Massive Data Exploration in Astronomy: What does Cognitive have to do with it?

Kirk Borne

@KirkDBorne

Principal Data Scientist, Booz Allen Hamilton

http://www.boozallen.com/datascience

Discovery in Science

Where does Discovery in Science start?

- Does it start with data?
- Does it start with a hypothesis?
- Does it start with a story?

http://palomarskies.blogspot.com/2010/03/astrophoto-friday-horsehead-nebula.html

Let us start with a story, by looking at data…

http://palomarskies.blogspot.com/2010/03/astrophoto-friday-horsehead-nebula.html

And now… we have this 21st century look with new data…

Source for image: http://hubblesite.org/image/3844/printshop

Zooming into this image … *"That's funny! We see galaxies!"*

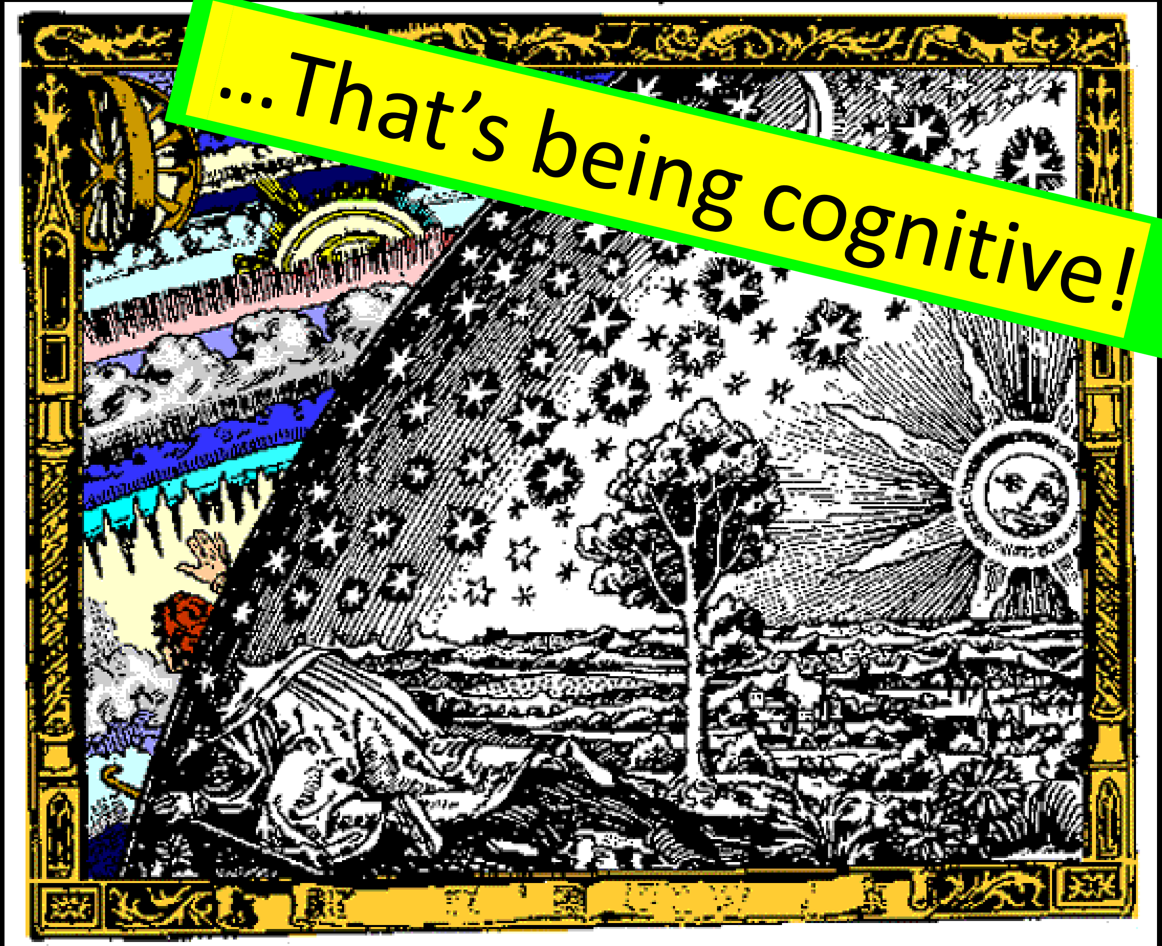
Source for image: http://hubblesite.org/image/3844/printshop

Ever since we first explored our world…
…we have asked questions about everything around us.

https://www.pinterest.com/pin/248683210647831264/

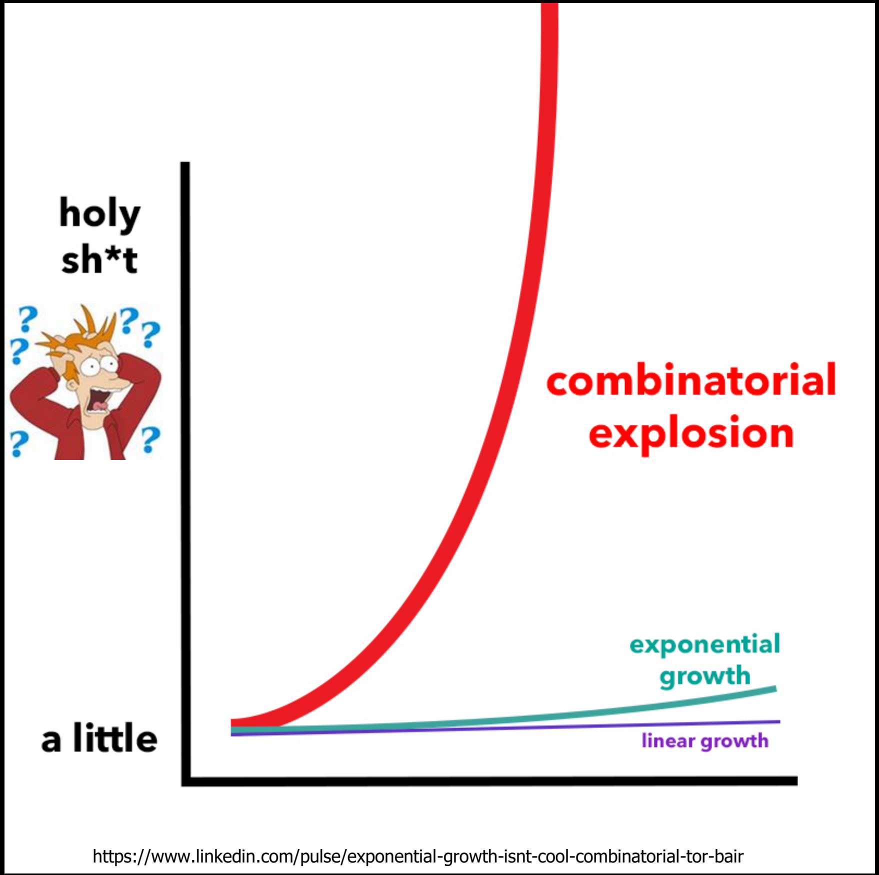
Ever since we first explored our world…

…we have asked questions about everything around us.



http pinterest.com/pin/248683210647831264/

So, we have collected evidence (data) to answer our questions, which leads to more questions, which leads to more data collection, which leads to more questions, which leads to **BIG DATA!**



Knowledge is about connecting the dots.

