THE STORY OF DATA

Why now?

1

- 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?

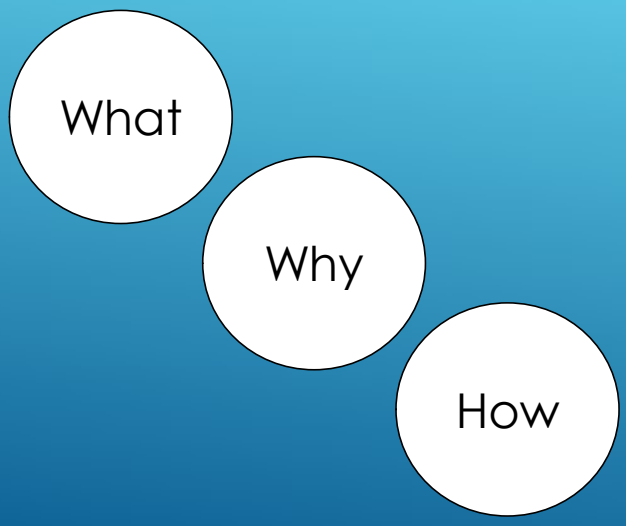
LUCK FAVORS THE PREPARED

2

GENERAL PURPOSE FORMULA FOR LIFE!

3

HOW TO LEARN EFFECTIVELY?



What are we seeking to learn?

Why should we learn?

How to learn?

4

RULE#1: ASK QUESTIONS

- 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/

5

QUESTIONS

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

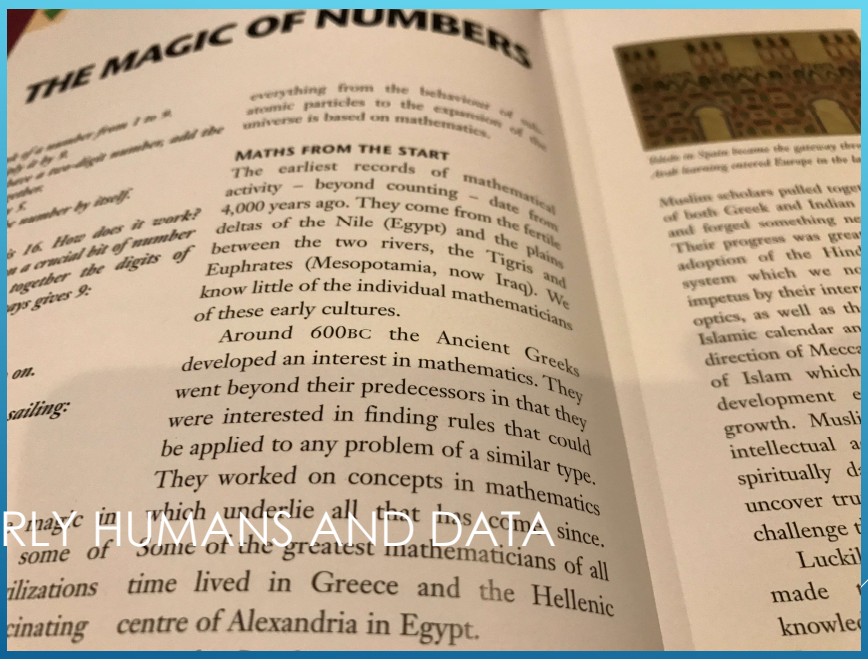
6

WHAT IS DATA?

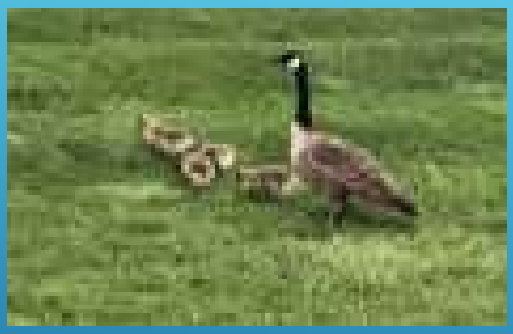
the concept of data as defined in the *IFIP Guide to Concepts and Terms in Data Processing*: "[Data is] a representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process."
--quoted from Forbes

