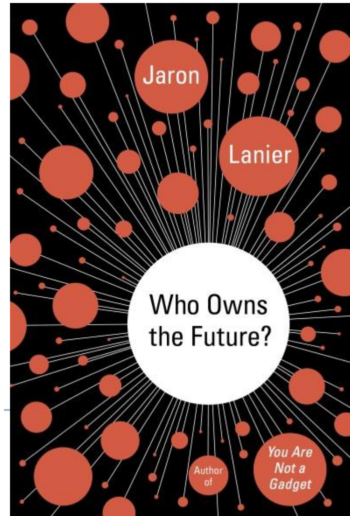
For the better or for the worse…?

Expect societal changes due to Web 3.0

- Privacy needs to be defined
  - Ownership of data
  - Right to be forgotten – in/alienability
- Intellectual property for data
  - Data security
  - Legal framework
- Radical transparency
  - Real-time visibility
  - Data integration, inference, prediction
- New business models, new platforms, new winners

➔ Who owns the future?

- We can’t expect to have everything for free online
- Currently we get it “free” in exchange for our personal information, attention
  - Huge industry brokering, spying, analyzing your personal data
- Maybe personal data should be explicitly traded
- Challenge: data governance – there is no IPR for data!
- J. Lanier



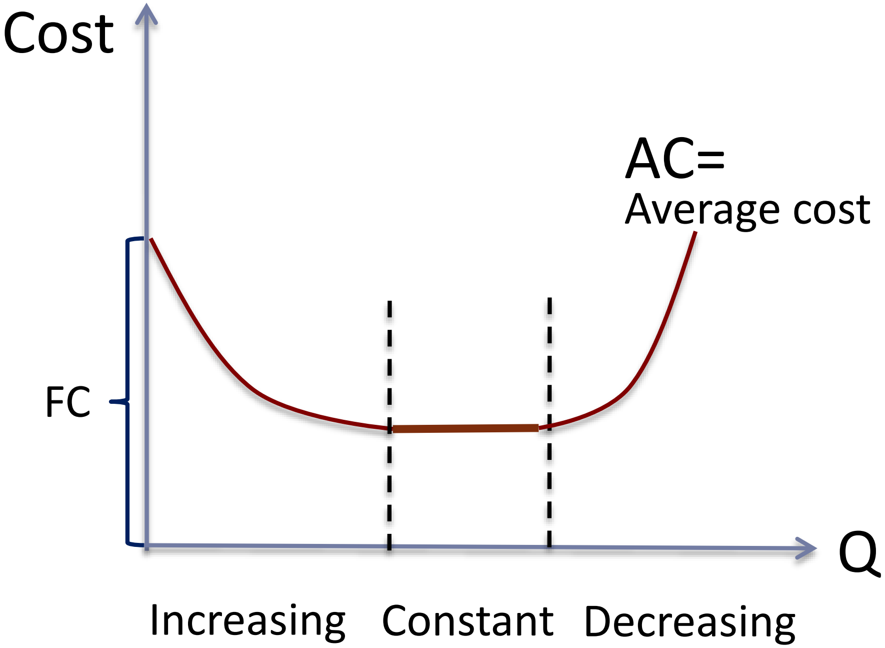
Economic characteristics of information (Romer 1990 etc)

1. Non-rival

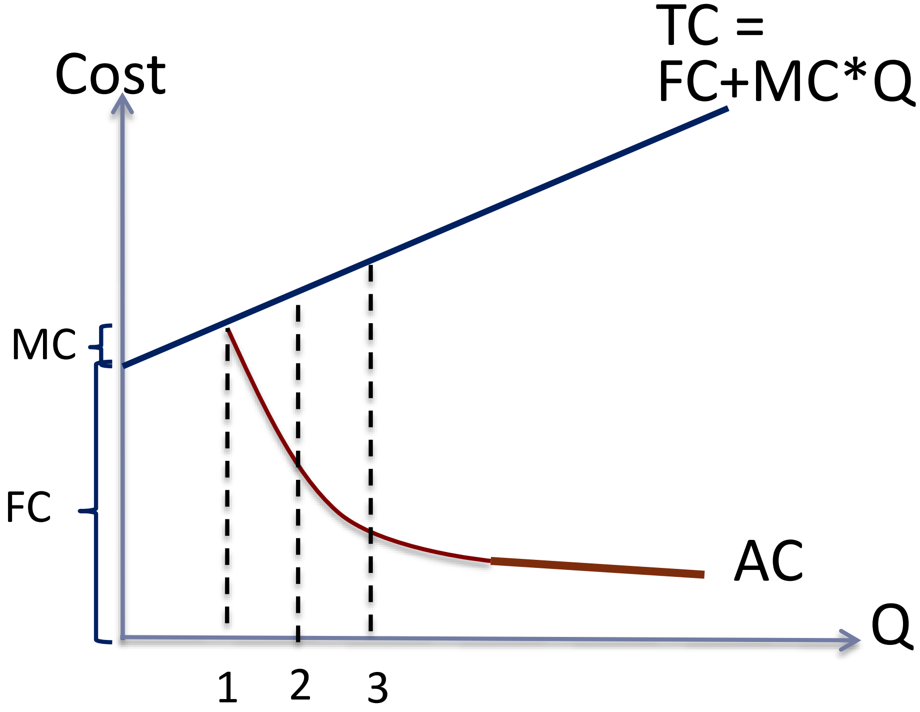
- Shareable
  - Use by one doesn’t preclude use by others
- Increasing returns to scale
  - High fixed (sunk) costs, low marginal costs → once created, easy to copy and distribute
- Pricing: How to recover the fixed cost?
  - P ≠ MC → 0