@KirkDBorne

$y \sim x! \approx x \wedge x$

→ Combinatorial Growth!
(all possible interconnections, linkages, and interactions)

$y \sim 2 \wedge x$ (exponential growth)

$y \sim 2 * x$ (linear growth)

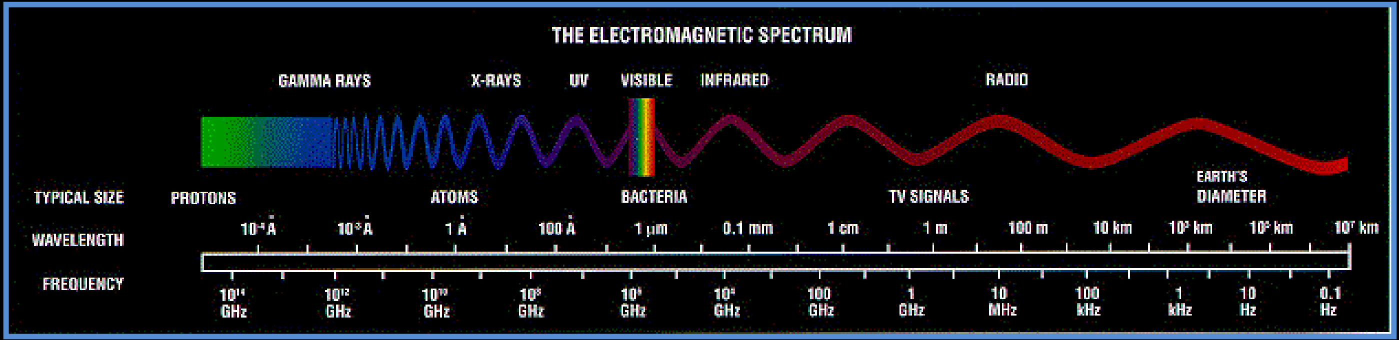
"Learn how to see. Realize that everything connects to everything else."

— Leonardo da Vinci

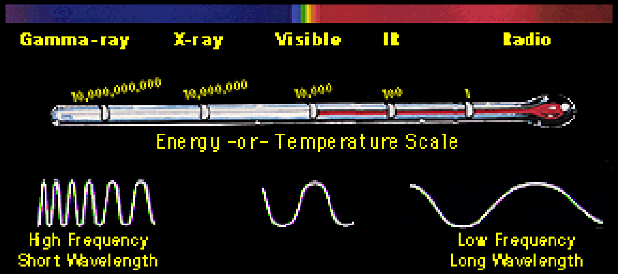
...That's cognitive!

Astronomy is a Forensic (evidence-based) Science

- The Electromagnetic Spectrum (complemented by Neutrinos, Cosmic Rays, and Gravitational Wave Radiation)



- **Radiation is the Astronomer's only source of information about the Universe!**
- And it is a remarkably rich & diverse source!
- Need multi-wavelength science instruments to observe our multi-wavelength Universe



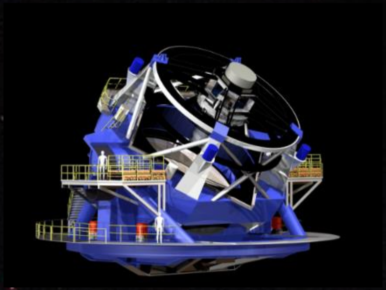
Astronomy is a Forensic (evidence-based) Science

- Discoveries are enabled by:
  - – data
  - – questions (and related stories)
  - – models, theories, and hypotheses
  - – hypothesis-testing with more data!

Discoveries have shown that the astronomical zoo is rich and diverse ...

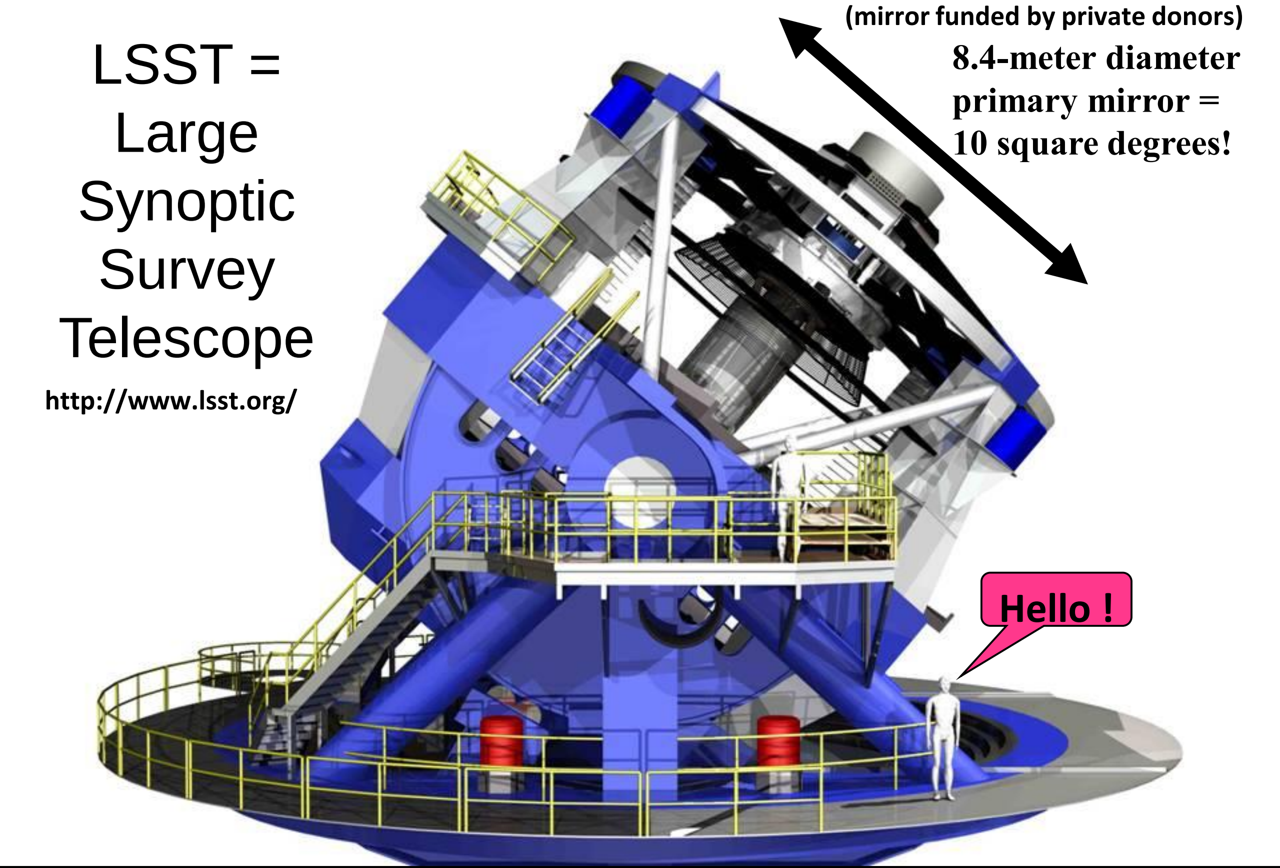
Black Holes
Quasars
Supernovae
Pulsars
Blazars
Tidal Streams
Colliding Galaxies
Magnetars
Gamma-ray bursts
Brown Dwarfs
Gravitational Lenses
Exo-planets
Serendipity !!
Incoming Killer Asteroid

Astronomy Big Data Example



The LSST (Large Synoptic Survey Telescope)
and the Dynamic Universe

100-200 Petabyte image archive
20-40 Petabyte database catalog



(construction started in 2014)

LSST Key Science Drivers: Mapping the Dynamic Universe

- Complete inventory of the Solar System (Near-Earth Objects; killer asteroids???)
- Nature of Dark Energy (Cosmology; Supernovae at edge of the known Universe)
- Optical transients (10 million daily event notifications sent within 60 seconds)
- Digital Milky Way (Dark Matter; Locations and velocities of 20 billion stars!)

