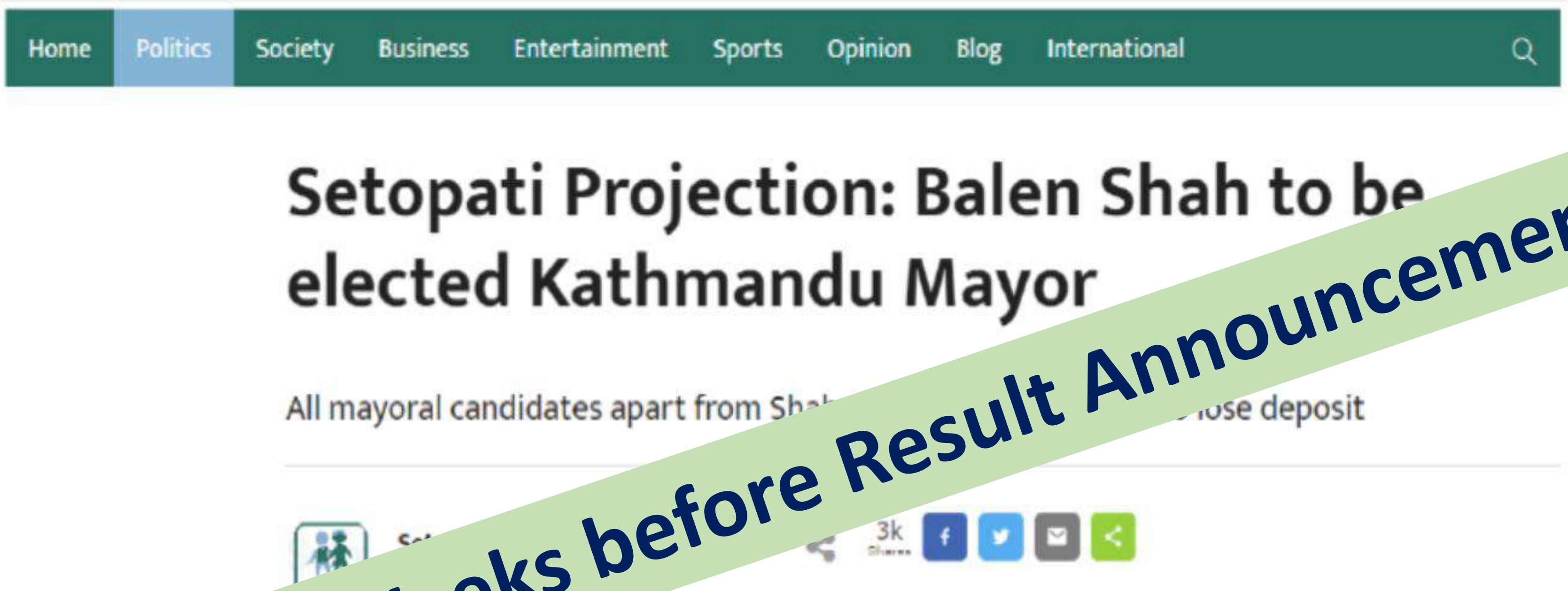
# Foundation of Data Science and Analytics

## 1. Introduction

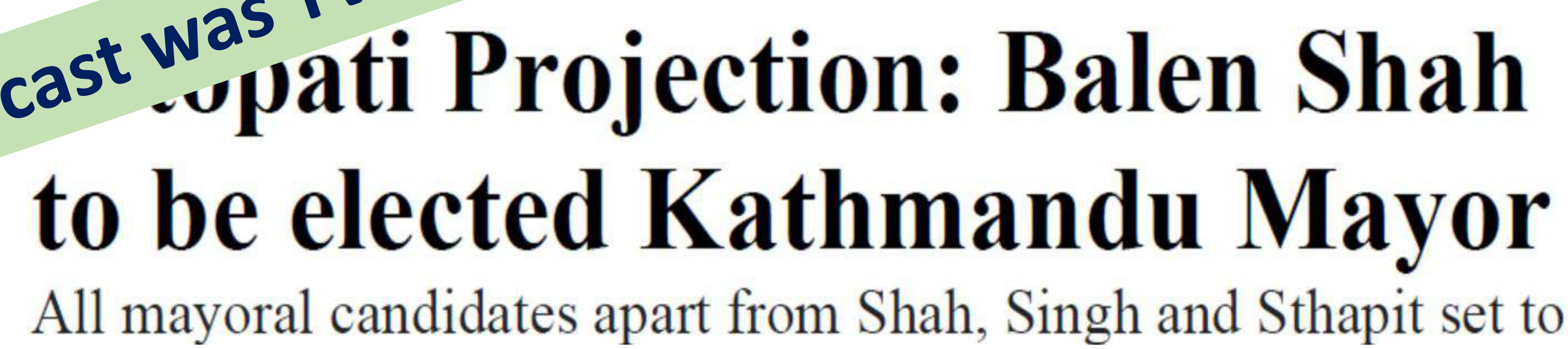
Arun K. Timalsina, PhD

# Forecast was TWO Weeks before Result Announcement!



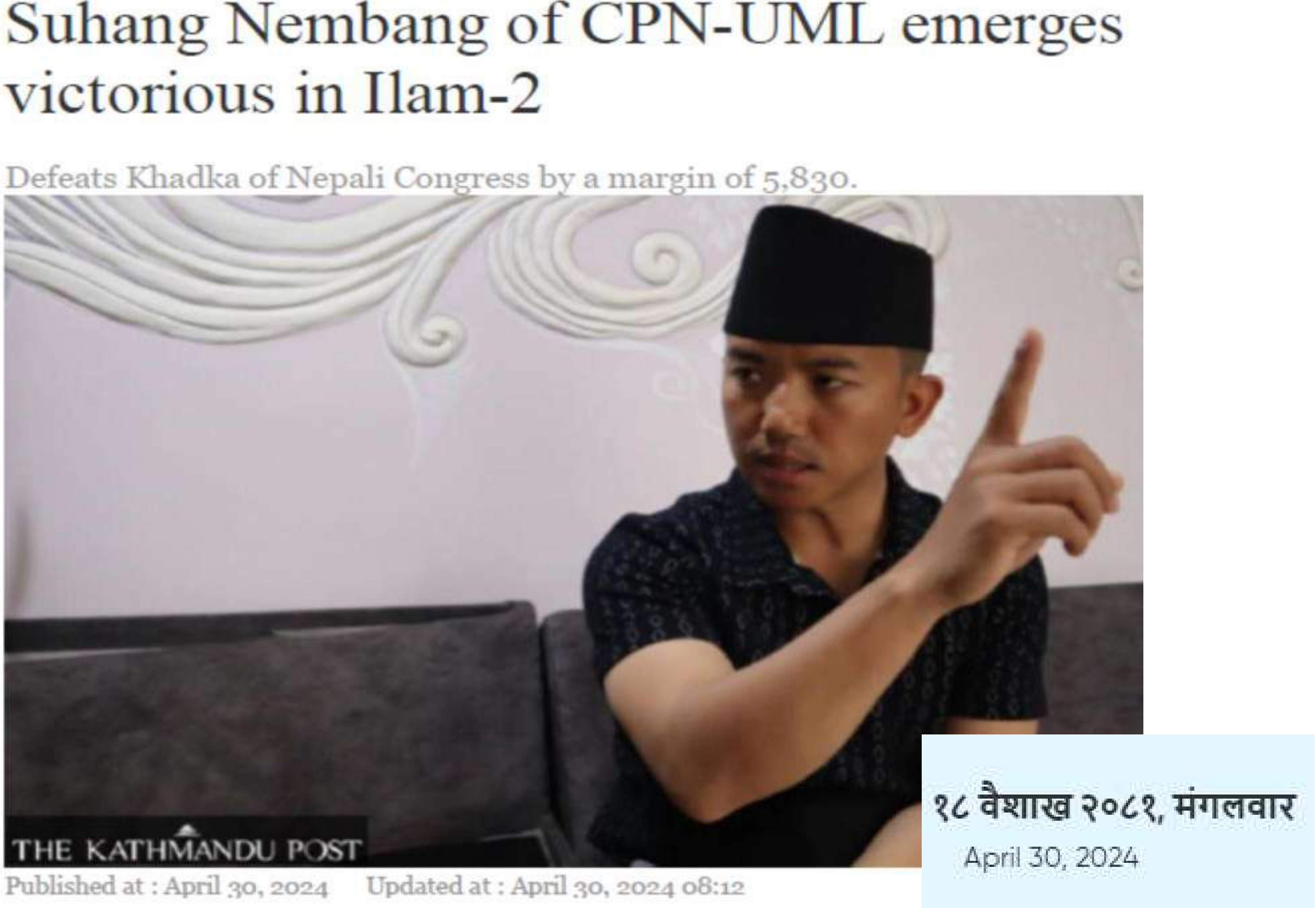
## Setopati Projection: Balen Shah to be elected Kathmandu Mayor

All mayoral candidates apart from Sh lose deposit



## opati Projection: Balen Shah to be elected Kathmandu Mayor

All mayoral candidates apart from Shah, Singh and Sthapit set to



# Suhang Nembang of CPN-UML emerges victorious in Ilam-2

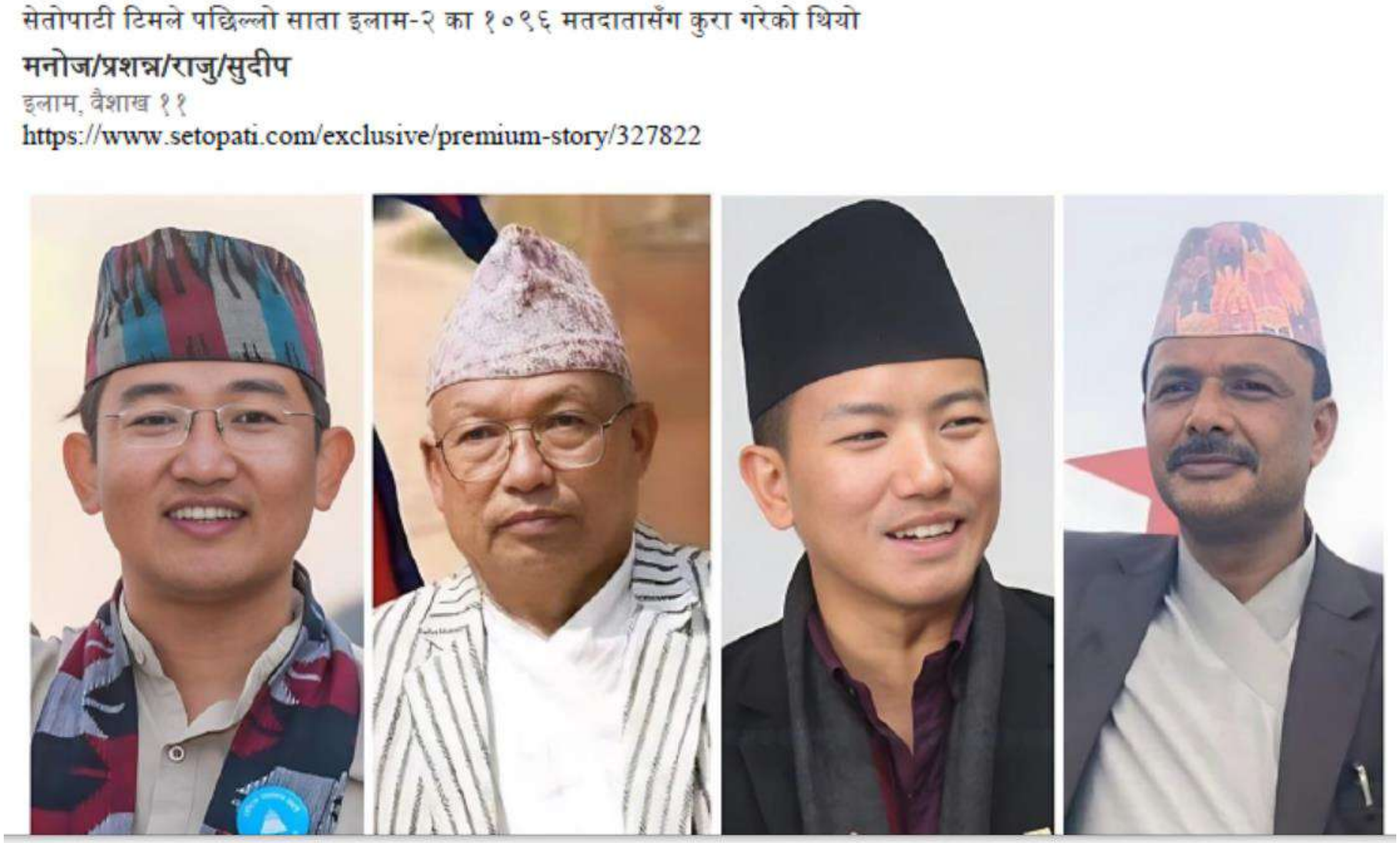
Defeats Khadka of Nepali Congress by a margin of 5,830.

