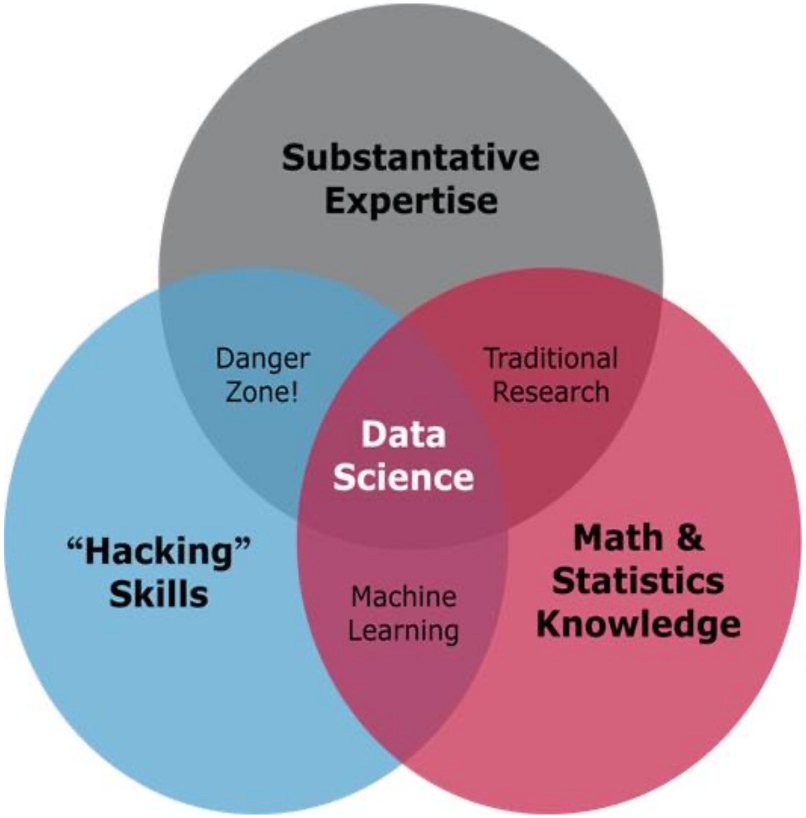
MIS 382N: Advanced Machine Learning

Prof. Joydeep Ghosh
ECE/UT
www.ideal.ece.utexas.edu/~ghosh
jghosh@utexas.edu

Data “Science”



- *Business Problem → Data Science sub-problems*
- *Additional AI modalities*
- *Enterprise Delivery Platform*
    - *(Software: Orchestration, monitoring.., e.g. Google Vertex AI)* *(MLOps)*
- *U/I and U/X: human in the loop*

https://cyborgus.com/2017/03/13/think-like-data-scientist/

Data Driven Modeling Approaches and Goals

- Types of Analytics:
  - 1. **Descriptive:** Find human-interpretable patterns that describe the data.
    - Provide large scale summary of data
      - e.g. characterize dominant customer types
    - Seek (local) patterns
      - Characterise a small portion of data, e.g. "rare patterns": fraud or intrusion detection
  - **2. Predictive:** Use some variables to predict unknown or future values of other variables.
    - **Regression**: predicting shelf life based on other attributes….
    - **Classification:** predicting what type of fruit is it? (class hierarchy!)
    - **Ranking and Recommendations**
- …..
  - 3. Prescriptive: (may need causal reasoning, domain expertise,…)
    - Reinforcement learning
    - SEM and other Causal models

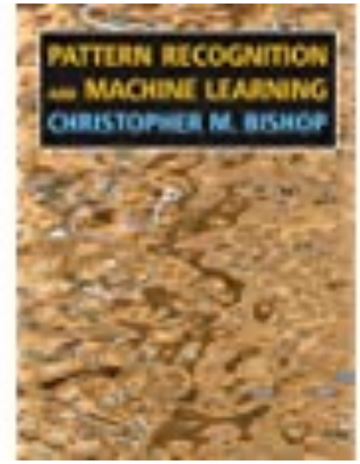
Analysis is often **retrospective (and not "prospective): data was not collected in a methodical way that is tailored for the analytical task.**

Course Scope and Sequencing

- Summer: Broad Intro to ML + Python
- Fall
  - APM (Advanced Predictive Modeling) → AML
  - Domain Courses
- Spring
  - Descriptive Modeling, Time Series
  - Deep Learning (new, optional)
  - Capstone

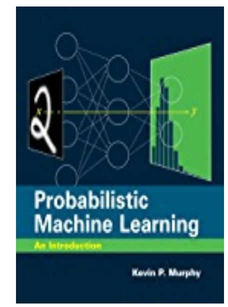
Texts

- **Main Text (only a few chapters; provided for you via canvas).**
  - **CB**: *Chris Bishop*, **Pattern Recognition and Machine Learnir (more mathematical, Bayesian)**

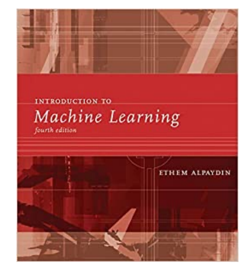


Supplementary

  - **KM**: *Kevin Murphy*, **Probabilistic Machine Learning: An Introduction**, MIT Press, March 2022. (Draft pdf file, 2022-05-09)
    - Code to recreate all the figures can be found in a series of colabs, one per chapter, stored here.



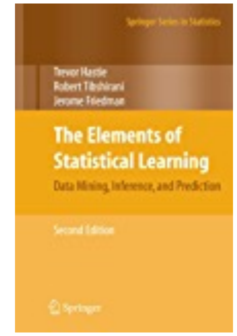
  - **EA**: *E. Alpaydin*, Introduction to Machine Learning, (4th Ed, 2020), MIT Press.



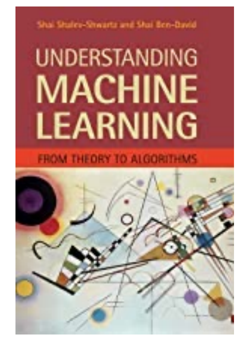
  - **AG**: *A. Geron*, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, O'Reilly, 2019

Other References

- (Basic) **JW:** **ISLR:** Intro to stats learning with R
  (Advanced) **HTF:** Hastie/Tibshirani/Friedman **(stats)**
  http://www-stat.stanford.edu/~tibs/ElemStatLearn/

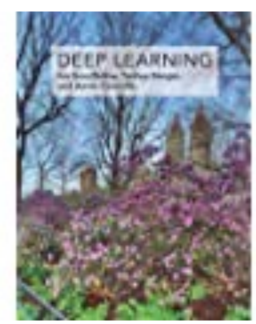


  - (Advanced): Understanding Machine Learning: From Theory to Algorithms, by Shai Ben-David and Shai Shalev-Shwartz (2014), Cambridge.



Deep Learning:

- Diving into Deep Learning, (online) Aston Zhang, Zack Lipton, Mu Li and Alex Smola (2019).
- Deep Learning, Ian Goodfellow, Yoshua Bengio and Aaron Courville (2016), MIT Press.