Increasing Returns to Scale

Average cost curve: large-scale production (e.g. steel)



Information goods: constant (or 0) marginal cost

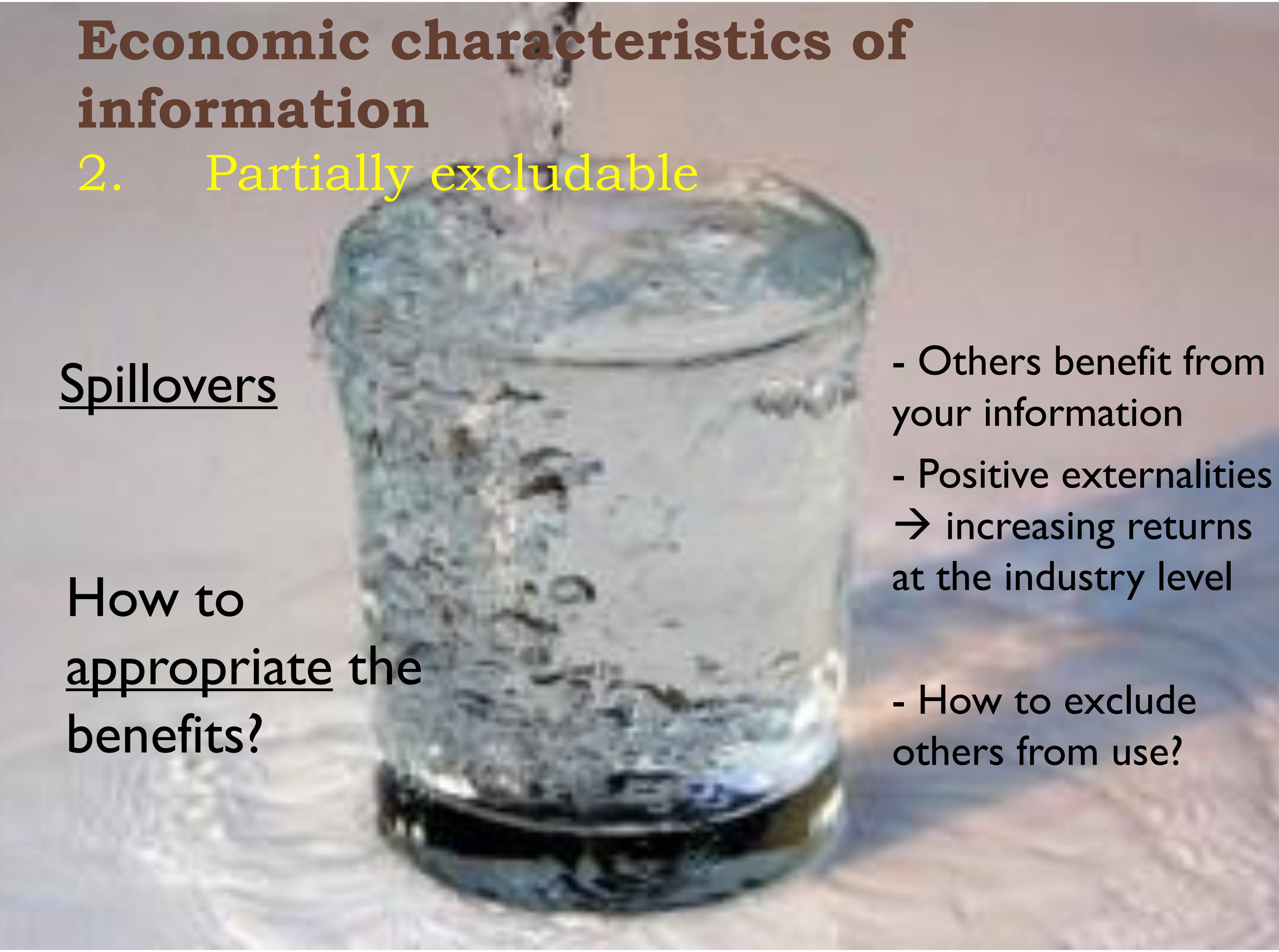


AC= Average cost = Total Cost/Quantity

IRS: doubling the inputs MORE than doubles the output!

Economic characteristics of information

2. Partially excludable



Spillovers

- Others benefit from your information
- Positive externalities → increasing returns at the industry level

How to appropriate the benefits?

- How to exclude others from use?

Appropriation mechanisms

- Intellectual Property Rights
  - Patent
  - Copyright
  - Trademark
  - Database right
- Secrecy
  - Trade secret
- Contracts
  - Employees, business partners

Economic characteristics of information

3. Experience good

- What is an experience good?
- What problems does it create?
  - Arrow's paradox
- Solutions?

Some solutions to the experience good problem

- Branding
- Reputation
- Sampling
- Recommendation
- Contracts
