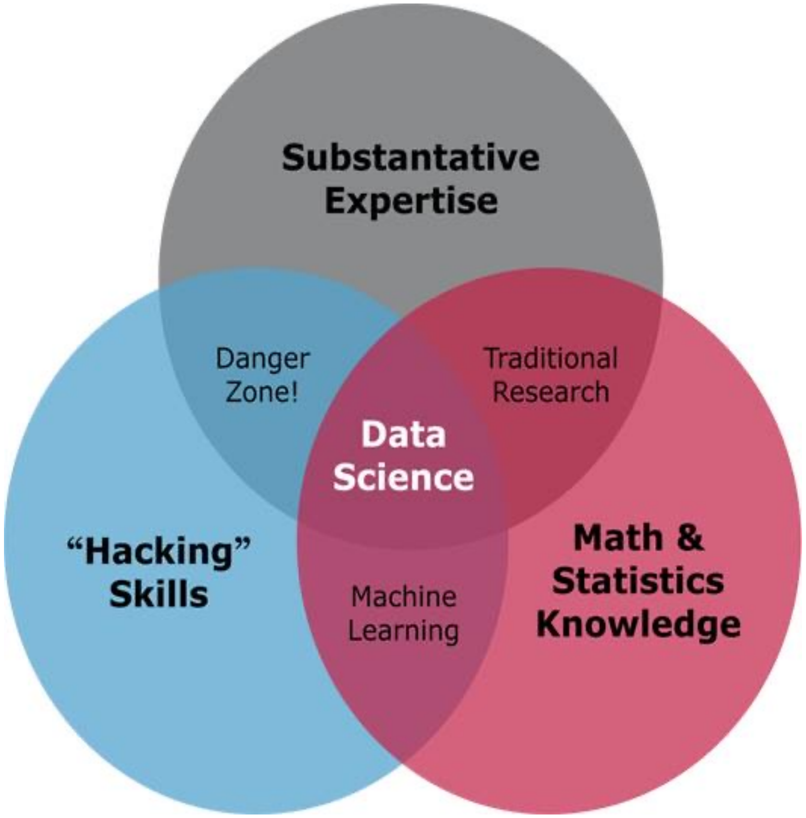
MIS 382N: Advanced Machine Learning

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,…)

Analysis is often **retrospective.**

(Backup) Texts

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

Other References:

- **Basic**
  - **JW:** **ISLR:** Intro to stats learning with R
  - **KJ:** Kjell and Johnson. Applied Predictive Modeling, Springer 2013.
  - http://appliedpredictivemodeling.com/
- **Advanced: HTF:** Hastie/Tibshirani/Friedman **(stats)**
  http://www-stat.stanford.edu/~tibs/ElemStatLearn/

  Math for Machine Learning (2020, Reference).

Machine Learning (ML) and Deep Learning

Machine Learning:

- **Introduction** to Machine Learning, Ethem Alpaydin (2014), MIT Press. [Book home page (3rd edition)]

- (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.
  - Coding oriented

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.

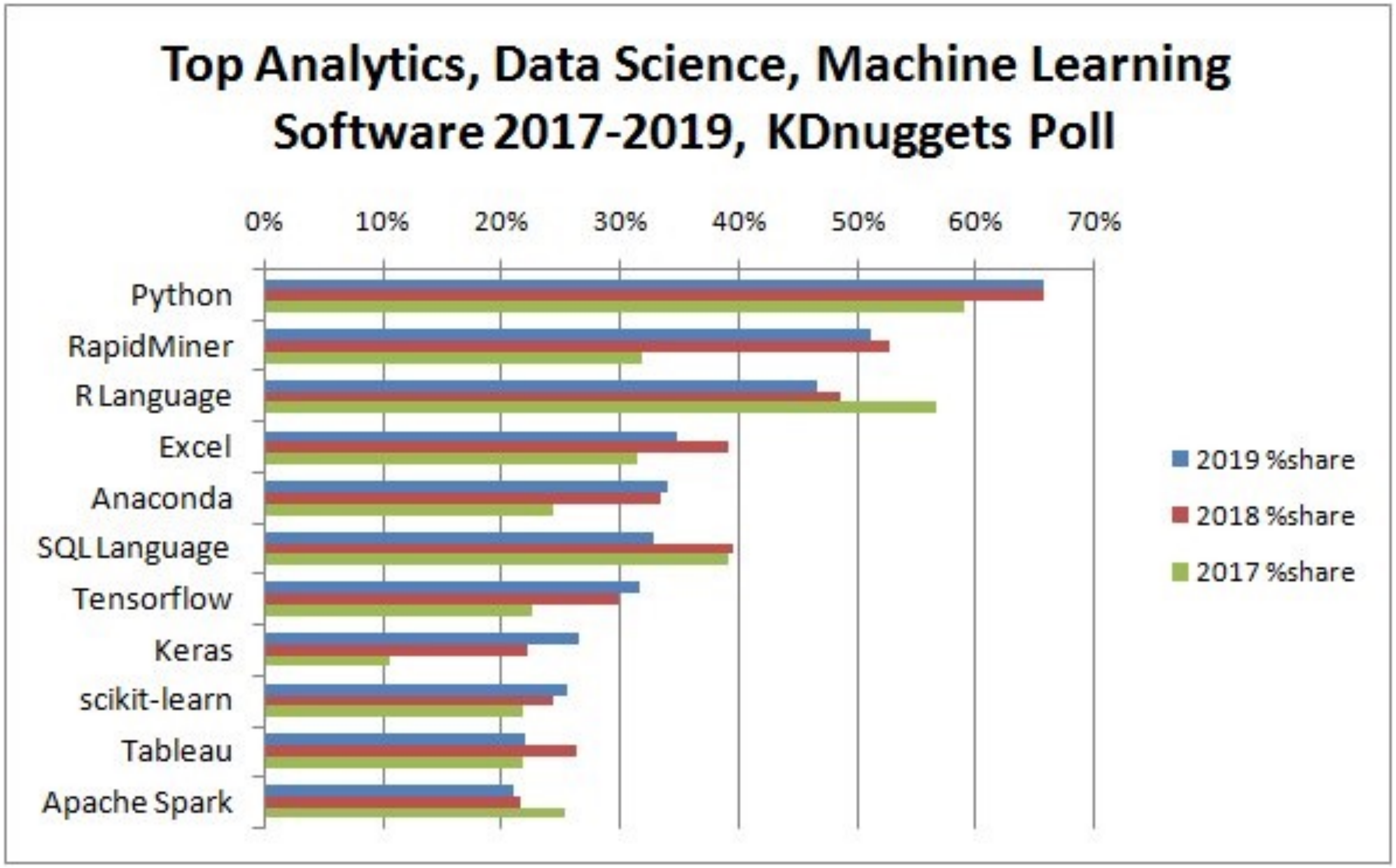
See: How Did Python Become A Data Science Powerhouse?

