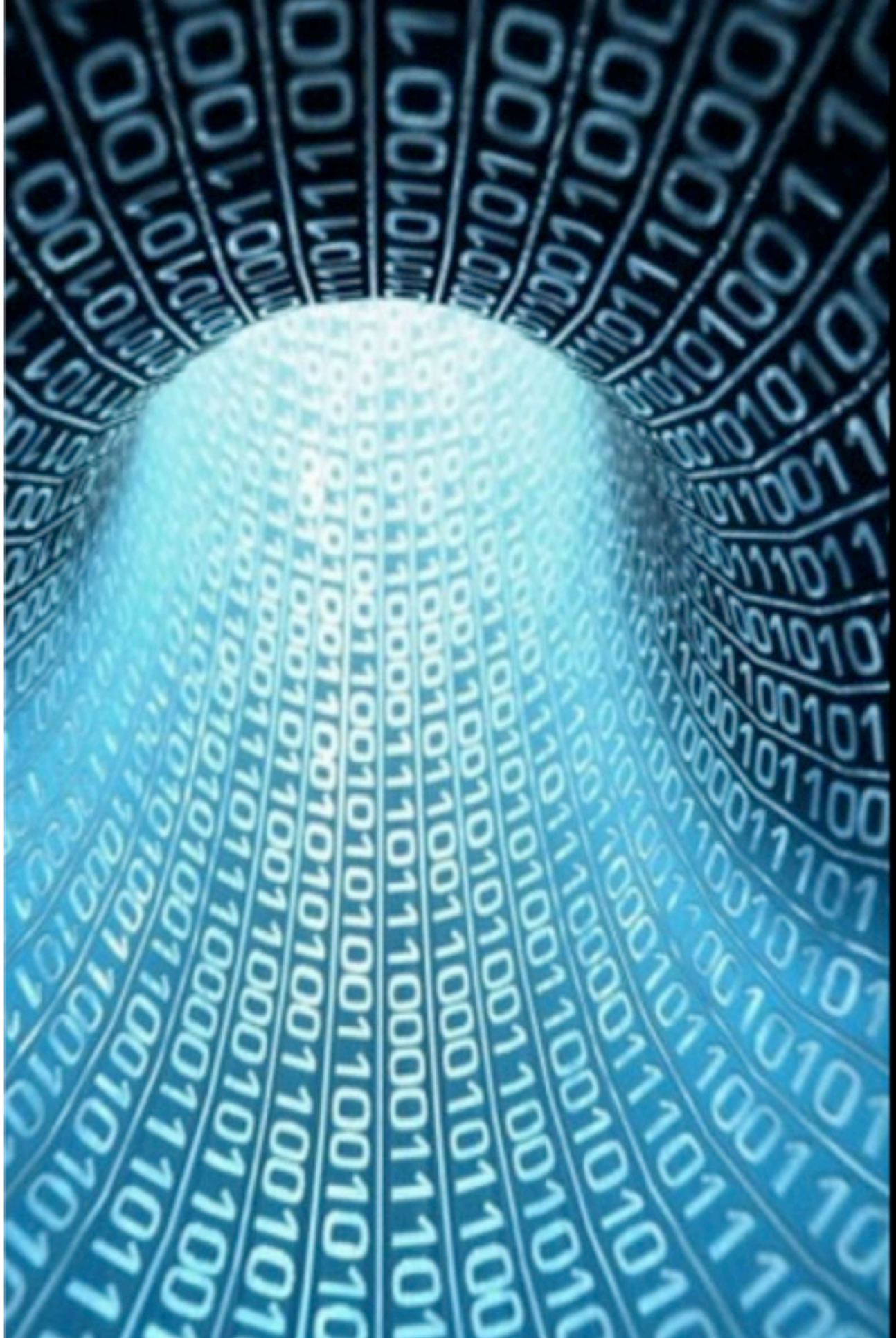
# Data Science: Theory

Ronojoy Adhikari
The Institute of Mathematical Sciences



# Axiom : your organization will benefit from data

# MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

## MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

## PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

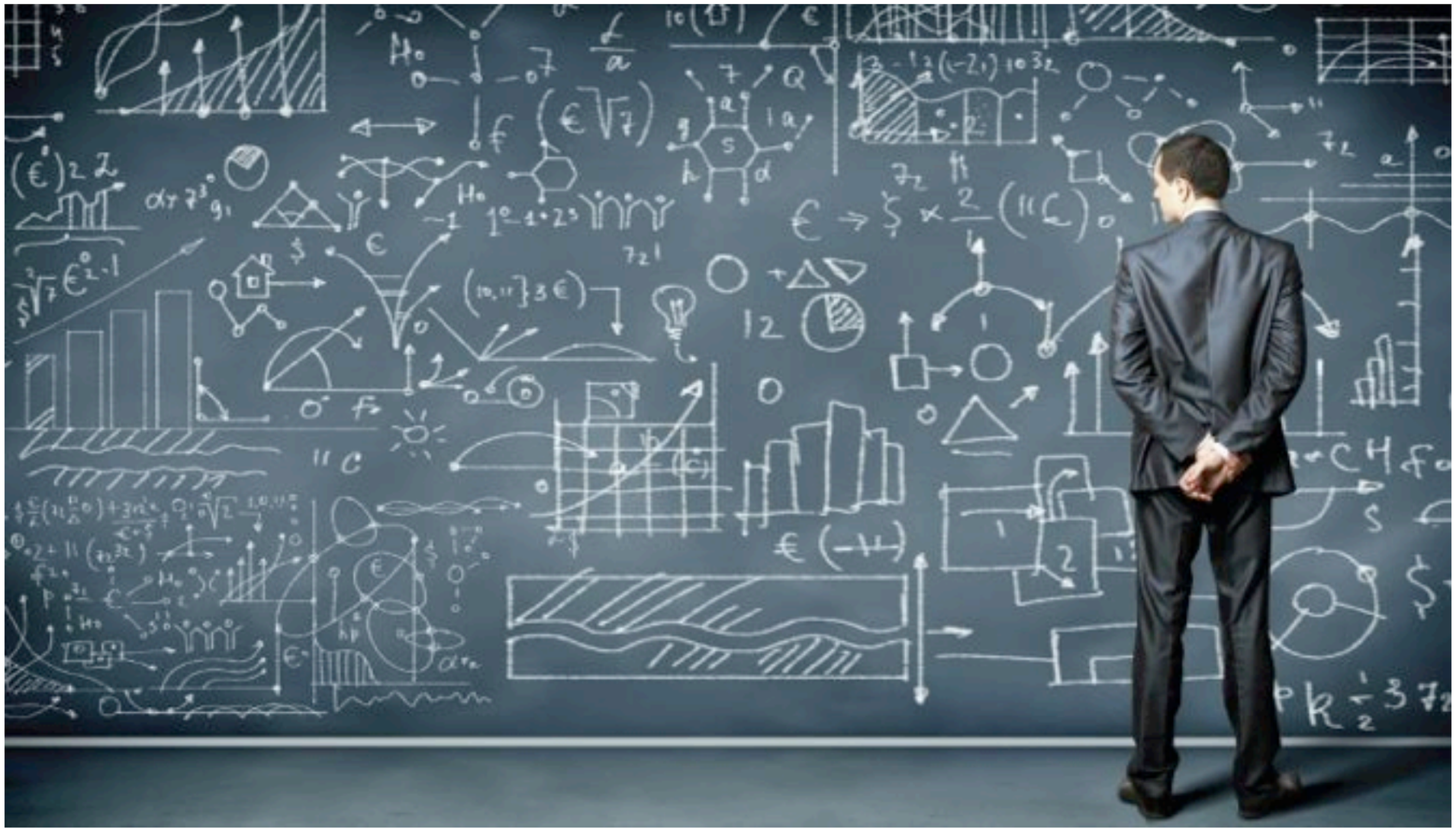
## DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

## COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

# Lots of data - where is the science ?

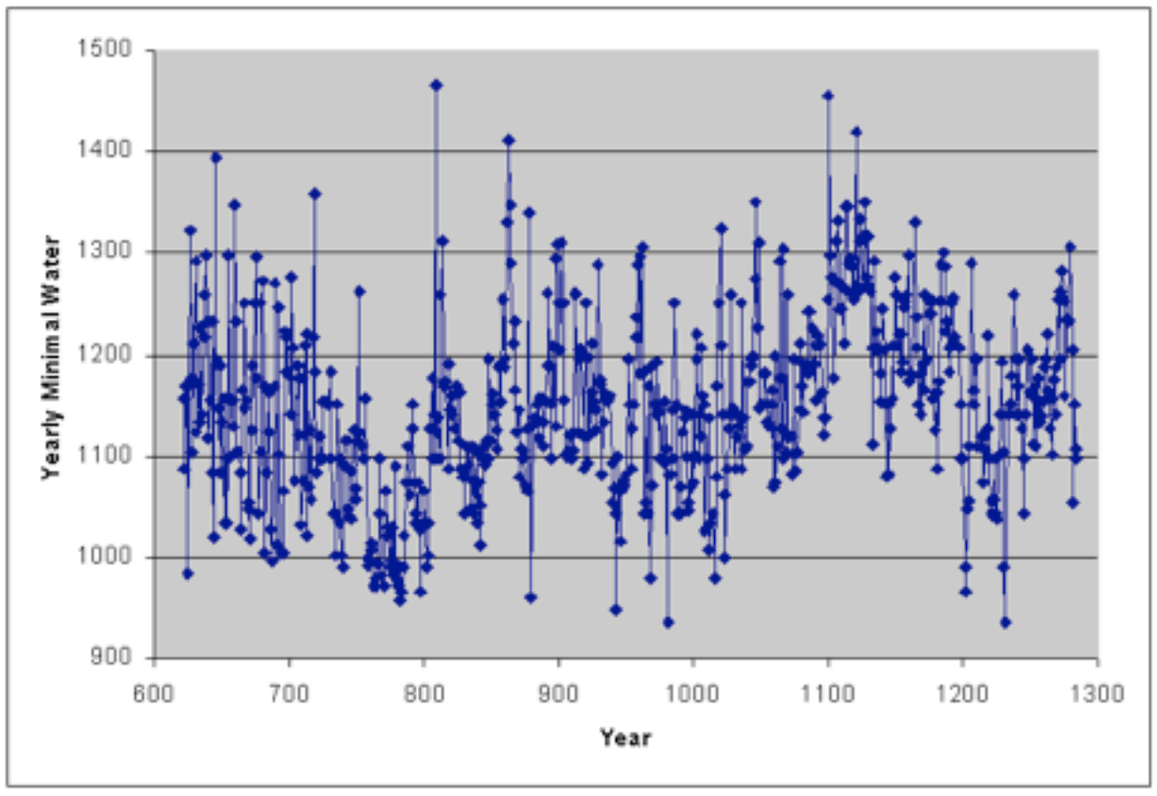


**Science : observation - hypothesis - experiment - theory**

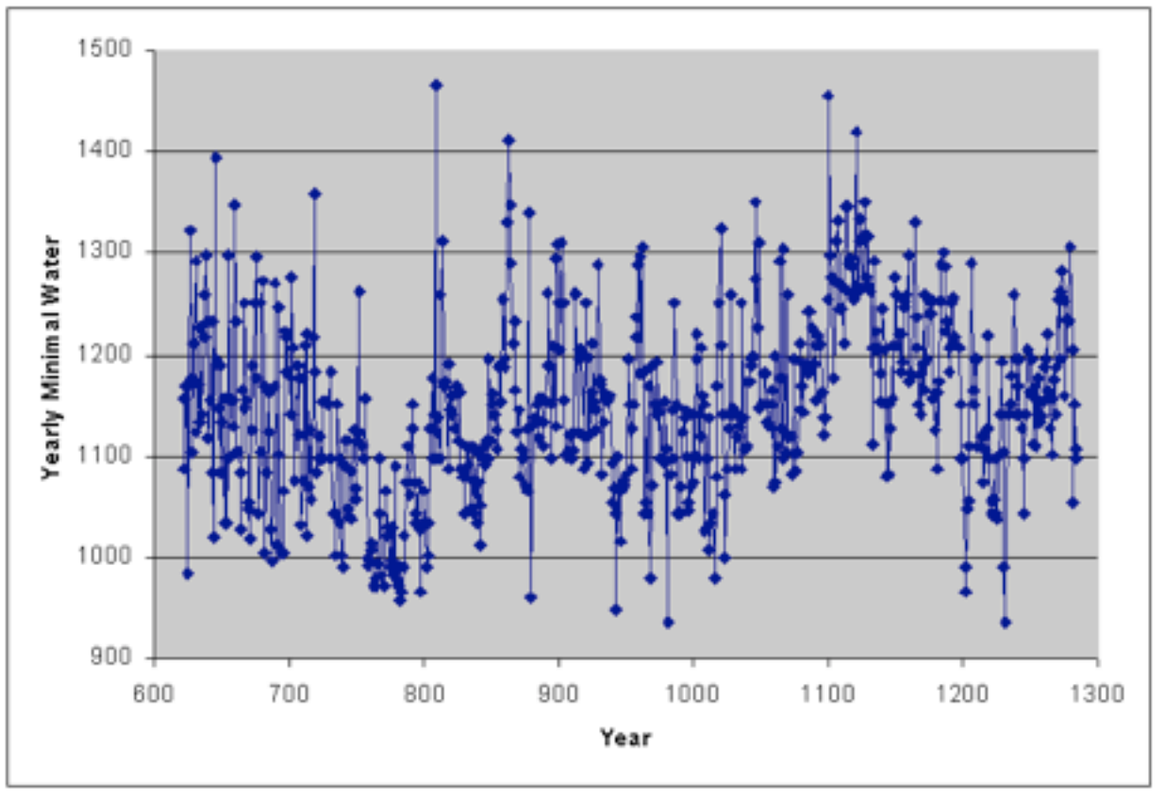
**What are we observing ?**
**What is our hypothesis ?**
**Can we experiment ?**
**Will there be a theory ?**


© Patricia Felix

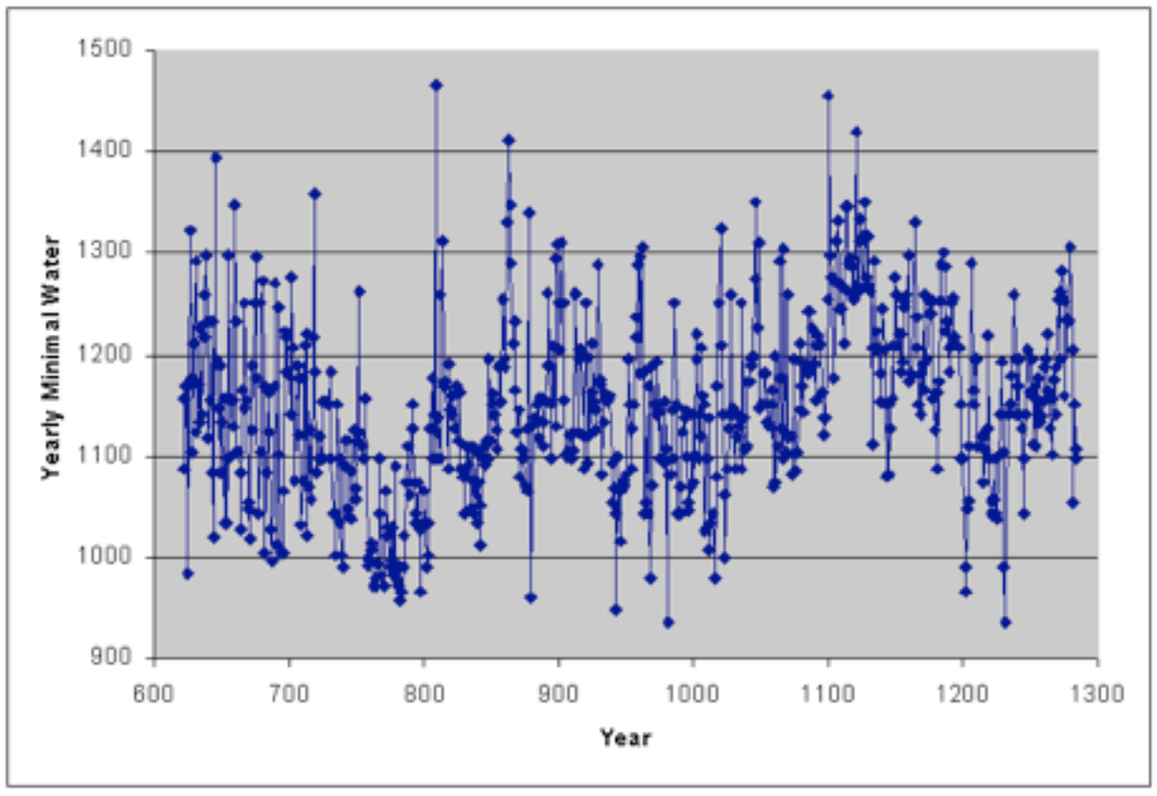

© Patricia Felix


1500
1400
1300
1200
1100
1000
900
Yearly Minimal Water
600
700
800
900
1000
1100
1200
1300
Year