  - Guarantees
  - Warranties
  - Incentives

Are data information goods?

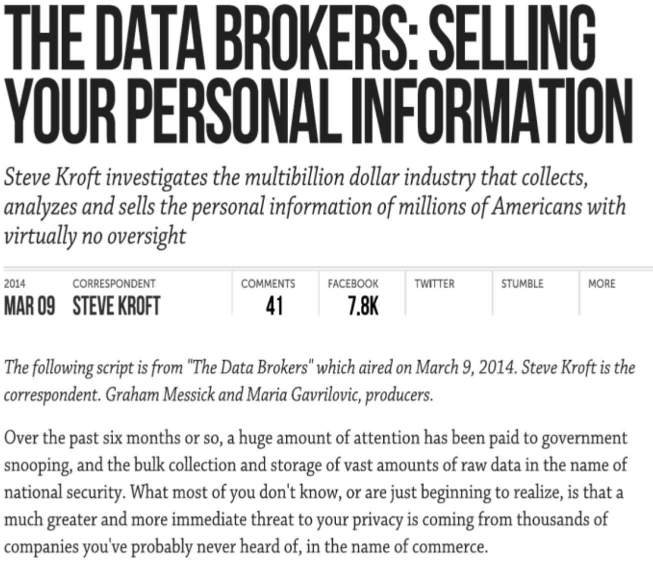
Nonrival?	Yes
Partially excludable?	NOT excludable
Experience good?	Yes if no metadata No if proper metadata
High fixed cost/low or constant marginal cost?	Varies: exhaust data vs. data collected for a purpose

Data Governance

Aija Leiponen
Cornell University

PUZZLE: How can data be commercially exploited?

- Data are not intellectual property
  - Individual data points have **no** legal protection
- Essentially needs to be controlled contractually (secrecy, organization forms, product design, non-compete and confidentiality contracts),
  - Not via intellectual property rights
- How can something so "leaky" be valuable, commercialized?



THE DATA BROKERS: SELLING YOUR PERSONAL INFORMATION

Steve Kroft investigates the multibillion dollar industry that collects, analyzes and sells the personal information of millions of Americans with virtually no oversight

