THE STORY OF DATA

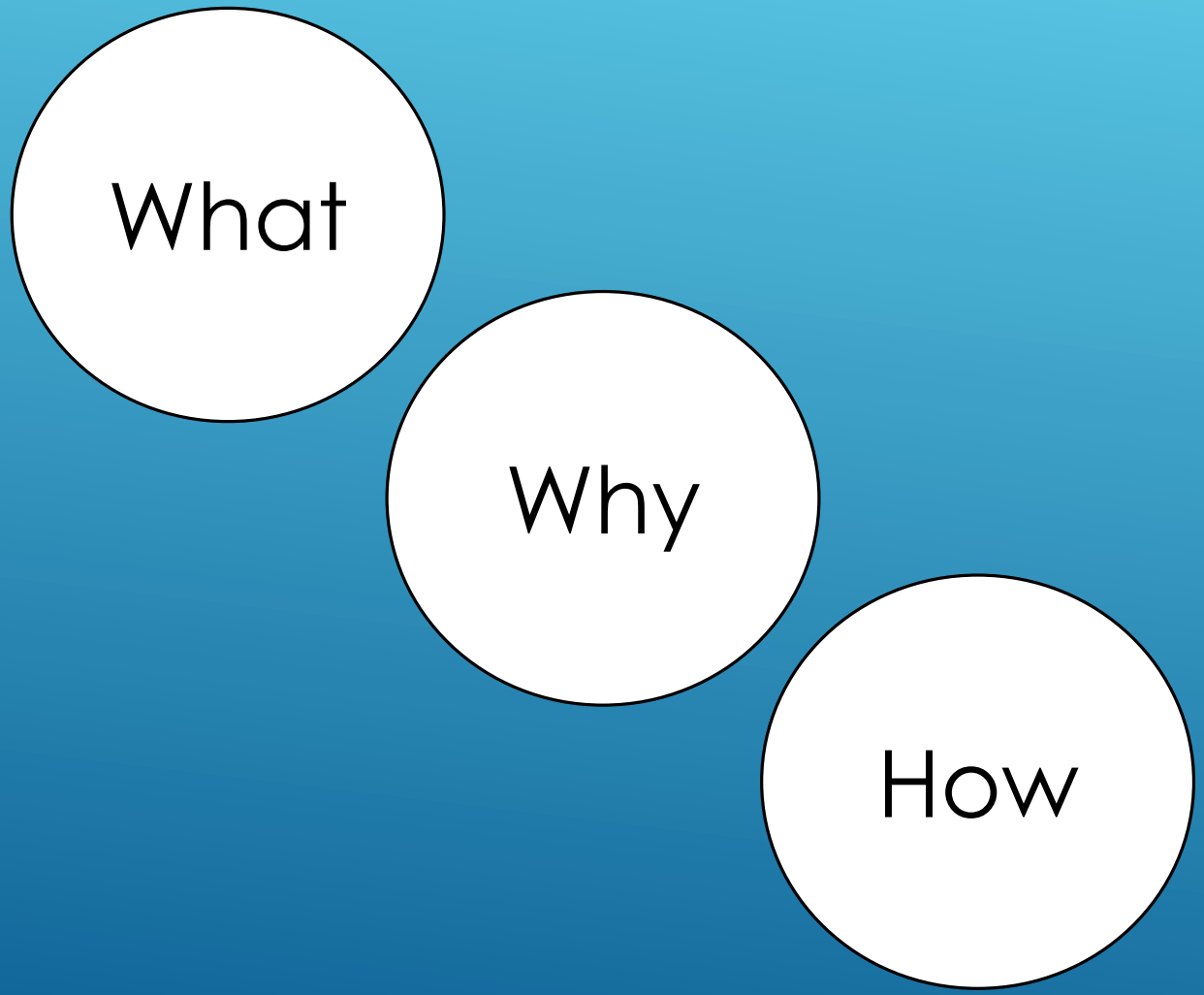
Why now?

- 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 are we seeking to learn?

Why should we learn?

How to learn?

