UNIVERSITY of HOUSTON
DEPARTMENT OF COMPUTER SCIENCE

Dr. Nouhad Rizk

713-743-3710 njrizk@uh.edu http://cs.uh.edu/~rizk/

Expertise

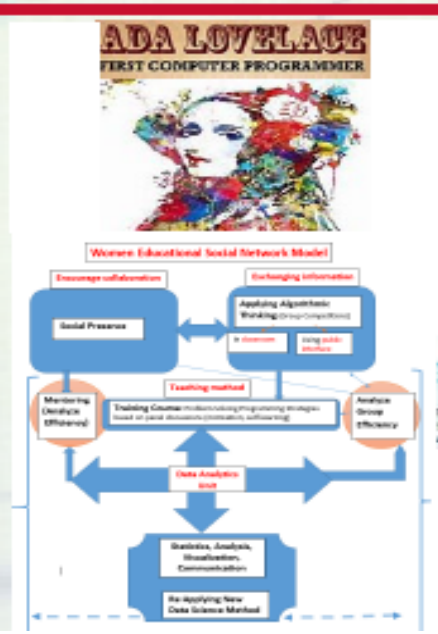
- Data Science and Data Mining
  - Educational Data Mining
  - Prediction
  - Clustering
  - Association rules
- Machine Learning
  - Supervised and Unsupervised Learning
  - Social Network Analysis
- Expert Recommendation Systems
  - Real-Time Skill Assessment
  - Decision Making Designs

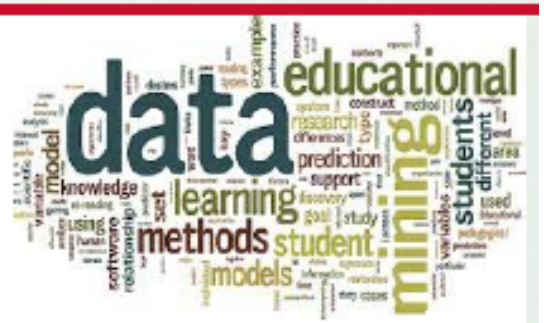
Research Projects

- Using coding skills to enhance Recruitment and Retention of Women in STEM
- Question Asking Strategies for Academic Early Warming Systems
- Data science techniques to Improve Academic Performance and Enhance Student Retention
- Data Analytics to Maximize Algorithmic Thinking
- Finding and evaluating group structure in classroom networks

Contributions & Impact

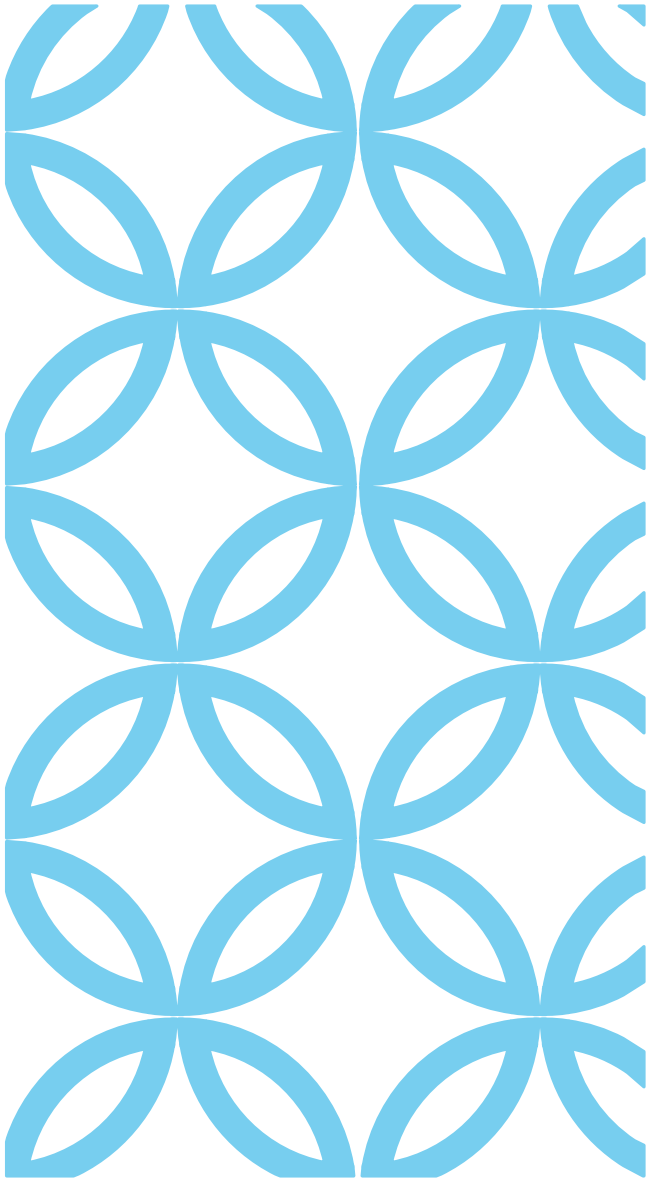
- Developed Machine Learning Models to Predict Student Performance Frameworks
- Developed Web-Based Expert System for Evaluating College Course
- Designed Performance Prediction and Self-Assessment Systems
- Developed Strategies for Academic Early Warming Systems





