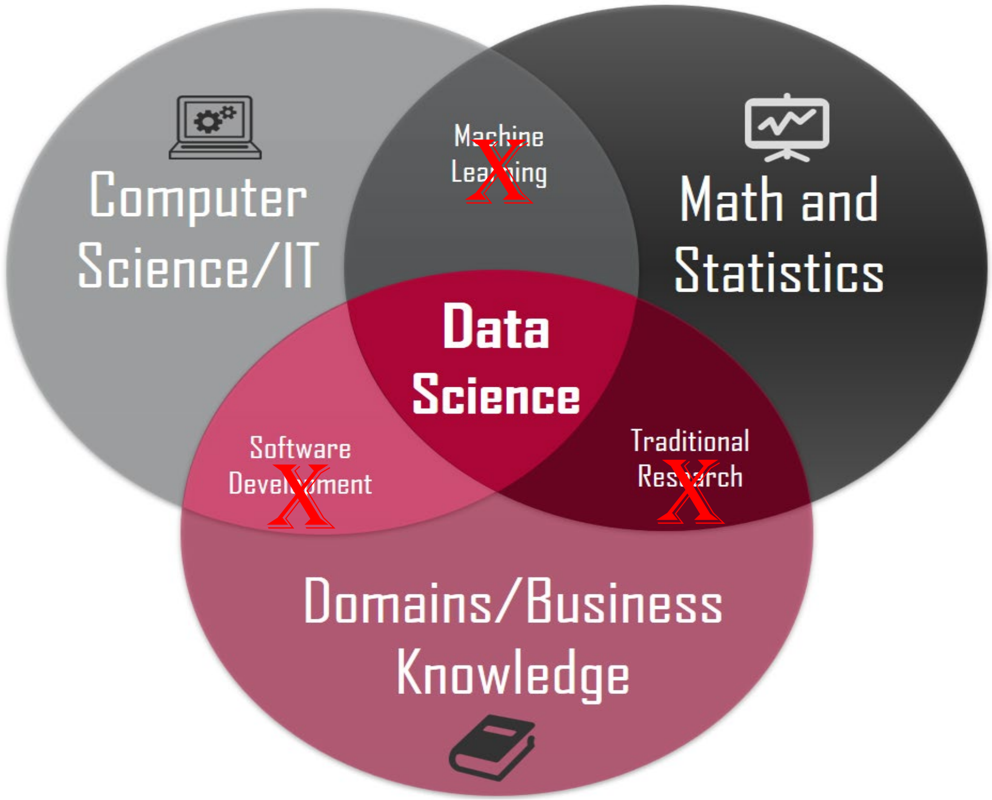
Defining a Field, a School and a Curriculum

