Selected Techniques in Machine Learning

Chi-kwan "CK" Chan, PHYS 305, 2025-02-27

Goals

- Overview
- Know where to look
- Terminology
- How to ask/answer the "right" questions
- The value of data
- Opportunities (in astronomy)
- Challenges

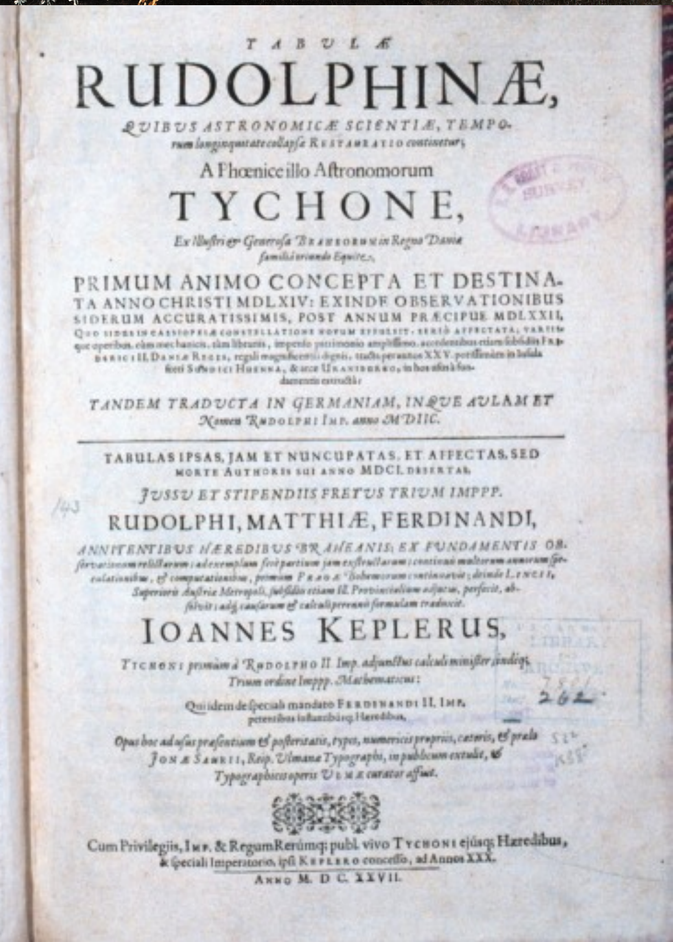
- How to answer questions with data?
- How to use data to come up with good questions?

Historical Context

- Data science is NOT NEW!
- Data-intensive/driven science:
  - Science that we do based on data, e.g., astronomy
- Data science:
  - Emergent discipline motivated by data-intensive/driven science.
  - The science of how to: collect, handle, clean, store, visualize, and model with data.

Tycho Brahe

- 1546–1601
- Collected rich data set of motion of planets
- Noticed inconsistencies of model at the time with what he observed
- Realized as a teenager to collect data to make a science of data
- Had a "science island" between Copenhagen and Sweden
- Knew the value of his data, didn't let other people (Kepler) have full access!



TABVLÆ

RUDOLPHINÆ,

QVIBVS ASTRONOMICÆ SCIENTIÆ, TEMPO-rum longinquitate collapsæ RESTAVRATIO continetur;

A Phœnice illo Astronomorum

TYCHONE,

Ex Illustri & Generosa BRAHEORUM in Regno Daniæ familia oriundo Equite;

PRIMUM ANIMO CONCEPTA ET DESTINATA ANNO CHRISTI MDLXIV: EXINDE OBSERVATIONIBUS SIDERUM ACCURATISSIMIS, POST ANNUM PRÆCIPUE MDLXXII, QUO SIDEREO CASSIOPEIÆ CONSTELLATIONE NOVUM EFFULSIT: SERIÒ AFFECTATA; VARIISque operibus, cùm uraniæ, cùm librariæ, impenso patrimonio amplissimo, accedentibus etiam fœlicibus FRIDERICI II. DANIÆ REGIS, regali magnificentiâ digni, tractu peractus XXV. per insulam sitam SVENDICI HUENNA, & arcem URANIBURGO, in hos usus fundatam exercitiis:

TANDEM TRADUCTA IN GERMANIAM, INQUE AULAM ET Nomen RUDOLPHI Imp. anno MDIIC.