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

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

KDD Nuggets Survey May 2019

- Animation



Also see analysis of Kaggle's 2020 ML/Cloud Computing Survey

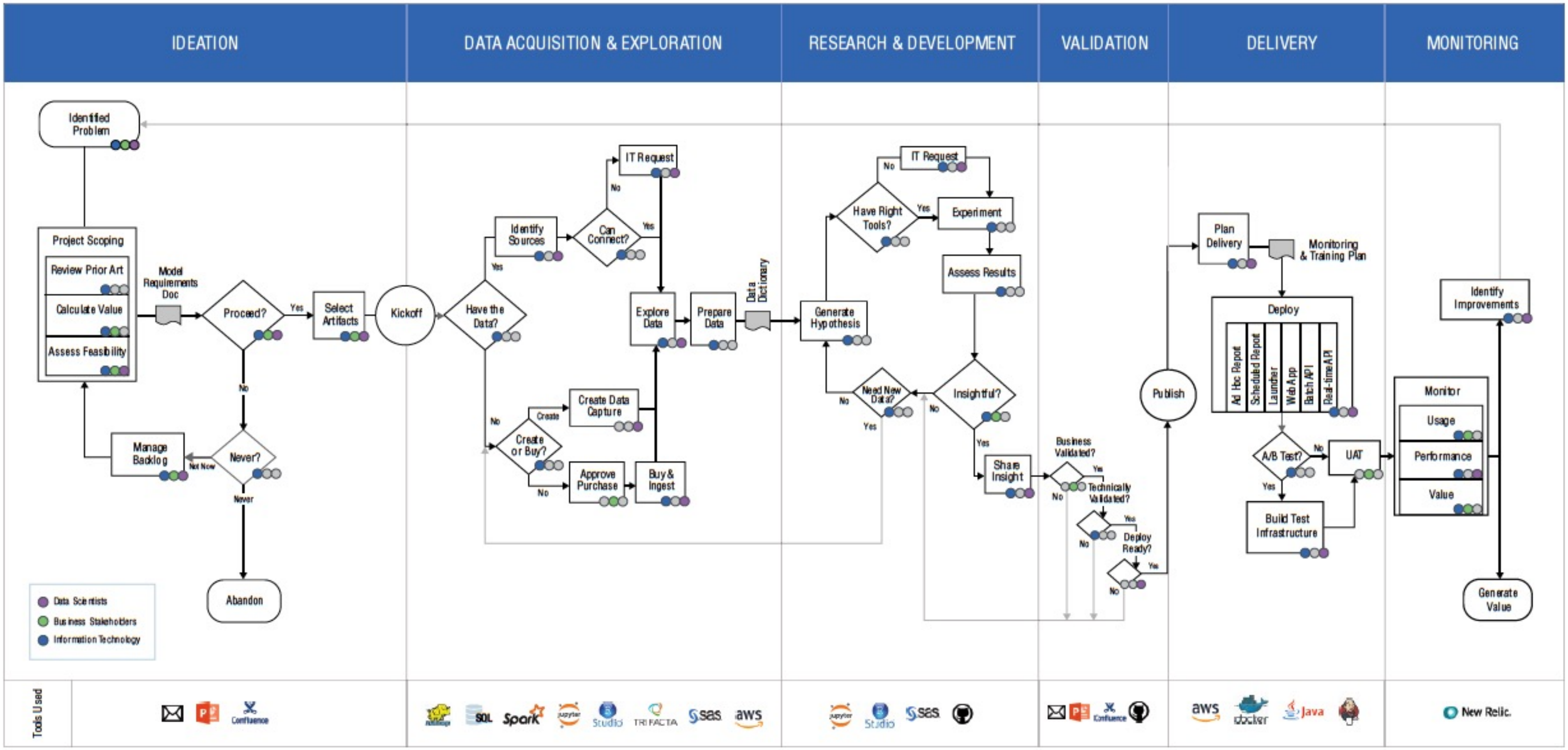
Trends

- MLOps; Integrating with Software Environment
  - Model lifecycle management
  - Kubernetes
  - Cloud Based Platforms (vs. On-prem vs. Edge)