Disclaimer

My slides are not comprehensive or self-contained (by design)

visual aids

keep track of key ideas and their ordering

Add:

+ issues discussed in class

+ readings indicated in the slides

ML Lifecycle

From: http://www.mlebook.com/wiki/doku.php. (2020)

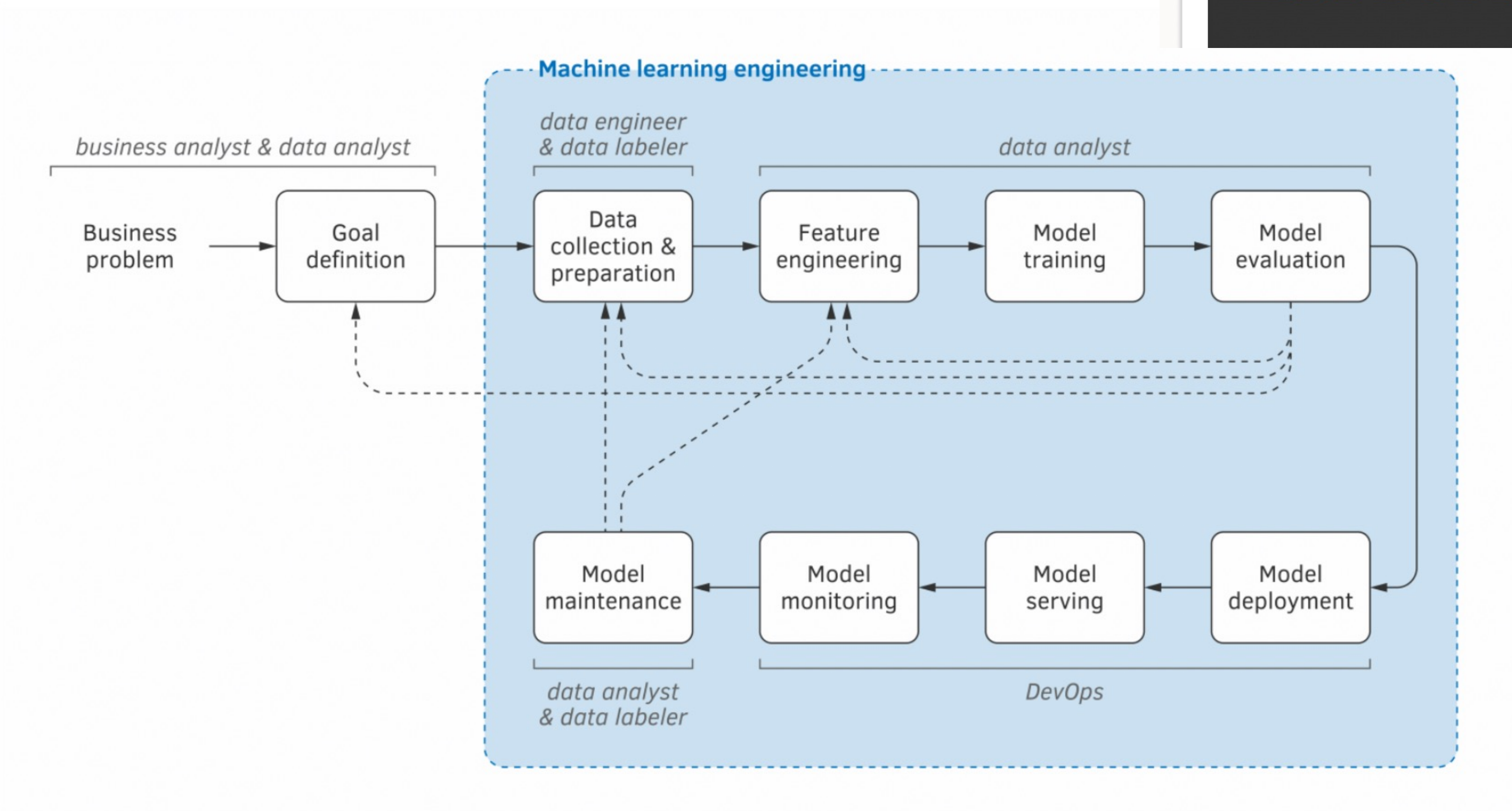


Figure 4: Machine learning project life cycle.

Also see AI-infrastructure.org work towards a Canonical ML Stack (2022)