EDUCATIONAL DATA MINING (DATA SCIENCE)

Dr Nouhad Rizk
njrizk@uh.edu



Educational Data Mining is an emerging discipline, concerned with developing methods for exploring data that come from educational settings and using algorithmic methods to better understand students, and the settings which they learn in.

DEFINITION

EVOLUTION OF EDUCATIONAL RESEARCH

Then

Educational research had to be conducted with small amounts of data

- Small data sets from a single school
- Limited number of researchers

EVOLUTION OF EDUCATIONAL RESEARCH

Now

Educational research is enhanced dramatically by:

- Access and use of large volumes of data
- Use of datamining methods to extract meaningful data and discover patterns in learning

  E.g. clustering

FEASIBLE MEANS OF DATA COLLECTION IN OUR RESEARCH

- Short-term outcome data based on student performance from COSC 1306-2430
- Long-term outcome data available though nation-wide database
  - National student clearinghouse
- Student-engagement data stored in scientific databases
  - Pittsburgh science of learning Center

Evaluate the correlation of student preparedness based on the *self-assessment* and actual letter grade obtained on exam at the University of Houston

SHORT-TERM OUTCOME DATA

Use scores obtained from 1306, 1430 and 2430 to predict future performance in higher level COSC classes based on different “classes” of students

SHORT-TERM OUTCOME DATA

DATA SCIENCE FOR SERVICE CHANGE....

WHAT IS DATA SCIENCE?

Data Science

Applying advanced statistical tools to existing data to generate new insights

Service Change

Converting new data insights into (often small) changes to business processes

Smarter Work

More efficient and effective use of resources

WHAT COMPLEMENTS DATA SCIENCE?

(and is really good stuff to do)

Approach	Process	Outcome	Examples
Performance Management	Define, visualize, often using dashboards	Meet goals and targets	SF Scorecard, PublicWorks Stat & Stat starter kit
Evaluation	Assess a project, program or policy design or results	Better investment of resources; Better policy decisions	Evaluation of transitional-
Policy Analysis	Define and assess alternatives using a broad range of tools	Report or memo with policy or program recommendations	Shape Up Policy Analysis
Open Data	Publish civic data for use by the City and the public	Easier data sharing and reporting, new tools or services built on data	SFPUC Adopt a Drain
DataScience	Identify insights using advanced statistics tied to a service change	Smarter work “on the ground” in real time	…..

WHAT COMPLEMENTS DATA SCIENCE?

(and is really good stuff to do)

Approach

Performance Management

Evaluation

Policy Analysis

Open Data

DataScience

All approaches can lead to service improvement. It's about choosing the right tool for the job (and sometimes combining them)!

WHAT'S IN THE DATASCIENCE TOOLKIT?

Statistical Methods | Tools | User Experience Research

Sentiment analysis

Time series analysis

Data mining

Multilevel modeling

Missing data imputations

Classification and clustering

Survival analysis

Pattern recognition

Principal component and factor analysis

AB testing

Machine learning

Forecasting

Propensity score matching

Logistic, multinomial and multiple linear regression techniques

Network analysis

WHAT'S IN THE DATA SCIENCESF TOOLKIT?

Statistical Methods | **Tools** | User Experience Research

Languages
- Python
- R
- SQL
- Javascript
- NodeJS

Libraries
- SciPy
- Pandas
- Scikit-learn
- GPText
- OpenNLP
- Mahout
- +many others

Data Engineering
- Profiling
- ETL
- Job notices
- APIs
- Optimized data pipelines
- Optimized data storage/access

Visualization
- D3.js
- Gephi
- R
- Leaflet
- PowerBI
- ggplot2
- shiny

WHAT'S IN THE DATASCIENCESF TOOLKIT?

