

Good Data; Bad Information

Why high quality raw data does not necessarily result in
good information and decisions

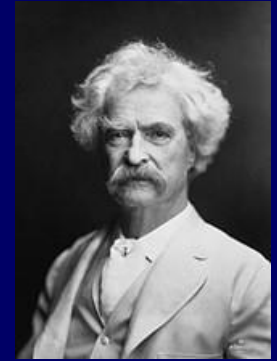
By Michael Scofield, M.B.A.

Asst. Professor, Loma Linda University

Vers. 38 Feb. 23, 2023

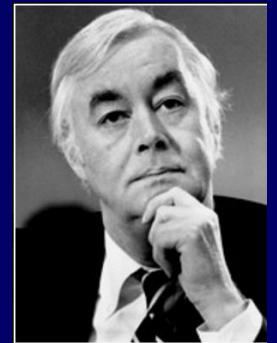
“The researches of many commentators have already thrown much darkness on this subject, and it is probable that if they continue, we shall soon know nothing at all about it.”

--Mark Twain



“You are entitled to your own opinions. But you are not to your own facts.”

-- Daniel Patrick Moynihan

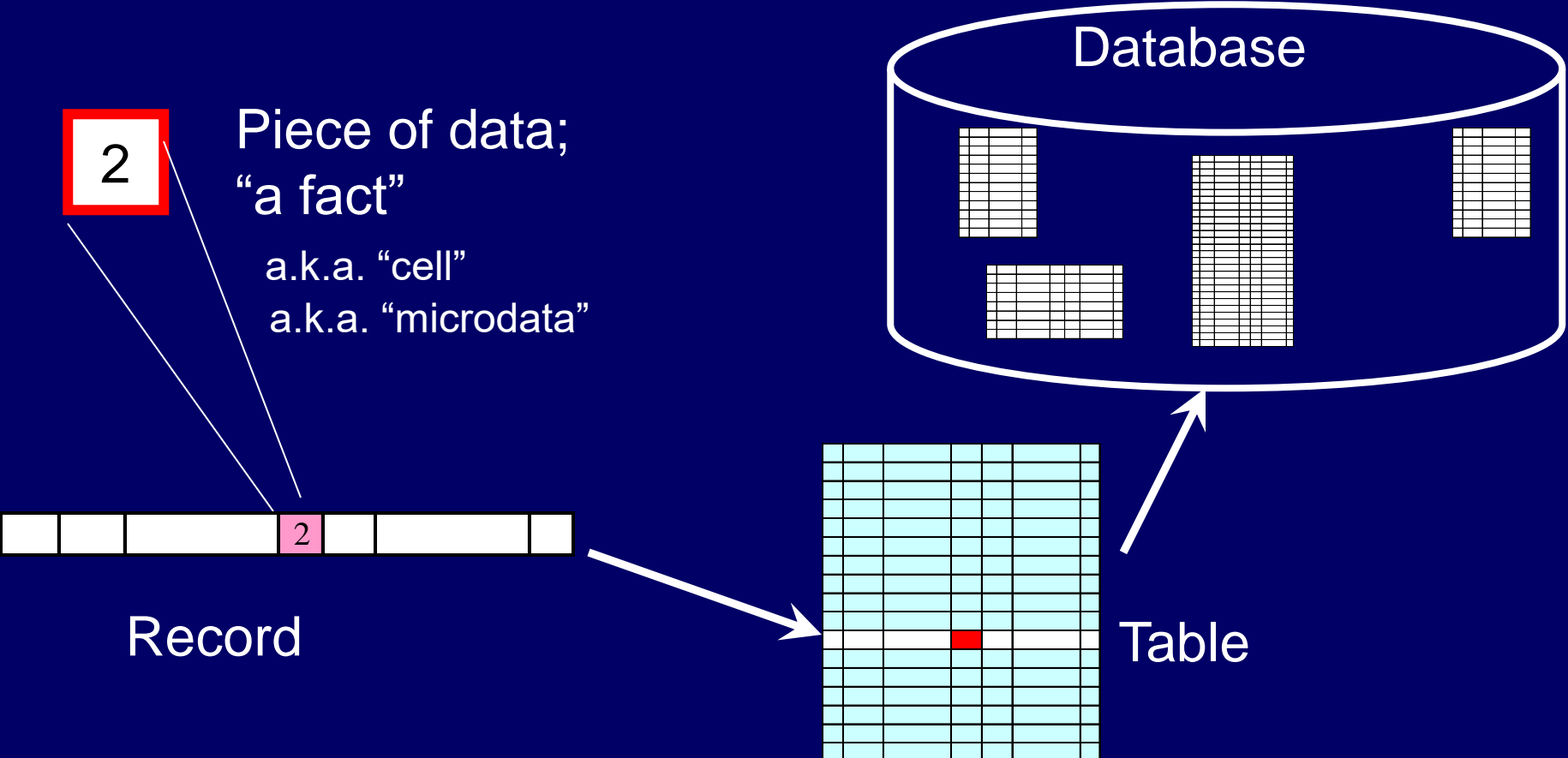


“Without data, you’re just another person with an opinion”.

--W. Edwards Deming

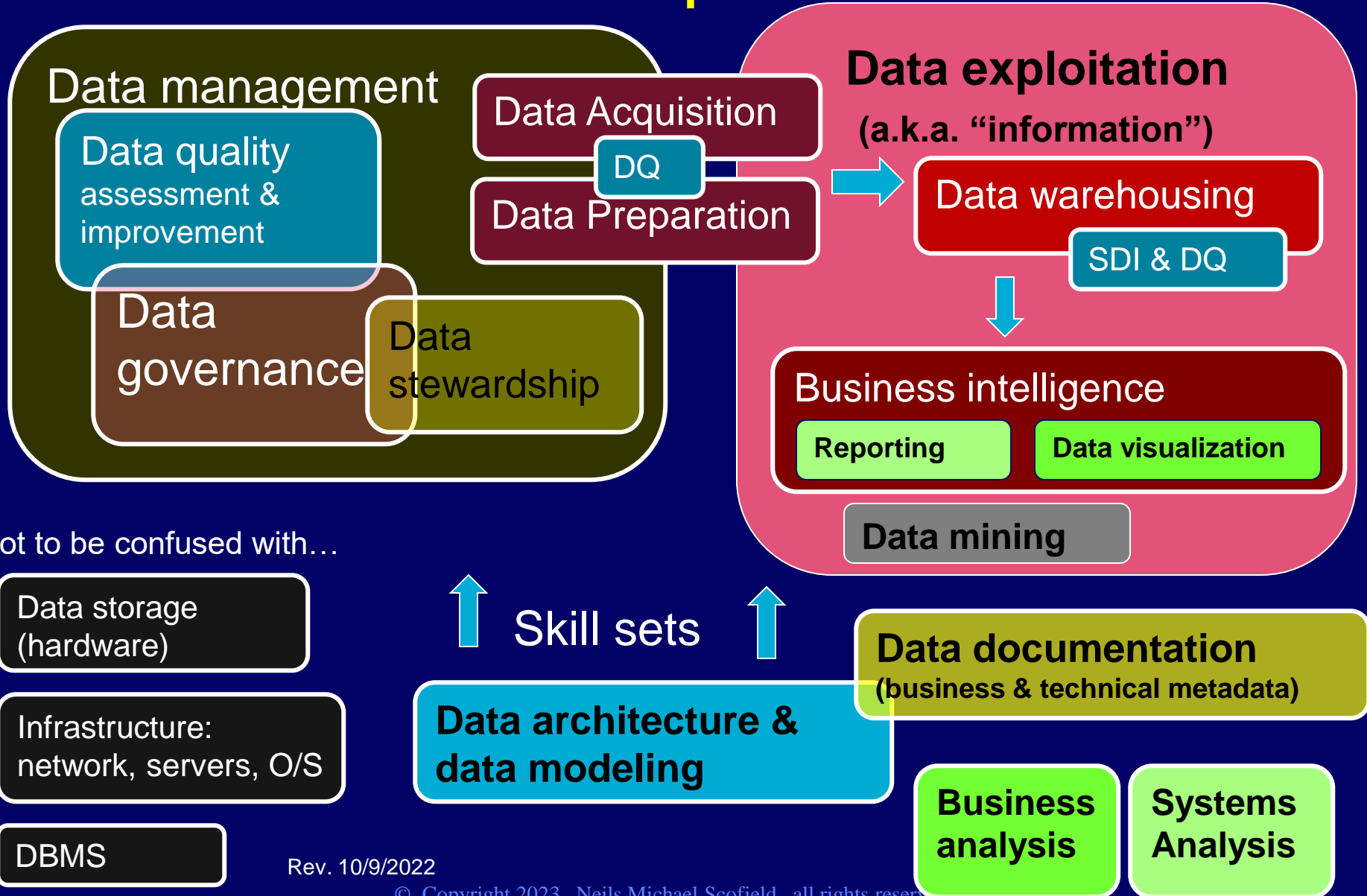


Structural elements of tabular “data”



When we say “data”, what do we mean?
A fact, a record, a table, or a database?

Data disciplines



Rev. 10/9/2022

© Copyright 2023 Neils Michael Scofield all rights reserved

Reality



Facts &
data



Information



Expression



Communication



Understanding
(of meaning)

Raw data



Information

Elemental or granular.

Requires logical keys.

When, where, etc.

Multiple observations
need to be consistent.

Human reader-oriented.

More meaningful.

Expressed in text and/or
graphics.

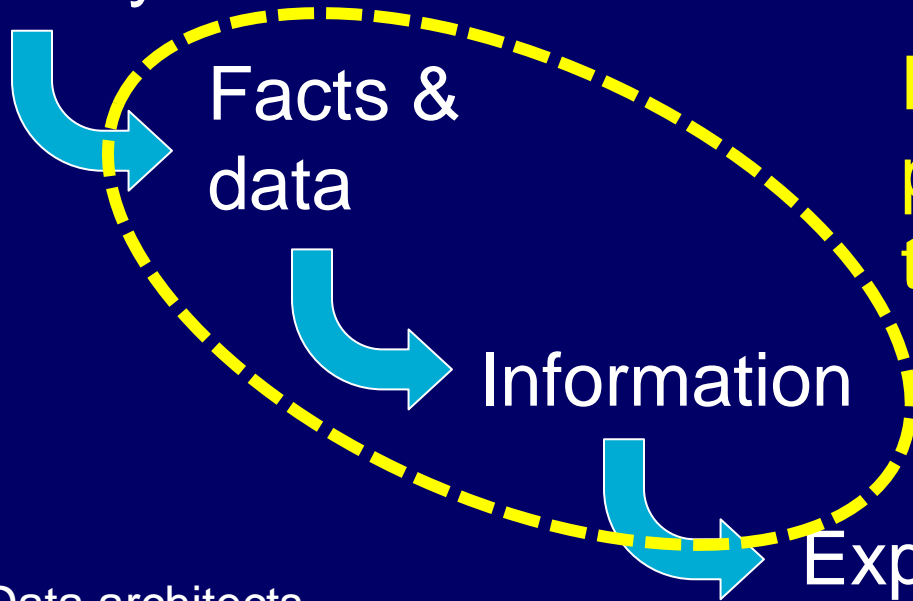
Often integrated

Often normalized

Often with interpretive bias.

Data-to-decision supply chain

Reality



Data management people focus mainly on these two.

Data architects
Database administrators
Data scientists
Master data management
Etc.



Communication



Understanding

(of meaning)

Path to executive decision-making

Reality



Facts &
data



Information



Expression



Communication



Understanding

(of meaning)

Knowledge workers
Business analysts
Marketing analysts
Process analysts
Etc.

Supporting executive decisions requires focus upon these parts of the chain.

Reality



Facts & data



Information



Expression



Communication



Understanding
(of meaning)

Impressions,
observations,
measurements

Human senses or devices
(mechanical sensors)

Facts: primary or second hand?

Granular data should accurately describe
reality...

Reality:
objective
verifiable

Models are not reality...
...but they help explain reality.

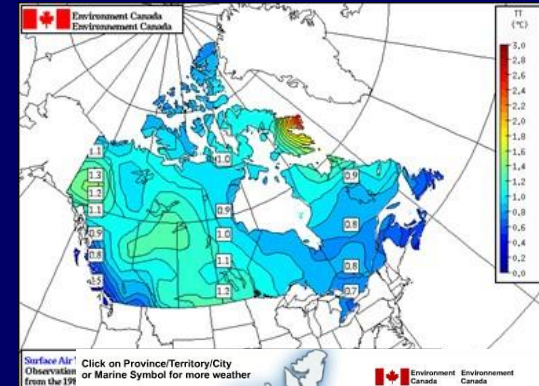
Maps help us understand and navigate reality, but
they should not be confused with reality.

Reality

Facts & data

Many observations:
wind
temperature
humidity

Basic wx map



Information

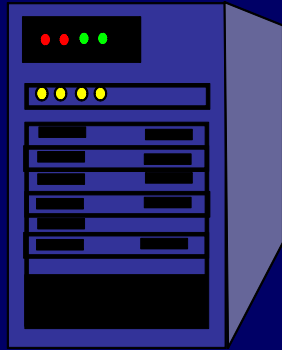
Expression



Map w/icons, symbols

Communication

Understanding
(of meaning)



Supercomputer

Tuesday

Tuesday
Night

Wednesday

Wednesday
Night

Thursday



Sunny

Chance Rain

Mostly Sunny
then Chance
Rain

Rain Likely

Rain Likely

High: 65 °F

Low: 36 °F

High: 51 °F

Low: 36 °F

High: 45 °F

Symbols / icons

Detailed Forecast

Today Sunny, with a high near 68. Calm wind becoming south around 5 mph in afternoon.

Tonight Partly cloudy, with a low around 44. West wind around 5 mph becoming

Human-readable
text

Washington's Birthday Sunny, with a high near 69. Calm wind becoming west around 5 mph in the afternoon.

Monday Night Mostly clear, with a low around 45. Calm wind becoming southeast around 5 mph after midnight.

Tuesday Sunny, with a high near 65. Southeast wind 5 to 10 mph becoming west in the afternoon.



Tuesday

Weather Bureau

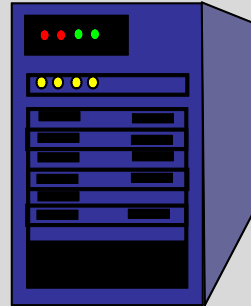
Wednesday
Night

Thursday



Meteorologist with advanced degree.

Text-generating software



Not necessarily A.I.



Rain Likely

High: 45 °F

Detailed Forecast

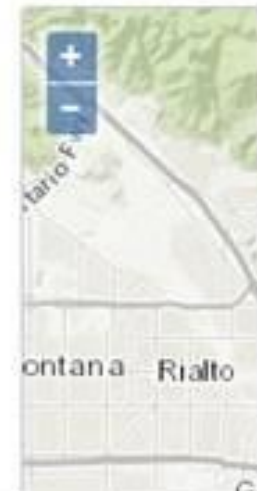
Today Sunny, with a high near 68. Calm wind becoming south around 5 mph in the afternoon.

Tonight Partly cloudy, with a low around 44. West wind around 5 mph becoming calm.

Washington's Birthday Sunny, with a high near 69. Calm wind becoming west around 5 mph in the afternoon.

Monday Night Mostly clear, with a low around 45. Calm wind becoming southeast around 5 mph after midnight.

Tuesday Sunny, with a high near 65. Southeast wind 5 to 10 mph becoming west in the afternoon.



BBC weather for London

Expression !

Today



13°
7°

Sunny and a gentle breeze

Mon 20th



14°
7°

Tue 21st



13°
6°

Wed 22nd



10°
3°

Thu 23rd



9°
1°

Fri 24th



1700	1800	1900	2000	2100	2200	2300	0000 Mon	0100	0200	0300	0400	0500
11°	10°	9°	9°	8°	8°	8°	8°	8°	8°	8°	8°	8°
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
9	9	10	11	11	11	11	11	12	12	12	12	12

BBC weather for London

Standard icons / symbols

Simple

Well-known to audience

Expression !



Today



13°
7°

Sunny and a gentle breeze

Mon 20th



14°
7°

Tue 21st



13°
6°

Wed 22nd



10°
3°

Thu 23rd



9°
1°

Fri 24th



1700	1800	1900	2000	2100	2200	2300	0000 Mon	0100	0200	0300	0400	0500
11°	10°	9°	9°	8°	8°	8°	8°	8°	8°	8°	8°	8°
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
9	9	10	11	11	11	11	11	12	12	12	12	12

Kinds of expression

Information



Expression



Communication



Understanding
(of meaning)

Her Northridge house did suffer considerable damage from the 1994 earthquake. Her brick chimney came crashing down through the roof into the den. I got to see the damage when I was up there shortly thereafter.

In 1994, Lela (single at the time) asked me to be her "date" to a wedding in South Pasadena. At this time, not before she had met the young couple at their company. Lela and I went for lunch in Glendale on July 17 before driving over to the Holy Smith Pasadena United Methodist Church just off Monterey Road. This was during world cup football at the Rose Bowl. It was not a large wedding. Looking after the ceremony, they wanted everyone to gather for group photos at the front of the church. So, in some hotel, somewhere in America is my likeness in their wedding photographs.

There was also a young couple at our table who had an interesting story, as here they met. They were both children of missionary parents and had played together as children. The girl went back to England with her parents, and the boy to America. After he graduated from college, he was going to take a tour of Europe. His parents managed him to look up the girl while he was in England. He did. They let it off, and much to their surprise, they fell in love.

Late in life, Lela purchased a larger house in West Hills. She showed me a king chair in the upper landing of the stairwell which she loved for a variety of business enterprises. Her son assisted in maintaining them. I thought this was very sensitive.

David L. Smith

Text

sales division	% sales to minorities
NORTHEAST	12.3
SOUTHEAST	39.1
MIDWEST	21.3
SOUTHWEST	17.6
PACIFIC	14.9
TOTAL U. S.	20.8

Quantitative
tabular



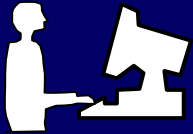
Quantitative
graphic

Student paper

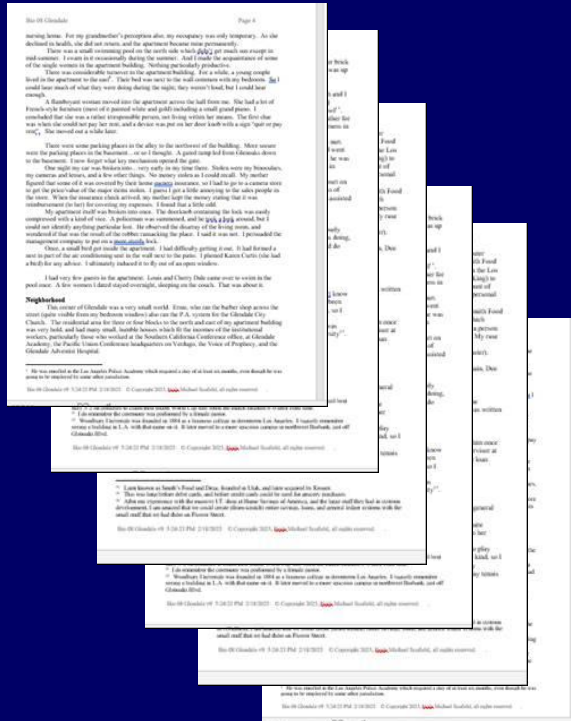
Textual expression of ideas.

Very little raw, quantitative data points

Traditional DQ tests do not work here.



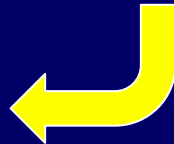
Student



Senior professor ran it through "Turn-it-In".

Came up 48% plagiarized.

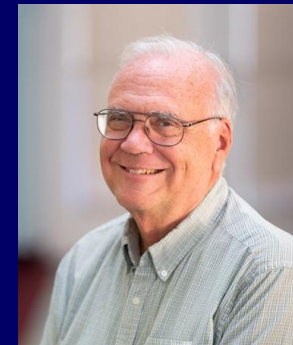
Software compared text, not its meaning!



Fails the "smell test".
Intuition: This is not a student's writing!

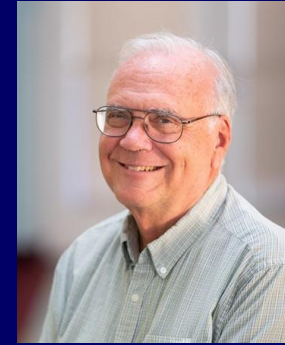


Tedious reading of text.



Tired underpaid teacher

Google search of 9 consecutive words in quotes.



Google found a match !

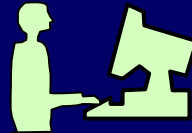
Turn-it-In probably uses brute force to find matches in text strings.

What if you stole the ideas, but expressed in different text?

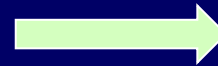
Can A.I. match ideas?

Text and words (expression) can be ambiguous and slippery.

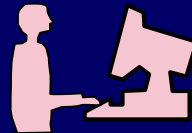
Russian text



Translator #1



One English version



Translator #2



Second (different) English version



Translator #3



Third (different) English version

Why?

Single words can have multiple meanings.

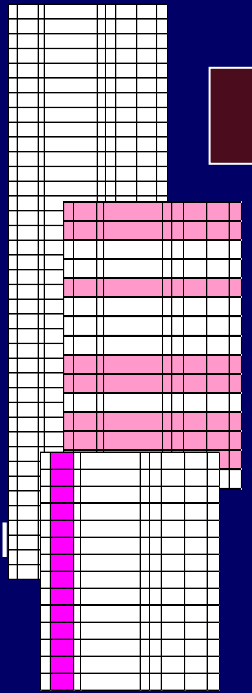
Ideas (collections of words) can have variety of expression.

Reality



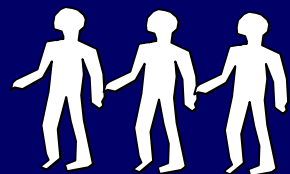
Observational mechanisms.
Human or automated

Raw data describing reality

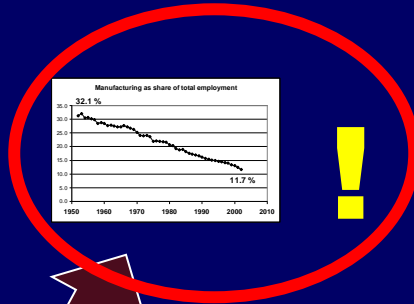


Flat files,
Spreadsheets,
databases

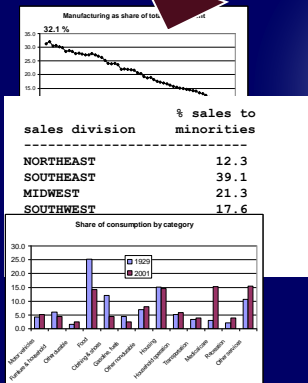
Data mining Data visualization



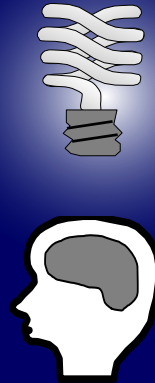
Operational management



Data query tools



Business analysis



Cultural
bubble.



Executive
decision-
makers

Awareness of
market (demand)
beyond your
cognitive horizon?

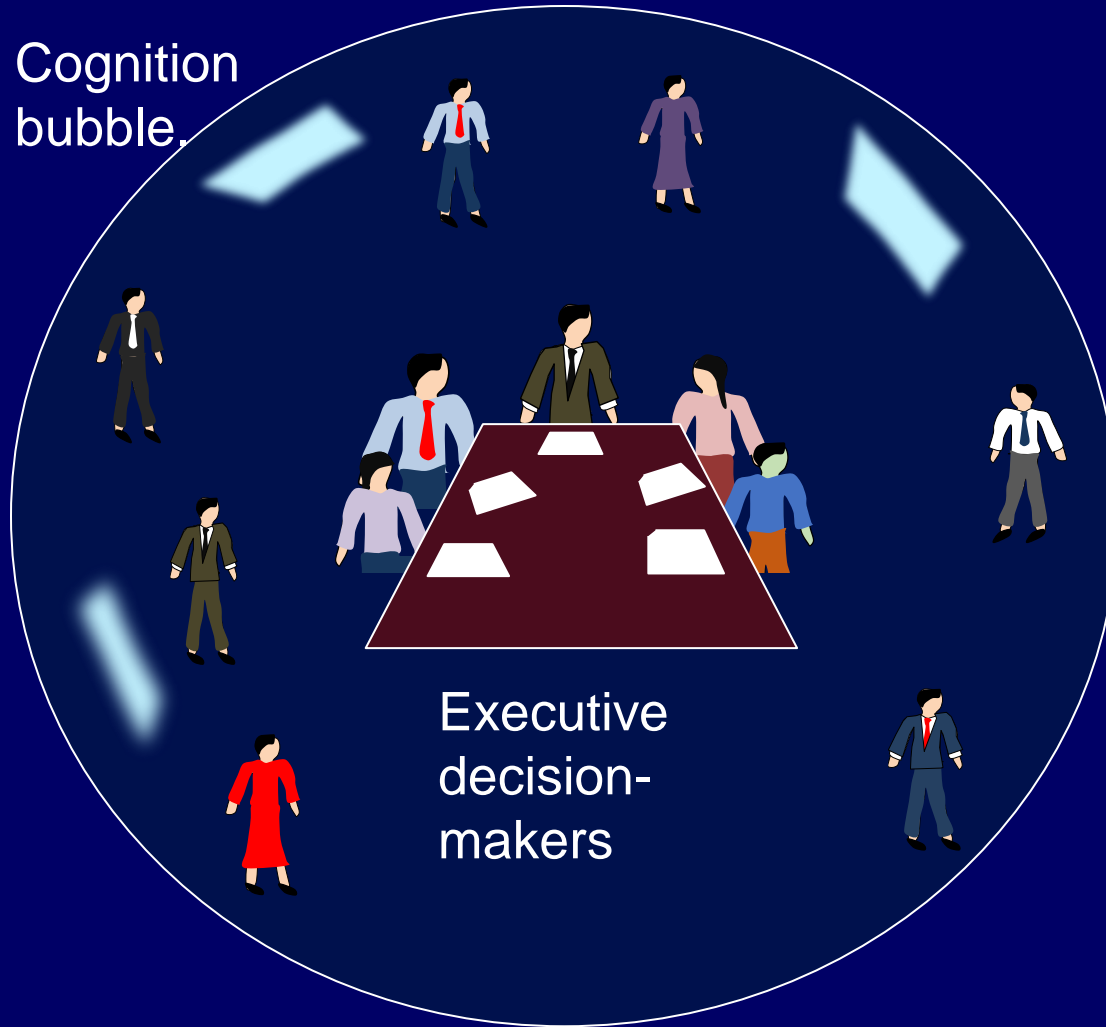
Awareness of
competition?