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/

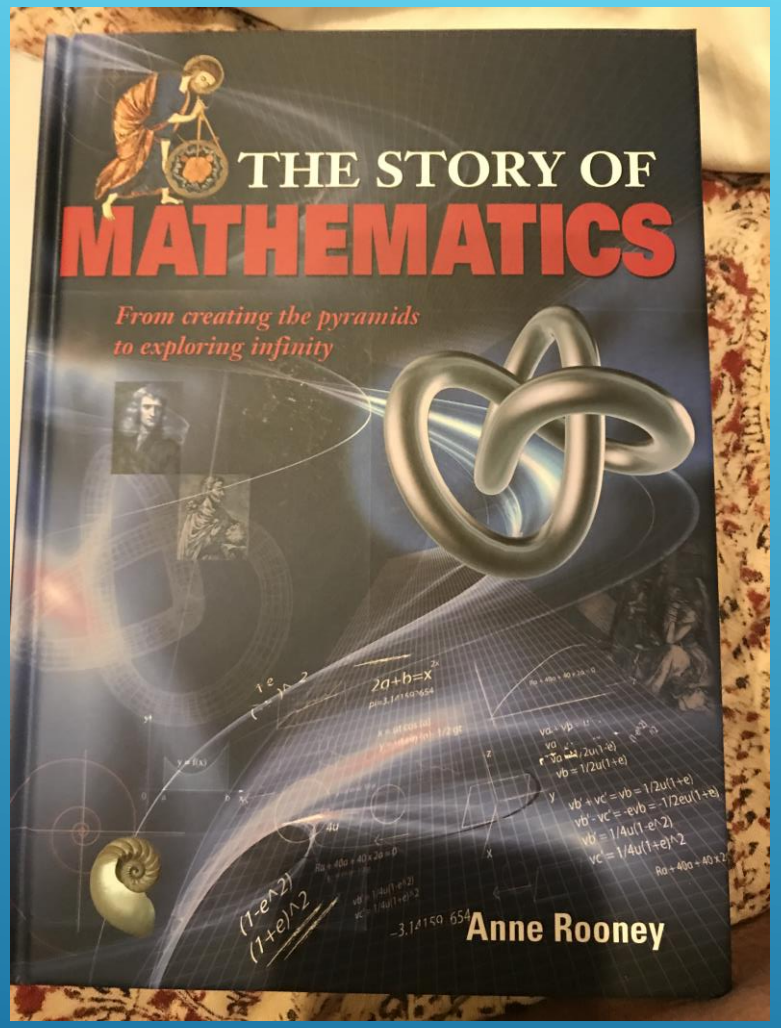
- What is data?
- Why do we care about data?
- Why now?
- How can I position myself?

QUESTIONS

WHAT IS DATA?

