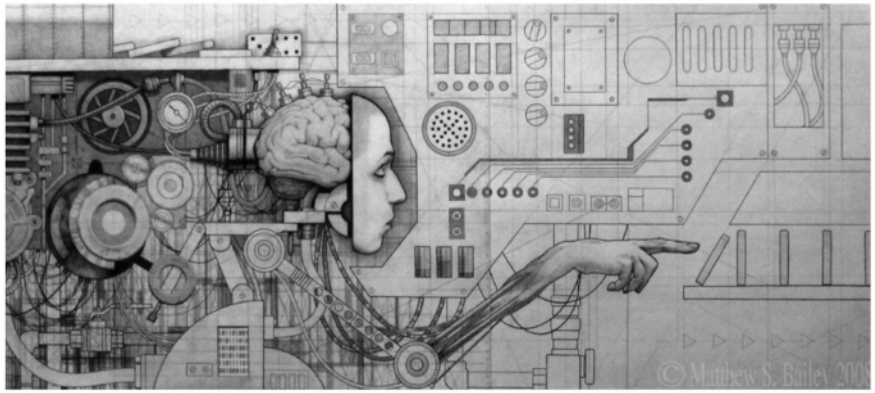
Introduction to Data Science

A Straightforward, Short, and Non - Academic Approach

By John Thomas Foxworthy

Data Scientist



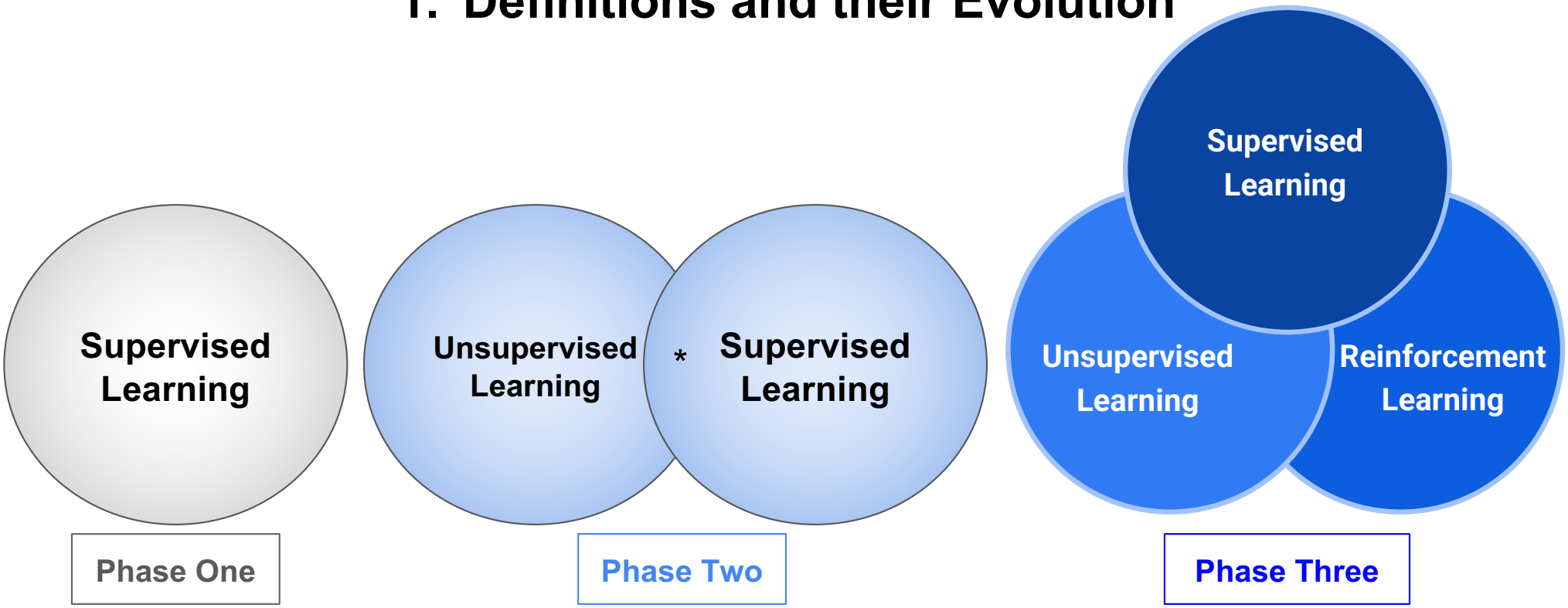
1. Definitions
2. Relabeling
3. Motto
4. Prior Job Titles
5. Origin
6. Models and Methodologies
7. Most Common Mistakes
8. Data Science Workflow
9. Learning Resources
10. The Future

1. Definitions

What's with all the Learning?