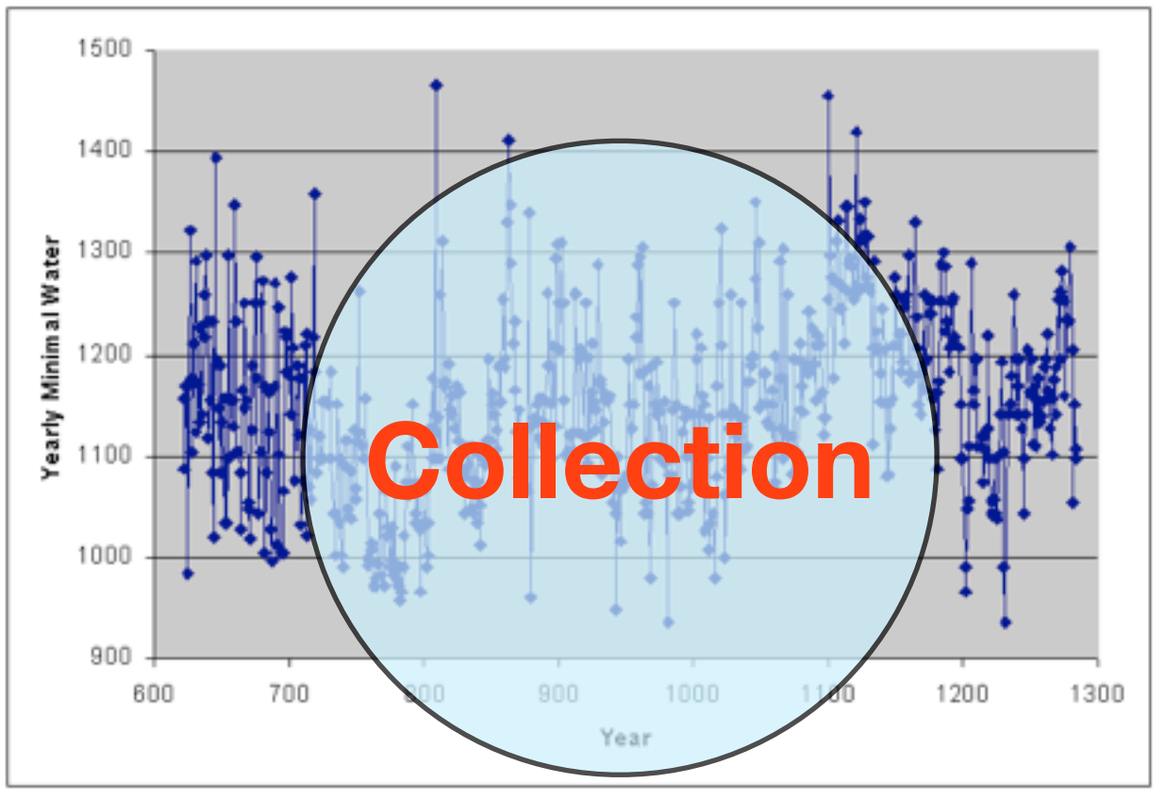

© Patricia Felix


Yearly Minimal Water
1500
1400
1300
1200
1100
1000
900
600
700
800
900
1000
1100
1200
1300
Year

