

# THE STORY OF DATA

Why now?



- ▶ Have a Plan in life!
- ▶ What do you want to be in life?
- ▶ Why do you want to be that?
- ▶ How can you make it happen?
- ▶ Don't let life run your life, you run your life!

# LUCK FAVORS THE PREPARED

GENERAL PURPOSE FORMULA FOR LIFE!



# HOW TO LEARN EFFECTIVELY?

What

What are we seeking to learn?

Why

Why should we learn?

How

How to learn?



- ▶ Why?
- ▶ What?
- ▶ How?

Growth and Improvement  
Reject naivette, Reject cynicism  
Doing the same thing –produces same result  
Be critically analytic!

<https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science>

<https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/>

## RULE#1: ASK QUESTIONS

- ▶ What is data?
- ▶ Why do we care about data?
- ▶ Why now?
- ▶ How can I position myself?

# QUESTIONS

the concept of data as defined in the IFIP *Guide to Concepts and Terms in Data Processing*: “[Data is] a representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process.”  
--quoted from Forbes

# WHAT IS DATA?

# THE MAGIC OF NUMBERS

everything from the behaviour of celestial bodies to the expansion of the universe is based on mathematics.

## MATHS FROM THE START

of 12 numbers from 1 to 12, which is by 12. The numbers add the same as new-digit numbers add the same by itself.  
5. How does it work?  
16. How does it work?  
in a circular bit of number  
together 9.

The earliest records of mathematical activity – beyond counting – date from 4,000 years ago. They come from the fertile deltas of the Nile (Egypt) and the fertile plains between the two rivers, the Tigris and Euphrates (Mesopotamia, now Iraq). We know little of the individual mathematicians of these early cultures.

Around 600BC the Ancient Greeks developed an interest in mathematics. They went beyond their predecessors in that they were interested in finding rules that could be applied to any problem of a similar type. They worked on concepts in mathematics

which underlie all that has come since. Some of the greatest mathematicians of all time lived in Greece and the Hellenic centre of Alexandria in Egypt.

Islamic scholars pulled together the best of both Greek and Indian and forged something new. Their progress was great. The adoption of the Hindu system which we now use, as well as the Islamic calendar and

direction of Mecca of Islam which development of growth. Muslim intellectual and spiritually discovered true challenge to Luckly made knowledge

## EARLY HUMANS AND DATA





This mother duck appears to know the number of ducklings it must Protect, guide and train.

Even dynasaurs did and that is why in Jurassic Park they are depicted to come looking for missing eggs.

# DATA IS EXISTENTIAL FOR ALL LIFE FORMS

 <https://www.rcseng.ac.uk/library-and-publications/library/blog/mapping-disease-john-snow-air-x-12,547-rk2153...>

break of cholera reached the district of Soho, London, in August 1854. This was the first time that cholera had reached London, having previously occurred in 1832 and 1849. In the mid-19th century, Soho was a densely populated area, and a lack of proper sanitary services: the filthy streets, the lack of clean water, and the poor drainage of the district. It was at this point and drainage was poor throughout London. It was common for most homes.

Local residents (with the help of the [Reverend Henry Whitehead](#)), Snow identified the contaminated public water pump on Broad Street (now Broadwick Street). He noted that they were mostly people whose nearest access to water was from the pump. (see map below from [On the Mode of Communication of Cholera, 2<sup>nd</sup> ed.](#)). His success in convincing enough to persuade the local council to disable the well pump has been credited with contributing significantly to the containment of the disease. It was later found that the water for the pump was polluted by sewage contaminated with cholera.

# JOHN SNOW — 1854 — VISUALIZING DATA

<https://www.rcseng.ac.uk/library-and-publications/library/blog/mapping-disease-john-snow-and-cholera/>

<https://www.iweather.net.com/educational/history-weather-forecasting>

Demand Forecasting (Walmart used data to forecast beer/popart demand spike)

UPS analyzed data to understand left turns resulted in lost productivity

UPS used data to deliver on the same side before switching to the other side

Eliminating or minimizing left turns – millions saved

# WEATHER FORECASTING

- ▶ Data and Analysis is at work...all the time...