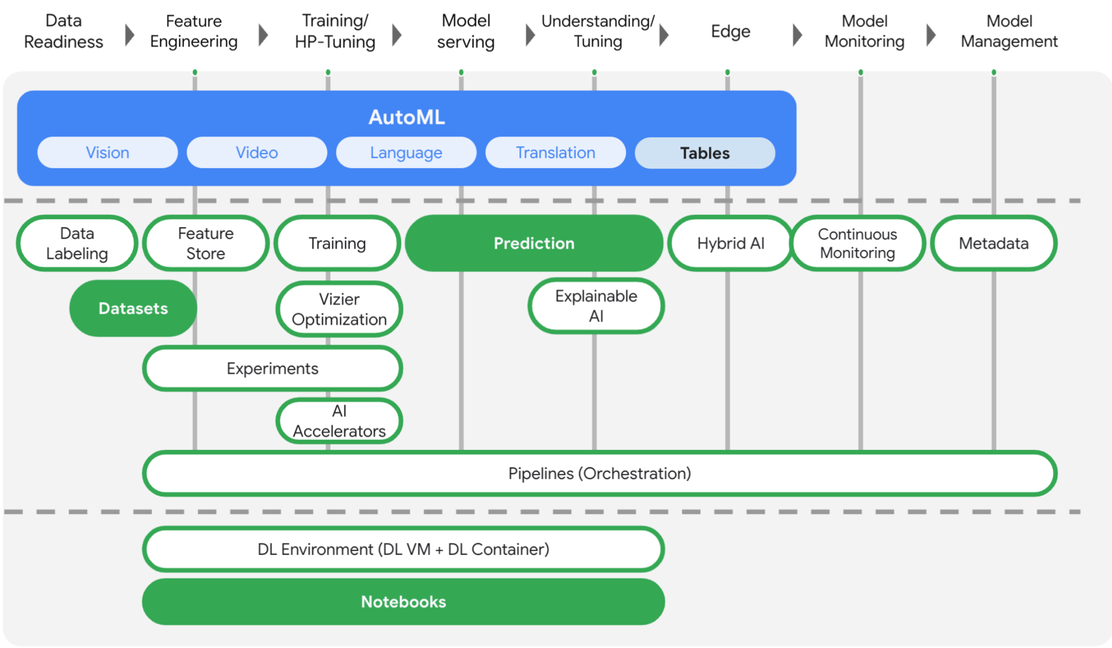
(Cloud-Based) ML Platforms

Amazon Sagemaker

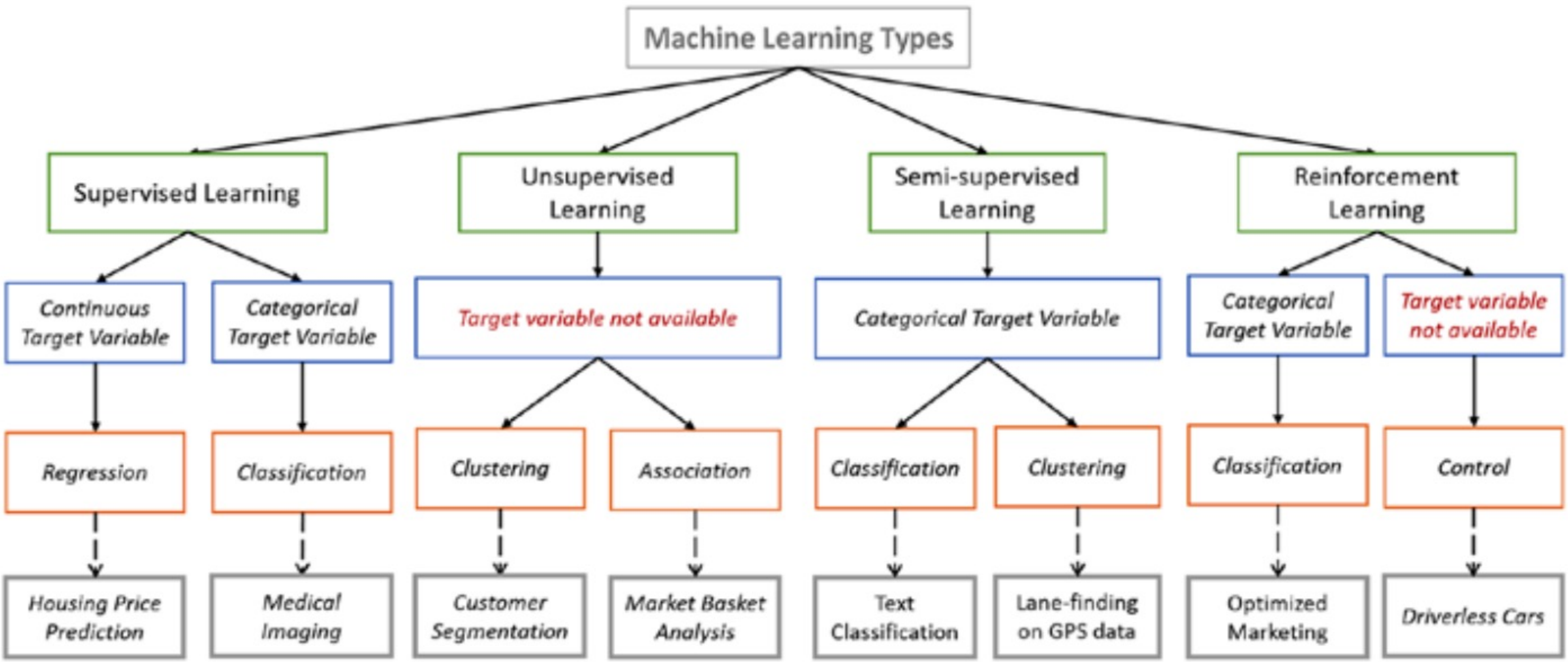
Microsoft Fabric : Microsoft Fabric is an end-to-end analytics solution with full-service capabilities including data movement, data lakes, data engineering, data integration, data science, real-time analytics, and business intelligence
—all backed by a shared platform providing robust data security, governance, and compliance.

Google Vertex AI

- ✓ Deploy more models, faster, with 80% fewer lines code required for custom modeling
- ✓ Use MLOps tools to easily manage your data and models with confidence and repeat at scale



Back to ML Models



Cold Start Problem: Mismatch → Online learning methods

No Free Lunch (NFL)

- No universally best model; so understand tradeoffs.
- Table from HTF

TABLE 10.1. *Some characteristics of different learning methods. Key:* ▲*= good,* ◆*=fair, and* ▼*=poor.*

Characteristic	Neural Nets	SVM	Trees	MARS	k-NN, Kernels
Natural handling of data of "mixed" type	▼	▼	▲	▲	▼
Handling of missing values	▼	▼	▲	▲	▲
Robustness to outliers in input space	▼	▼	▲	▼	▲
Insensitive to monotone transformations of inputs	▼	▼	▲	▼	▼
Computational scalability (large N)	▼	▼	▲	▲	▼
Ability to deal with irrelevant inputs	▼	▼	▲	▲	▼
Ability to extract linear combinations of features	▲	▲	▼	▼	◆
Interpretability	▼	▼	◆	▲	▼
Predictive power	▲	▲	▼	◆	▲

It Depends

“all models are wrong, but some are useful”

- George Box, 1987

- All statistical models make assumptions
  - (Let’s pretend…)
  - Given the situations, some assumptions are plausible, others are not

  Visualize: http://setosa.io/ev/ordinary-least-squares-regression/

"***Lies, damned lies, and statistics***”

Deep Nets and other Complex Models

- Very general & Powerful: Few or no assumptions
- Breakthrough results in
  - Images/video recognition
  - Language
  - Speech

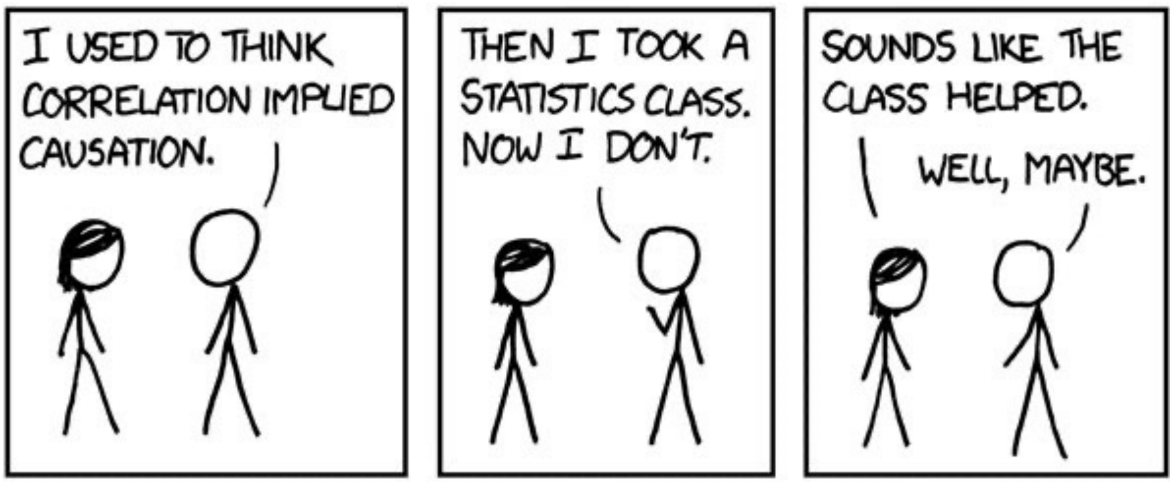
- but..
  - Lots of data (or transfer learning)
  - Lots of hyperparameter optimization
  - Lots of compute
  - Little statistical or human insights
  - Solution may not be robust

"no free lunch"