TABULAS IPSAS, JAM ET NUNCUPATAS, ET AFFECTAS, SED MORTE AUTHORIS SUI ANNO MDCI. DESERTAS,

JUSSU ET STIPENDIIS FRETUS TRIUM IMPPP.

RUDOLPHI, MATTHIÆ, FERDINANDI,

ANNITENTIBUS HÆREDIBUS BRAHEANIS; EX FUNDAMENTIS OBservationum stellarum adexemplum sitæ partium jam exstellarum; continuâ multorum annorum speculatione, & computationibus, primùm PRAGÆ Bohemorum continuavit; deinde LINCII, Superioris Austriæ Metropoli, subsidiis etiam II. Provincialium adjutus, perfecit, absolvit; adq; canonum & calculi præsentis formulam traduxit.

IOANNES KEPLERUS,

TYCHONI primùm à RUDOLPHO II. Imp. adjunctus calculi minister indictis, Trium ordine Imppp. Mathematicus:

Qui idem de speciali mandato FERDINANDI II. Imp. præsentibus inhærebat Hæredibus,

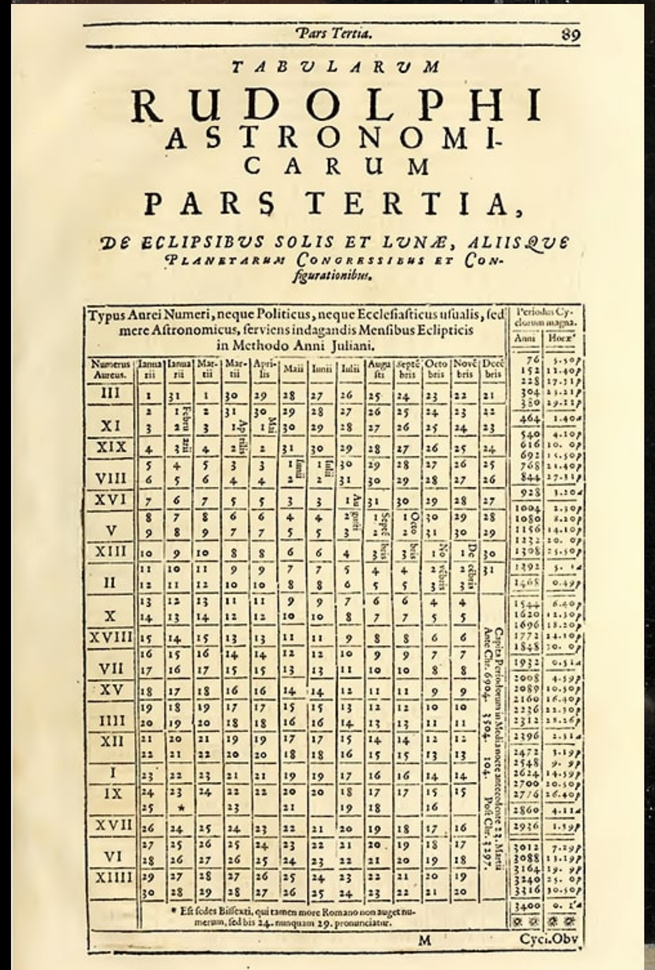
Opus hoc ad usus præsentium & posterorum, typis, numericis propriis, cæteris, & præli JONÆ SAUBII, Reip. Ulmanæ Typographi, in publicum extudit, & Typographicis operis VLMÆ curator assist.

Cum Privilegiis, Imp. & Regum Rerúmq; publ. vivo TYCHONI ejúsq; Hæredibus, & speciali Imperatoris, ipsi KEPLERO concesso, ad Annos XXX.

ANNO M. D C. XXVII.