WHEN DID YOU LEAVE HOME TO GET  
HERE TODAY?



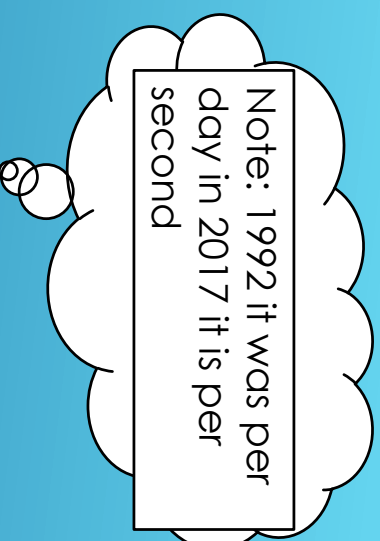
- ▶ Types (what kind of values are allowed .. Business rules → range of value)
  - ▶ Structural: Unstructured/Structured
  - ▶ Function: Transactional (Operational)/Fundamental
  - ▶ Models: Hierarchical/Network/Relational Data
- ▶ Another Slice (Enterprise Data Management)
  - ▶ – Master
  - ▶ – Metadata
  - ▶ – Reference
- ▶ <http://msdn.microsoft.com/en-us/library/bb190163.aspx>

## WHAT TYPES OF DATA

- ▶ In the beginning everything was hand-written, even books
- ▶ Then came printing press – print media
- ▶ Then came computers – digital media
  - ▶ Highly structured – transactional, point of sale
    - ▶ (Station, Date, Time,SKU,Qty,UnitPx,totalCost)
- ▶ Then came networks – first computers got connected
- ▶ Then with HTML/Social Media applications – People got connected
  - ▶ Human Communication is patently “unstructured”

## STRUCTURED/UNSTRUCTURED

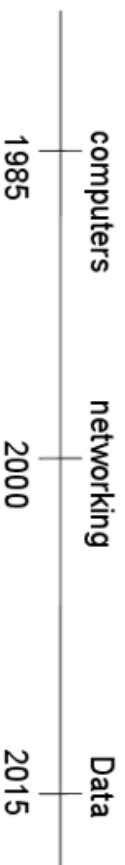
Year	Global Internet Traffic
1992	100 gigabytes per day
1997	100 gigabytes per hour
2002	100 gigabytes per second
2007	2,000 gigabytes per second
2012	12,000 gigabytes per second
2017	35,000 gigabytes per second



How long will it take to process all the tweets?  
 Entire wiki?  
 Watch all the youtube videos?

# EXPONENTIAL GROWTH

## Perfect Storm



We are now in  
the zone –  
hockey stick

✓ IBM estimates 2.5 quintillion bytes of data are generated each day.

✓ Ninety percent of the data in the world is less than two years old.



A quintillion – 18 zeros

Billion – 9 zeroes

Quintillion – billion billion

Big Data For Dummies by Alteryx

# PERFECT STORM – WHY EXP GROWTH?



#### INSIGHT 07

## Mastering data to drive outcomes creates competitive advantage

The problem for businesses is no longer the absence of data. In a time when they are flooded with new data, the problem becomes the absence of the *right* data, which is what will produce the sharp insights that spur the most actionable outcomes. And those outcomes, in turn, create competitive advantage.

<http://www.accenture.com/in-en/landing-pages/advertising/Documents/PDF/Accenture-High-Performance-IT-1.pdf>

Business needs actionable insight.

