Evolution of Terms:
Data Science, A.I., Analytics, Statistics, Operations Research, Machine Learning, etc.

NISS webinar on
"**What's in a Name** –
Data Analytics, Machine Learning, Artificial Intelligence and What Else?"

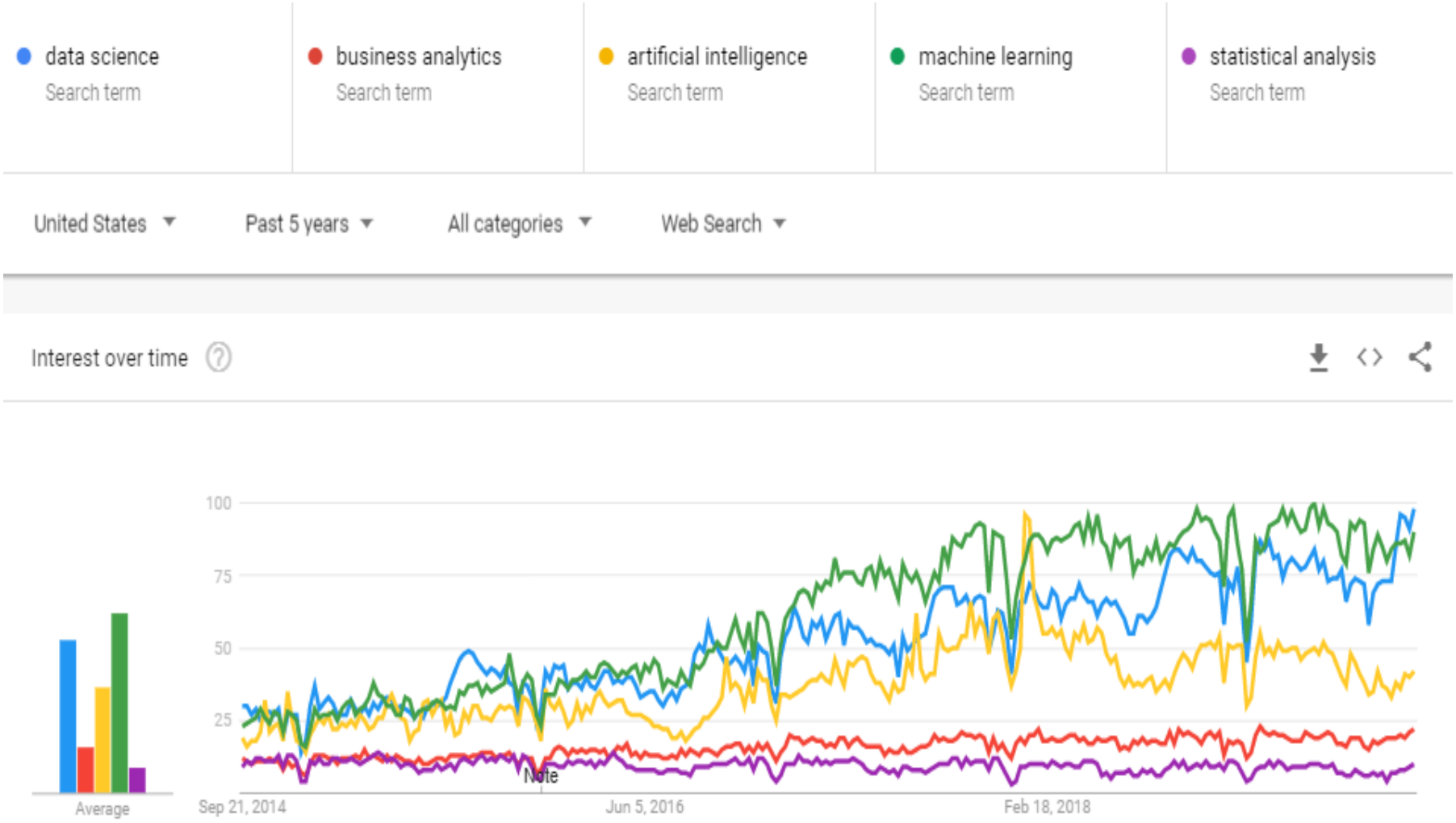
Victor S.Y. Lo

September 2019

Content

- **Data Scientist and Data Science**
- **Artificial Intelligence & Machine Learning**
- **Descriptive, Predictive, and Prescriptive Analytics**
- **Operations Research (Optimization)**
- **Illustrative Example: How These Techniques Work Together**