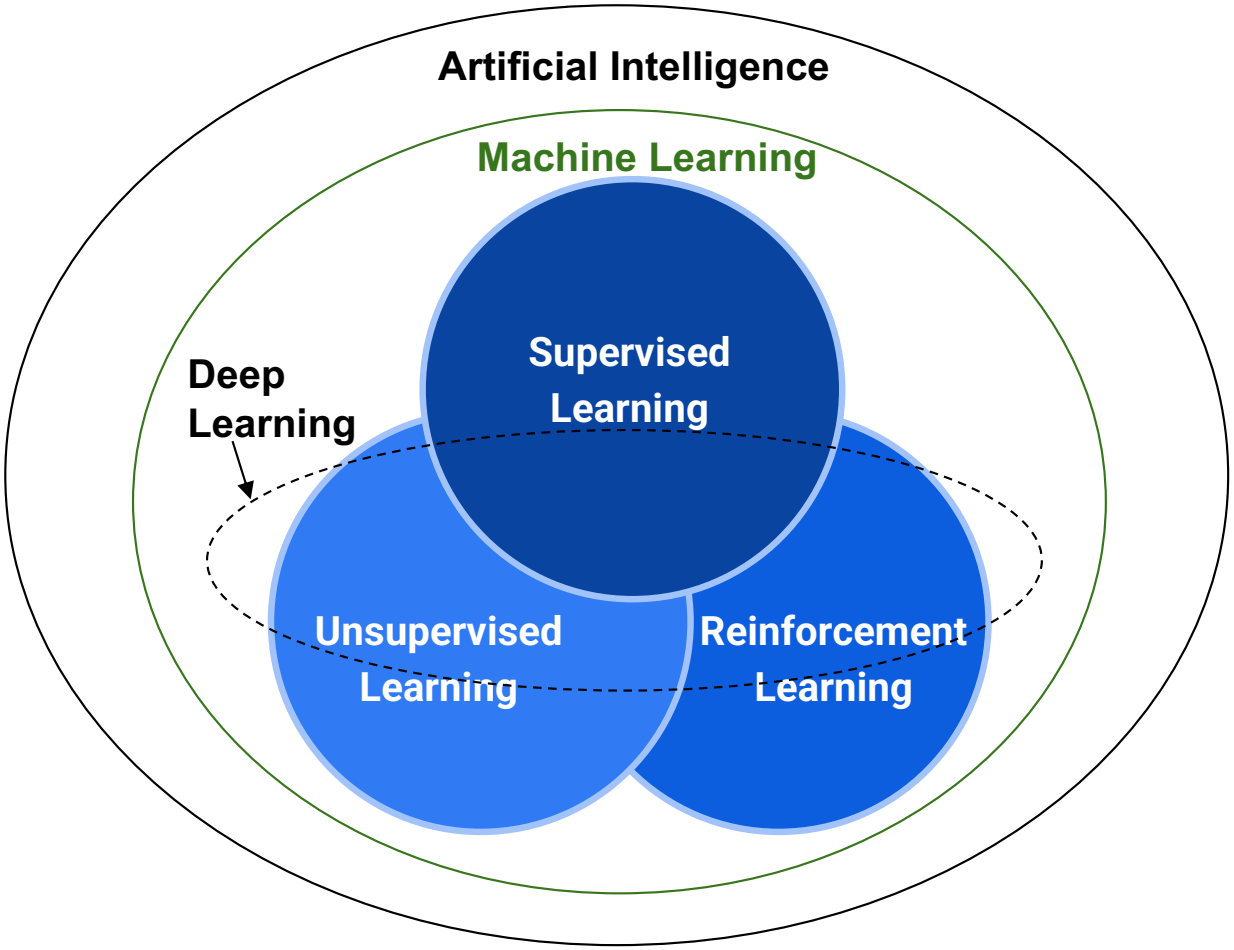
Data Science	The re - packaging of Statistics with more computing resources and techniques
Machine Learning (ML)	A software program that makes decisions without explicit programing
Deep Learning (DL)	Machine Learning with depth or various layers such as a Neural Network
Supervised Learning	Predictive Analytics
Semi - Supervised Learning	Both Predictive and Descriptive Analytics
Unsurprised Learning	Descriptive Analytics
Reinforcement Learning (DL)	A software agent that observes, then acts to receive rewards
Artificial Intelligence (AI)	Anything that is not biological that behaves biological

1. Definitions and their Evolution



* Semi-Supervised Learning

1. The Current State of Definitions of Data Science



1. Definition Breakdown with (Un)Supervised Learning

Subject Area	Unsupervised Learning	Supervised Learning
Business	Inputs	Inputs & Outputs
Engineering	Drivers	Drivers & Outcomes
Mathematics	Regressors	Regressors & Regressands
Statistics	Independent Variables	Independent Variables & Dependent Variables
Psychometrics	Predictors	Predictors & Responses
General Science	Explanatory	Explanatories & Focuses
Linguistics	Descriptive	Descriptives & Predictive
Machine Learning	Unlabeled Training Data	Unlabeled Training Data & Labeled Training Data

