TEXAS TECH UNIVERSITY

Introduction to Data Science, Machine Learning, and Decision-Making for Higher Education

2021 TAIR Annual Conference (Virtual)

Luciano Boas
Institutional Research
Texas Tech University

March 2nd, 2021

AGENDA

01 Data Science & Data Scientist

- History, Definitions etc.
- Big Data

02 Data Science in Higher Education

- Advanced Analytics
- Data Governance

03 Machine Learning

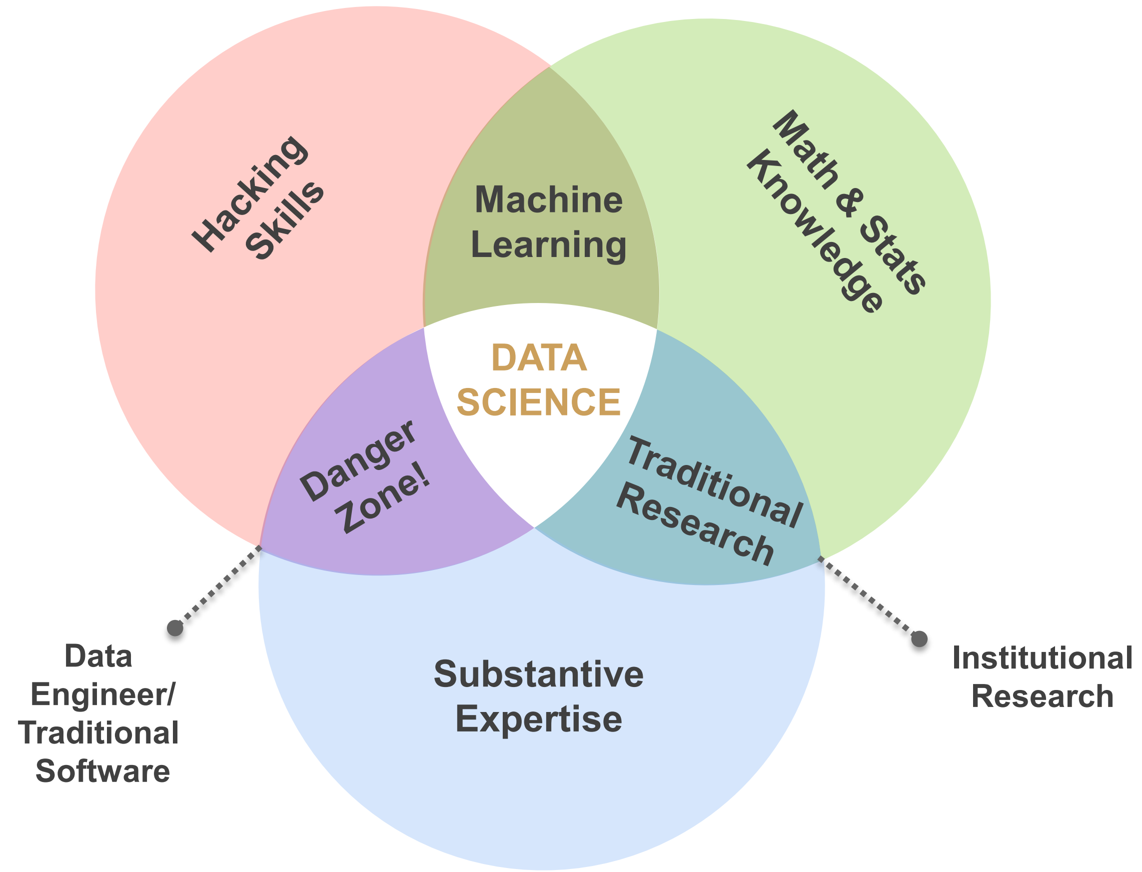
- Best Practices
- ML Pipeline Codes in R

04 Decision-Making

- Data-Driven Culture
- Basics of Decision-Making

WHAT IS DATA SCIENCE?

Venn Diagram created by Drew Conway (Data Scientist)



(Conway, 2010)

WHAT IS A DATA SCIENTIST?