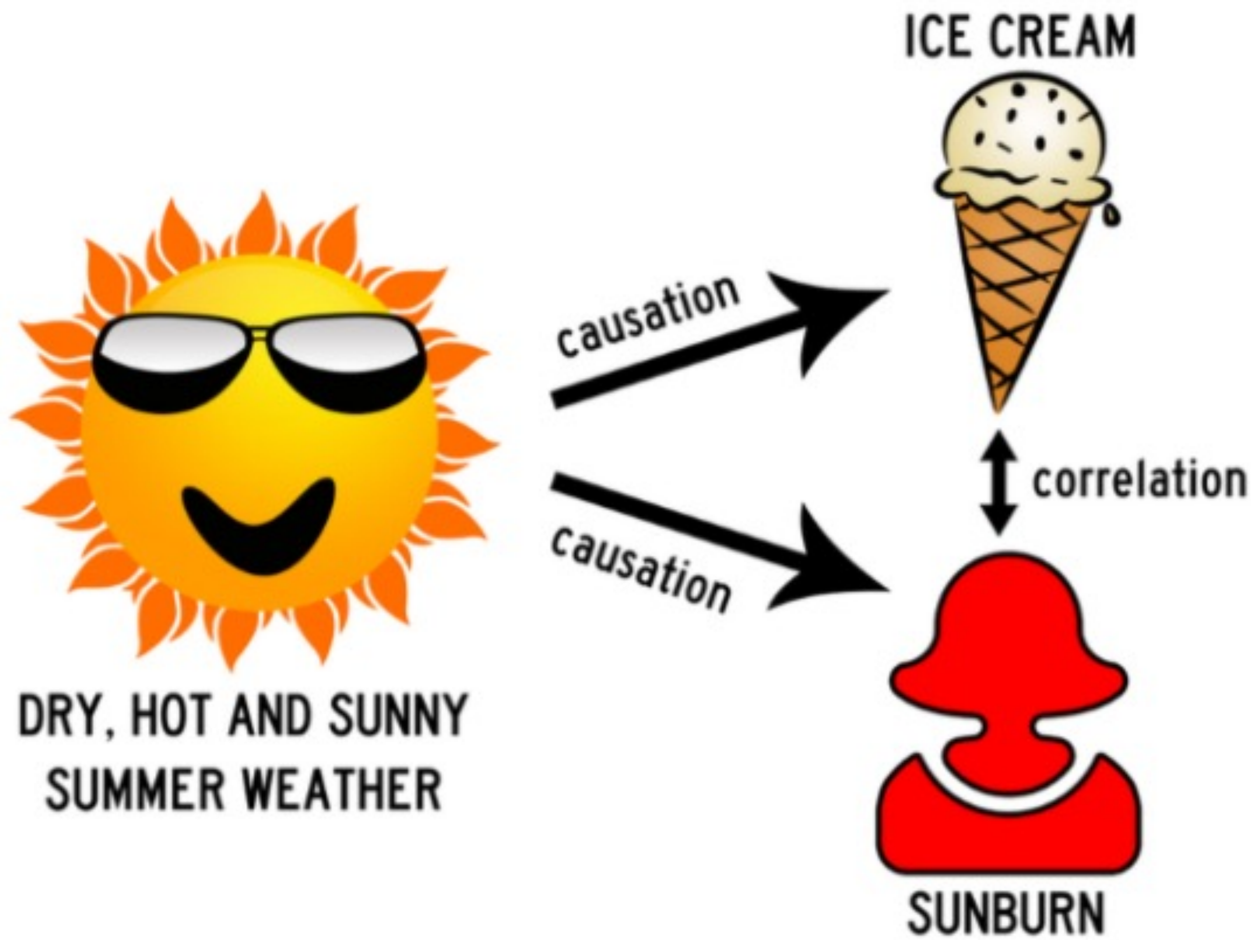
Course Goals

- study different predictive models for a given task
  - Properties, pros and cons
  - Evaluation metrics
    - Business relevance
  - Build predictive models in Python
- Process-oriented viewpoint
- Introduction to issues of scale and real data considerations

Broader Goals: Reason about data, its analysis and the "results" obtained



XKCD


ICE CREAM
causation
causation
correlation
DRY, HOT AND SUNNY
SUMMER WEATHER
SUNBURN

Causal??

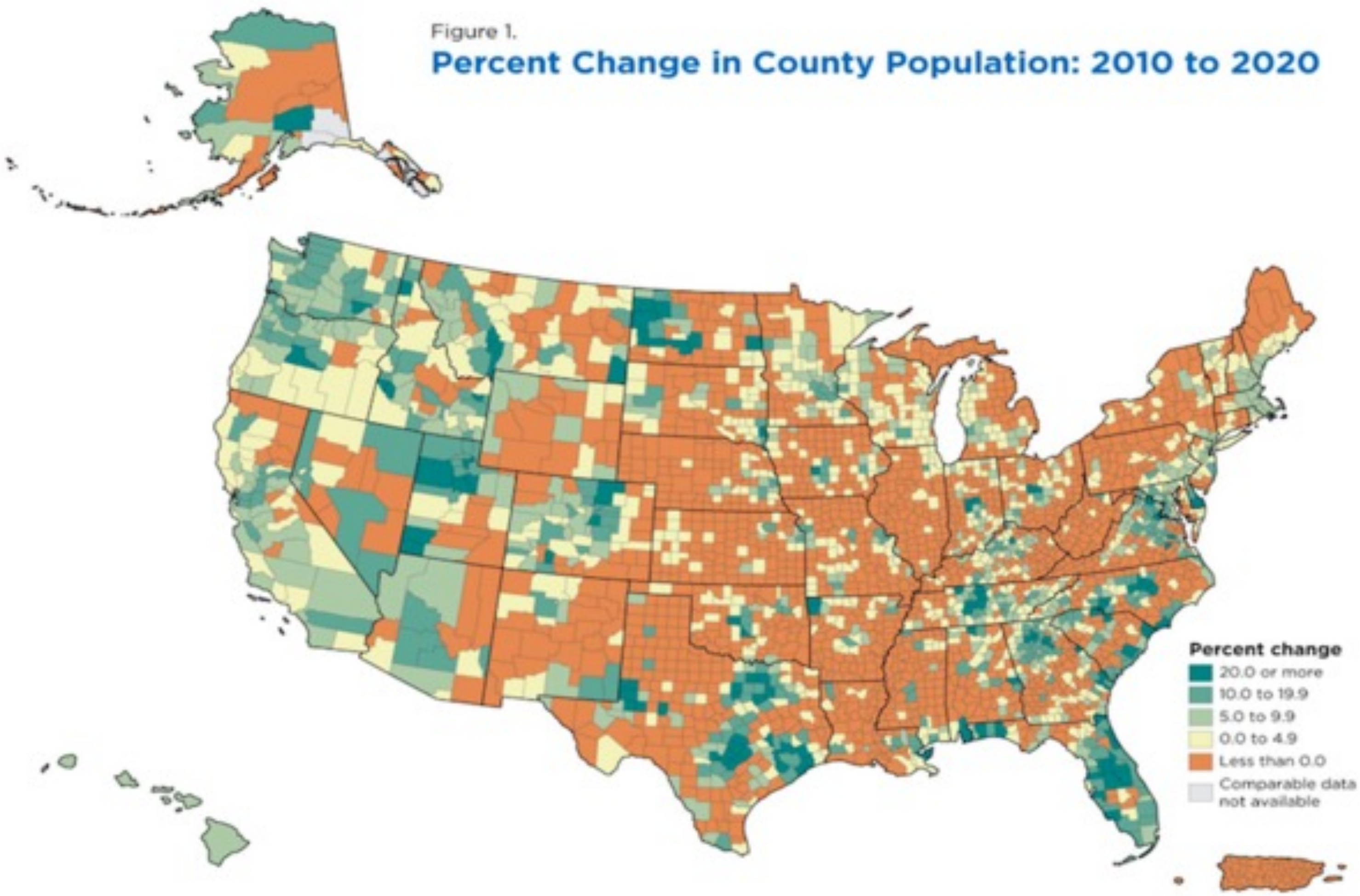
POLITICS

Vast Stretches of America Are Shrinking. Almost All of Them Voted for Trump.

