Big Data Will Kill Survey Research - Truth or Myth?
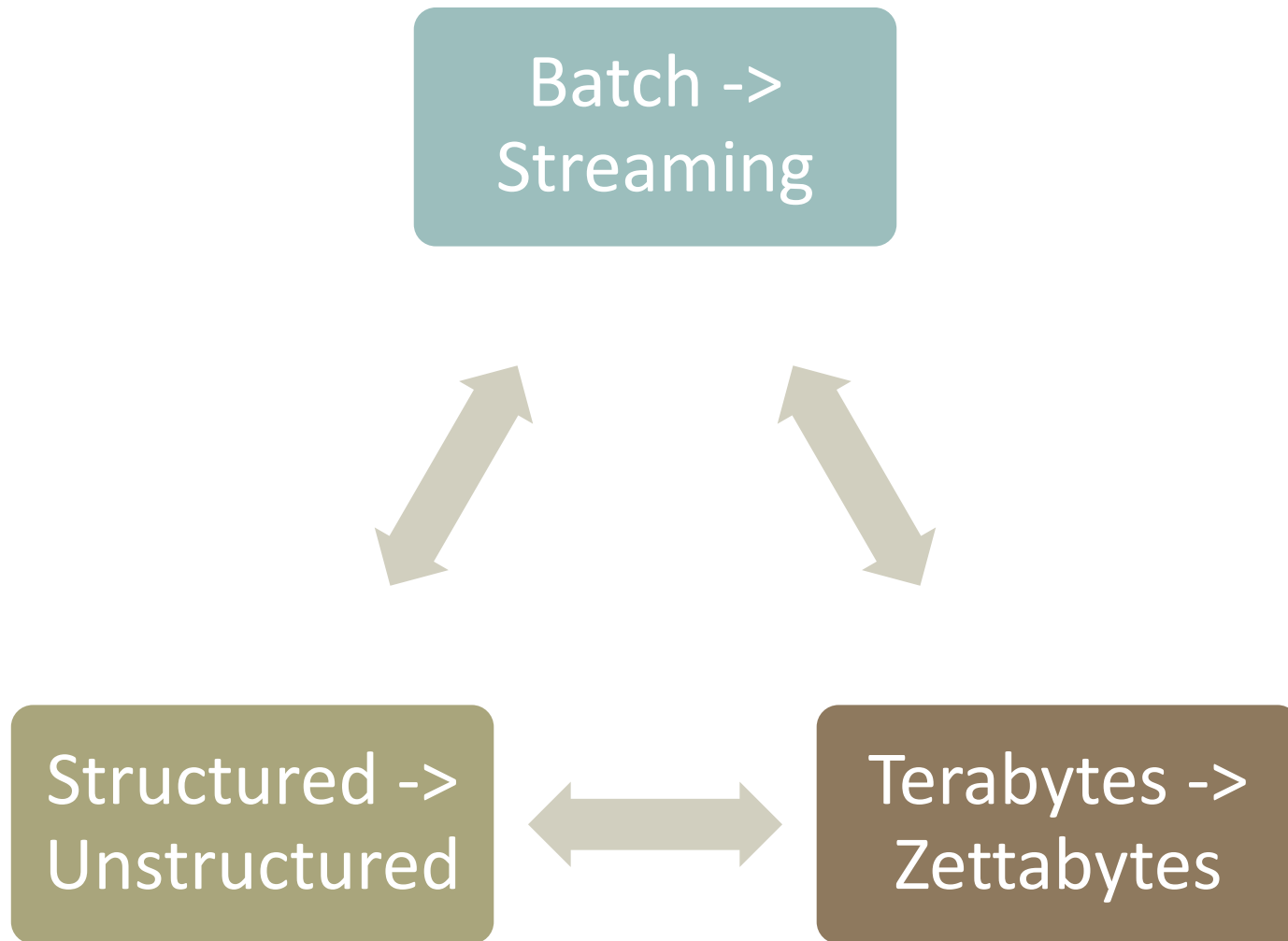
Shankar Sahai

About Me

- Have worked with data/big data for over 15 years

- Cloud technologies for over 8 years

- Not an expert on survey research

- Have a big data blog (infoivy.com) and also share on twitter

Follow @shankar_sahai   6,408 followers

Data Evolution

Batch -> Streaming

Structured -> Unstructured

Terabytes -> Zettabytes

What is big data?