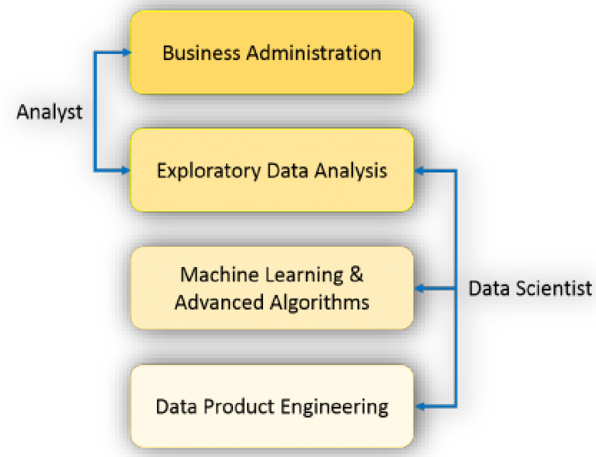
Five Traits of a Data Scientist: Scientist, Hacker, Quantitative Analyst, Trusted Advisor, Business Expert.
(Davenport, 2014)

How Does Data Science Differ from B.I.?

Features	Business Intelligence (BI)	Data Science
Data Sources	Structured (Usually SQL, often Data Warehouse)	Both Structured and Unstructured (logs, cloud data, SQL, NoSQL, text)
Approach	Statistics and Visualization	Statistics, Machine Learning, Graph Analysis, Neuro- linguistic Programming (NLP)
Focus	Past and Present	Present and Future
Tools	Pentaho, Microsoft BI, QlikView, R	RapidMiner, BigML, Weka, R

(Sharma, 2020)

Analyst vs. Data Scientist

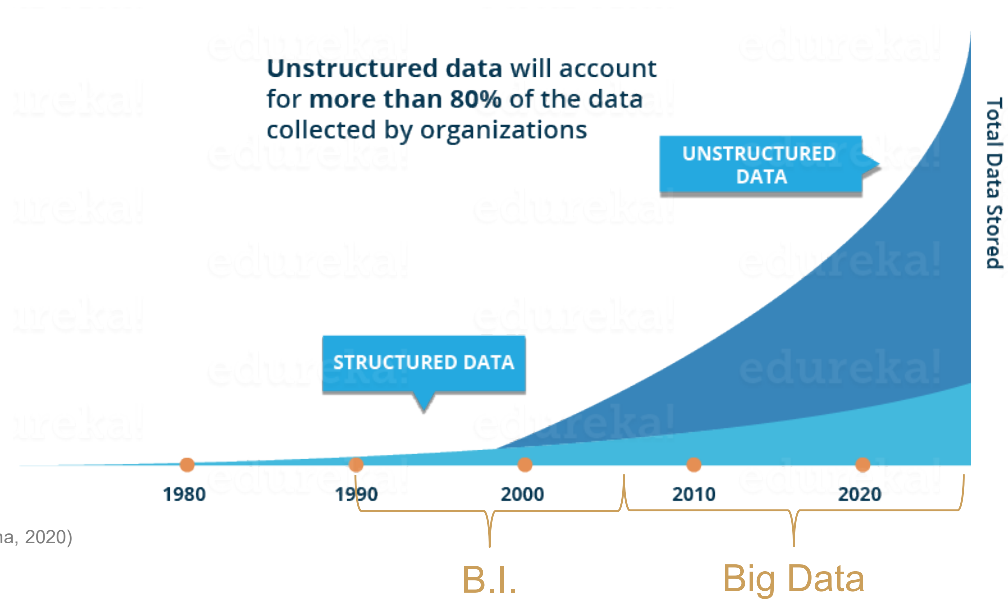


(Sharma, 2020)

FROM TABLES TO UNSTRUCTURED DATA

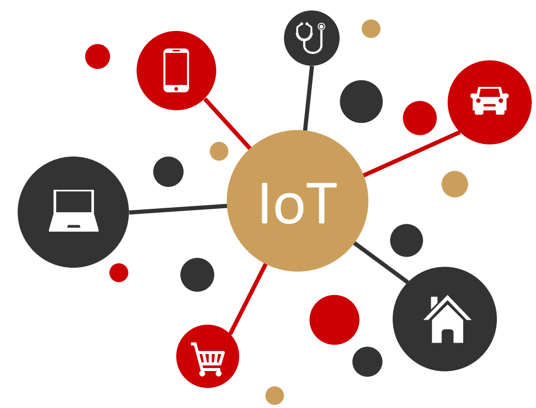
Unstructured Data Demands New Expertise to Mining Value out of Data.

