HONEYBEE
CAPITAL

DECEMBER 2013

BIG DATA ISSUE

Over this last year or two it's seemed that a Big Data publication is required of every sort of researcher. Though one of my mottos is "don't be a sheep", here we are, with the Honeybee Capital Big Data Issue. As usual, we have tried to avoid repetition and instead have highlighted a few key topics plus a few key resources in this area, aided greatly by attendance at events hosted by the Santa Fe Institute, PopTech, and NECSI, amongst others.

QUOTES OF THE MONTH:

> *We are in an age of knowledge and information, which has led to new and often anonymous kinds of power.*
> – *Pope Francis*
>
> *Big data and whole data are not the same.*
> *– danah boyd*
>
> *Despair is self-fulfilling.*
> \- *Lesley Hazelton*
>
> *We're desperate for predictability.*
> *– Cris Moore, SFI*
>
> *Economics is not an exact science; it is in fact, or ought to be, something much greater: a branch of wisdom.*
> – *E.F. Schumacher*
>
> *Data needs theory, and data needs stories.*
> – *Sean Gourley*
>
> *The presentation of data is a moral act.*
> – *Edward Tufte*

Torture the data, and it will confess to anything.

- *Ronald Coase (Nobel laureate)*

We need more pi-shaped people – the ones with deep expertise in multiple areas, with connection in-between.

- *Alex Szalay*

It is a capital mistake to theorize before one has data.

- *Sherlock Holmes*

We've got no money, so we've got to think.

- *Lord Ernest Rutherford*

Ughhh! Anything but big data.

- *Business editor, when asked about her favorite story ideas*

BIG DATA THEMES: WHAT'S SO NEW, ANYWAY?

As we've heard and read input on big data from dozens of experts, a few key themes have emerged. One of the most central questions is, what's so new, anyway? Isn't big data just, you know, more data? It's not. This is a case where the whole (or the big) is more than the some of the parts - or at least different from the previous sum. As many researchers have noted, ***big data is not just bigger. It's different. (I would also add, it's not magic pixie dust.)***

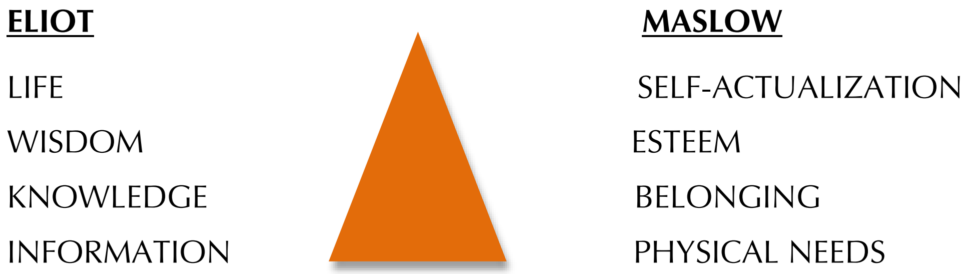
Simon DeDeo describes big data as "bigger than a person with deep expertise can hold in mind." Which leads to a mega-question, what happens when your data is bigger than your theories?

RELEVANCE FOR INVESTORS:

1. DATA VERSUS WISDOM.

This distinction is essential, as it keeps data, no matter how big, in its proper place. I have often turned to TS Eliot's "The Rock" as a sort of Maslow's hierarchy for wisdom *("Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?")*. When I was heading up research efforts, we'd always try to help new analysts move from "reporter" mode (information) to "analyst" mode (knowledge) to "investor" mode (wisdom), though of course that's never a linear exercise.

Extrapolating a bit from the poem, and simplifying a bit from Maslow, we end up with something like this:



ELIOT	MASLOW
LIFE	SELF-ACTUALIZATION
WISDOM	ESTEEM
KNOWLEDGE	BELONGING
INFORMATION	PHYSICAL NEEDS

So **where is data** in this lineup? In its unprocessed form, it's way down at the information level, maybe even lower. But the promise of big data is that by *asking the right questions*, with thoughtful and nuanced analysis, we can move up to knowledge and maybe even wisdom. But check it out - Life, that still trumps it all!