- Volume – exabyte (1000^6)to Zettabyte (1000^7) range
- Velocity – real time
- Variety – multiple sources including support for multimedia
- *Bonus Vs*
  - *Veracity*
  - *Value*

Ngram results – "Friday 13"

Ngram Results– "Friday 12/13"
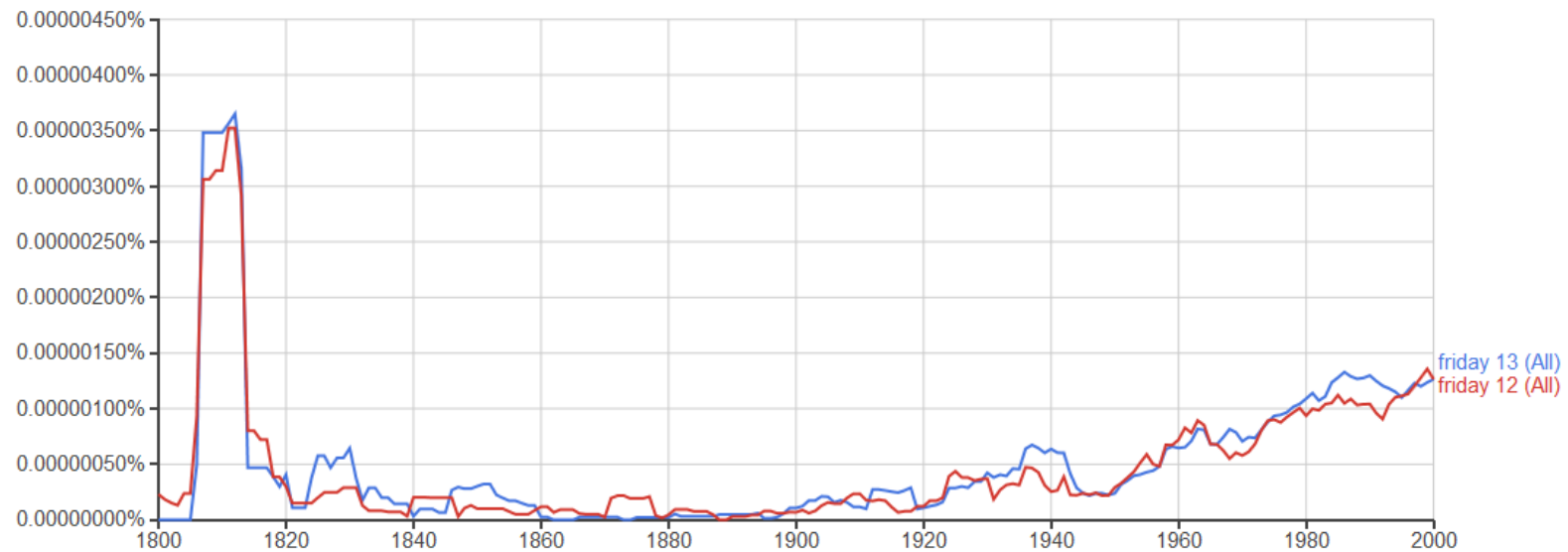
Why should you care?