South America

Chile

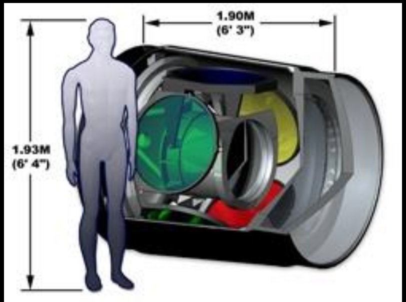
Region de Coquimbo

<u>LSST in time and space:</u>

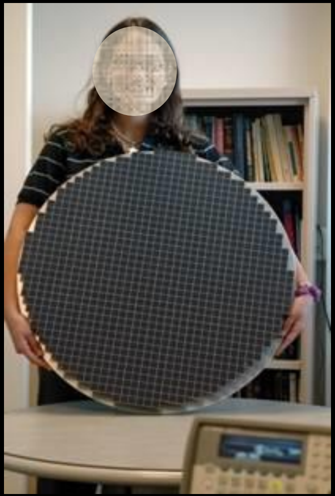
- When? ~2022-2032
- Where? Cerro Pachon, Chile

LSST Summary:
Big Data and Data Science

http://www.lsst.org/



- 3-Gigapixel camera
- One 6-Gigabyte image every 20 seconds
- 20 Terabytes every night for 10 years
- Repeat images of the entire night sky every 3 nights:
  - Celestial Cinematography
- 100-200 Petabyte final image data archive:
  - **all data are public!**
- 20-40 Petabyte final database catalog:
  - ~20 trillion sources with 200+ database attributes each
- ~10M events per night, every night, for 10 years:
  - Real-time event detection, triage, response, classification



But…

the LSST is not the biggest Big Data Astronomy project being planned …

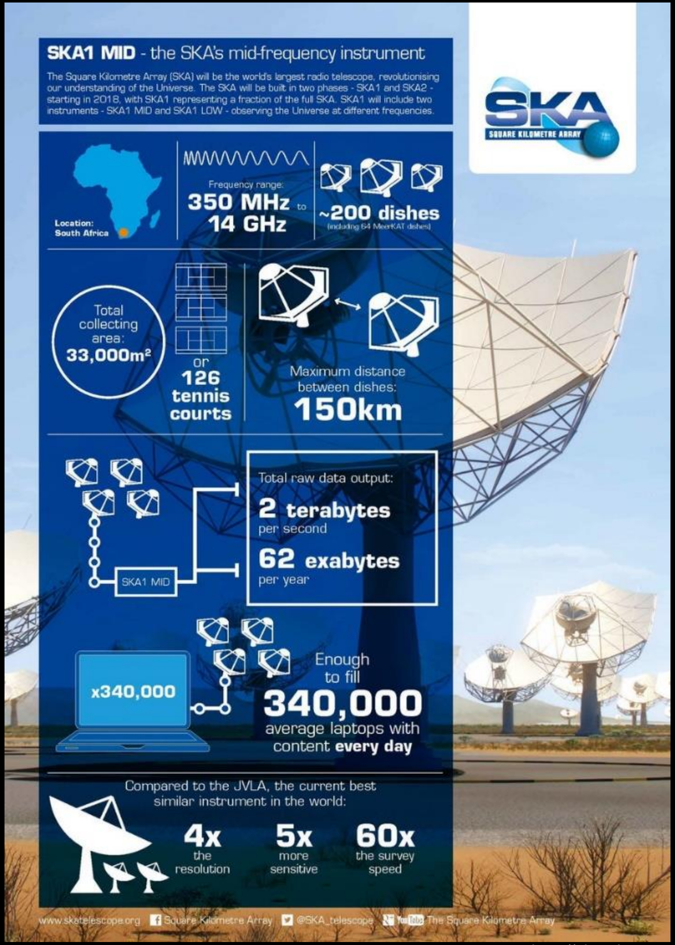
SKA

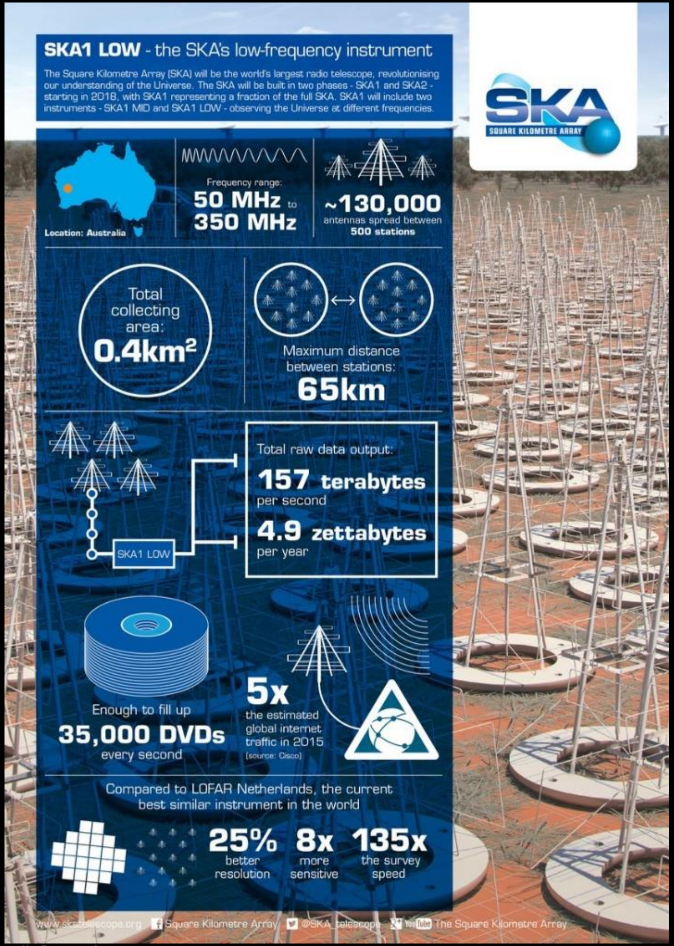
(starting in 2024)

SKA = Square Kilometer Array http://www.ska.gov.au/
(Joint project: Australia and South Africa)
= Discovery at Petascale and Exascale!

SKA = Square Kilometer Array http://www.ska.gov.au/

= Discovery at Petascale and Exascale!



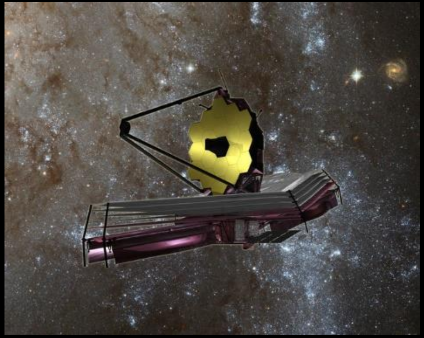


http://skatelescope.org/multimedia/image/

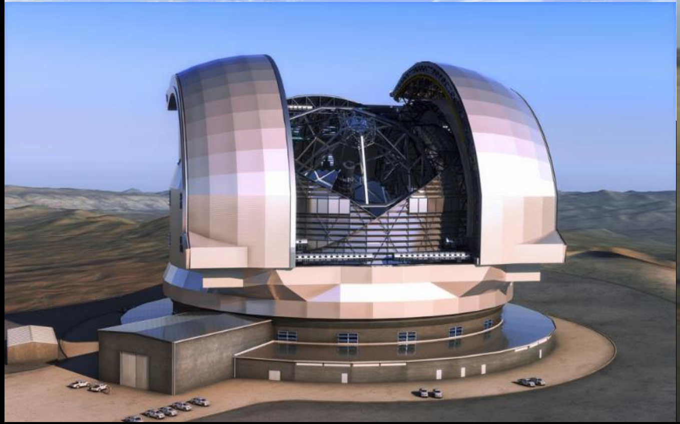
Why so many Telescopes?

Why so many Telescopes?