There is a deluge of data

Raw Data ->Information ->Knowledge

Information Management is key

[http://www.allanalytics.com/radio.asp?doc\\_id=269199&gateway\\_return=true](http://www.allanalytics.com/radio.asp?doc_id=269199&gateway_return=true)

# WHY?

importing data (finding sources, exploring, refining/cleansing data)  
Analyzing data (modeling, extracting patterns, knowledge)  
Reporting explaining what was done (explaining to the world around)

It is relevant to us while importing/analyzing/reporting using  
real world data is the focus,  
using established/core information management principles, because  
we want reproducible and repeatable experiments. One time results are not  
useful

[http://www.allanalytics.com/author.asp?doc\\_id=269883&f\\_src=AllAnalytics\\_finalanalysis](http://www.allanalytics.com/author.asp?doc_id=269883&f_src=AllAnalytics_finalanalysis)

# HOW

Data --> Information --> Knowledge

Event ID : 1	
HR	68
PR	181
QRSD	86
QT	388
QTc	413
--Axis--	
P	64
QRS	-29
T	0

01/17/2018 13:25:11	
Vent. Rate:	75 bpm
RR Interval:	800 ms
PR Interval:	176 ms
QRS Duration:	88 ms
QT Interval:	382 ms
QTc Interval:	408 ms
QT Dispersion:	56 ms
P Axis:	68 deg
QRS Axis:	-1 deg
T Axis:	-42 deg

Abnormal ECG

Sinus rhythm,  
Possible left  
Extensive T

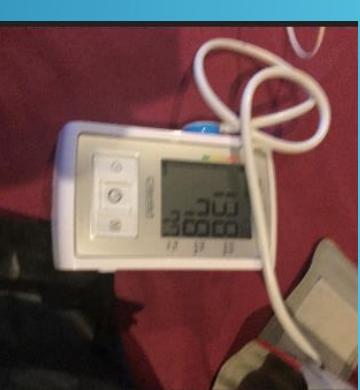
Data

Context

Visit a  
cardiologist  
Doctor says  
this ECG is  
normal.

Patient walks  
out with  
knowledge  
ACTION-  
NOTHING TO  
DO

138  
78



DATA->INFORMATION->KNOWLEDGE

## 10

### Rise of metadata catalogs helps people find analysis-worthy big data

For a long time, companies threw away data because they had too much to process. With Hadoop, they can process lots of data, but the data isn't generally organized in a way that can be found.

Metadata catalogs can help users discover and understand relevant data worth analyzing using self-service tools. This gap in customer need is being filled by companies like **Alation** and **Waterline** which use machine learning to automate the work of finding data in Hadoop. They catalog files using tags, uncover relationships between data assets, and even provide query suggestions via searchable UIs. This helps both data consumers and data stewards reduce the time it takes to trust, find, and accurately query the data. In the coming year, we'll see more awareness and demand for self-service discovery, which will grow as a natural extension of self-service analytics.

# META DATA: TOP TEN TREND

Consider

AAA,1891,330440,435

FFF,1975,109000,20000

ZZZ,1812,440000,3700

If you get this collection of data, what sense can you make out of this?

Meta data helps you to understand what the data is? Use it consistently with those who created the data.

Again it is not that easy if we do not have a standard DDL

## HOW META DATA(EDM)

Data about data.

Now, let us make a small change .

Consider

IBM,1891,330440,435

CSCO,1975,109000,20000

C,1812,440000,3700

If you get this collection of data, what can you now make out of this?

Meta data helps you to understand what the data is?

## HOW META DATA – 02 (CONTEXT)

IBM,1891,330440,435  
CSCO,1975,109000,20000  
C,1812,440000,3700

This is data

There are four fields:

Company Name, Year Established, NumberOfEmployees, Locations

This is meta-data

Data about data, not data

## HOW META-DATA - 03

- ▶ Keeping data, separate from meta data
  - ▶ Allows mis-interpretation
- ▶ How to prevent
  - ▶ Self Describing Format
    - ▶ XML → XBRL
    - ▶ JSON (to an extent)

```
IBM,1891,330440,435  
CSCO,1975,109000,20000  
C,1812,440000,3700
```

```
<Corporation>  
<Symbol>IBM</Symbol>  
<YearOfIncorporation>1891</YearOfIncorporation>  
<NumberOfEmployees>330440</NumberOfEmployees>  
<NumberOfLocations>435</NumberOfLocations>  
</Corporation>
```

## SELF DESCRIBING DATA



- ▶ Meta data then describes format, business connotation and
- ▶ range of values (aka domain)
- ▶ Context, rules of use and interpretation, units of measure
- ▶ Temperature is 32
  - ▶ Is it cold or hot?
  - ▶ Depends if it is Celsius or F...