- Limon & Nada Vending In Spain – How to succeed with social media (and smart vending machines)?
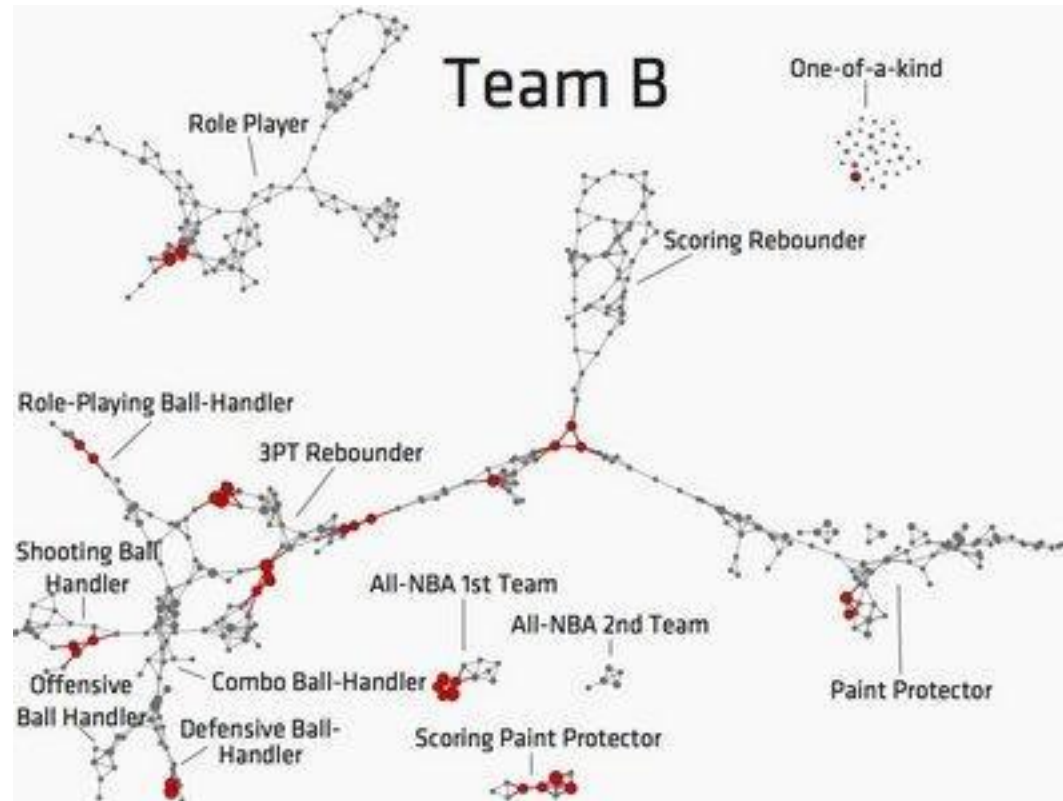
Big Data Analytics in NBA

The above shows 10 positions rather than the customary well established 5 positions.

Big data in Politics

- Obama victory

    "single massive system that could merge the information collected from pollsters, fundraisers, field workers and consumer databases as well as social-media and mobile contacts with the main Democratic voter files." - WSJ

    Rayid Ghani - chief data scientist for Obama for America 2012

Promise of Big Data

- Can help solve any problem out there?
  - World hunger?
  - Traffic accidents?
- Reasons for hype
  - Media stories
  - Greed (If Google/LinkedIn/Netflix could use it to become rich so can we)
  - Lack of understanding

Where does big data fall short? (1/2)

- Correlation implies causality - Just because there is a correlation does not mean that there is causality

- Individuals follow the group - Big data many times can provide group level summary data but that should not be treated as true at the individual level

- Unanswered "Why?" - "Big data is good at reading lines but not good at reading between the lines" - Kevin Roberts *CEO Worldwide Saatchi & Saatchi*

Where does big data fall short? (2/2)

- "Lay of the land" which requires representative samples that cannot be provided in an analysis of data

- Context Awareness – Retweets can be a good measure to prove popularity or dislike. It's just hard to know which one it is.

- Experiments – It is not easy to have meaningful data for an experimental situation that does not really exist in the real world.

Other Challenges

- Data Analysis and Inferences
  - Lack of tools (HW, SW)
- Lack of skilled, trained and experienced personnel

The Promise of Big Data

- Passive Measurement - Can result in more accurate picture since "human error" not a factor

- Cost – Much cheaper than hiring and conducting research work

- Speed – Much faster turn around time than research work

Comparison Chart

	Big Data	Surveys
Coverage of total market	Low	High
Passivity of measurement	High	Low
Granularity	High	Low
Transparency	Low	High
Managed, balanced samples	Low	High
Speed of Delivery	High	Low
Availability to all	Low	High
Data collection cost	Low	High

Big Data "Sweet Spot"

- Can shed light into usage behavior, shopping behavior, customer lifetime value (from transactional data)

- Answer the questions "how many?" and "how much?"

- Can serve as a first step in a study, which can be followed by analyses of sub-samples on a much smaller scale.

Conclusion

- Big data approach to opinion research might have issues but it is here to stay

- Researchers need to embrace the new tool

- It will *not* replace the existing tools but will supplant it

- Big data tools are improving every single year

Something to chew on…

Harvard political scientist Stephen Ansolabehere envisions a time when so much data will be available on individuals that a pollster could forgo lengthy interviews ("Age? Sex? Marital status?") and simply contact people via a social network to ask, on a given subject: "What are you thinking right now?"

- **Harvard Business Review**

THANKS