Awareness of
changing
technology?

(Al Crosson)

What is their cognitive horizon ?

Cognition
bubble.



Executive
decision-
makers

The “little people”

There are smart,
knowledgeable people
on the edge of your
communication bubble.

They know stuff you
don't know.

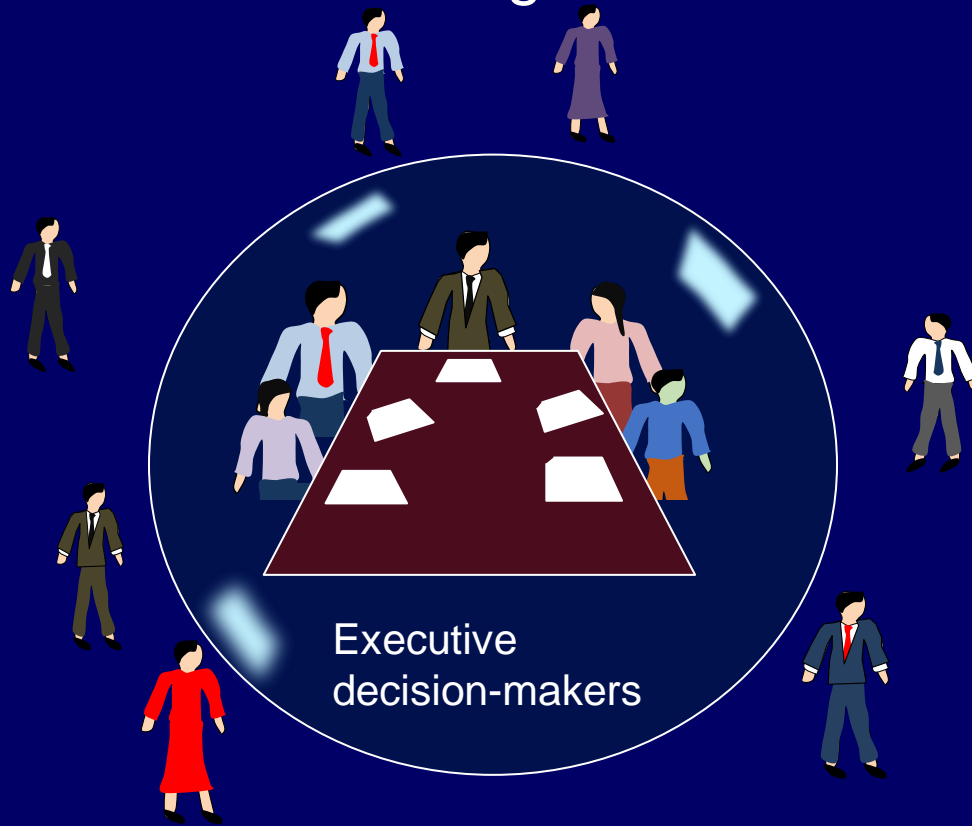
W. Edwards Deming
says, “Listen to them”.

Seek them out !

But if an executive is
insecure, he won't
reveal his ignorance.

What happens in a pandemic isolation?

The bubble gets smaller!



All these people get excluded.

Executives lack casual contact with them.

Executives run risk of making bad decisions.

You don't know what you don't know !

The delusion of high executive compensation.

You think you are worth it !

You don't need to listen to people closer to the business.

They know stuff you don't know !



What causes bad business decisions?

Inadequate model of reality – assumptions & paradigm

Cognition
bubble.



Vladimir
Putin

Sycophants and “yes” men.

Bad assumptions:

Ukrainians will welcome us
with open arms

Russian military is well
prepared and effective.

The war will be over quickly.

Russian cost will be low.

The West will not respond
(as in Syria)

Oil and gas revenue will
continue to flow into Russia.

**A competent executive
isn't afraid to hear
negative information.**

What causes bad business decisions?

Inadequate model of reality – assumptions & paradigm

Absence of adequate data

- Didn't seek it out; didn't know it existed; too costly

- Inability to resolve conflicting data

- Focus on anecdotes

Failure to understand the weakness of the data available

- Inadequate metadata

Bias in the raw data (known or unknown)

Data converted to information incorrectly

- Bad expression of derived data and information

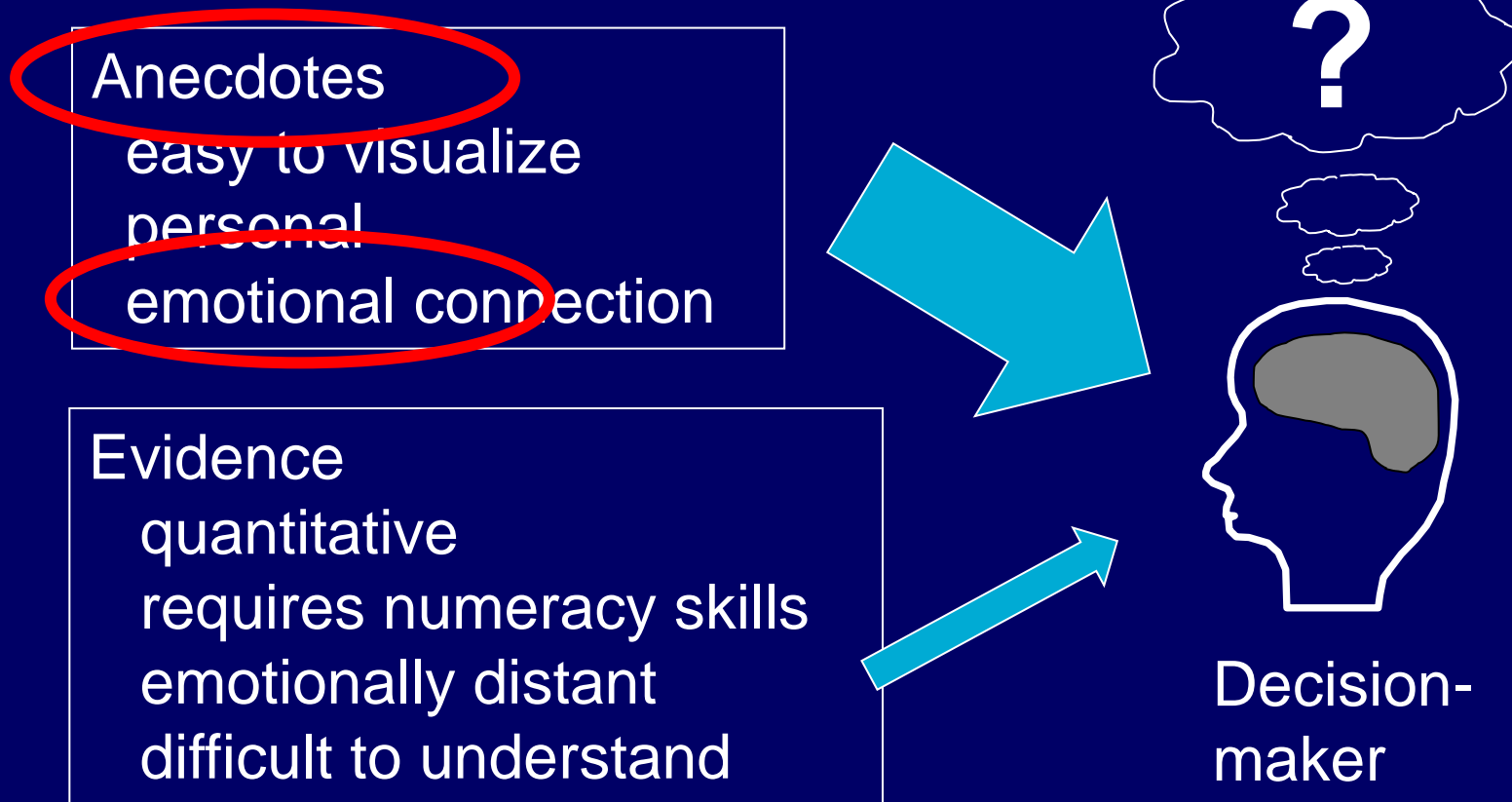
- Bad techniques or deliberate bias in expression

- Inadequate context leads to misinterpretation.

What causes bad business decisions? (cont.)

Decision-maker pre-conceived ideas.

Decision-maker preference for easy-to-understand anecdotes.



What causes bad business decisions? (cont.)

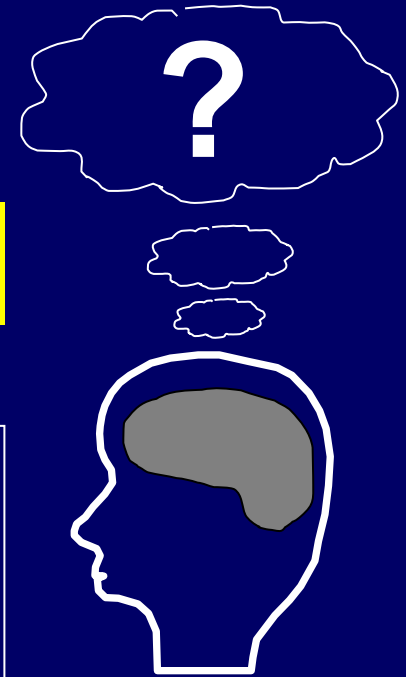
Reality is always more complex than we want !

Many readers / decision-makers don't have the patience to understand complex problems,

...or simple problems with complex causes.

Reality can be complex !

The best information in the world is useless if the decision-maker can't grasp / understand it.



Decision-maker

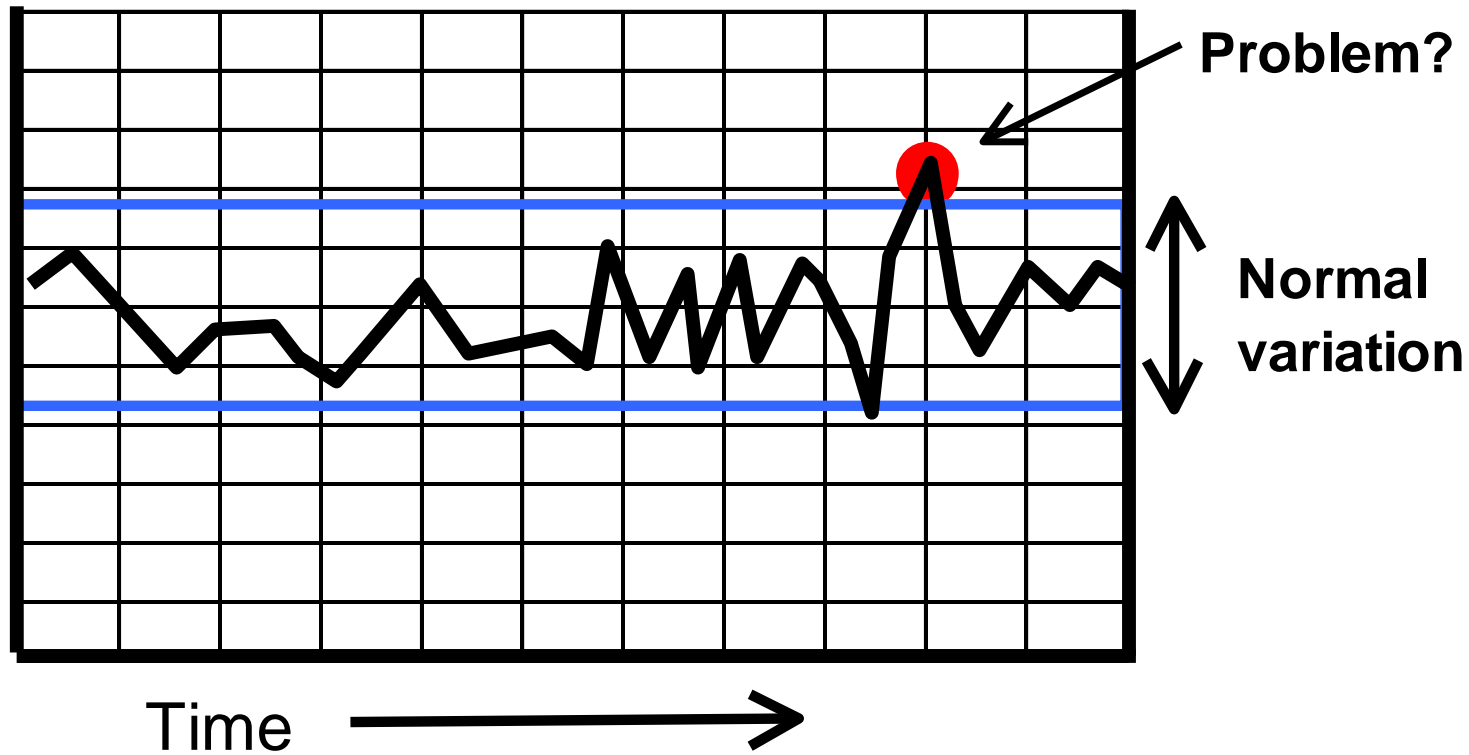
3-hour topic

Sometimes...

...low-level data anomalies

require high level response

Automated surveillance should ignore common causes (statistical noise) but signal when an outlier occurs on the metric.



Not all threats are external.

What behavior anomalies are lurking in your daily operations?



All accused of flooding 2 Ohio counties with addictive opioids.

Convicted; significant penalties expected.

Opioid epidemic cost > \$1 billion for many counties.

Where is raw (granular) data created?

Conscious human activity

- Deliberate business actions (e.g. purchase)

 - Equities, currency, real estate, etc.

- Secondary business behavior (e.g. browsing, product movement)

- Categorizing the world (reference tables, dimensions)

Natural world

- Seismic, weather, pollution, animal behavior, etc.

- Biological census (animals, flora, etc.)

Non-business human behavior

- Media viewership (radio, TV, web)

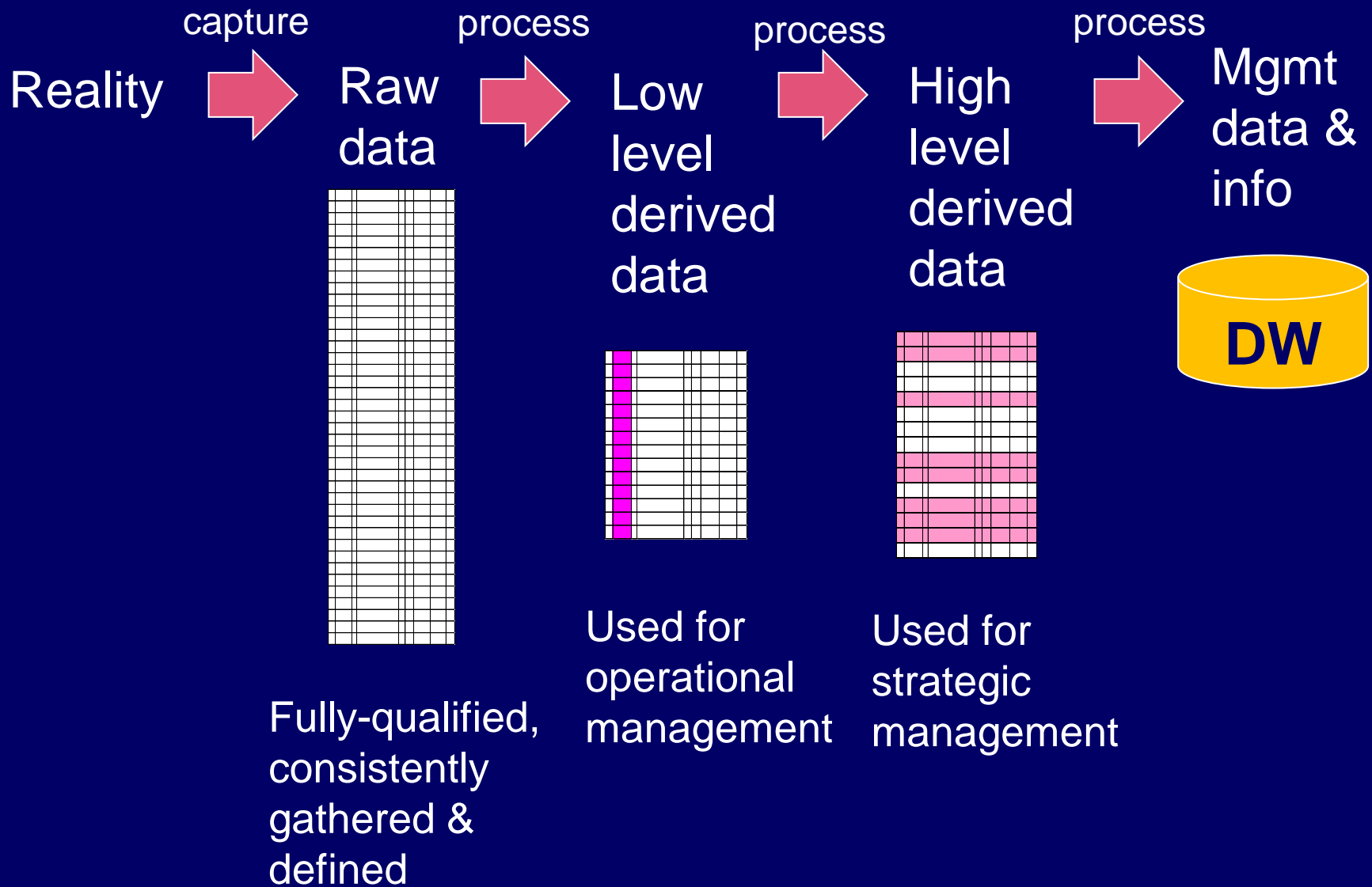
- Mobility (vehicle movement, etc.)

- Surveillance (communication, web text, cameras, sensors, etc.)

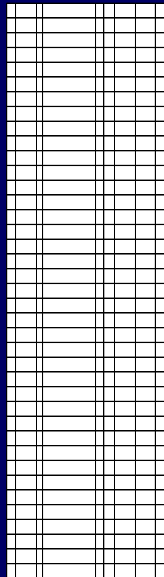
Granular data should accurately describe *reality*...

...except when data describes *fantasy*.

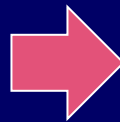
Fantasy:
Budgets
Sales forecasts
Crowd counts



Reality **capture** → Raw data

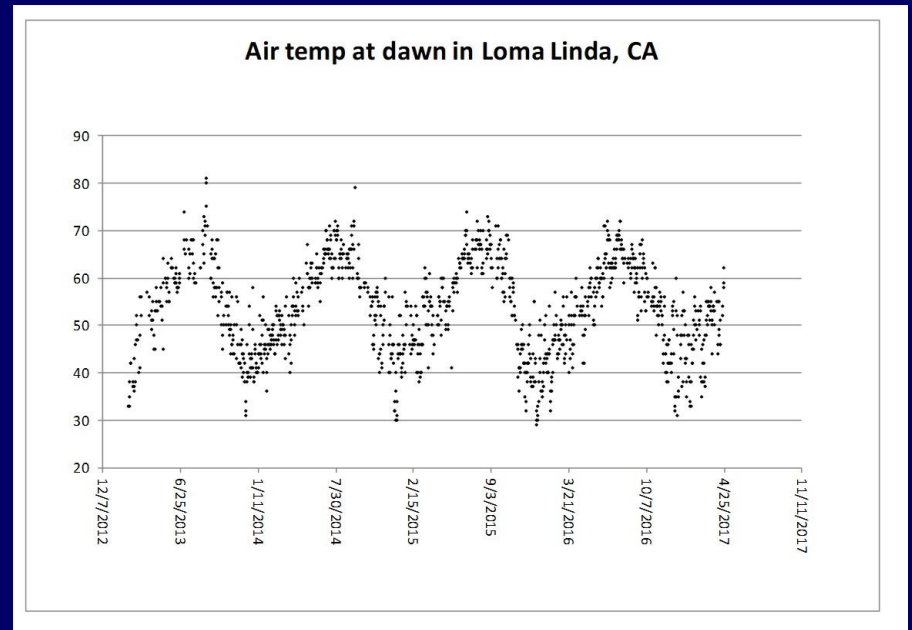


expression →

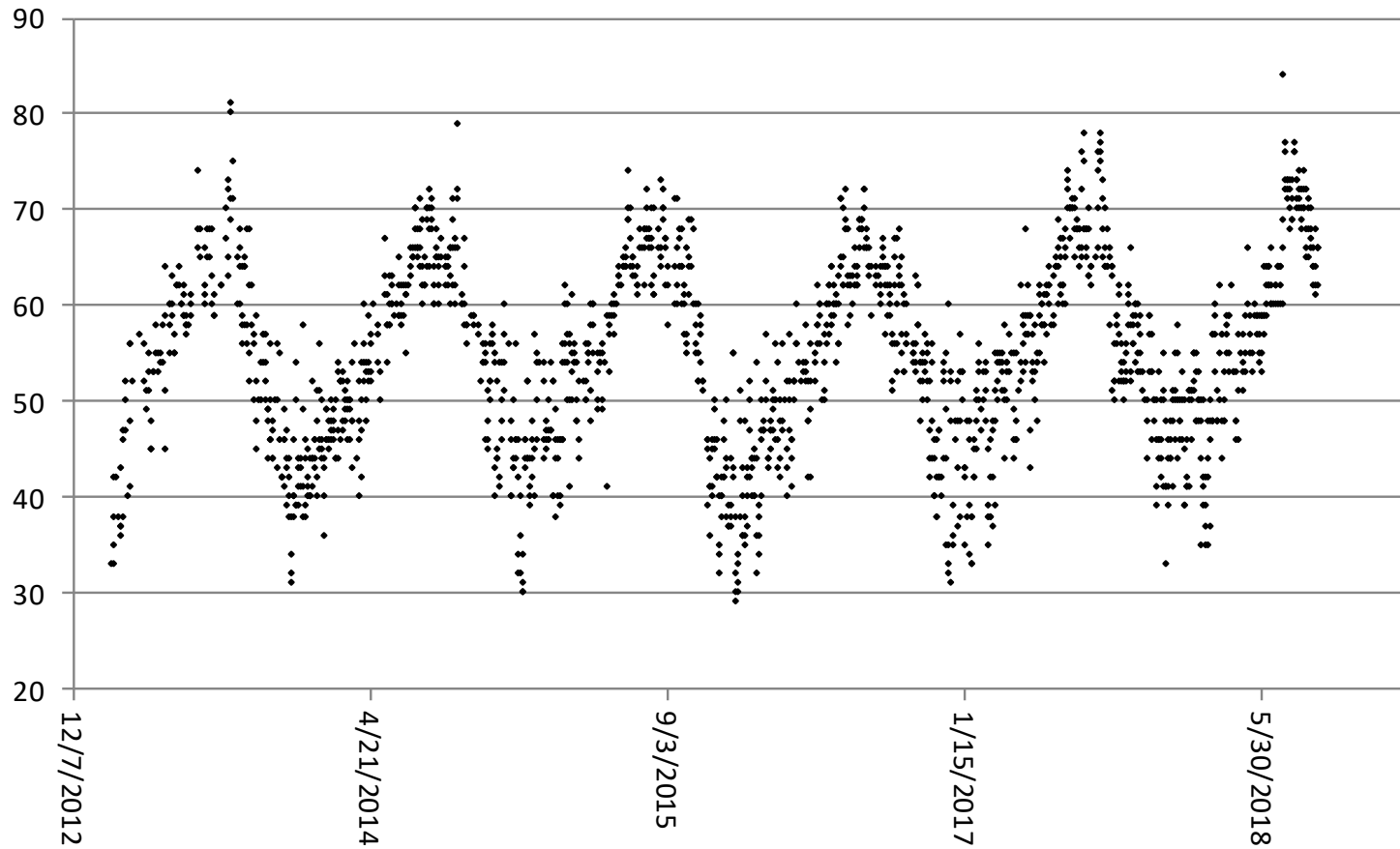


Fully-qualified,
consistently
gathered &
defined

Information !



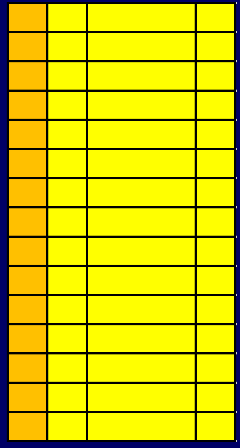
Air temp at dawn in Loma Linda, CA



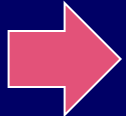
Semantic data integration

Phenomenon #1

Raw data

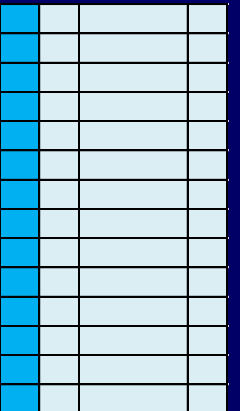


capture

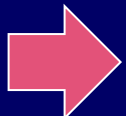


Raw data

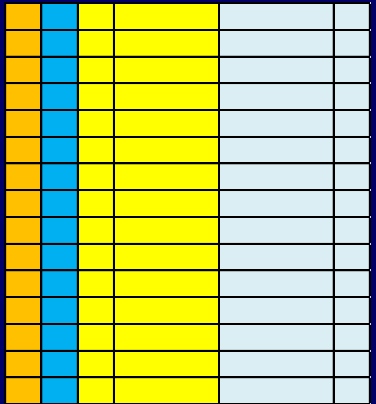
Phenomenon #2



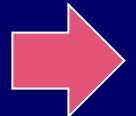
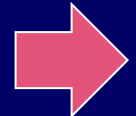
capture



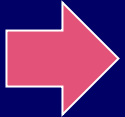
Level 1 derived data



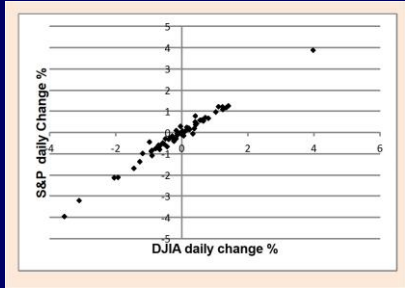
JOIN



expression



Scatter diagram
2 quantitative metrics



Information:
correlation

Not necessarily
cause & effect !

We intuitively bring our own context to raw data.

Medical vital signs:

Most medical professionals know the “normal” limits of a metric (pulse, blood pressure, blood chemistry, etc.)

A single metric (fact)...

...plus personal experience (context)...

...yields intuitive information (even if not documented or expressed)...

... “That doesn’t look good to me.”



Problem:
Most patients
lack that context.

“132 over 70”

Definition and meaning
are culturally understood.



Crafting useful information from raw data

Creating “derived” data

Tuning to a specific audience

Visualization & graphic techniques

Importance of clear definitions

Meaning & context may require integration from multiple sources.