Disclaimer: This presentation does not represent any opinions from Fidelity Investments



Extracted on Sep 15, 2019

Data Scientist in the News

McKinsey Quarterly, 2009

Hal Varian on Statistics and Data

Posted to Quotes, Statistics

> I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?

Hal Varian, The McKinsey Quarterly, January 2009

HBR, 2012

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

LinkedIn, 2016 - 2019

Statistical Analysis & Data Mining:
#2 Hottest Skill (2016 – 2018)

A.I. #2 Hard Skill 2019

Glassdoor, 2016 - 2019

Data Scientist:
#1 Best Job since 2016

Data Scientist Is …

“‘Data Scientist’ is a Data Analyst who lives in California”

"A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician."

“A data scientist is a business analyst who lives in New York.”

"A data scientist is a statistician who lives in San Francisco."

"Data Science is statistics on a Mac."

Source: https://datascopeanalytics.com/blog/what-is-a-data-scientist/

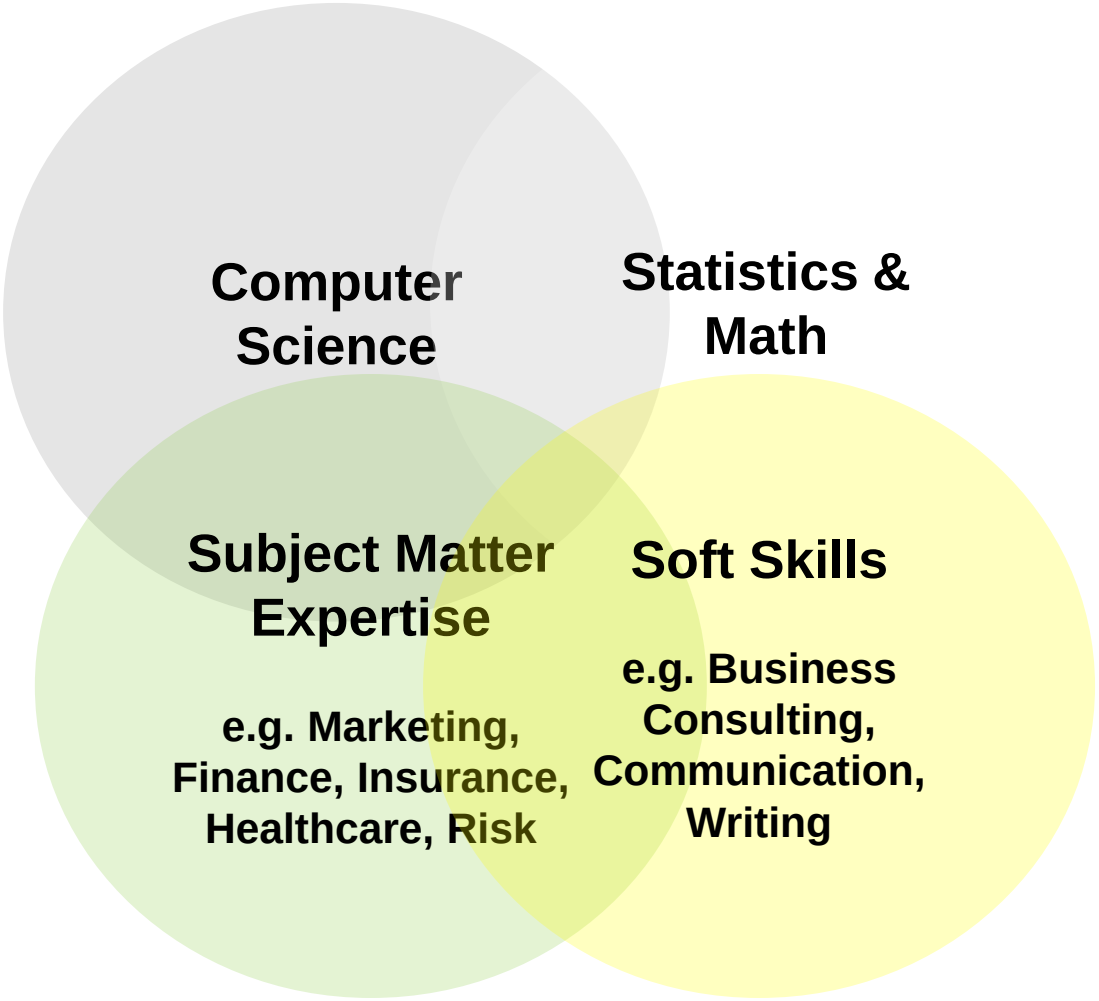
History of Data Science

- Wu 1997, proposed:
  - Statistics → Data Science
  - Statistician → Data Scientist

- Cleveland 2001, proposed:
  - Enlarge the major areas of Statistics → Data Science

Source:
https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#5a5a13cd55cf
https://course.ccs.neu.edu/cs7280sp16/CS7280-Spring16_files/50YearsOfDataScience.pdf