Big Data Overtime



(Sharma, 2020)

Big Data Ecosystem: Internet of Things (IoT)



Sensor data is here to stay - connected to IoT, creating data 24/7.

AGENDA

01 **Data Science & Data Scientist**

- History, Definitions etc.
- Big Data

02 **Data Science in Higher Education**

- Advanced Analytics
- Data Governance

03 **Machine Learning**

- Best Practices
- ML Pipeline Codes in R

04 **Decision-Making**

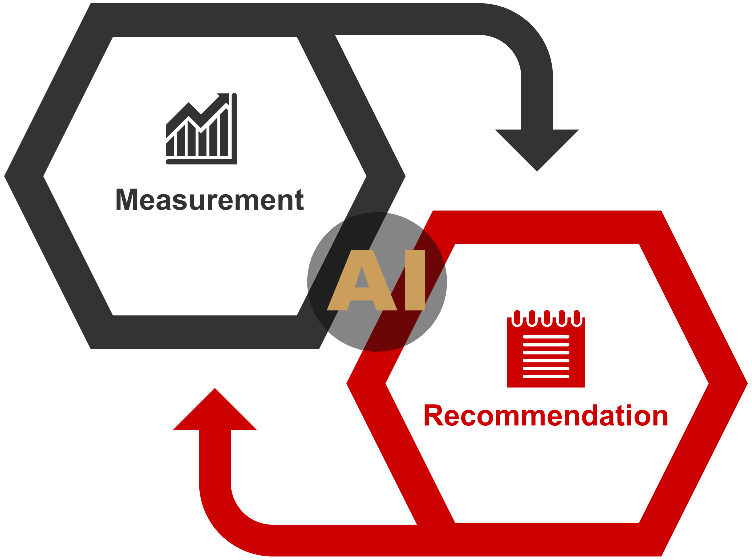
- Data-Driven Culture
- Basics of Decision-Making

Artificial Intelligence (AI) as an Autonomous Tutoring System

Some researchers predict that **Big Data for Education** will be a major focus in near future.

The creation of "**Auto Tutors**" – machines that are capable to measure student's performance and suggest recommendation pro-actively.

Not to be confused with E-learning.



Looking further, well, not really:

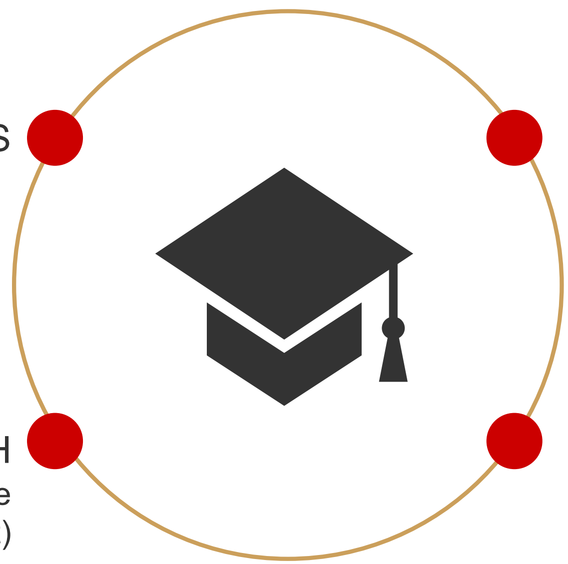
- AI Tutor Talking with Humans
  Source: http://ace.autotutor.org/IISAutotutor/index.html
- AI's Body Language Reader/Detector
  Source: https://www.scienceofpeople.com/body-language-ai/
- Behavioral Biometrics
  (Cowley, 2018)
- GPT-3 (Generative Pre-Trained Transformer 3)
  (Johns Hopkins University & Open AI, 2020)

DATA SCIENCE IN HIGHER EDUCATION

Partner with Internal and/or External Stakeholders – Medium to Long Term Strategic Projects

ADVANCED PREDICTIVE ANALYTICS

RESEARCH
(shed light on complex data and increase human insight)