Brian Wright
brianwright@virginia.edu

Defining a Field

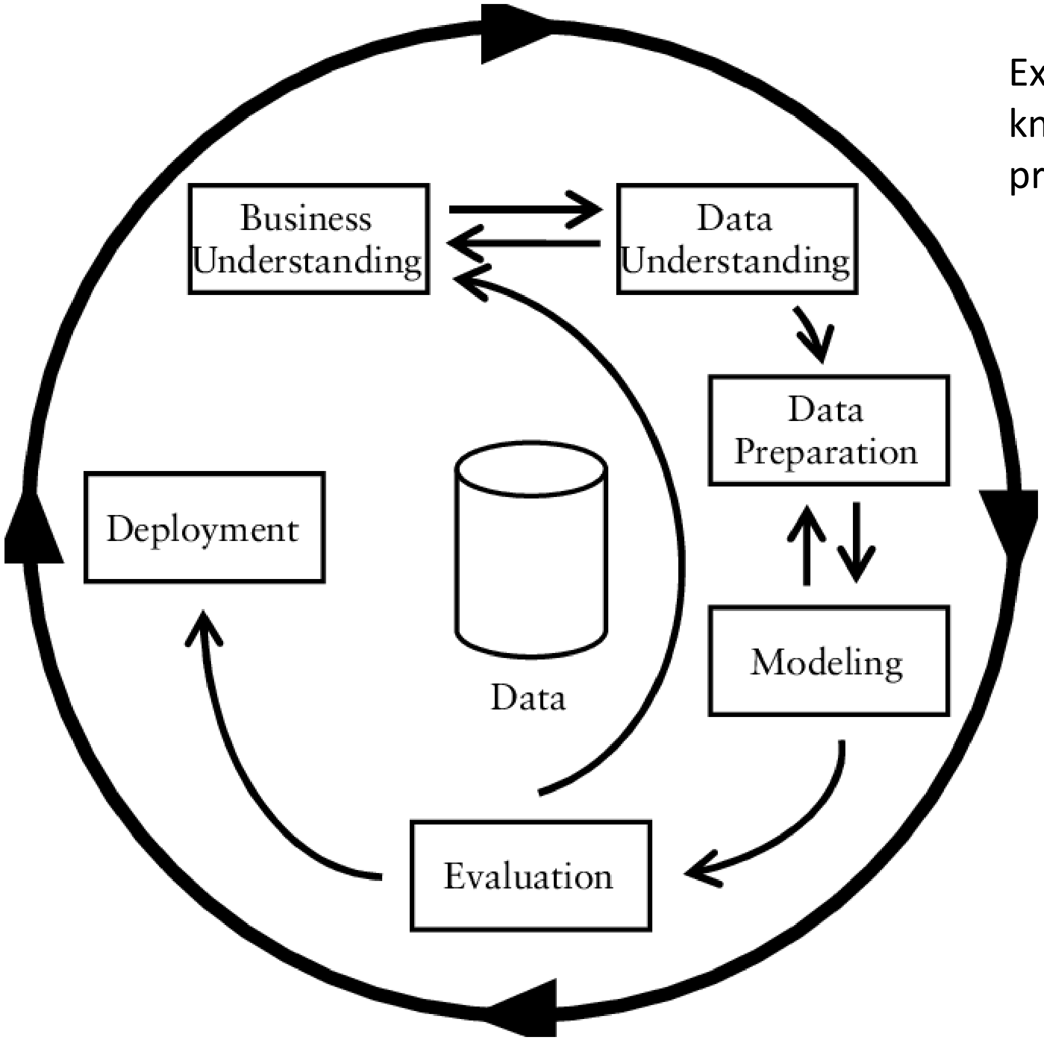
The “Classic” View



Source: https://honingds.com/blog/intro-to-data-science/

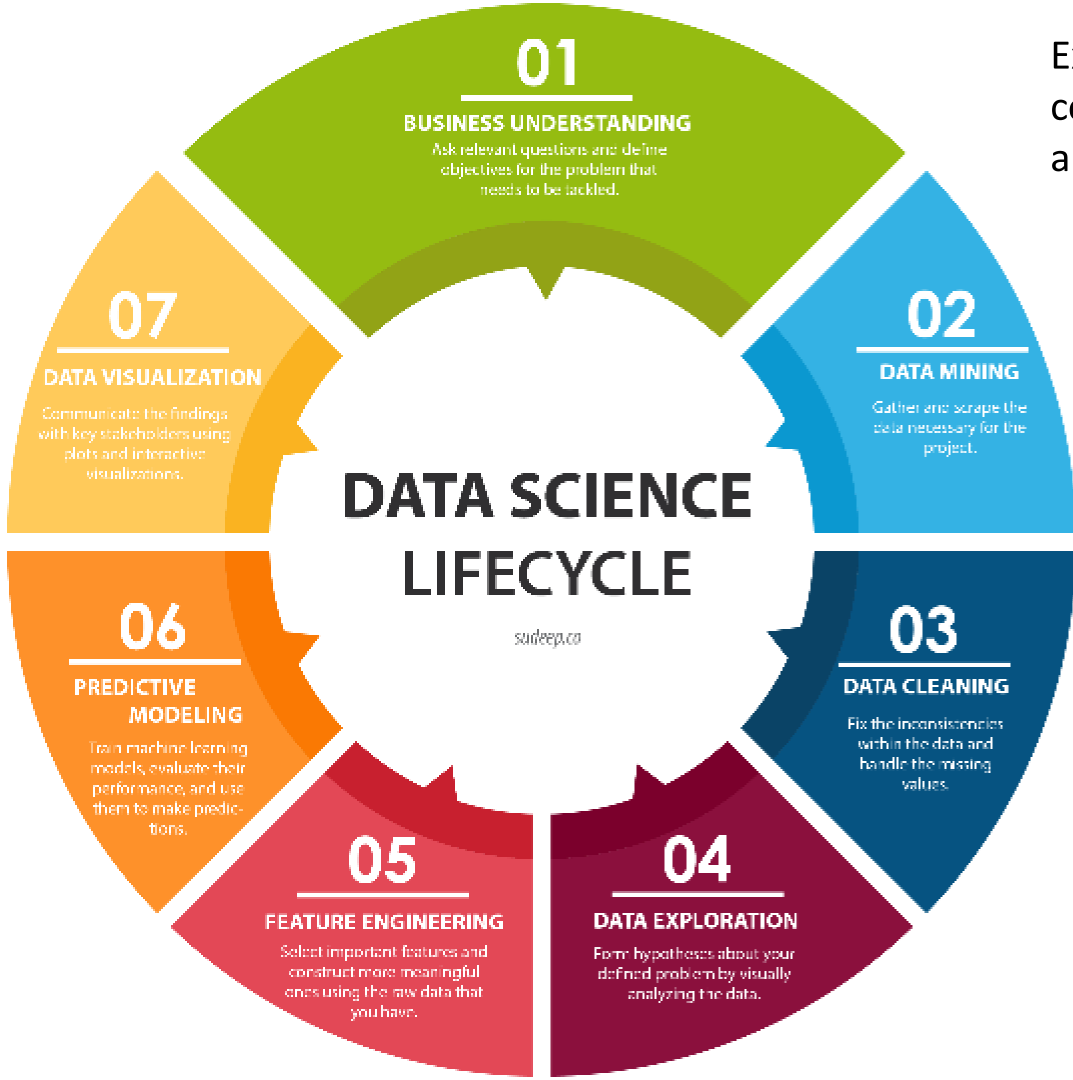
A common reference for the field of Data Science. The “Xs” represent a perceived over extension of this model in that ML does require domain knowledge, CS does conduct traditional research and doing software development of course requires math.