Data Science Venn Diagram

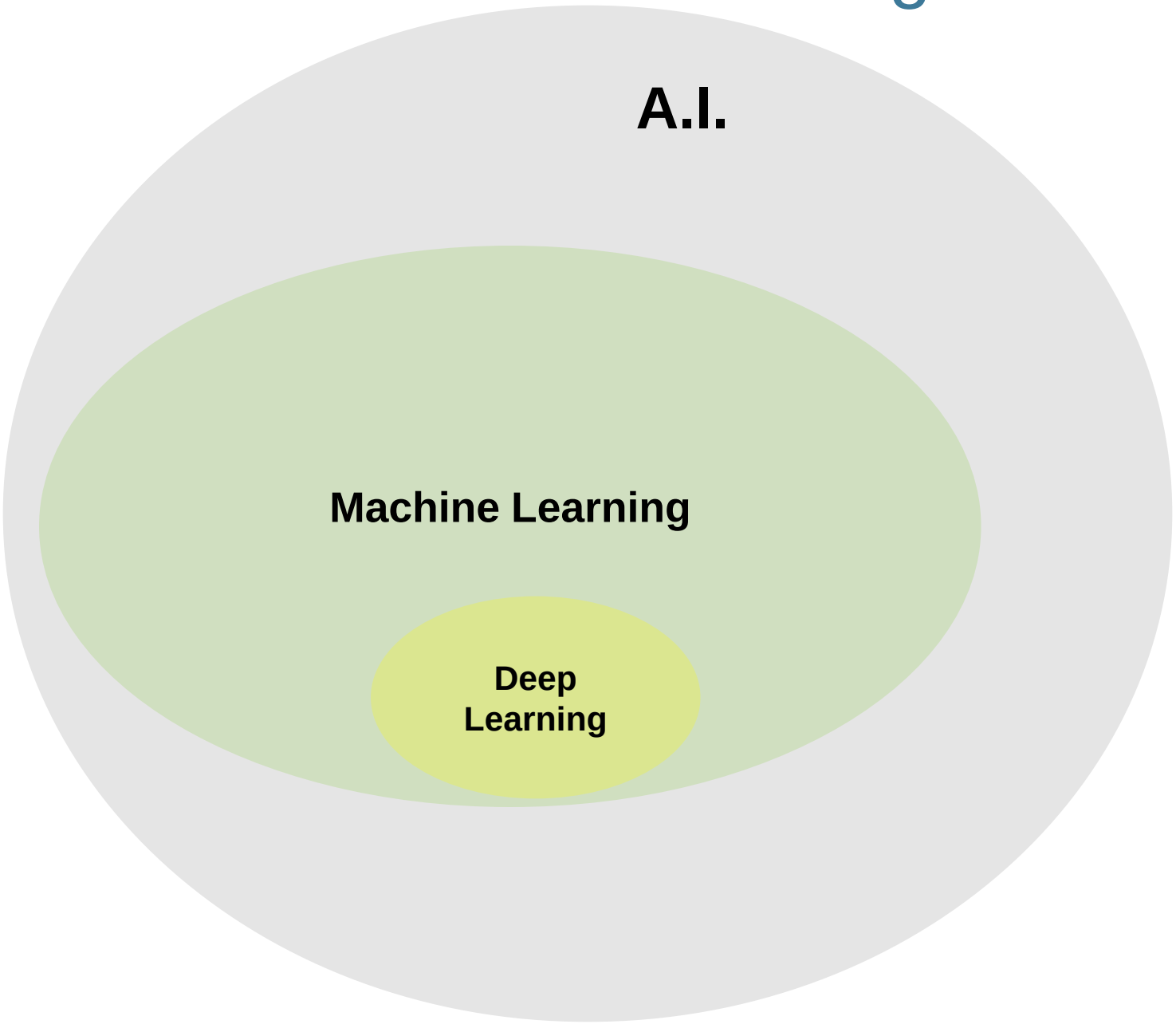


See also: https://pubsonline.informs.org/do/10.1