Statistical Methods

Tools

User Experience Research

Iterative Prototyping

Photo journaling and documenting

Service blueprinting

Journey mapping

Ride-alongs

Process mapping

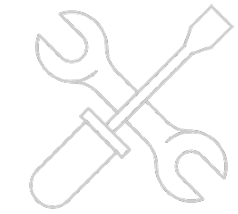
Ethnographic field research and user observation

Usability testing

WHAT IS **NOT** DATA SCIENCE?

✓ This

Service change

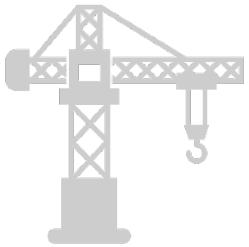


Small changes

Use existing data

✖ Not that

Academic research

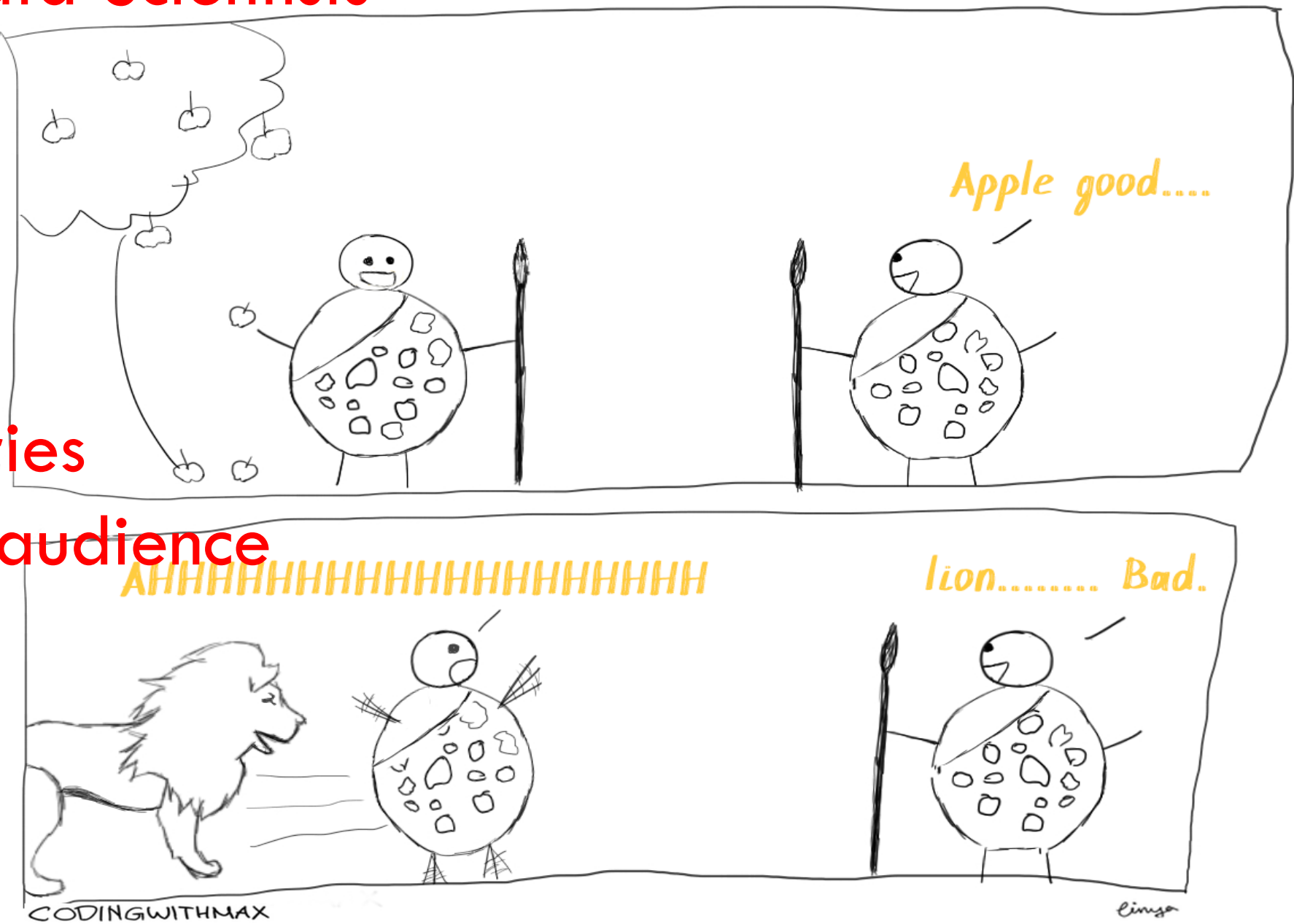


Major overhauls / service disruptions

Collecting new data (mostly ;)