- Integrating with Business KPIs, and with other Decision Making Systems

- Human in the loop
- Trustworthy AI (Fairness/Bias, Explanability, Robustness,..)
- AutoML
  - https://www.topbots.com/automl-solutions-overview/

DATA SCIENCE LIFECYCLE



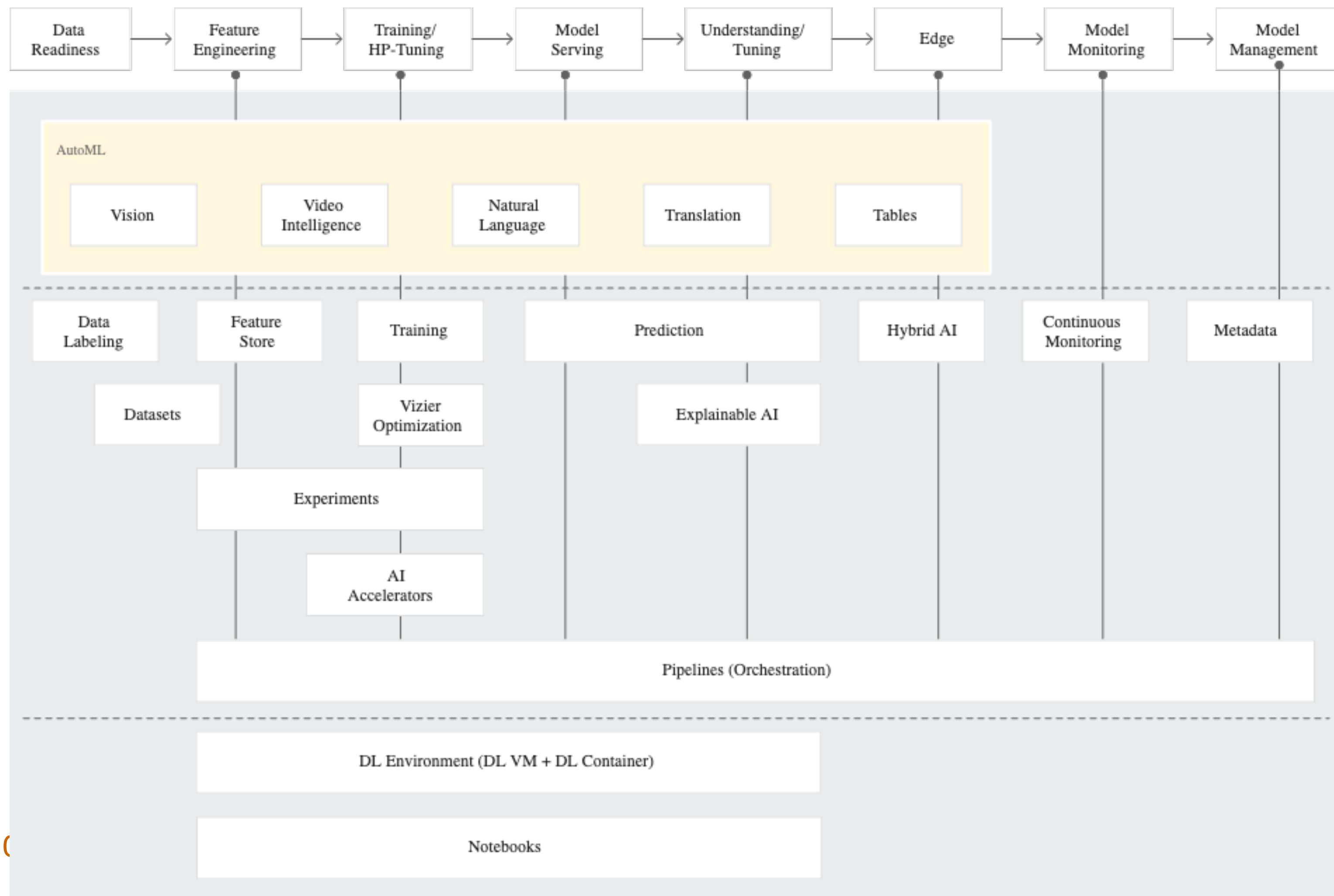
Read the "Domino" article before/while doing your project