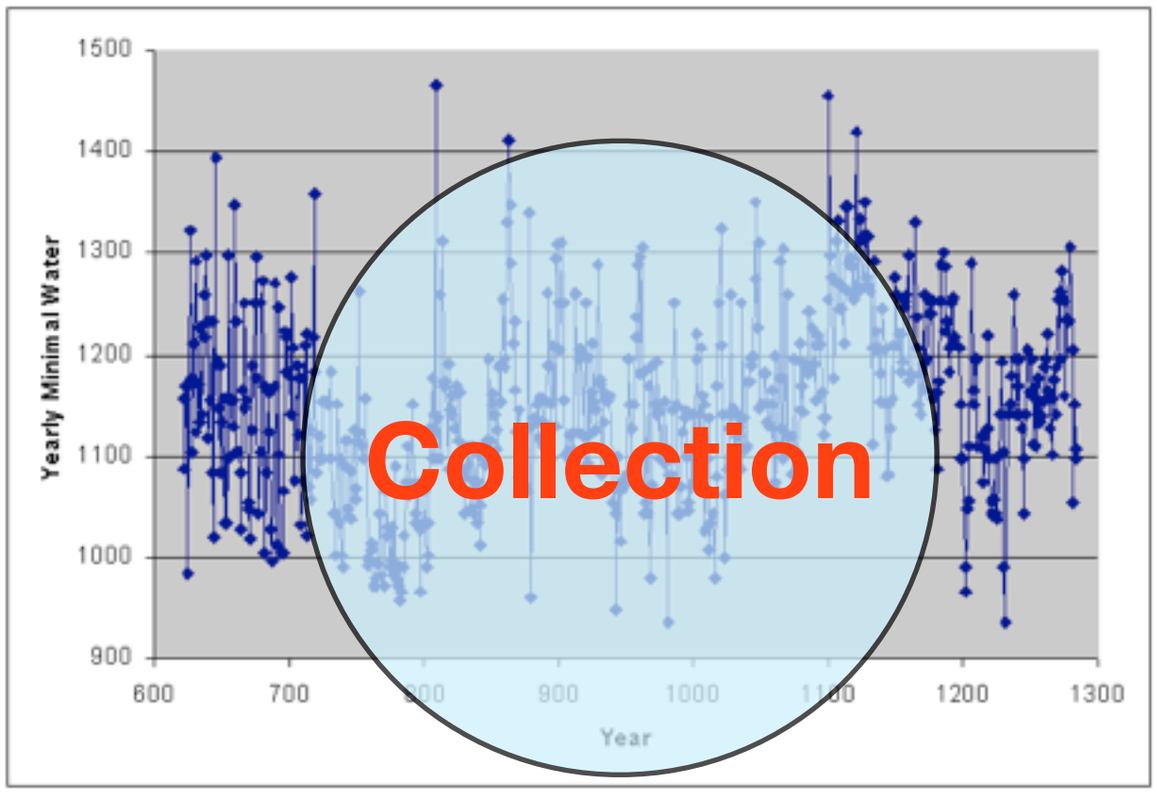

Need
© Patricia Felix

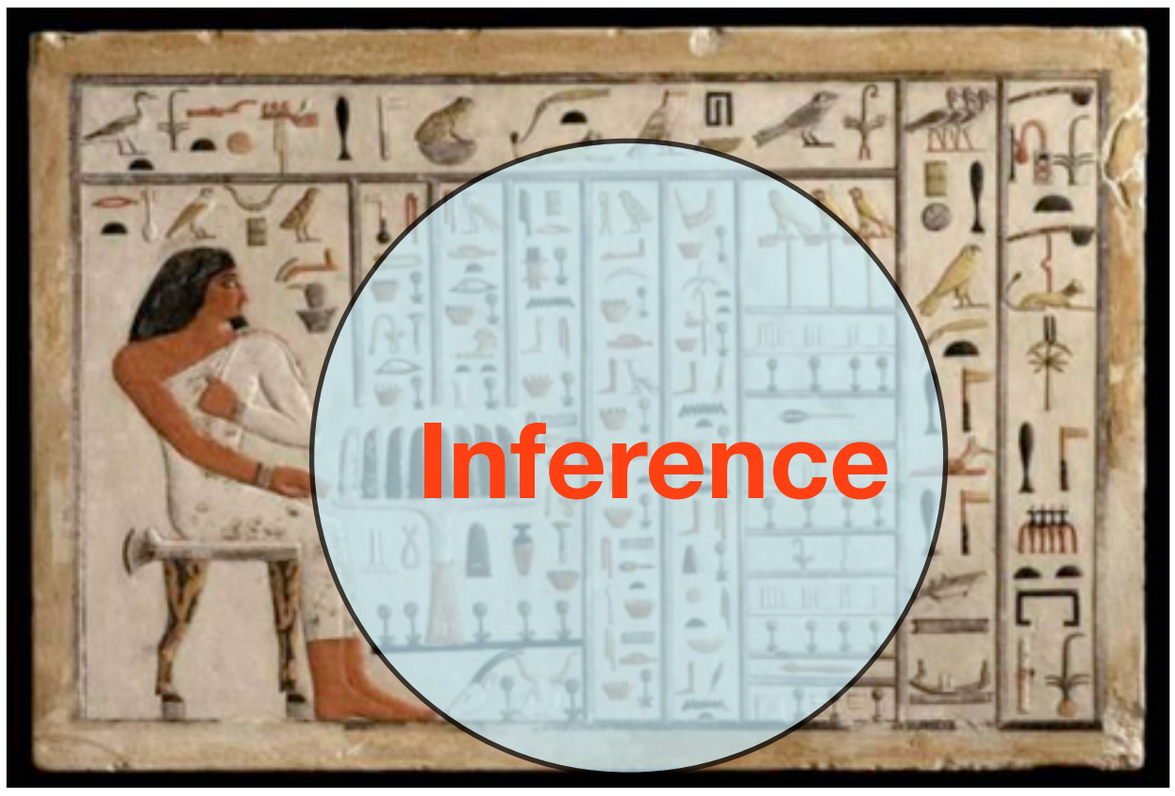

Yearly Minimal Water
1500
1400
1300
1200
1100
1000
900
600
700
800
900
1000
1100
1200
1300
Year