Storytelling for Data Scientists

Turn data into stories
to persuade your audience



Storytelling

Data analysis is only half of the story

Data analysis is a very important skillset for scientists, because models are built on the results that we see in experiments, and if we are able to properly analyze our experimental data, we are able to formulate models that better represent reality.

The other half is that you also need to be able to:

1. Communicate your findings to others
2. Convince others that what you've found is indeed correct

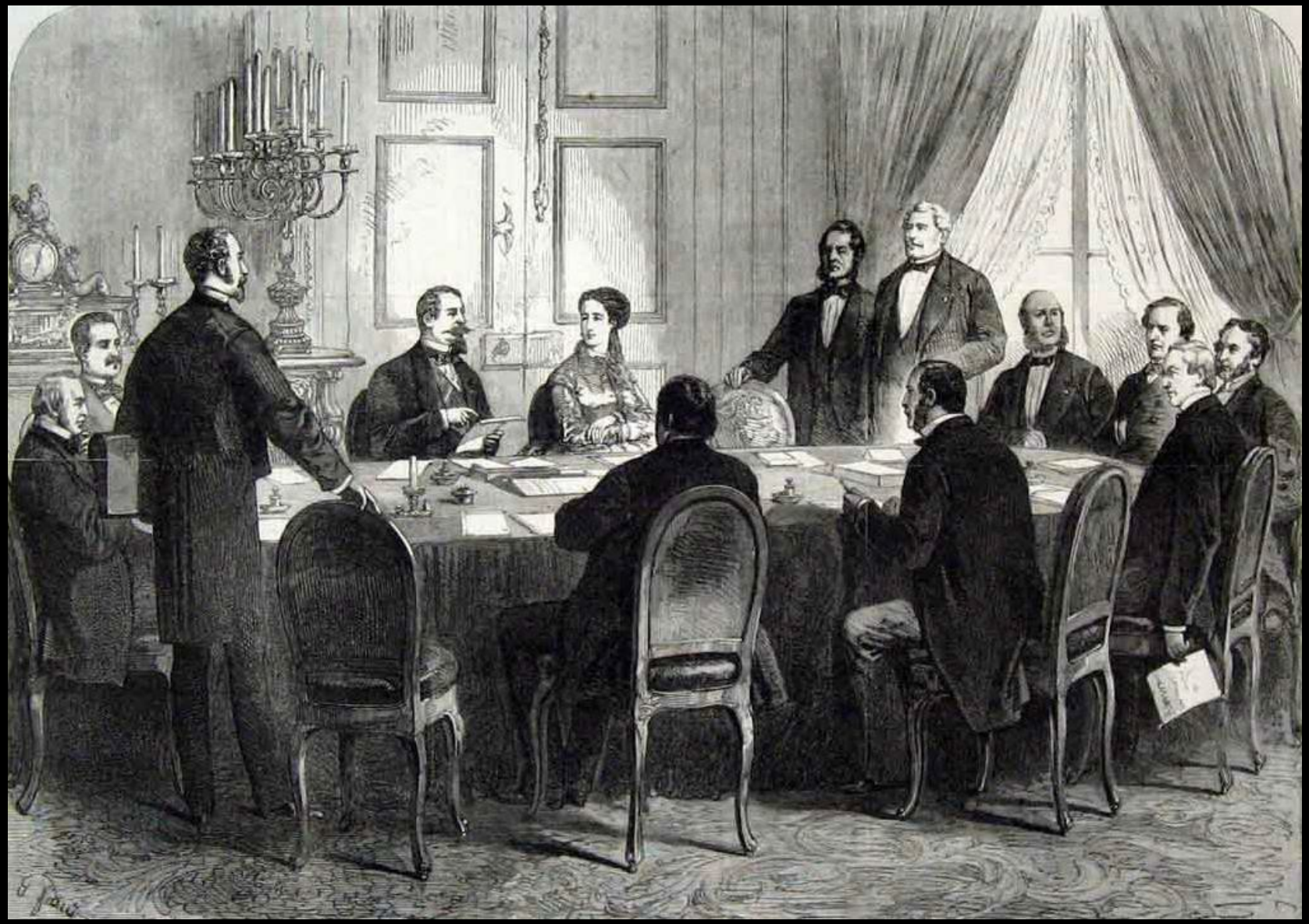
During a 2009 interview, Google's Chief Economist Dr. Hal R.Varian stated, "The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades." Fast forward to 2016 and many businesses would agree with Varian's astute assessment.