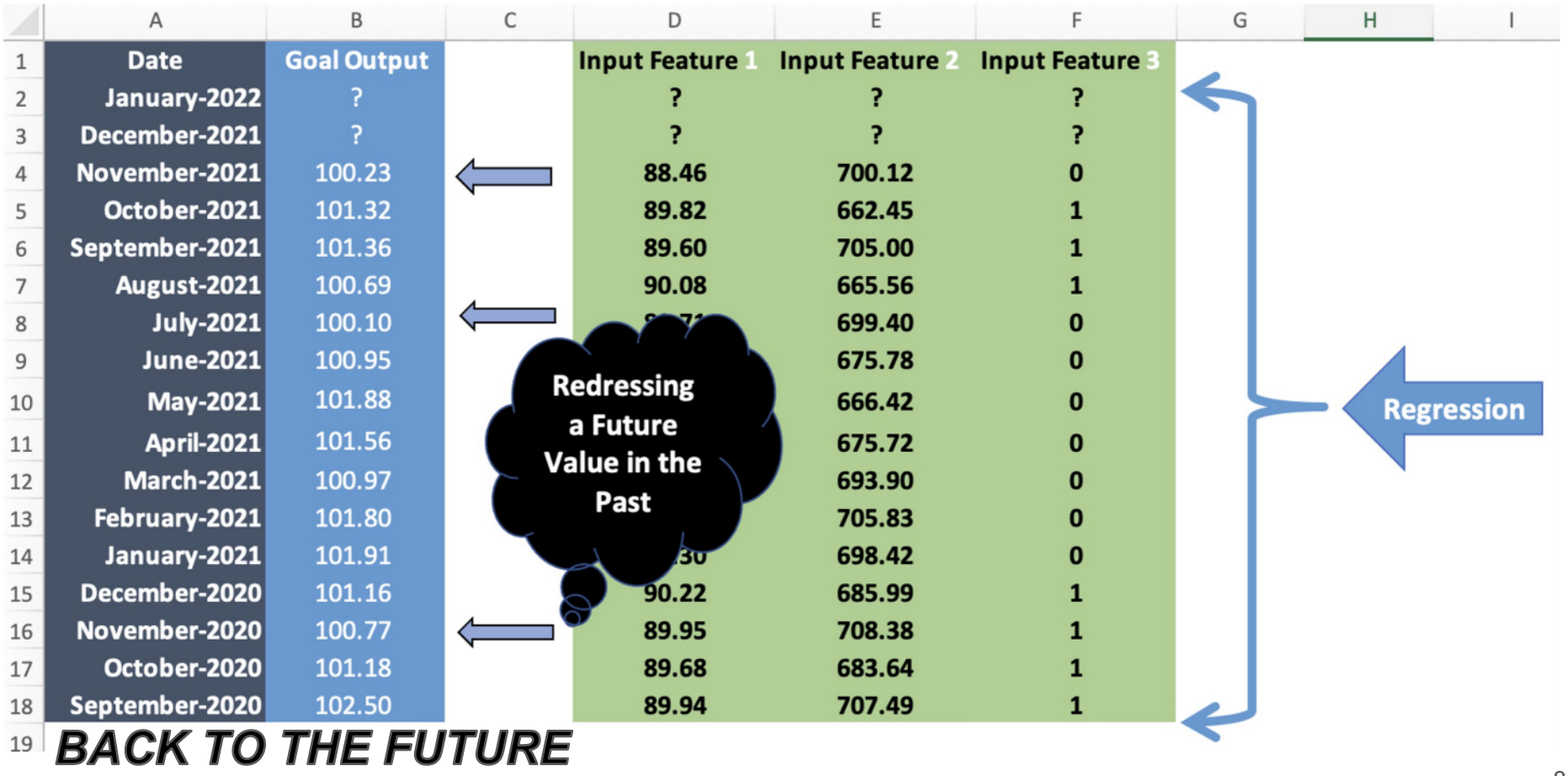
2. Relabeling Defines Data Science and its Purpose

Statistical Learning		Machine Learning
Fitting Equations	⇨	Learning Process
Model	⇨	Model (ML) or Network (DL) or No Model (RL)
Regression or Classification	⇨	Supervised Learning (Predictive Analytics)
Density Estimation or Clustering	⇨	Unsupervised Learning (Descriptive Analytics)