Need
© Patricia Felix

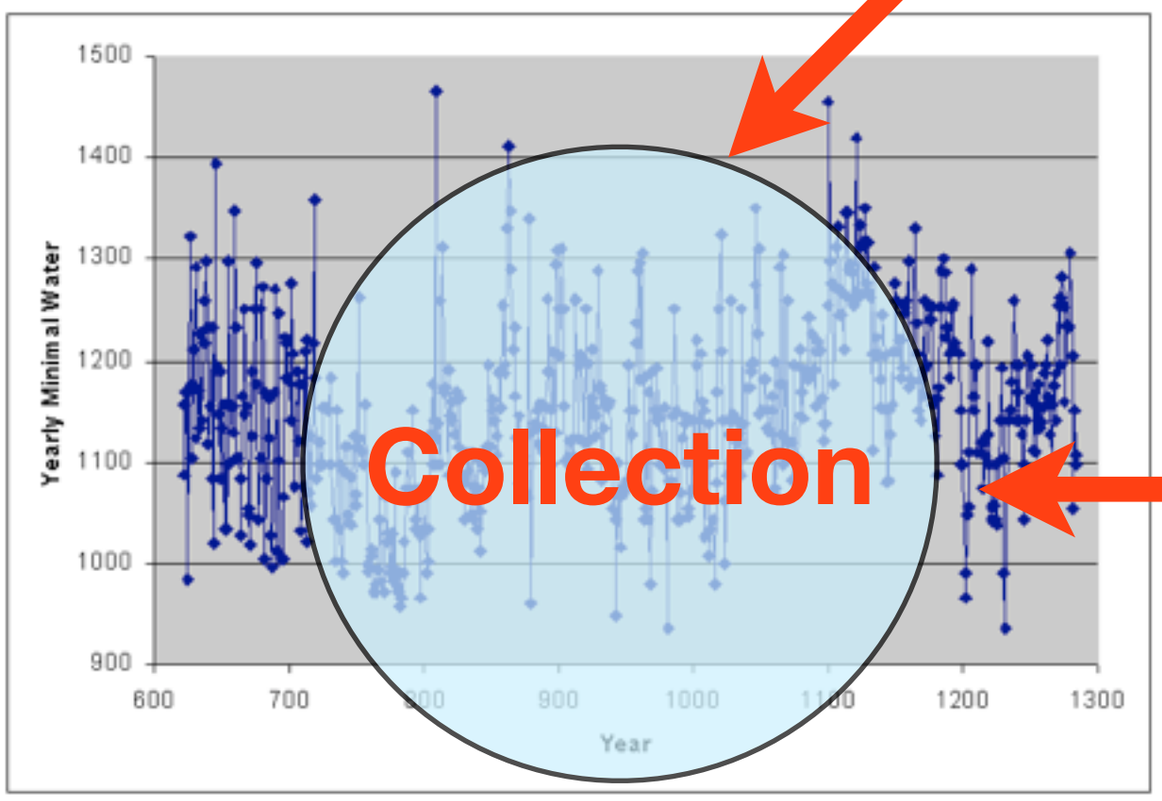

Yearly Minimal Water
1500
1400
1300
1200
1100
1000
900
600
700
900
1000
1200
1300
Year
Collection