LET ME TELL YOU A (DATA) STORY.

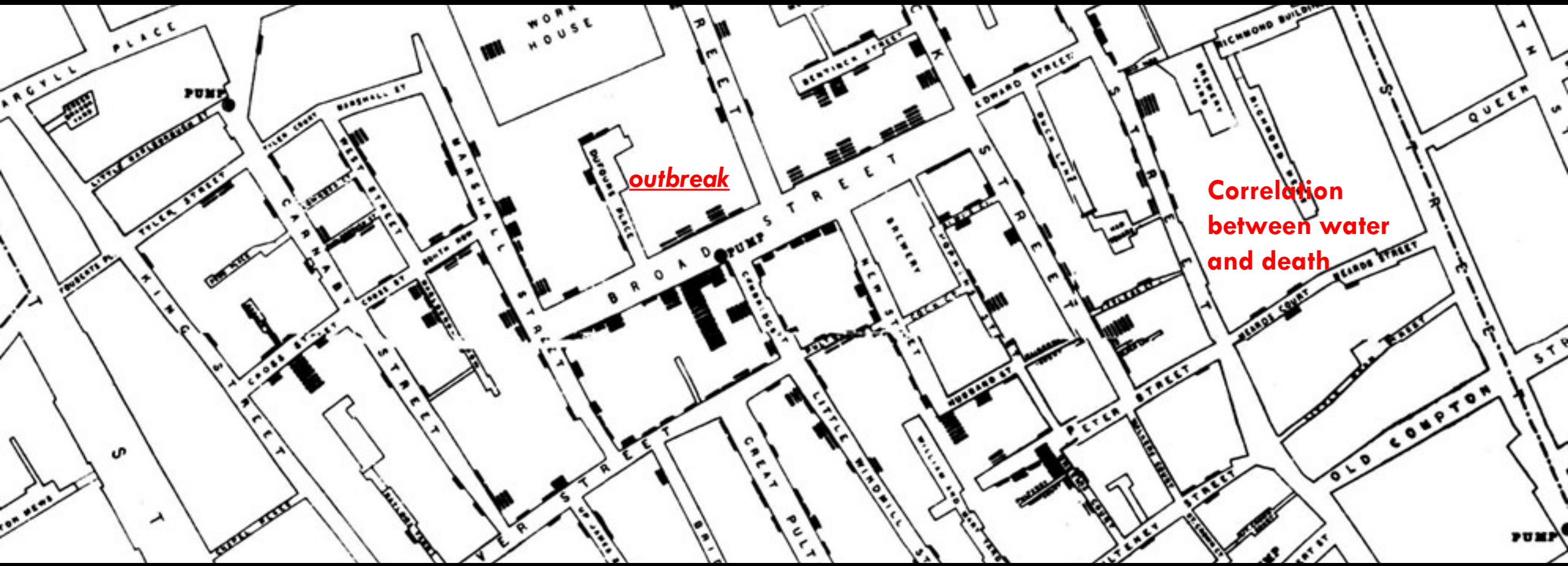
The Red Granite kerbstone
marks the site of the historic
BROAD STREET PUMP
associated with Dr. John Snow's
discovery in 1854
that Cholera is conveyed by water

LOND ON 1854

Within 3 weeks 127 died …why this disease is being spread



Stories are powerful. Cholera
Dr. John Snow told a *story* to prompt an action.



It's a myth that he used this map to convince people to remove the pump handle at Broad Street.

Story ? time, data, valgus (how many death),trend

Collect information
Who lives there?
When they died ?
Why they did not die?
Why they died?