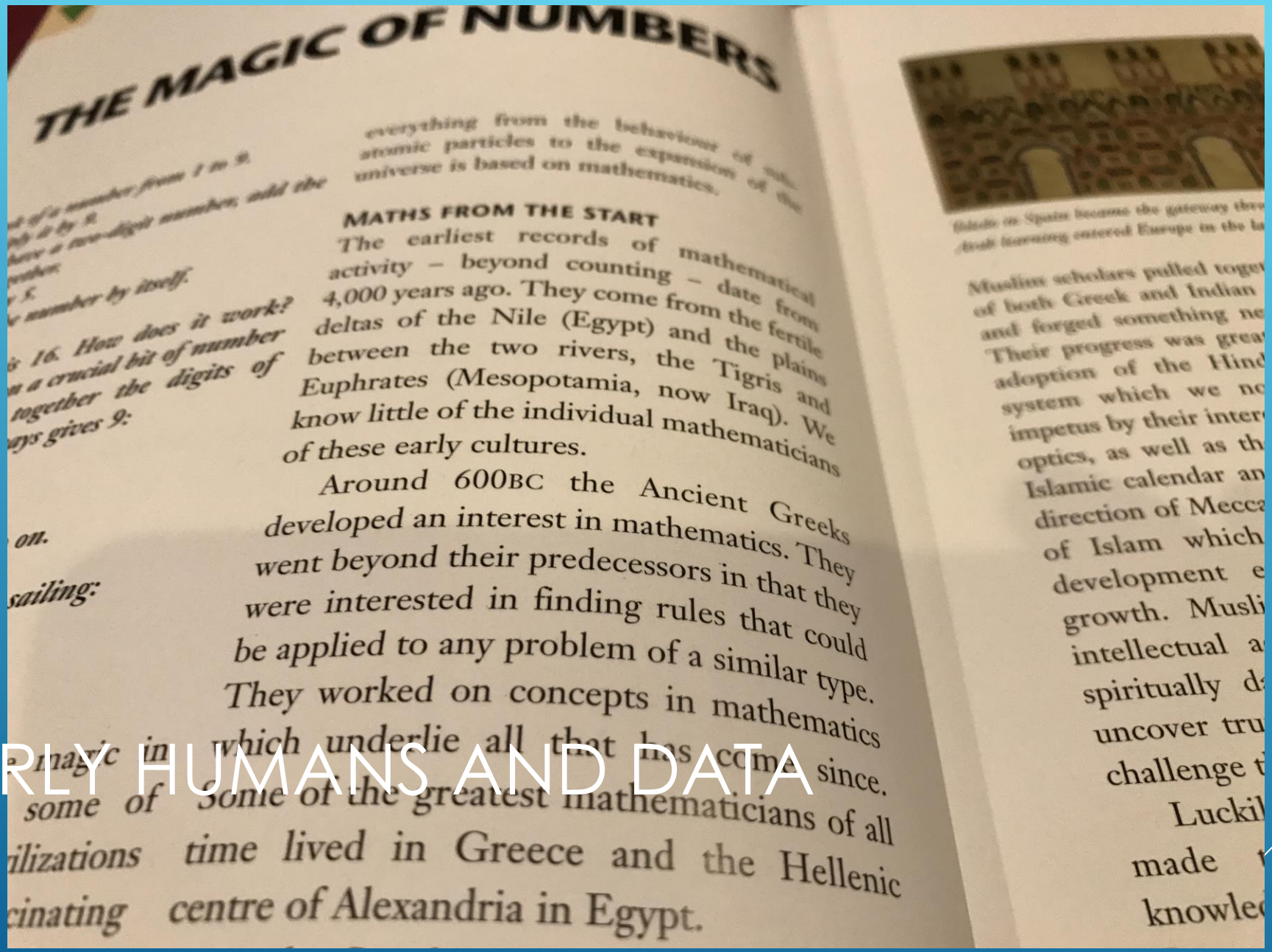
the concept of data as defined in the *IFIP Guide to Concepts and Terms in Data Processing*: "[Data is] a representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process."

--quoted from Forbes



HIGHLY RECOMMENDED

EARLY HUMANS AND DATA



THE MAGIC OF NUMBERS

everything from the behaviour of sub-atomic particles to the expansion of the universe is based on mathematics.

number from 1 to 9.
by 9.
a two-digit number, add the
number by itself.
16. How does it work?
a crucial bit of number
together the digits of
ays gives 9:
on.
sailing:

MATHS FROM THE START

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 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.

magic in
some of
ilizations
cinating

Some of the greatest mathematicians of all time lived in Greece and the Hellenic centre of Alexandria in Egypt.

Muslim scholars pulled toge
of both Greek and Indian
and forged something ne
Their progress was grea
adoption of the Hind
system which we n
impetus by their inter
optics, as well as th
Islamic calendar an
direction of Mecca
of Islam which
development e
growth. Musli
intellectual a
spiritually d
uncover tru
challenge t
Luckil
made
knowle

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

DATA IS EXISTENTIAL FOR ALL LIFE FORMS

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

WHEN DID YOU LEAVE HOME TO GET HERE TODAY?

WHAT TYPES OF DATA

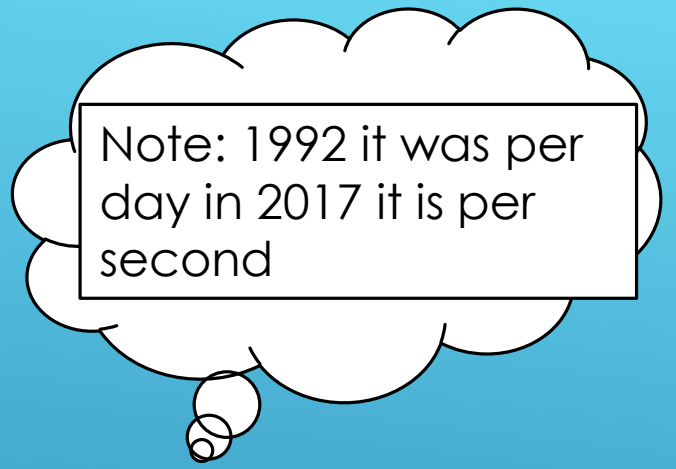
- Types (what kind of values are allowed .. Business rules → range of value)
- – Unstructured/Structured
- – Transactional (Operational)/Fundamental
- – Hierarchical/Network/Relational Data
- Another Slice (Enterprise Data Management)
- – Master
- – Metadata
- – Reference
- http://msdn.microsoft.com/en-us/library/bb190163.aspx