# CONSISTENT MEANING

## Quantitative

- Numerical
- Integer/double
- Precision
- Ratio (division, zero)

## Qualitative

- Categorical
- Nominal (values, Chicago, NYC, Boston, LA)
- Ordinal (LOW,HIGH)
- Interval (Temperature)

# TYPES OF DATA

Variety, not volume or velocity,  
drives big-data investments

**Gartner** defines big data as the three Vs: high-volume, high-velocity, high-variety information assets. While all three Vs are growing, variety is becoming the single biggest driver of big-data

Ask the Question: Why might that be?

# BIG DATA: THE NEW KID

- ▶ Weather data has always been voluminous – not a recent phenomena
- ▶ Financial Services has always handled transactions at very high rate
  - ▶ <https://www.nasdaq.com/aspx/dailymarketstatistics.aspx>
  - ▶ <http://www.nasdaqtrader.com/Trader.aspx?id=DailyMarketSummary> (10mm trades)
- ▶ Credit Card transactions

#### Visa transactions per second

VisaNet handles an average of 150 million **transactions** every day and is capable of handling more than **24,000 transactions per second**. **Visa** has invested heavily in advanced fraud-fighting technologies, so you can assure your customers that their card information is safe.

~ approx 40 MICRO

# VOLUME AND VELOCITY ARE NOT NEW...



Prior to people oriented conversation, data was entirely generated by computers – with a definite format – aka structured data

Unstructured data dominates structured data.

We just don't know how to stop talking, even though we have one mouth, two ears!

# VARIETY IS NEW

<https://www.nyse.com/data/transactions-statistics-data-library>

[http://www.nyxddata.com/nysedata/asp/factbook/viewer\\_edition.asp?mode=table&key=3141&category=3](http://www.nyxddata.com/nysedata/asp/factbook/viewer_edition.asp?mode=table&key=3141&category=3)

Date Shares, Trades, USD

1/2/2015	891175786	3969459	33253336431
1/5/2015	1167614439	5049475	44299075404
1/6/2015	1338735158	5974051	49062304563
1/7/2015	1104507004	4942803	40680944878
1/8/2015	1165175679	4724036	44757928499
1/9/2015	1035301255	4526313	39108246670
1/12/2015	1106969304	4718908	41560740908
1/13/2015	1265891339	5714159	46180555406
1/14/2015	1346417157	5822538	48745211277
1/15/2015	1285191043	5562173	45749131945
1/16/2015	1341580612	5302701	51952925517
1/20/2015	1211541615	5020222	45205949674

So, volume, velocity is  
nothing new. We have  
always known it

# NYSE, LET US LOOK AT SOME REAL DATA

Mining large amounts of structured and unstructured data to identify patterns that can help an organization rein in costs, increase efficiencies, recognize new market opportunities, understand and predict customer behavior and increase an organization's competitive advantage.

DATA → DATA SCIENCE

*More than 50 years ago, John Tukey called for a reformation of academic statistics. In 'The Future of Data Analysis', he pointed to the existence of an as-yet unrecognized science, whose subject of interest was **learning from data**, or 'data analysis'.*

<http://courses.cscill.mit.edu/18.337/2015/docs/50YearsDataScience.pdf>

## Prediction and Inference

cal Modeling: **The Two Cultures**, Breiman described two cultural outlooks about extracting value from data.

*Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables  $x$  (independent variables) go in one side, and on the other side the response variables  $y$  come out. Inside the black box, nature functions to associate the predictor variables with the response variables ...*

*There are two goals in analyzing the data:*

- **Prediction.** To be able to predict what the responses are going to be to future input variables;
- **Inference**.<sup>23</sup> To [infer] how nature is associating the response variables to the input variables.