DATA: PRIVACY, INTEGRITY, ACCURACY, ADVOCACY, DOCUMENTATION, ETC

DATA GOVERNANCE

DATA GOVERNANCE

Breaking the Silos | Integrating

ISSUES

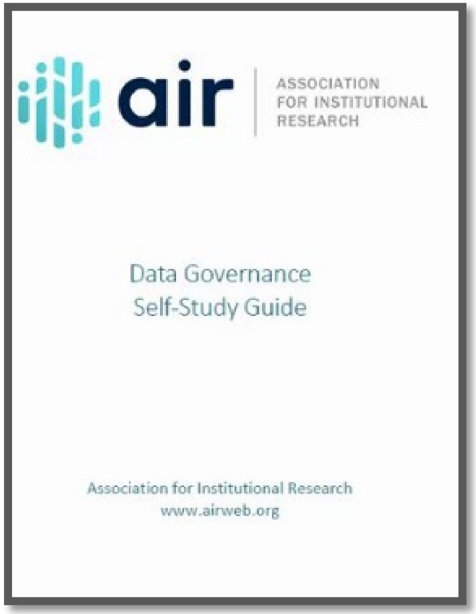
- Data collection from different departments – independent databases
- Silos and barriers can hold back school's performance
- **Fragmented "governance"**
- Not optimal use of resources

REALIZATION

- Successful Big Data initiatives will demand Data Governance
- Data Governance as a **competitive advantage**
- Implementation of "big data college management mode"

CHALLENGES

- **Scarce resources** on how to apply Data Science in Higher Ed.
- Few publications about usage of Big Data in Higher Ed.
- Small numbers of Data Governance examples in Higher Ed.
- Politics within the institution


air
ASSOCIATION FOR INSTITUTIONAL RESEARCH

Data Governance
Self-Study Guide

Association for Institutional Research
www.airweb.org

Source: AIR – Data Governance Self-Study Guide

RESOURCE

- AIR's Data Governance Self-Study Guide
- "The Data Governance Imperative" book by Steve Sarsfield

AGENDA

01 Data Science & Data Scientist

- History, Definitions etc.
- Big Data

02 Data Science in Higher Education

- Advanced Analytics
- Data Governance

03 Machine Learning

- Do we need it?
- ML Pipeline Codes in R

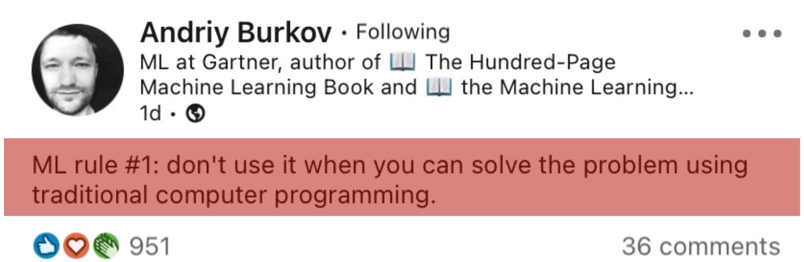
04 Decision-Making

- Data-Driven Culture
- Basics of Decision-Making

DO I NEED MACHINE LEARNING TO SOLVE THIS PROBLEM?

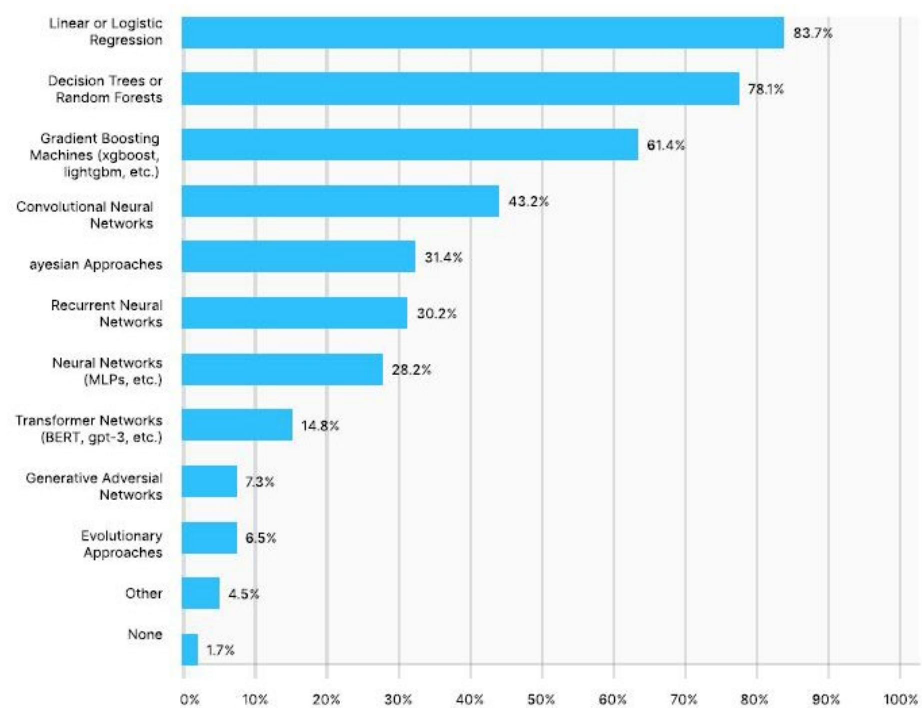
"Machine Learning learns patterns from historic data (training) and tries to generalize that in unseen (new) data."
Mathangi Sri