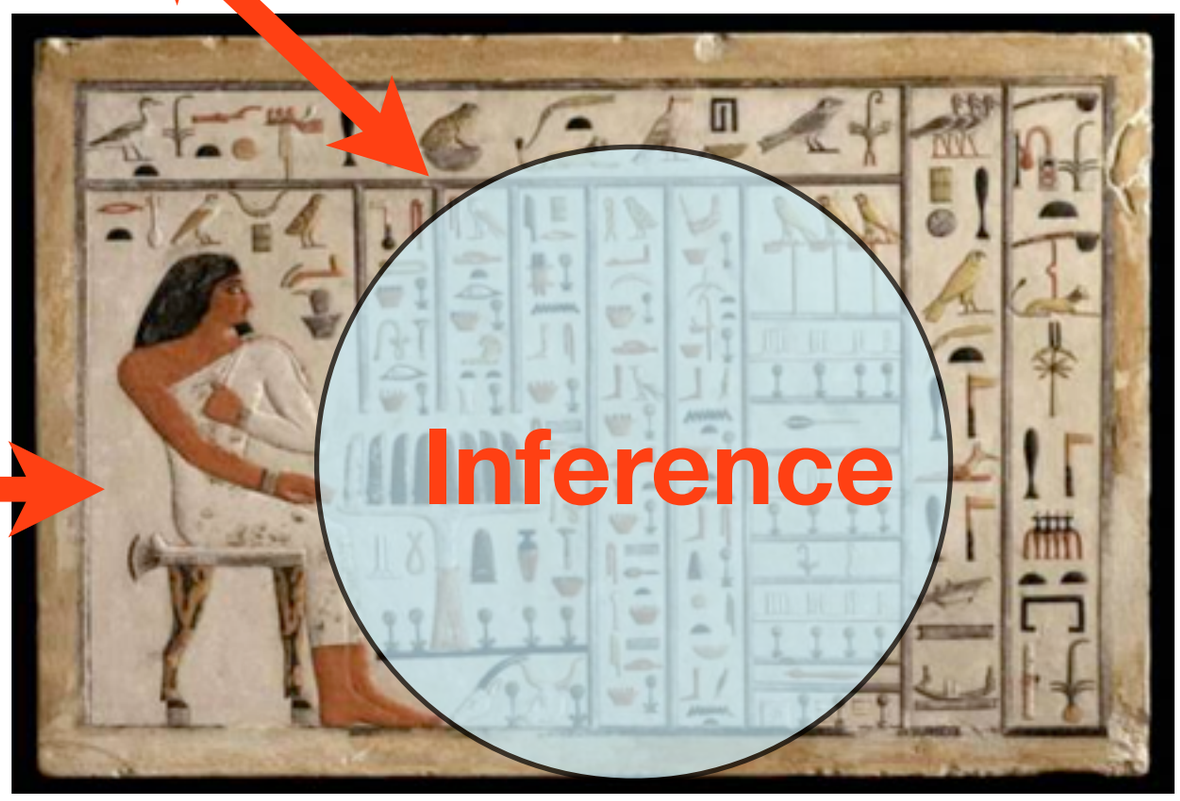

Need
© Patricia Felix

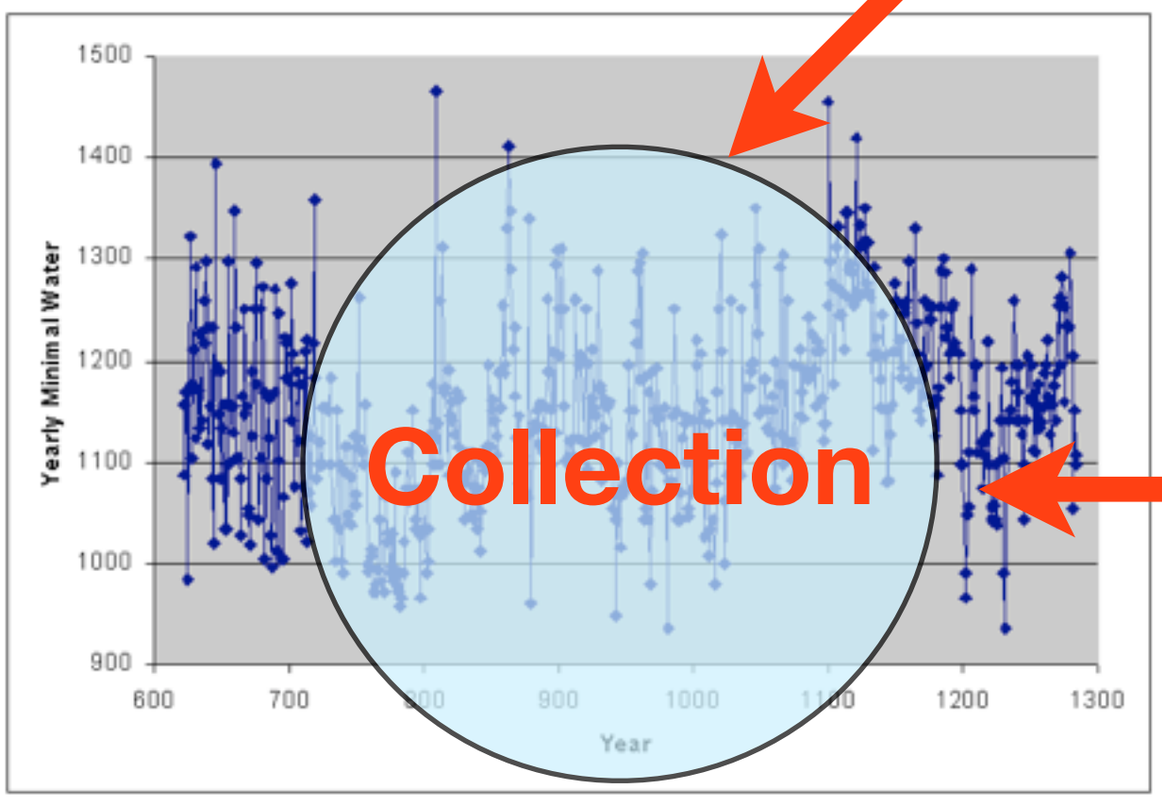

Yearly Minimal Water
1500
1400
1300
1200
1100
1000
900
600
700
800
900
1000
1100
1200
1300
Year