THEORY

APPLIED

2. What is a Regression Model? A Clarification



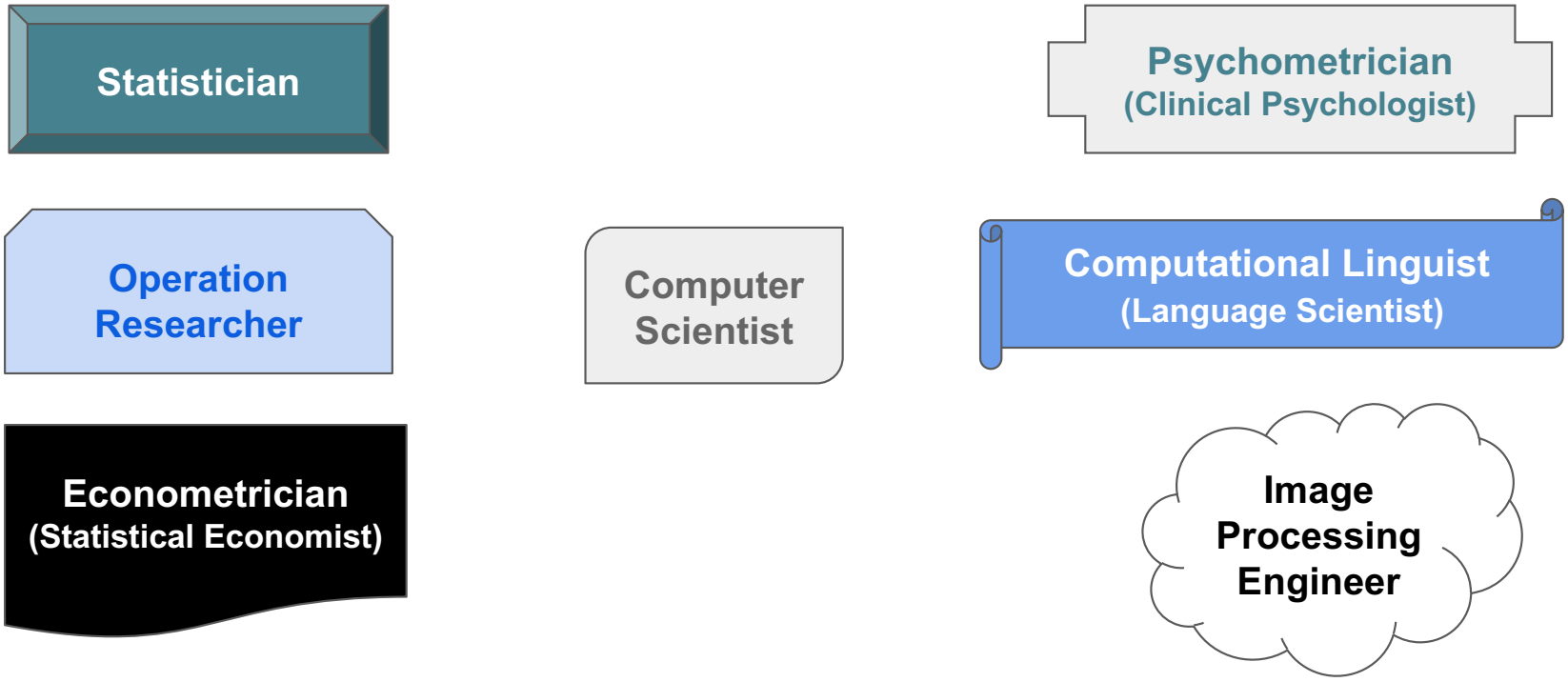
	A	B	C	D	E	F	G	H	I
1	Date	Goal Output		Input Feature 1	Input Feature 2	Input Feature 3			
2	January-2022	?		?	?	?			
3	December-2021	?		?	?	?			
4	November-2021	100.23		88.46	700.12	0			
5	October-2021	101.32		89.82	662.45	1			
6	September-2021	101.36		89.60	705.00	1			
7	August-2021	100.69		90.08	665.56	1			
8	July-2021	100.10			699.40	0			
9	June-2021	100.95			675.78	0			
10	May-2021	101.88			666.42	0			
11	April-2021	101.56			675.72	0			
12	March-2021	100.97			693.90	0			
13	February-2021	101.80			705.83	0			
14	January-2021	101.91			698.42	0			
15	December-2020	101.16		90.22	685.99	1			
16	November-2020	100.77		89.95	708.38	1			
17	October-2020	101.18		89.68	683.64	1			
18	September-2020	102.50		89.94	707.49	1			
19									

3. Data Science Motto

All models are wrong,
but some are useful.

George Box (1976)