# DATA SCIENCE/ANALYTICS



# Many Names: Science, Mining, Learning

## Parametric method

Statistical: Central Tendencies, Measures of Dispersion, Correlation, Covariance,  
Probability Distribution, Joint, Conditional probability, Bayesian

## Non-parametric

### Nearest Neighbor Based

birds of a feather flock together

tell me who your friends are, I will tell you, who you are

### Logic Based

### Topology/geometry based

### Perceptron based neural and deep learning

- ▶ It is cloudy, take an umbrella (prior experience)
  - ▶ It is highly probable it will rain, but clouds can come and go without rain
- ▶ Onset of pandemic
  - ▶ Markets likely to decline
- ▶ Probabilistic reasoning is centuries old, robust formalism exist
  - ▶ Gaussian (normal distribution, bell curve) ( $\mu$  and  $\sigma$  are the parameters),
    - ▶ Observed in sufficiently large numbers every natural process asymptotically follows the bell curve
  - ▶ Binomial and Poisson distribution and there are other exotic distributions

# PARAMETRIC METHODS (WE DO IT ALL THE TIME)

```

> cdataset<-rbind(birds2,data.frame(nlegs=rep(2,5),can_fly=rep(0,5), height=hchicks
+ color=sample(chickencolors,5,replace=T),species=rep('chicken',5)),
+ data.frame(nlegs=rep(2,5),can_fly=rep(1,5), height=hvulture,
+ color=sample(vulturecolors,5,replace=T),species=rep('vulture',5)),
+ data.frame(nlegs=rep(2,5),can_fly=rep(1,5), height=hparrot,
+ color=sample(parrotcolors,5,replace=T),species=rep('parrot',5)),
+ data.frame(nlegs=rep(2,5),can_fly=rep(0,5), height=hostrich,
+ color=sample(ostrichcolors,5,replace=T),species=rep('ostrich',5)),
+ data.frame(nlegs=rep(2,5),can_fly=rep(1,5), height=hsparrow,
+ color=sample(sparrowcolors,5,replace=T),species=rep('sparrow',5))
+ )
> cdataset<-rbind(birds2,data.frame(nlegs=rep(2,5),can_fly=rep(0,5), height=hchicks
+ color=sample(chickencolors,5,replace=T),species=rep('chicken',5)),
+ data.frame(nlegs=rep(2,5),can_fly=rep(1,5), height=hvulture,
+ color=sample(vulturecolors,5,replace=T),species=rep('vulture',5)),
+ data.frame(nlegs=rep(2,5),can_fly=rep(1,5), height=hparrot,
+ color=sample(parrotcolors,5,replace=T),species=rep('parrot',5)),
+ data.frame(nlegs=rep(2,5),can_fly=rep(0,5), height=hostrich,
+ color=sample(ostrichcolors,5,replace=T),species=rep('ostrich',5)),
+ data.frame(nlegs=rep(2,5),can_fly=rep(1,5), height=hsparrow,
+ color=sample(sparrowcolors,5,replace=T),species=rep('sparrow',5))
+ )

```

```

> cdataset
  nlegs can_fly  height color species
1      2      0 25.000000 black chicken
2      2      1 40.000000 black vulture
3      2      1 20.000000 blue  parrot
4      2      0 150.000000 black ostrich
5      2      1 10.000000 brown sparrow
6      2      0 30.742869 black chicken