Johannes Kepler

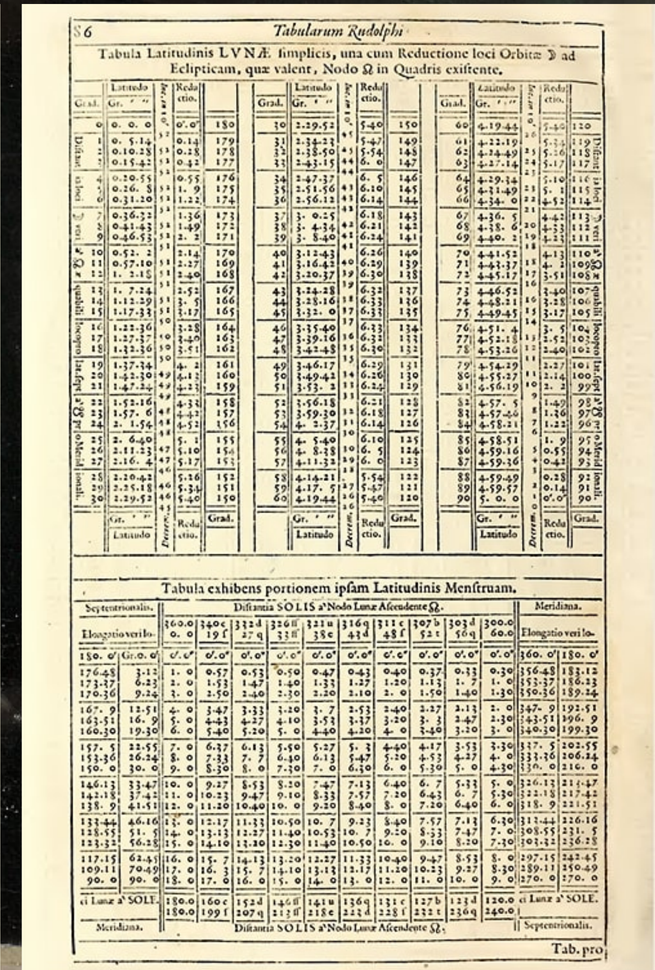
- 1571–1630
- An assistant to the Tycho Brahe
- Discovered three (geometrical) laws of planetary motion:
  - Planets move in elliptical orbits with the Sun as a focus
  - A planet covers the same area of space in the same amount of time no matter where it is in its orbit
  - A planet's orbital period is proportional to the size of its orbit


Pars Tertia. 89

TABULARUM
RUDOLPHI
ASTRONOMI-
CARUM
PARS TERTIA,

DE ECLIPSIBUS SOLIS ET LUNÆ, ALIISQUE PLANETARUM CONGRESSIBUS ET CONfigurationibus.

Typus Aurei Numeri, neque Politicus, neque Ecclesiasticus usualis, sed mere Astronomicus, serviens indagandis Mensibus Ecliptici in Methodo Anni Juliani.