4. Prior Job Titles of Today's Data Scientist



5. Where did Data Science come from?

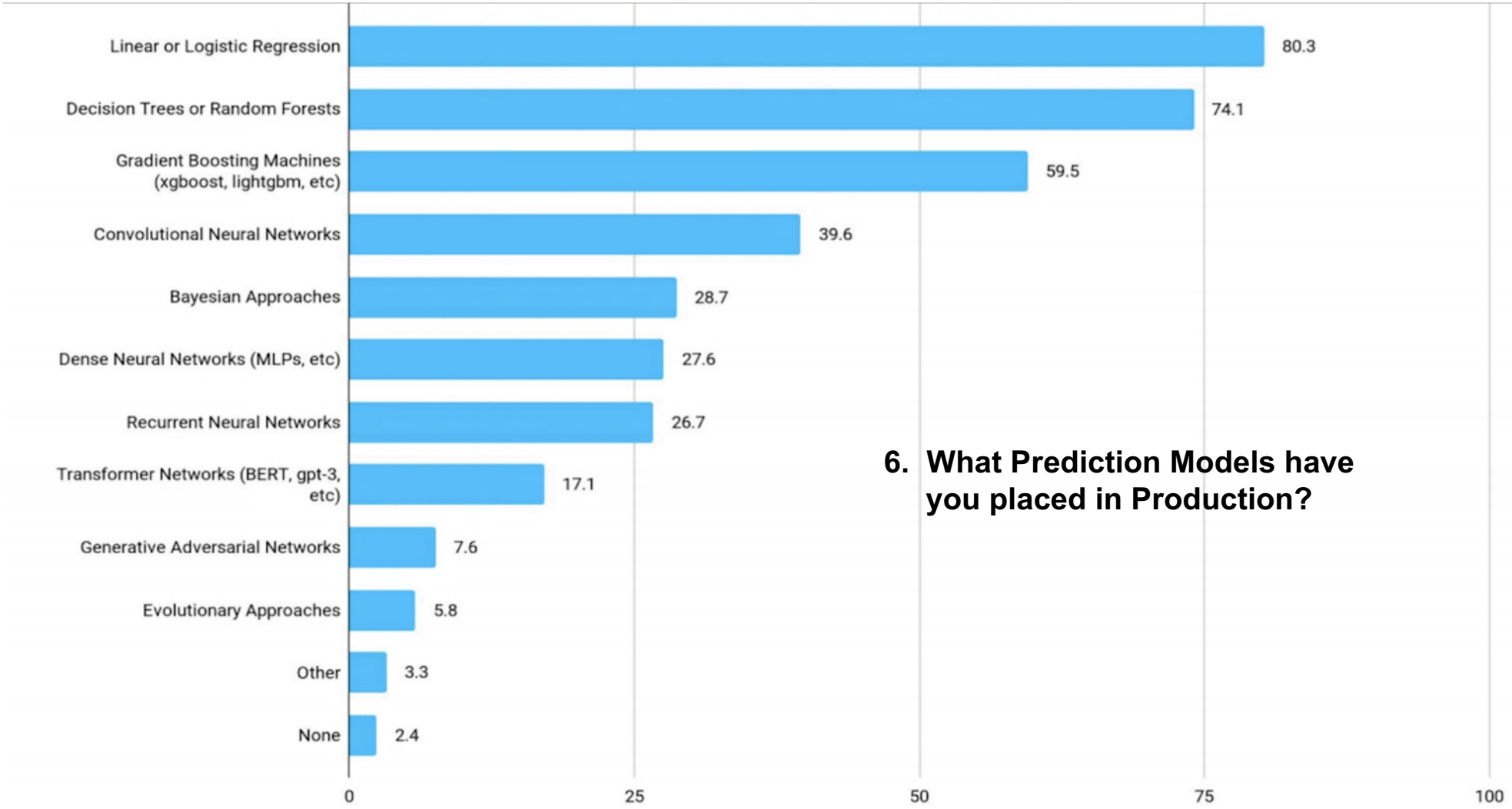
William S. Cleveland

- **Professor of Statistics at Purdue University, Indiana**
- **Vote at the Statistical Symposium with various University Professors in 2001**
  - **Processing power of computers increasing exponentially**
  - **Exponential growth of the quantity and quality of data, esp unstructured**

Question

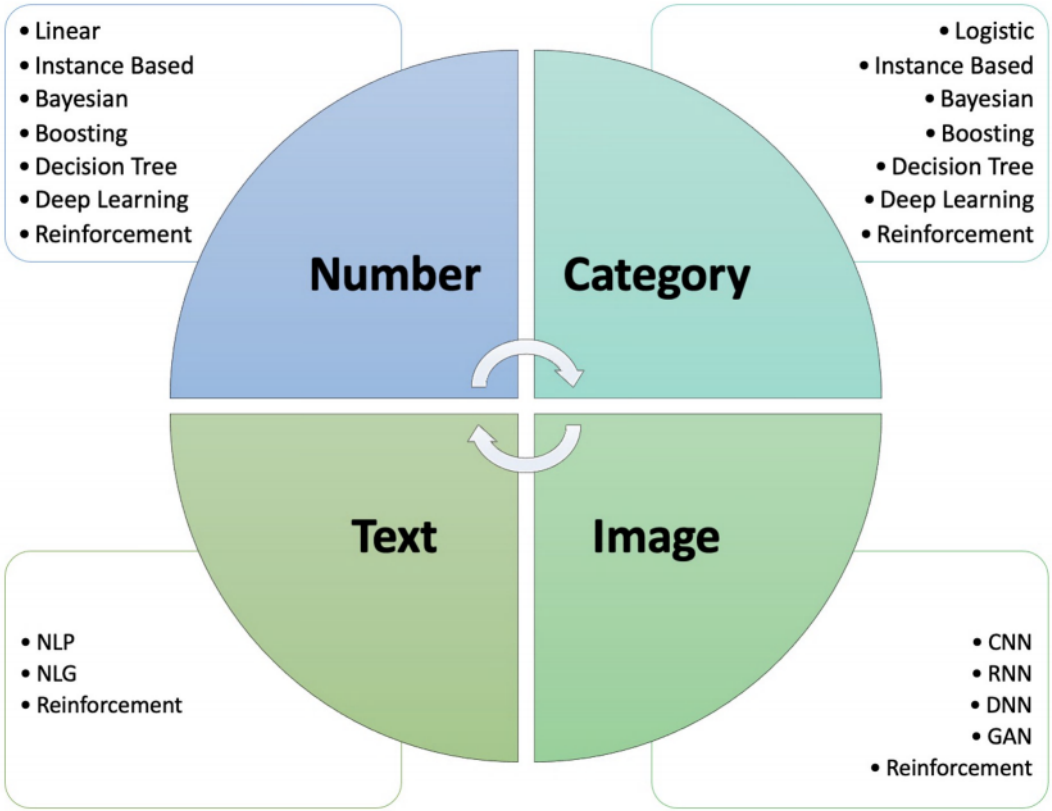
If you clone a human being, then

does the clone have artificial intelligence?