287/LYTX.2019.04.02/full/

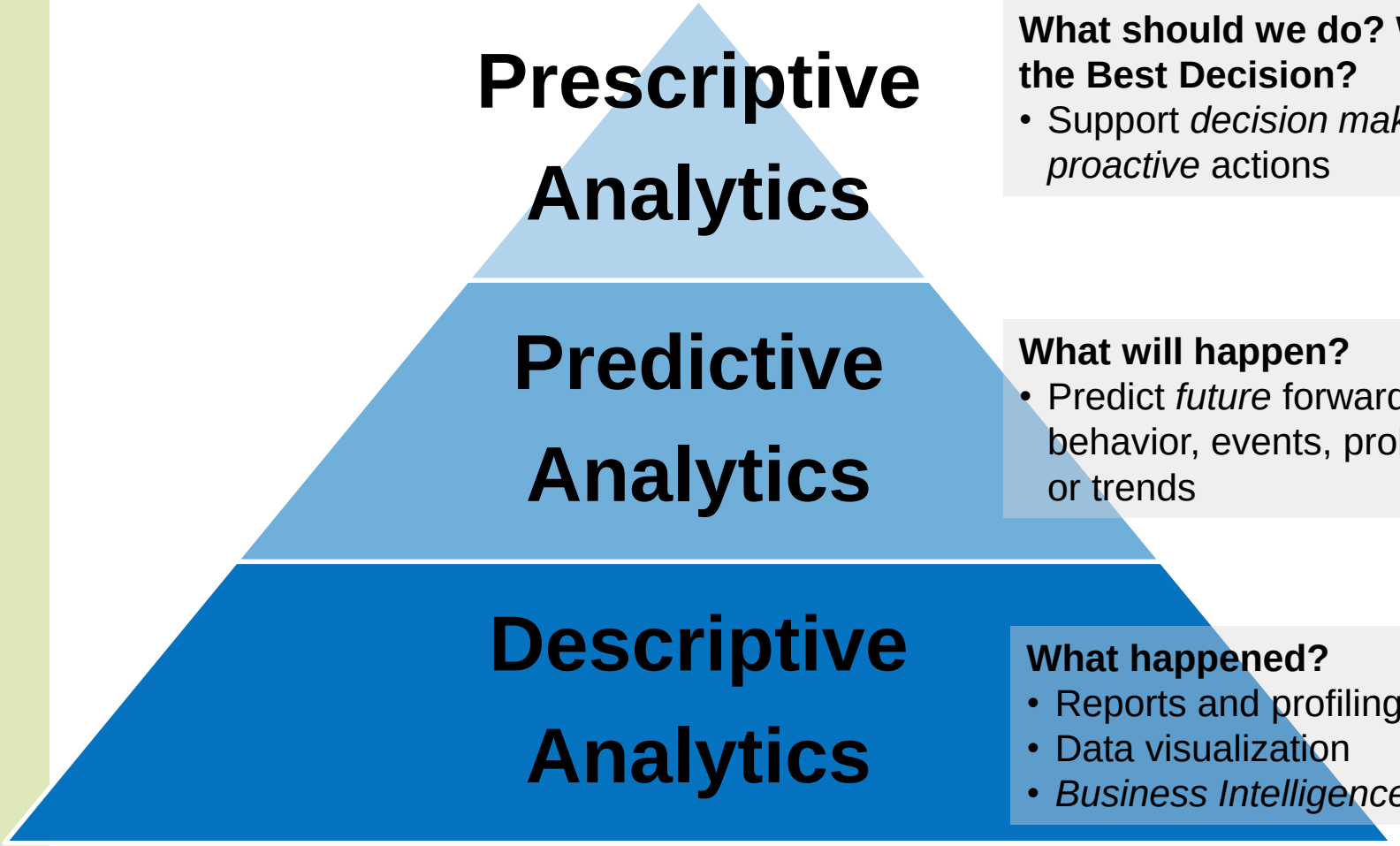
History & Definition of A.I. and Machine Learning