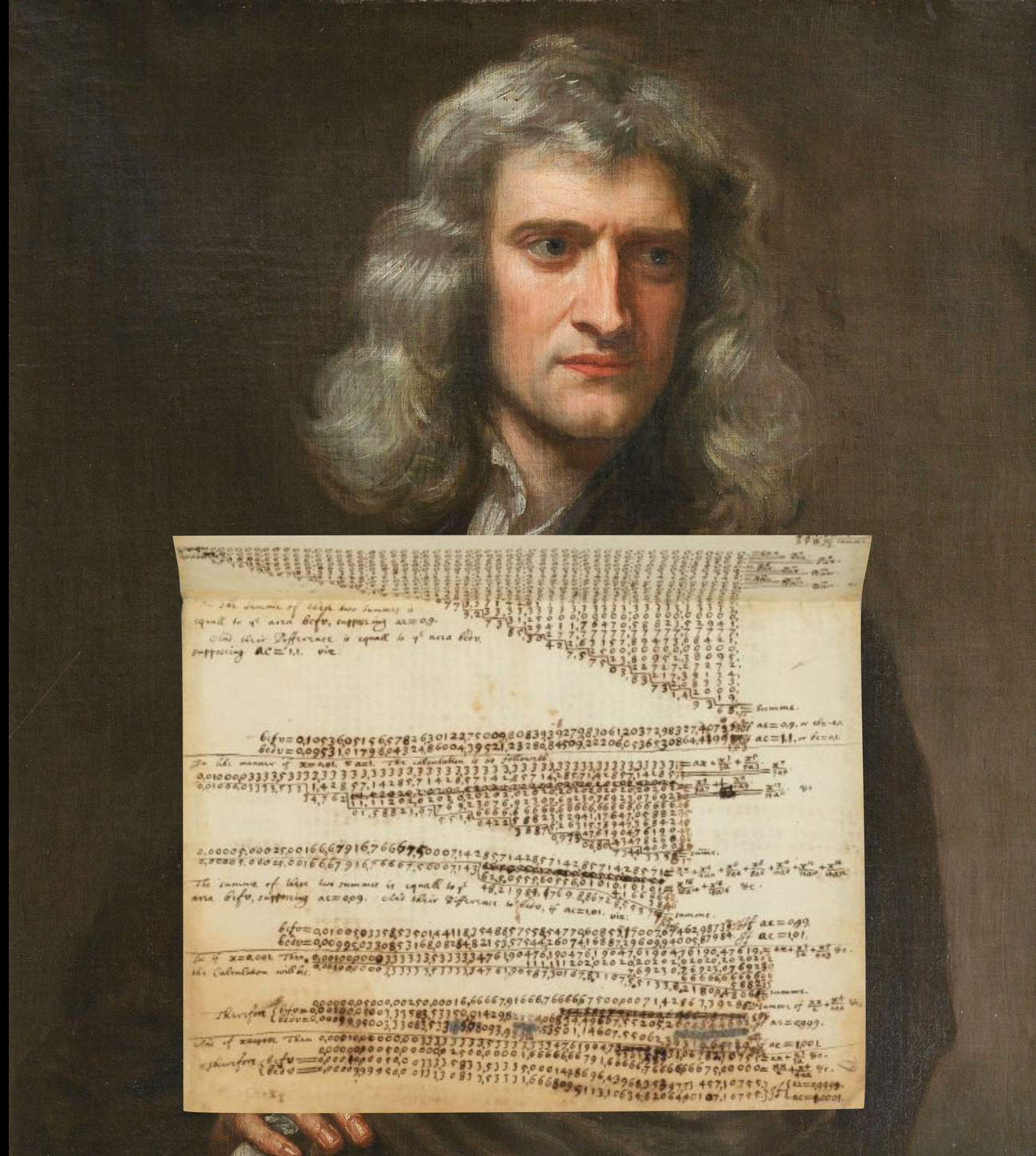

S6 Tabularum Rudolphi

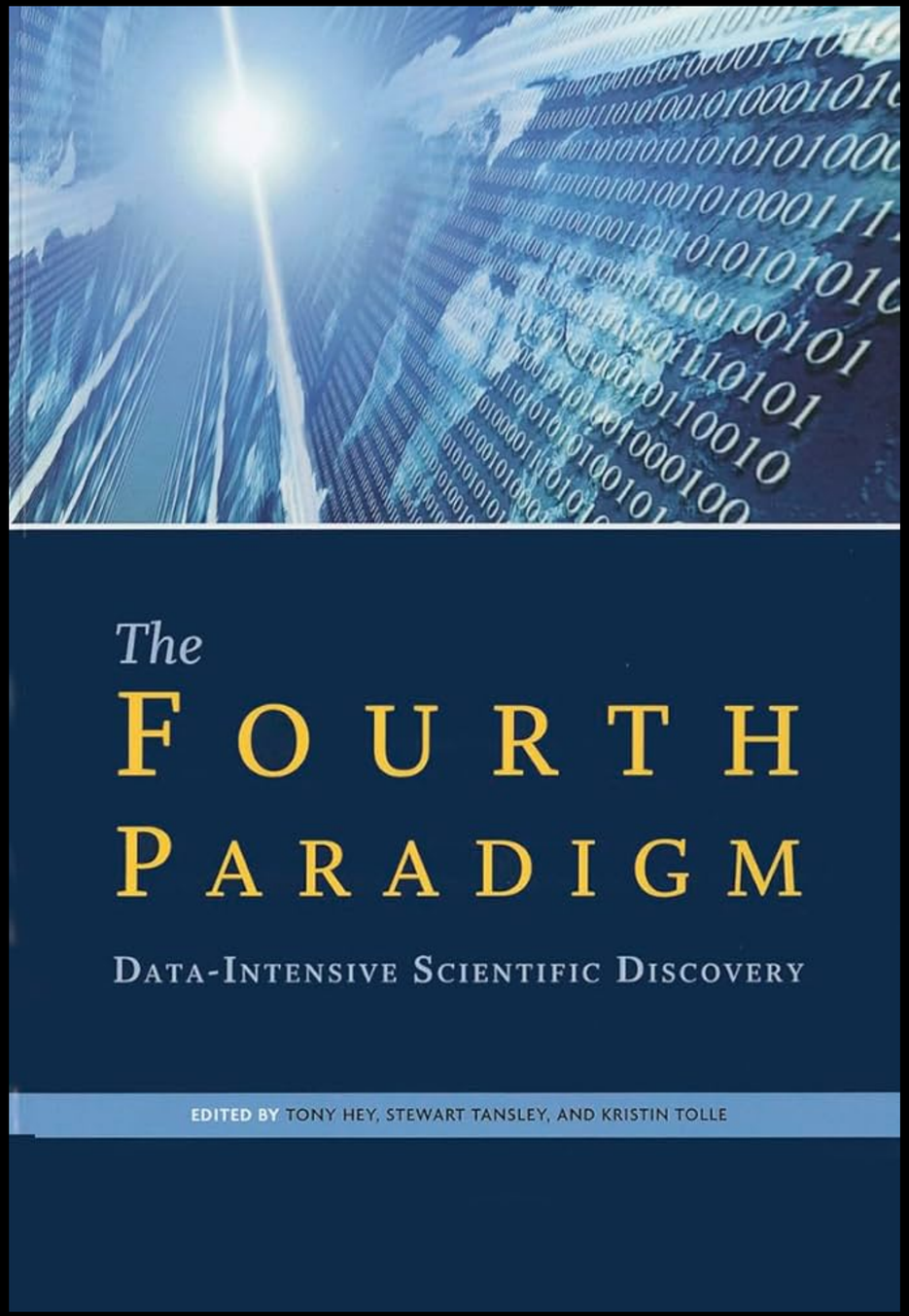
Tabula Latitudinis LVNÆ simplicis, una cum Reductione loci Orbitæ ☽ ad Eclipticam, quæ valent, Nodo ☊ in Quadris existente.