2. CORRELATION VERSUS CAUSALITY.

I know, you are rolling your eyes – really? Are we still going over this same old ground from STAT101?

[For those who don't think about these things all day long, here's a quick anecdotal refresher: One of the first serious data gatherers in retail was WalMart, and way back in the mid-1990's they were beginning to do some neat analysis. When they did the first product placements based on that analysis, some weird combinations appeared at the stores. The bananas were next to the cereal. And also next to the milk. And also next to the hammers and socks. Turns out, most households buy bananas every single week – so bananas purchases are correlated with absolutely everything else. Bananas + cereal = possible causation. Bananas + hammers = likely just correlation].

Surprisingly a lot of big data exercises completely ignore this element, and many do so deliberately, even willfully.

Chris Anderson from Wired magazine says, "Petabytes allow us to say: 'Correlation is enough….With enough data, the numbers speak for themselves.'" Ken Cukier from the Economist says, "what is more important than why." This is intriguing – what if we could skip all the pesky steps about understanding mechanisms and processes and skip right to application or intervention? But it's intriguing, not factual; to state these provocative ideas as truth is a dangerous proposition.

I actually felt queasy when I heard Cukier say, "what is more important than why," and it's not because I spent decades being trained as an experimental scientist (obviously I have not). Here are several objections to this kind of thinking:

- **It does not consider context.** Maybe it does not matter to Google why one web page is more popular than another, and maybe that's okay. But it definitely matters why one village is experiencing more disease than another – and moreover, we can probably figure that out.
- **It holds the danger of perpetuating existing, harmful biases.** There is a strong correlation between old white guys and successful senior corporate executives. I really hope that this correlation does not drive an algorithm that influences the trajectory of my own career.
- **It does not provide wisdom.** It does not even allow for the development of wisdom. For example, Cukier tells a story about data signals that alert us to be able to treat babies in distress more effectively. It is great that the data shows new signals, and even greater that some babies might be saved. But it does not absolve us of the effort to still figure out *why*. Nathan Eagle of Jana describes a seemingly similar story, where data showed that mobility declined before cholera outbreaks. But it turned out that mobility was down because of flooding; it was the water that was relevant, and movement was just a symptom. What if we could figure out why those babies were in distress in the first place? Which brings us to point #4…
- **It is not thinking at all. It lacks curiosity.** How can you not want to know? Or at least to see what you'll discover when you try to figure it out?

So, instead of saying "correlation is enough," perhaps we can say, **"big data illuminates new questions."** Or maybe even "in some contexts, understanding mechanism is less important than understanding correlation." These still jumble up and challenge the traditional scientific method (hypothesis-experiment-theory), but in a way that might be useful.

Here are links to Anderson and Cukier's writing on this topic:
http://www.wired.com/science/discoveries/magazine/16-07/pb_theory
http://amzn.to/1dy3Ghs

3. NEW CONSIDERATIONS.

- **POWER, AGENCY, AND OWNERSHIP**: What constitutes informed consent? Who decides how data is collected, used, compensated?
- **QUESTIONS:** How to develop the capability to ask better questions? And who gets to ask them?

- **INTERPRETATION & USE:** A large volume of data does not mean it's representative, or relevant to the questions at hand. There are lots of "false positives", like the data from Hurricane Sandy tweets, which disproportionately represented Manhattan versus other affected areas. Potential for discrimination and misinterpretation are gigantic.
- **INCLUSION:** There's risk of promoting a technical discourse that crowds out local participation: emphasizing technology rather than the multiplicity of approaches and perspectives (see Bellagio report, below).
- **CONTEXT:** is increasingly important. "It depends" needs to be a more acceptable & well-understood answer in all of our institutions and discussions.
- **PREDICTION VERSUS UNDERSTANDING**: As implied above, I believe that prediction models always, ultimately, fail. Without the understanding that goes with them, they are eventually useless.

PATHWAYS FORWARD:

These two frameworks provide helpful synthesis of all of the above, plus they begin to chart a constructive, thoughtful path for engagement.

1. DATA SCIENCE → DATA INTELLIGENCE (SEAN GOURLY):

Sean Gourley has described priorities in research and application as moving from artificial intelligence to augmented intelligence. He makes the distinction between data science and data intelligence: while data science is focused on optimizing and automating in a tactical way, data intelligence aims for creating change in a strategic way, with *people* as the decision makers.