Data is not information

MyChart



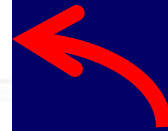
Component Results

Component	Your Value	Standard Range
WBC	7.41 bil/L	4.80 - 11.80 bil/L
RBC	5.26 tril/L	3.80 - 5.30 tril/L
Hgb	15.5 g/dL	11.0 - 16.0 g/dL
Hct	44.2 %	33.5 - 47.0 %
MCV	84.0 fL	77.0 - 96.0 fL
MCH	29.5 pg	24.5 - 32.0 pg
MCHC	35.1 g/dL	30.0 - 35.0 g/dL
RDW	13.4 %	12.0 - 15.0 %
Plts	211 bil/L	140 - 340 bil/L
MPV	10.2 fL	0.0 - 15.0 fL

General Information

Component Results

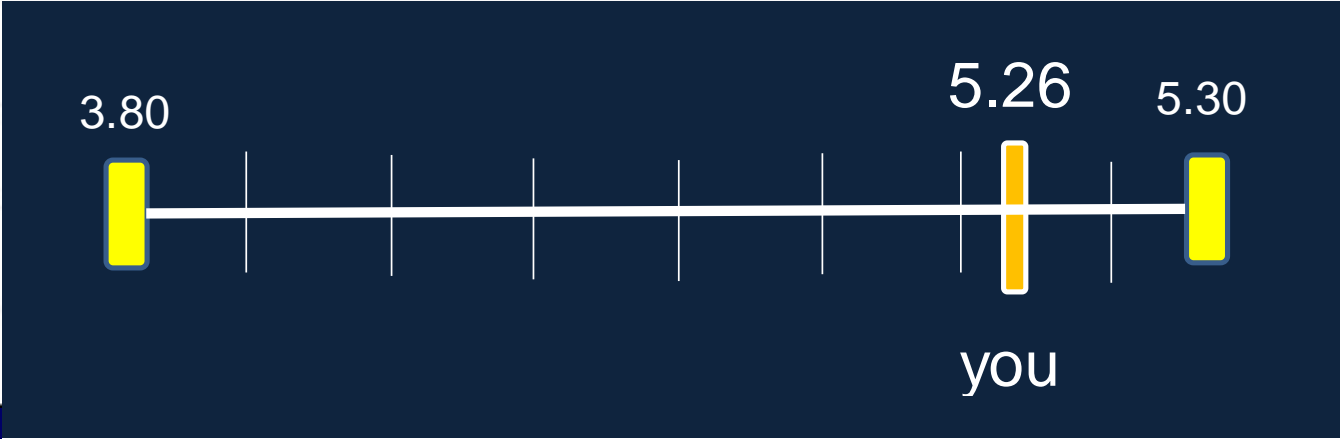
Component	Your Value
WBC	7.41 bil/L
RBC	5.26 tril/L
Hgb	15.5 g/dL
Hct	44.2 %
MCV	84.0 fL
MCH	29.5 pg
MCHC	35.1 g/dL
RDW	13.4 %
Plts	211 bil/L
MPV	10.2 fL



“5.26” Is this good or bad?

Component Results

Component	Your Value	Standard Range
WBC	7.41 bil/L	4.80 - 11.80 bil/L
RBC	5.26 tril/L	3.80 - 5.30 tril/L
Hgb	15.5 g/dL	11.0 - 16.0 g/dL
Hct	44.2 %	33.5 - 47.0 %
MCV	84.0 fL	77.0 - 96.0 fL
MCH	29.5 pg	24.5 - 32.0 pg
MCHC		
RDW		
Plts		
MPV		



General Information

Collected:
09/17/2018 12:53 PM

When the sample was taken

Resulted:
09/17/2018 3:48 PM

When the test was run.

Ordered By:
Fabrizio Luca, MD

For whom

Result Status:
Final result

This test result has been released by an automatic process.

“Metadata”

Converting data to information

Create metrics which capture the whole situation
(not merely anecdotal)

Express them in **context**. Context includes...

- historical values for same metric

- normalization to remove known factors

- ratios and share of total activity

- visualization techniques (graphics, etc.)

Data....but not information



Annual sales conference

Adjectives about sources

Reliable

Unbiased, objective
subjective (hidden agenda)

Irresponsible

Credible

Amateur – crowdsourced

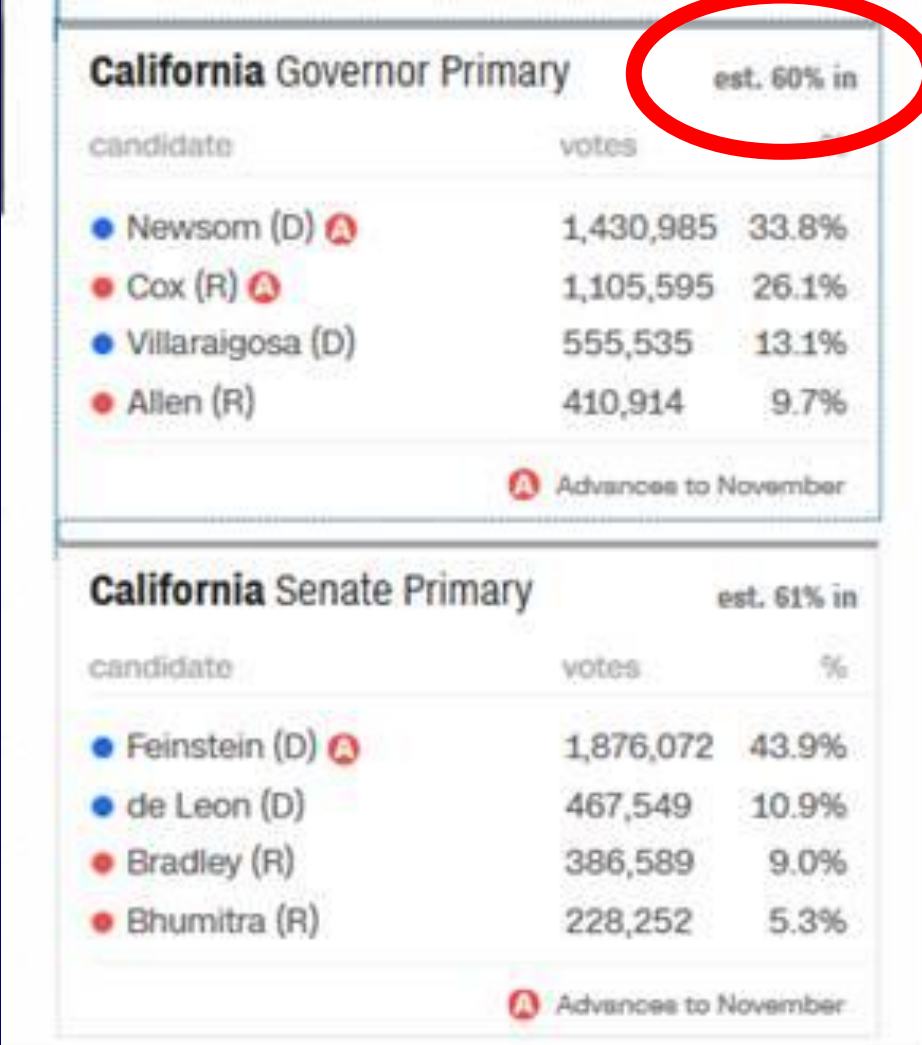
Motive for data capture

Motive for supplying to us



Telling the “truth” requires full disclosure of qualifications and conditions.


What's included.
What's excluded.
When the observation
was made.
Other conditions.


Election results
with % vote
counted.




The image shows two tables of election results. The top table is for the California Governor Primary, and the bottom table is for the California Senate Primary. Both tables list candidates, their party affiliation, and their vote counts and percentages. A red circle highlights the text 'est. 60% in' in the top right corner of the Governor Primary table.

California Governor Primary		est. 60% in	
candidate	votes		%
● Newsom (D) 	1,430,985		33.8%
● Cox (R) 	1,105,595		26.1%
● Villaraigosa (D)	555,535		13.1%
● Allen (R)	410,914		9.7%

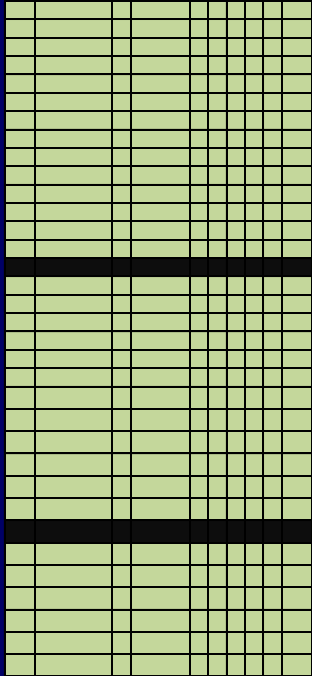
 Advances to November

California Senate Primary		est. 61% in	
candidate	votes		%
● Feinstein (D) 	1,876,072		43.9%
● de Leon (D)	467,549		10.9%
● Bradley (R)	386,589		9.0%
● Bhumitra (R)	228,252		5.3%

 Advances to November

Danger of Dashboards

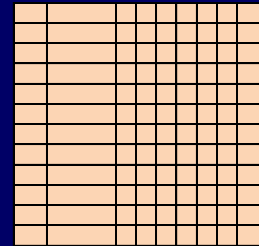
Raw granular data



Extract, filter, transform join, load.

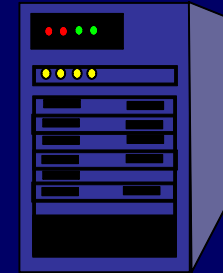
Aggregation program

Summary data

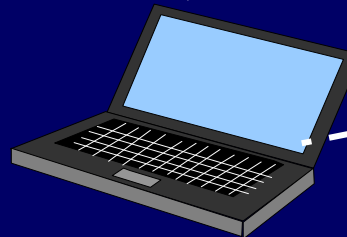
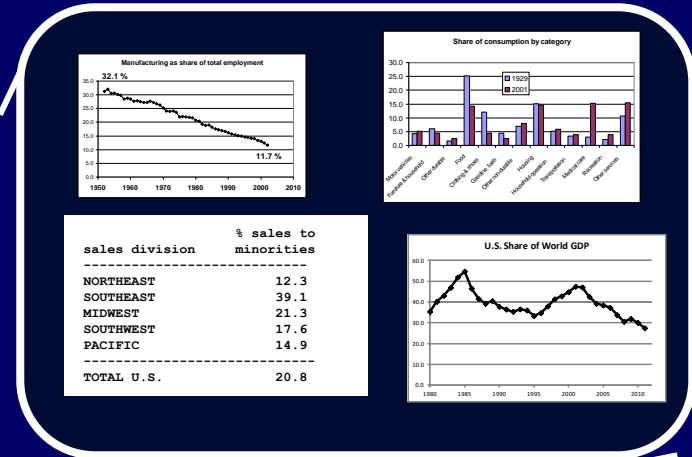


Often called a data mart.

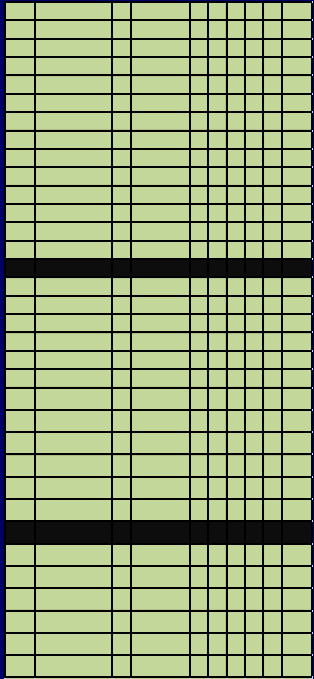
Decision-support server



Dashboard



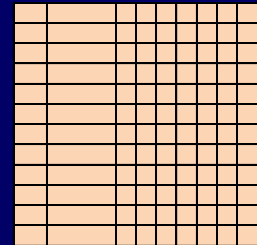
Raw transactional data



Extract, filter, transform join, load.

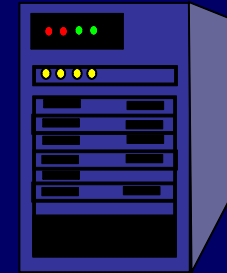
Aggregation program

Summary data

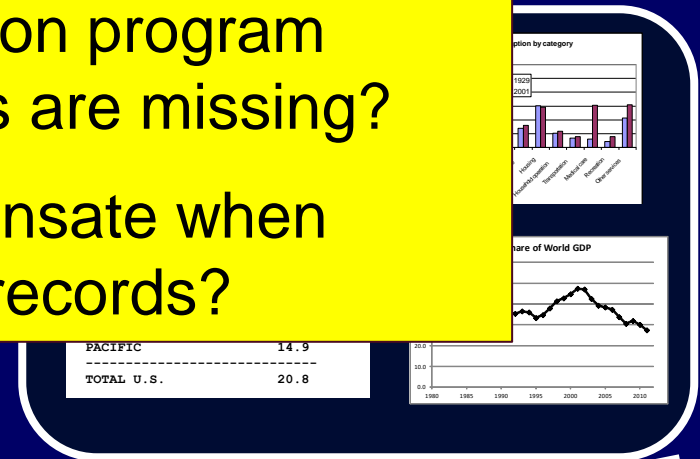


Often called a data mart.

Decision-support server

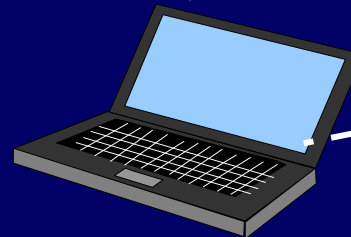


Dashboard

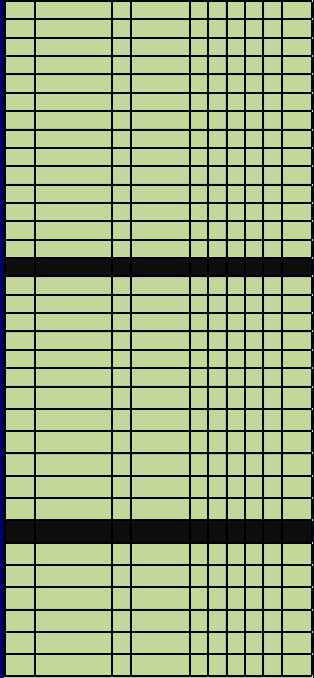


Does the aggregation program recognize that rows are missing?
How does it compensate when creating summary records?

Missing rows



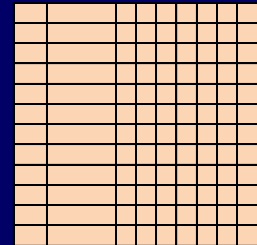
Raw transactional data



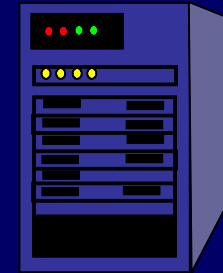
Extract, filter, transform join, load.

Aggregation program

Summary data



Decision-support server



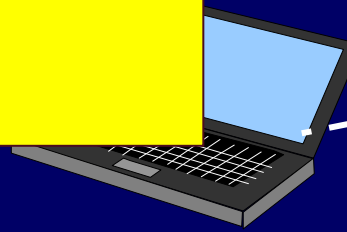
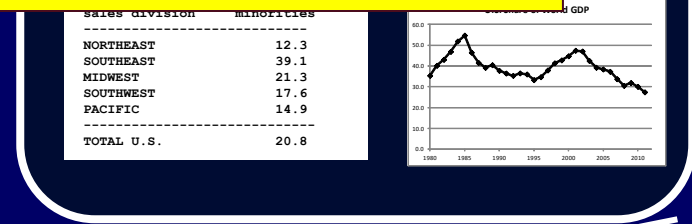
Dashboard

...then the summary data (information) is not reliable !

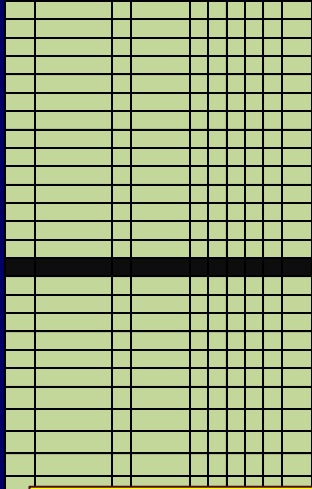
If the raw data has DQ problems...

missing rows

inconsistencies in definitions & scope...



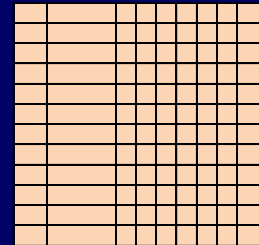
Raw transactional data



Extract, filter, transform join, load.

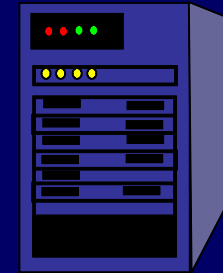
Aggregation program

Summary data



Often called a data mart.

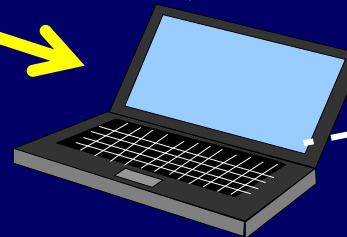
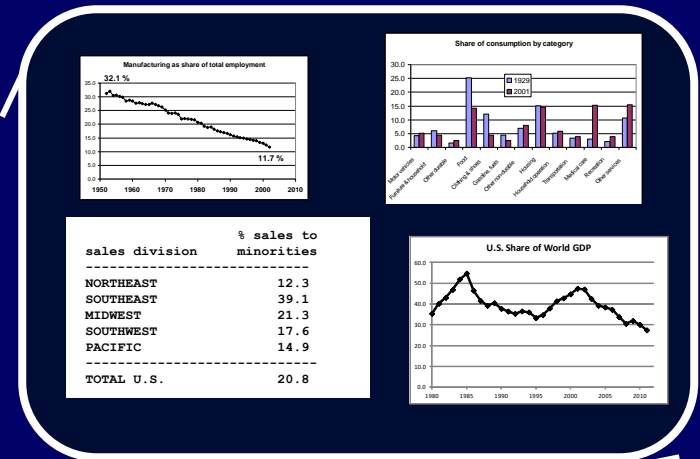
Decision-support server

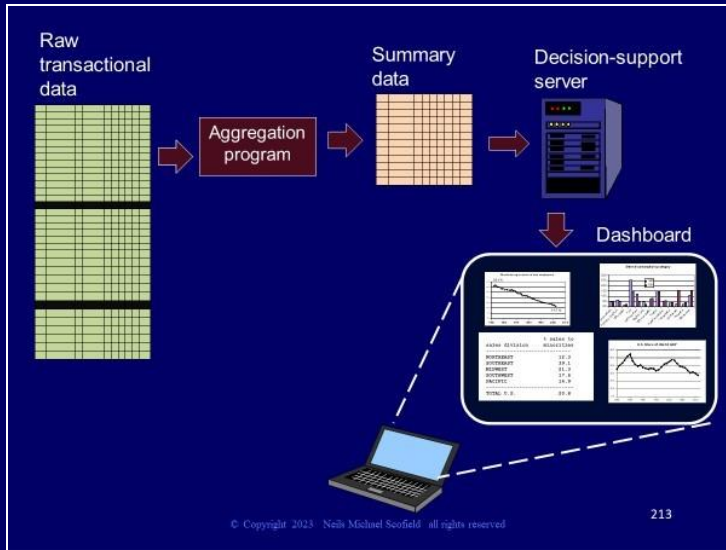


Dashboard

The dashboard user has no clue that the original raw data is bad.

Metadata about quality is not provided.





Information (metadata) which the end (business) user should know about the raw data and the information:

Scope: What's included?
What's excluded?

Consistency:

Is data captured consistently over time?

Is data defined consistently over time and space?

Currency:

Is the summary information current?

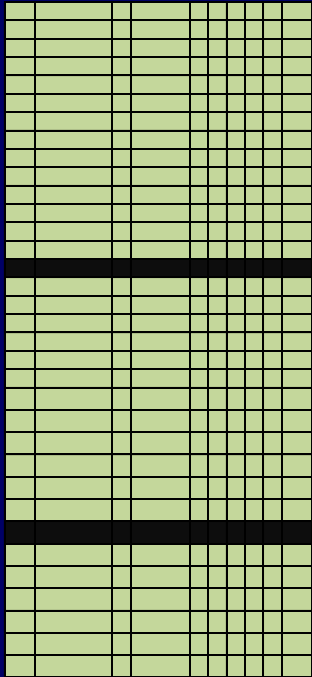
Question posed to dashboard designer:

“Who is responsible for the quality of the data?”

A: “It is the responsibility of the end-user to assess the quality of the data.”

...and this guy got promoted!

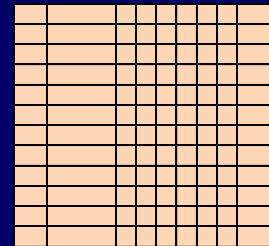
Raw transactional data



Extract, filter, transform join, load.

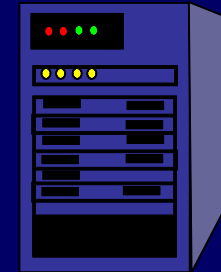
Aggregation program

Summary data

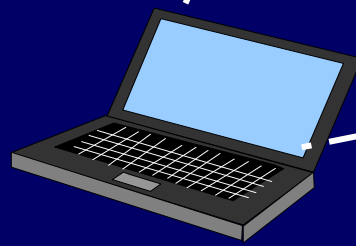
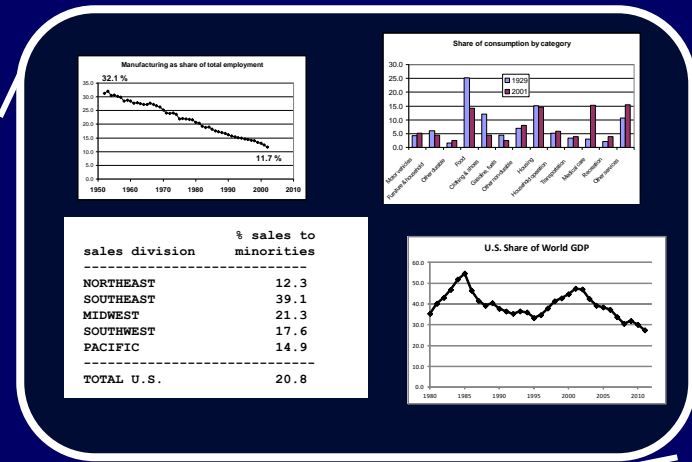


Often called a data mart.

Decision-support server



Dashboard



Reality is complex

Brief assertions usually be

Understanding a complex

...for an executive or a v

Oversimplifying a complex
(e.g. Iraq is one homo

...yet...

Audience (readers, liste

Deny complexity ("it car



Reality is complex (cont.)

Any assertion about reality needs to be adequately qualified.

There may be assumptions not made explicit.

There may be conditions not made explicit.

The audience (reader) may assume definitions which are not so.

Withholding definitions is a common way of falsifying.



How to misrepresent the truth in information construction.

Select only favorable data series

Use only favorable data points (“cherry pick”)

Failure to normalize data

Failure to disclose sources of bias and other disconnects

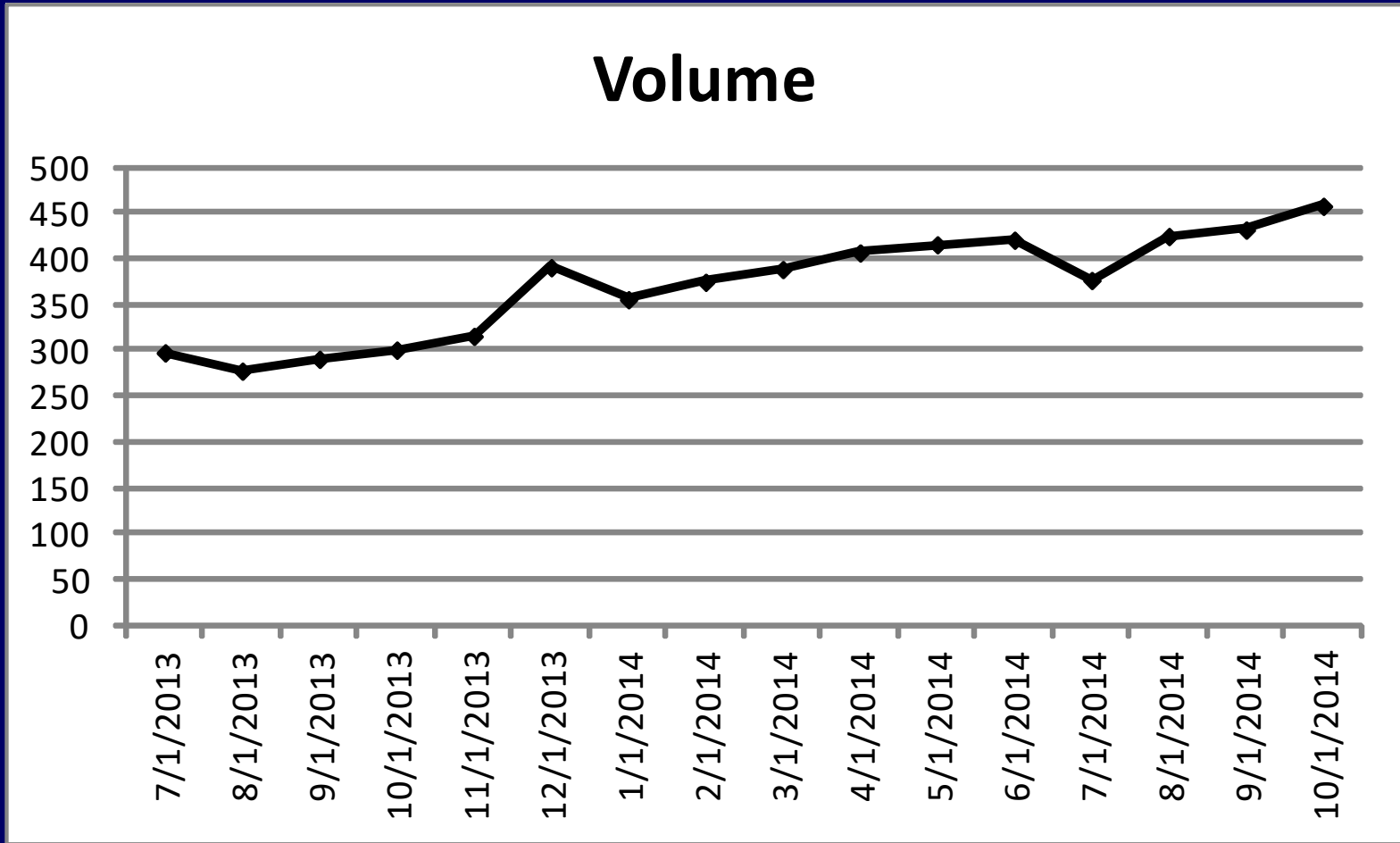
Failure to provide appropriate context (history or peer)

Select only advantageous peers

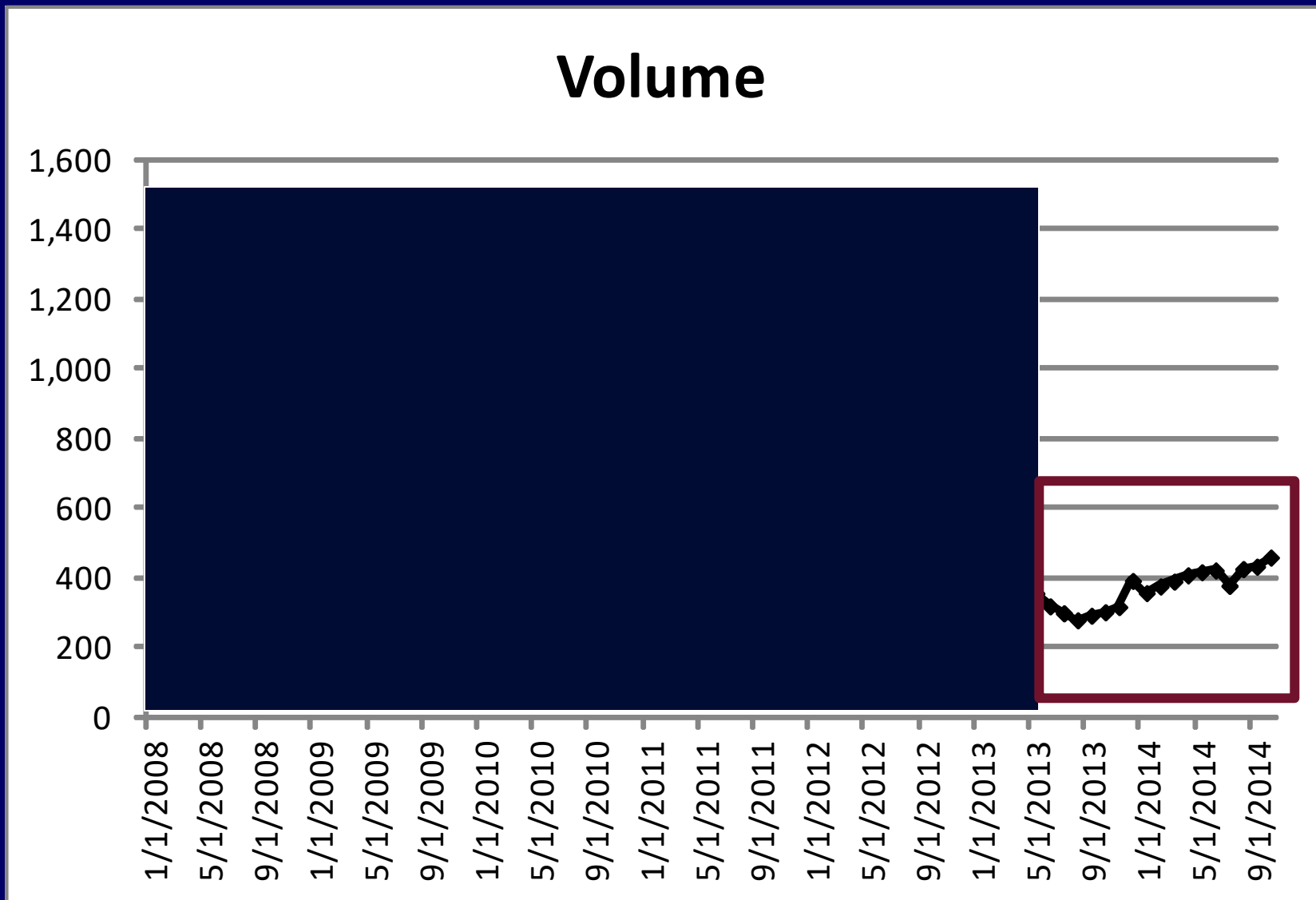
Misrepresentation in graphic form

“...more than Delaware and Rhode Island combined...”