- Can a Machine Learning approach be undertaken to predict **student retention**?
- Does the amount of **effort** to develop an ML pipeline is justified by the **value** that it can potentially bring?
- How well an ML model performs in comparison with a **standard statistical** approach?
- Does the **explainability** of the ML model allow for stakeholders to understand and trust the model?
- Are **decision-makers** willing to invest in the implementation of an ML project?



Source: Andriy Burkov on LinkedIn

Methods and Algorithms Usage



Source: Kaggle – State of Machine Learning and Data Science 2020

MACHINE LEARNING CODES IN R

towards data science | DATA SCIENCE | MACHINE LEARNING | PROGRAMMING | VISUALIZATION | VIDEO | ★ | ABOUT | CONTRIBUTE

Data Science

Crafting a Machine Learning Model to Predict Student Retention Using R

Student Retention is one of the most important indicators in Higher Education. Therefore, Predictive Analytics plays a crucial role in that

Luciano Vilas Boas
Jul 29 · 19 min read ★

(Vilas Boas, 2020)

Dataset was very challenging (very small, and variables not really statistically significant).

The article was featured in the First page of the “Data Science” section.

Scan this QR Code
to quickly access the article:

AGENDA

01 **Data Science & Data Scientist**

- History, Definitions etc.
- Big Data

02 **Data Science in Higher Education**

- Advanced Analytics
- Data Governance

03 **Machine Learning**

- Best Practices
- ML Pipeline Codes in R

04 **Decision-Making**

- Data-Driven Culture
- Basics of Decision-Making

HUMANS vs MACHINES IN DECISION-MAKING

Hybrid Approach: Not Us And Them

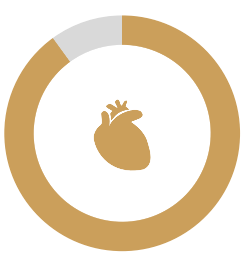
HUMANS

Decisions are made based on “**gut feelings**”, emotions, and experience.

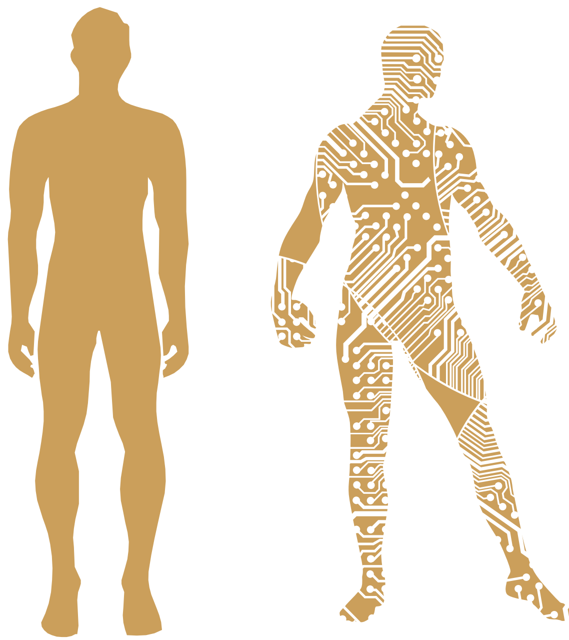
“humans, unlike computer models, have the ability to **recognize** when something isn't quite right.“

(Heingartner, 2006)

90% Emotions



(Natarelli, 2017)



MACHINES

Unemotional decisions – accuracy and consistency.

“mathematical models generally make more **accurate** predictions than humans do.”