Tabula exhibens portionem ipsam Latitudinis Menstruam.

Sir Isaac Newton

- 1642–1726
- Formulated the laws of motion and universal gravitation
- Derived Kepler's laws of planetary motion
- Explained "why"
- Generalizable; enabled us to put people on the moon
- "Higher order" theory
- In data-driven science, in astrophysics, maybe an ultimate goal is push **from Kepler to Newton**





Experiments/
Observations

Data Science

Science

Theory

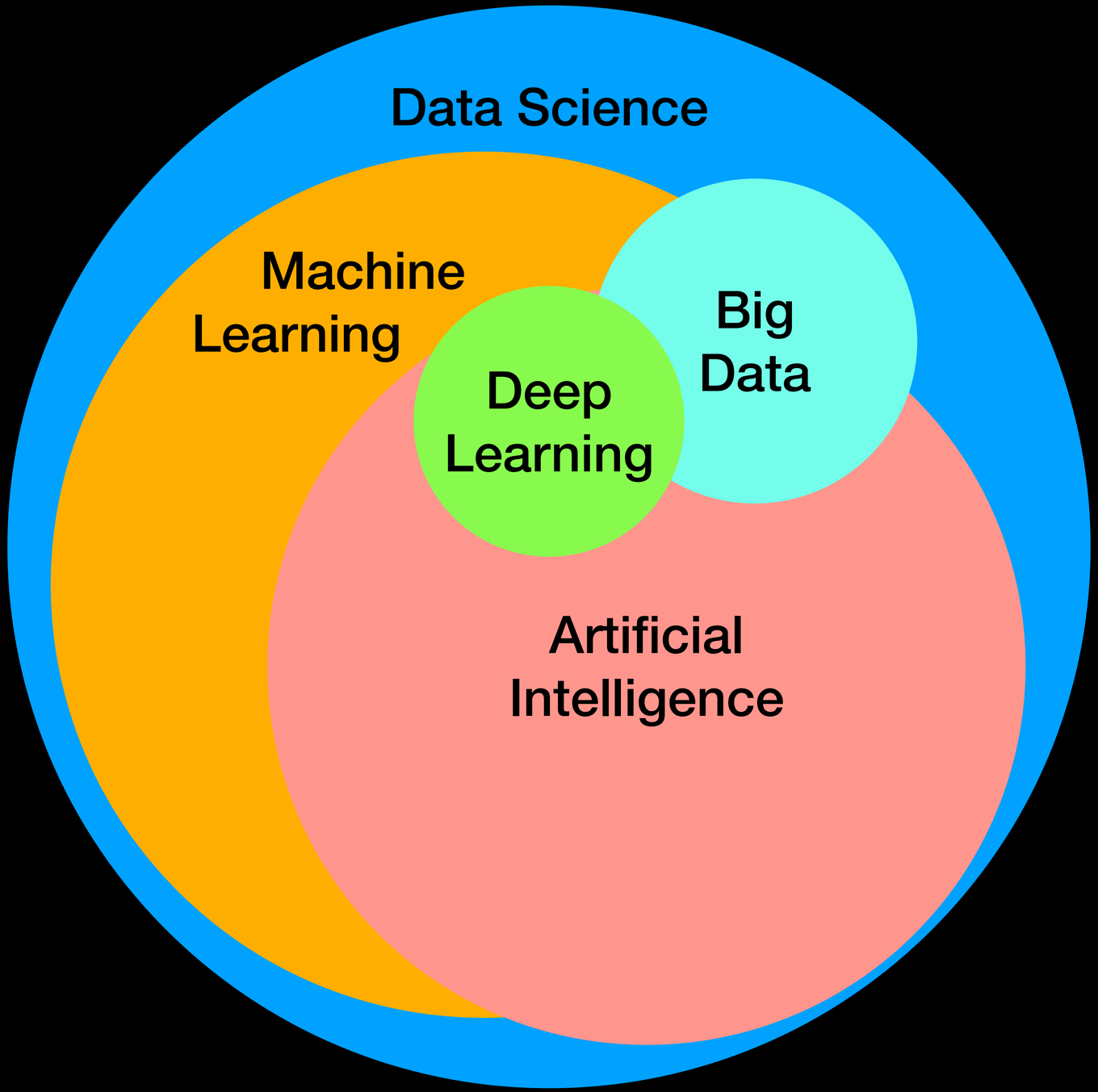
Computational
Science

Computational Science

- “Traditional” simulation
- Forward problem:
  - model -> solution
- Equations is (relatively) simple
- Algorithms are well understood, often designed to preserve equation properties
- Complexity is in the many DoF and non-linearity
- Codes are usually developed by a small team in University or national labs

Data Science

- “Cool” machine learning
- Inverse problem:
  - data -> model
- Data is usually messy
- The most powerful algorithms are less well understood, as long as “they work”
- Complexity is in the coefficients learned from the data
- Codes are developed by large communities, many support by big tech (Tensorflow, PyTorch, JAX)


Data Science
Machine Learning
Deep Learning
Big Data
Artificial Intelligence

Advancement of Machine Learning

- Some of the most exciting applications of deep learning
  - DQN Breakout
  - AlphaGo (core algorithm mctx, documentary)
  - AlphaZero
  - AlphaFold3
  - DALL-E3
  - ChatGPT
  - Segmenting Anything 2
  - AlphaGeometry
  - GraphCast
- What will the “ChatGPT moment” be like in astronomy?

Machine Learning

- Building useful models from data
- Models hopefully improve with increase volume of data
- Misconcept 1: ML is magic
- Reality: built on foundational advances in math and computer science in decades:
  - optimization & regression
- Most useful background fields: linear algebra, optimization, statistics
- Misconcept 2: ML is just Neural Network (Deep Learning)
- Reality: outside NN/DL, there are many ML techniques that are:
  - effective, efficient, and very often understandable/interpretable

Machine Learning Goals

- Model be "useful" in the future
- Accurate
- Fast
- Generalizable (f = ma)
- Interpretable/explainable
- Certify/guarantee

Cross-Validation

- Past: training data
- Future: testing data
- Split available to 80% for training and 20% for "testing"
- K-fold validation for increase statistics
- **Never trust/publish any ML results without cross-validation!!!**