Ninety percent of counties that lost population in the last decade backed the ex-president.

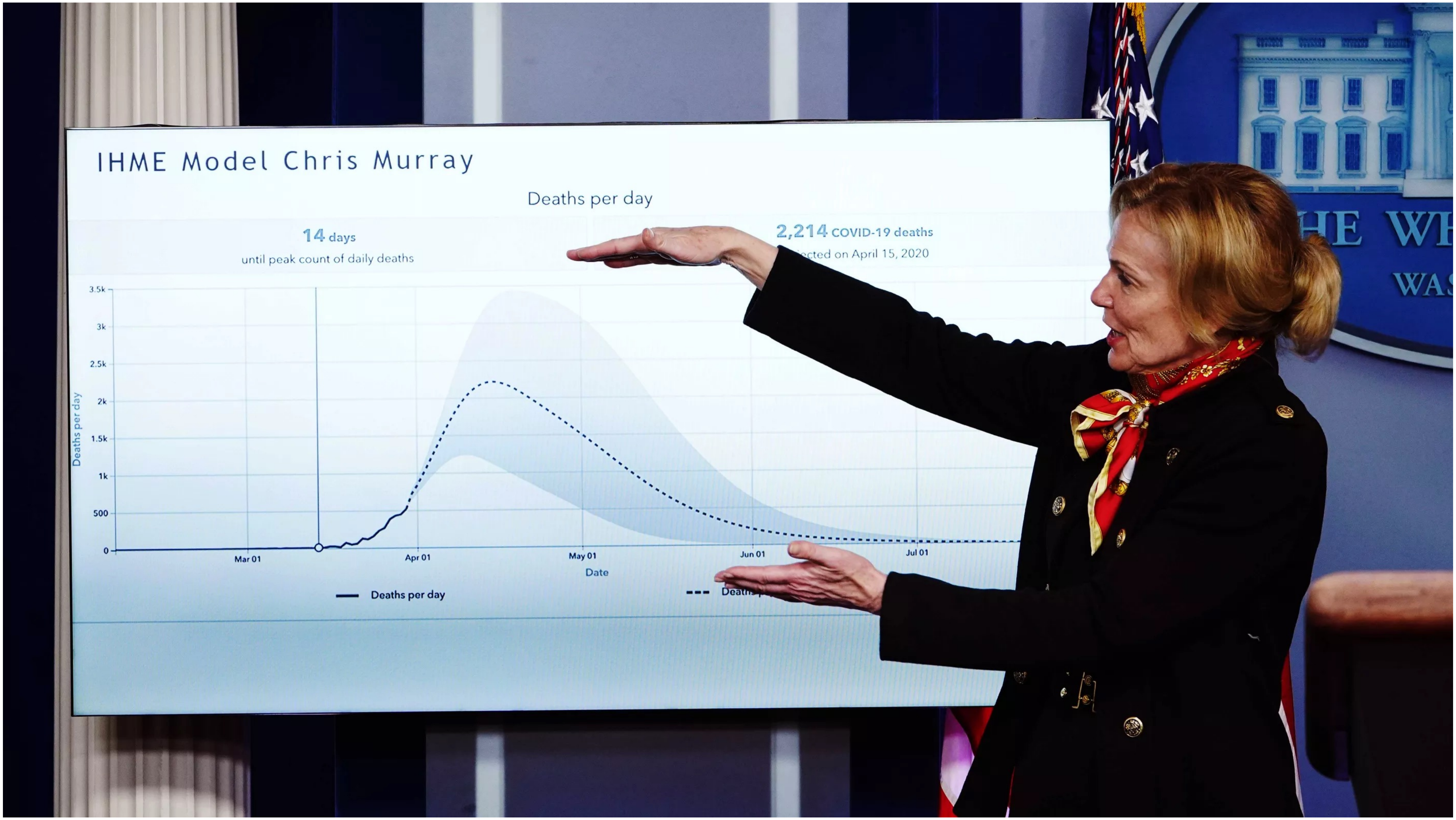
BY JORDAN WEISSMANN

AUG 14, 2021 • 5:40 AM



This coronavirus model keeps being wrong. Why are we still listening to it?

in early April, it revised its projections to say that the total death toll through August was "projected to be 60,415" (though it acknowledged the range could be between 31,221 and 126,703).



- https://www.vox.com/future-perfect/2020/5/2/21241261/coronavirus-modeling-us-deaths-ihme-pandemic

Sanity Checks

- **One analysis of the IHME model found** that its next-day death predictions for each state were outside its 95 percent confidence interval 70 percent of the time — meaning the actual death numbers fell outside the range it projected 70 percent of the time.

Towards Good Predictive Models

- Use data driven models to complement domain expertise and intuition
  - Understand problem context
  - Get relevant data
  - Use versatile toolbox and select appropriately
    - Prediction vs. interpretation tradeoff
    - Tailor to data properties
      - » But do not overfit
  - Convey results effectively

(End of Broad Intro)