Published at : April 30, 2024 Updated at : April 30, 2024 08:12

१८ वैशाख २०८१, मंगलवार

April 30, 2024

# इलाम-२ को निर्वाचनबारे सेतोपाटी विश्लेषण



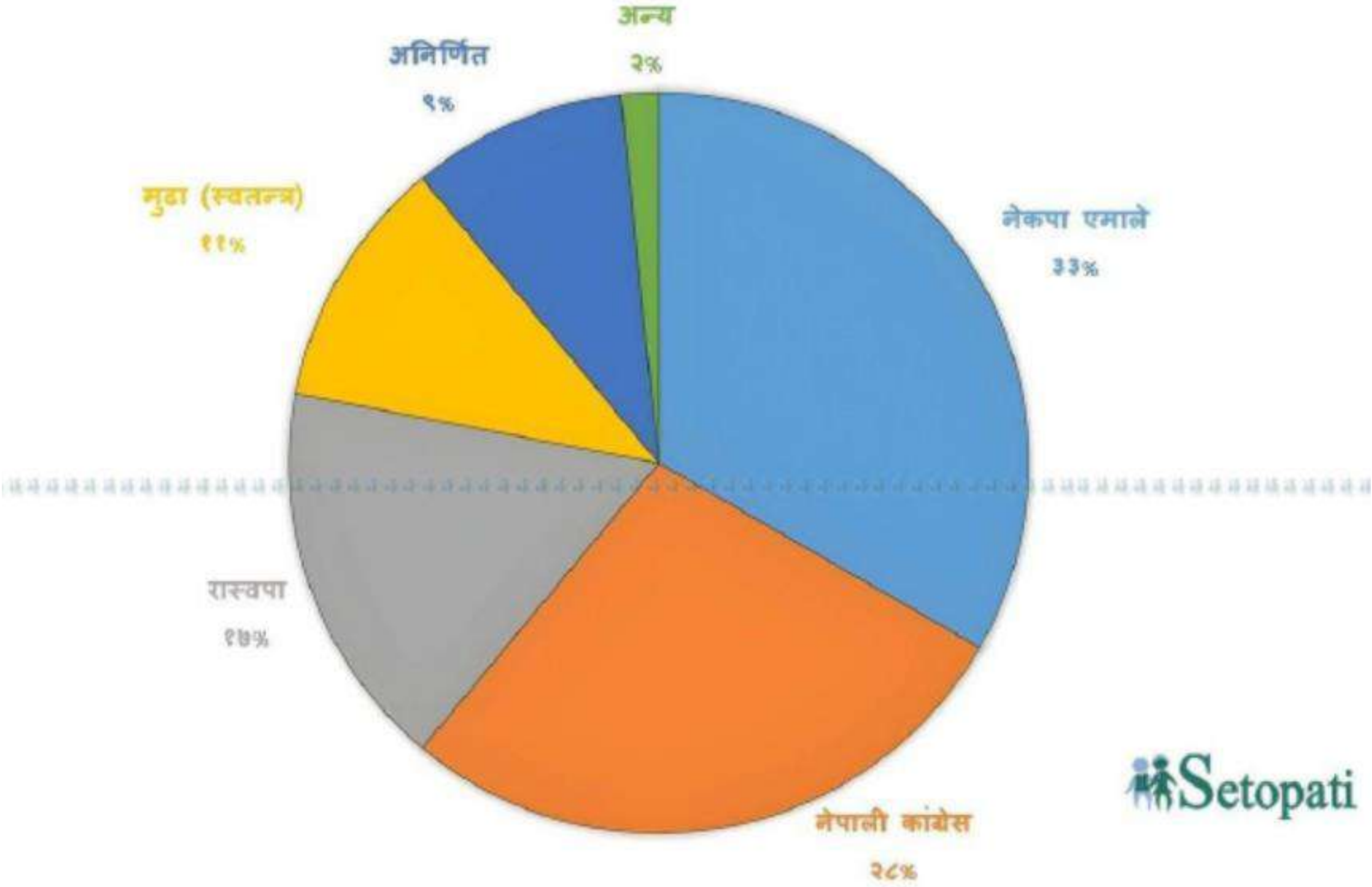
सेतोपाटी टिमले पछिल्लो साता इलाम-२ का १०९६ मतदातासँग कुरा गरेको थियो

**मनोज/प्रशन्न/राजु/सुदीप**

इलाम, वैशाख ११

https://www.setopati.com/exclusive/premium-story/327822

# इलाम-२ को चुनावी विश्लेषण



https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did

Feb 16, 2012 11:02am EST

# How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill Former Staff

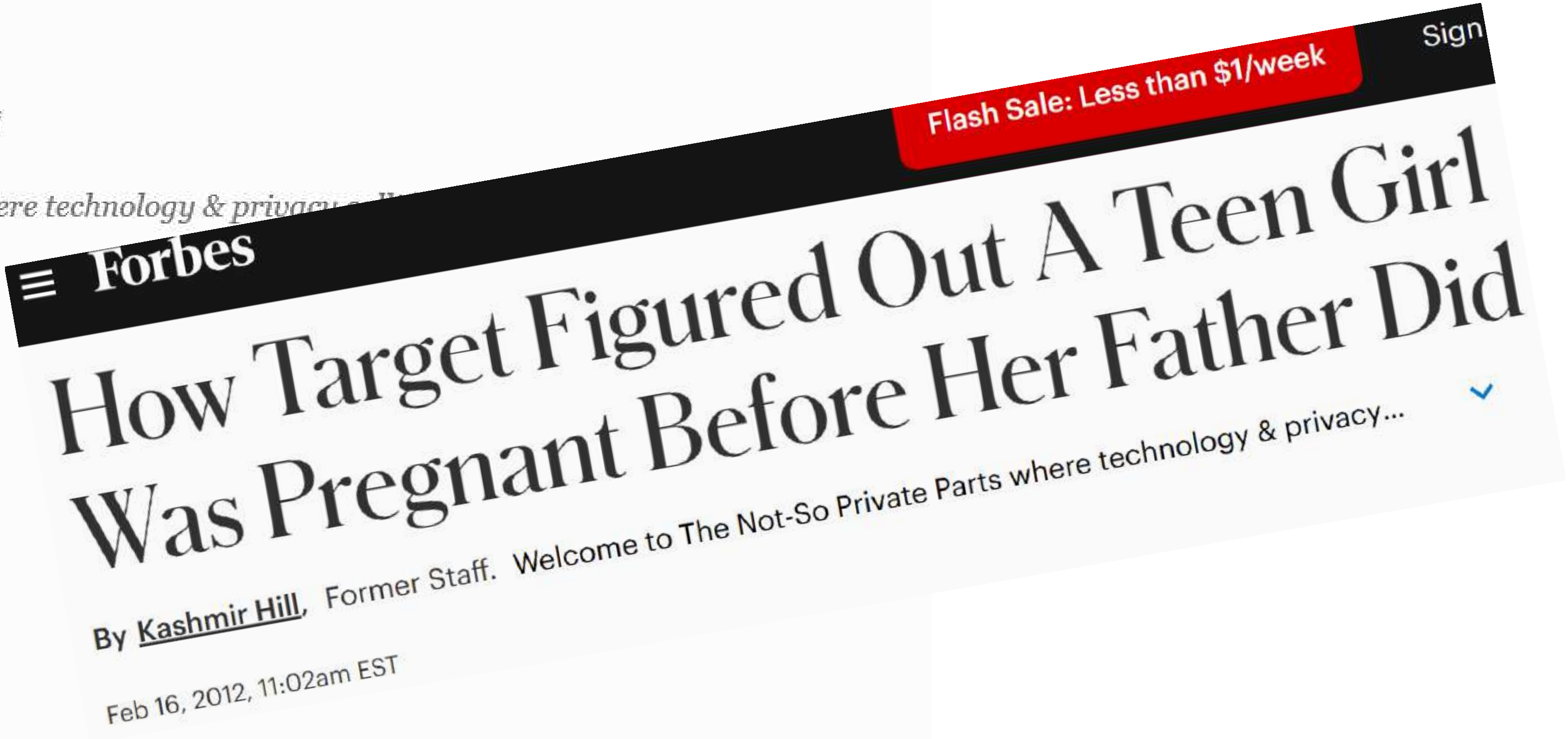
Tech

*Welcome to The Not-So Private Parts where technology & privacy*

**This article is more than 8 years old.**

Target has got you in its aim

Every time you go shopping, you share intimate details about your consumption



How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

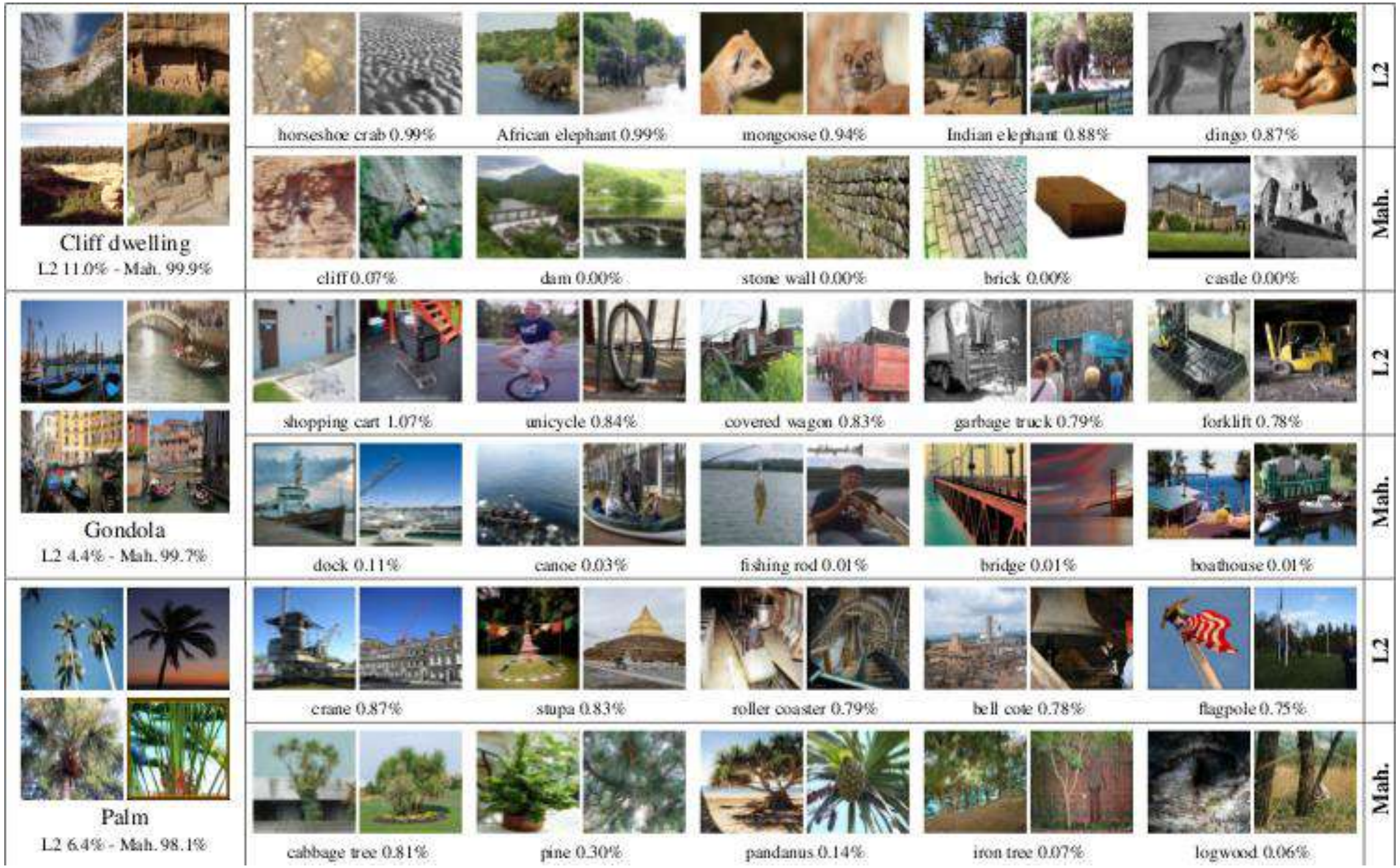
By Kashmir Hill, Former Staff. Welcome to The Not-So Private Parts where technology & privacy...

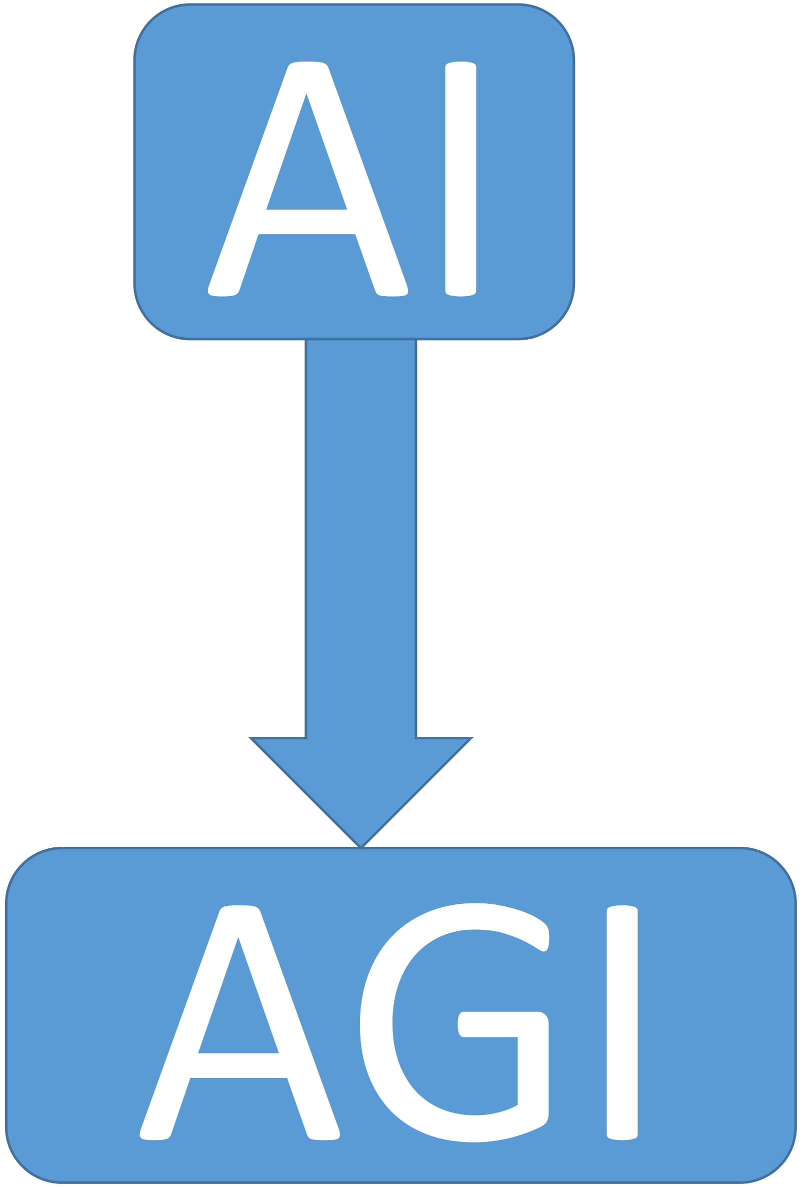
Feb 16, 2012, 11:02am EST

# Plausible Effectiveness of Deep Learning

2012 Imagenet challenge:

Classify 1 million images into 1000 classes.