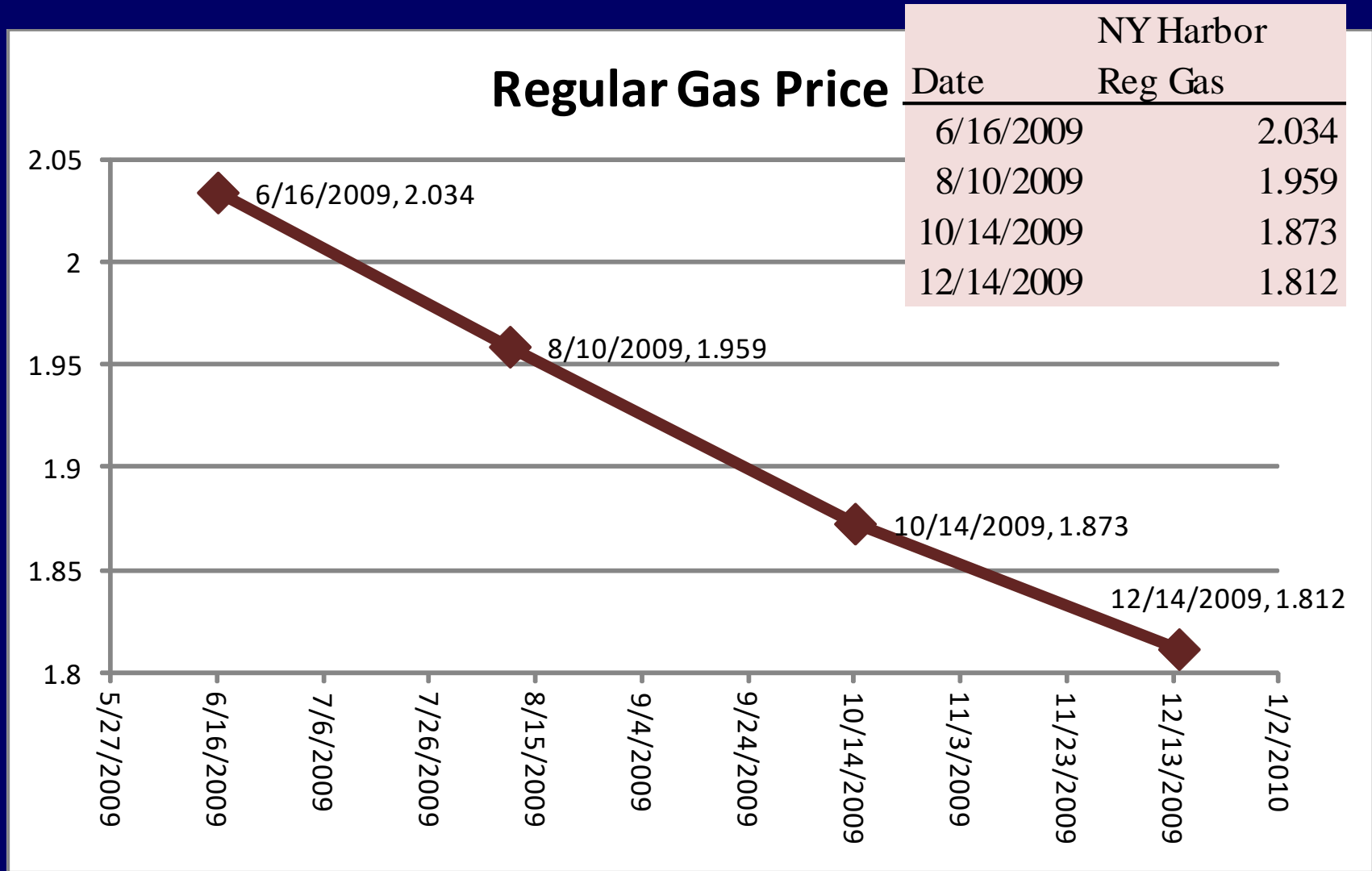
Context may come from history



Why lots of history?

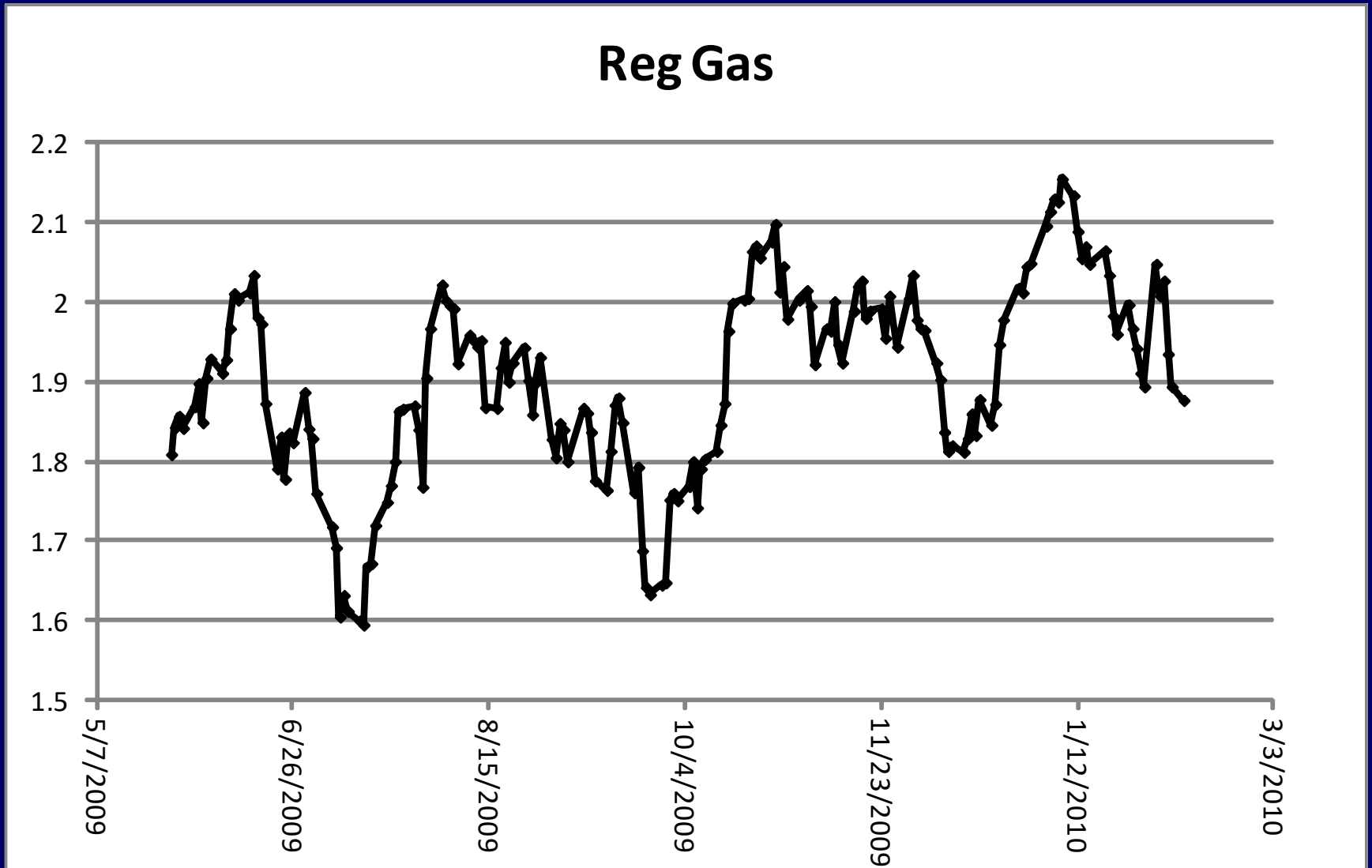


Ask for the most granular data available!



Source: U.S. Energy Info. Admin., Spot Prices for Crude Oil and Petroleum Products

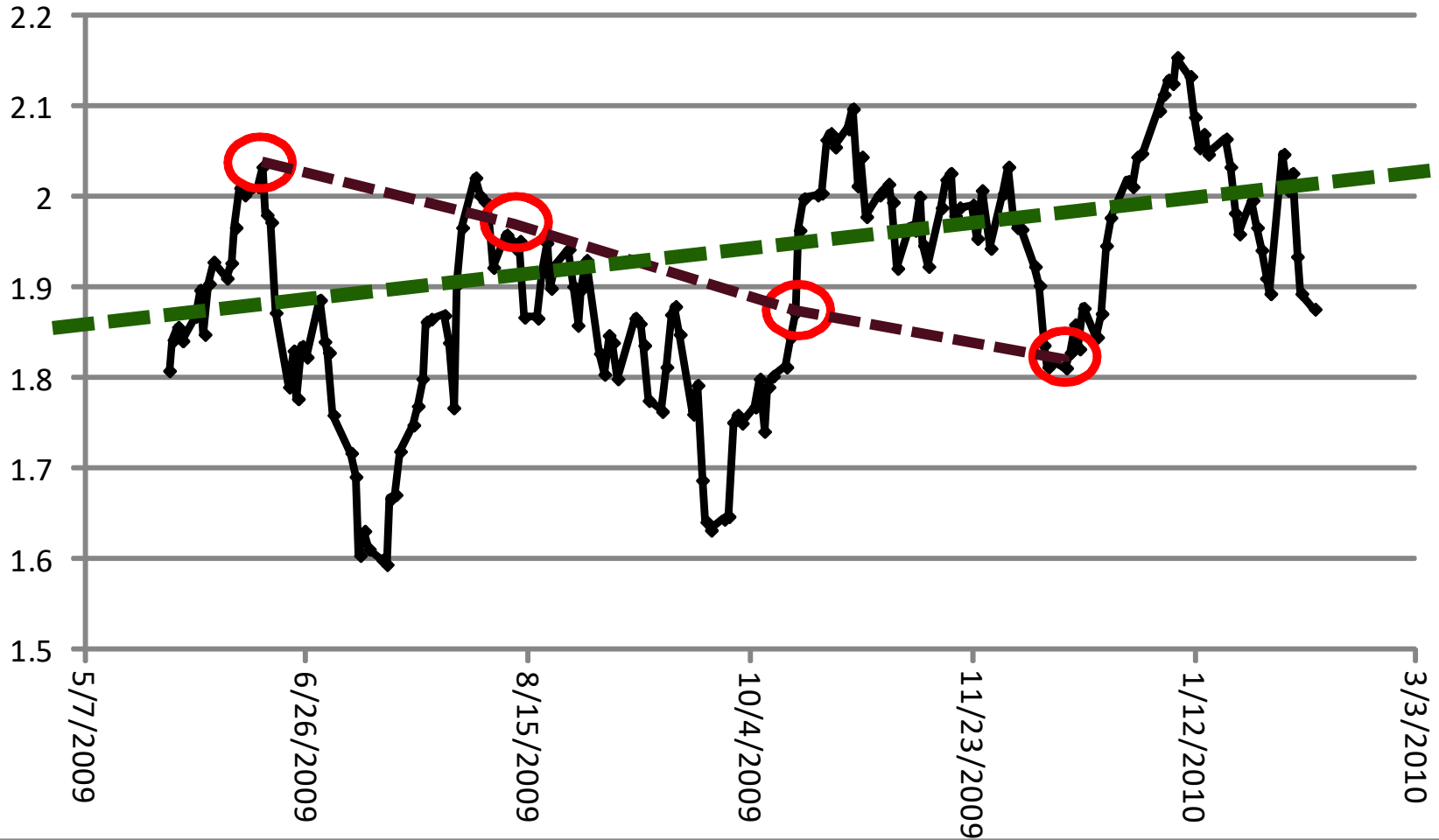
With more data points!



Source: U.S. Energy Info. Admin., Spot Prices for Crude Oil and Petroleum Products

© Copyright 2023 Neils Michael Scofield all rights reserved

Reg Gas



Source: U.S. Energy Info. Admin., Spot Prices for Crude Oil and Petroleum Products

Rev. 10/9/2022

© Copyright 2023 Neils Michael Scofield all rights reserved

Context !

Cactus Valley, NV Pop. 1,403

Fact: Dr. Smith has 1 letter of reprimand.



Letter of reprimand.
CA Medical Board



Dr. Eaton



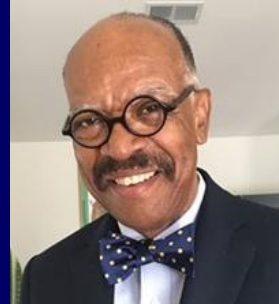
Dr. Lee



Dr. Smith



Dr. Jones



Dr. Lang

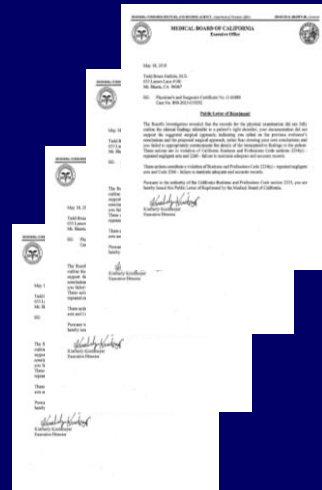


Dr. Kelly

Context !

Cactus Valley, NV Pop. 1,403

More complete Fact:
Dr. Smith has fewer letters than the rest.



Dr. Eaton



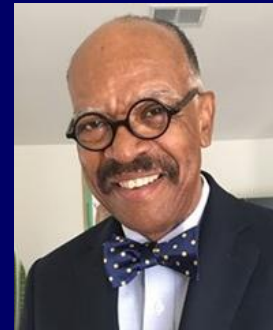
Dr. Lee



Dr. Smith



Dr. Jones



Dr. Lang



Dr. Kelly

Creating vital information from raw data.

*A distress radio signal
in the North Atlantic.*

Bearings are all raw data.

North Atlantic

Shannon, Ireland
52 42 N 8 55 W

Brest, France
48 26 N 4 21 W

Porto, Portugal
41 14 N 8 41 W

Ship in distress:
41 N 14 W

Derived data !!

235 ° radial

261 ° radial

304 ° radial

1500

1600

1700

1800

1900

2000



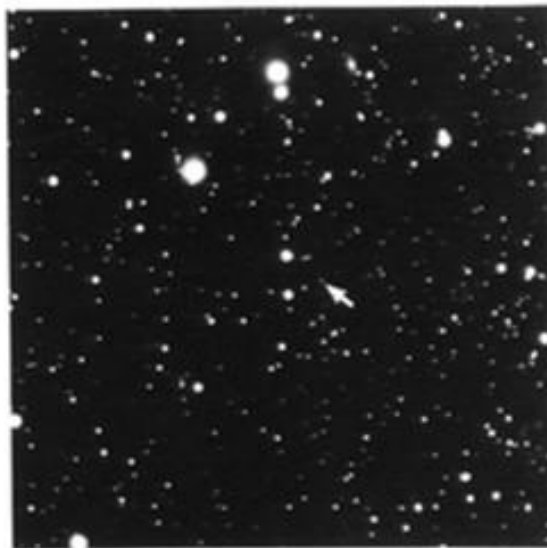
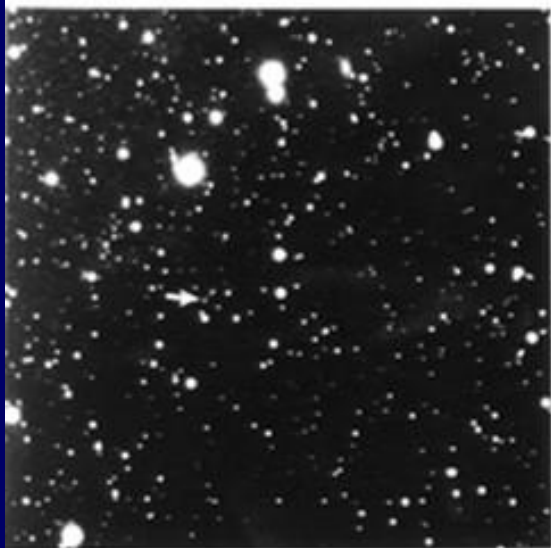
Galileo