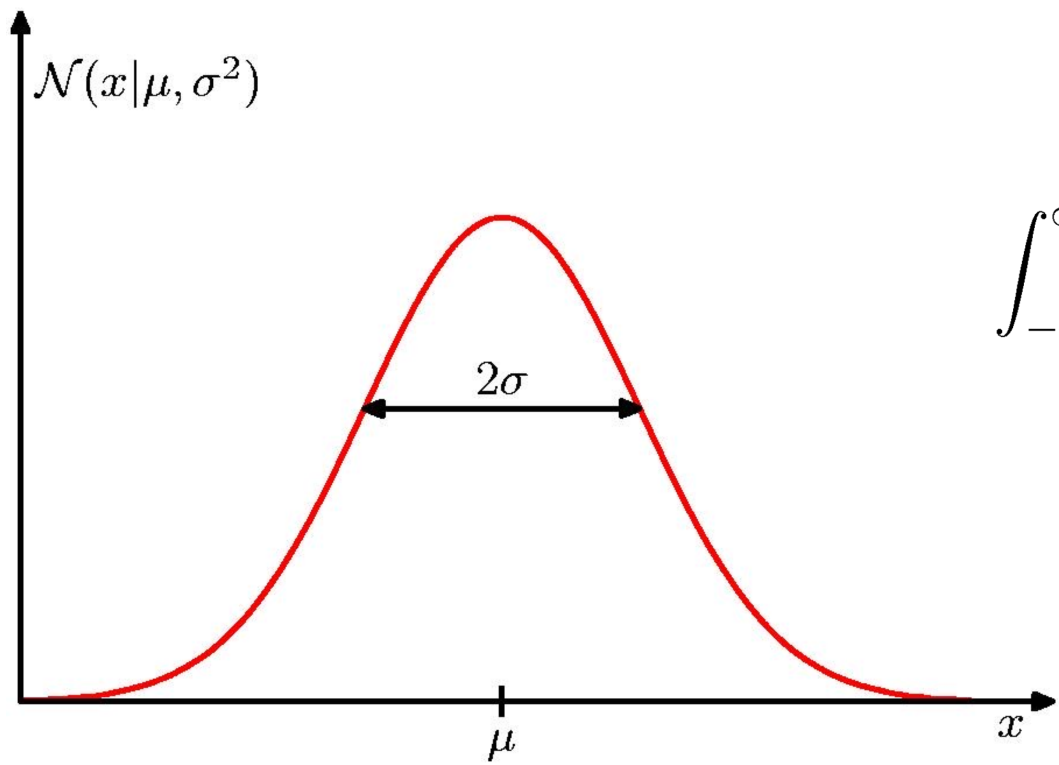
Read: CB:1.2 to 1.2.4; KM: 2.2

- Basic Concepts:
  - Discrete vs. Continuous Variables
  - Joint distribution of Multiple Variables
    - Marginal distribution
    - Conditional distribution (Video)
    - Covariance

Visualize: http://setosa.io/ev/conditional-probability/

The Gaussian (or "Normal") Distribution

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) \, \mathrm{d}x = 1$$

Gaussian Mean and Variance

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x \, \mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

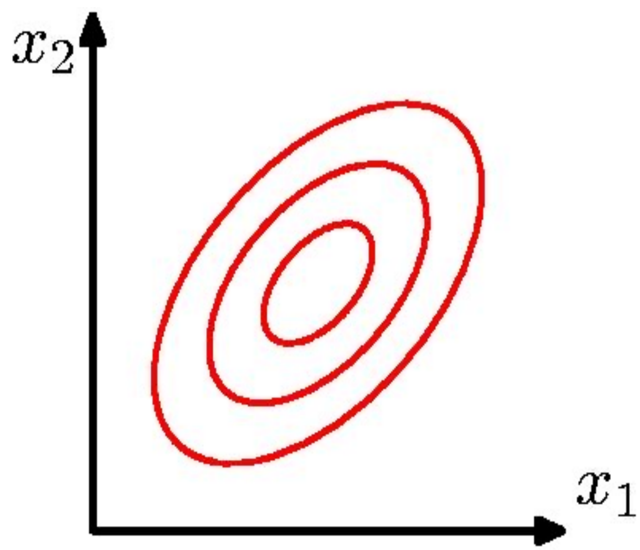
$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

Denotes the "expectation" operator

The Multivariate Gaussian (in D dimensions)

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Vector Mean *D-by-D Covariance Matrix* *Determinant of the covariance matrix*



Marginals and *conditionals* of multivariate Gaussians?

Gaussian Parameter Estimation

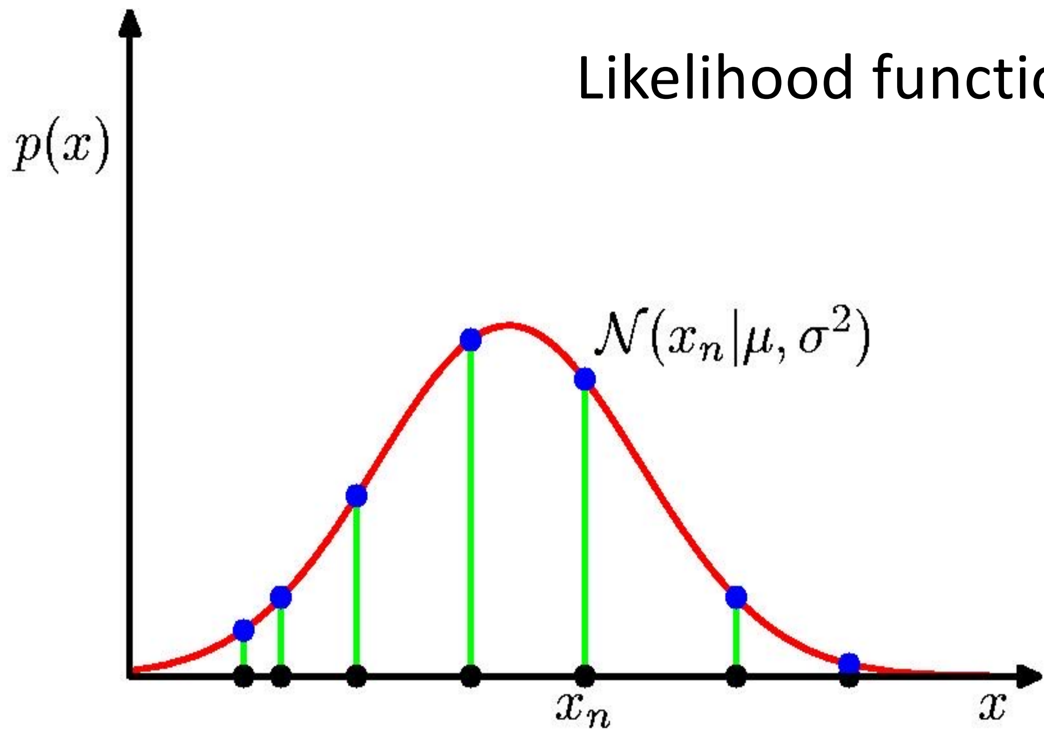
- **Given**: a datset $\mathcal{X}$ of size N, (assumed to be) obtained i.i.d. from an unknown Gaussian Distribution
- **Goal**: obtain your best estimate of the parameters of this Gaussian

Maximum Likelihood Principle provides a general and principled way of obtaining such an estimate.

Read: CB 2.3 to 2.3.4, KM: 4.2, EA: 4.2

Gaussian Parameter Estimation

- What is the probability that a datset $\mathcal{X}$ with N i.i.d. points was obtained from a specified Gaussian?



Likelihood of observing the given set of points $\mathcal{X}$ $= \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right)$

Maximum (Log) Likelihood Principle

- **Apply ML principle to select the Gaussian that most likely produced the given dataset.**

$$\boxed{\textit{Log Likelihood} =} -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

(Note: for fixed σ, NLL is equivalent to using sum/mean squared error cost function to estimate the mean.)

- **Maximized when**

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

Why know about ML?

- **Are your ML estimates *biased*?**

Parametric Estimation

- $\mathcal{X} = \{ x_n \}$, where $x_n \sim p(x)$
- Parametric estimation:

  Assume a parametric form for $p(x \mid \theta)$ and estimate θ, its parameters (sufficient statistics), using $\mathcal{X}$

  e.g., $\mathcal{N}(\mu, \sigma^2)$ where $\theta = \{ \mu, \sigma^2 \}$

Maximum Likelihood Estimation

- Likelihood of θ given the sample $\mathcal{X}$

$$l(\theta|\mathcal{X}) = p(\mathcal{X}|\theta) = \prod_n p(x_n|\theta)$$

- Log likelihood

$$\mathcal{L}(\theta|\mathcal{X}) = \log l(\theta|\mathcal{X}) = \sum_t \log p(x_n|\theta)$$

- Maximum likelihood estimator (MLE)

$$\theta^* = \text{argmax}_\theta \, l(\theta|\mathcal{X})$$
$$= \text{argmax}_\theta \, \mathcal{L}(\theta|\mathcal{X})$$

(or minimize negative log-likelihood (NLL), i.e. treat NLL as a cost function)

Videos: https://youtu.be/XepXtl9YKwc and https://youtu.be/Dn6b9fCIUpM

Extras

Four Trends

- MLOps; Integrating with Software Environment
  - Model lifecycle management
  - Kubernetes
  - Enterprise grade services, e.g. **Feathr** – An Enterprise-Grade, High Performance Feature Store, open-sourced by LinkedIn, Apr 2022.
- Integrating with Business KPIs, and with other Decision-Making Systems. ("AI Engineering" or Enterprise AI)
  - Human in the loop
  - Trustworthy AI (Fairness/Bias, Explanability, Robustness,..)
- AutoML
  - https://www.topbots.com/automl-solutions-overview/
- AI function as a service (often Deep Learning oriented)

What is MLOps?