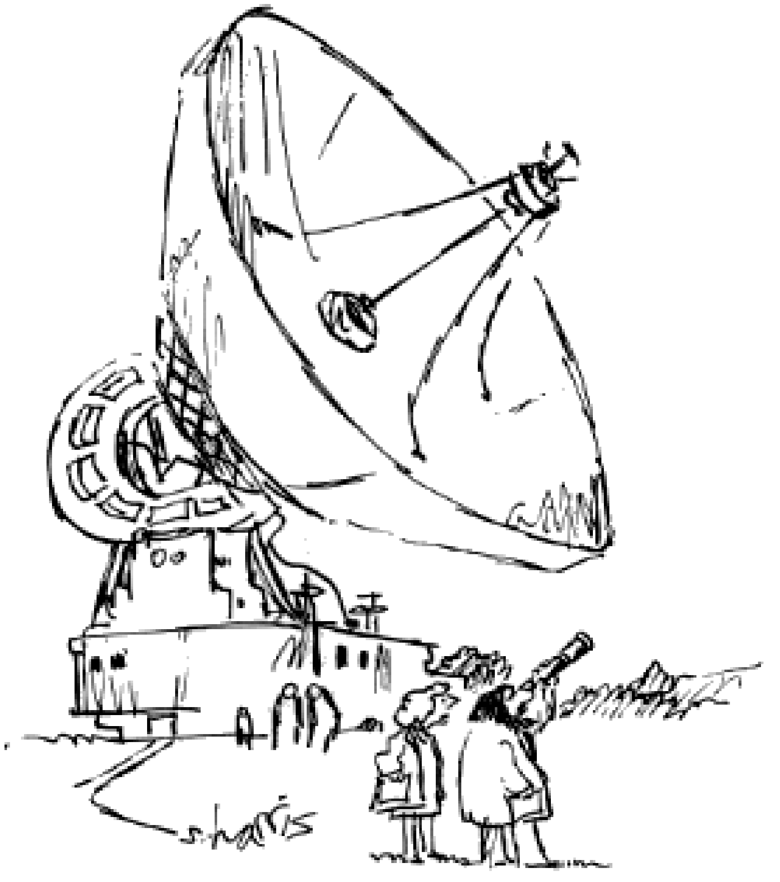


(on the Earth, and in space)


Why so many Telescopes?



"Just checking."

Because …

Many great astronomical discoveries have come from inter-comparisons of new objects and sources observed in different energy bands:

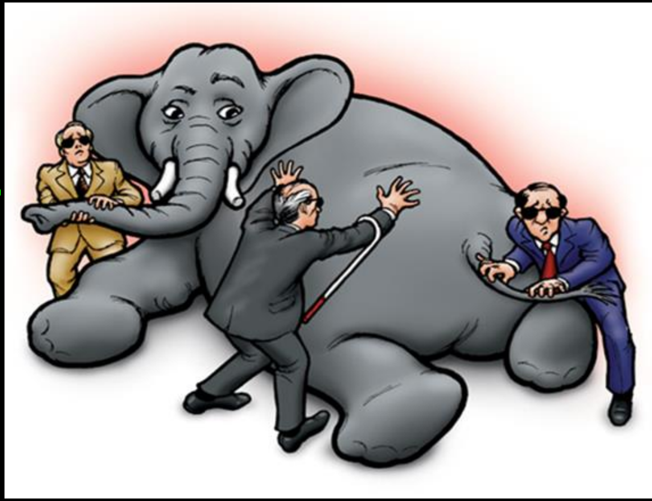
- Quasars
- Gamma-ray bursts
- Ultraluminous IR galaxies
- X-ray black-hole binaries
- Radio galaxies
- Neutrino oscillations
- . . .

https://www.aps.org/publications/apsnews/199511/backpage.cfm

A more universal reason for collecting data from many different sensors and instruments:

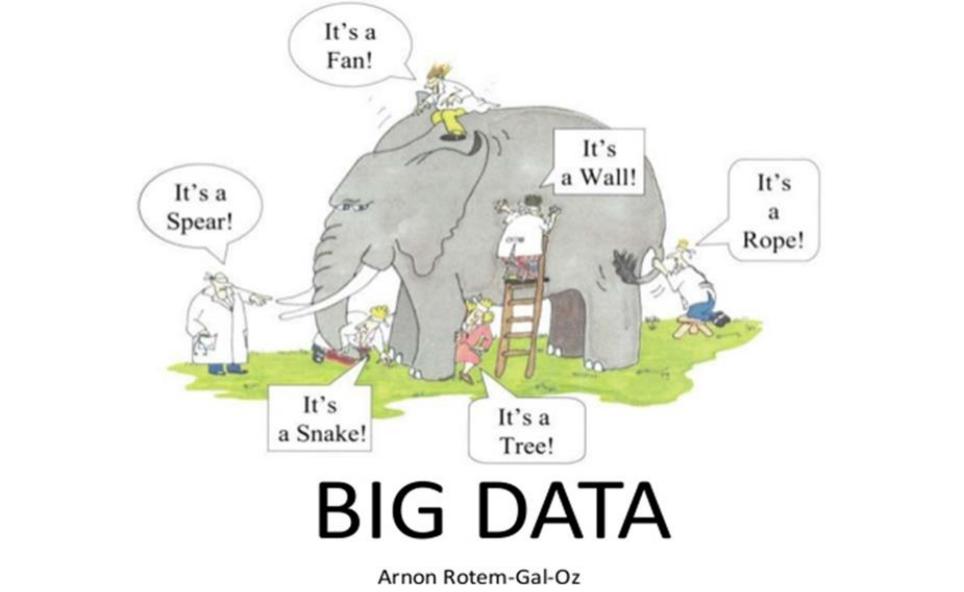
1. We collect many different sources of data.
2. But we usually store diverse data in separate silos.
3. Therefore, we cannot easily integrate the data to combine them for unified insight.

Consider the Blind Men
and the Elephant…



Source for graphic: http://www.vitria.com/blog/Big-Data-Analytics-Challenges-Facing-All-Communications-Service-Providers/

Adding more data doesn't necessarily help…



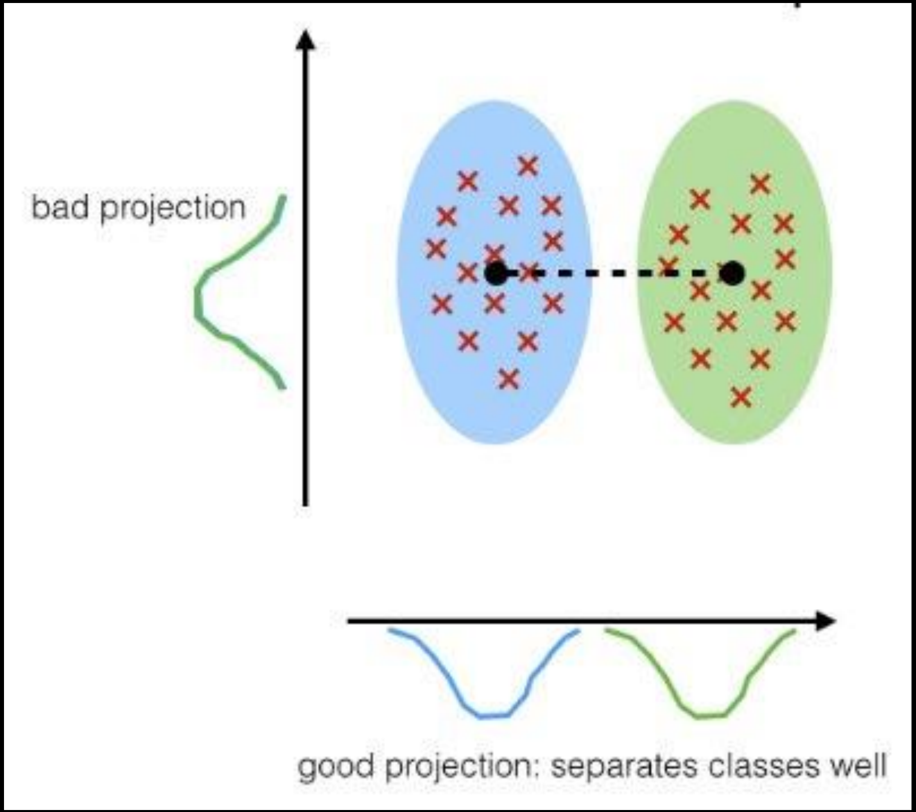
https://paulmead.com.au/blog/understand-perceptions/

Unless we can combine and integrate the different signals into a "single view" of the thing, there will continue to be many possible interpretations of what the source is!