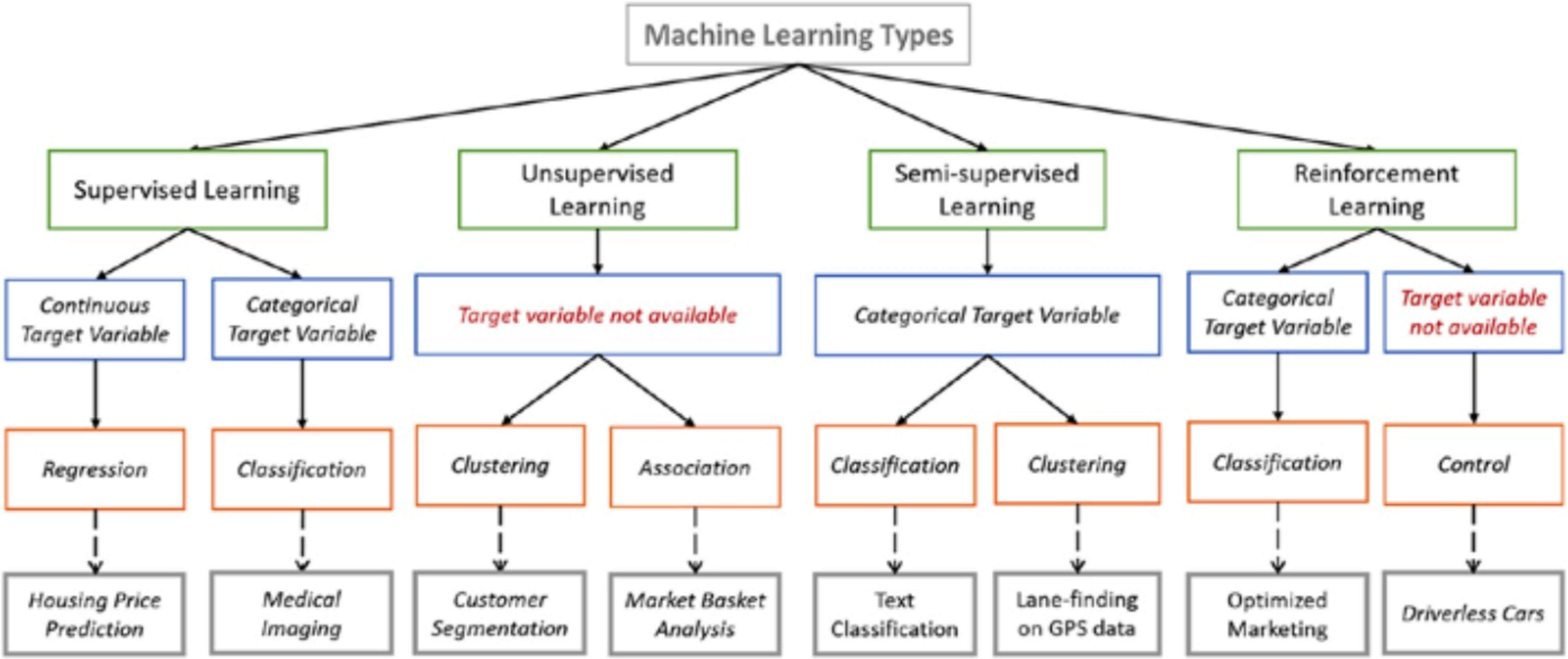
Google’s Vertex AI (May 2021)

- https://cloud.google.com/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





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

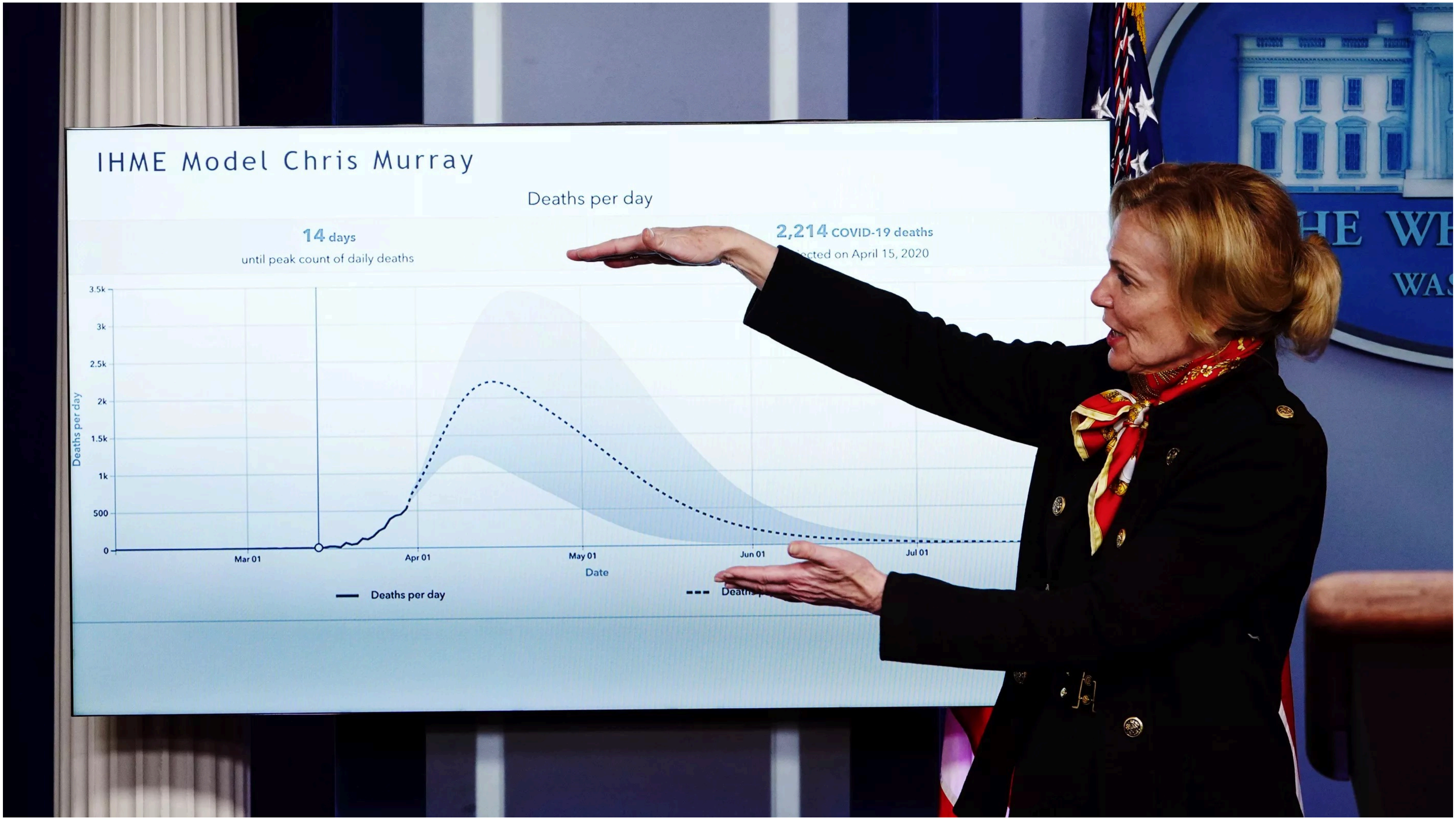
- All statistical models make assumptions
  - (Let’s pretend…)
  - Given the situations, some assumptions are plausible, others are not