This distinction begs a big question – ***what is the purpose?*** The goal of all this data development was not to put the machines in charge (at least we hope not) – it was to put the machines in **service**, and hopefully in service to noble goals. Instead of maximizing revenue for advertising placements on websites, Gourley's work focuses on things like avoiding military attacks – maybe even avoiding wars in the first place.

Video & publications with more detail on Gourley's work are available here:

- http://gigaom.com/2013/03/20/data-science-is-not-enough-we-need-data-intelligence-too/
- http://seangourley.com
- http://youtu.be/mKZCa_ejbfg
- http://youtu.be/V43a-KxLFcg

2. PRIMACY OF ETHICS (BELLAGIO FRAMEWORK):

The Bellagio framework is one developed by a group of fellows convened by the Rockefeller Foundation and PopTech last summer. This group was especially focused on big data and its role in resilience.

Perhaps the most surprising, and most encouraging theme of all that has emerged is recognition of the ***centrality of ethics and human judgment*** when it comes to big data. When the Rockefeller/PopTech group summarized their framework from the Bellagio gathering, ethics was at the very top of the chart. Here's why:

> *"In the popular discourse resilience often focuses on resources and infrastructure and overlooks issues of power, ethics and accountability…Resilience is not a normative term: systems characterized as resilient may be either desirable or undesirable. As a result, if ethics is not consciously considered at the inception of data projects, steps taken to increase community resilience could in fact create more vulnerability and thus do harm. This explains why our focus is on ethical resilience – and how we might apply this idea to data-driven community projects."*

This framework notes that principles should include informed consent, data ownership, accountability, transparency, data protection and data access.

More from the Bellagio report can (and should) be accessed here:

- http://poptech.org/system/uploaded_files/66/original/BellagioFramework.pdf

MORE BIG DATA RESOURCES:

- Patrick Meier also has written (and worked) in depth in this area, including with SFI, Ushahidi and Harvard's Program on Crisis Mapping. His blog can be found here: http://irevolution.net/2013/01/11/disaster-resilience-2-0/

- danah boyd's research, rooted in analysis of social media, has also frequently addressed big data: http://www.danah.org/papers/2012/BigData-ICS-Draft.pdf

- Of course Nassim Taleb's work also relates, including his latest "beyond resilience" work, Antifragile (where the core concept is of systems that are strengthened – not weakened - by disruptions): http://amzn.to/1ghrkBf

- The Fourth Paradigm: Jim Gray's work at Microsoft is constantly referenced by researchers. I have not read the full account given in Tony Hey's book yet, but am looking forward to it: http://amzn.to/1iqCpDl

- We've referenced Andrew Zolli's Resilience book on these pages before: terrific on its own, and also as a home base for all sorts of other resilience-related resources. http://amzn.to/1cipiff

- Jaron Lanier's books have raised provocative and important questions for the collection and use of big data. The latest, Who Owns the Future?, challenges the ethics of data collection and use, especially unpaid, non-consenting data collection, even more especially from the poor. http://amzn.to/Jqc5dL

BEYOND DATA: OTHER BOOKS AND MEDIA

POPTECH 2013 – SPARKS OF BRILLIANCE

As noted above, we find PopTech to be an endless source of insight and inspiration. Many of the talks from the October 2013 gathering are now online, including some of my favorites:

http://poptech.org/popcasts/ellen_langer_mindfulness_over_matter
http://www.poptech.org/popcasts/helen_marriage_art_interventions
http://www.poptech.org/popcasts/lisa_servon_better_banking
http://www.poptech.org/popcasts/david_robertson_creative_constraint
http://www.poptech.org/popcasts/rodney_mullen_getting_back_up

POPE-TECH – TALK ABOUT SPARKS!