```

```

> require(rpart)
> require(rpart.plot)
> sp_rpart<-rpart(species~.,data=cdset,minsplit=2)
> rpart.plot(sp_rpart,main=' Logic driven Decision Tree')
> |

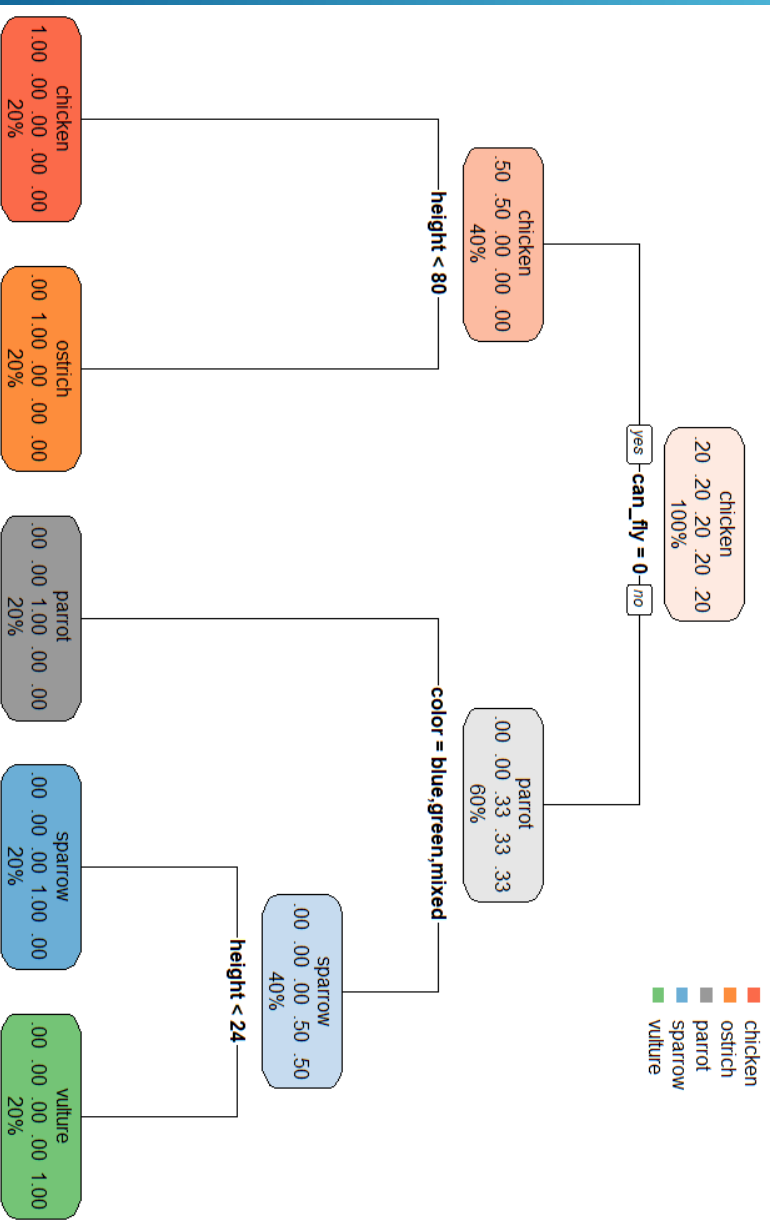
```

```

> require(rpart)
> require(rpart.plot)
> sp_rpart<-rpart(species~,data=cset,minsplit=2)
> rpart.plot(sp_rpart,main=' Logic driven Decision Tree')
> |

```

## Logic driven Decision Tree



- ▶ Nearest neighbor (similarity based) reasoning
  - ▶ <https://www.slideshare.net/rk2153/chapter-05-k-nn>



BIRDS OF A FEATHER FLOCK TOGETHER

Instance based methods are awaiting for renaissance from your generation...

https://en.wikipedia.org/wiki/Minkowski\_distance

80%

...

↑

↓

↻

🔍

👤

🔒

Raman Kaman | New ... Raman Kaman ProfileprofiteK mess... Raman Kaman at Hun... 10.1 Panel Data | Intro...

Minkowski distance

From Wikipedia, the free encyclopedia

Not to be confused with the pseudo-Euclidean metric of the Minkowski space.

The **Minkowski distance** or **Minkowski metric** is a metric in a normed vector space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance. It is named after the German mathematician Hermann Minkowski.

**Definition** [ edit ]

The Minkowski distance of order  $p$  (where  $p$  is an integer) between two points  $X = (x_1, x_2, \dots, x_n)$  and  $Y = (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$  is defined as:

$$D_p(X, Y) = \left( \sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}}$$

For  $p \geq 1$ , the Minkowski distance is a metric as a result of the Minkowski inequality. When  $p < 1$ , the distance between (0,0) and (1,1) is  $2^{1/p} > 2$  but the point (0,1) is at a distance 1 from both of these points. Since this violates the triangle inequality, for  $p < 1$  it is not a metric. However, a metric can be obtained for these values by simply removing the exponent of  $1/p$ . The resulting metric is also an  $F$ -norm.