Combining, connecting, and linking diverse data makes data "smart"!

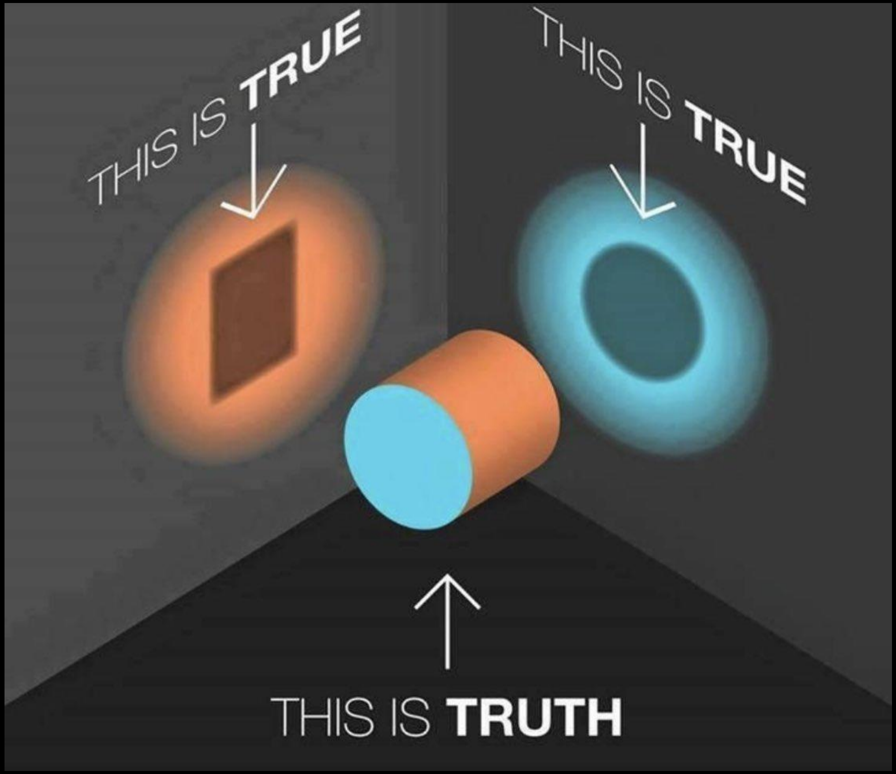
Think of data <u>not as information</u>, but as <u>measurements that encode knowledge</u>.

Feature Selection and Projection



Feature Selection is important in order to disambiguate different classes.
More importantly, **Class Discovery** depends on choosing the right projection and selecting the right features!

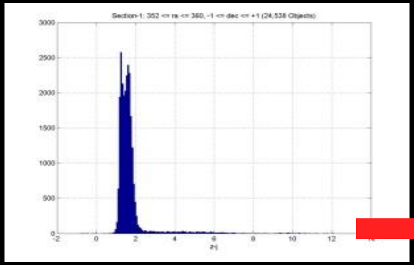
Projection Matters



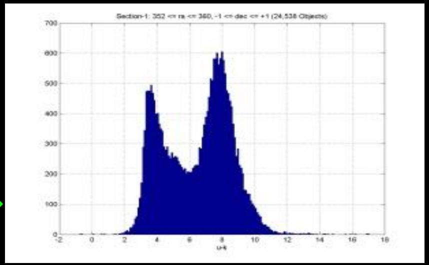
Your chosen data attributes represent a low-dimension projection of the full truth – the feature space (dimensions) in which you explore your data is a form of cognitive bias – it matters!

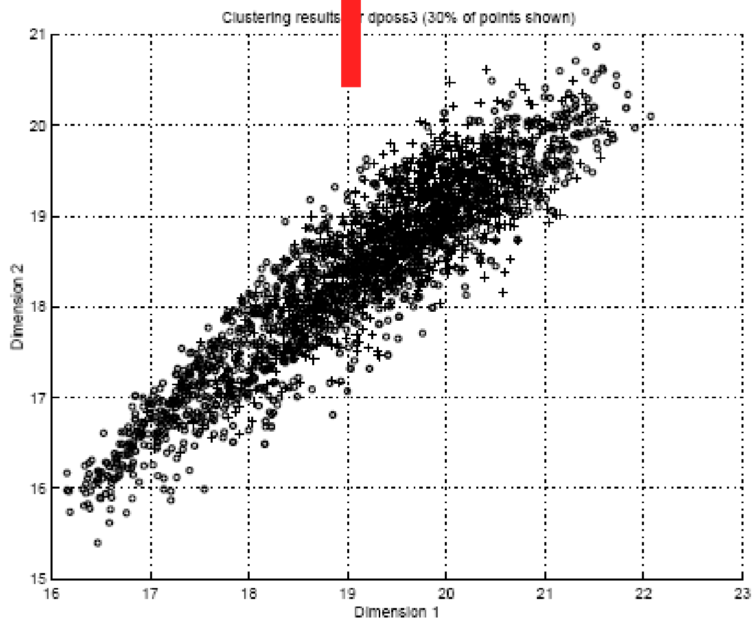
The 5 important D's of Data Variety:

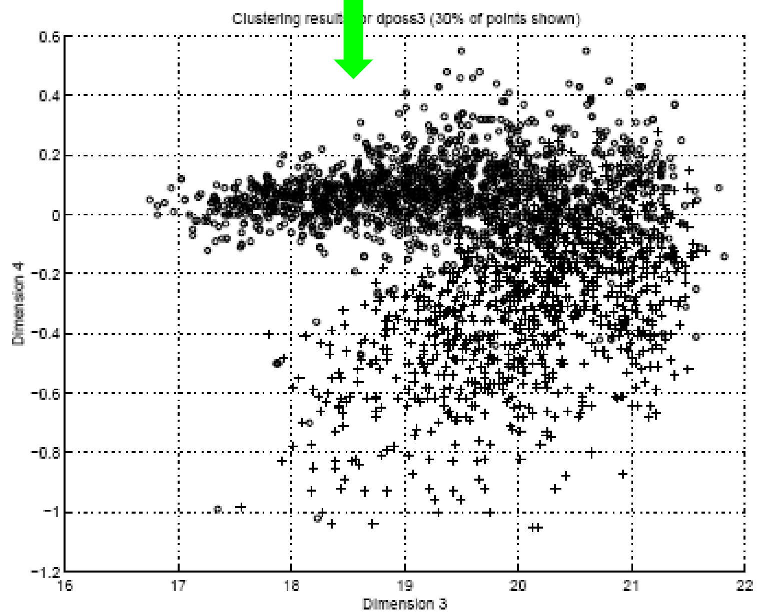
Entity **D**isambiguation, Entity **D**eduplication, **D**iscrimination between multiple classes, **D**iscovery of new classes, and **D**ecreased model bias (underfitting).