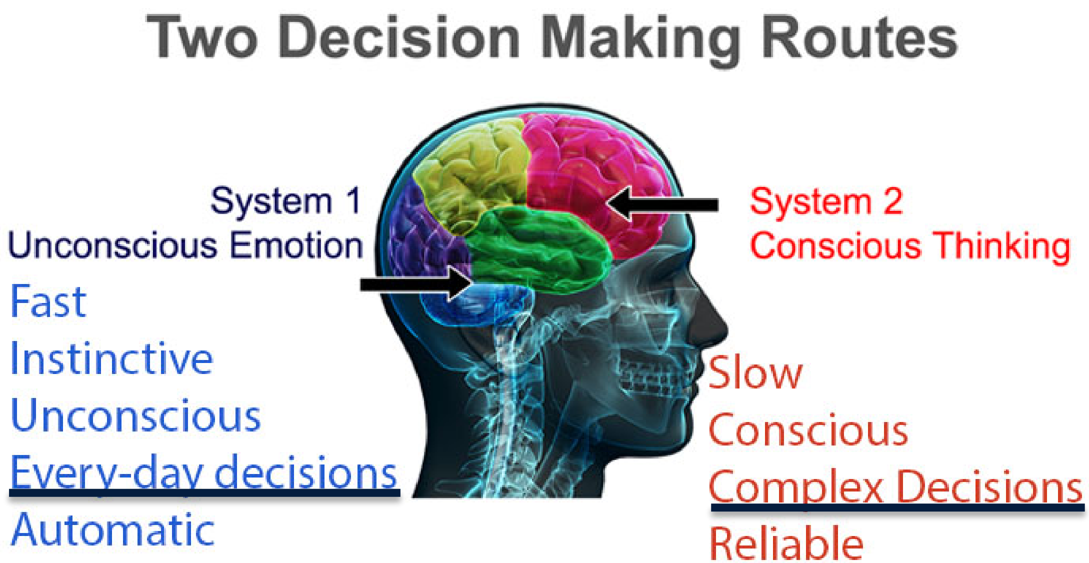
(Heingartner, 2006)

100% Unemotional

SYSTEM 1 AND SYSTEM 2 OF THINKING

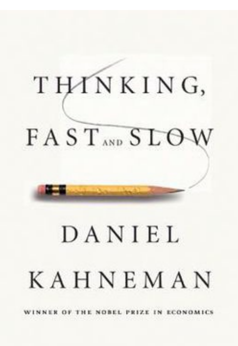
System 1 it's not Optimal When Making Complex Decisions

No effort
2 + 2 = ?



(B2B Web Marketing Hub, 2018)

Some effort
23 x 17 = ?



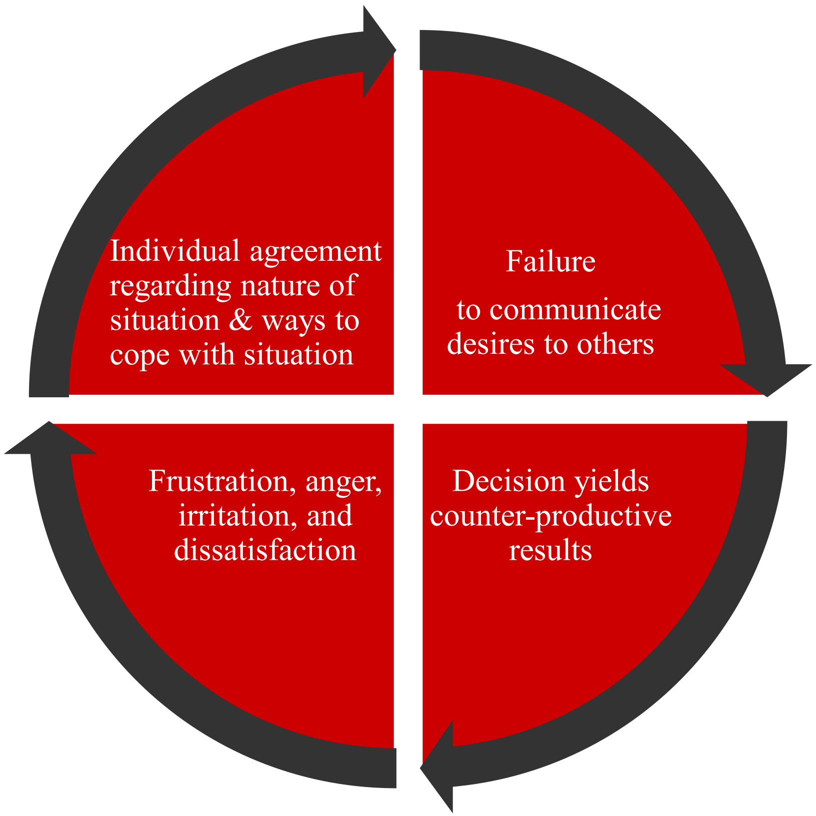
(Kahneman, D., 2011)

"System 1 runs ahead of the facts in constructing a rich image on the basis of **scraps of evidence**." – D. Kahneman

We may end up making a "good enough" decision (sub-optimal).