1930: Pluto discovered

Clyde Tombaugh
1906 – 1997

DISCOVERY OF THE PLANET PLUTO



Janu

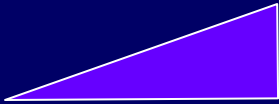
ry 29, 1930

Blink comparator

ils Michael Scofield all rights



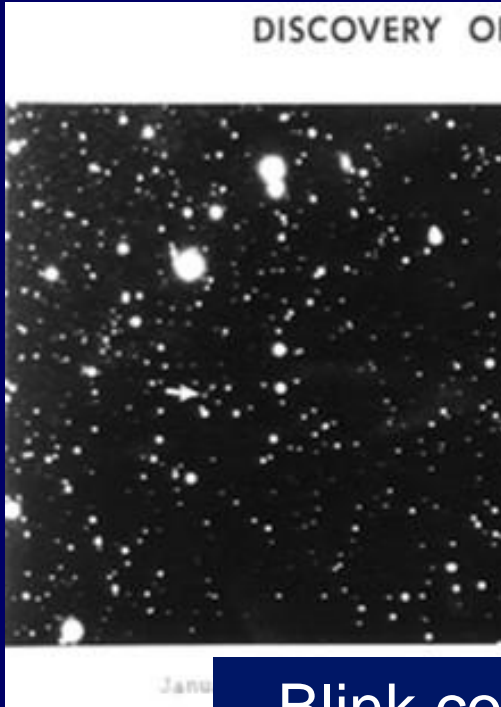
1500 1600 1700 1800 1900 2000



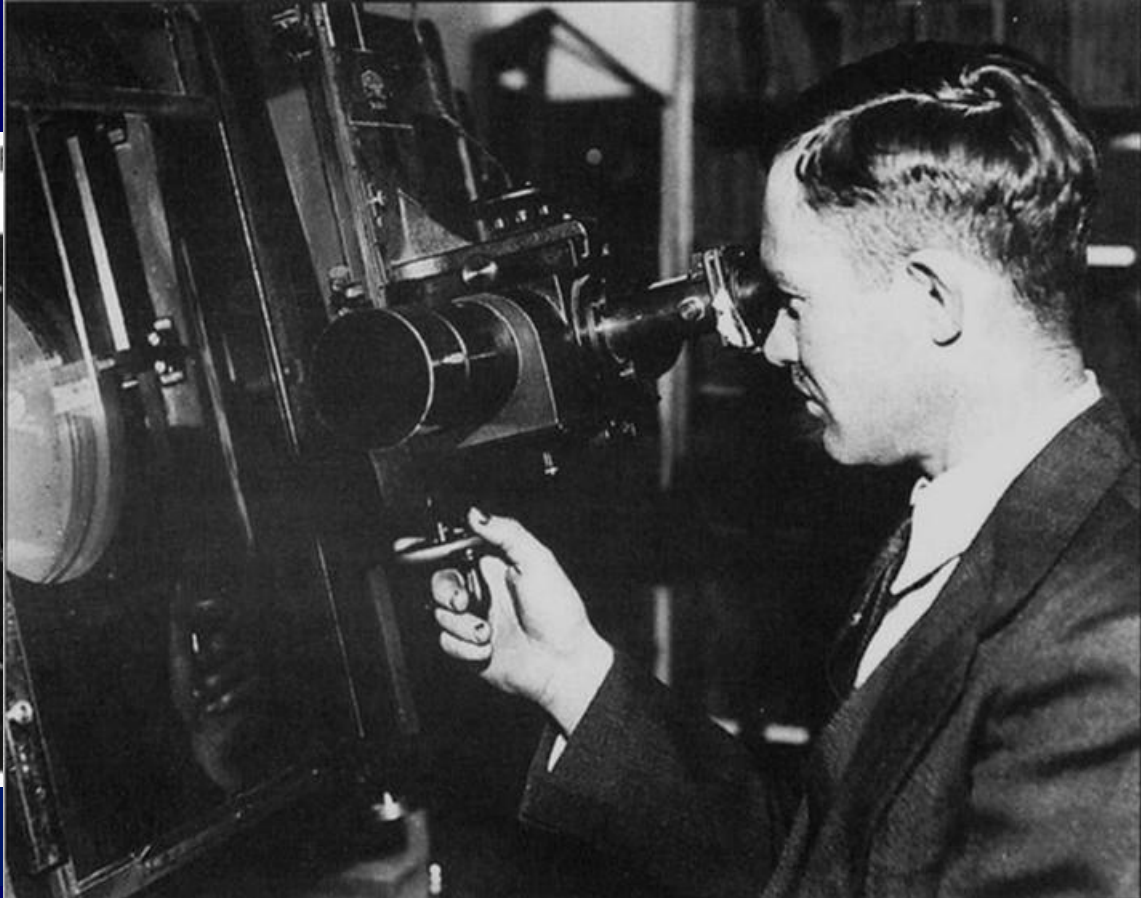
Galileo

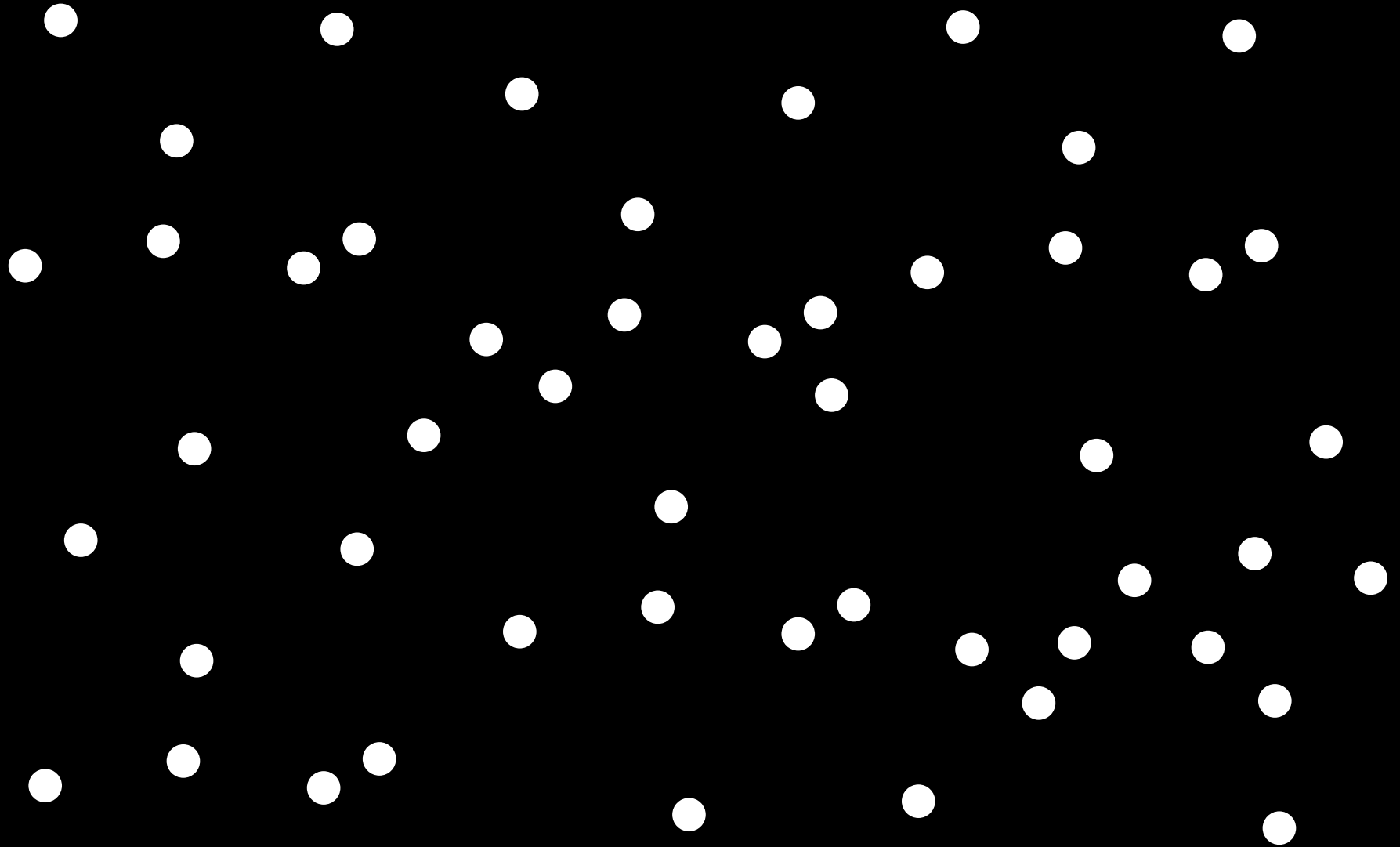


1930: Pluto discovered

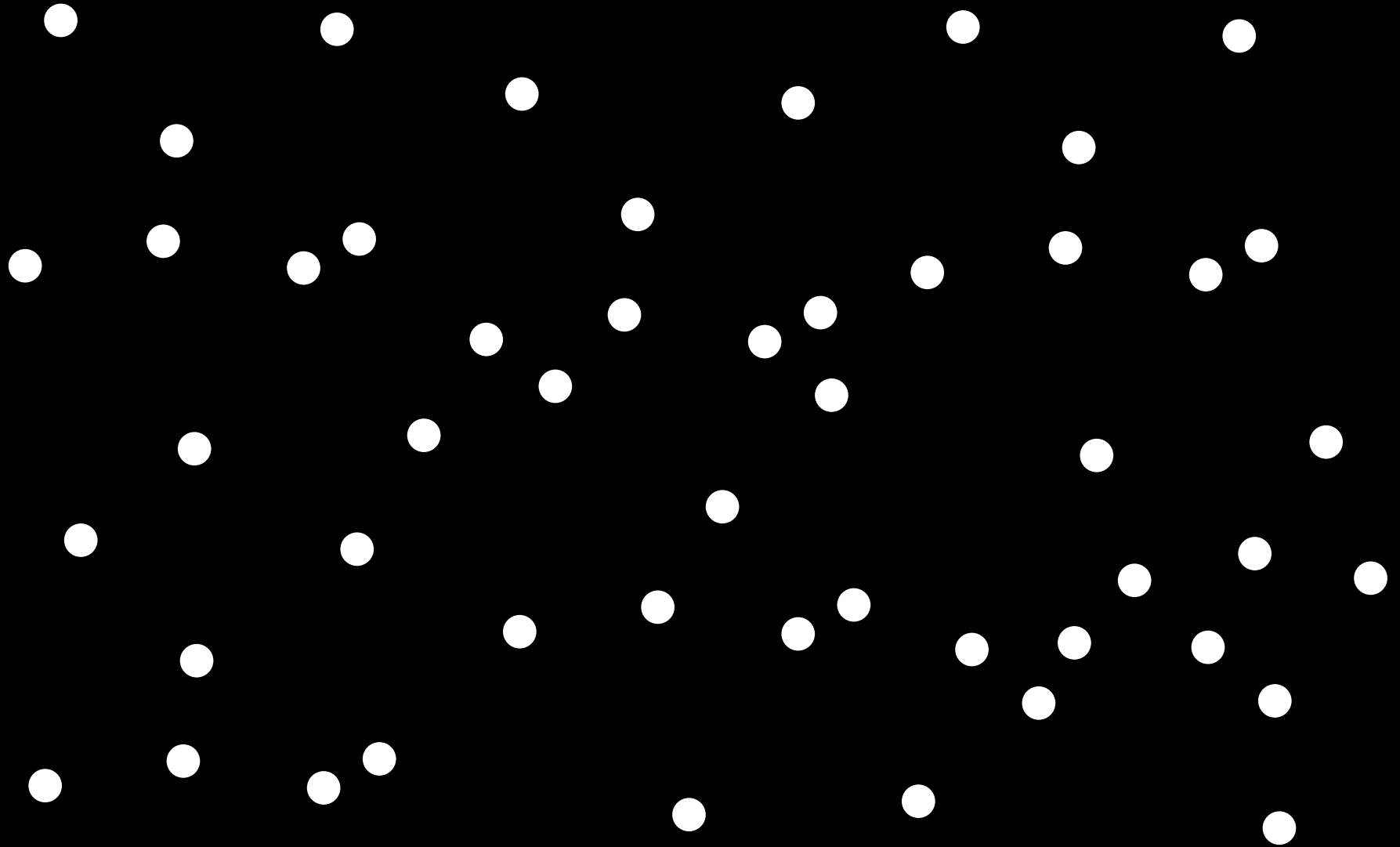


Blink co





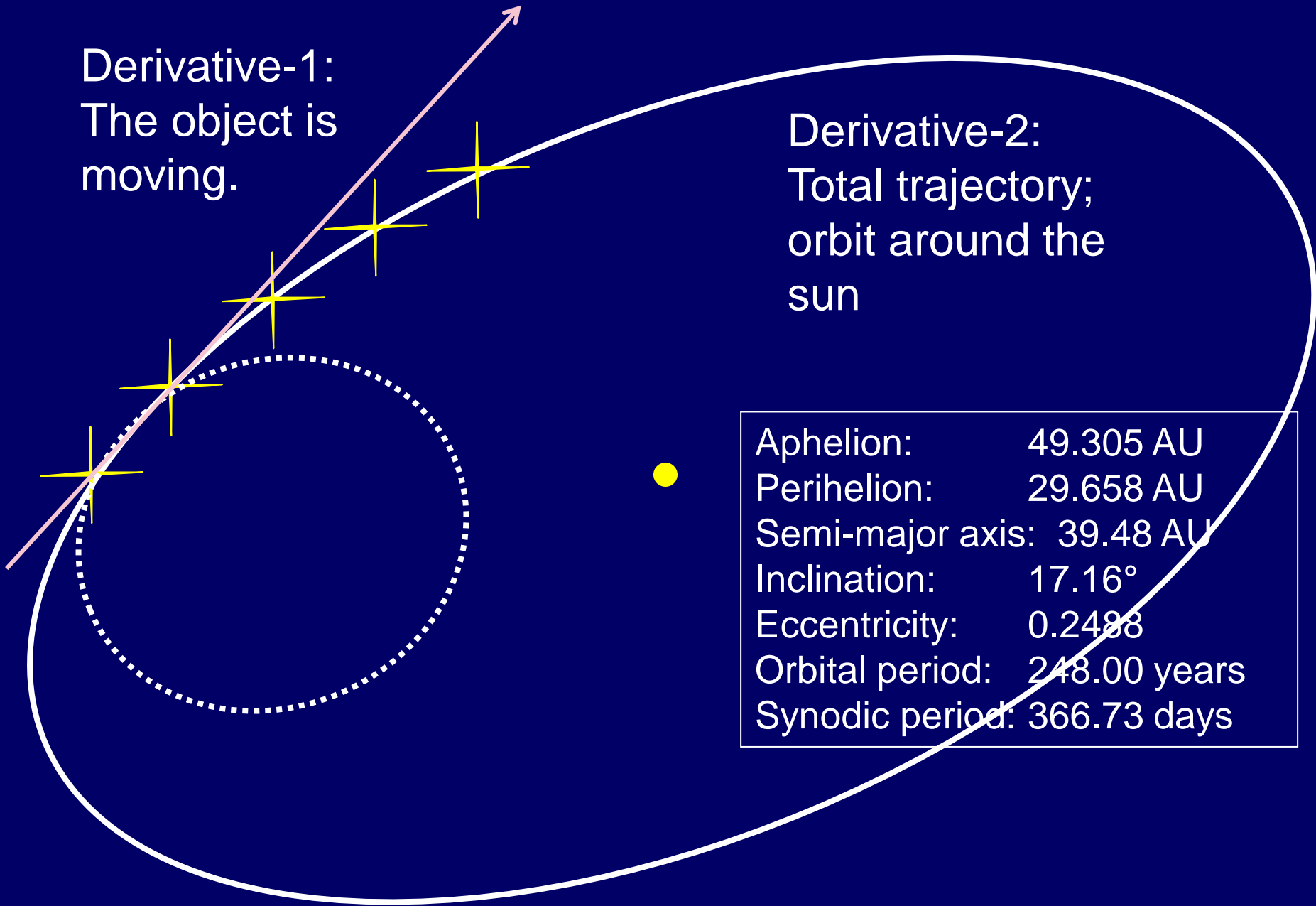
January 23



January 29

Derivative-1:
The object is
moving.

Derivative-2:
Total trajectory;
orbit around the
sun



Aphelion:	49.305 AU
Perihelion:	29.658 AU
Semi-major axis:	39.48 AU
Inclination:	17.16°
Eccentricity:	0.2488
Orbital period:	248.00 years
Synodic period:	366.73 days

Are there other explanations (models) for the data?

Is it really the same object (planet or star)?

Spectral analysis might answer that.

Or it might not?

Alternative hypothesis:

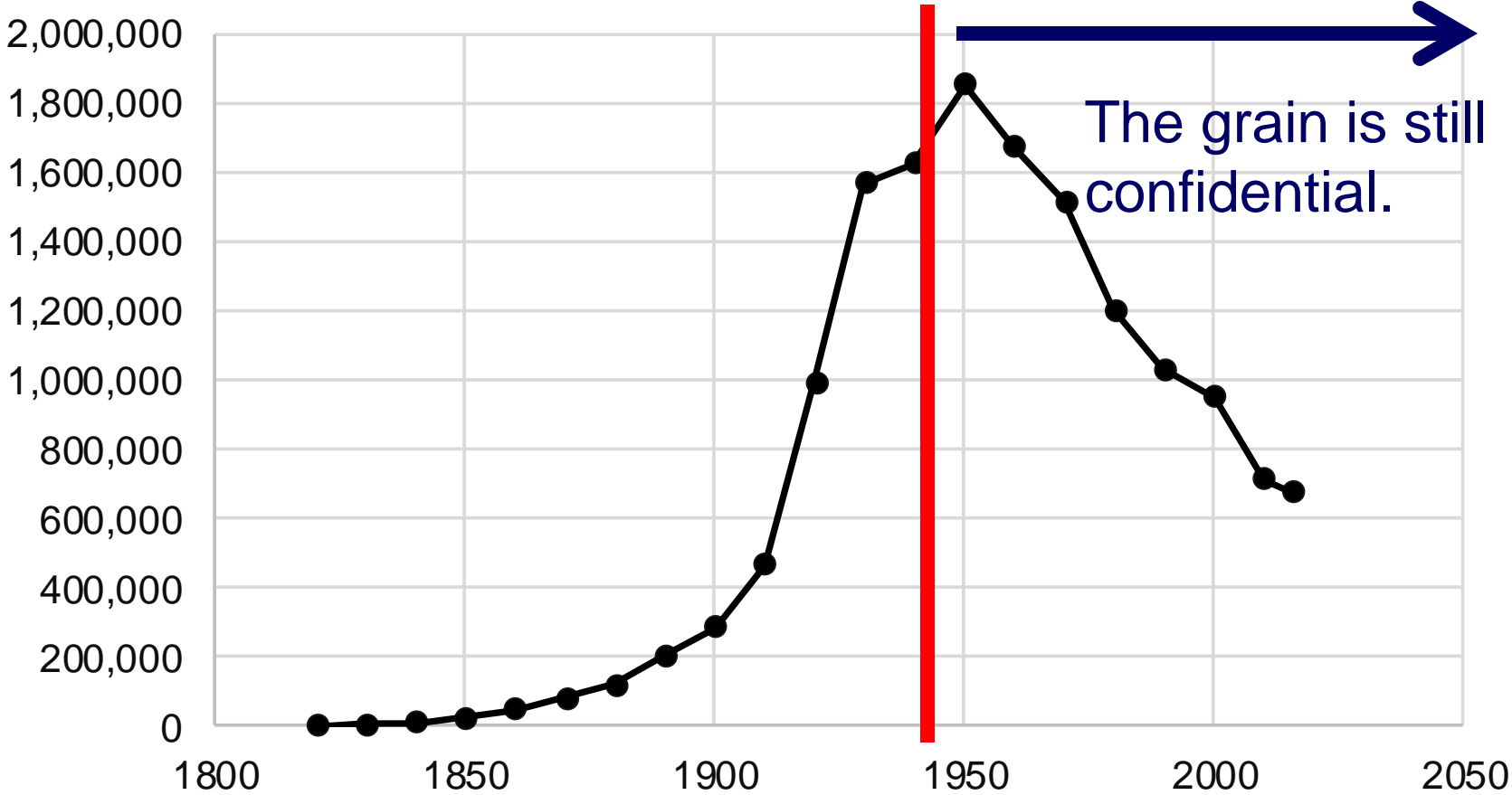
Two stars, blinking alternately.

Year	Phoenix population	
1830	0	
1840	0	
1850	0	
1860	0	
1870	240	—
1880	1,708	611.70%
1890	3,152	84.50%
1900	5,544	75.90%
1910	11,314	104.10%
1920	29,053	156.80%
1930	48,118	65.60%
1940	65,414	35.90%
1950		
1960		
1970	581,572	32.40%
1980	789,704	35.80%
1990	983,403	24.50%
2000	1,321,045	34.30%
2010	1,445,632	9.40%
2016	1,615,017	11.70%

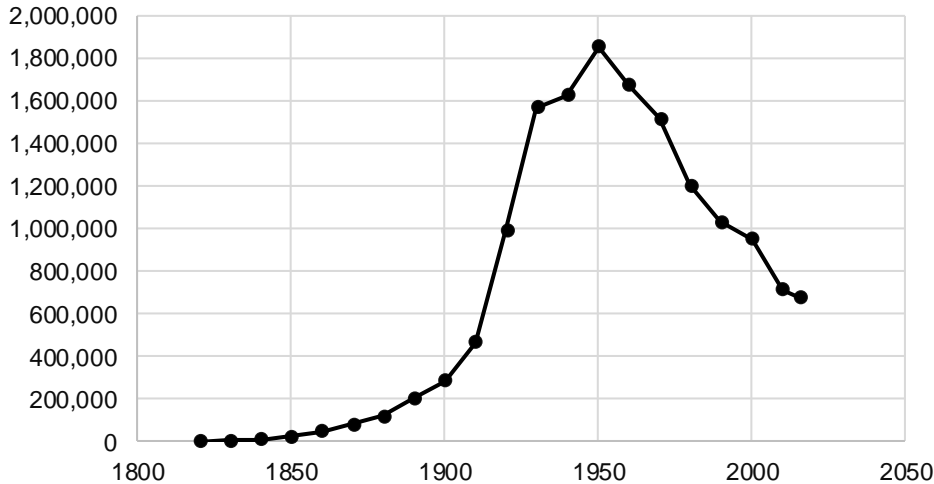
Year	Detroit population	
1820	1,422	—
1830	2,222	56.30%
1840	9,102	309.60%
1850	21,019	130.90%
1860	45,619	117.00%
1870	79,577	74.40%
1880	116,340	46.20%
1890	205,876	77.00%
1900	285,704	38.80%
1910	465,766	63.00%
1920	993,678	113.30%
1930	1,568,662	57.90%
1940		3.50%
1950		13.90%
1960		-9.70%
1970	1,514,063	-9.30%
1980	1,203,368	-20.50%
1990	1,027,974	-14.60%
2000	951,270	-7.50%
2010	713,777	-25.00%
2016	672,795	-5.70%

First Derivative

Detroit population



Detroit population

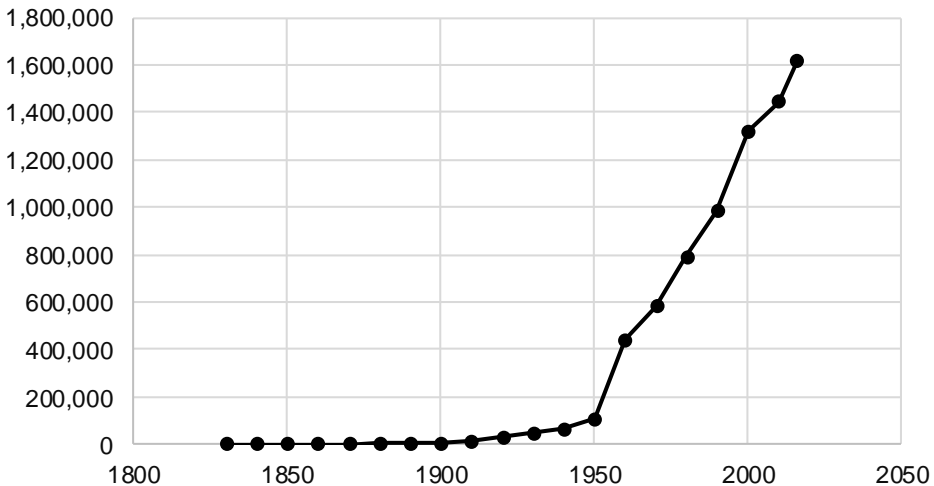


Raw Census data (individual household) has privacy issues.

The yearly totals (first derivative) do not.

Each chart is a “Second Derivative”

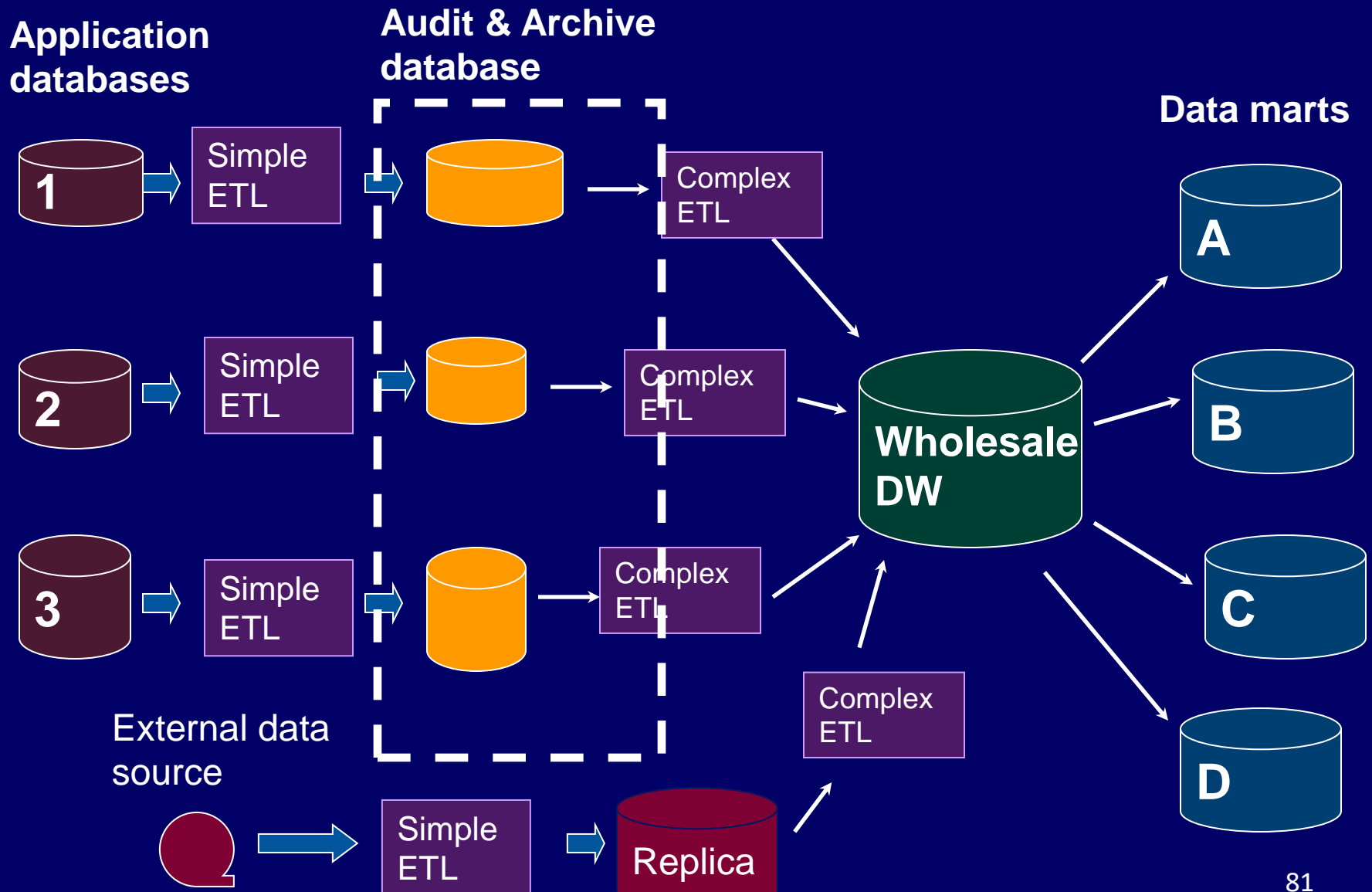
Phoenix population



Most recent total with previous decades providing context.

The shape tells a story!

Stages of data warehousing



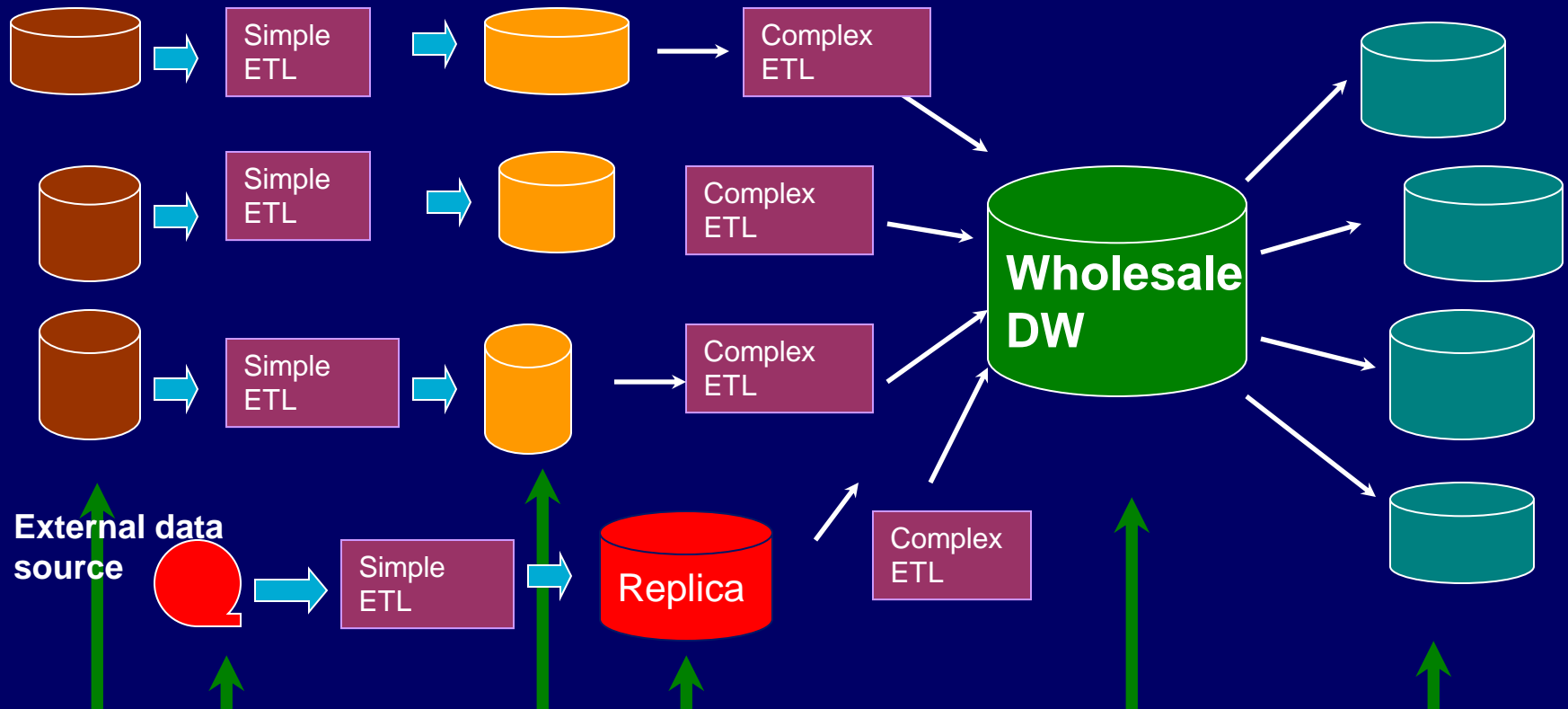
Total DW strategy

4GL Query Tool

Application databases

Audit & Archive database

Data marts



Meta Data Repository

Problem: Expectation of precision and absolute reliability

Artificial intelligence for professionals and for consumers—different.

Medical A.I.:

Doctors want to know what is behind an A.I.-generated diagnosis.

What was the source data?
May wish to “drill down”.



Ginny Rometty,
IBM

Many answers (“derived data”) involve confidence levels. (a granular metadata)

Inform user of the likelihood of error.