AI

AGI

# What  the course FDSA is about ………?

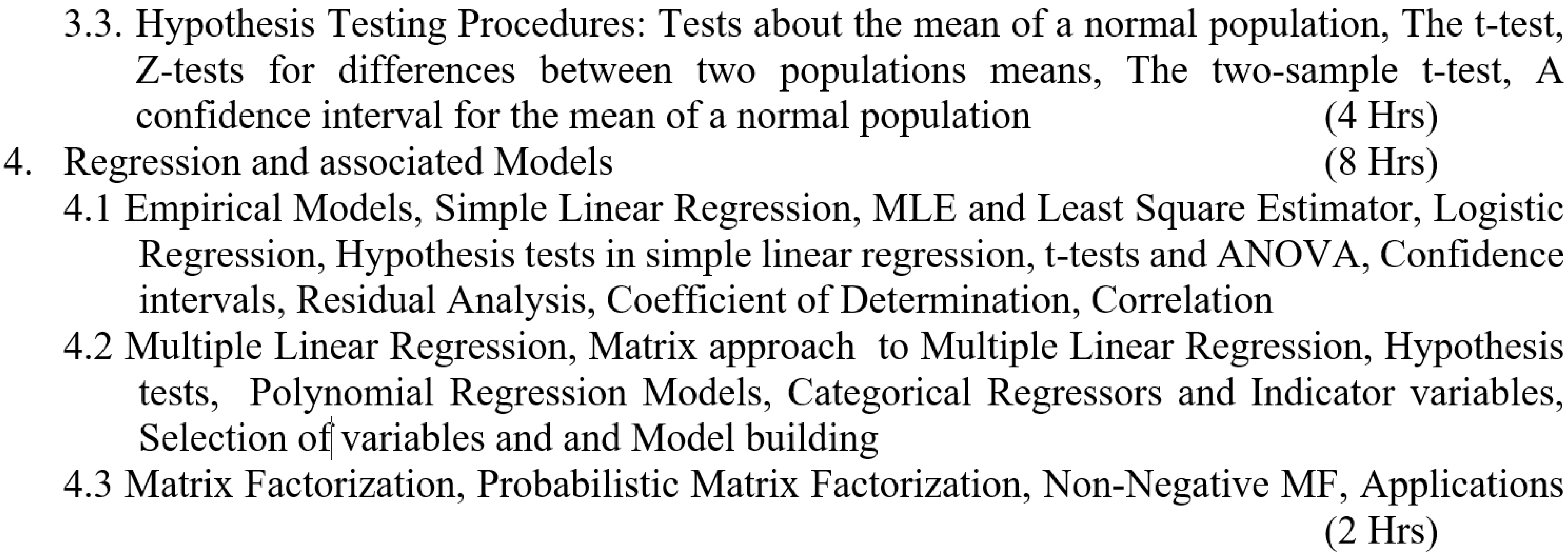
# Foundation of Data Science & Analytics

- Overall Summary of Data Science & Analytics
- Mathematics of Data Analysis
  - Basic Statistics , Regression, Matrix factorization
- Data Wrangling /Cleaning (EDA)
- Model and Evaluation specifics and setups
- OLTP/OLAP – NoSQL Specifics
- Related Research Trends

# Course Contents

1. Introduction to Data Science (3 Hrs)
   Data Science Hype, Why data science, Getting Past the Hype, The Current Landscape, Role of Data Scientist
2. Data Types and Data Science Processes (7 Hrs)
   2.1. Facets of data:Structured data,Unstructured data, Natural language, Machine-generated data , Graph-based or network data, Audio, image, and video, Streaming data
   2.2. Process Overview, Defining goals, Retrieving data, Data preparation, Exploratory Data Analysis, Data Wrangling & Cleaning, Data Integration and Transformation, Data Reduction, Data modeling and Result Presentation
3. Mathematical Foundation for Data Science (20 Hrs)
   3.1. Introduction and Descriptive Statistics : An overview of probability and statistics, Pictorial and tabular methods in descriptive statistics, Measures of central tendency, dispersion, and direction, Joint and conditional probabilities, Central limit theorem (4 Hrs)
   3.2. Random Variables and Probability Distributions: Random variables, Probability distributions for random variables, Expected values of discrete random variables and continuous distributions, The binomial probability distribution, Hypothesis testing using the binomial distribution, The Poisson probability distribution (4 Hrs)

# Course Contents



3.3. Hypothesis Testing Procedures: Tests about the mean of a normal population, The t-test, Z-tests for differences between two populations means, The two-sample t-test, A confidence interval for the mean of a normal population (4 Hrs)

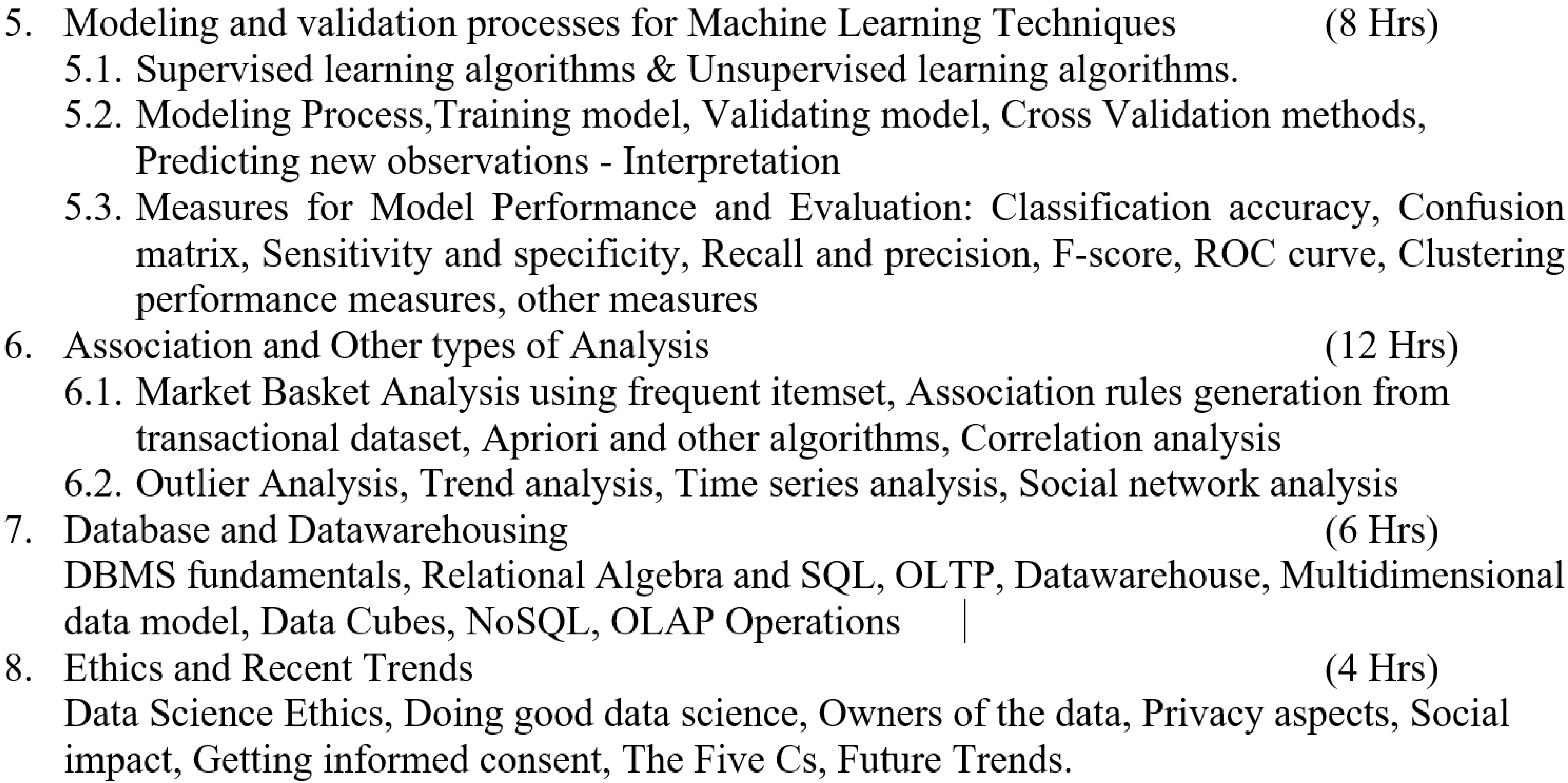
4. Regression and associated Models (8 Hrs)

4.1 Empirical Models, Simple Linear Regression, MLE and Least Square Estimator, Logistic Regression, Hypothesis tests in simple linear regression, t-tests and ANOVA, Confidence intervals, Residual Analysis, Coefficient of Determination, Correlation

4.2 Multiple Linear Regression, Matrix approach to Multiple Linear Regression, Hypothesis tests, Polynomial Regression Models, Categorical Regressors and Indicator variables, Selection of variables and and Model building

4.3 Matrix Factorization, Probabilistic Matrix Factorization, Non-Negative MF, Applications (2 Hrs)

# Course Contents



5. Modeling and validation processes for Machine Learning Techniques (8 Hrs)

   5.1. Supervised learning algorithms & Unsupervised learning algorithms.

   5.2. Modeling Process,Training model, Validating model, Cross Validation methods, Predicting new observations - Interpretation

   5.3. Measures for Model Performance and Evaluation: Classification accuracy, Confusion matrix, Sensitivity and specificity, Recall and precision, F-score, ROC curve, Clustering performance measures, other measures

6. Association and Other types of Analysis (12 Hrs)

   6.1. Market Basket Analysis using frequent itemset, Association rules generation from transactional dataset, Apriori and other algorithms, Correlation analysis

   6.2. Outlier Analysis, Trend analysis, Time series analysis, Social network analysis

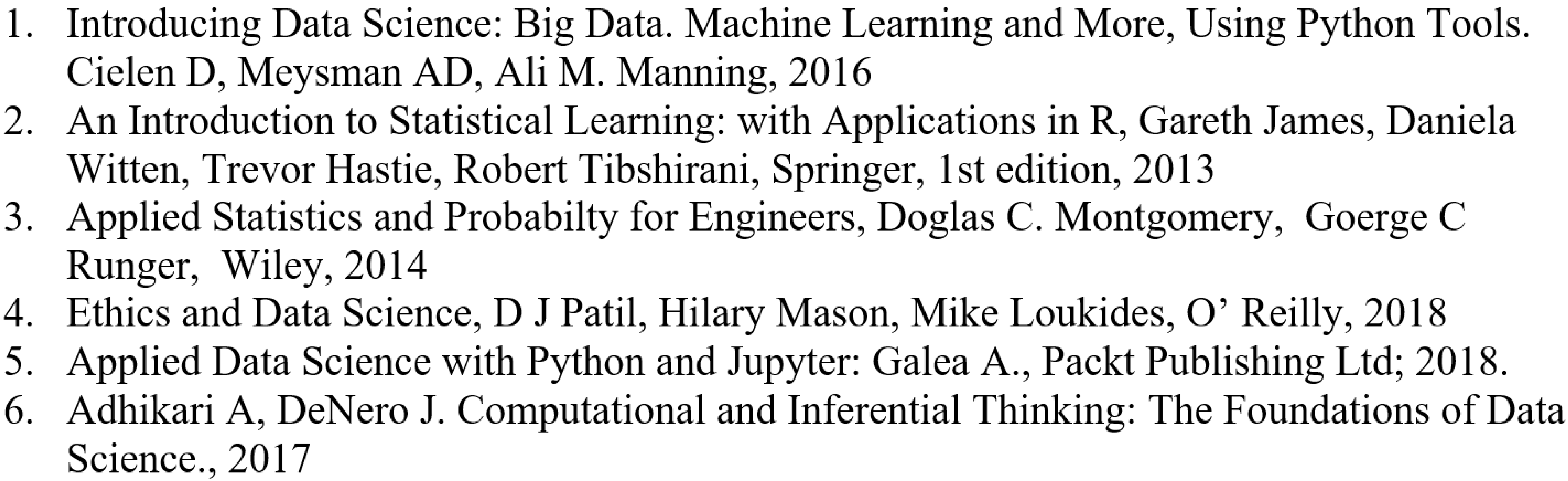
7. Database and Datawarehousing (6 Hrs)

   DBMS fundamentals, Relational Algebra and SQL, OLTP, Datawarehouse, Multidimensional data model, Data Cubes, NoSQL, OLAP Operations

8. Ethics and Recent Trends (4 Hrs)

   Data Science Ethics, Doing good data science, Owners of the data, Privacy aspects, Social impact, Getting informed consent, The Five Cs, Future Trends.