7

EARLY HUMANS AND DATA



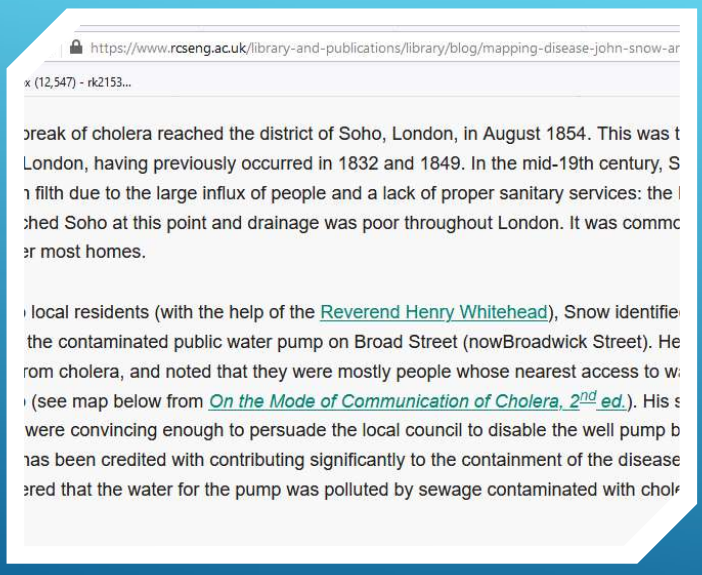
8



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

DATA IS EXISTENTIAL FOR ALL LIFE FORMS

9


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

x (12,547) - rk2153...

reak of cholera reached the district of Soho, London, in August 1854. This was t
London, having previously occurred in 1832 and 1849. In the mid-19th century, S
filth due to the large influx of people and a lack of proper sanitary services: the
ched Soho at this point and drainage was poor throughout London. It was commc
er most homes.

local residents (with the help of the Reverend Henry Whitehead), Snow identifie
the contaminated public water pump on Broad Street (nowBroadwick Street). He
rom cholera, and noted that they were mostly people whose nearest access to w
(see map below from *On the Mode of Communication of Cholera, 2nd ed.*). His s
were convincing enough to persuade the local council to disable the well pump b
has been credited with contributing significantly to the containment of the disease
ered that the water for the pump was polluted by sewage contaminated with chol

JOHN SNOW – 1854 – VISUALIZING DATA

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

10

WEATHER FORECASTING

https://www.iweathernet.com/educational/history-weather-forecasting

Demand Forecasting (Walmart used data to forecast beer/poptart 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

11

WHEN DID YOU LEAVE HOME TO GET HERE TODAY?

- Data and Analysis is at work…all the time…

12

- Types (what kind of values are allowed .. Business rules → range of value)
- – Unstructured/Structured
- – Transactional (Operational)/Fundamental
- – 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

13

- 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

14

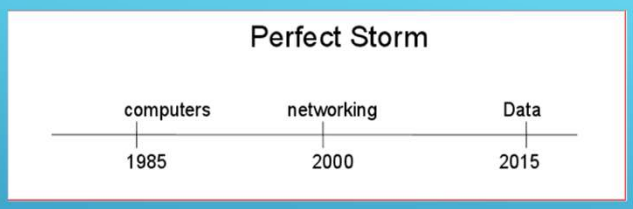
Year	Global Internet Traffic
1992	100 gigabytes per day
1997	100 gigabytes per hour
2002	100 gigabytes per second
2007	2,000 gigabytes per second
2012	12,000 gigabytes per second
2017	35,000 gigabytes per second

Note: 1992 it was per day in 2017 it is per second

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

EXPONENTIAL GROWTH

15



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 quinitillian – 18 zeros
Billion – 9 zeroes
Quintillian – billion billion

Big Data For Dummies by Alteryx

PERFECT STORM – WHY EXP GROWTH?

16


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

WHY?

17

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

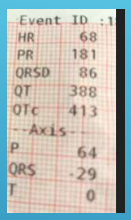
It is relevant to us while importing/analyzing/reporting using
real world data is the focus,
using established/core information management principles, because
we want reproducible and repeatable experiments. One time results are not useful

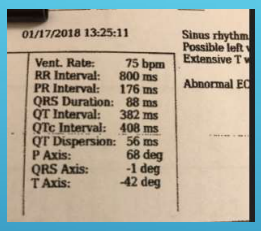
http://www.allanalytics.com/author.asp?doc_id=269883&f_src=AllAnalytics_finalanalysis

HOW

Data --> Information --> Knowledge

18





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

19

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

20

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)

21

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)