- See https://towardsdatascience.com/ml-ops-machine-learning-as-an-engineering-discipline-b86ca4874a3f. Also see Google's take

ML Ops is a set of practices that combines Machine Learning, DevOps and Data Engineering, which aims
to deploy and maintain ML systems in production reliably and efficiently.

Practice	DevOps	Data Engineering	ML Ops
Version control	Code version control	Code version control Data lineage	**Code version control + Data versioning + Model versioning (linked for reproducibility)**
Pipeline	n/a	Data pipeline/ETL	**Training ML Pipeline, Serving ML Pipeline**
Behavior validation	Unit tests	Unit tests	**Model validation**
CI/CD	Deploys code to production	Deploys code to data pipeline	**Deploys code to production + training ML pipeline**
Data validation	n/a	Format and business validation	**Statistical validation**
Monitoring	SLO-based	SLO-based	**SLO + differential monitoring, statistical sliced monitoring**

SLO = service level objective

Languages and Software

- Stats oriented: R, Python (with packages)
  - Commercial: SAS, IBM SPSS,..
  - Open: GUI oriented: Knime, RapidMiner
- General purpose (Java for text analysis)
- Distributed/bigdata
  - Hadoop/Spark/MapReduce/PigLatin
  - HIVE (SQL like for Hadoop)
  - Various NoSQL
- New (2018): AUTO-ML (DataRobot, H2O,…); (2019): ML in the cloud. (2020): ML-OPS, AI Engineering

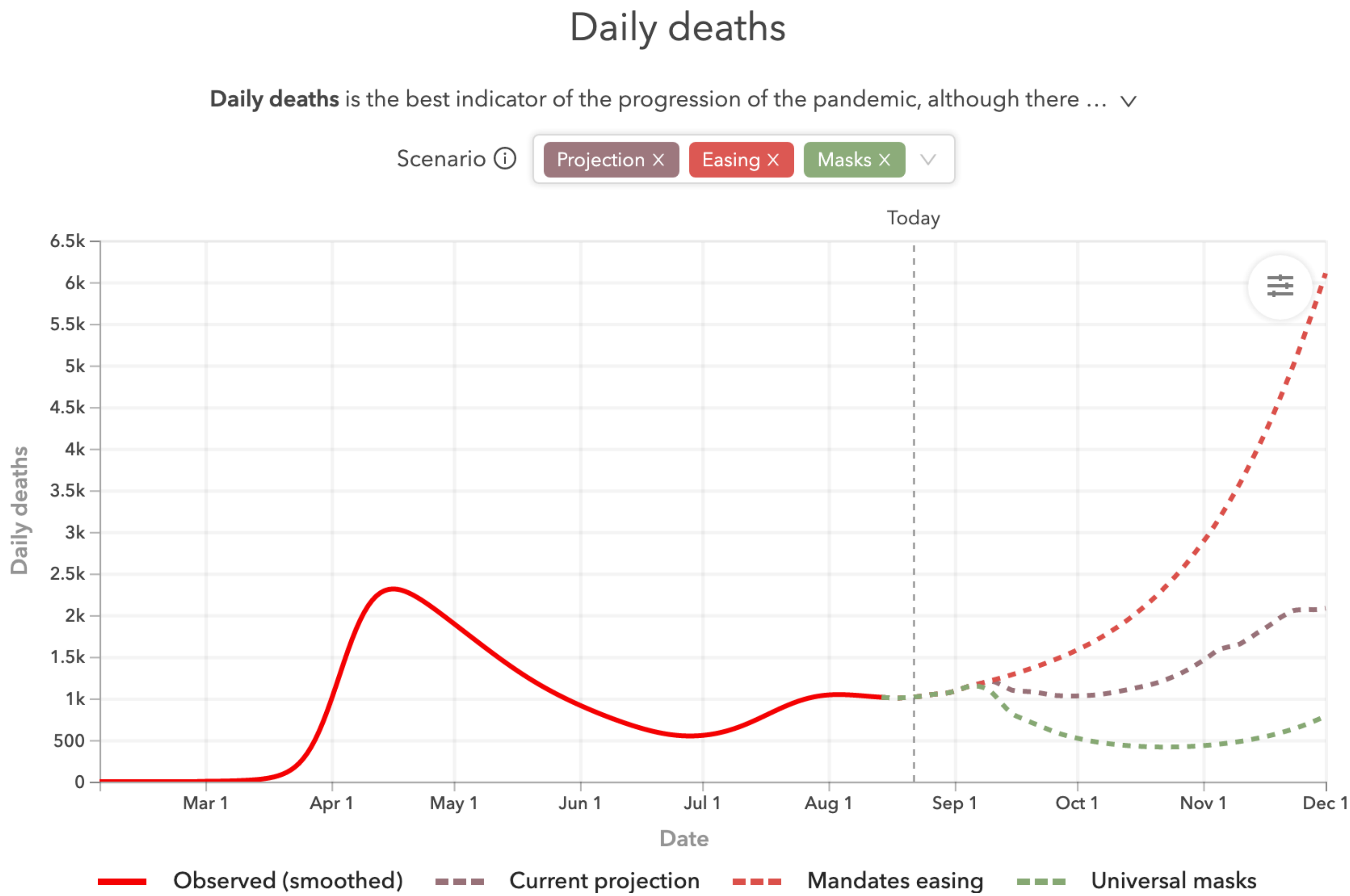
See: How Did Python Become A Data Science Powerhouse?