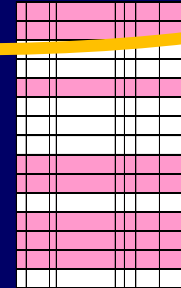
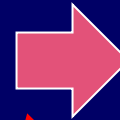
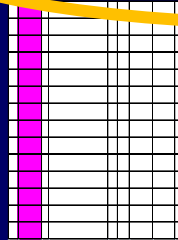
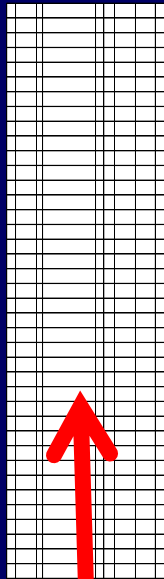
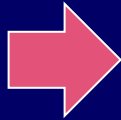
Classical
data
warehouse

Raw
data

Low
level
derived
data

High
level
derived
data

Reality

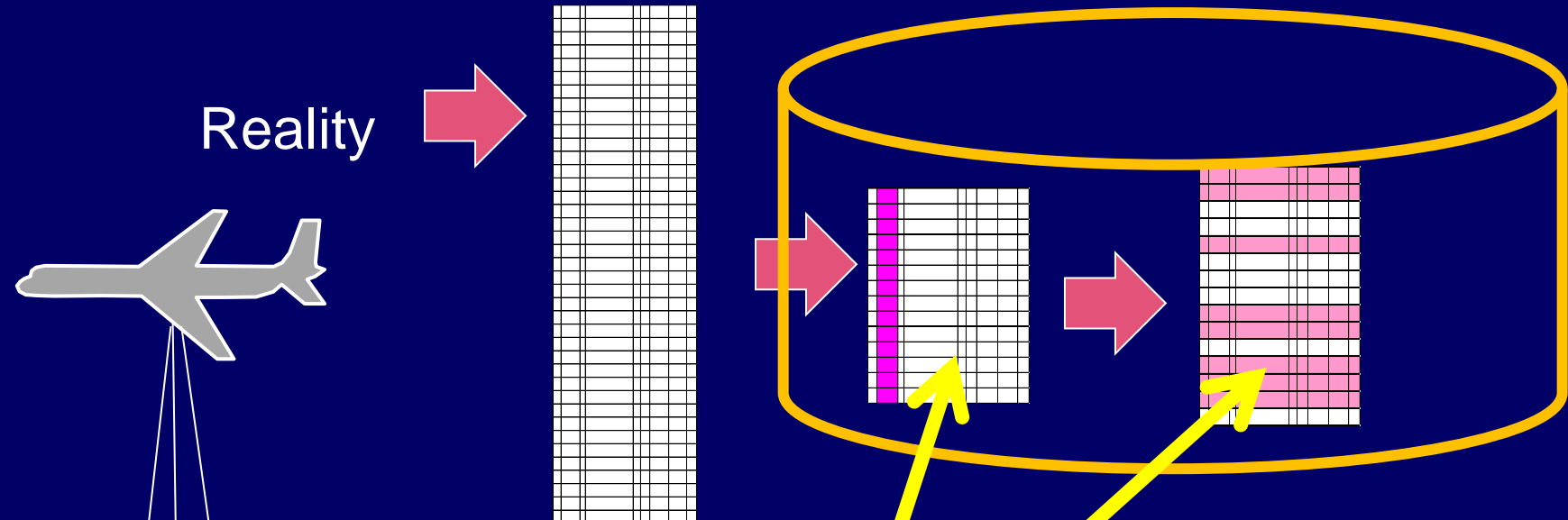


DW

Where does the raw
data come from?

What's in this process
or algorithm?

Imagery analysis for military



These derivative processes are really difficult. Usually rely on human analysts.

Detect patterns in movement of objects on the ground, and deduce strategic intent.

Stages of imagery data (towards information)

How things appear

What objects are

How many objects (in class)

How situation changes

Why things change

raw data

1st level derivative data

2nd level derivative data

3rd level derivative data

4th level derivative data



pixels

A kind of aircraft

How many aircraft

Planes coming & going.

What are they up to?

Analysis of object

Color
Heat
Condition

Intent

Increased analysis & “processing”

More room for error and/or speculation

Hostile?
Benign?
Urgent?

Stages of imagery data (towards information)

How things appear

raw data



pixels

Image interpretation



Assumptions:
Objects are as they appear to be.

What objects are

1st level derivative data

A kind of aircraft



What about decoy aircraft?

Stages of imagery data (towards information)

How things appear

raw data



What objects are

1st level derivative data

A kind of aircraft

Counting like objects

How many objects (in class)

2nd level derivative data

How many aircraft

Image interpretation

Assumptions:
No planes hiding in hangars or elsewhere.

Stages of imagery data (towards information)

How things appear

What objects are

How many objects (in class)

How situation changes

Why things change

raw data

1st level derivative data

2nd level derivative data

3rd level derivative data

4th level derivative data



pixels

A kind of aircraft

How many aircraft

Planes coming & going.

What are they up to?

Analysis of object

Color
Heat
Condition

Intent

Increased analysis & “processing”

More room for error and/or speculation

Hostile?
Benign?
Urgent?

This is aggregation and integration of non-quantitative information.

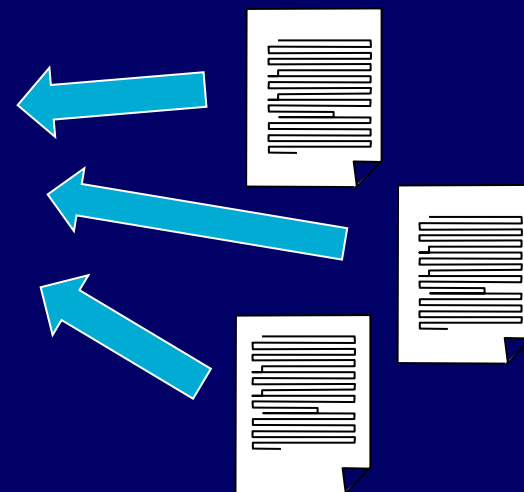
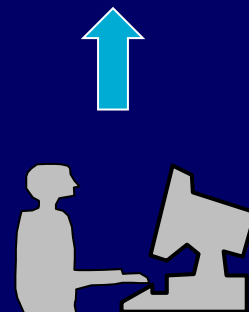
Humans now. Future... A.I.?

An "information warehouse"
(parallel to a data warehouse)

Higher-level analyst.




Executive summary:
strategic assessment



Statement of intent is an "educated guess"
Only part of the big picture

Assessments and estimates from other intelligence sources

Stages of imagery data (towards information)

How things appear	What objects are	How many objects (in class)	How situation changes	Why things change
raw data	1st level derivative data	2nd level derivative data	3rd level derivative data	4th level derivative data
 pixels	A kind of aircraft	How many aircraft	Planes coming & going.	What are they up to?
	Analysis of object	Color Heat Condition	Intent	Hostile? Benign? Urgent?
Increased analysis & "processing"				
More room for error and/or speculation				

This is aggregation and integration of non-quantitative information.

Humans now. Future A.I.?

An "information warehouse" (parallel)

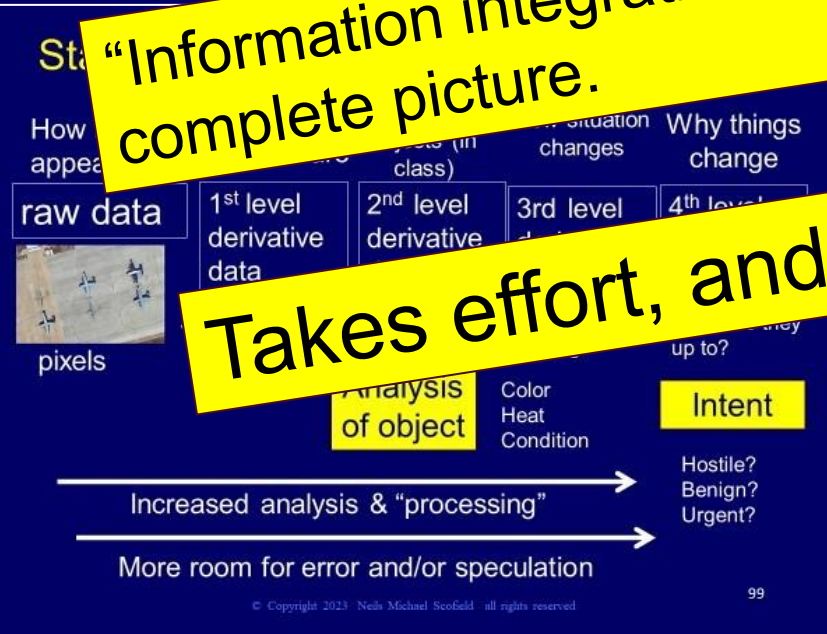


Executive summary: strategic assessment

Raw data from one source does not tell the whole story !

"Information integration" from many sources gives a more complete picture.

Takes effort, and costs money !



Statement of intent is an "educated guess" Only part of the big picture

Other intelligence sources

This is aggregation and integration of non-quantitative information.



Executive summary: strategic assessment

Human Future A.I.?

An (parallel)

Executive summary is an estimate...a forecast!
It is not absolute!

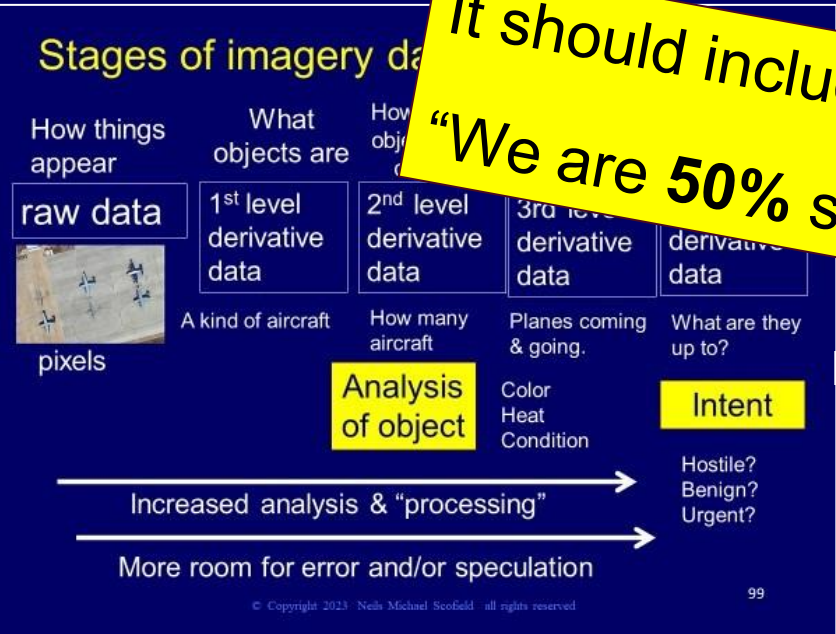
High analyst.

It should include metadata...
"We are 50% sure this will happen."



Other intelligence sources

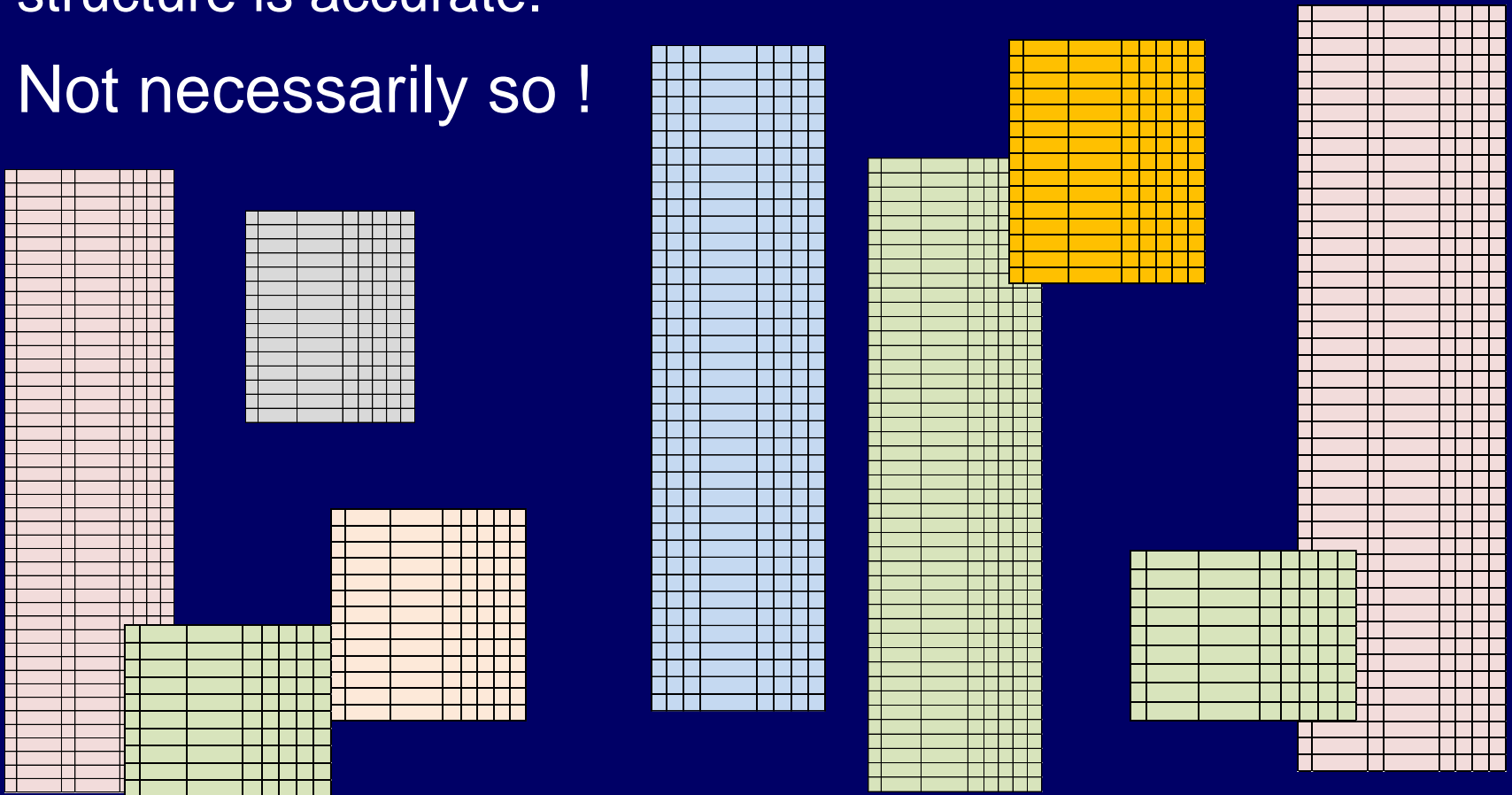
State intent is an "educated guess"
 Only part of the big picture



99% of all tabular data lacks cell-level metadata.

Hence, typical users assume that all data in a tabular structure is accurate.

Not necessarily so !



Cell-level metadata.

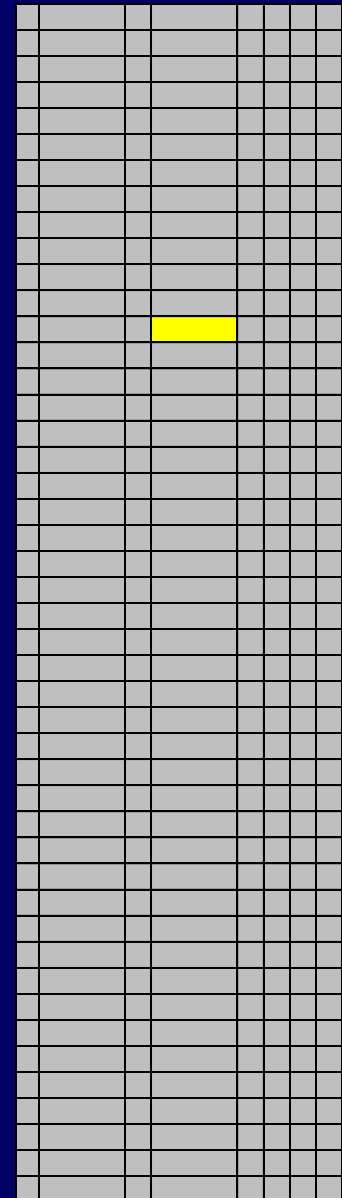
What we know about a single fact.

When we got it.

Where we got it from.

When was it observed.

Confidence level about the fact.



Business data pedigree

Showing where a fact or record came from, and when.

a.k.a. Audit trail

a.k.a. “Imbedded metadata”

Table pedigree

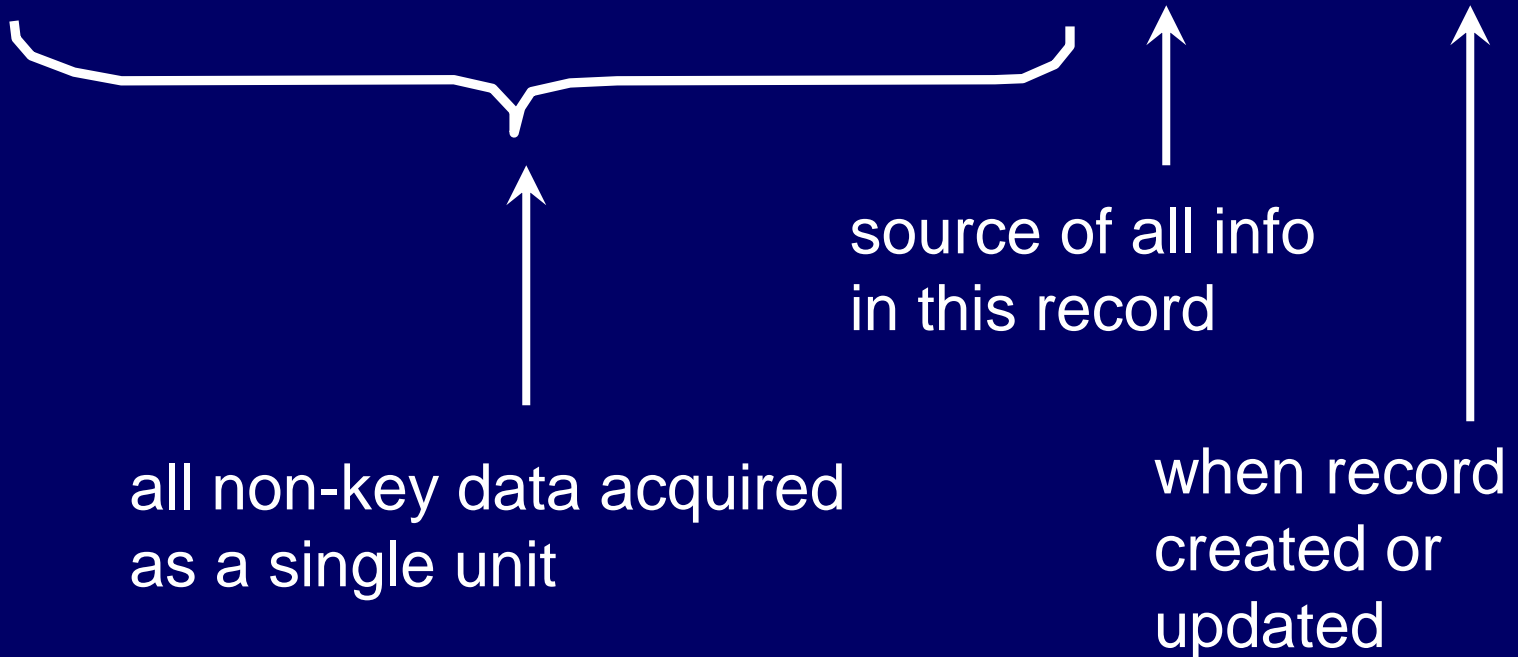
Row pedigree

Fact pedigree

Record-level metadata

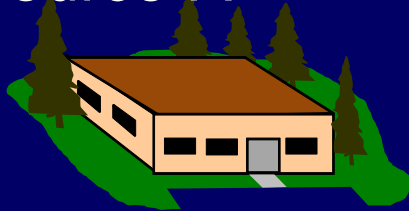
Source and update date for the whole record (all fields)

ID	Name	Street Addr	City / St	Source	Updt Dt
489735	John Smith	971 Pine Drive	Portland, ME	CA DMV	8/2/1997
489735	Mary Allard	6174 Huron St.	Albany, NY	NY DMV	4/13/2003
489735	Ty Kobb	572 Ottawa	Boston, MA	US Army	4/14/2003



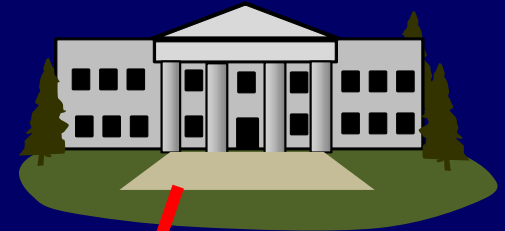
Individual facts may come from different sources.

Source-A

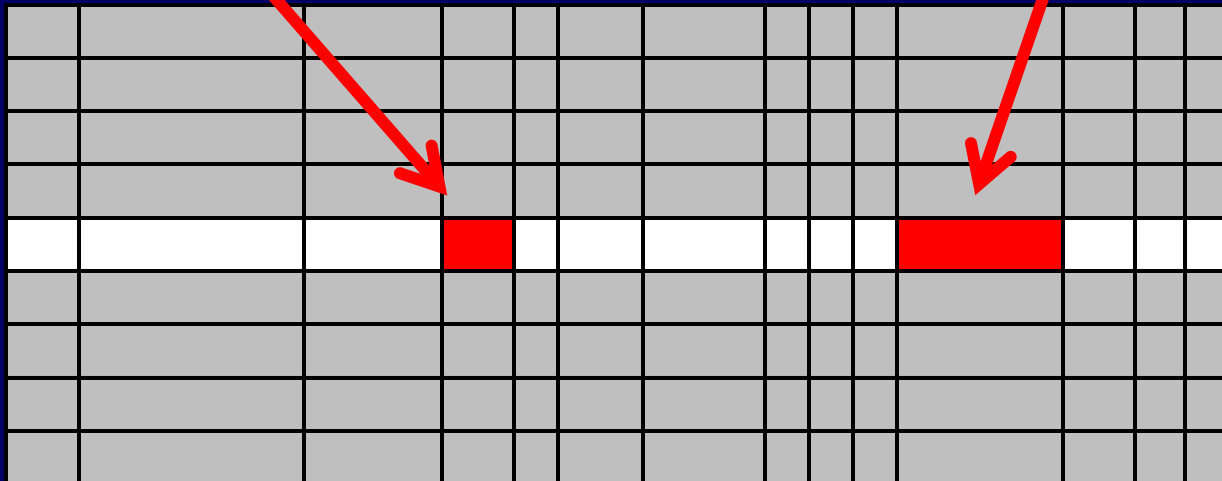


Fact A

Source-B



Fact B



Imbedded metadata – cell level

Some facts are acquired individually, unrelated to peer cells in a record.

Credit bureau record on person

Person ID	Name	SSN	SSN src	SSN updt	DOB	DOB src	DOB updt
489735	John Smith	587-98-1473	US Army	4/15/2001	4/3/1952	CA DMV	8/2/1997
489735	Mary Allard	589-88-8891	CitiBank	2/2/1997	3/9/1972	NY DMV	4/13/2003
489735	Ty Kobb	433-52-8743	Chase 57	6/2/2004	4/15/1978	US Army	4/14/2003

↑
fact

↑
where we got
the fact

↑
when we got
the fact

These 3 data
elements belong
together.

In academic papers, we use footnotes for fact-level metadata....at least citing where we got it.

a lab classroom on north end of the top floor of the physics building. We worked from a syllabus which Dr. Lee had written. One important thing it emphasized (and this was more of a mathematics issue) was expressing error and deviations by percentages rather than absolute amounts. I repeatedly listened to Don's same old example about an inch variance was tolerable when shooting a rocket to the moon, but was not tolerant when fitting two pieces of machinery together.

In my final year at La Sierra, I ended up teaching the lab myself. Some of my students included Joannie Hoatson, Dan Rathbun, and many religion majors¹³.

Oak Ridge National Laboratory

It may have been in my sophomore year that the Physics Department was visited for a week by a special education program sponsored by the Oak Ridge National Laboratory. They supplied a special truck and driver (who was also the lab technician for the operation) which came out from Tennessee¹⁴. This was part of a program which visited many smaller colleges; Dr. Riggs arranged

¹⁰ Donald Lee was registrar of the college. He taught this class on the side, probably to keep his hand in it.

¹¹ I remember in the Criterion, the college newspaper a poem which talked about long periods of time, and something to the effect that "...men grow old, old men grow older, a very long time, Physical Science class"

¹² Other people on campus described it as a "very responsible position" which, to my mind then, made it sound bigger than it really was.

¹³ As an aside, I think it is good that religion majors have some basic fluency in science and the scientific method so they don't make fools of themselves in sermons, particularly when talking about creation. Yet, as evidenced in the pages of the Adventist Review, many ministers and church leaders are ignorant of the scientific method, and make unreasonable claims about the nature of the universe, seismic activity, and other issues. One might have hoped that at least the editors of the Adventist Review would run some of these articles past credible scientists, but it seems that they have not.

¹⁴ Or at least had Tennessee plates.

Other data sourcing issues.

Meaning and context may require multiple sources and semantic data integration

Cause and effect may be measured in different systems or organizations.

Intuitive data integration.

Clue #1



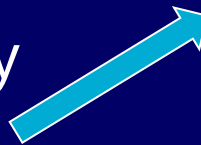
Clue #2



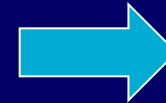
Clue #3



Seemingly
irrelevant



Clue #4



Information:
Solution to
the crime!

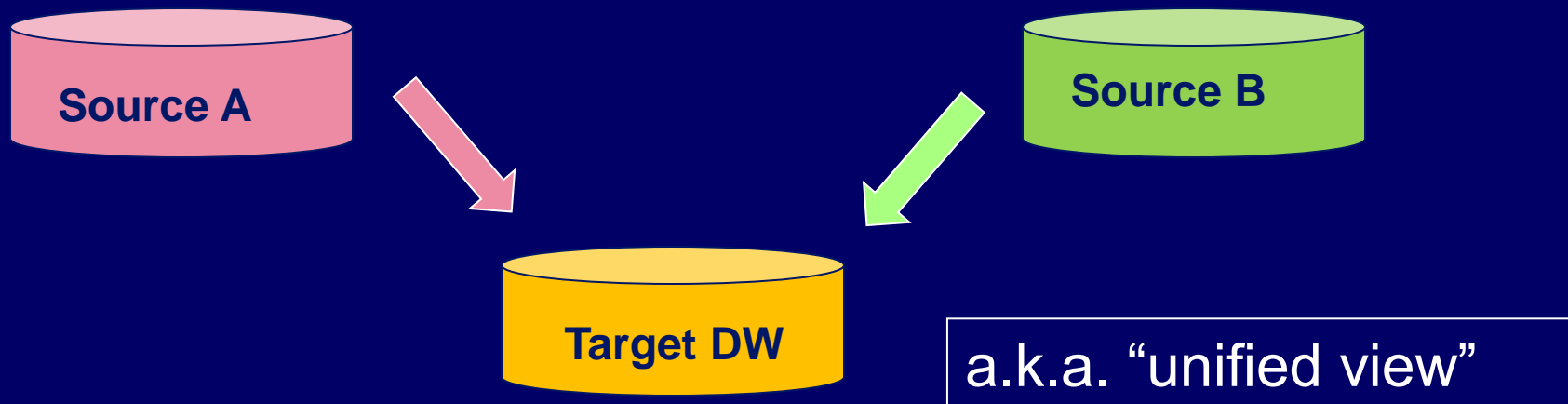
“A-hah!”