2014 MAR 09 | CORRESPONDENT STEVE KROFT | COMMENTS 41 | FACEBOOK 7.8K | TWITTER | STUMBLE | MORE

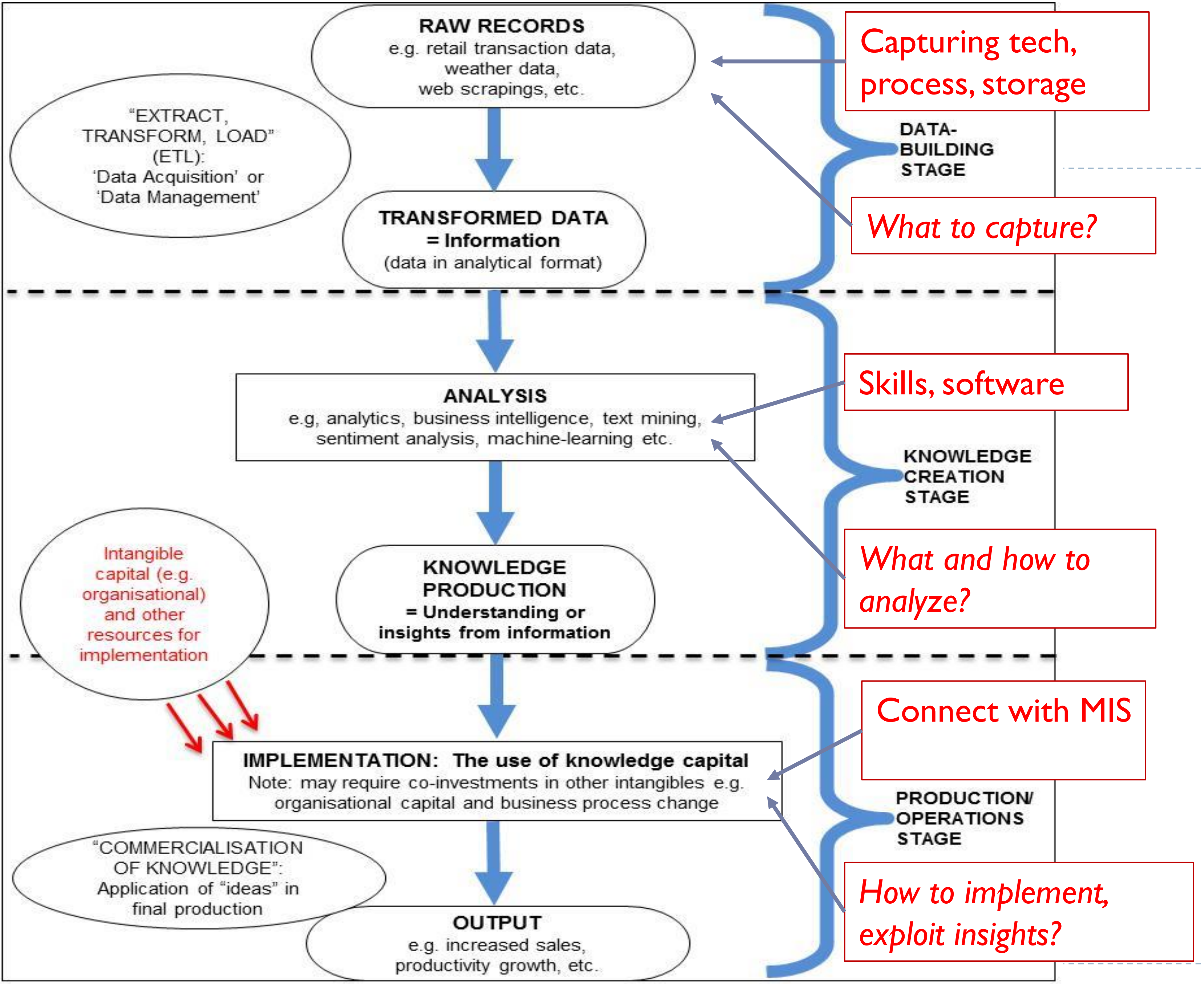
The following script is from "The Data Brokers" which aired on March 9, 2014. Steve Kroft is the correspondent. Graham Messick and Maria Gavrilovic, producers.

Over the past six months or so, a huge amount of attention has been paid to government snooping, and the bulk collection and storage of vast amounts of raw data in the name of national security. What most of you don't know, or are just beginning to realize, is that a much greater and more immediate threat to your privacy is coming from thousands of companies you've probably never heard of, in the name of commerce.

I COLLECTED OPTIMISTIC DATA, PUT IT IN THE CONTEXT OF BAD ANALOGIES, SEASONED IT WITH SALIENCY BIAS. . .
. . . ADDED HERD INSTINCT, A PINCH OF CONFIRMATION BIAS. . . AND HERE'S YOUR STRATEGY.
JUST ADD LEADER-SHIP.
WHY DO I ALWAYS GET THE HARD PART?

Dilbert.com DilbertCartoonist@gmail.com

5-21-10 © 2010 Scott Adams, Inc./Dist. by UFS, Inc.


RAW RECORDS
e.g. retail transaction data, weather data, web scrapings, etc.
"EXTRACT, TRANSFORM, LOAD" (ETL): 'Data Acquisition' or 'Data Management'
TRANSFORMED DATA = Information (data in analytical format)
DATA-BUILDING STAGE
Capturing tech, process, storage
What to capture?
ANALYSIS
e.g, analytics, business intelligence, text mining, sentiment analysis, machine-learning etc.
Skills, software
KNOWLEDGE CREATION STAGE
What and how to analyze?
KNOWLEDGE PRODUCTION = Understanding or insights from information
Intangible capital (e.g. organisational) and other resources for implementation
IMPLEMENTATION: The use of knowledge capital
Note: may require co-investments in other intangibles e.g. organisational capital and business process change
Connect with MIS
PRODUCTION/ OPERATIONS STAGE
How to implement, exploit insights?
"COMMERCIALISATION OF KNOWLEDGE": Application of "ideas" in final production
OUTPUT
e.g. increased sales, productivity growth, etc.