https://www.youtube.com/watch?v=9by46AAqz70

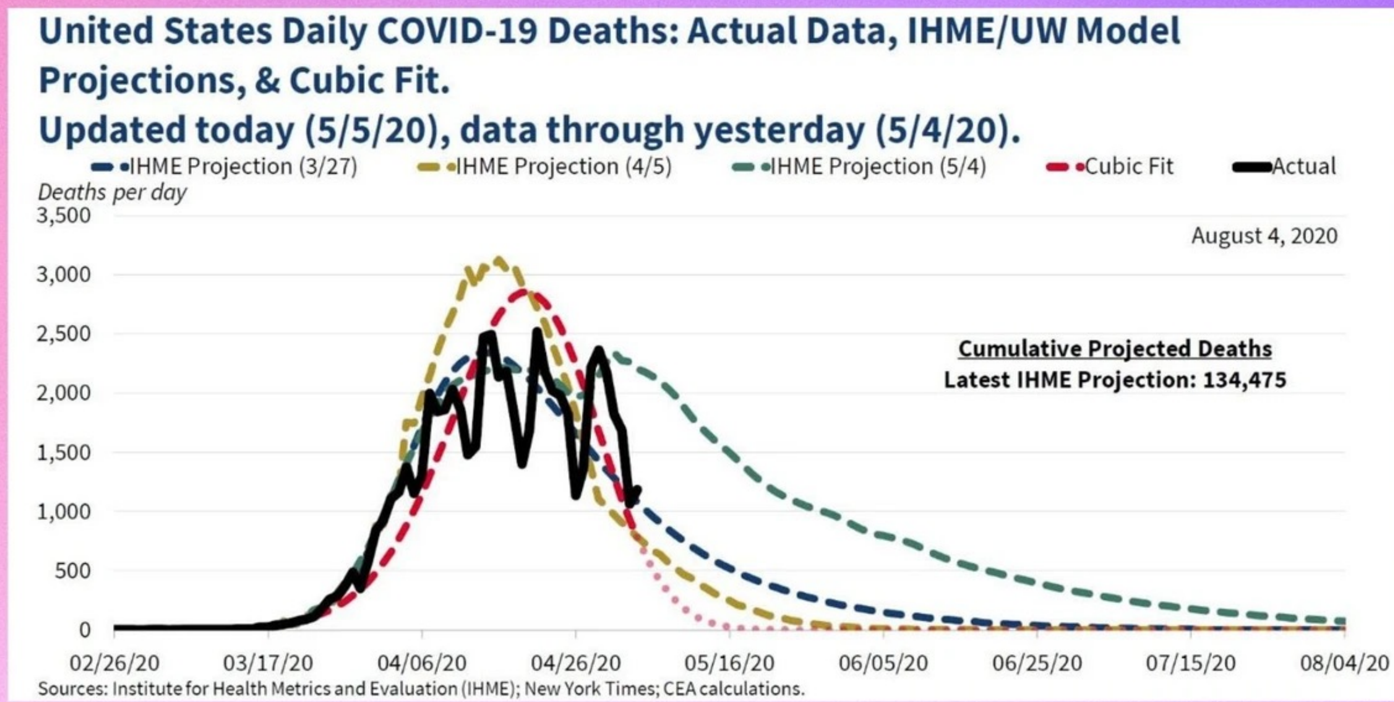
https://www.datanami.com/2019/08/15/is-python-strangling-r-to-death/

A Bit More on IHME Model

- [Projection as of 8/22/2020](#) (brown curve = ~310K deaths by Dec 1, 2020)
- github repo - https://ihmeuw-msca.github.io/CurveFit/methods/.



Cubic Model from White House Council of Economic Advisors



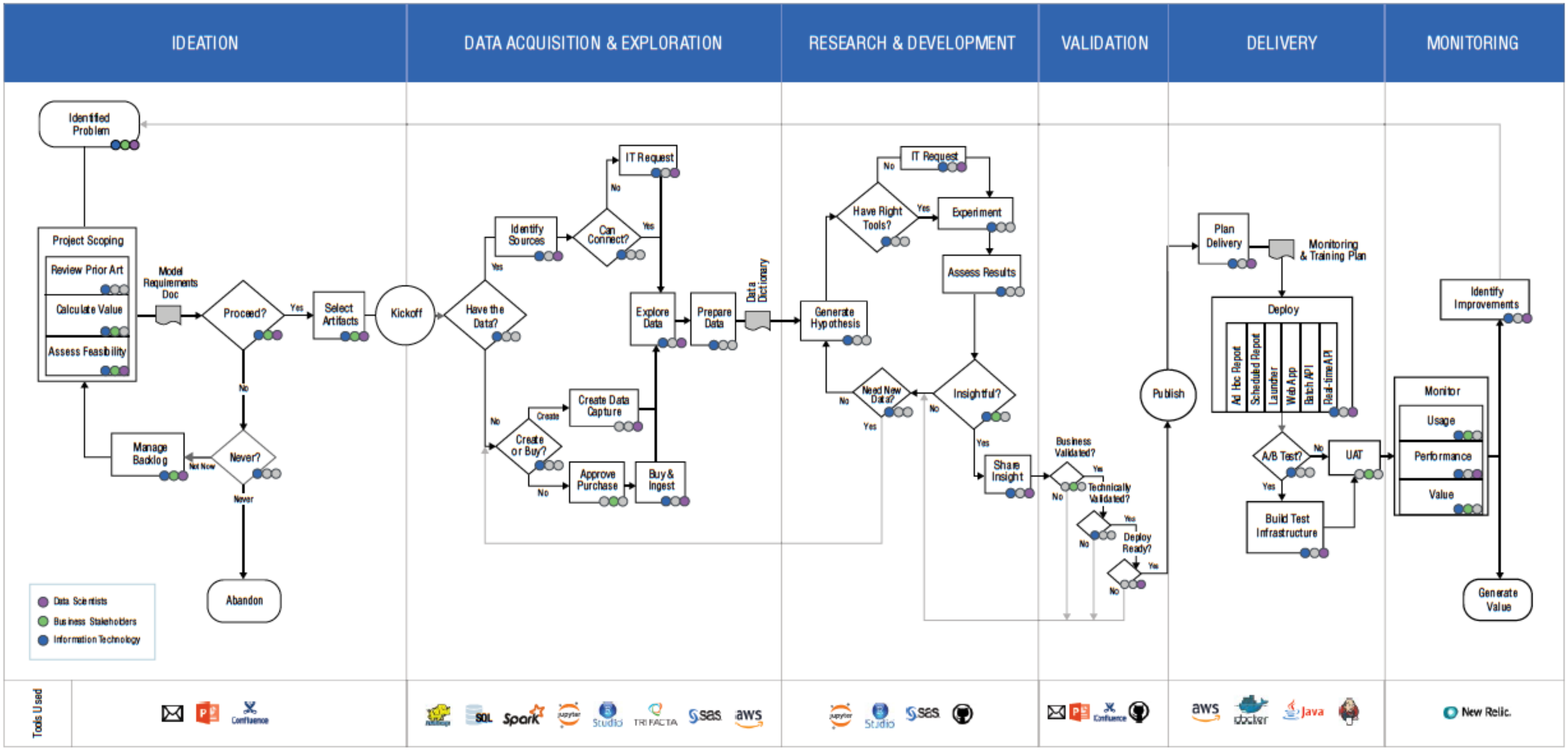
https://www.vice.com/en_us/article/bv8gym/amateur-hour-white-house-graph-shows-covid-19-deaths-hitting-0-in-10-days

Simpson's Paradox

- https://www.covid-datascience.com/post/israeli-data-how-can-efficacy-vs-severe-disease-be-strong-when-60-of-hospitalized-are-vaccinated
- Pfizer efficacy in Israel

Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	vs. severe disease
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	**67.5%**
<50	1,116,834 23.3%	3,501,118 73.0%	43 3.9	11 0.3	**91.8%**
>50	186,078 7.9%	2,133,516 90.4%	171 91.9	290 13.6	**85.2%**

DATA SCIENCE LIFECYCLE



Read the "Domino" article before/while doing your project