CRISP-DM



Example of a well-known data science process map

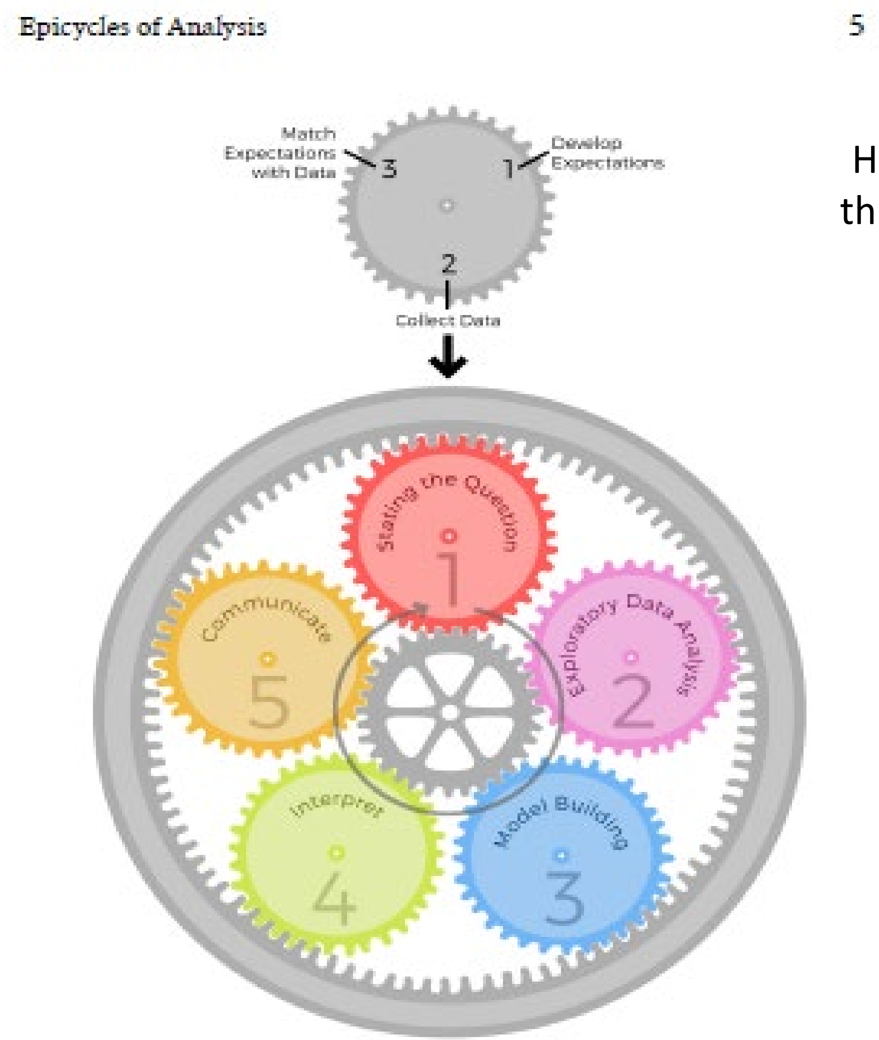
https://www.researchgate.net/figure/CRISP-DM-process-model_fig3_261307514