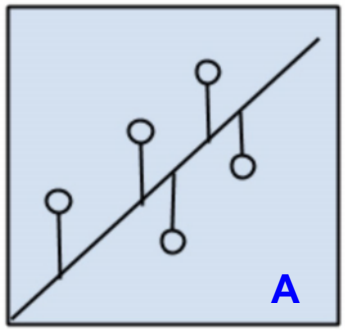


6. What Prediction Models have you placed in Production?

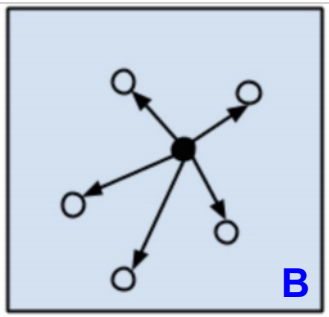
6. Supervised Learning and Predictive Analytics



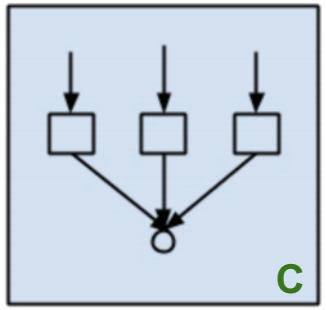
6. The Three Major Methodologies for Prediction



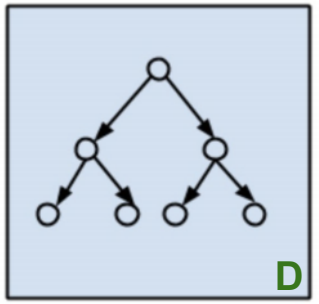
Regression



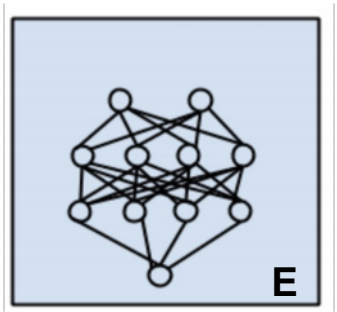
Instance - Based



Ensemble



Decision Tree



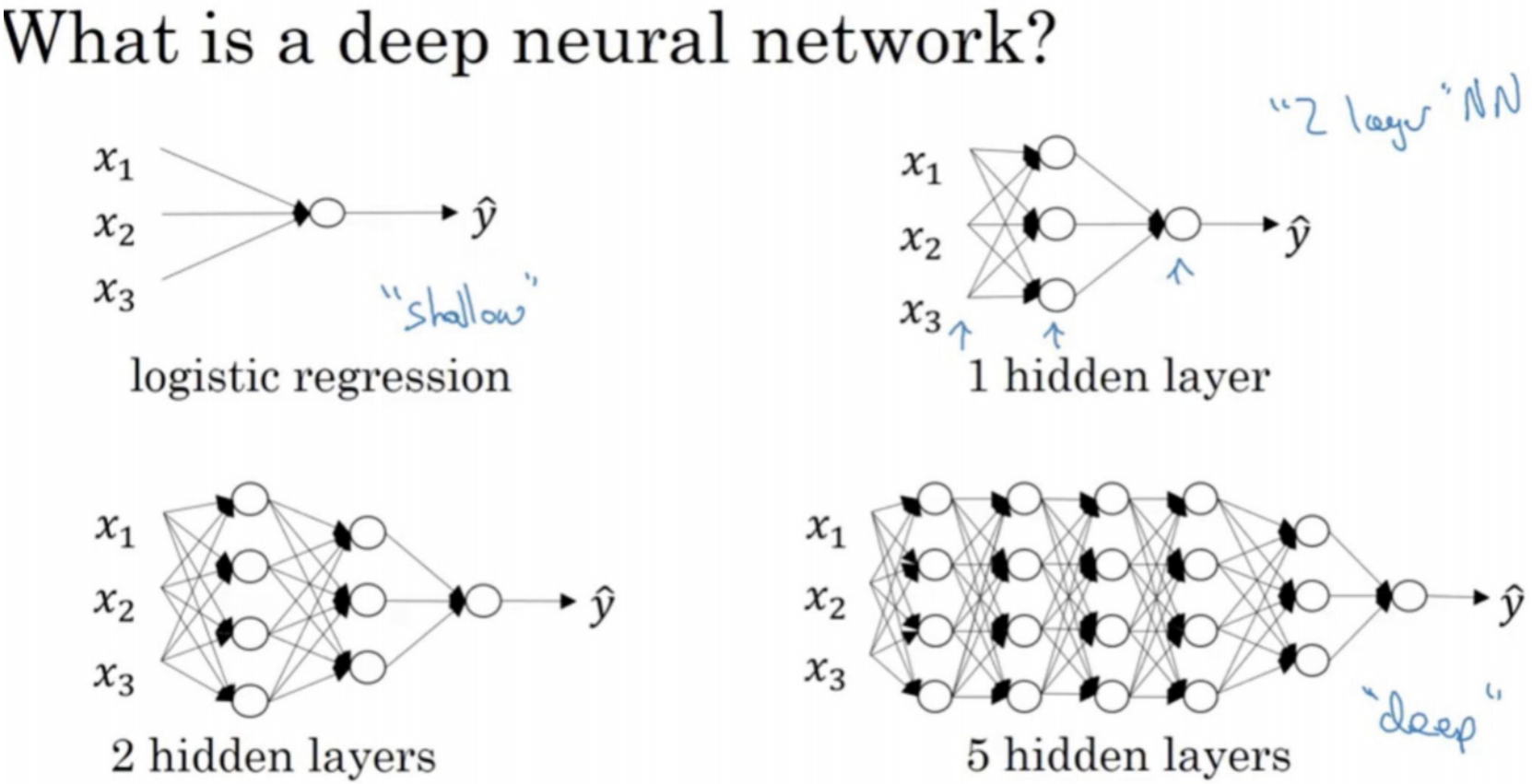
Deep Learning

Static and Sequential
(1) Data
(2) Model