THE ABILENE PARADOX: THE MANAGEMENT OF AGREEMENT

"The first rule of decision-making is that one does not make a decision unless there is disagreement"
Peter Drucker



(Ahokiecares, 2016)

The 5 Dysfunctions of a Team:

(Lencioni, 2002)

Setting up to fail: "75% of respondents admit that their projects are either always or usually "**doomed right from the start**."
(Geneca, 2017)

REVERSE DECISION

Sunk Cost Fallacy

"***Sunk Costs*** are investments of time, energy and resources that **can't be recovered** once they're made. Continuing to invest in a project to recoup lost resources doesn't make sense - throwing "good money after bad" is not a winning strategy." – Josh Kaufman, The Personal MBA: Master the Art of Business.

Reverse a decision based on the most current information and not what was known at the time of the (first) decision.

Change, change, change the plan.
Be flexible. It's still respectable.

Ernie (Sesame Street)

THREE KEYS FOR MAKING GOOD DECISIONS

"The fact is, decision making is not an event. It's a process, one that unfolds over, weeks, months, or even years"
(Garvin and Roberto, p.1)

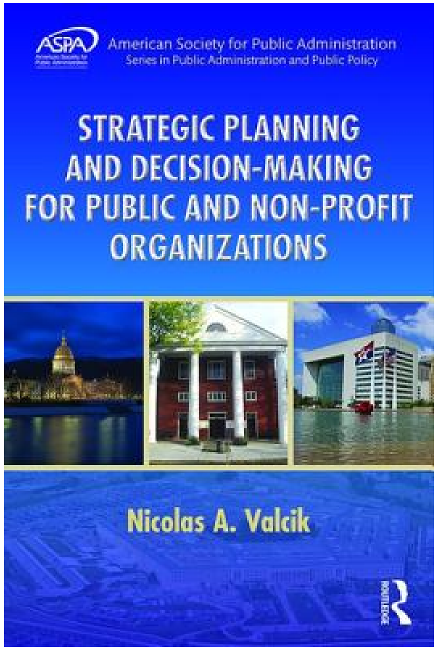
TRAINING — Become **aware** of how you make decisions, biases you may have, etc
Take advantage of decision tools, processes, and methods.

PREPARATION — Use **data**, gather feedback, research, benchmark, accumulate experience, choose right methods, etc

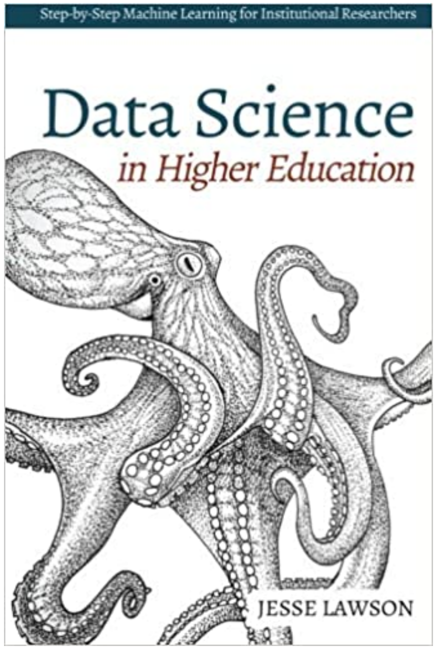
PRACTICE — Gain expertise and skill and iterate with the decision-making processing.