Example of a more contemporary view of a Data Science process

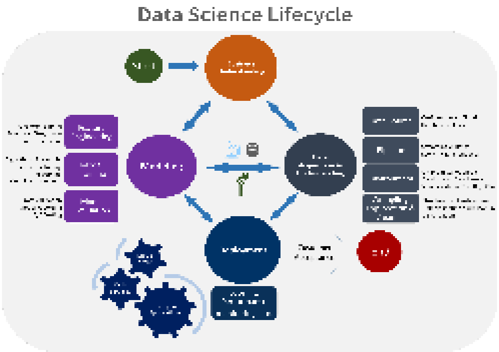
http://sudeep.co/data-science/Understanding-the-Data-Science-Lifecycle/

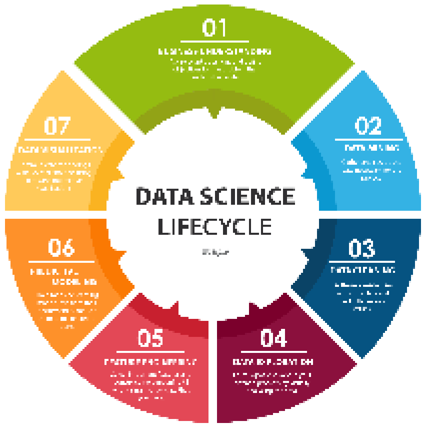
Roger Ping: Art of Data Science

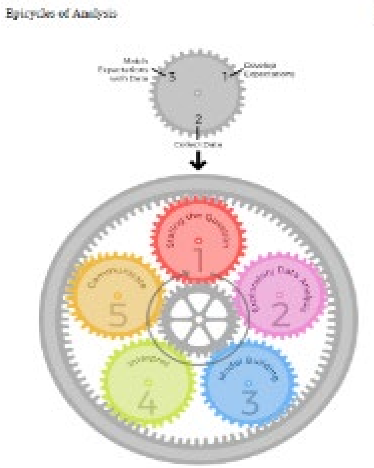


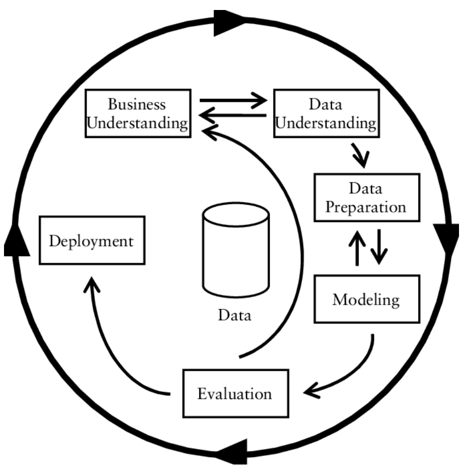
Here a bit more accurate as this reflects the forward and backward process that is typical with most Data Science projects.