- Definition
- 1956 Dartmouth Workshop
- Rule-Based: Tell a Machine Exactly What to Do
- Machine Learning: Let the Machine Learn by Itself
  - Feed data and set a goal
  - Highly related to Statistics

A.I. and Machine Learning Relationship



Analytics (Data Analytics, Business Analytics)



What should we do? What is the Best Decision?
- Support *decision making* and *proactive* actions

What will happen?
- Predict *future* forward-looking behavior, events, probabilities, or trends

What happened?
- Reports and profiling
- Data visualization
- *Business Intelligence*

Source: http://www.sas.com/news/sascom/2008q4/column_8levels.html, and https://www.informs.org/Community/Analytics

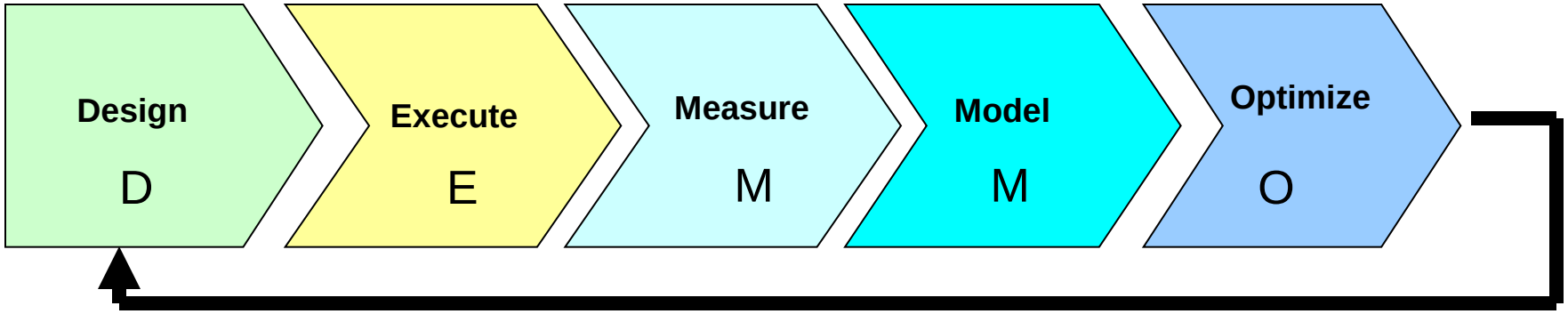
9

Operations Research (OR)