Still, good decisions can have bad outcomes

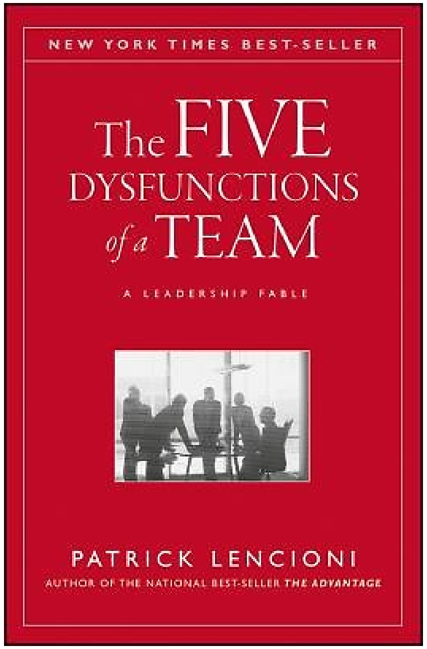
RECOMMENDED BOOKS



(Valcik, 2016)



(Lawson, 2015)



(Lencioni, 2002)

REFERENCES

Conway, D. (2010, September 30). The data science venn diagram. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
Thomas H. Davenport. 2014. Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. *Harvard Business Review Press*, Boston, MA, USA.
Sharma, H. (2020, November 25). What is data science? A beginner's guide to data science. *Edureka!* https://www.edureka.co/blog/what-is-data-science/
Science of People. How AI will revolutionize how we use body language. *Science of People. https://www.scienceofpeople.com/body-language-ai/*
Cowley, S. (2018, August 13). Banks and retailers are tracking how you type, swipe and tap. *The New York Times. https://www.nytimes.com/2018/08/13/business/behavioral-biometrics-banks-security.html*
John Hopkins University & Open AI. (2020, May 28). Language models are few-shot learners. *Cornell University. https://arxiv.org/abs/2005.14165*
Kaggle (2020). State of machine learning and data science 2020. *Kaggle.* https://www.kaggle.com/kaggle-survey-2020
Brownlee, J. (2016, February 13). How to evaluate machine learning algorithms with R. *Machine Learning Mastery.* https://machinelearningmastery.com/evaluate-machine-learning-algorithms-with-r/
Amidi, A. & Amidi, S. (n.d.). Machine learning tips and tricks cheatsheet. *Stanford.* https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-machine-learning-tips-and-tricks
Vilas Boas, L. (2020, July 29). Crafting a machine learning model to predict student retention using R. *Towards Data Science.* https://towardsdatascience.com/crafting-a-machine-learning-model-to-predict-student-retention-using-r-5eb009dcb1ec
Ambraz, P. (2017, December 29). How to be data-driven and act on your analytics vision. *IBM.* https://www.ibm.com/blogs/business-analytics/data-driven-analytics-vision/
Heingartner, D. (2006, July 16). Trust an algorithm with your business? *The New York Times.* https://www.nytimes.com/2006/07/18/technology/18iht-model.2233108.html
Natarelli, M. (2017, November 14). How emotion drives brand choices and decisions. *Branding Strategy Insider. https://www.brandingstrategyinsider.com/how-emotion-drives-brand-choices-and-decisions/#.YAnw8uhKiUI*
B2B Web Marketing Hub (2018, March 27). B2B decision making process: the role of emotions. https://www.the-b2b-webmarketing-hub.com/index.php/2018/03/27/b2b-decision-making-process-the-role-of-emotions/
Kahneman, D. (2011). Thinking, fast and slow. New York: Farrar, Straus and Giroux.
Ahokiecares (2016, February 17). My trip to Abilene. https://ahokiecares.wordpress.com/2016/02/17/my-trip-to-abilene/
Harvey, J. B. (1974). The Abilene paradox: the management of agreement. *Organizational Dynamics*, 3(1), 63-80. https://www.sciencedirect.com/science/article/pii/0090261674900059?via%3Dihub
Geneca (2017, January 25). Why up to 75% of software projects will fail. https://www.geneca.com/why-up-to-75-of-software-projects-will-fail/
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263-291.
Harley, A. (2016, June 19). Prospect theory and loss aversion: how users make decisions. *Nielsen Norman Group. https://www.nngroup.com/articles/prospect-theory/*
Valcik, Nicolas. (2016). Strategic Planning and Decision-Making for Public and Non-Profit Organizations. 10.4324/9781315640686.
Lawson, Jesse. (2015). Data Science in Higher Education: A Step-by-Step Introduction to Machine Learning for Institutional Researchers.
Lencioni, Patrick. (2002). The Five Dysfunctions of a Team: A Leadership Fable. Jossey-Bass.

TEXAS TECH UNIVERSITY®

From here, it's possible.