The separation and discovery of classes improves when a sufficient number of "correct" features are available for exploration and testing, as in the following two-class discrimination tests:







Reference: http://www.cs.princeton.edu/courses/archive/spr04/cos598B/bib/BrunnerDPS.pdf

The Analytics Maturity Scale and Unsupervised Discovery

Source: https://resources.zilliant.com/blog/smart-pricing-part-3-unsupervised-learning-algorithms

Levels of Analytics Maturity in Data-Driven Applications

1) Descriptive Analytics

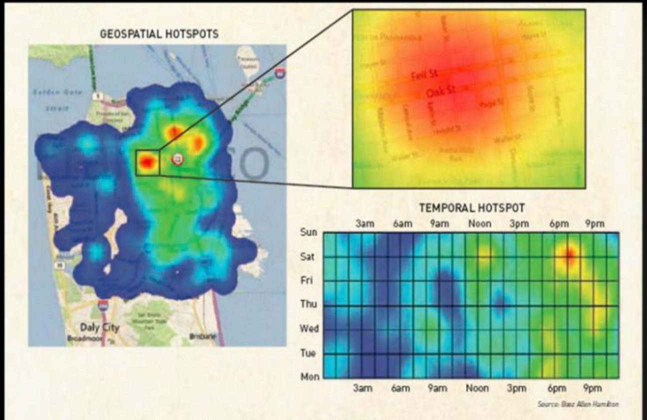
– **Hindsight** (What happened?)

2) Diagnostic Analytics

– **Oversight** (real-time / What is happening? Why did it happen?)

3) Predictive Analytics

– Foresight (What will happen?)



Geospatial and Temporal Car Theft Hotspots

5 Levels of Analytics Maturity in Data-Driven Applications

1) Descriptive Analytics

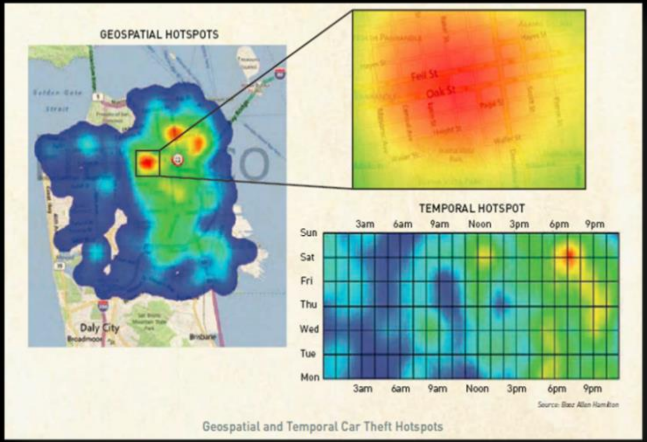
– **Hindsight** (What happened?)

2) Diagnostic Analytics

– **Oversight** (real-time / What is happening? Why did it happen?)

3) Predictive Analytics

– Foresight (What will happen?)



Geospatial and Temporal Car Theft Hotspots

4) Prescriptive Analytics

– **Insight** (How can we optimize what happens?) (Follow the dots / connections in the graph!)

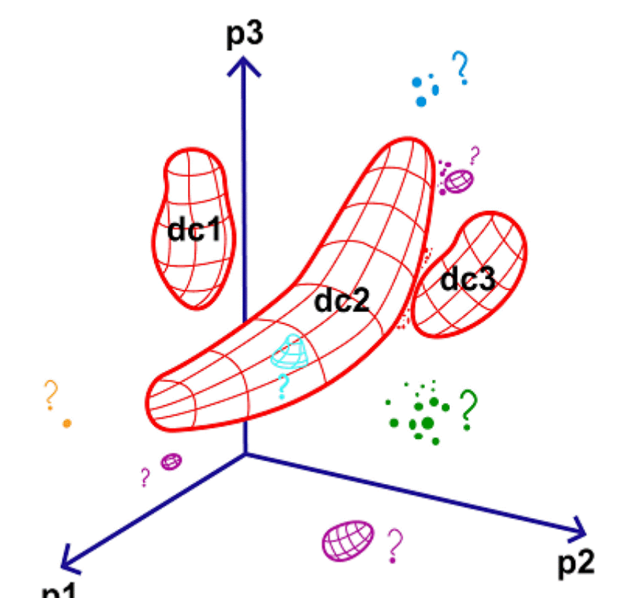
5) Cognitive Analytics

– **Right Sight** (the 360 view , **what is the right question to ask for this set of data in this context** = Game of Jeopardy)

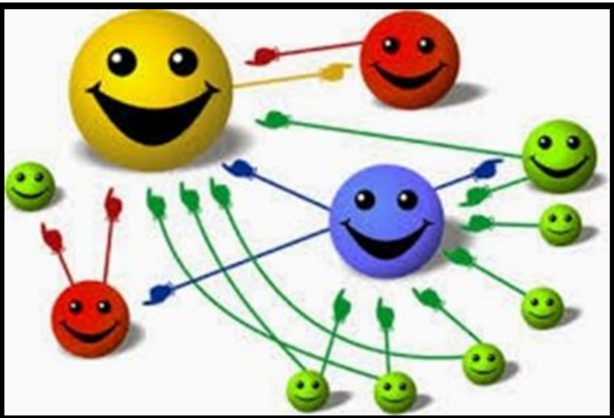
– Moves beyond simply providing answers, to generating new questions and hypotheses.

…That's cognitive!

4 Types of Discovery from Data:



(Graphic by S. G. Djorgovski, Caltech)



1) **Class Discovery:** Find the categories of objects (population segments), events, and behaviors in your data. + Learn the rules that constrain the class boundaries (that uniquely distinguish them).
2) **Correlation (Predictive and Prescriptive Power) Discovery:** (**INSIGHT DISCOVERY**) – Find trends, patterns, and dependencies in data that reveal the governing principles or behavioral patterns (the object's "DNA").
3) **Outlier / Anomaly / Novelty / Surprise Discovery:** Find the new, surprising, unexpected one-in-a-[million / billion / trillion] object, event, or behavior.
4) **Association (or Link) Discovery:** (Graph and Network Analytics) – Find both the typical (usual) and the atypical (unusual, interesting) data associations / links / connections in your domain.

Data Characterization, Contextualization, and Curation for Cognitive Discovery

Source: https://it.semrush.com/blog/content-curation-migliorare-posizionamento-case-study/

Data Characterization

Extraction, Exploration, Eureka!

- Identify and **Characterize** forensic features in the data:
  - Machine-generated
  - Human-generated
  - Crowdsourced? (**Citizen Science** = **Tap the Power of Human Cognition to find patterns and anomalies in massive data!**)
- Extract the **Context** of the data: the instrument, the time, the scientific use cases, extracted results, re-uses … where, when, who, how, what, why = *Metadata!*
- **Curate** these features for search, re-use, exploration, and new question-generation = ***Cognitive Discovery!***
  - Include other parameters and features from other data sources and databases

Data Contextualization

Extraction, Exploration, Eureka!

- Identify and **Characterize** forensic features in the data:
  - Machine-generated
  - Human-generated
  - Crowdsourced? (**Citizen Science** = **Tap the Power of Human Cognition to find patterns and anomalies in massive data!**)
- Extract the **Context** of the data: the instrument, the time, the scientific use cases, extracted results, re-uses … where, when, who, how, what, why = *Metadata!*
- **Curate** these features for search, re-use, exploration, and new question-generation = ***Cognitive Discovery!***
  - Include other parameters and features from other data sources and databases

Data Curation for Cognitive Discovery

Extraction, Exploration, Eureka!