- WWII: Need for making better decisions
- **Optimization**
- Business Applications
- **Management Science** and **Business Analytics**
- Link to **Machine Learning** and **Reinforcement Learning**:
  - Optimizing *weights* in models
  - Reward maximization and self-training

Application: Customer Relationship Management (CRM)

- **Track and optimize contacts with customers**
- **Improve services provided to customers**
- **Use historical customer contact data and apply analytics to improve future customer interactions:**



Looks familiar?

- Industrial product design
- Six Sigma
- Clinical trials

Application: CRM Optimization

Individual Characteristics

Treatments

Channel

Message/Offer

Individuals

Best deal…

No risk free trial…

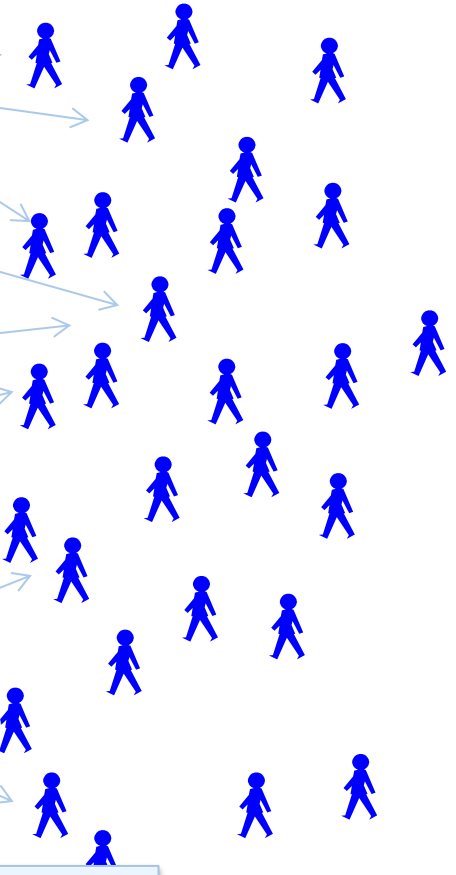
Lowest price ever…

Super benefits…

Unbeatable service…

Top performance…

Scientifically proven…



Lots of Possible Treatment Combinations

APPENDIX

Translation Between Statistics and AI / ML:

Same or Similar Terminology

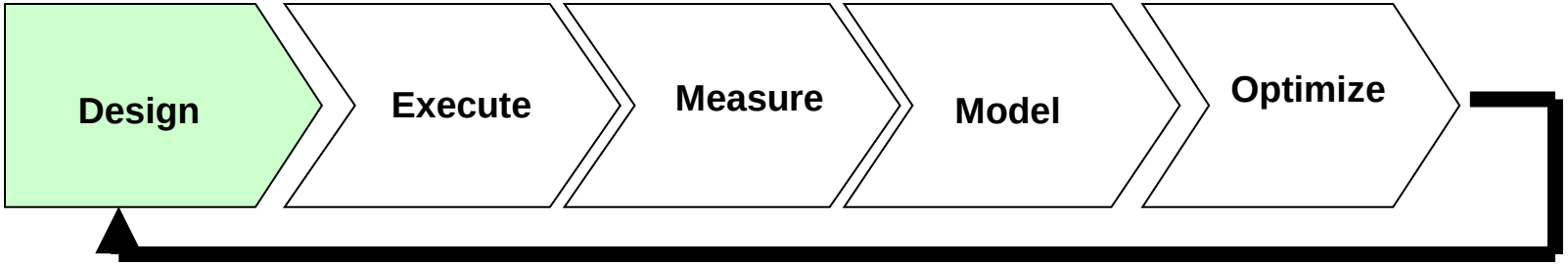
Statistics / Economics / Epidemiology / Math	Data Science / AI / Data Mining
Statistical modeling	Machine Learning
Dependent Variable / Response Variable	Target Variable / Label
Independent Variable	Feature[1]
Parameters / coefficients	Weights
Intercept	Bias[2]
Estimation	Training
Out-of-Sample / Holdout Sample	Test Data
Regression / Classification	Supervised Learning
Cluster Analysis / PCA / Factor Analysis / SVD	Unsupervised Learning
Variable Selection	Feature Selection
Dimension Reduction	Feature Reduction
Data point / observation	Instance / Sample[3] / Example
Outlier Detection	Anomaly Detection
Log likelihood function of a binary variable	Cross Entropy
Logistic function	Sigmoid function
Multinomial Logit	Softmax
Dummy Coding	One-hot Coding
Misclassification Table	Confusion Matrix
Bayesian Computation	Probabilistic Programming
Approximate Dynamic Programming/Markov Decision Process	Reinforcement Learning
Randomized Controlled Trial (RCT)	A/B Testing
Factorial Design	Multivariate Testing (MVT)
Time series data	Sequential data
Classification Matrix	Confusion Matrix
Power [P(Reject H0 | H1 is true) or 1-P(Type II error)]	Recall
False Discovery Rate (FDR)	1 – Precision
Average Treatment Effect (ATE)	Lift (Marketing)
Heterogeneous Treatment Effect (Econ.)	Uplift Modeling
Or Conditional Average Treatment Effect (CATE; Econ.)	Uplift Modeling
Or, Effect Modification (Epidemiology)	Uplift Modeling
Or, Impactibility Modeling (Health)	Uplift Modeling
Or, Subgroup Analysis (Biostat)	Uplift Modeling

[1] A feature can also be a function of original variables.

[2] The standard statistical definition of Bias is the discrepancy between the actual value of an unknown parameter and the expected value of its estimator. Such definition is also used in machine learning, which is totally different from the Intercept-equivalent meaning in neural networks.

[3] The traditional definition of a sample refers to a subset of the population, which is a collection of observations. In some AI/ML literature, a single observation is sometimes called a sample.

Customer Relationship Management (CRM) – **Campaign Design**



Business Problem

- Utilize past learning to maximize campaign response
- Design campaign to maximize learning for next time

Solution

- Find best targets and best treatment using predictive modeling and optimization
- Experimental Design
- Sample Size Determination
- Sampling Techniques

Experimental Design Approach to Campaign Design

For Direct Marketing - testing particular treatment combinations of product and marketing features:

Treatment Combinations for a Banking Credit Card Email Campaign

Hypothetical

Attributes

Re-mail	APR	Credit Limit $	Color of Font	Offer	Subject Line
yes	7.9%	2000	Blue	Incentive	Introducing...
no	12.9%	5000	Black	No Incentive	Hi from XYZ Bank...
		8000			Great value from...
		12,000			Don't miss...
Attribute Levels: 2	2	4	2	2	4