(3) Output

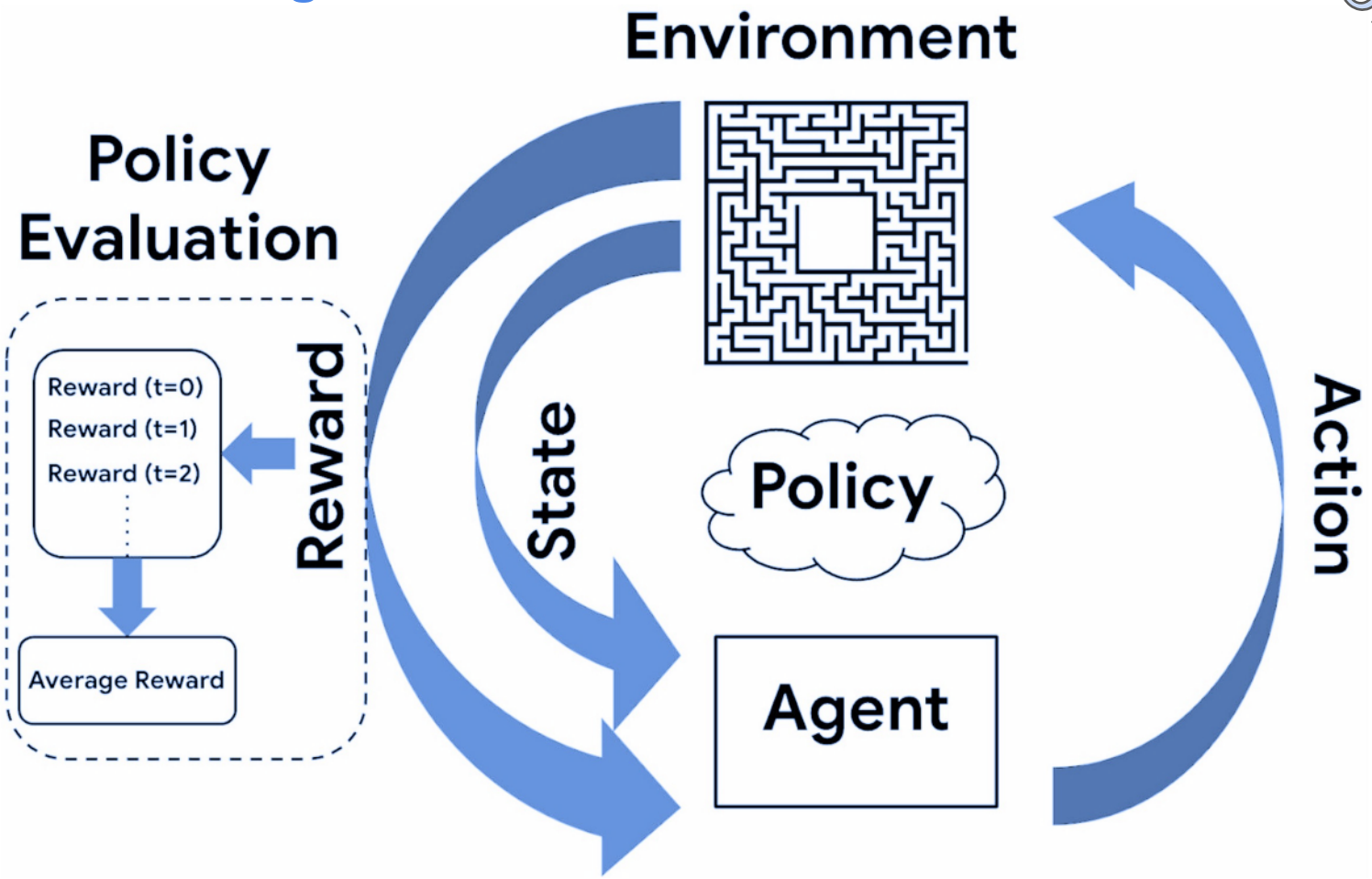
Iterative or Flow Chart
(1) Data
(2) Model
(3) Output, back to (2)

Neural Network as a Set of Models

6. Deep Learning Method for Artificial Intelligence



6. Reinforcement Learning Method for Artificial Intelligence



Grok

Grokking

6. Unsupervised Learning and Descriptive Analytics

Clustering

Data Reduction Techniques

Anomaly Detection

Feature Selection Automation

Autoencoder

Group Segmentation

Should you ask your Supervisor about Unsupervised Learning? Yes!