Minkowski distance is typically used with  $p$  being 1 or 2, which correspond to the Manhattan distance and the Euclidean distance, respectively. In the limiting case of  $p$  reaching infinity, we obtain the Chebyshev distance:

$$\lim_{p \rightarrow \infty} \left( \sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}} = \max_{i=1}^n |x_i - y_i|.$$

Similarly, for  $p$  reaching negative infinity, we have:

$$\lim_{p \rightarrow -\infty} \left( \sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}} = \min_{i=1}^n |x_i - y_i|.$$

BRAIN MAKES IT LOOK EASY – GORY  
MATH LURKS BELOW

[https://en.wikipedia.org/wiki/Problem\\_of\\_induction](https://en.wikipedia.org/wiki/Problem_of_induction)

- ▶ Russell, Locke, Hume thought deeply about generalization or induction
- ▶ <https://www.jstor.org/stable/pdfplus/27744698>

The **problem of induction** is the philosophical question of what are the justifications, if any, for any growth of knowledge understood in the classic philosophical sense—knowledge that goes beyond a mere collection of observations!<sup>11</sup>—highlighting the apparent lack of justification in particular for:

1. Generalizing about the properties of a class of objects based on some number of observations of particular instances of that class (e.g., the inference that "all swans we have seen are white, and, therefore, all swans are white", before the discovery of black swans) or
2. Presupposing that a sequence of events in the future will occur as it always has in the past (e.g., that the laws of physics will hold as they have always been observed to hold). Hume called this the principle of uniformity of nature.

# NOT A NEW PROBLEM

It will be untenable to be data illiterate...  
Data Literacy and proficiency are imperative

BE WHERE THE PUCK WILL BE! W.G.





**Table 1.1** Example Analytics Applications

Marketing	Risk Management	Government	Web	Logistics	Other
Response modeling	Credit risk modeling	Tax avoidance	Web analytics	Demand forecasting	Text analytics
Net lift modeling	Market risk modeling	Social security fraud	Social media analytics	Supply chain analytics	Business process analytics
Retention modeling	Operational risk modeling	Money laundering	Multivariate testing		
Market basket analysis	Fraud detection	Terrorism detection			
Recommender systems					
Customer segmentation					

# INDUSTRY –APPLICATIONS

- ▶ Develop critical thinking, ask questions, be curious and be open
- ▶ Data has a story to tell, be open minded – avoid biases
- ▶ Develop strong analytical and statistical analytical skills
- ▶ Be a doer – this is not for spectators
- ▶ Communication skills – you need them

## HOW TO BECOME PROFICIENT

- ▶ Technical – Scaling solution –
  - ▶ parallel, distributed solutions
  - ▶ Data is moving, make algorithms move to where data is
- ▶ Semantics (age old problem in NLP meaning...)
- ▶ Astronomy – Habitable planets
- ▶ Earthquake/Volcano -- Prediction is hard
- ▶ Stock market movements
- ▶ Brain wave/cyborg territory
  - ▶ Man/machine interface
  - ▶ <https://www.slideshare.net/rk2153/augmented-11022020ieee>

# MANY GRAND CHALLENGES REMAIN

GIVE ME 6 HOURS TO CUT DOWN A TREE  
AND I WILL SPEND THE FIRST FOUR HOURS  
SHARPENING MY AXE  
...LINCOLN

<https://www.goodreads.com/quotes/83633-give-me-six-hours-to-chop-down-a-tree-and>

You have taken the first step toward sharpening your axe!

# PLAN PREPARE PERSIST AND PERSEVERE

- ▶ Thank you
- ▶ Get addicted to lifelong learning
- ▶ [rk1750@nyu.edu](mailto:rk1750@nyu.edu)
- ▶ Raman Kannan
- ▶ <https://www.slideshare.net/rk2153/documents>
- ▶ <https://www.slideshare.net/rk2153/presentations/>