Peter Goodridge & Jonathan Haskel (2015)

What are economic properties of data?

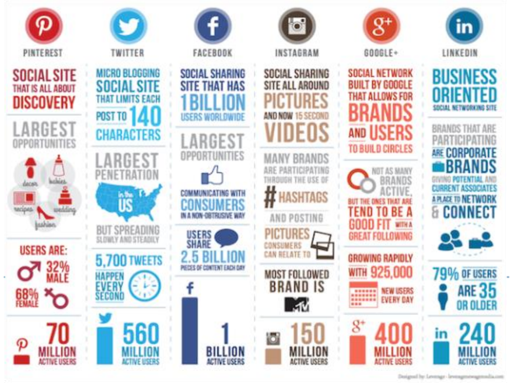
- **Data as records** of actions, events and situations
- **Intangible information** good – non-rival and mostly non-excludable.
  - High fixed cost of *structuring*; low marginal cost of sharing
- **Intermediate & final** good
- Information and insights sometimes created **cumulatively** in long chains of merging and analyzing records – **Data Supply Chains**
- Who **owns** the combined, analyzed data?
  - **Provenance** is hard to track.
  - Individuals have some **claims**.
  - **Contractual** data partners have other claims.

Money vs. Data

- Data is viewed as the "new oil", an asset class
- Digital currency is data on a fundamental level – streams of bits
- Currencies rely on **trust** in the medium – data have **intrinsic value**.
- Increasing **subjectivity** of data goods as we go from raw to tagging/cleaning, aggregating, combining, processing
- **Provenance** is hard to prove for data, currencies are verifiable
- **Non-exchangeability** of data – there is no quantum of data with a minimum value
- Nevertheless CS researchers starting to consider "data as money"; developing conceptual models of a "central bank for data"

Content vs. Data

- Both have (some) **intrinsic value**
- Both governed by **copyright**
- On a fundamental level, **content IS data** and subject to analytics (Natural Language Processing)
- But the value of record data largely comes from combination with other data and algorithms (models, statistics, prediction, deep learning…)
- And as a result, copyright is very weak on data



Economic features of digital goods – all controversial and legally contested

	Record Data	Content	Software	Currency
Information Type	Raw records or structured databases	Knowledge (insights)	Knowledge (instructions)	Pure value
Good Type	Intermediate/ Final	Final	Final	Final
Alienability	Variable	Medium	High	High
Inferability	High	Low	Low	Zero
Excludability	None	Variable	Variable	High
Fungibility	Variable	Low	Low	High
Protection Method	Secrecy	Copyright	Copyright or patents in some cases	Blockchain or other verification technology
Protection Aspect	Reuse	Expression (patterns)	Expression (patterns) or insight (invention)	Transaction value

Characteristics of different data sources

Source of data	Privacy implications	Alienability	Duration/ useful life	Sampling frequency	Inferrability
Health care	High	Low (health, retail, social network, locational)	>50 years	Very low	Low
Public sector administration	Medium	Medium (public sector) – these usually have specific data protection protocols (confidential, etc)	>50 years	Low	Low
Manufacturing/ Operations (sensor networks)	Medium	Medium (manufacturing) - these usually have specific data protection protocols (confidential, etc)	10-20 years	Medium	Low
Individual behavior	High	Low (health, retail, social network)	1-5 years	High	High
Personal Locational Data	Medium	Medium	1-5 years	Very high	Medium

What have we learned?

- The economics of data goods depend on an analysis of data characteristics
  - Data are very heterogeneous
- Description, classification of data and its institutional framework is necessary for understanding its commercialization potential
- Overall, data goods substantially differ from other information goods
  - Excludability (protection)
  - Transparency (metadata)
  - Alienability (ongoing implications for individuals)
  - Inferrability (implications of data integration for individuals)

Emergence of data markets?

- Data markets will work differently in different industries
- The legal framework is evolving → data attributes
- Competitive strategies & outcomes will depend particularly on the fungibility, excludability, alienability/inferrability of the data in question
  - Business model design with determine profit potential of fungible, poorly excludable, alienable data