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

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

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

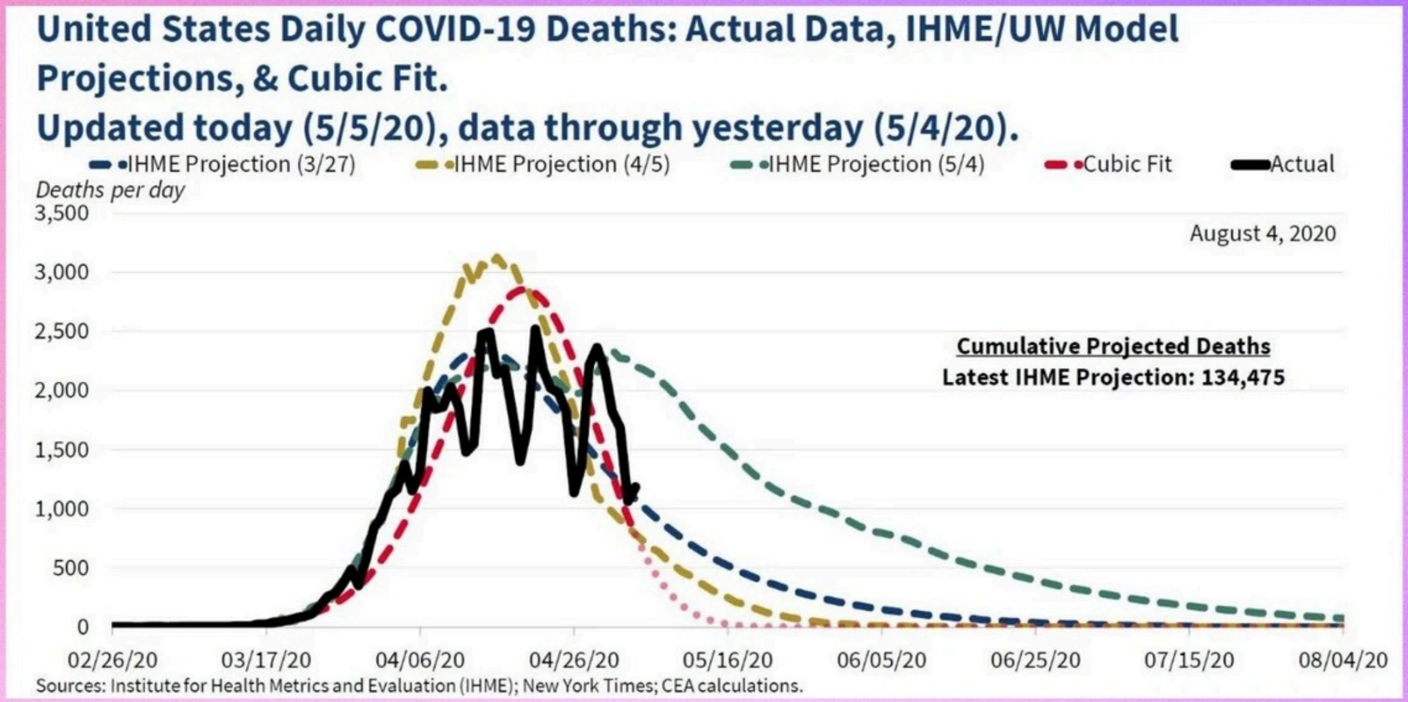
in early April, it revised its projections to say that the total death toll through August was “projected to be 60,415” (though it acknowledged the range could be between 31,221 and 126,703).