# References



1. Introducing Data Science: Big Data. Machine Learning and More, Using Python Tools. Cielen D, Meysman AD, Ali M. Manning, 2016
2. An Introduction to Statistical Learning: with Applications in R, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Springer, 1st edition, 2013
3. Applied Statistics and Probabilty for Engineers, Doglas C. Montgomery, Goerge C Runger, Wiley, 2014
4. Ethics and Data Science, D J Patil, Hilary Mason, Mike Loukides, O' Reilly, 2018
5. Applied Data Science with Python and Jupyter: Galea A., Packt Publishing Ltd; 2018.
6. Adhikari A, DeNero J. Computational and Inferential Thinking: The Foundations of Data Science., 2017

# Data Analysis : Timeline

1935: “The Design of Experiments”

R.A. Fisher

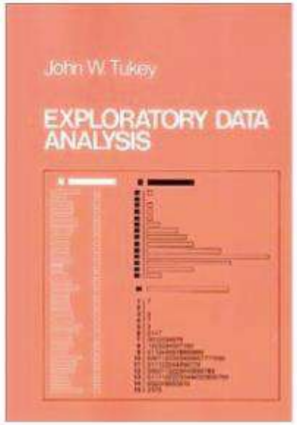
1939: “Quality Control”

W.E. Demming

1958: “A Business Intelligence System”

Peter Luhn

1977: “Exploratory Data Analysis”



1989: “Business Intelligence”

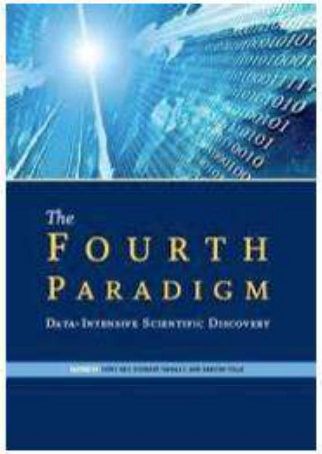
Howard Dresner

# Data Analysis : Timeline

1996: Google

1997: “Machine Learning”

2007:“The Fourth Paradigm”

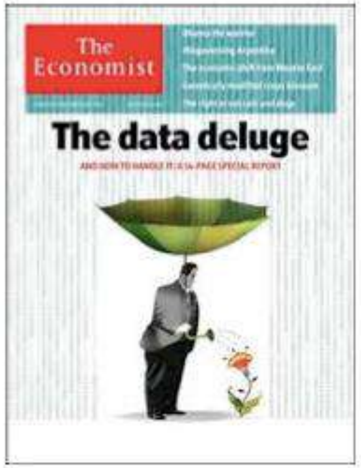


First 3 paradigms of science : Empirical, Theoretical and Simulation.
4th Data Driven Science

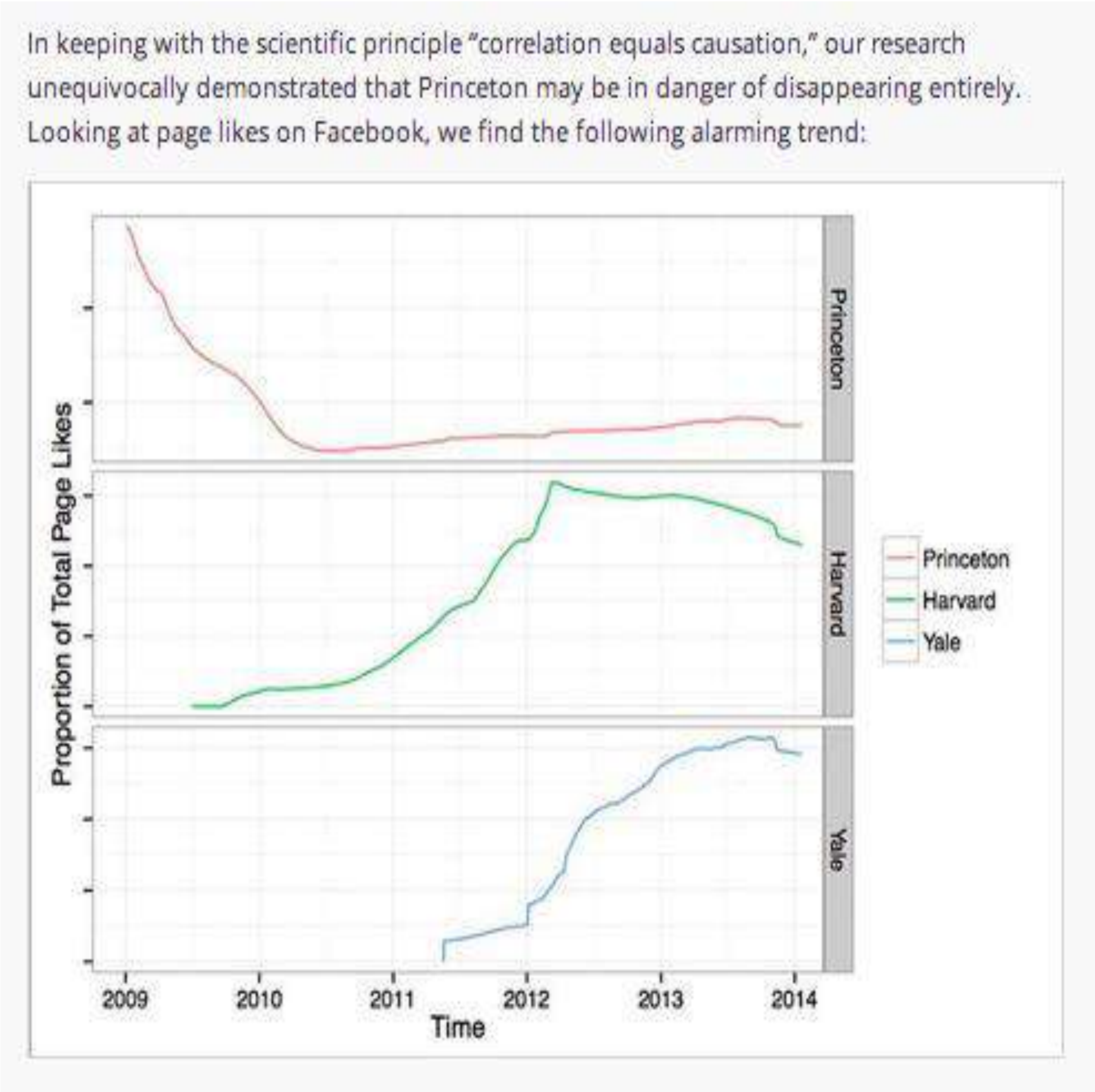
2009:“The Unreasonable Effectiveness of Data”

Peter Norvig :
Simple Model
+ Voluminous Data
➔ Complex Model

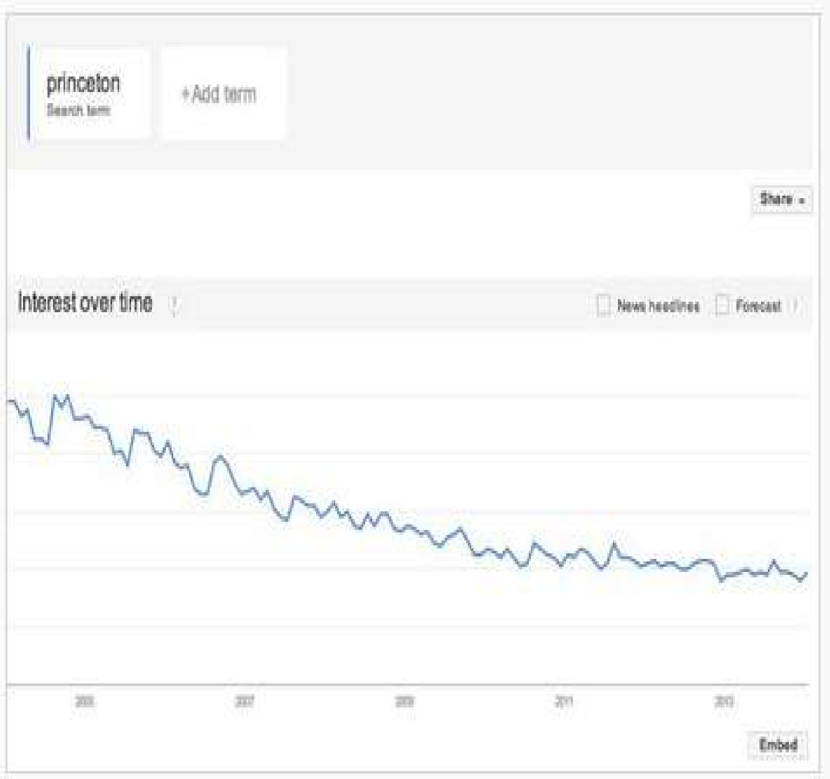
2010:“The Data Deluge”



# Data Makes Everything Clearer



In keeping with the scientific principle "correlation equals causation," our research unequivocally demonstrated that Princeton may be in danger of disappearing entirely. Looking at page likes on Facebook, we find the following alarming trend:



and based on Princeton search trends:

"This trend suggests that Princeton will have only  half its current enrollment by 2018, and by 2021 it will have no students at all,…

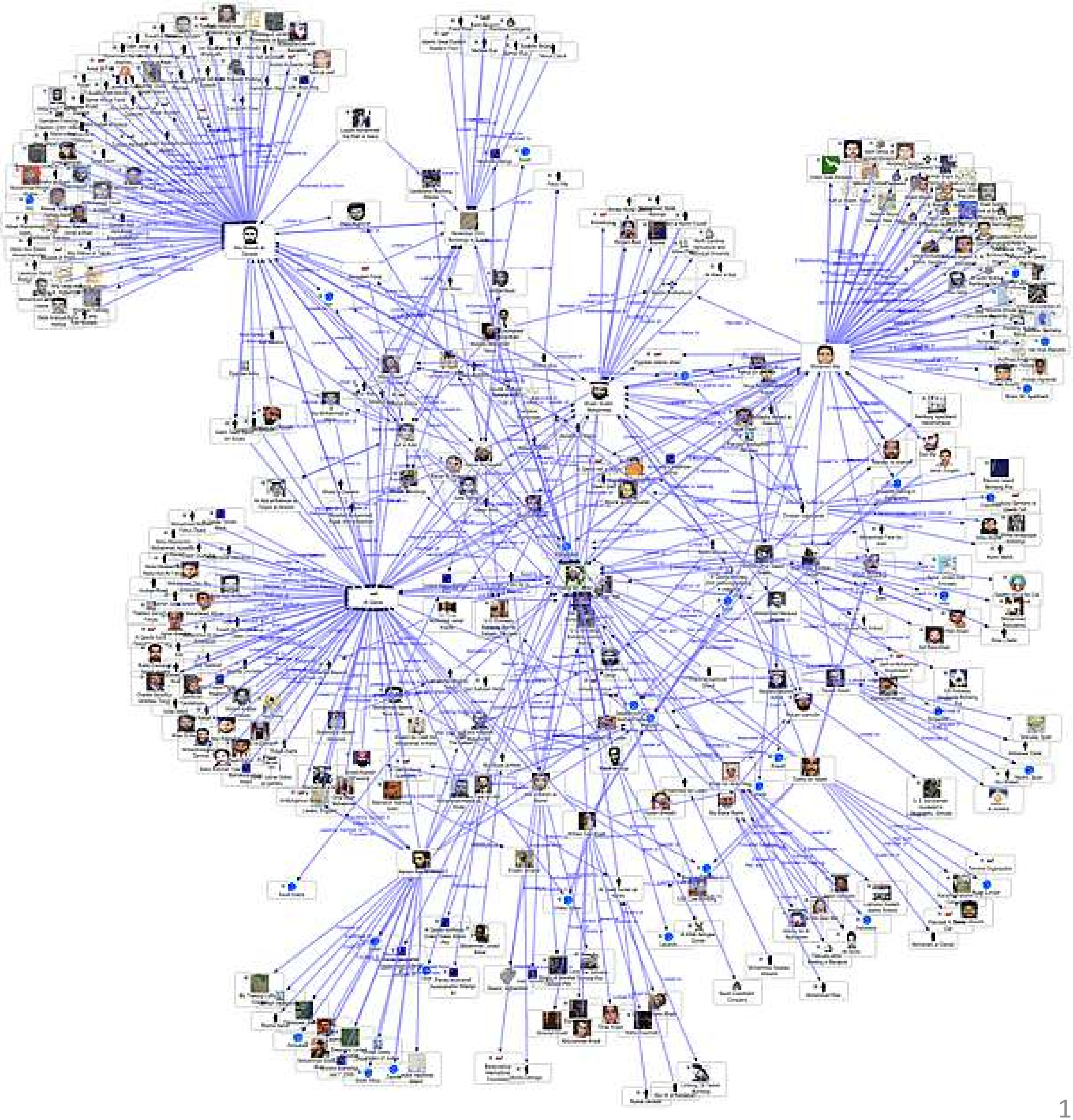
http://techcrunch.com/2014/01/23/facebook-losing-users-princeton-losing-credibility/

# Graph Data

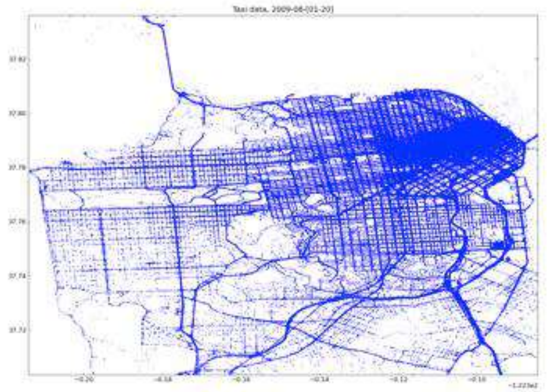
Lots of interesting data has a graph structure:

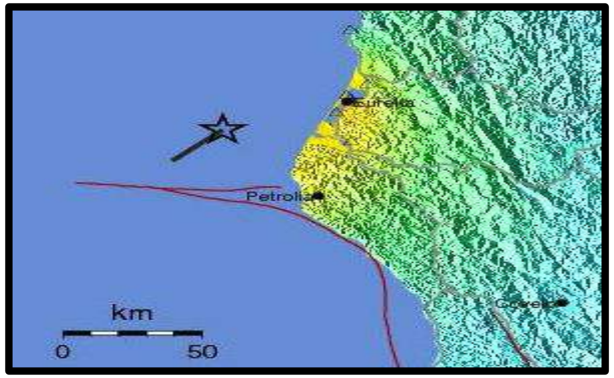
- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- …

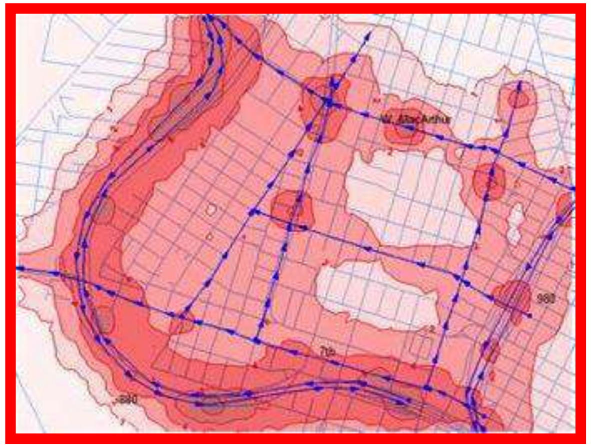
Some of these graphs can get quite large (e.g., Facebook* user graph)