Commonality

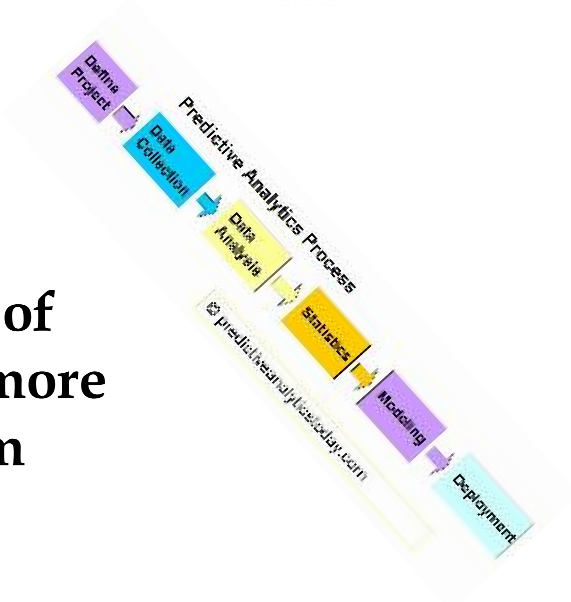






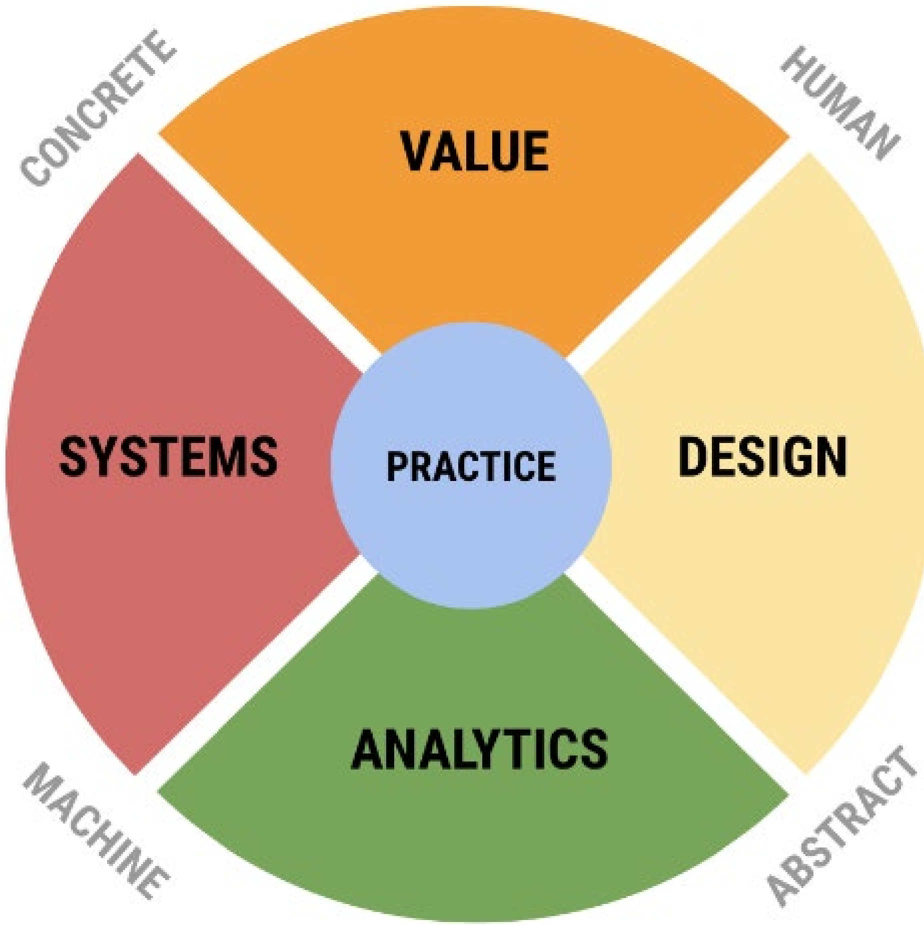


Commonality of all of these is that they are more Function than Form



SDS View on Data Science

UVA SDS needed not just pipeline/process but a conceptual model of field. Rafael Alavardo created this framework of the field.



Design

Human machine interaction as it appears at the points of both consuming data and producing **data products**. Activities here include the representation and communication of human reality as data for the work of analytics, e.g. in database design, the curation of data, and of complex data and analytical results to humans to **drive decision-making and influence behavior.**

- Key tensions: discovery vs product, user-facing vs analyst-facing.
- Common theme: Communication.
- Keywords: communication, design thinking, representation, use interface, communication, visualization, visual and object languages, informatics, ontology, curation, HCI, informating, data product design, data discovery.
- Values: openness, authenticity, beauty, form and function.

Systems

Infrastructure systems and architectures to **support working with big data** — big in terms of volume, velocity, and variety — and building high performance pipelines in both development and production environments. It includes the broad areas of **hardware and software (programming)** as such — computer technology as opposed to computer science.