Evidence : Brewery and workhouse survived

The ways in which organizations deliver business intelligence and analytics insights are evolving, notably in the rising use of what is called data storytelling.

This trend is an extension of the now dominant self-service model of Business Intelligence, combining explorative data visualization with narrative techniques to deliver insights in a way that engages with decision makers in a compelling and easily assimilated form.

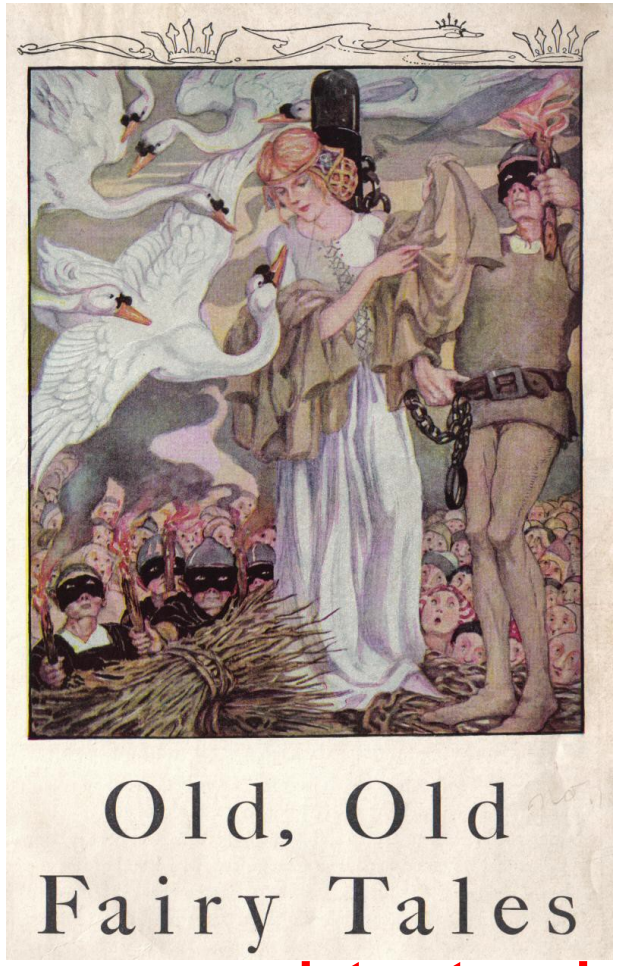
What data storytelling is?
How it is evolving and how to best use it to go beyond reporting and dashboarding.

KEY ISSUES

1. Beginning: What data storytelling is and why it matters
2. Middle: The how of data stories
3. End: Moderating the dark side of data narratives
   ➔Presenting complex data using a multidimensional aspects of visualization ➔discovery of correlation between water pump and the death of many people ➔prompt for an action

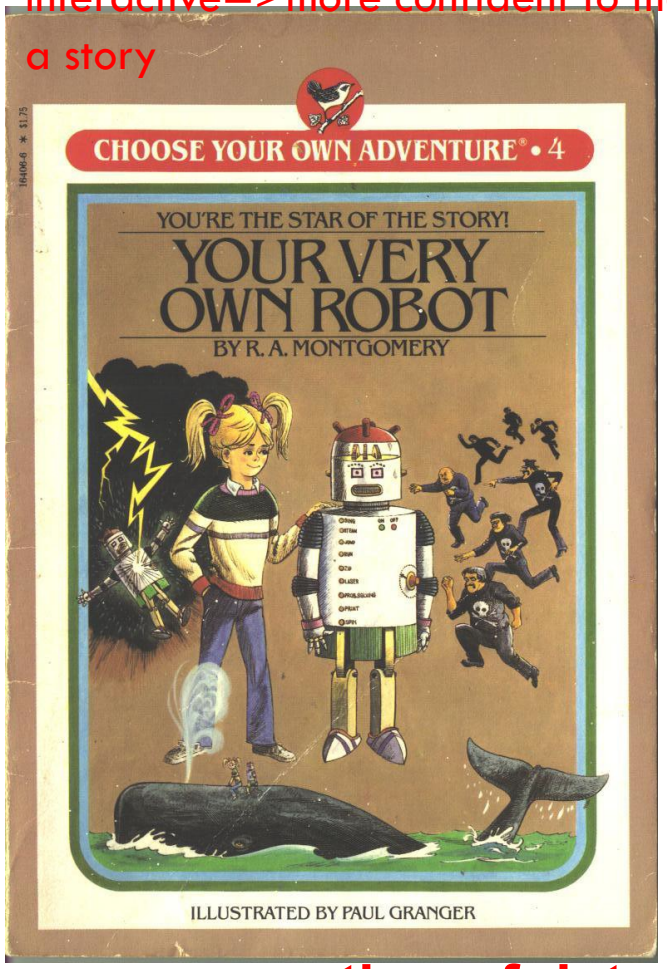
Story ➔Deliver information to get an action