STRUCTURED/UNSTRUCTURED

- 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”

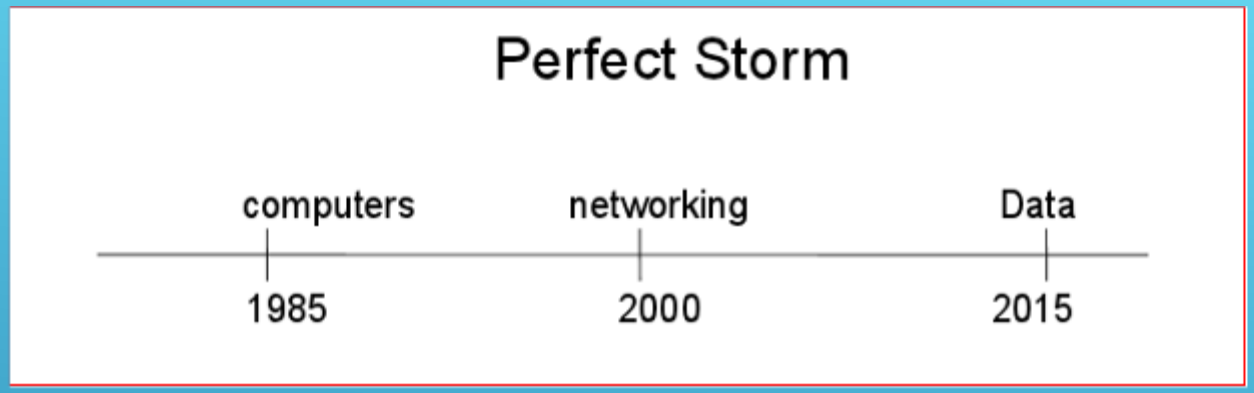
EXPONENTIAL GROWTH

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?

PERFECT STORM – WHY EXP GROWTH?

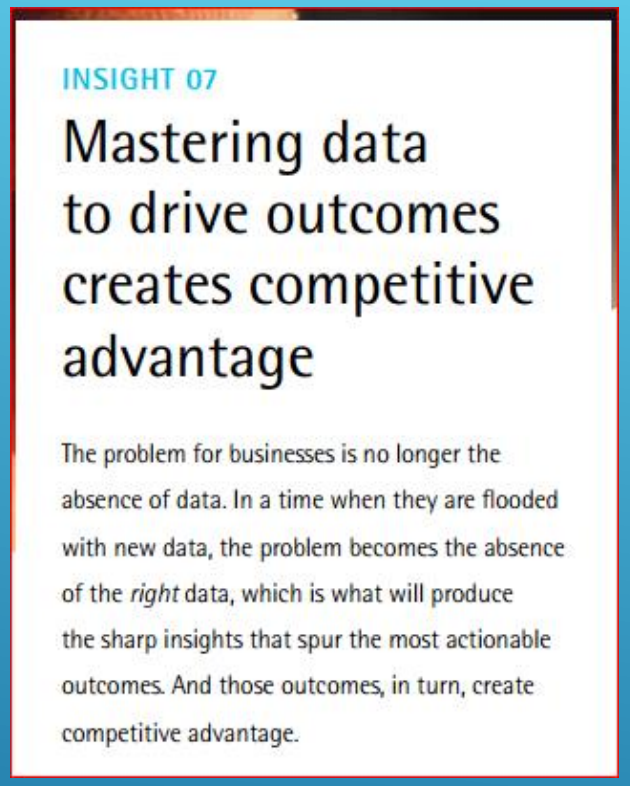


We are now in the zone – hockey stick

- 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



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

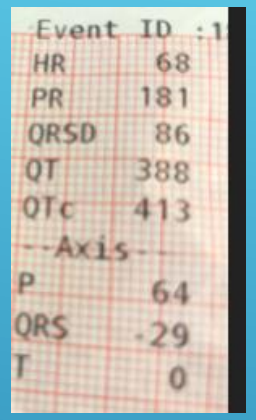
importing data (finding sources, exploring, refining/cleansing data)
Analyzing data (modeling,extracting patterns, knowledge)
Reporting explaining what was done (explaining to the world around)