- Key tensions: development vs production, volume vs speed.

- Common theme: building

- Keywords: infrastructure, data systems, data engineering, the cloud, networks, hardware, software, programming languages, big data management, benchmarking, continuous integration, availability, cybersecurity.

- Values: speed, stability, robustness, resilience, uptime.

Analytics

Analytics includes what most consider to be the heart of data science, the combination of **statistical methods with machine learning**, along with optimization, signal processing, network analysis, and other rigorous quantitative methods from a variety of fields.

- Key tensions: inference vs prediction, analysis vs simulation.
- Common theme: Mathematical models and methods.
- Keywords: prediction, inference, machine learning, statistics, operations research, AI, experimental design, causality, optimization, knowledge, models, feature engineering, data mining.
- Values: accuracy, precision, validity, truth, convergence, explainability.

Value

Combines the traditional **discipline of ethics** with the professional activities of **business planning**, policy making, developing motivations for scientific research, and other activities that have a direct impact on **people** and the planet.

- Key tensions: enterprise vs ethics, private interest vs public good (see Plato's *Republic*).

- Common theme: human value.

- Keywords: ethics, justice, wealth, value, social good, motivation, meaning,

- Values: responsibility, diversity, inclusion, flourishing, excellence

Ref: Digital Epidemiology

Practice

This area consists of actual activities that brings **people together to combine expertise** from each of the four areas. It is characterized by data science teams working together and with external parties to develop solutions and projects that are responsible, authentic, effective, and efficient.

Many activities that are considered an essential part of data science, such as **data wrangling, actually exist only in practice**, combining expertise in systems, design (representation), and analytics, and they are not usually taught in distinct classes.

Practice is also where the core areas of data science come into contact with **domain knowledge and real world problems**.

Mapping of Categories

1. Data description and Curation →	Design
2. Mathematical Foundations →	Analytics
3. Computational Thinking →	Design and Analytics
4. Statistical Thinking →	Analytics
5. Data Modeling →	Design
6. Communication, reproducibility, and ethics →	Design, Value

Defining a School

Structure of the School (Academic)

- Areas Practice
- Collaboratories

Structure of the School

- Areas of Practice – They differ from departments in that they have a **five-year life span** in keeping with the dynamic nature of the field and are considered by SDS leadership for renewal or modification based on recommendations of an external committee to the SDS towards the end of their fourth year. They are established and retired in alignment with **the goals** of the SDS.
  - Director, deputy director, faculty and administrators (budgets, goals, etc.)