No Fairy Tales: Data Stories Aren't About "Happy Ever After" but Options and decisions



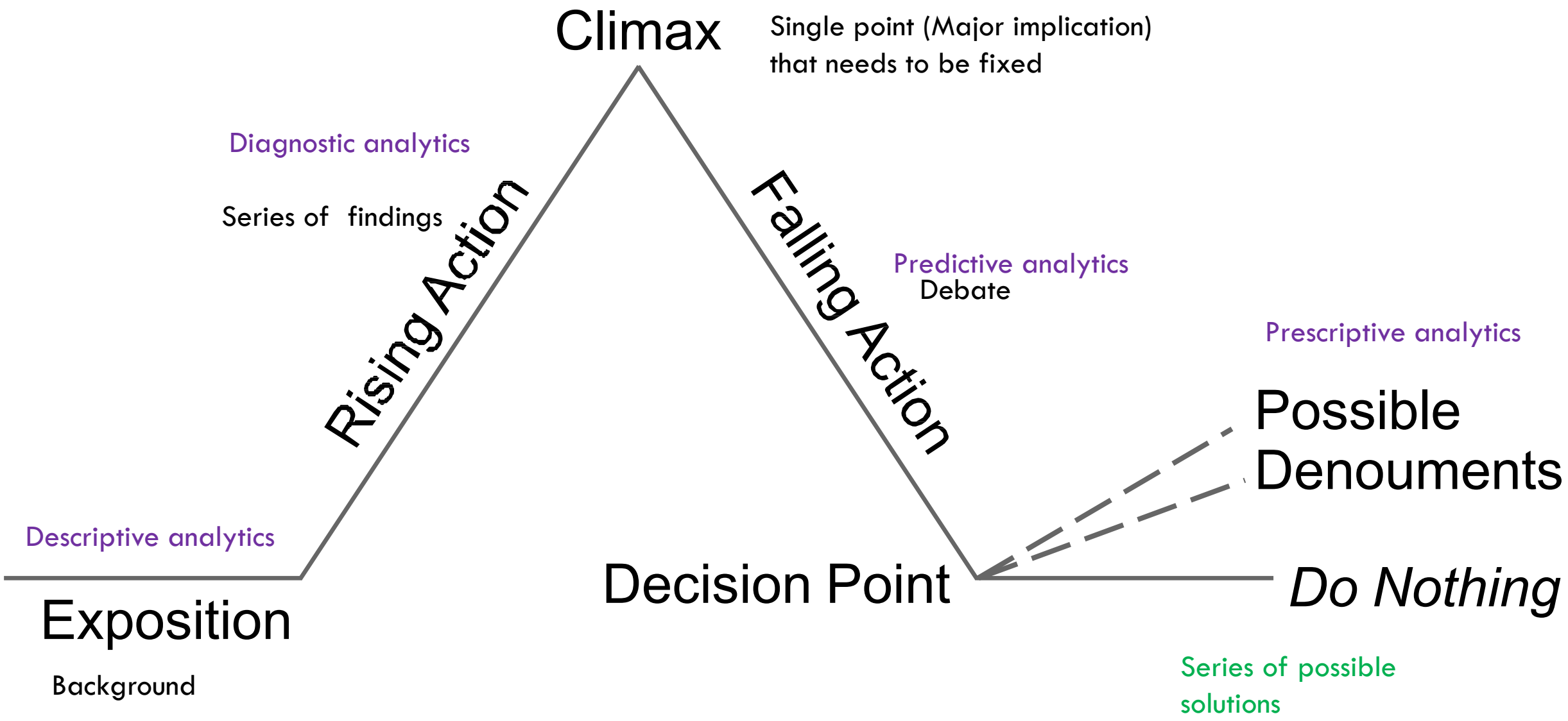
VS.