- Identify and **Characterize** forensic features in the data:
  - Machine-generated
  - Human-generated
  - Crowdsourced? (**Citizen Science** = **Tap the Power of Human Cognition to find patterns and anomalies in massive data!**)
- Extract the **Context** of the data: the instrument, the time, the scientific use cases, extracted results, re-uses … where, when, who, how, what, why = *Metadata!*
- **Curate** these features for search, re-use, exploration, and new question-generation = ***Cognitive Discovery!***
  - Include other parameters and features from other data sources and databases

Computer Vision for Cognitive Discovery

Extraction, Exploration, Eureka!

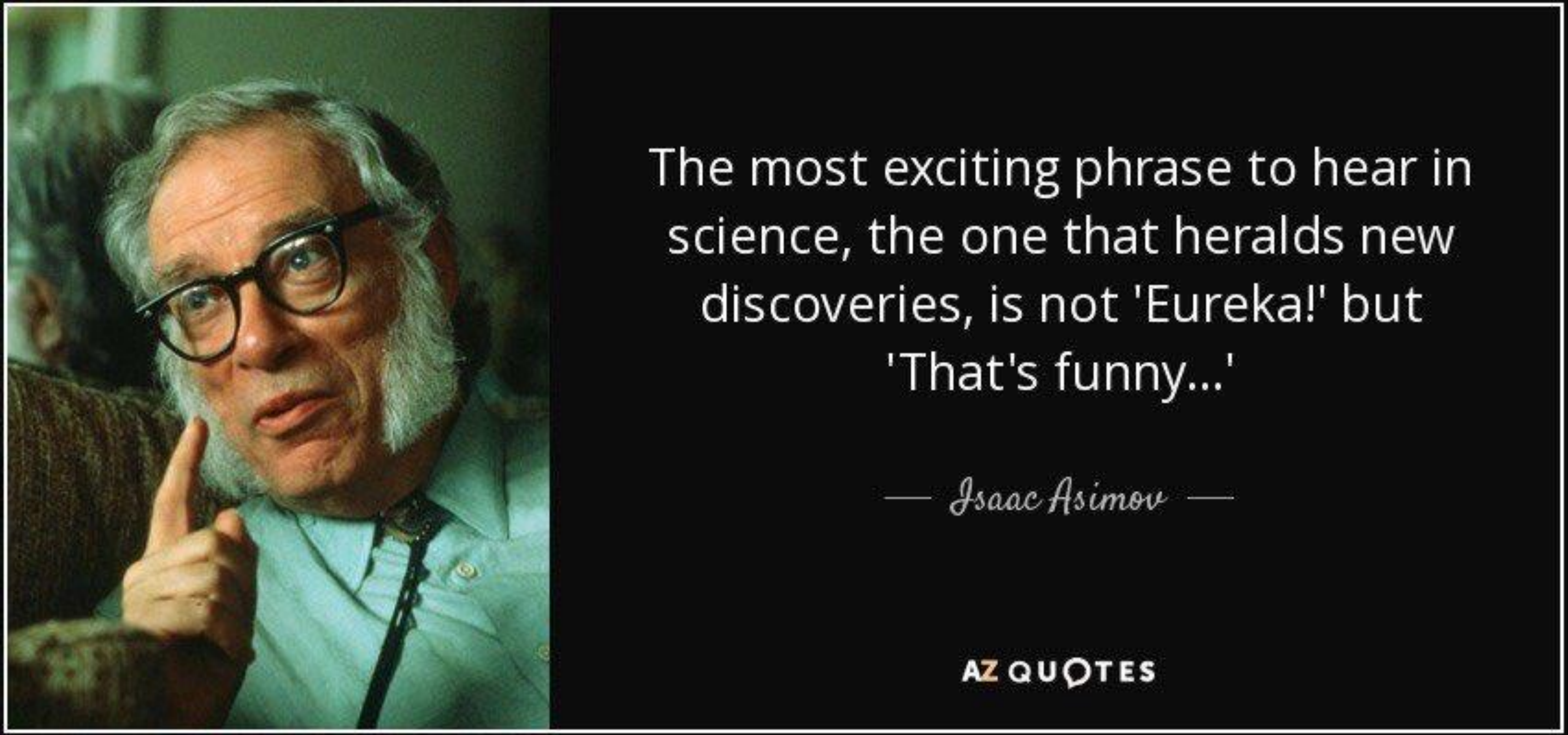
- Identify and **Characterize** forensic features in the data:
  - Machine-generated
  - Human-generated
  - Crowdsourced? (**Citizen Science** = **Tap the Power of Human Cognition to find patterns and anomalies in massive data!**)
- Extract the **Context** of the data: the instrument, the time, the scientific use cases, extracted results, re-uses … where, when, who, how, what, why = *Metadata!*
- **Curate** these features for search, re-use, exploration, and new question-generation = ***Cognitive Discovery!***
  - Include other parameters and features from other data sources and databases
- **2 examples: Computer Vision** "interesting feature" extraction from (a) "Google Maps" zoom views ; (b) Grand Tour sweeping views [**]

[**]Reference: https://link.springer.com/chapter/10.1007/978-1-4612-2856-1_16

Where does Discovery in Science start?

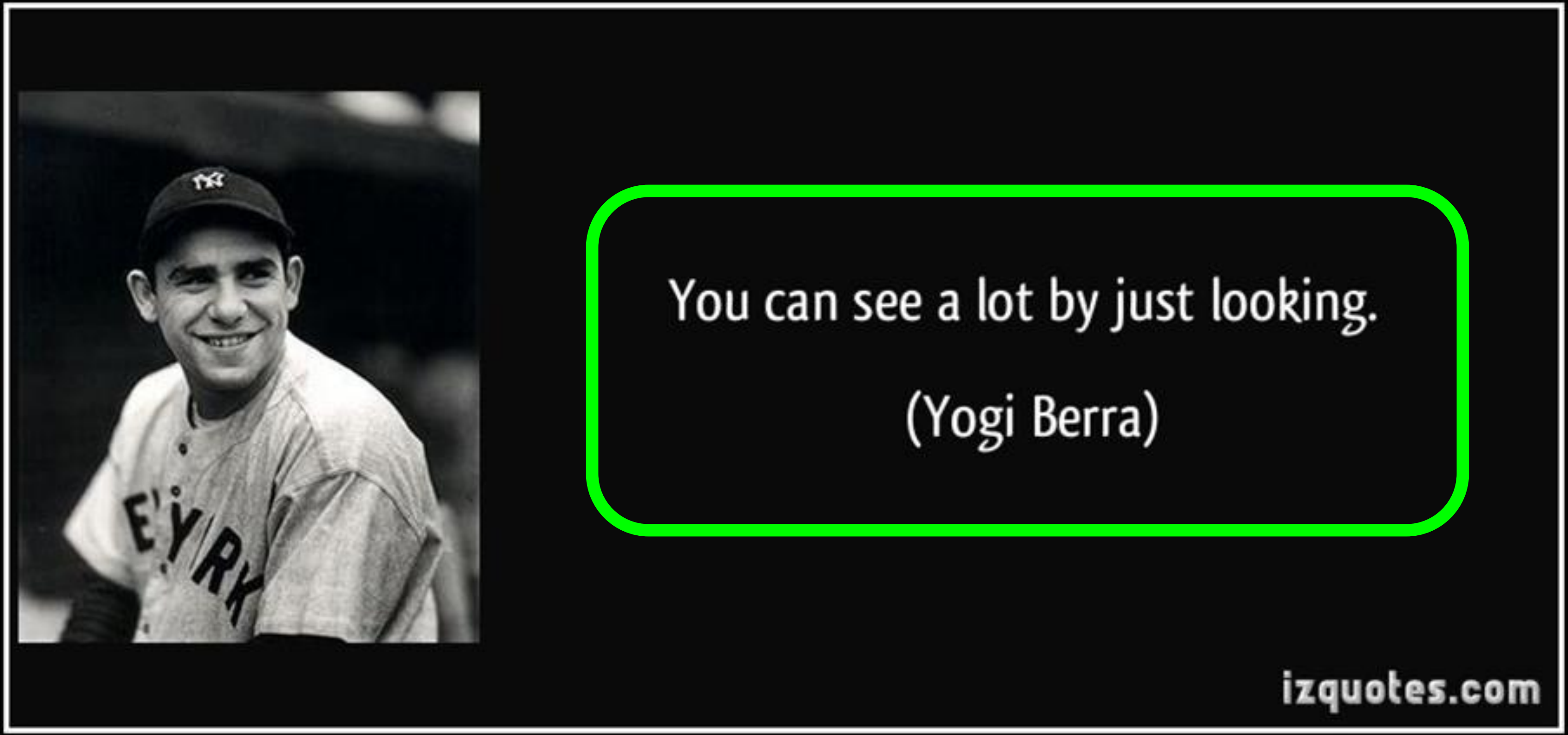
- Does it start with data? … YES!
- Does it start with a hypothesis? … not really, but as an inference from data (observation)!
- Does it start with a story? … inspired by data!