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

Sanity Checks

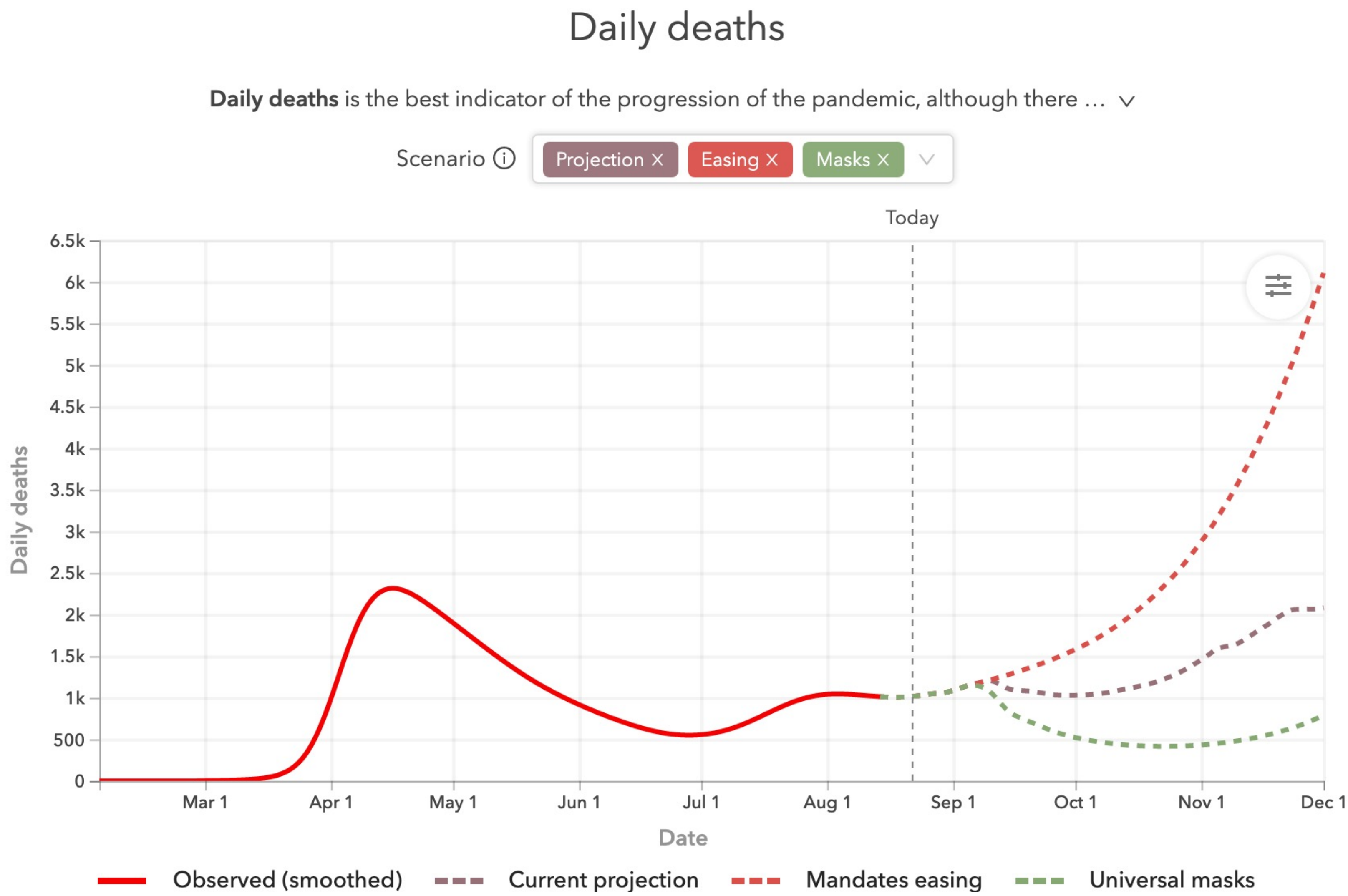
- **One analysis of the IHME model found** that its next-day death predictions for each state were outside its 95 percent confidence interval 70 percent of the time — meaning the actual death numbers fell outside the range it projected 70 percent of the time.



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

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/.



Deep Nets and other Complex Models

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

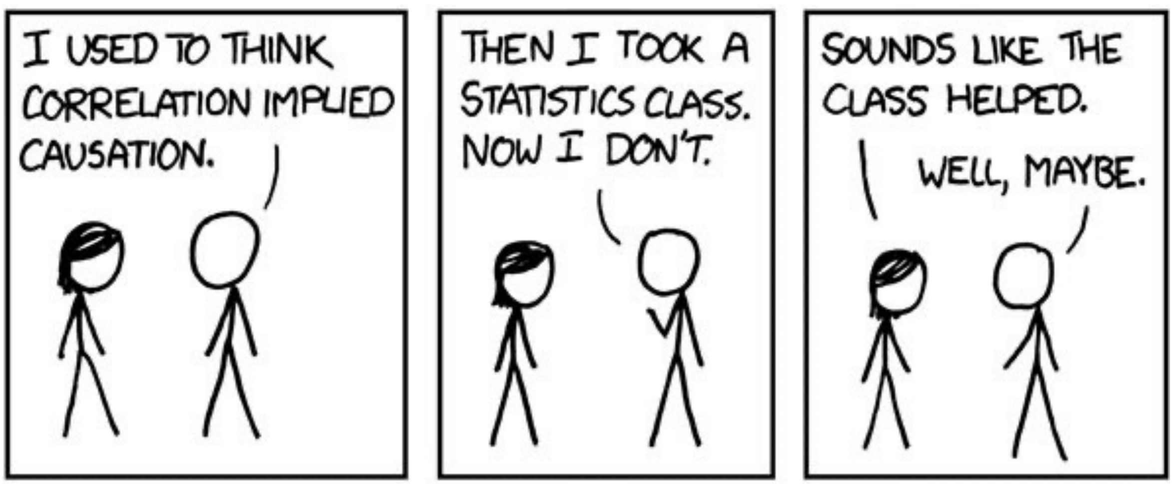
- but..
  - Lots of data (or transfer learning)
  - Lots of hyperparameter optimization
  - Lots of compute
  - Little statistical or human insights
  - Solution may not be robus

"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 analysis and the “results” obtained



XKCD

Towards Good Predictive Models

- Use data driven models to complement domain expertise and intuition (see quotes in KJ 1.2)
  - 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