Interactive=>more confident to make a story



data story is not about the passive compensation of data (representation) =>it is interaction between audience and the data to get them closest to make decisions

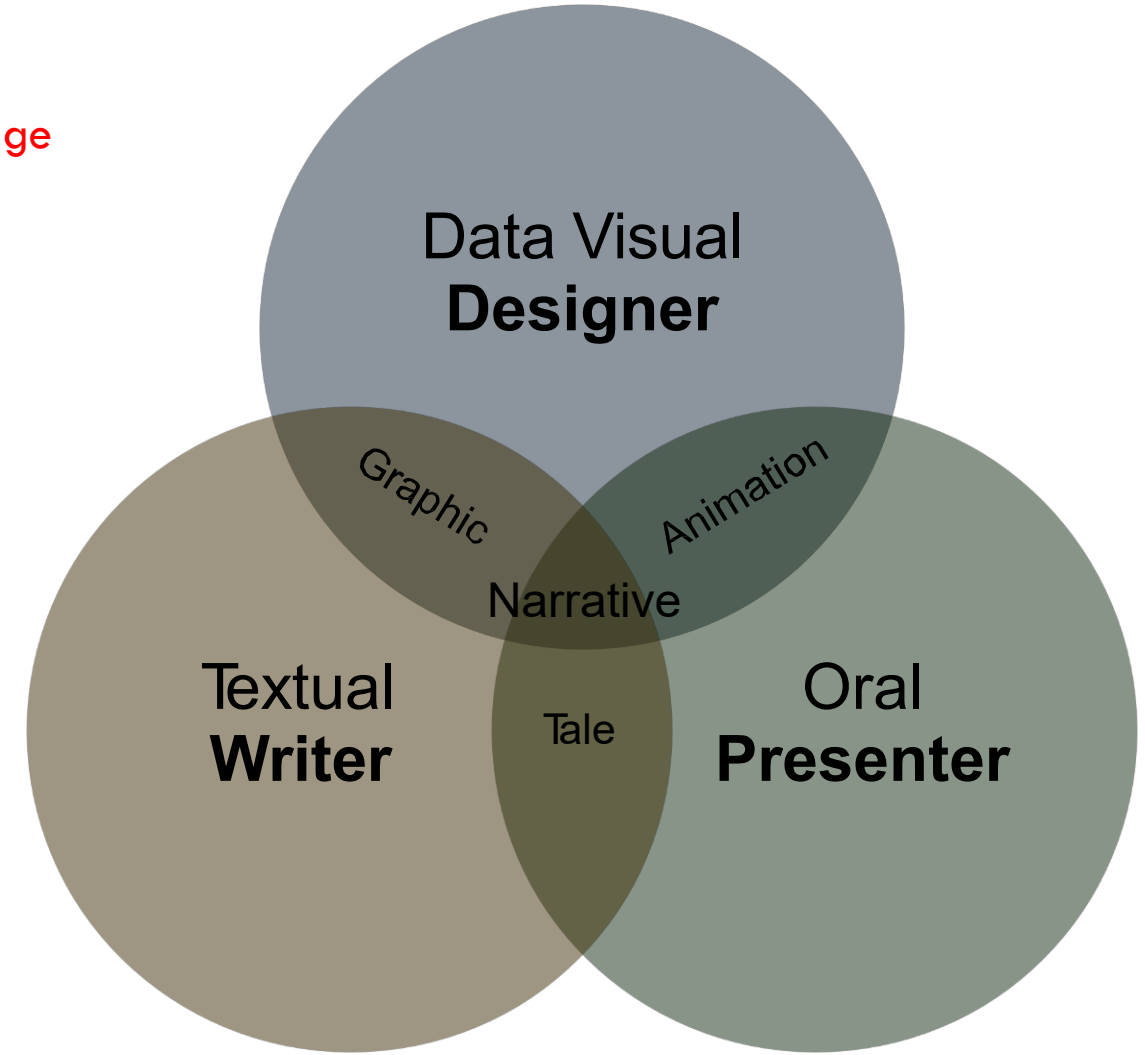
Stick to the Plot — Apply Freytag's Pyramid to Data Stories



HIRE OR TEACH DATA STORYTELLING'S COMBINED SKILLS

Designer (no knowledge of major impact)

Textual (data story is edited (analyst who like to talk and focus more on the core message to deliver)



Right presenter with the right profile

Great stories are:

Tell yours.