Where does Discovery in Science start?



That's Cognitive!

Where does most Discovery start?



That’s also Cognitive!

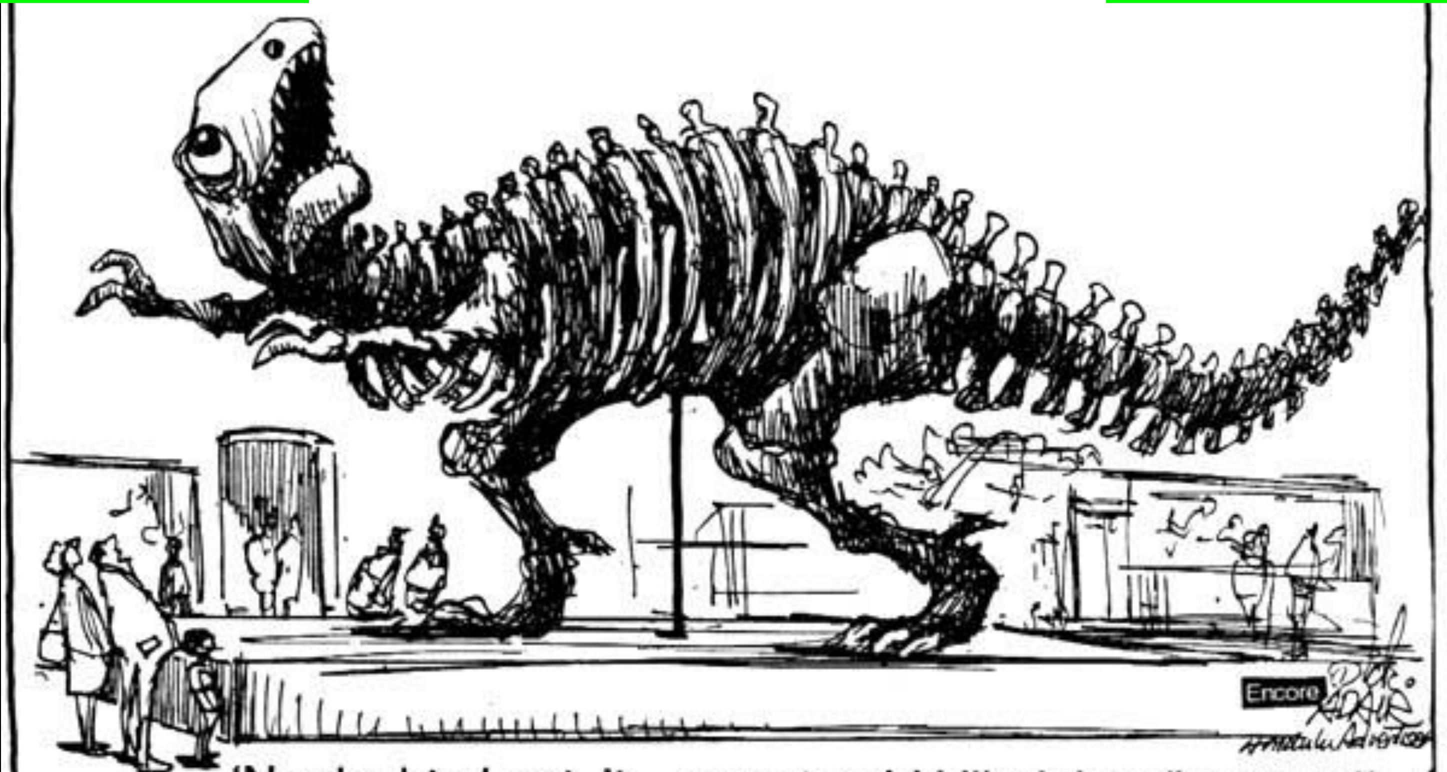
The Data, the Hypothesis, and the Story…

"No doubt about it ... an asteroid killed the dinosaurs!"

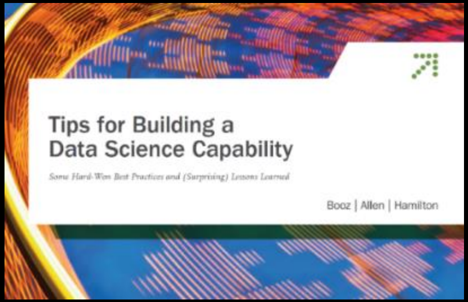
…That's forensic!

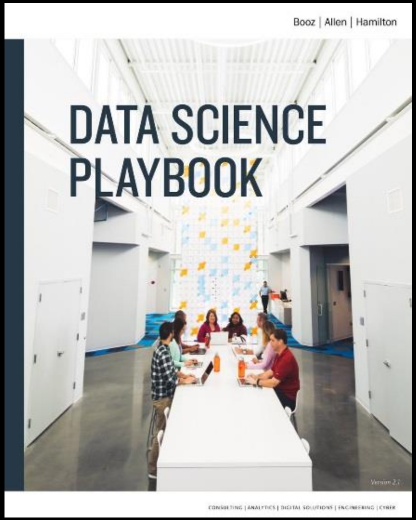
https://online.science.psu.edu/bisc002_activeup002/node/5264

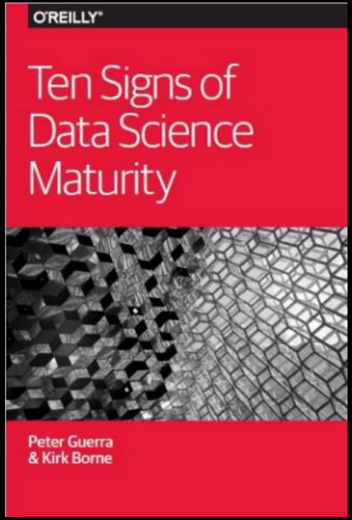
…That's cognitive!

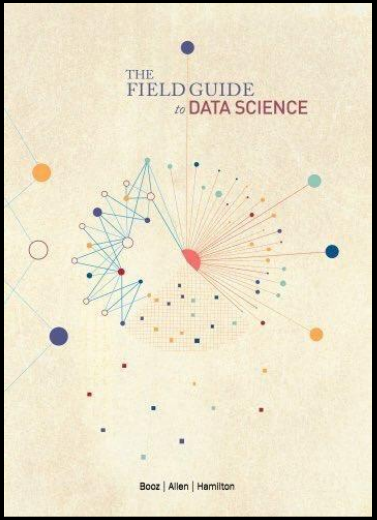


'No doubt about it... an asteroid killed the dinosaurs!'









Come for the data. Stay for the Science!

Thank you!

Twitter: @KirkDBorne or Email: kirk.borne@gmail.com

Get slides here: http://www.kirkborne.net/ADASS2018