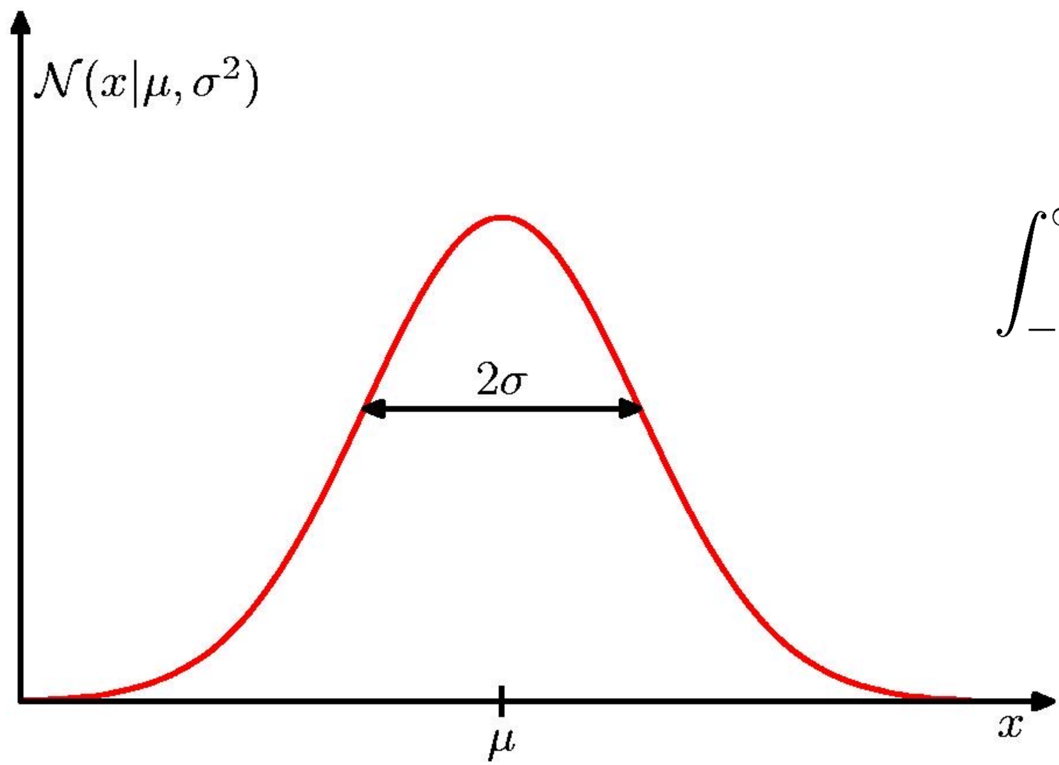
Probability Recap

- Basic Concepts:
  - Joint distribution
  - Marginal distribution
  - Conditional distribution
  - Variance and covariance

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

The Gaussian 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

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

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

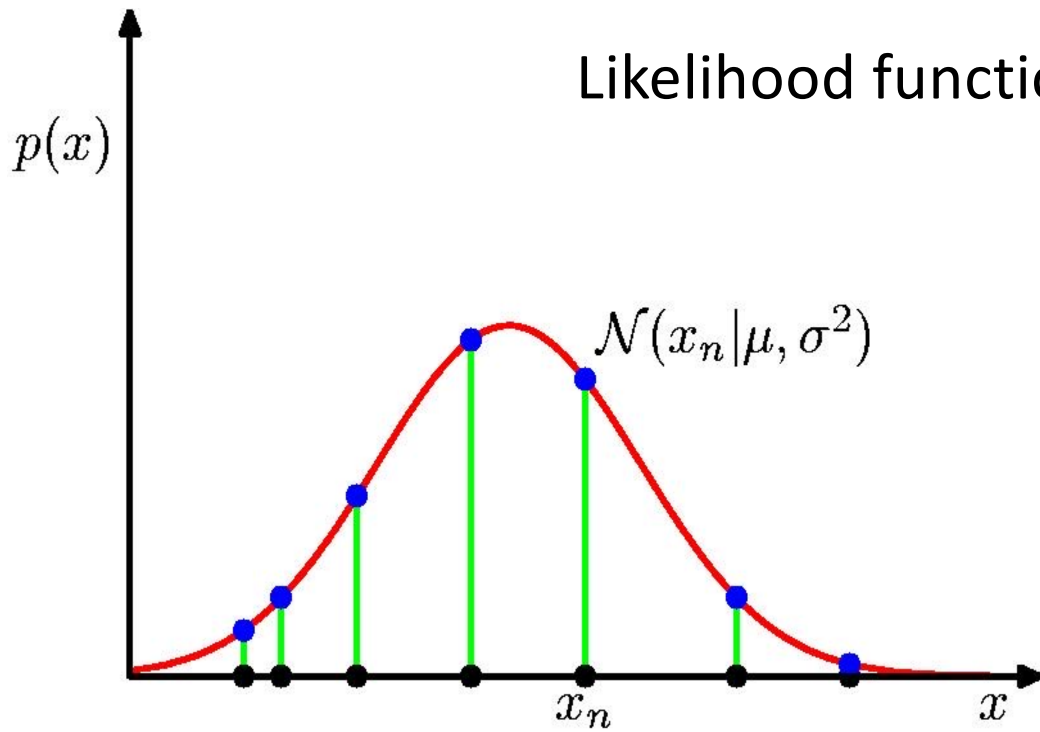
$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

Denotes the "expectation" operator

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 function



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

Maximum (Log) Likelihood Principle

- **Applied 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, "cost" is sum/mean squared error)

- **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 ML estimates *biased*?**

Extras

What is MLOps?