Collection

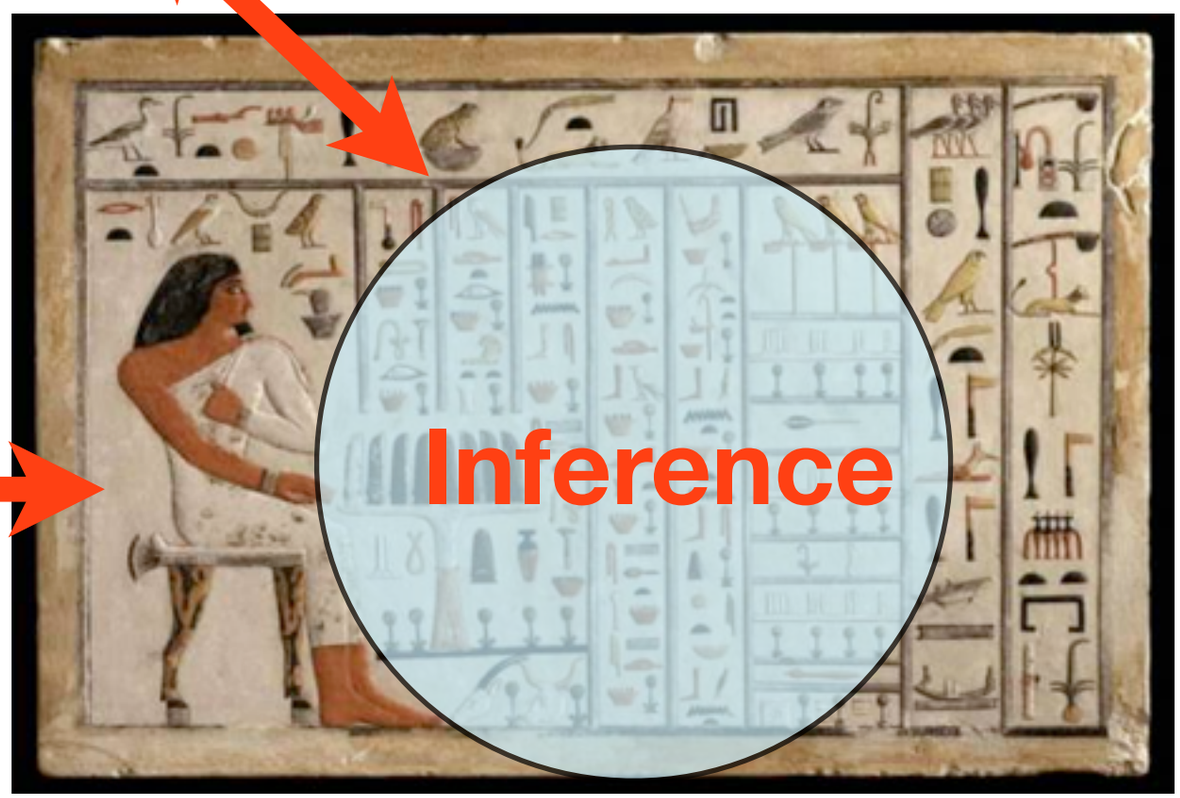

Inference


Need
© Patricia Felix


1500
1400
1300
1200
1100
1000
900
Yearly Minimal Water
600
700
800
900
1000
1100
1200
1300
Year
Collection