# What *can not be done* with the data?

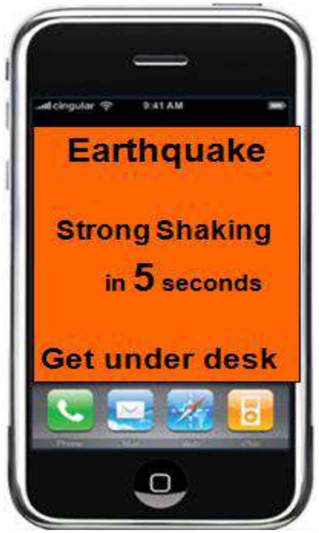




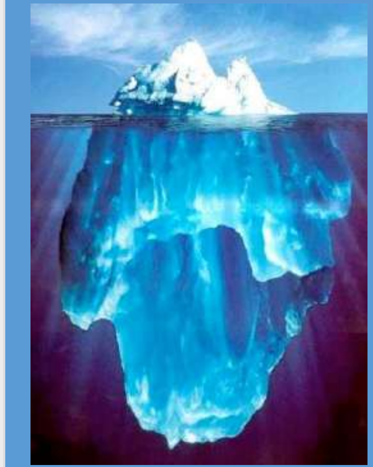


Crowdsourcing + physical modeling + sensing + data assimilation

to produce:



## It's All Happening On-line



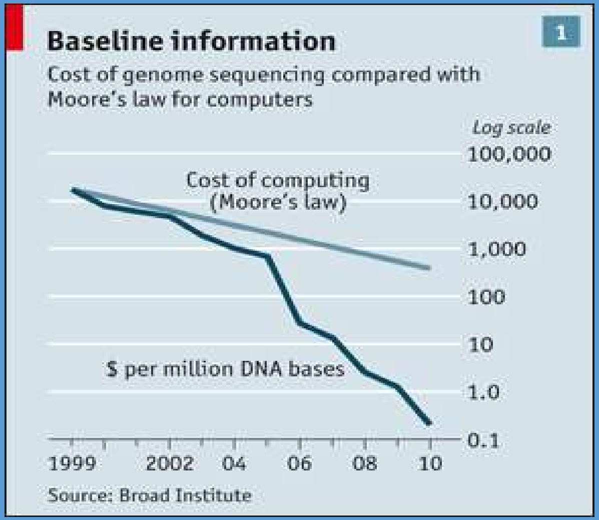
Every
Click
Ad impression
Billing event
Fast Forward, pause,…
Server request
Transaction
Network message
Fault
…

## User Generated (Web & Mobile)

…

## Internet of Things / M2M

## Health/Scientific Computing



**Baseline information**

Cost of genome sequencing compared with Moore's law for computers

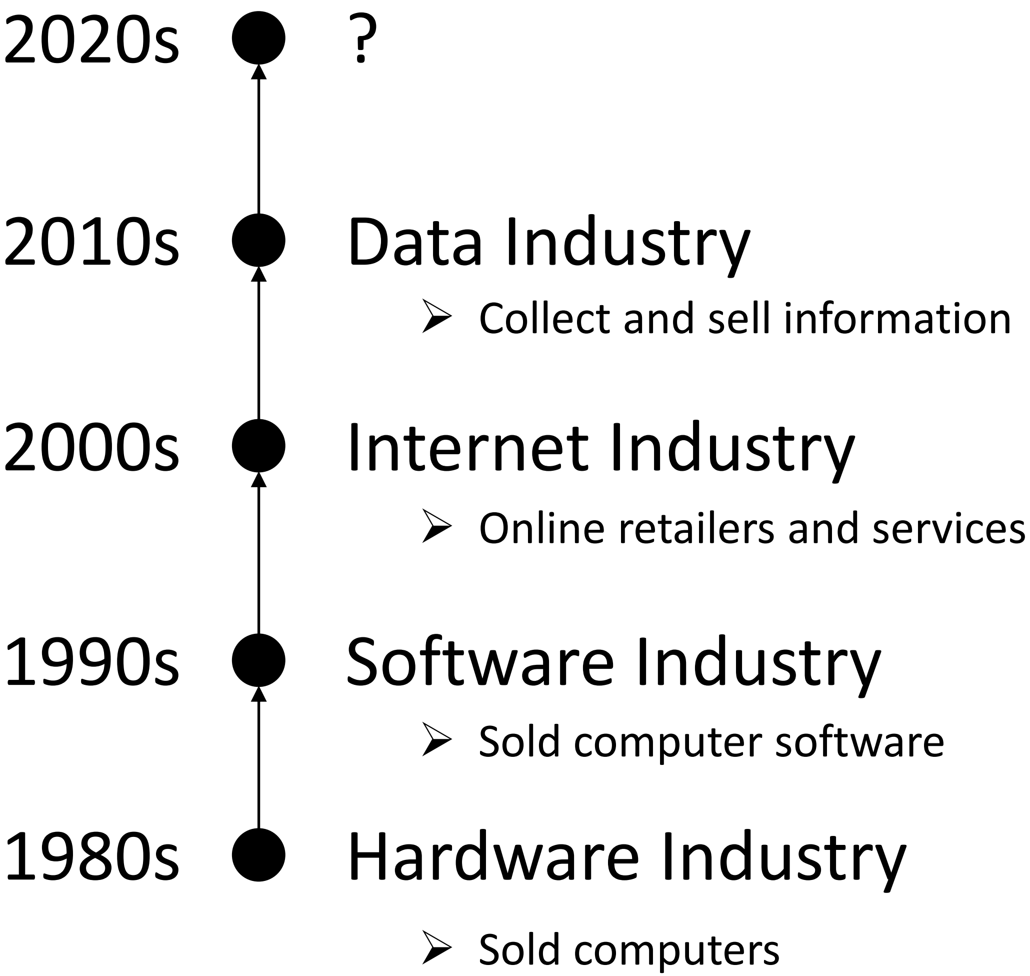
Source: Broad Institute

# Data

# To

# "Big Data"

# Technology Trends



2020s ?

2010s Data Industry

- ➢ Collect and sell information

2000s Internet Industry

- ➢ Online retailers and services

1990s Software Industry

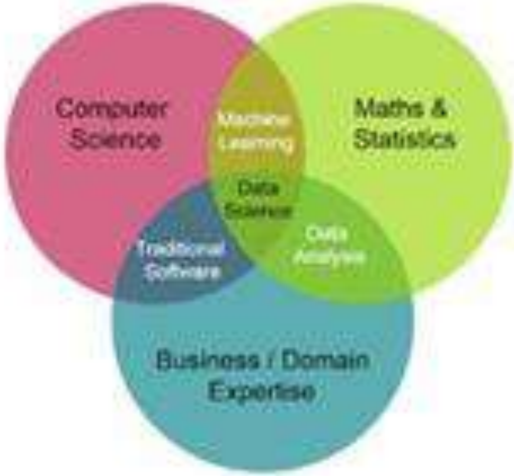
- ➢ Sold computer software

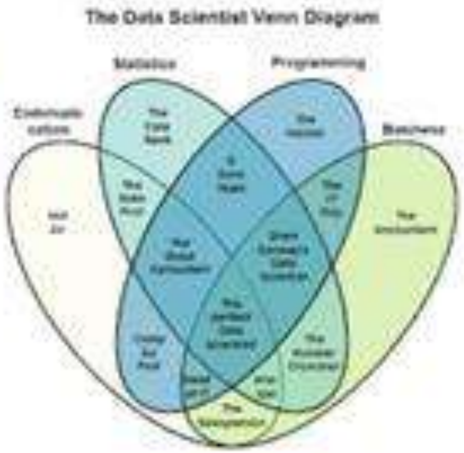
1980s Hardware Industry

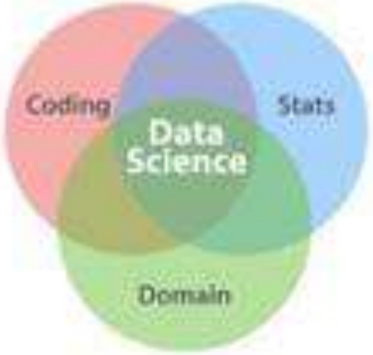
- ➢ Sold computers

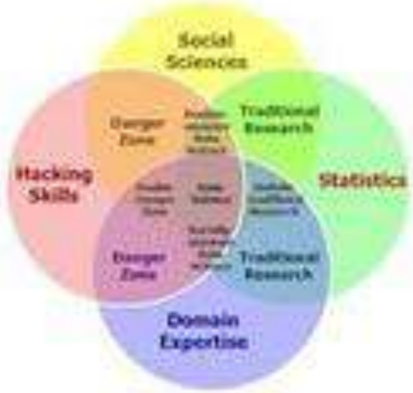
# Data Science: Multiple Domain

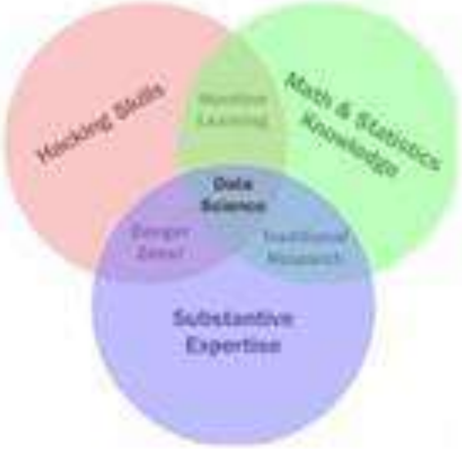
(2010 SIAM Article )

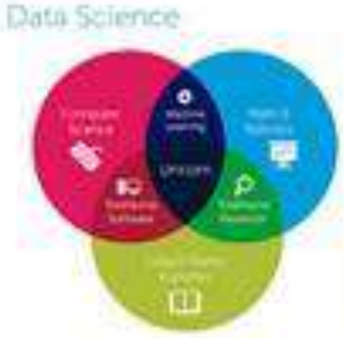


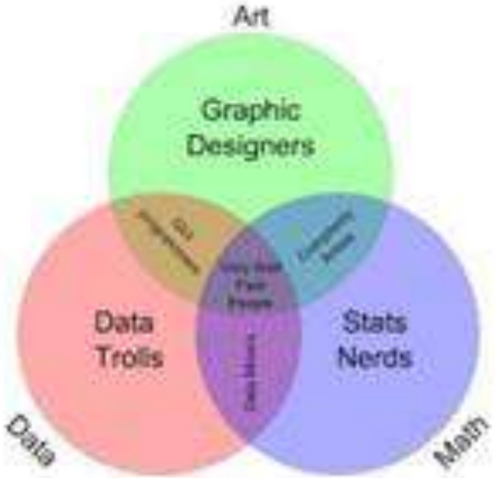




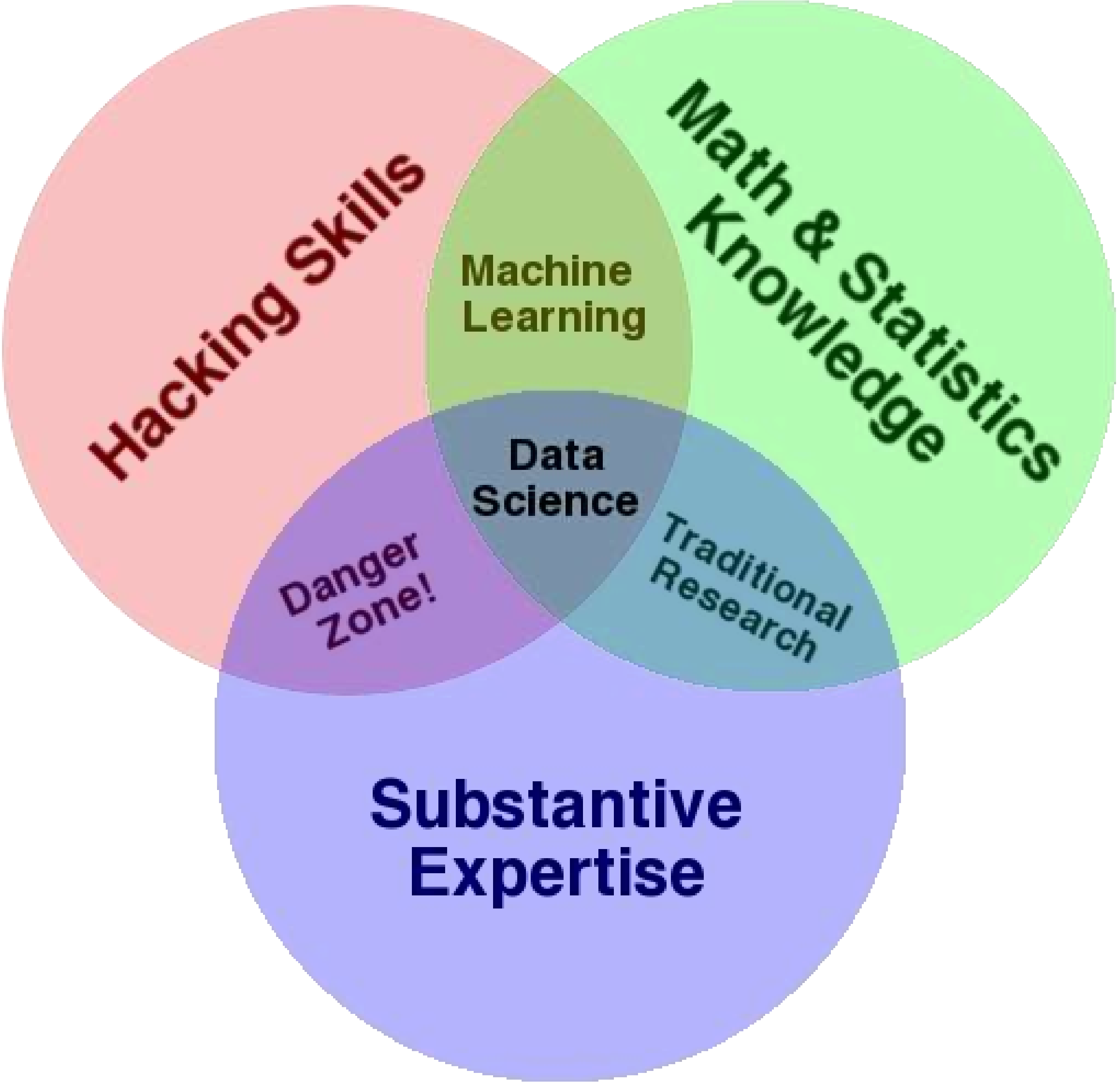








# Data Science : One Definition



# Why “Danger Zone?”

Ronny Kohavi* keynote at KDD 2015

- People are incredibly clever at explaining “very surprising results”. Unfortunately most very surprising results are caused by data pipeline errors.