22

HOW META-DATA - 03

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

23

SELF DESCRIBING DATA

- 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

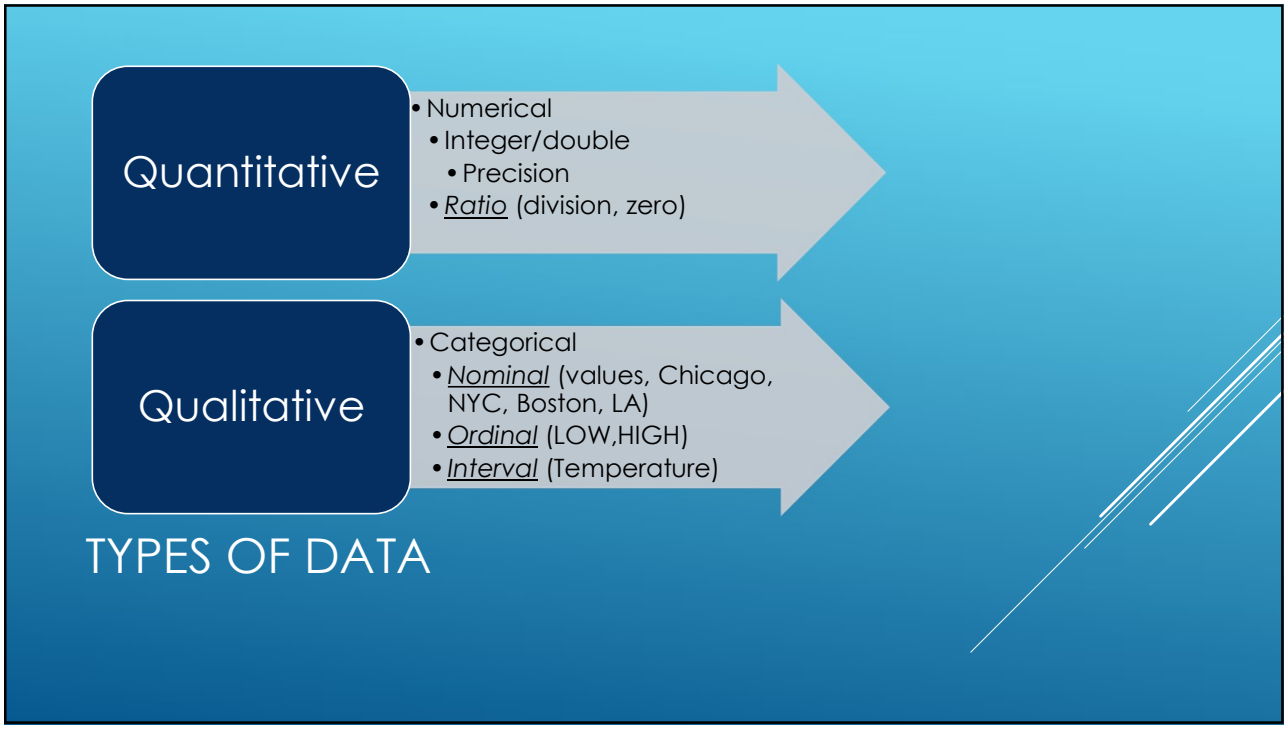
```
<Corporation>
<Symbol>IBM</Symbol>
<YearOfIncorporation>1891</YearOfIncorporation>
<NumberOfEmployees>330440</NumberOfEmployees>
<NumberOfLocations>435</NumberOfLocations>
</Corporation>
```

24

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

CONSISTENT MEANING

25



TYPES OF DATA

26

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

27

- Weather data has always been volumonous – 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.3. **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…

28

- Images
- Audio
- video

+

Human generated content (emails/blogs etc)

Aka unstructured data

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

29

https://www.nyse.com/data/transactions-statistics-data-library
http://www.nyxdata.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

30

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

31

*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.csail.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].*[23] *To [infer] how nature is associating the response variables to the input variables.*

DATA SCIENCE/ANALYTICS

32

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

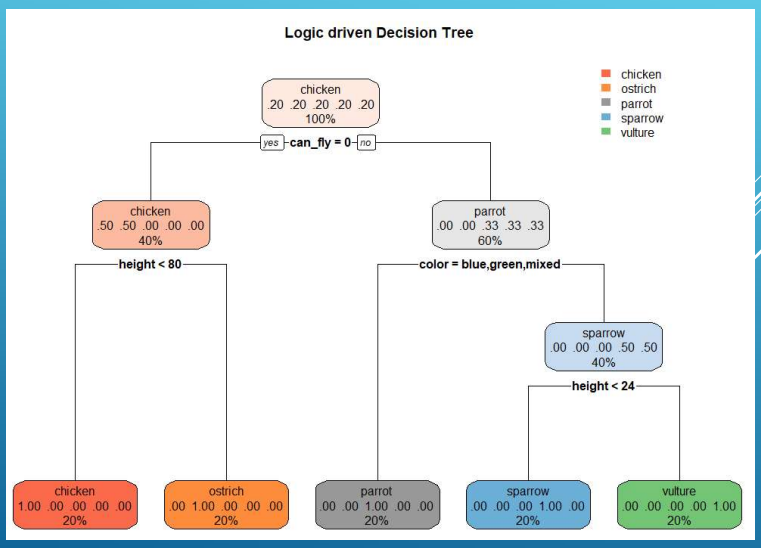
33

```
> cdset<-rbind(birds2,data.frame(nlegs=rep(2,5),can_fly=rep(0,5), height=hchick$
+ 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))
+ )
> cdset<-rbind(birds2,data.frame(nlegs=rep(2,5),can_fly=rep(0,5), height=hchick$
+ 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))
+ )
> cdset
  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')
>
```

34

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



35

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[1]—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

36

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

BE WHERE THE PUCK WILL BE! W.G.

37

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

38

- 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

39

- 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

MANY GRAND CHALLENGES REMAIN

40

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!

41

PLAN
PREPARE
PERSIST AND PERSEVERE

42