Supervised

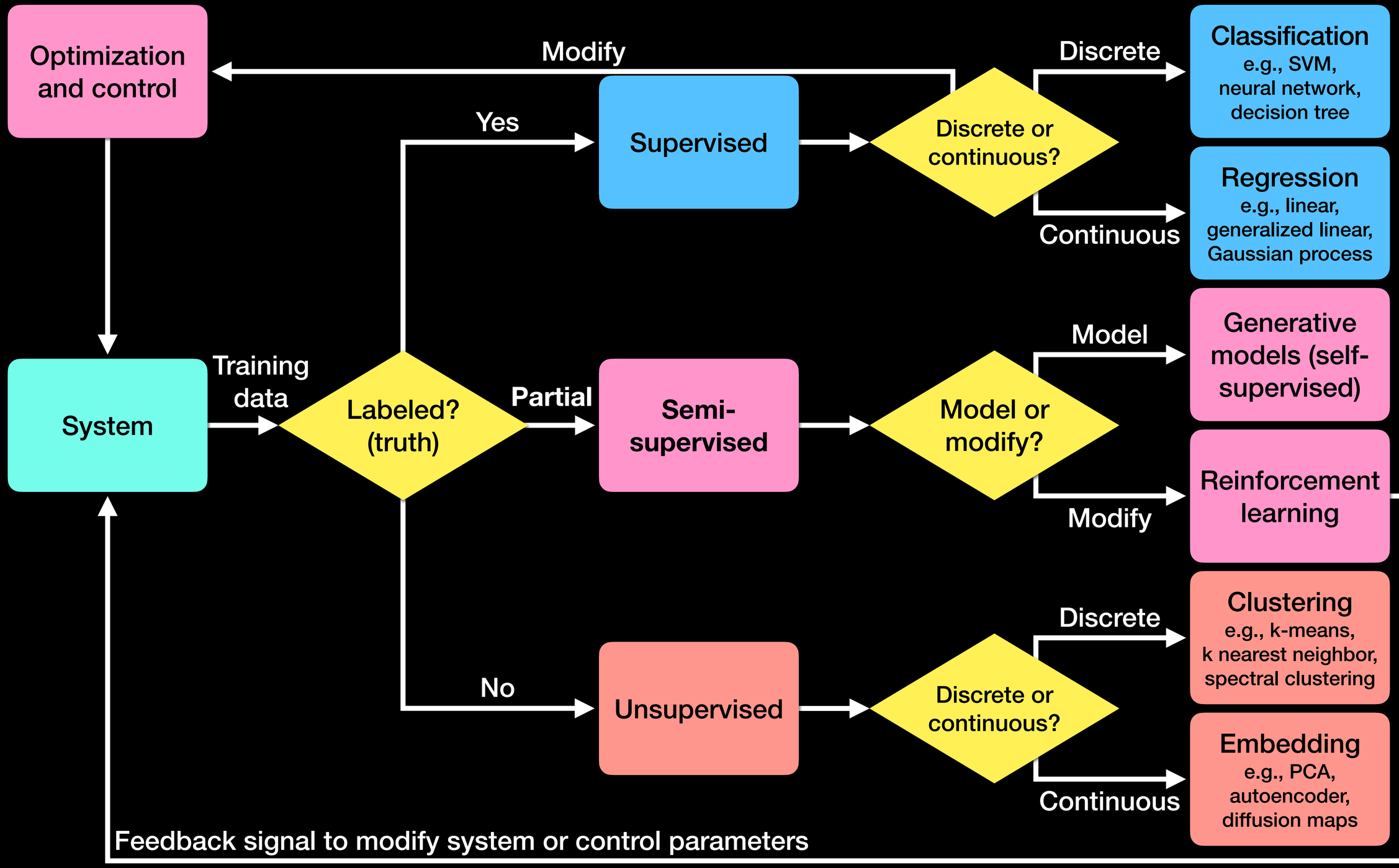
- With labels
- Classification (discrete): classify type I and II supernovae
  - SVM
  - neural network
  - decision tree
- Regression (continue): light curve of AGNs
  - linear,
  - generalized linear,
  - Gaussian process
- Optimization and control

Semi-supervised

- Some feedbacks
- Generative models (self-supervised)
  - Generative adversarial networks (GAN)
- Reinforcement learning
  - Deep reinforcement learning

Unsupervised

- No labels, "data mining"
- Cluster (discrete)
  - k-means
  - k nearest neighbor
  - spectral clustering
- Embedding, dimensionality reduction, pattern extraction
  - PCA
  - autoencoder
  - diffusion maps


Optimization and control
Modify
Discrete
Classification
e.g., SVM, neural network, decision tree
Yes
Supervised
Discrete or continuous?
Regression
e.g., linear, generalized linear, Gaussian process
Continuous
Model
Generative models (self-supervised)
System
Training data
Labeled? (truth)
Partial
Semi-supervised
Model or modify?
Reinforcement learning
Modify
Discrete
Clustering
e.g., k-means, k nearest neighbor, spectral clustering
No
Unsupervised
Discrete or continuous?
Embedding
e.g., PCA, autoencoder, diffusion maps
Continuous
Feedback signal to modify system or control parameters

Resources

- Scikit-learn: https://scikit-learn.org
- Keras: https://keras.io
- AstroML: https://www.astroml.org