- Beware “HiPPOs” (Highest Paid-Person’s Opinion)

* General Manager for Microsoft’s Analysis & Experimentation Team

# Succinct Definition of Data Science

The application of **data centric**, **computational**, and **inferential thinking** to

*understand the world* & *solve problems*

**Science** **Engineering**

- ➢ *Data science is fundamentally <u>interdisciplinary</u>*

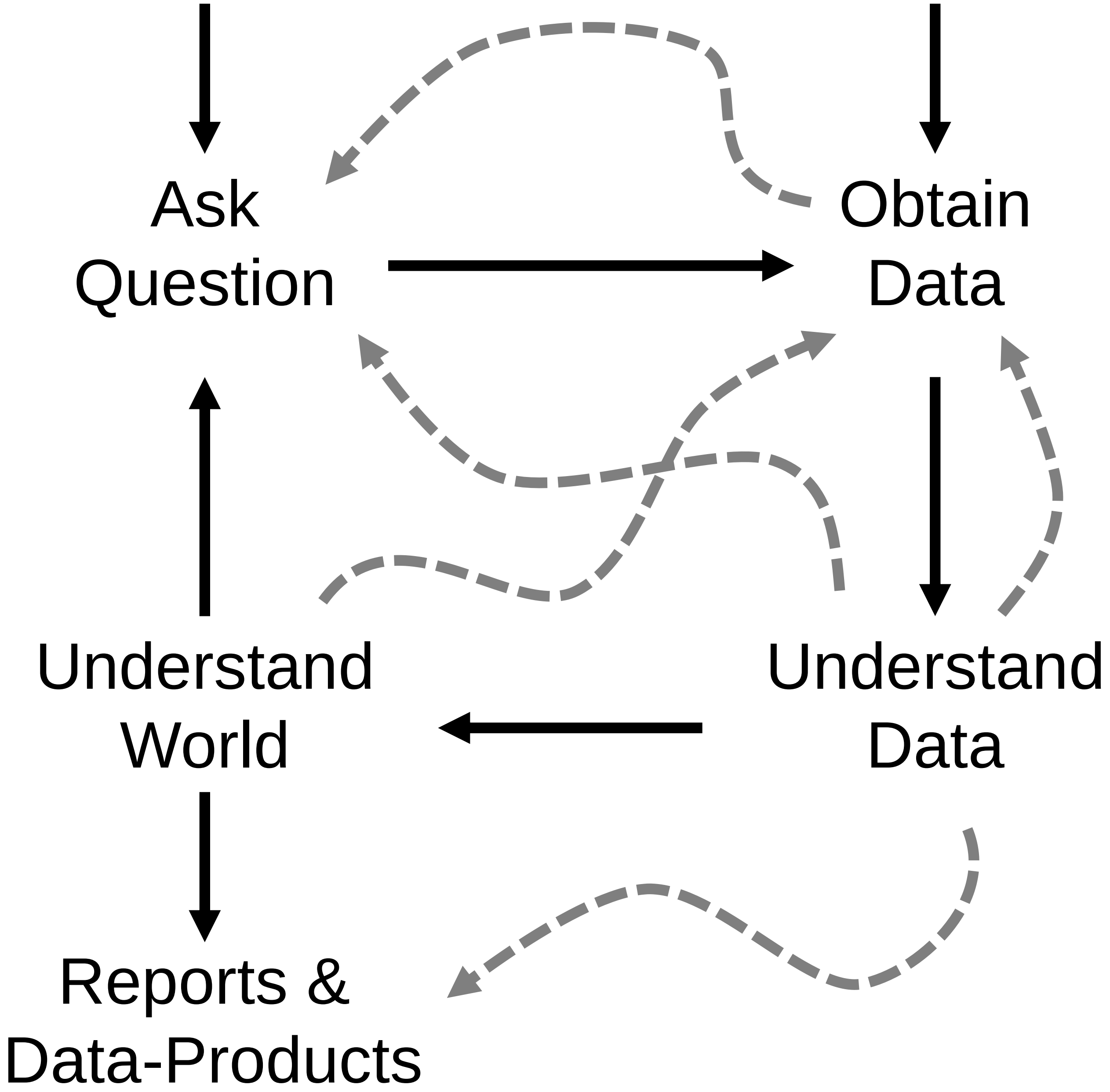
**Joseph E. Gonzalez** from CS.Berkeley

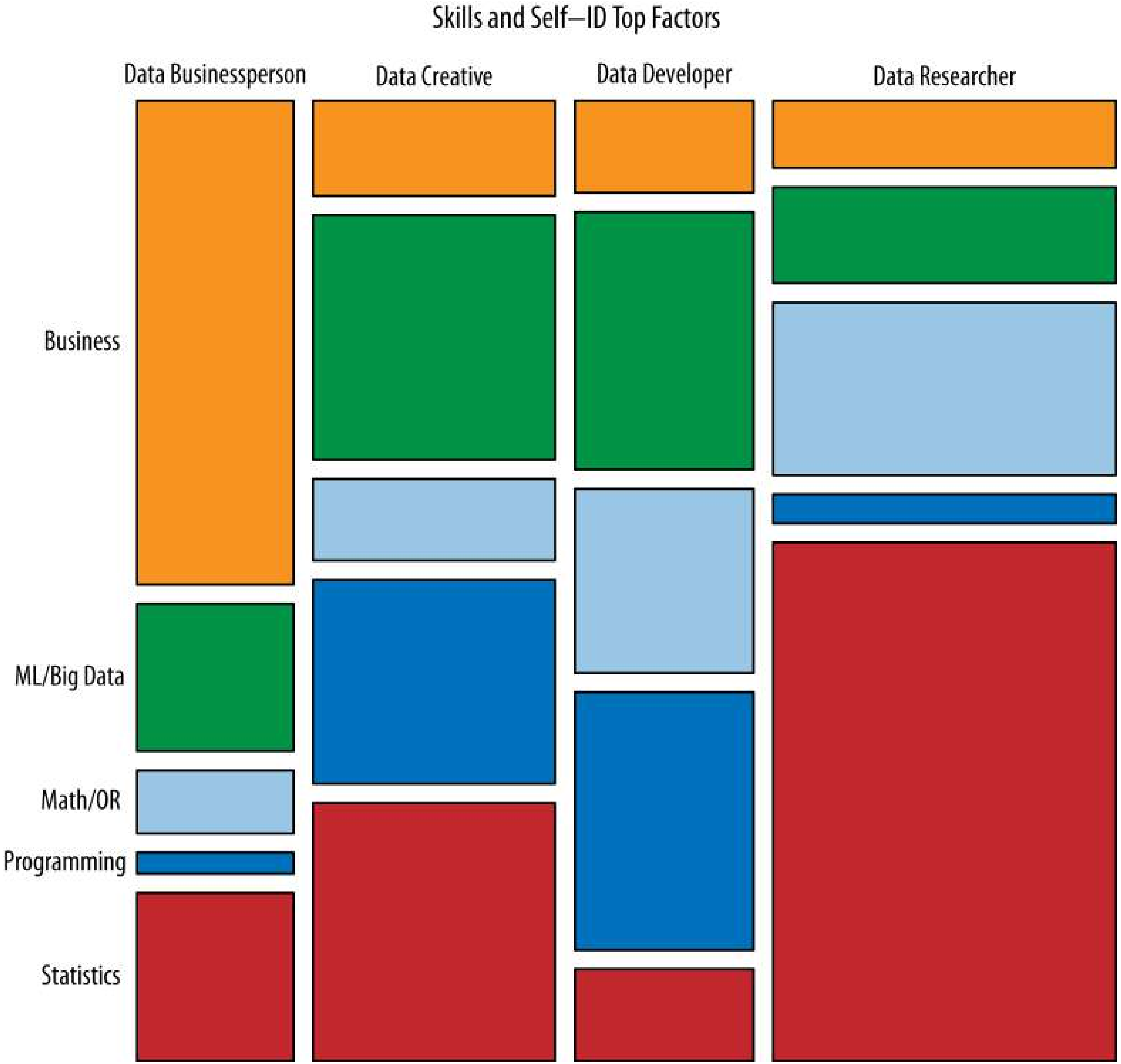
# Data Science Lifecycle

*High-level description of the data science workflow*

- Frame questions & design experiments
- Obtain and clean data
- Summarize and visualize data
- Inference and prediction

continuous process …





# Skill Patterns

- Different skill profiles
  - Business = Domain Knowledge.
  - Data Creative /Developer

# Contrast: Databases

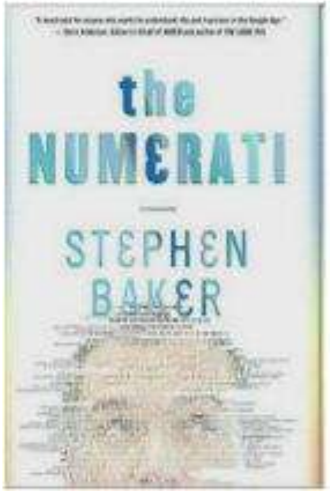
| | Databases | Data Science |
|---|---|---|
| Data Value | "Precious" | "Cheap" |
| Data Volume | Modest | Massive |
| Examples | Bank records,<br>Personnel records,<br>Census,<br>Medical records | Online clicks,<br>GPS logs,<br>Tweets,<br>Building sensor readings |
| Priorities | Consistency,<br>Error recovery,<br>Auditability | Speed,<br>Availability,<br>Query richness |
| Structured | Strongly (Schema) | Weakly or none (Text) |
| Properties | Transactions, ACID* | CAP* theorem (2/3),<br>eventual consistency |
| Realizations | SQL | NoSQL:<br>Riak, Memcached,<br>Apache River,<br>MongoDB, CouchDB,<br>Hbase, Cassandra,… |

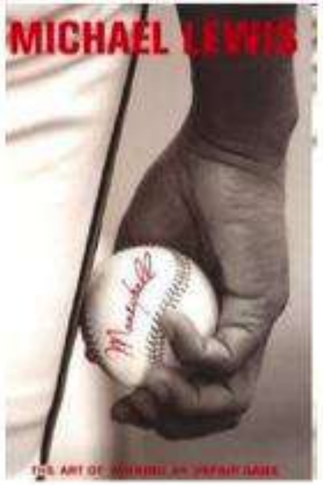
ACID = Atomicity, Consistency, Isolation and Durability

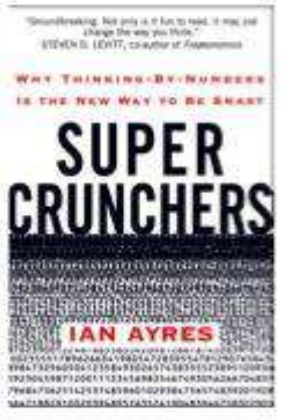
CAP = Consistency, Availability, Partition Tolerance

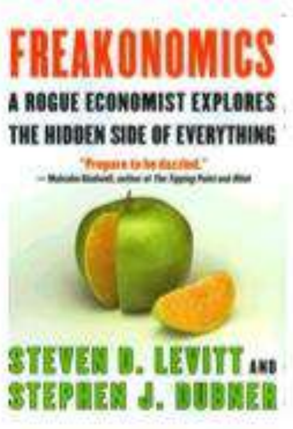
| Databases | Data Science |
|---|---|
| Querying the past | Querying the future |

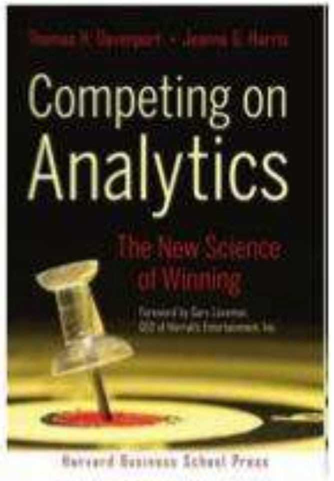
# Contrast: Databases








FREAKONOMICS
A ROGUE ECONOMIST EXPLORES THE HIDDEN SIDE OF EVERYTHING
STEVEN D. LEVITT AND STEPHEN J. DUBNER

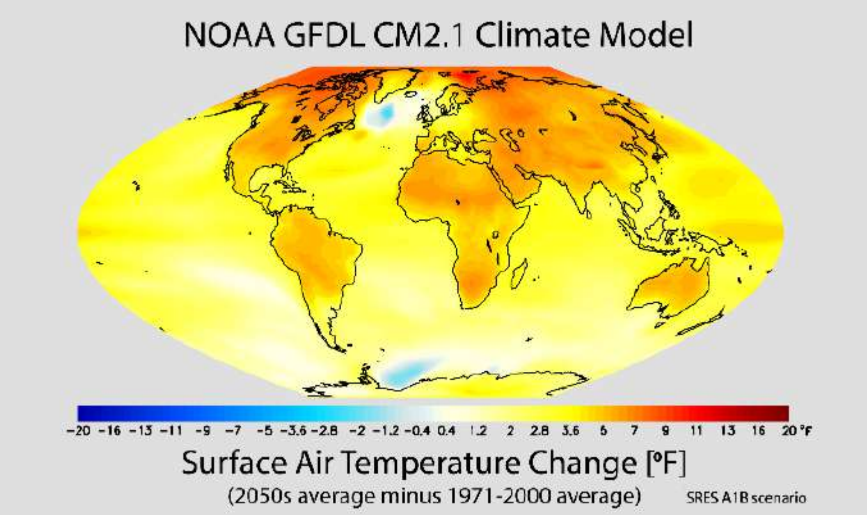

Thomas H. Davenport · Jeanne G. Harris
Competing on Analytics
The New Science of Winning
Harvard Business School Press