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

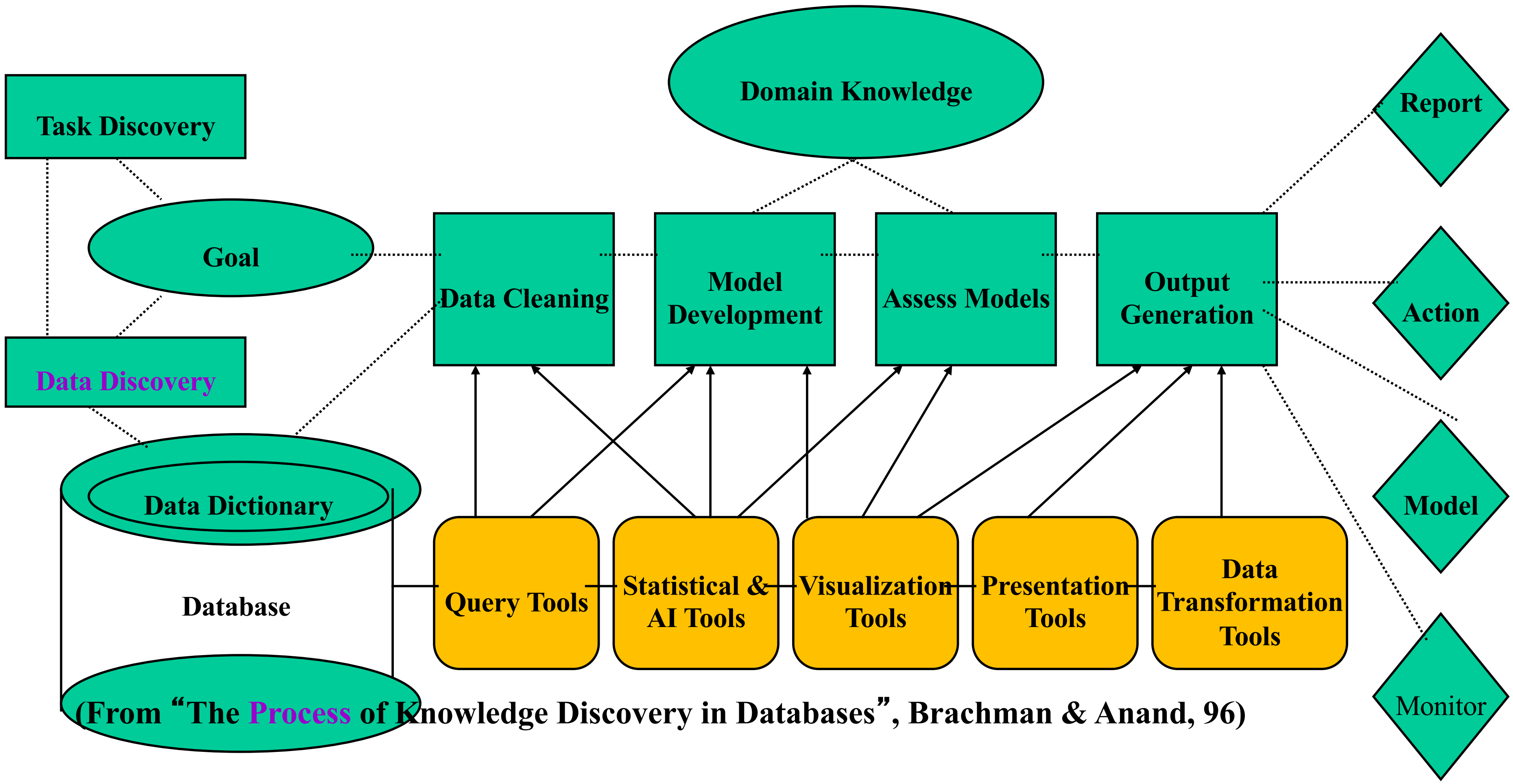
ML Ops is a set of practices that combines Machine Learning, DevOps and Data Engineering, which aims to deploy and maintain ML systems in production reliably and efficiently.

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

(Inter-disciplinary) Process of Data Mining

Collect *Prepare* *Analyze* *Consume*



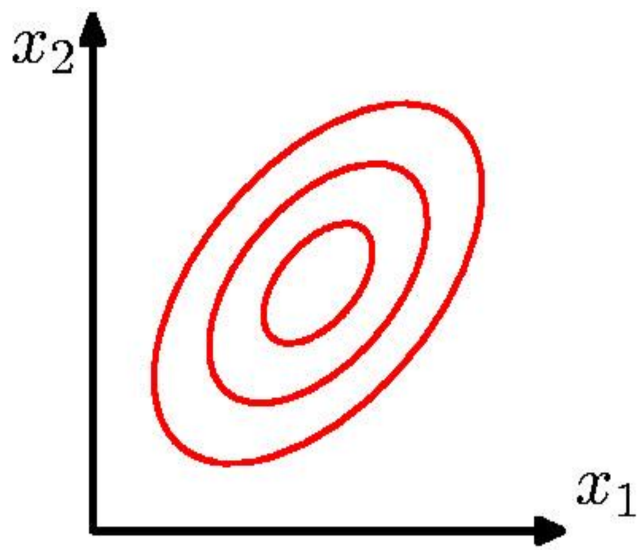
(From "The Process of Knowledge Discovery in Databases", Brachman & Anand, 96)

Iterative Process at multiple levels

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?

Nate Silver's 2012 model

- Uncertainly is everywhere
  - even in the "perfect" model
  - (used weighted ensemble)