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

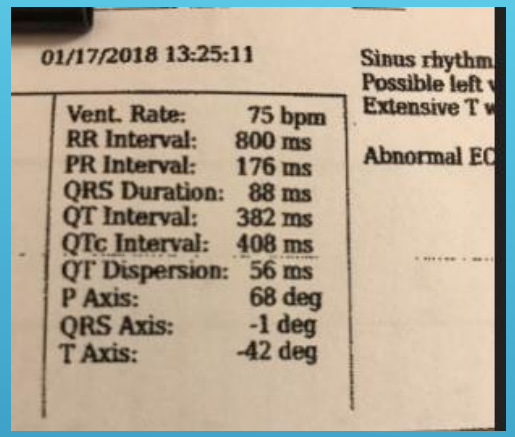
http://www.allanalytics.com/author.asp?doc_id=269883&f_src=AllAnalytics_finalanalysis

HOW

Data Information Knowledge



Data

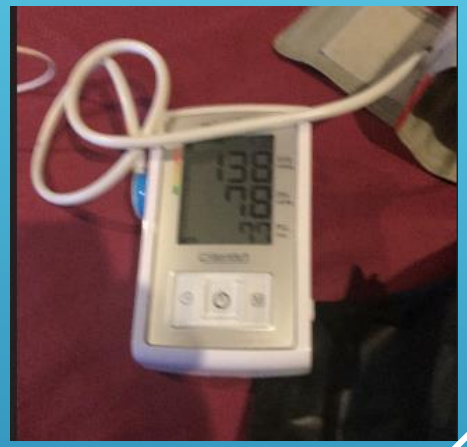


Context

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

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>
```

CONSISTENT MEANING

- 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…

TYPES OF DATA

Quantitative

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

Qualitative

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

I cnduo't bvleiee taht I culod aulaclty uesdtannrd waht I was rdnaieg. Unisg the icndeblire pweor of the hmuan mnid, aocdcrnig to rseecrah at Cmabrigde Uinervtisy, it dseno't mttaer in waht oderr the lterets in a wrod are, the olny irpoamtnt tihng is taht the frsit and lsat ltteer be in the rhgit pclae. The rset can be a taotl mses and you can sitll raed it whoutit a pboerlm. Tihs is bucseae the huamn mnid deos not raed ervey ltteer by istlef, but the wrod as a wlohe. Aaznmig, huh? Yaeh and I awlyas tghhuot slelinpg was ipmorantt! See if yuor fdreins can raed tihs too.

Can you read this?
How old are you? Do you think a 4-6 year old can read ?
What happened?
https://www.ecenglish.com/learnenglish/lessons/can-you-read

LET US LOOK AT EXAMPLES (NOT SO OBVIOUS)

BIG DATA: THE NEW KID

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?

- 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…

VARIETY IS NEW

- 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!

https://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/

[[https://www.bing.com/search?q=proportion+of+unstructured+to+structured+data]]
[[https://sherpasoftware.com/blog/structured-and-unstructured-data-what-is-it/]]

Unstructured data is raw and unorganized and organizations store it all. Ideally, all of this information would be converted into structured data however, this would be costly and time consuming. Also, not all types of unstructured data can easily be converted into a structured model. For example, an email holds information such as the time sent, subject, and sender (all uniform fields), but the content of the message is not so easily broken down and categorized. This can introduce some compatibility issues with the structure of a relational database system.

Social Media Posts
Looking at the list, you may be wondering what these files have in common. The files listed above can be stored and managed without the format of the file being understood by the system. This allows them to be stored in an unstructured fashion because the contents of the files are unorganized.

UNSTRUCTURED 01

UNSTRUCUTRED 02

- In case you're still not quite sure what we mean, here is a limited list of types of unstructured data:

- Emails
- Word Processing Files
- PDF files
- Spreadsheets
- Digital Images
- Video
- Audio

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.

WHAT: 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.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.*

WHAT: DATA SCIENCE/ANALYTICS

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

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					

WHY: INDUSTRY –APPLICATIONS

Here we come Homo-Connexus

TECH

Smartphone Use while Walking Is Painfully Dumb

Distracted walking is the new hip reason for an ER trip

SCIENTIFIC AMERICAN

MOVE OVER HOMO-SAPIENS

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

TO BE CONTINUED

You have taken the first step toward sharpening your axe!