7. Data Science Prediction Mistakes

Always Inaccurate
Reselect Features
Change Model

Confusing Cause and Effect
Endogeneity

The Old Way does not work anymore
Regime Switching

Too Many Similar Features
Multicollinearity or try DL

Same Model, Different Results
High Variance

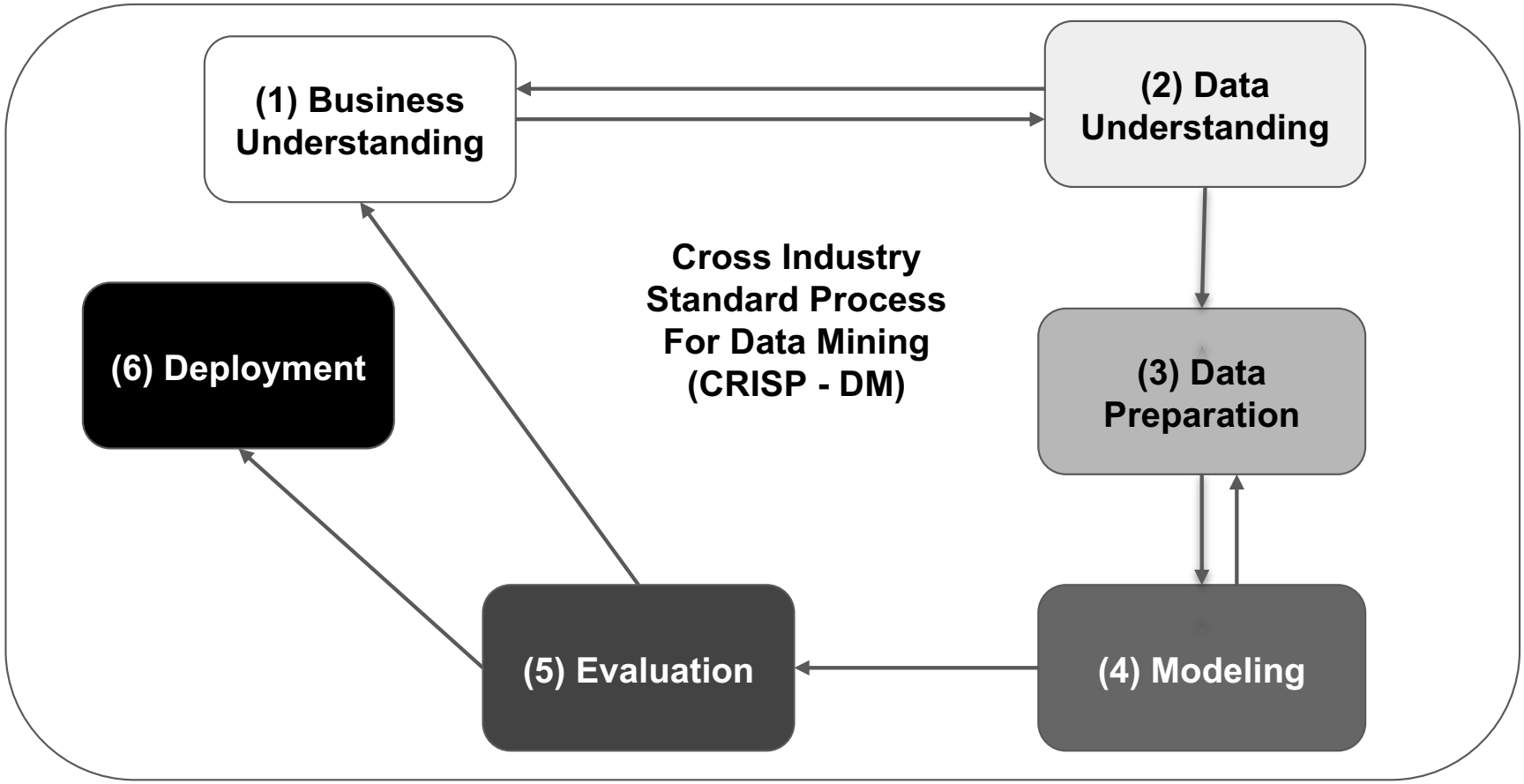
Overgeneralizing
Overfitting

Missing Data causes No Prediction
Wrong Model

More Data, Less Usefulness
High Bias

Misclassification
Type I, False Positive

8. Generalized Data Science Workflow



9. Data Science Learning Resources

- **Most Popular Course in the World for Data Science**
  - **Professor Andrew Ng, Stanford University**
  - https://www.coursera.org/learn/machine-learning

- **Effective Data Analytics**
  - **Cole Nussbaumer Knaflic, Former People Analytics Team Manager at Google**
  - **https://www.storytellingwithdata.com**

10. The Future

- **Boosting will overtake Linear and Logistic Regression**
- **Reinforcement Learning Software Agents will overtake ML**
- **Neural Networks will expand as data grows exponentially**
- **Just about everything in our lives will be Artificial Intelligence**

ANY QUESTIONS?