Attribute Levels

256 possible offer combinations – A Full Factorial Design

From 256 to 32 cells

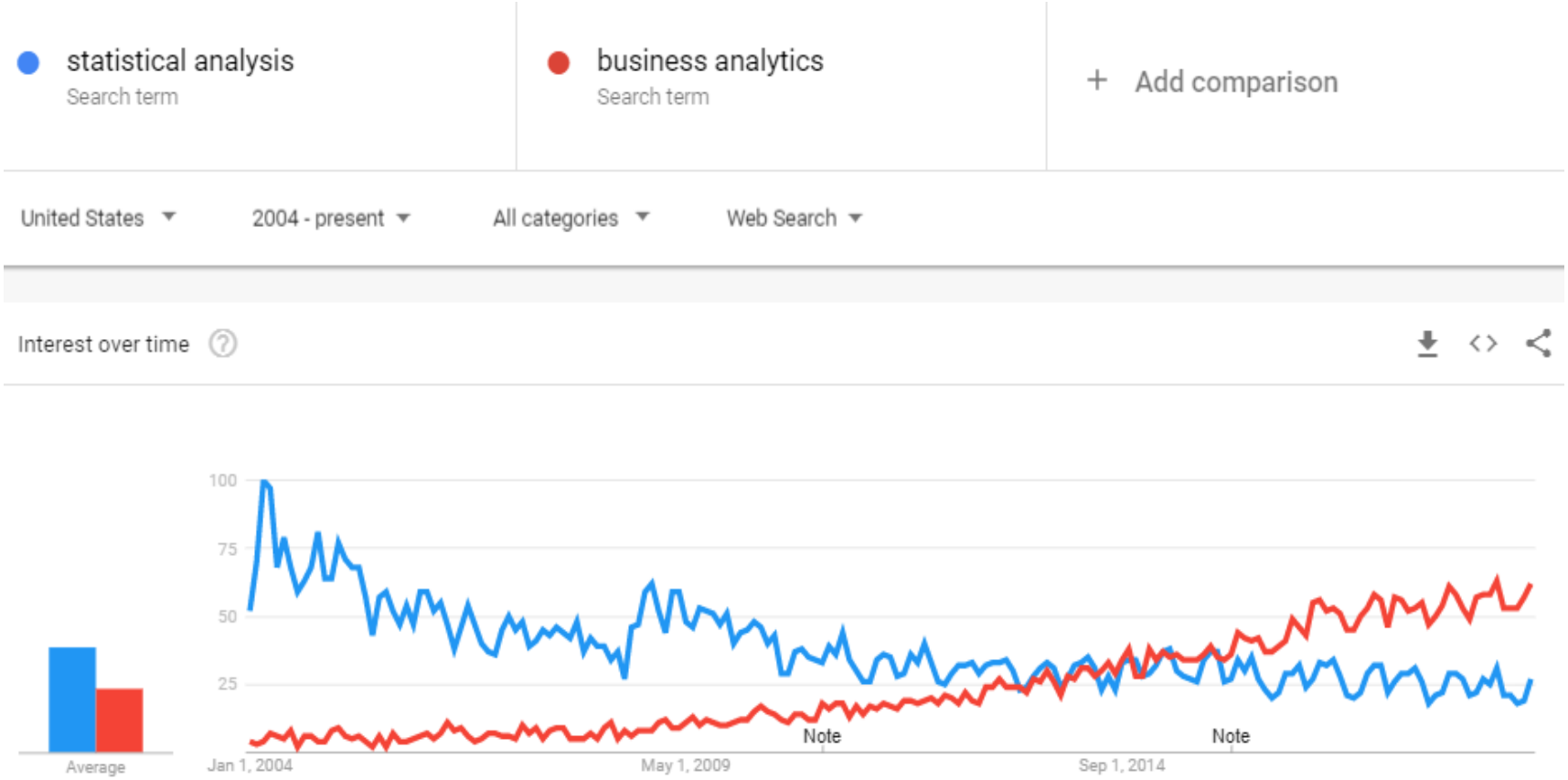
Fractional Factorial Design

Statistically Reduced to

cell	remail	APR	Credit limit	color	offer	subject
1	no	7.90%	2000	black	no incentive	Introducing...
2	no	7.90%	2000	blue	with incenti	Do not miss...
3	no	7.90%	5000	black	with incenti	Great value fr
4	no	7.90%	5000	blue	no incentive	Hi from XYZ Bank…
5	no	7.90%	8000	black	with incenti	Hi from XYZ Bank…
6	no	7.90%	8000	blue	no incentive	Great value fr
7	no	7.90%	12000	black	no incentive	Do not miss...
8	no	7.90%	12000	blue	with incenti	Introducing...
:	:	:	:	:	:	:
:	:	:	:	:	:	:

An 87.5% savings in # cells

Search Patterns of Two Terms



Extracted on Sep 22, 2019