Inference

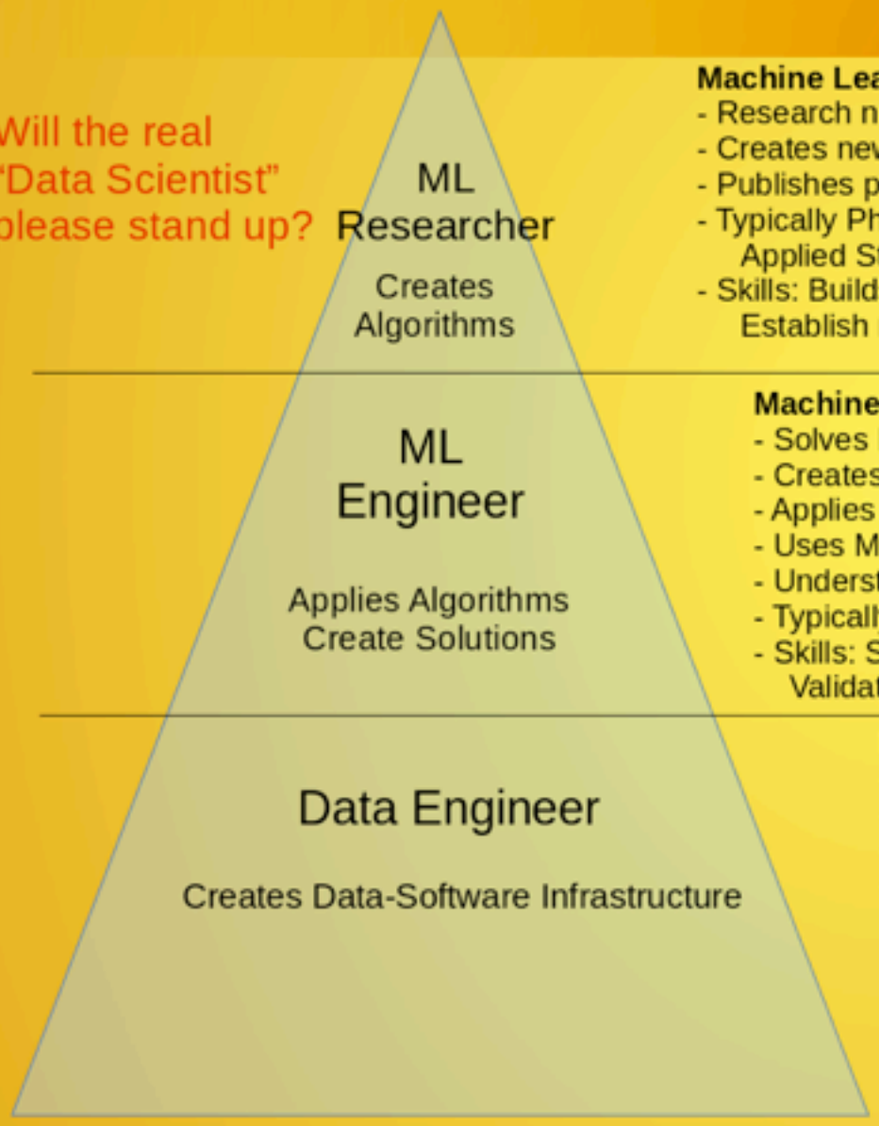




# Automated Inference ~ Machine Learning

# Machine Learning Skills Pyramid v1.0

Will the real "Data Scientist" please stand up?




**Machine Learning Researcher/Scientist**:
- Research novel machine learning problems
- Creates new mathematical models and algorithms
- Publishes papers on research results
- Typically PhD/MA Level: Robotics, Machine Learning, Cognitive Science, Applied Statistics, Engineering, Operations Research, Math, etc.
- Skills: Builds mathematical models, Breaks ground in research, Establish new paradigms, Scientific Formalism, Experiment design

**Machine Learning Engineer**:
- Solves business/data learning problems
- Creates ML solutions to achieve an organization's objective
- Applies established algorithms
- Uses ML algorithm libraries
- Understands strengths and weaknesses of different algorithms
- Typically BS/MA Level: Computer Science, Math, Other Technical
- Skills: Software Eng. PLUS Data Analysis, ML Algorithm Selection, Cross Validation, Metrics/Scoring, Feature Engineering

**Data Engineer**:
- Develops code in support of Machine Learning Solutions
- Data extraction, transformation, scraping, joining, cleaning
- Summary Statistics, counting, sampling on request
- Skills: Platform/DB/Language specific expertise, Performance, Parallel and Distributed Computing, Quality, Reliability, Map/Reduce-Hadoop, VMs/Cloud, SQL/noSQL, Production Scaling etc.

## Uncertainty

- how much will the Nile flood ?
- when will this equipment fail ?
- is this email spam ?
- is this applicant a good hire ?

## Decisions

- should we invest in dams ?
- should we build redundancy ?
- should i delete without reading?
- should we look more ?

# We need to make reasoned decisions in the face of uncertainty

Reasoning : Logic - Boolean algebra
Uncertainty : Chance, probability
Combine : Bayesian probability

# Six steps to value from data

1. **Identify the problem**
2. **Find relevant data sources**
3. **Preprocess the data**
4. **Apply the algorithm (ML)**
5. **Visualize the process**
6. **Tell your story and maintain**