**Business intelligence** (**BI**) is the transformation of raw data into meaningful and useful information for business analysis purposes. BI can handle enormous amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities - Wikipedia

# Contrast: Machine Learning

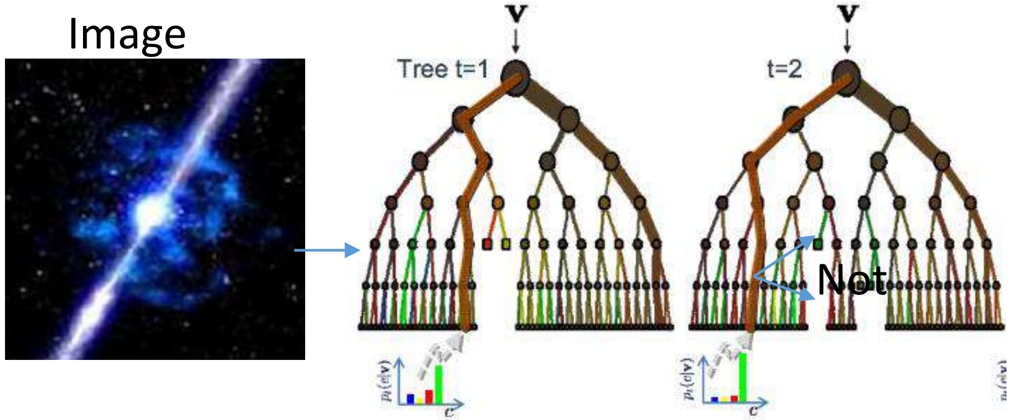
| Machine Learning | Data Science |
|---|---|
| Develop new (individual) models | Explore many models, build and tune hybrids |
| Prove mathematical properties of models | Understand empirical properties of models |
| Improve/validate on a few, relatively clean, small datasets | Develop/use tools that can handle massive datasets |
| Publish a paper | Take action! |

# Contrast: Scientific Computing


NOAA GFDL CM2.1 Climate Model
-20 -16 -13 -11 -9 -7 -5 -3.6 -2.8 -2 -1.2 -0.4 0.4 1.2 2 2.8 3.6 5 7 9 11 13 16 20 °F
Surface Air Temperature Change [°F]
(2050s average minus 1971-2000 average)
SRES A1B scenario

General purpose classifier

Image



Supernova

Nugent group / C3 LBL

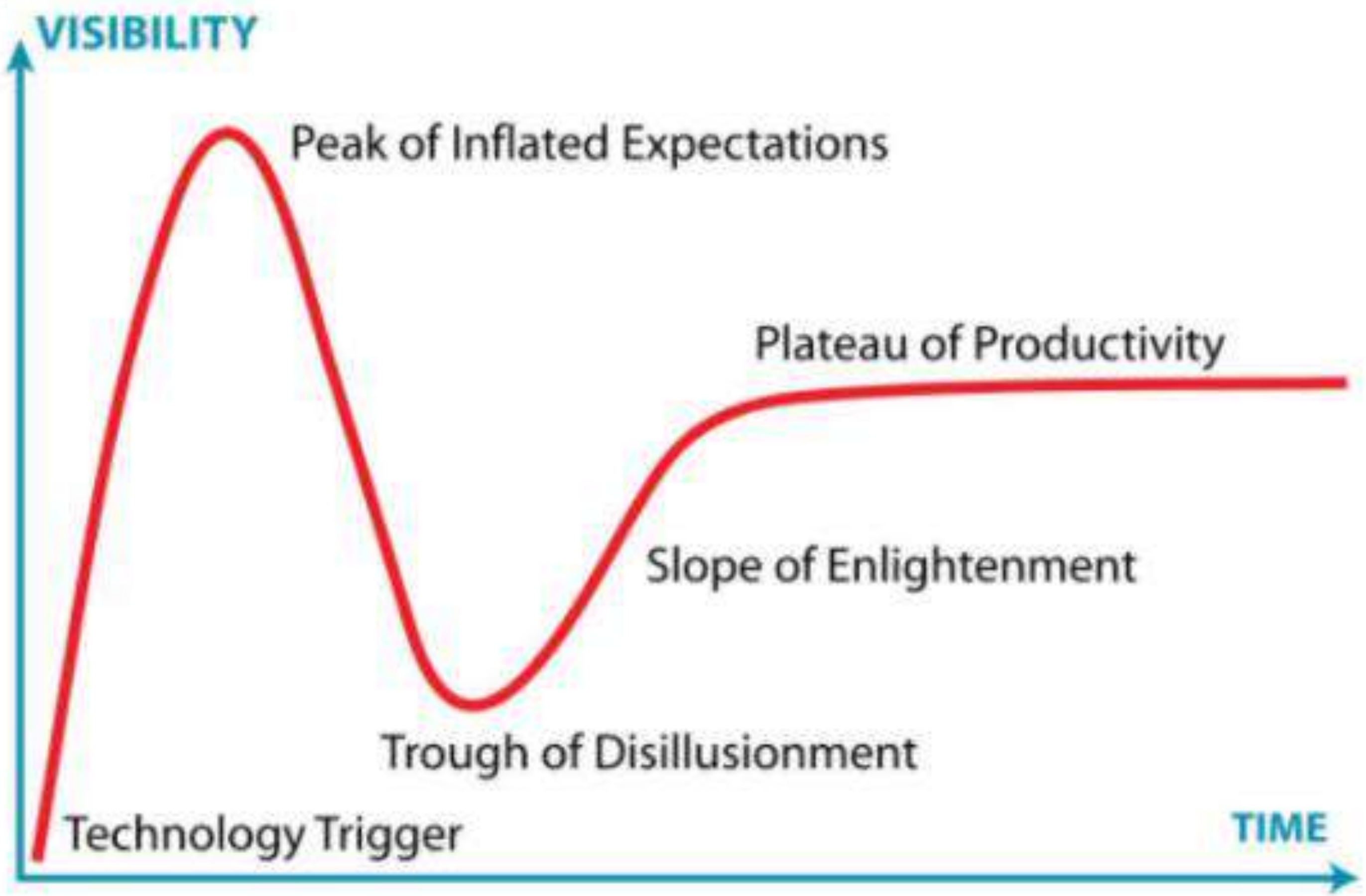
| Scientific Modeling | Data-Driven Approach |
|---|---|
| Physics-based models | General inference engine replaces model |
| Problem-Structured | Structure not related to problem |
| Mostly deterministic, precise | Statistical models handle true randomness, and **unmodeled complexity**. |
| Run on Supercomputer or High-end Computing Cluster | Run on cheaper computer Clusters (EC2) |

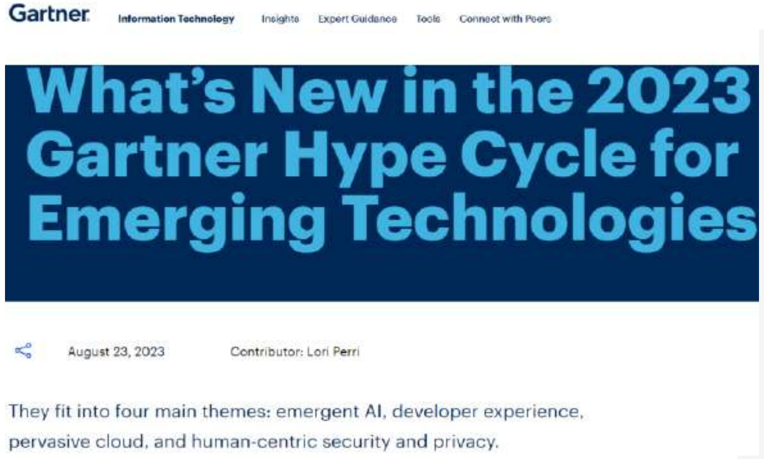
# Hype Cycle

Gartner

The **five** phases in the Hype Cycle are

1. **Technology Trigger**
2. *Peak of Inflated Expectations*
3. **Trough of Disillusionment**
4. *Slope of Enlightenment*
5. **Plateau of Productivity**





Gartner Information Technology Insights Expert Guidance Tools Connect with Peers

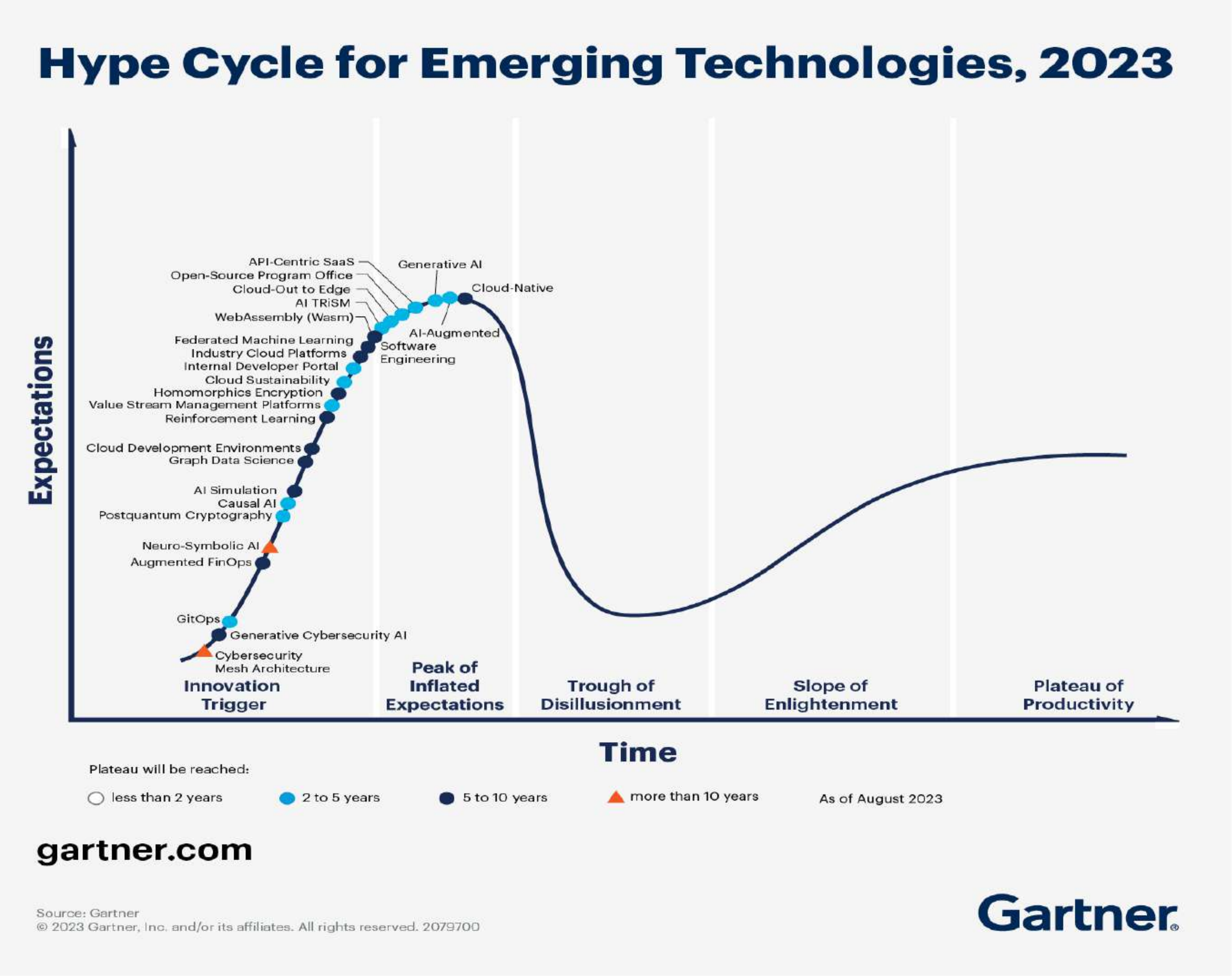
# What’s New in the 2023 Gartner Hype Cycle for Emerging Technologies

August 23, 2023 Contributor: Lori Perri

They fit into four main themes: emergent AI, developer experience, pervasive cloud, and human-centric security and privacy.

They fit into **4** themes:

- o Emergent AI
- o Developer experience
- o Pervasive clou
- o Human-centric security and privacy



gartner.com

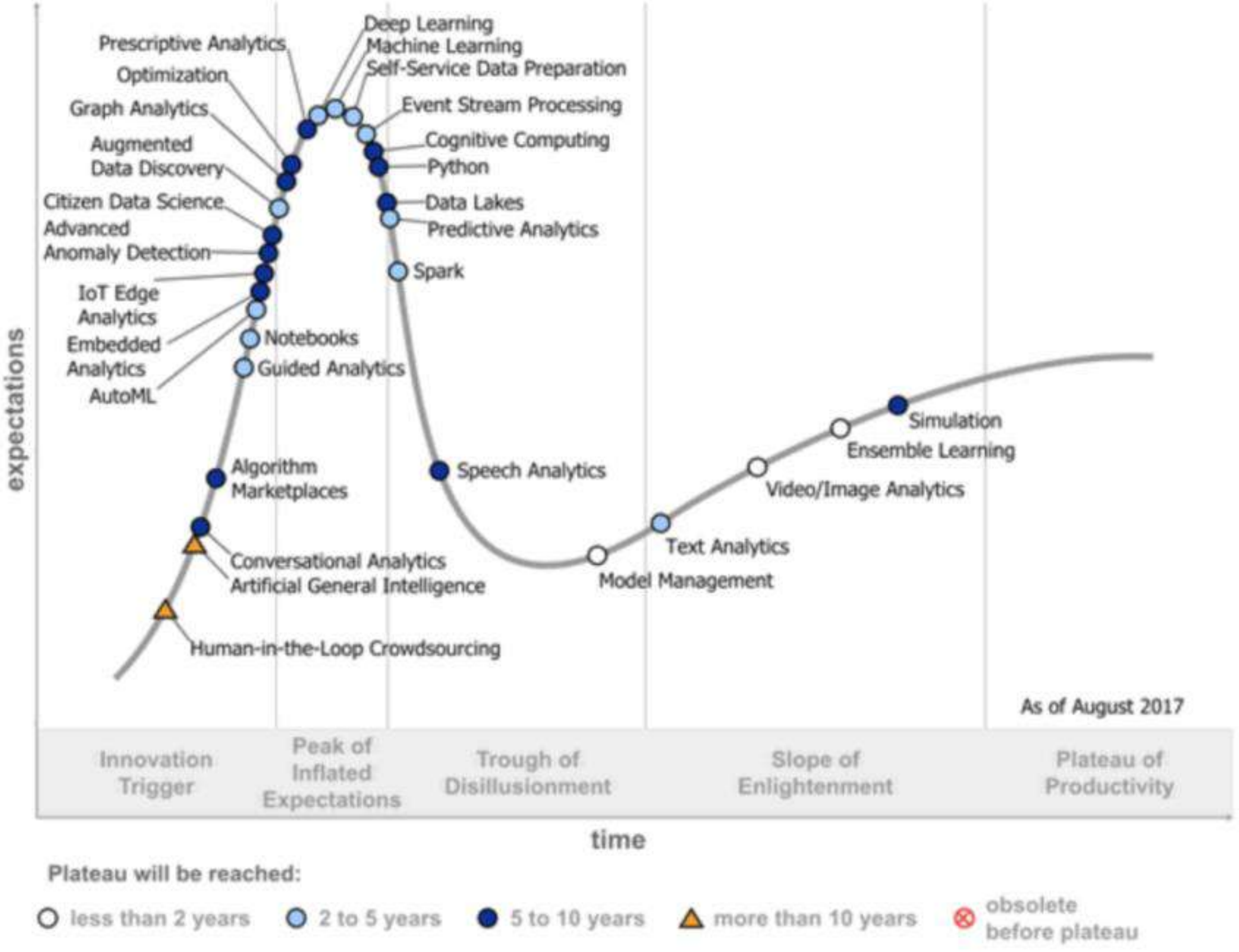
Source: Gartner

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. 2079700

Gartner

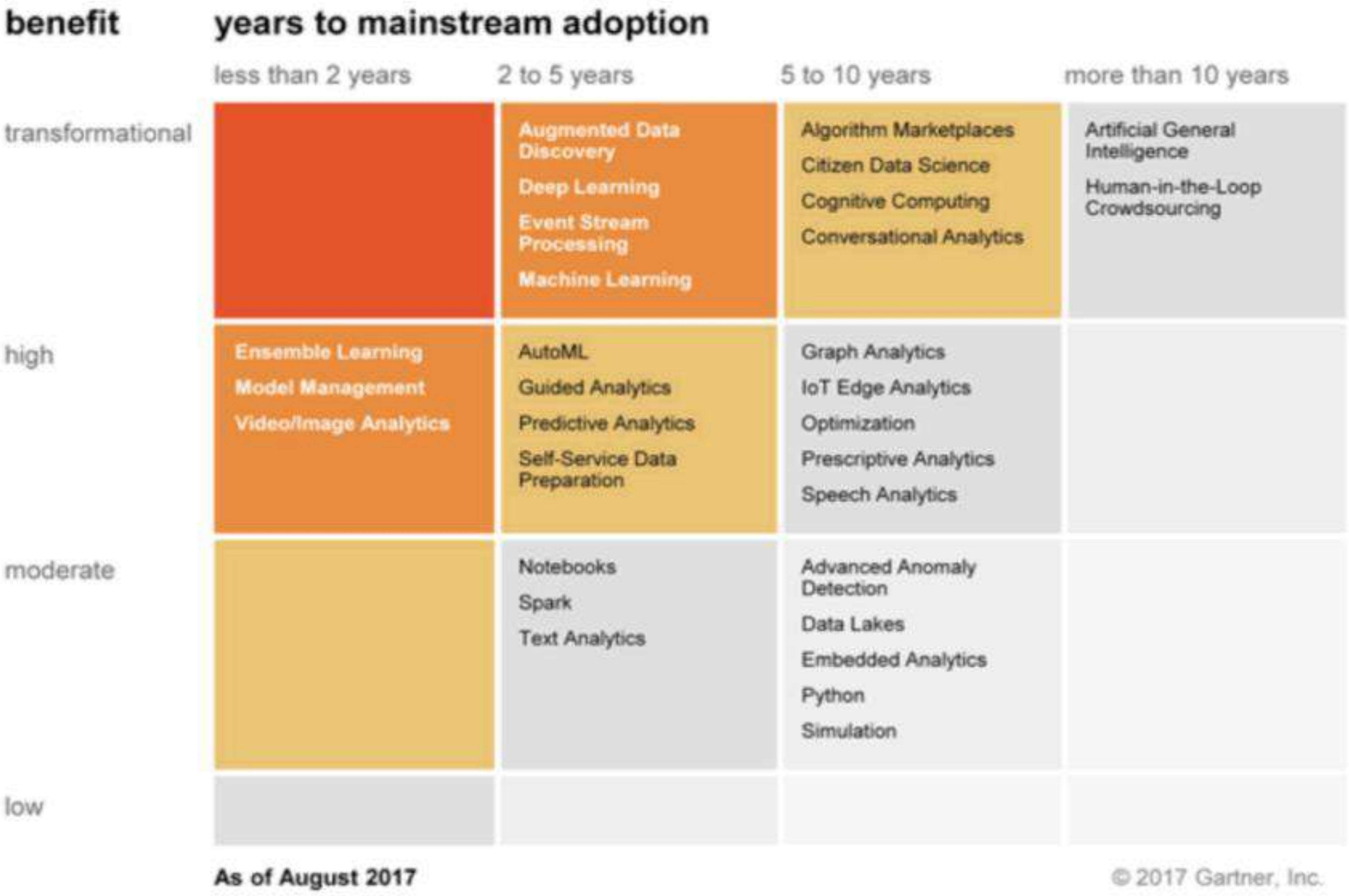
# Hype Cycle

# Data Science & Machine Learning



© 2017 Gartner, Inc.

# Data Science & Analytics

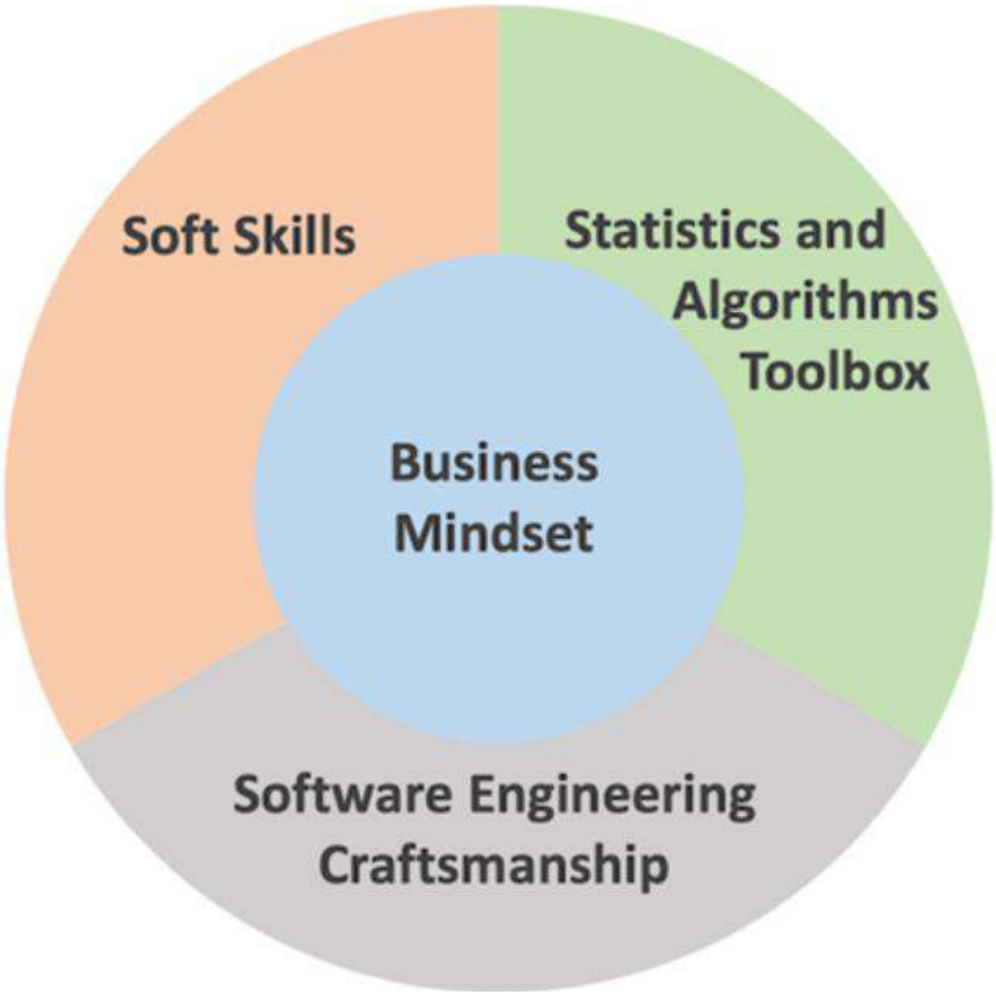


| benefit | years to mainstream adoption | | | |
|---|---|---|---|---|
| | less than 2 years | 2 to 5 years | 5 to 10 years | more than 10 years |
| transformational | | Augmented Data Discovery<br>Deep Learning<br>Event Stream Processing<br>Machine Learning | Algorithm Marketplaces<br>Citizen Data Science<br>Cognitive Computing<br>Conversational Analytics | Artificial General Intelligence<br>Human-in-the-Loop Crowdsourcing |
| high | Ensemble Learning<br>Model Management<br>Video/Image Analytics | AutoML<br>Guided Analytics<br>Predictive Analytics<br>Self-Service Data Preparation | Graph Analytics<br>IoT Edge Analytics<br>Optimization<br>Prescriptive Analytics<br>Speech Analytics | |
| moderate | | Notebooks<br>Spark<br>Text Analytics | Advanced Anomaly Detection<br>Data Lakes<br>Embedded Analytics<br>Python<br>Simulation | |
| low | | | | |

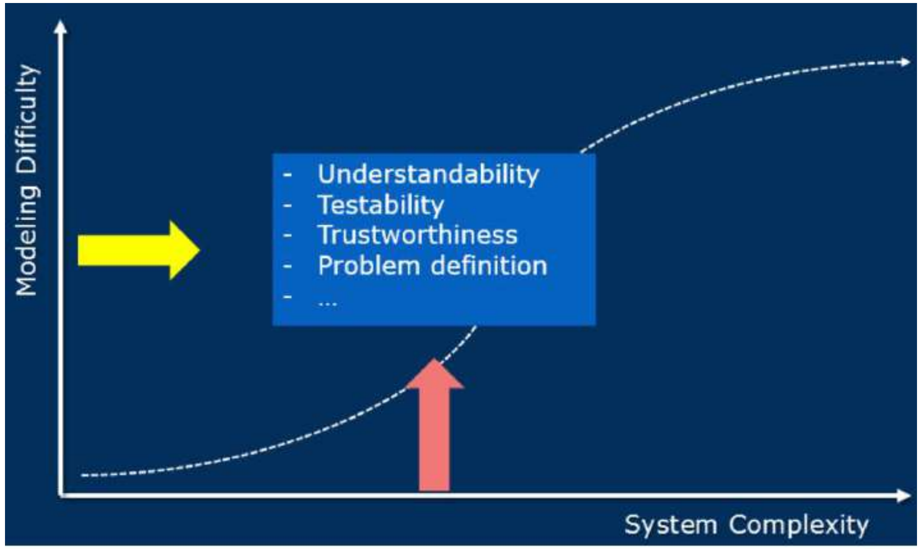
As of August 2017

© 2017 Gartner, Inc.

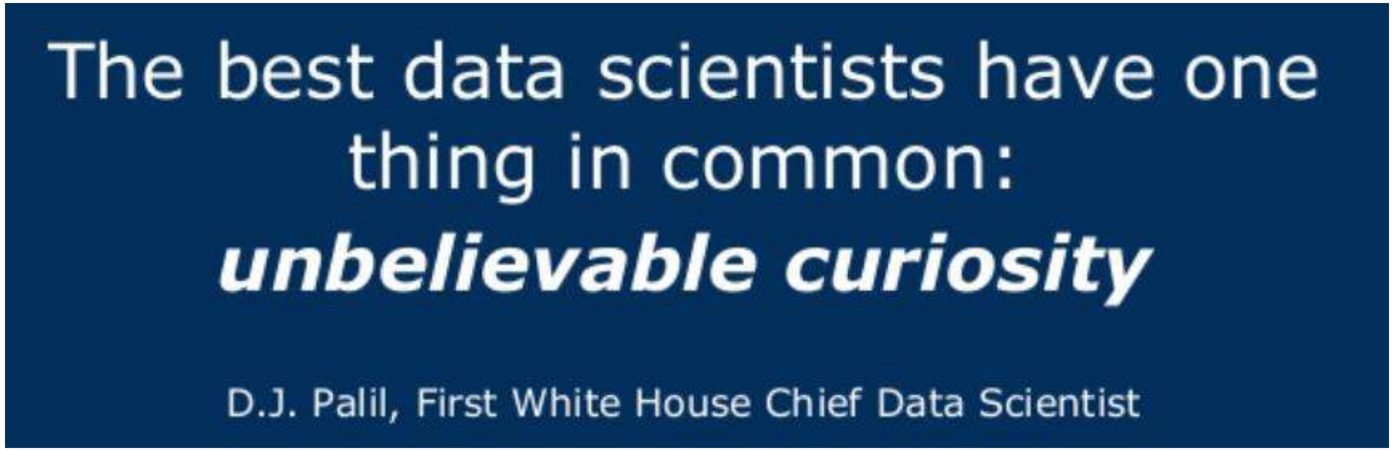
# Path to be a successful Data Scientist /Analyst


Soft Skills
Statistics and Algorithms Toolbox
Business Mindset
Software Engineering Craftsmanship

“Skill portfolio of the third wave data scientist.” Dominik Haitz



The data science landscape with the dimensions system complexity and modeling difficulty (cf. Ramanathan, 2016)



The best data scientists have one thing in common:
***unbelievable curiosity***

D.J. Palil, First White House Chief Data Scientist