Integrating data to create information

Semantic data integration

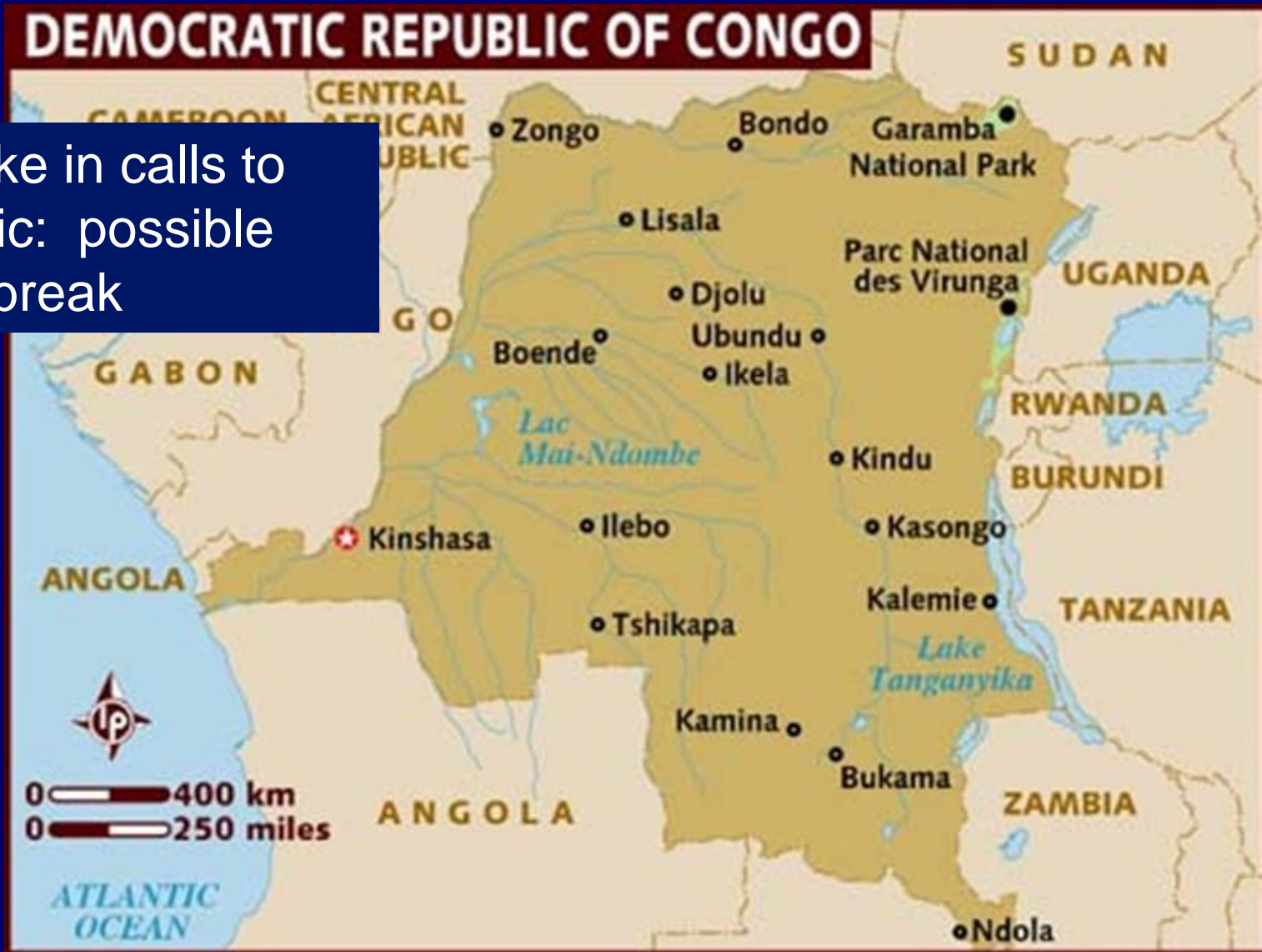
Bringing data together from two or more sources in a way that makes sense.

In a way that makes the combined results meaningful, reliable, and useful.



Virologists look for unexpected patterns of cell phone calls.

Spike in calls to clinic: possible outbreak



A photograph of a public library interior. The scene shows rows of bookshelves filled with books, with two green upholstered chairs in the foreground. The ceiling has fluorescent lights and exposed pipes. Three yellow text boxes are overlaid on the image.

A public library is an “information warehouse”.

Yes, some raw data, but mostly information (text, maps, photos, expression, sketches, art, illustration, etc.).

Findability is critical ! Google changed all that !

Crowd-sourced data integration.



Photo of Air Force One on a “secret” trip to Iraq.
Photographed by amateur in Sheffield, England.
Posted on web, went viral. Dec. 26, 2018

Reality



Facts &
data



Information

Shared
vocabulary

Shared vocabulary:

meaning of words, individually & in context.

def. of metrics and scope

meaning of phrases, sentences & metaphors

shared context

assumptions about “know-ability”

aesthetics



(of meaning)

Physician talking to a semi-lucid patient

Giving post-op instructions



“What did he say?”

Will patient remember it all?

Did patient understand it in the first place?

Reality



Facts &
data



Information



Expression



Communication



Understanding

(of meaning)

Reality



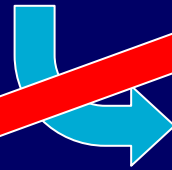
Facts &
data



Information



Expression



Communication



Understanding
(of meaning)

Expressor and audience
must share lexicon.

Good day Dearest one:
...I am a dying woman
in Nigeria...

Context gives credibility

Sender's track record, as perceived by the audience.

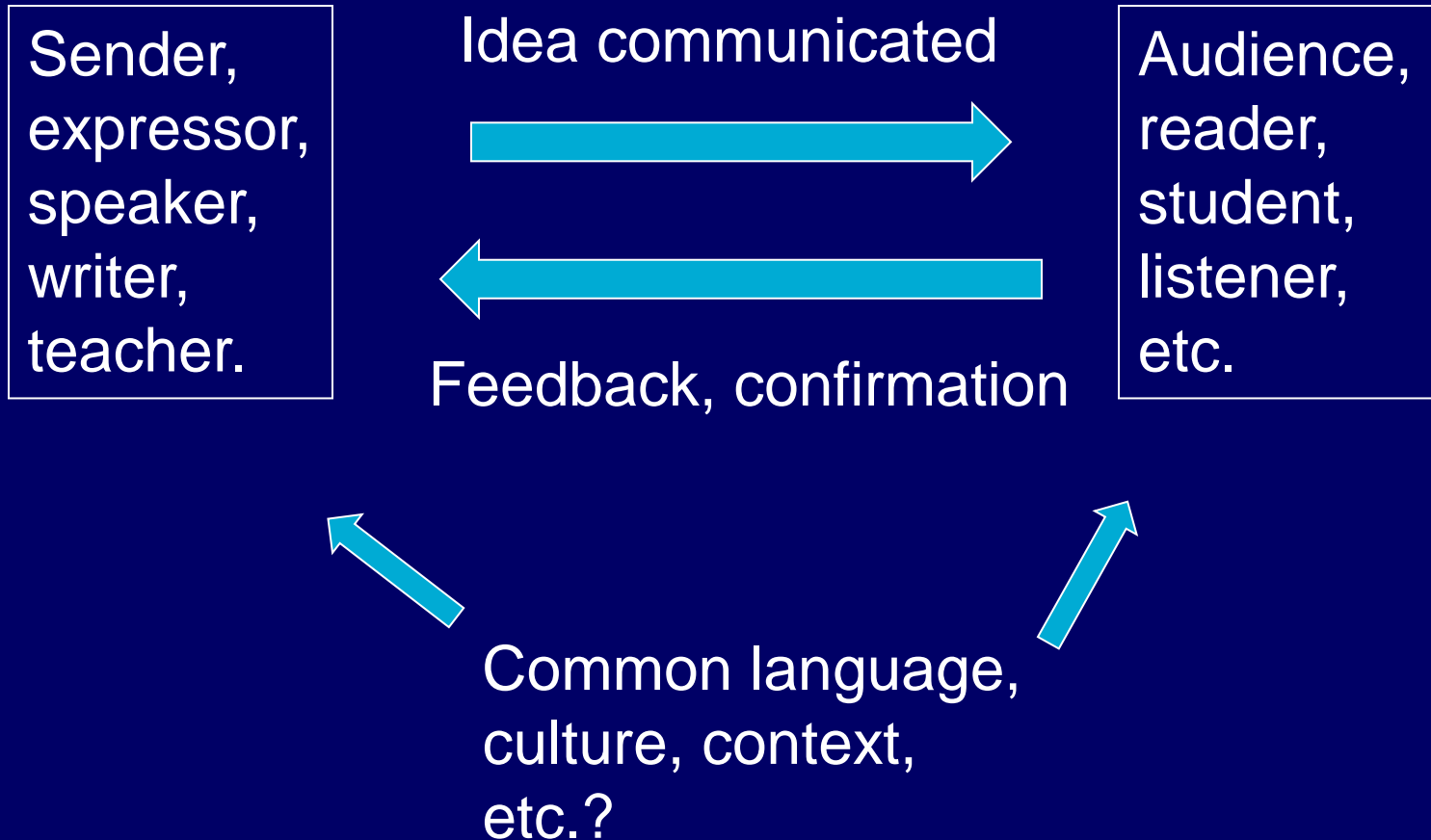
Warning sounds like previous unrealized warnings.



“Yea, we’ve been told to evacuate before. Nothing came of it.”

NOAA

“Communication has not occurred unless desired results of the communication have taken place.” --FAA



April 21, 2021

Expression design

brevity



Tedious
length



Ambiguity &
misunderstanding



Precision and clarity
of meaning

ALERT !

IMMEDIATE DANGER TO BUILDING OCCUPANTS

For more information, go to:

http://employee.alerts/building/safety/m384g12p_corp.com

Enter your Employee ID, location number, payroll code.

Click on the **RECENT ALERTS** button.

Re-enter your Employee ID, User_ID and password .

Select warning message appropriate.

Click on **MESSAGE READ** button.

Evacuate building quickly.

Walk, do not run, to exit.

Expression of information for decision-making

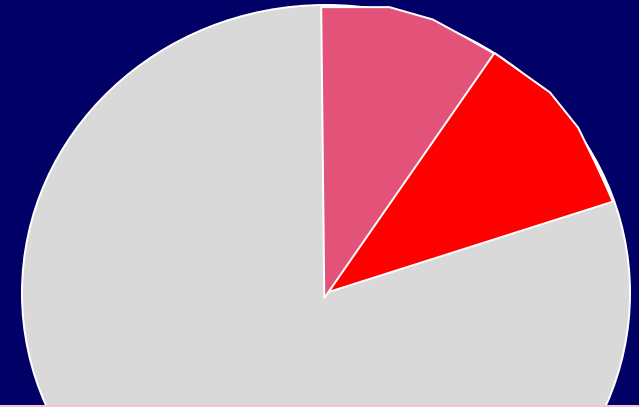
Two purposes of graphics

1. For *you* to figure out what is going on

Exploration, analysis, trying different correlations, etc.

2. Explain to *decision-makers* what they need to know about reality.

“5-second rule”

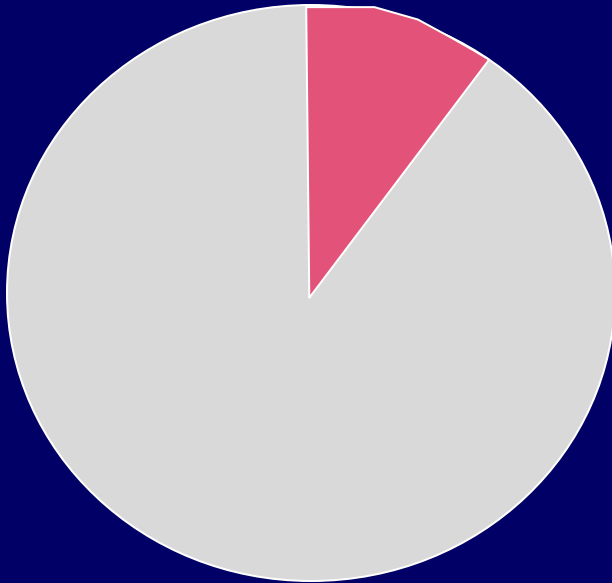


1. Understand scope, metric, and what you are seeing.

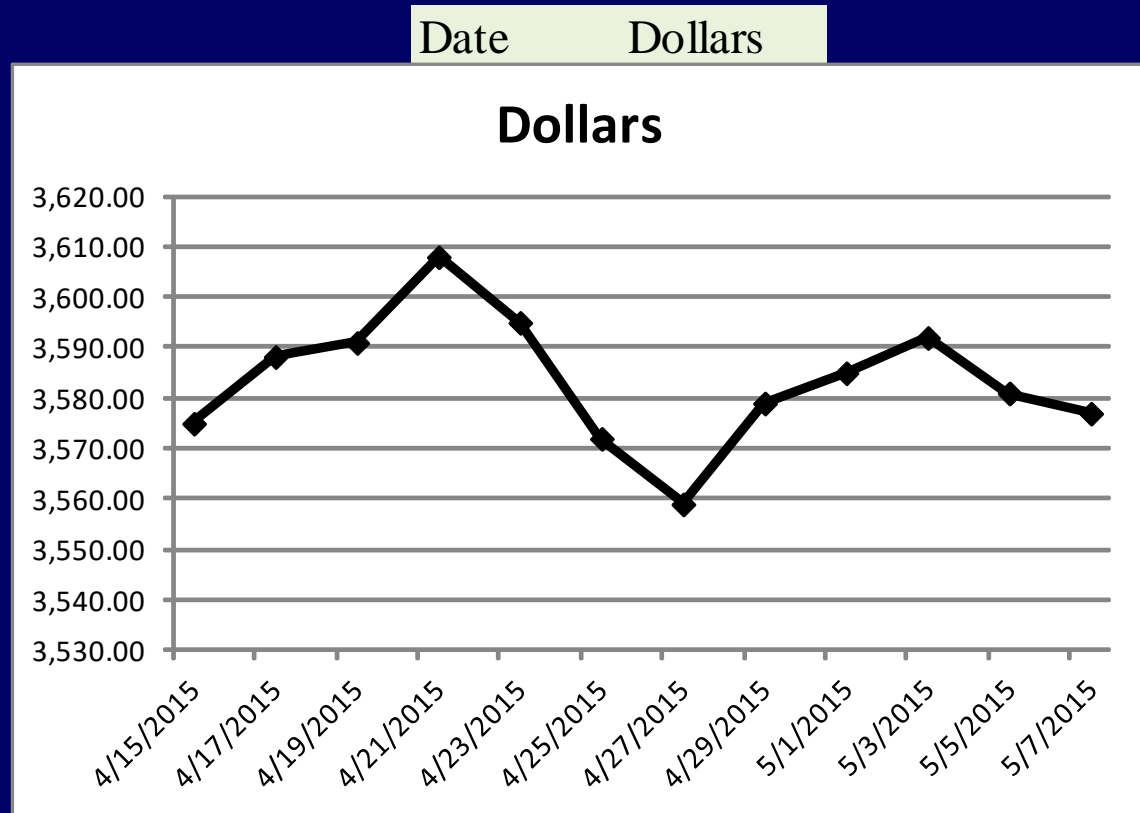
2. Understand the behavior being emphasized...why this is important.

Clarity

“5-second rule”

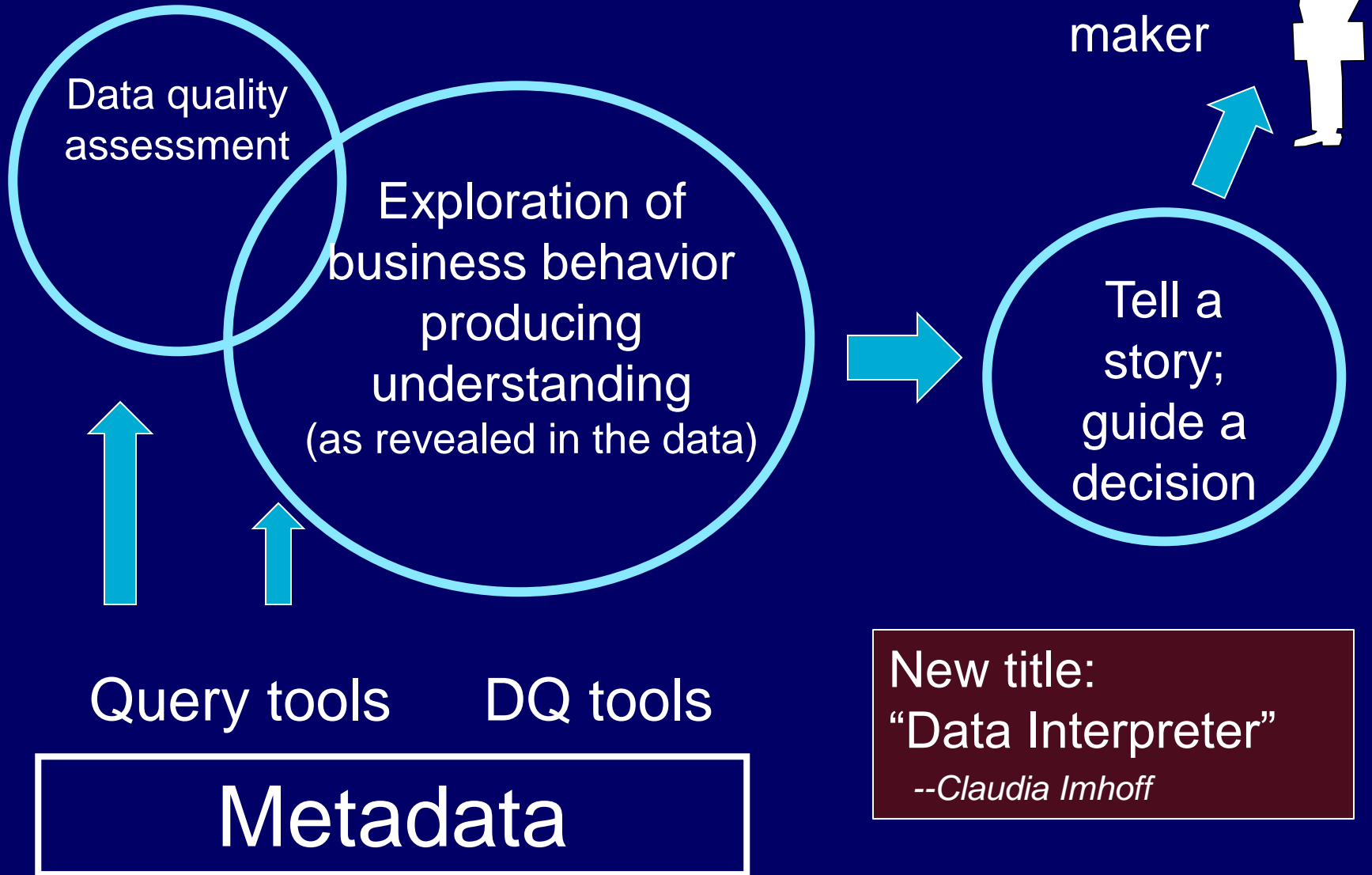


1. Understand scope, metric, and what you are seeing.

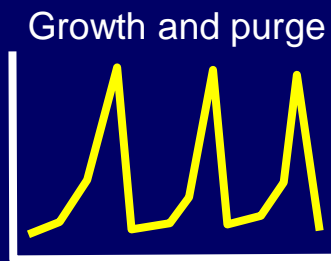
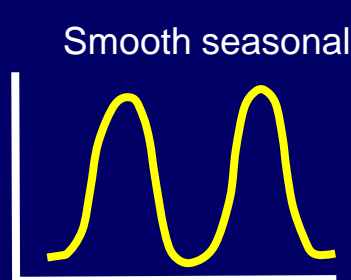
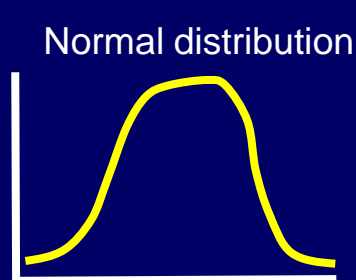
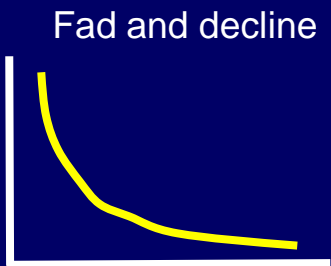
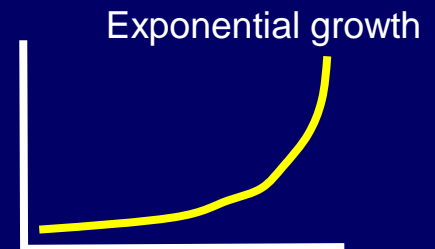
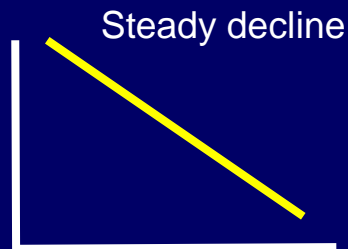
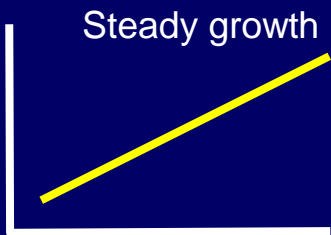


What's wrong with this chart?

Looking at the data



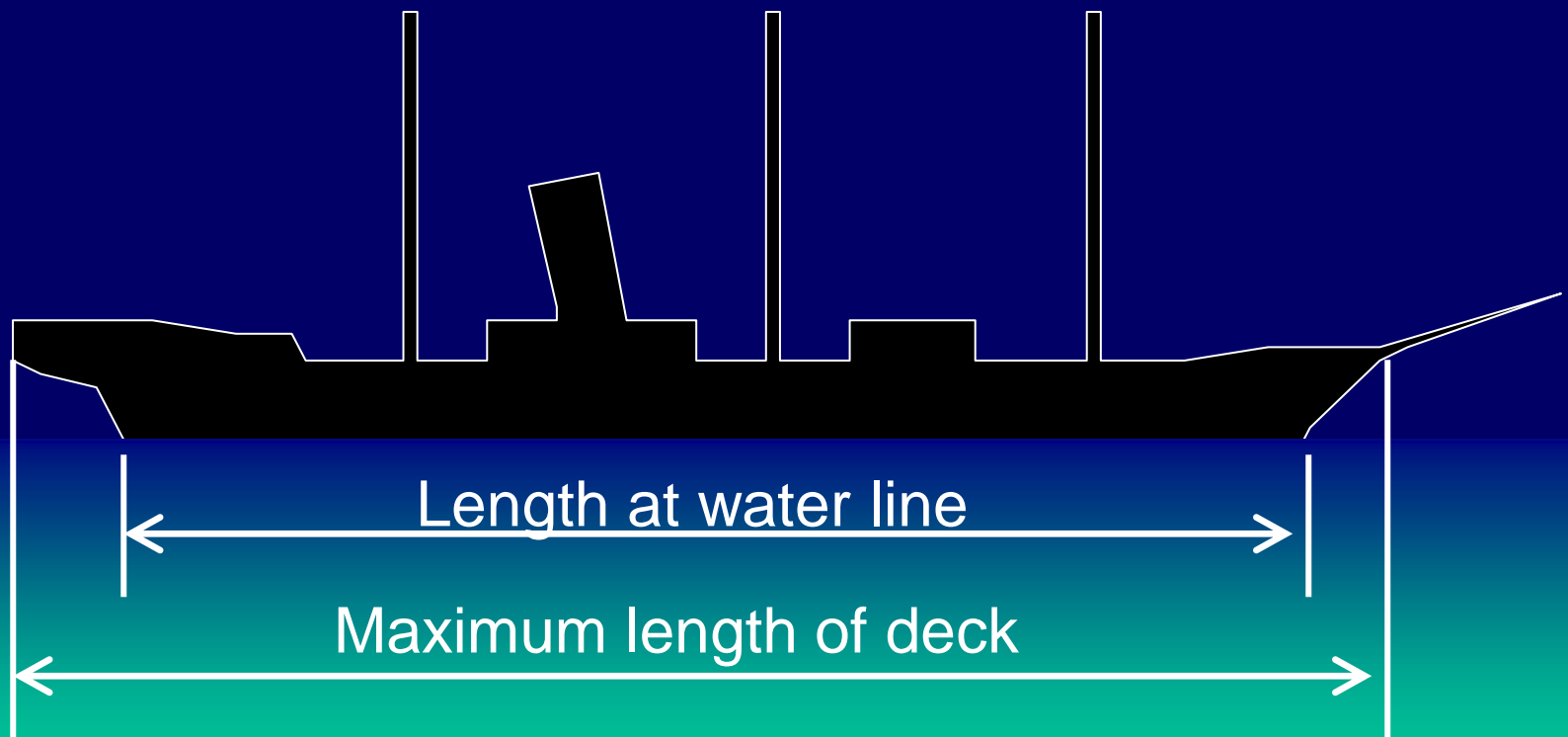
Histograms and line charts both imply shapes. And shapes can tell stories.



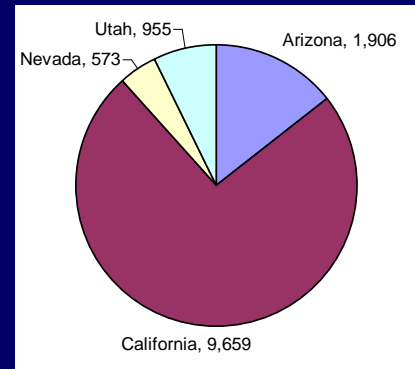
Ambiguities of most measures

Any measure must be clearly defined

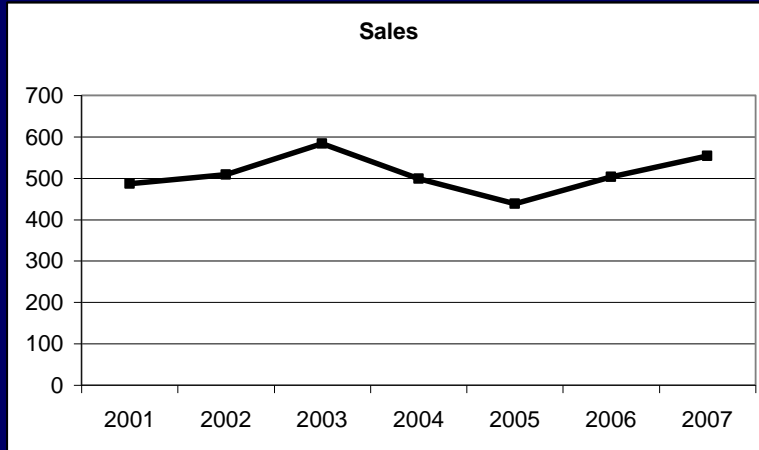
With more tediousness than most people tolerate.