How 'bout this Pope? By definition, world leaders are serious investors. They invest time, energy, and resources of all sorts, and guide others to do the same. They have capital. As one President bluntly noted, "I earned capital in the campaign, political capital, and now I intend to spend it." Pope Francis, spiritual leader of 1.2 billion Catholics, recently released his first long publication, the apostolic exhortation *Evangelii Gaudium (The Joy of the Gospel),* and it is a doozy. Even if you have no interest whatsoever in the Catholic Church, this is worth reading, I promise. And if you don't have time for his longer document, you can read the summary in our recent blog post (http://wp.me/p31aPs-7W). To see humble, authentic moral leadership elevated in some of our largest institutions is a source of great hope.

http://www.catholicculture.org/culture/library/view.cfm?recnum=10390

LITTLE FREE LIBRARY

Oh, how I love this venture! You put a little structure up in front of your home, school, or workplace and fill it with books – then people take one, give one, so you have, well, just like it says, a Little Free Library. They're like beautiful giant birdfeeders, but with books. People food.

http://www.littlefreelibrary.org

GAPMINDER

This is the amazing Hans Rosling's site, where you can see lots more about his work and even download some of his basic "moving graphs". Take the Ignorance Survey and review awesome time series like "Africa is Not a Country!" You will feel dumber and smarter, in quick succession.

http://gapminder.org

TOP PET PEEVES, by Clifford Asness

This is a thoughtful, detailed, heartfelt, and strongly presented list from the CFA Financial Analysts Journal, not a simple one-pager. Do terms like "smart beta" drive you bananas? Hop aboard, you've got company here.

http://www.cfapubs.org/doi/pdf/10.2469/faj.v70.n1.2

LET'S EXPLORE DIABETES WITH OWLS, by David Sedaris

HOLIDAYS ON ICE, by David Sedaris

I'm betting you have a little downtime right now, at the turn of the year – c'mon, at least a little. Even if it's just 1 hour, David Sedaris can make it feel like a holiday. I recently read this on a plane and woke up my passenger-neighbor, I was laughing so hard. If you're in a particularly jolly mood, Holidays on Ice also hits the spot (and the fantastic Macy's elf chronicles are available in audio form on NPR as well):

http://amzn.to/1k1qcXw

http://www.npr.org/2012/12/24/167716732/david-sedaris-reads-from-his-santaland-diaries)

STITCHES, by Anne Lamott

It's true, I am one of those annoying people who *really* likes the holiday season: I have indeed been known to wear socks with jingle bells on them, until I realized that I sounded like a giant cat. But year-end is not so festive for lots of people, and let's face it, it's cold and dark and we all have those days where we just want to burrow under the covers and not come out for a good long while. Whether you are a jingle-beller or a curl-upper (or a little bit of both), take this book with you. I promise it will help.

http://amzn.to/1ghp9xz

AND TWO FASCINATING NEW PUBLICATIONS:

NAUTILUS:

"Nautilus lets science spill over its usual borders. We are science, connected."

http://nautil.us/issue/7/waste/the-science-of-gratitude

ELEMENTA JOURNAL: Science of the Anthropocene

"Open Science for Public Good. We aim to facilitate scientific solutions to the challenges presented by this era of accelerated human impact."

http://elementascience.org

FINALE, PART I:

T.S. ELIOT, THE ROCK

The Eagle soars in the summit of Heaven,
The Hunter with his dogs pursues his circuit.
O perpetual revolution of configured stars,
O perpetual recurrence of determined seasons,
O world of spring and autumn, birth and dying
The endless cycle of idea and action,
Endless invention, endless experiment,
Brings knowledge of motion, but not of stillness;
Knowledge of speech, but not of silence;
Knowledge of words, and ignorance of the Word.
All our knowledge brings us nearer to our ignorance,
All our ignorance brings us nearer to death,
But nearness to death no nearer to God

Where is the Life we have lost in living?
Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?
The cycles of Heaven in twenty centuries
Bring us farther from God and nearer to the Dust.

FINALE, PART II:

I do love data, yet…

Still, what I want in my life
is to be willing to be dazzled—
to cast aside the weight of facts
and maybe even to float a little
above this difficult world.

I want to believe I am looking
into the white fire of a
great mystery.

I want to believe that the imperfections are nothing—
that the light is everything—
that it is more than the sum of each flawed blossom
rising and fading.

And I do.

— Mary Oliver, *House of Light*

With best wishes to all for a joyful, peaceful,
dazzling new year.