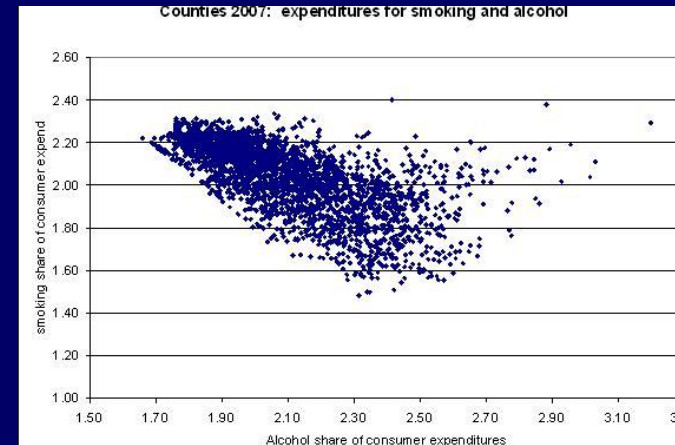
Kinds of charts



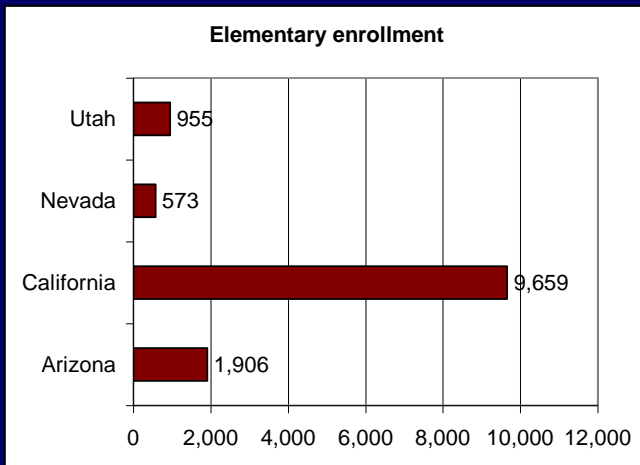
Pie



Line

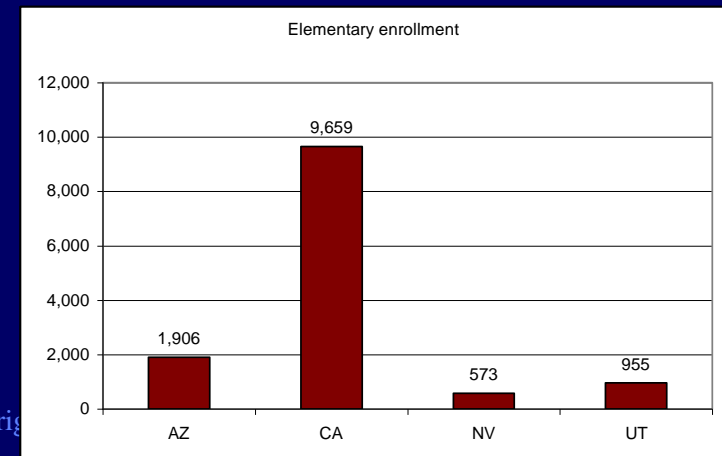


Scatter



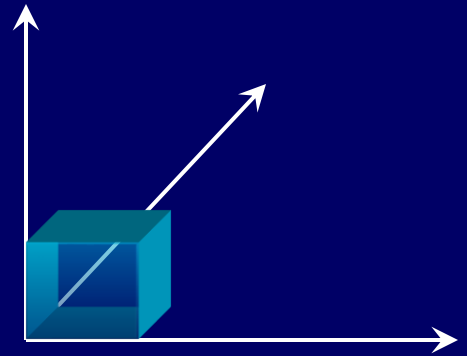
Bar

Column



Dimensions

A way of subdividing the whole of activity, behavior, or existence.



Categorical

- Categories, classes, types, kinds
- Geography or political units
- Organizational units
- Product groupings

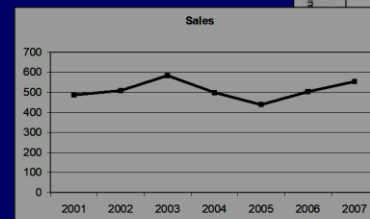
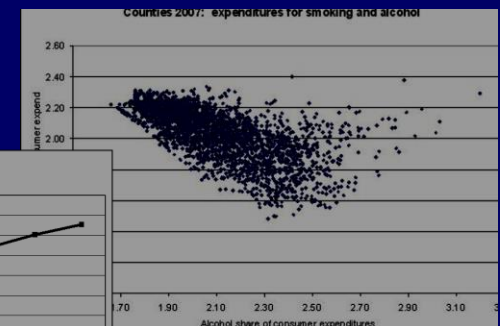
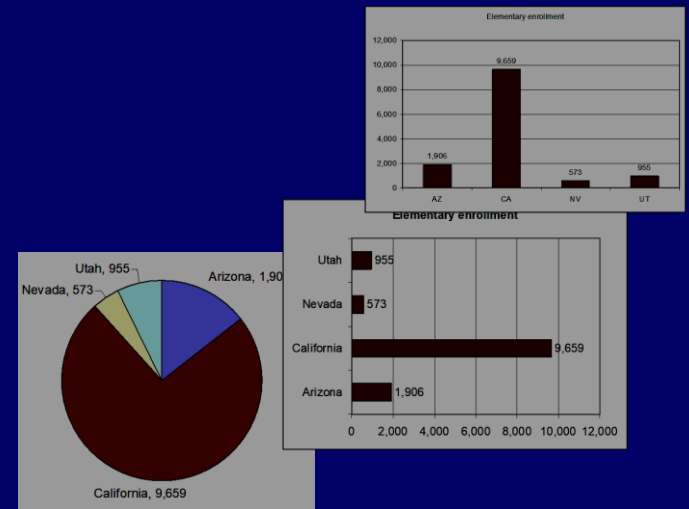
Linear (or “quantitative”)

- Time and periods of time
- Size of things

Other linear attributes

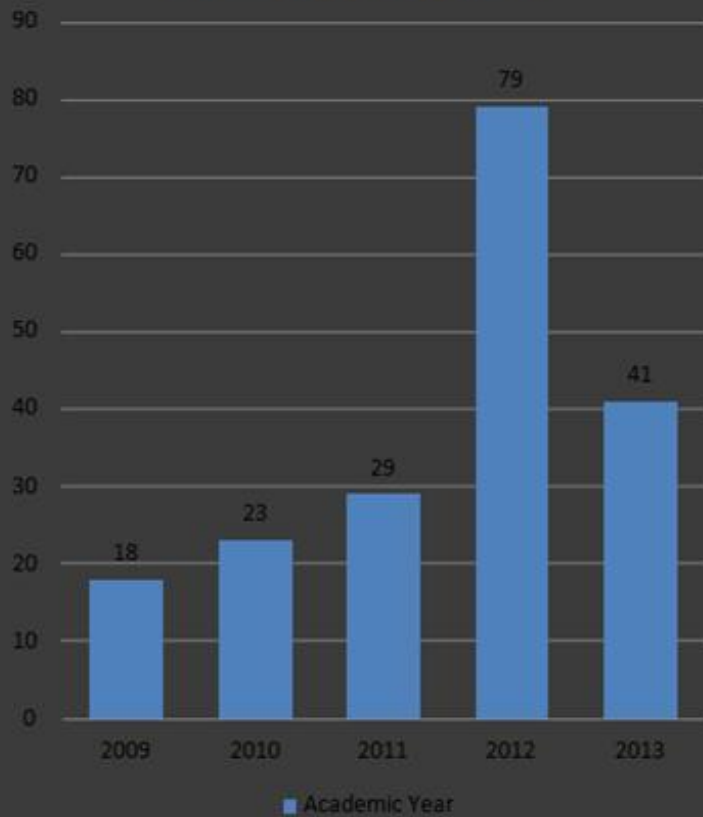
Sequential steps

Categories, events, or concepts with a natural order, but not exactly linear

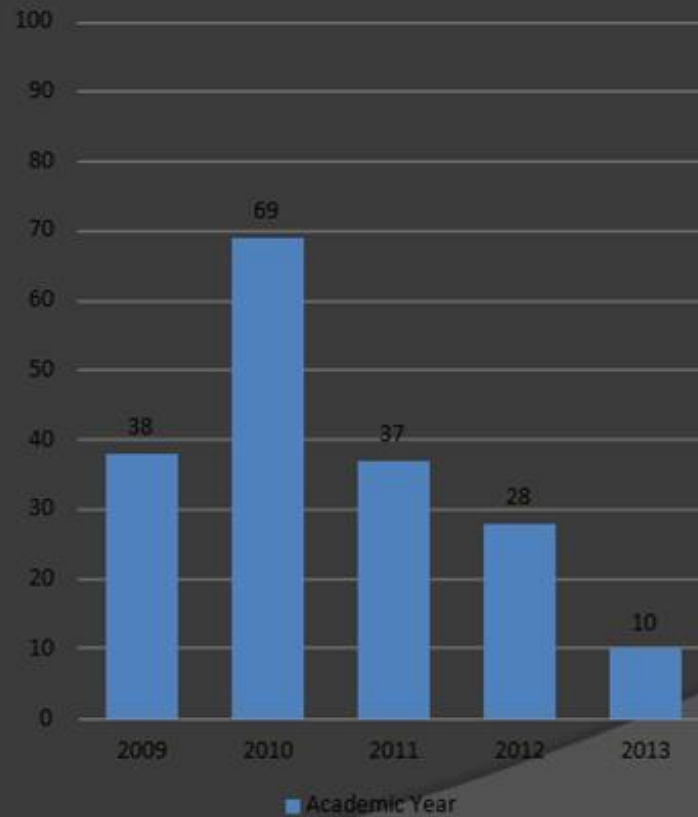


Online and Off-campus Enrollment

Online Programs Enrollment



Off-campus Programs Enrollment



Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.

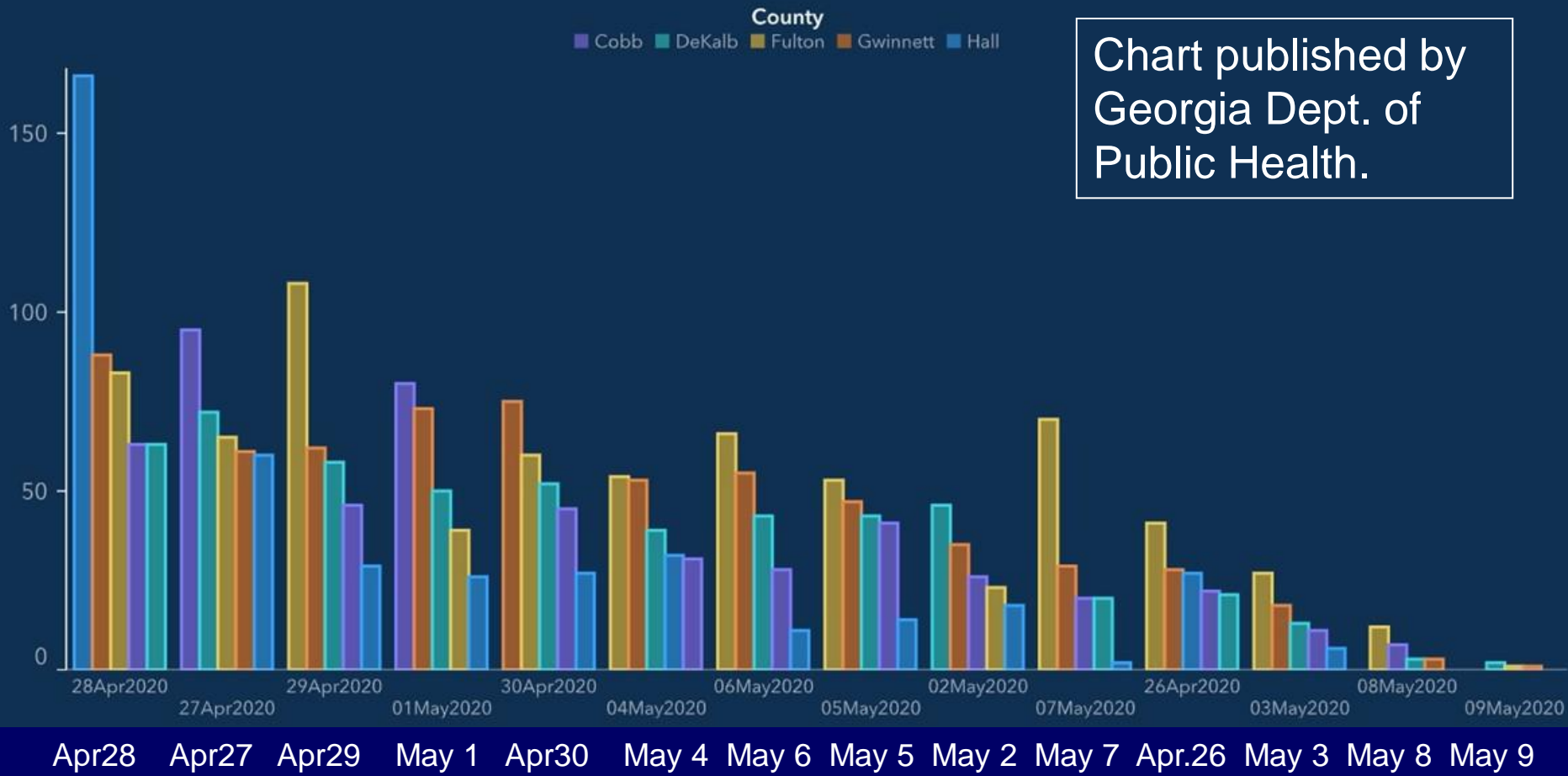
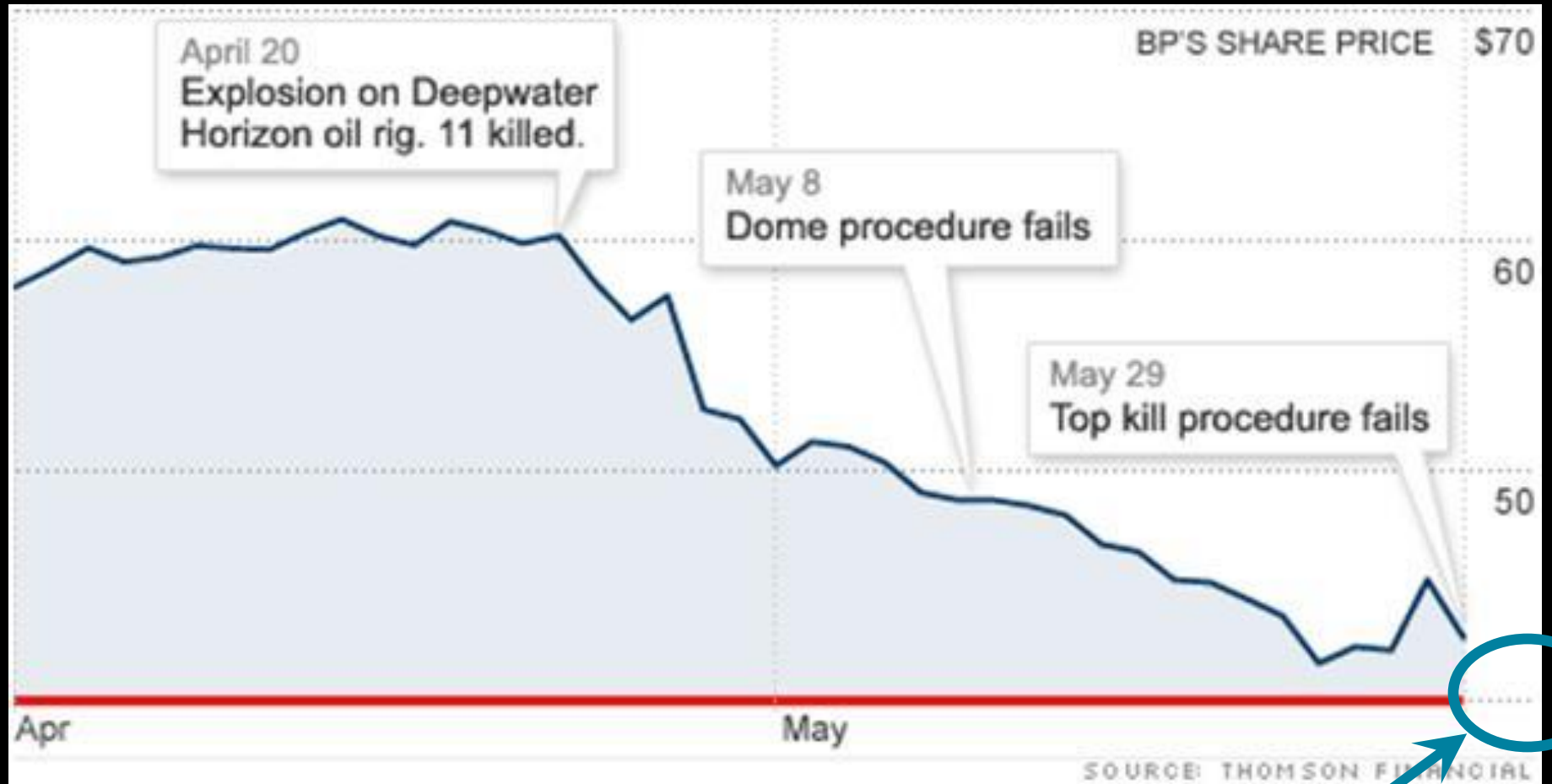


Chart published by Georgia Dept. of Public Health.

Deliberate misrepresenting reality. Implying a downward trend.

Current problem: CNN web site



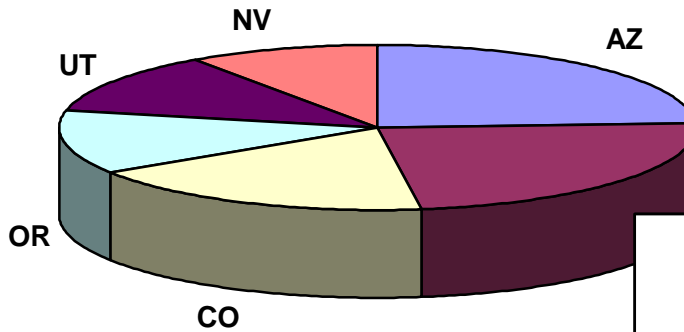
Non-zero base of chart

Bill Colwell's rule of technology:

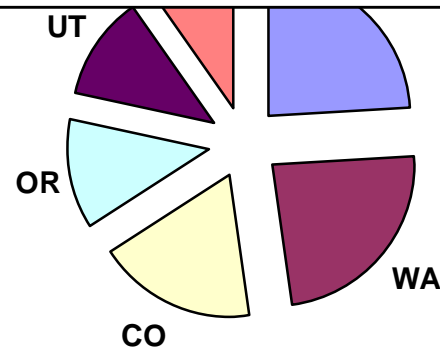
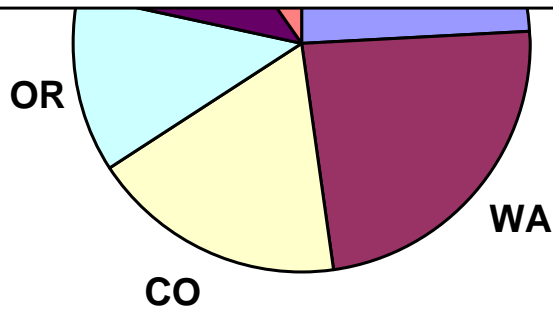
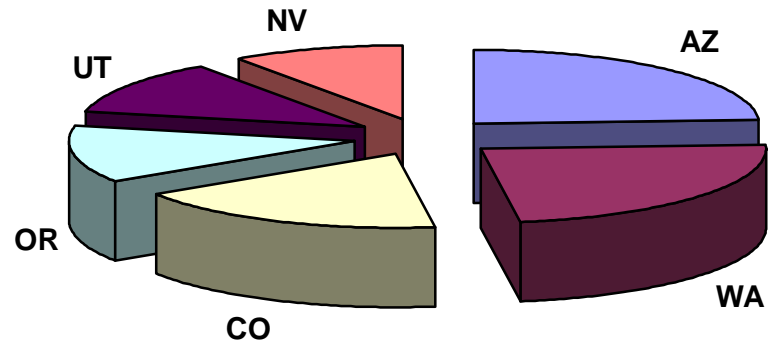
Just because you *can*
do something,
doesn't mean you
should!

Advanced “artistic” techniques

Elementary enrollment



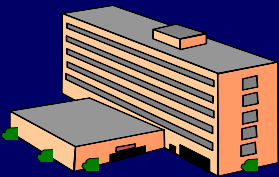
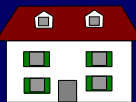
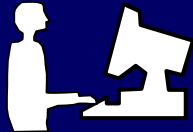
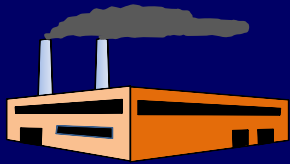
Elementary enrollment



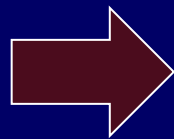
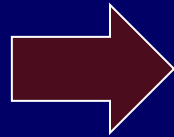
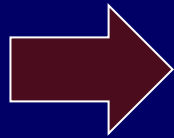
The end...

...unless we keep going.

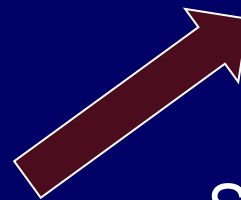
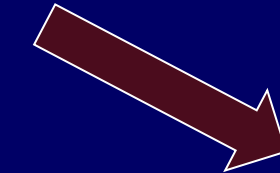
Reality



Raw, original research



Secondary ("survey") research



Real tests, raw data

Surveying previous research, seeking patterns and trends

Raw data

Census records



School yearbooks

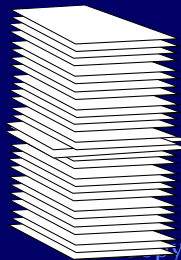
Birth & death

Draft & military

Immigration & naturalization

Property rec'ds

Ship manifests



Personal papers

Raw data

Census records



School yearbooks

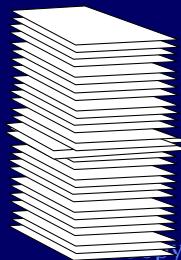
Birth & death

Draft & military

Immigration & naturalization

Property rec'ds

Ship manifests



Personal papers

Potential problems:

Mis-identity:
name homonym

Errors in raw data:
spelling

census enumeration
clerical date errors

Missing data
destroyed documents

Raw data

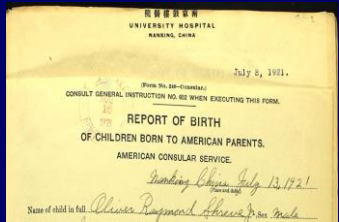
Derived data



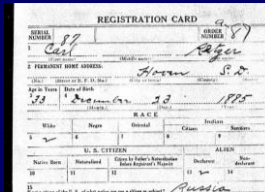
Census records



School yearbooks



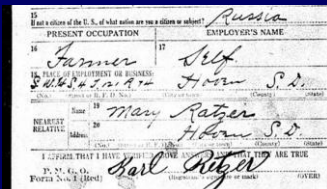
Birth & death



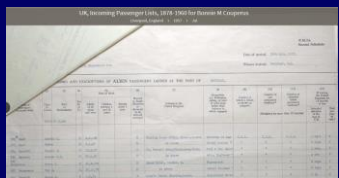
Draft & military



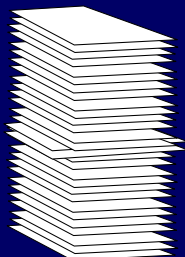
Immigration & naturalization



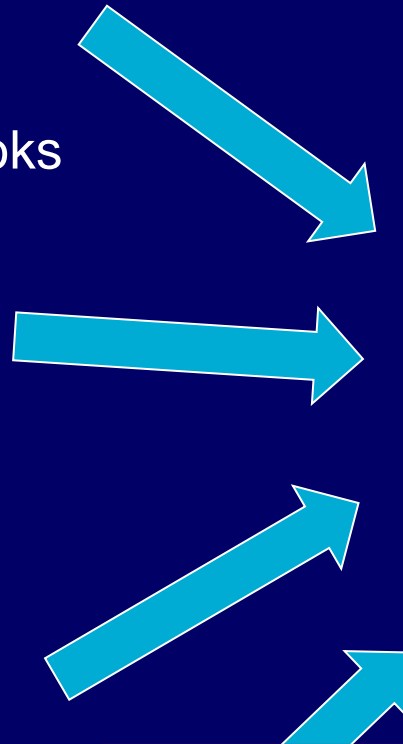
Property rec'ds



Ship manifests



Personal papers



Life narrative

Security clearance

Obituary

Secondary error checking:
Do dates make sense?
Do places make sense?

Raw data

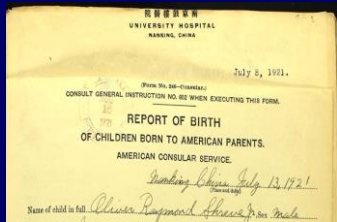
Derived data



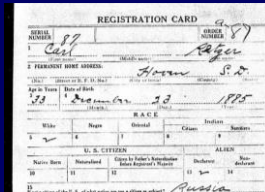
Census records



School yearbooks



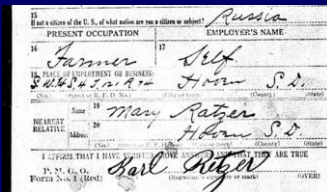
Birth & death



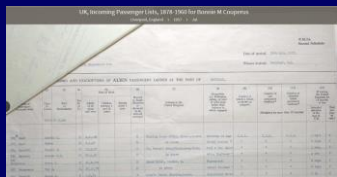
Draft & military



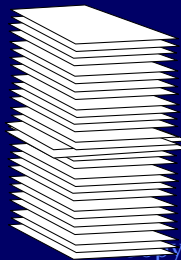
Immigration & naturalization



Property rec'ds



Ship manifests



Personal papers



Integrated data:
Raw data
in context

Secondary DQ
test for
reasonableness

Raw data

Derived data



Census records



Birth & death



Immigration & naturalization



Ship manifests



School yearbooks



Draft & military



Property rec'ds



Personal papers

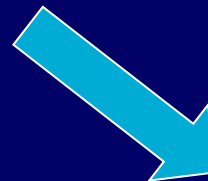
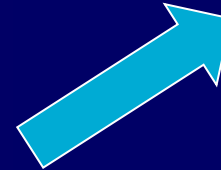


Secondary error checking:
Do dates make sense?
Do places make sense?

Life narrative

Security clearance

Obituary



Hagiography.

A biased narrative.

Perhaps with contrived dialogue.

Screenplay.

“based on a true story...”



The end...

...unless we keep going.