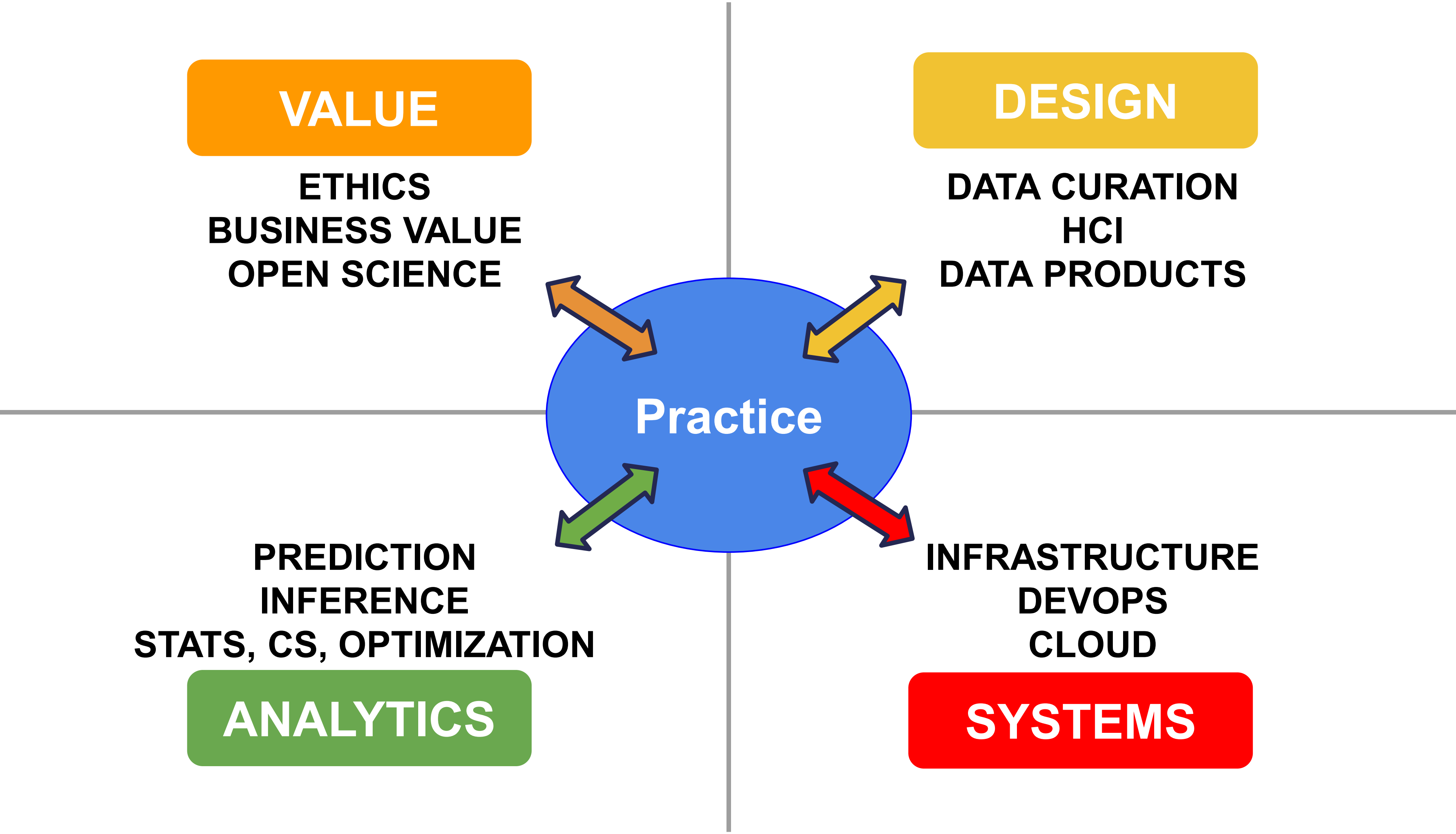
Structure of the School

- Collaboratories – Each satellite exists as a physical presence in another school and brings **data science expertise** to that school.

  - Collabs will be directed by a team member (50:50 appointment) who is responsible for developing the satellite to **meet the needs of the discipline** while at the same time ensuring that best practices, data, analytics etc. Each satellite is expected to **evolve differently** to meet the respective needs of the school.

  - Faculty director, joint appointed faculty, post docs, PhD students, etc.

Defining a Curriculum (Undergrad)


VALUE
ETHICS
BUSINESS VALUE
OPEN SCIENCE
DESIGN
DATA CURATION
HCI
DATA PRODUCTS
Practice
PREDICTION
INFERENCE
STATS, CS, OPTIMIZATION
ANALYTICS
INFRASTRUCTURE
DEVOPS
CLOUD
SYSTEMS

Data Science Minor (Currently in Place at UVA)

1. Programming for Data Science SYS
2. Intermediate Analytics Course A
3. Introductory Systems/Infrastructure SYS Des
4. Data Design or Value V Des
5. Capstone/Data Driven Domain Course PRA

Emphasis on Systems/Design versus Analytics

Course offering example, not in practice at UVA, (assumes completion of a two-year liberal arts education)

1. Logic and Problem Solving with Data V PRA
2. Programming for Data Science SYS
3. Computer Science Foundations SYS Des
4. Data Ethics V
5. Practice of Data Science V A SYS Des
6. Data Warehousing SYS Des
7. Data Mining A SYS
8. Data Communications Des
9. Cloud Computing SYS

Again, emphasis on "Systems/Design" components

Upper Level, Electives and Concentrations (Example)

1. Upper Level Courses (General Application and Data Specific Courses)
   a. General Application (Required, 9 credits)
      i. Machine Learning I and II A
      ii. Bayesian ML A
      iii. Text Mining A

   b. Data Electives (Select 2, 6 credits)
      i. Natural Language Processing II A
      ii. Time Series/Forecasting A
      iii. Spatial Data/GIS Modelling A
      iv. Image Modelling A
      v. Streaming Data (Social Media/IOT/Real-Time) A
      vi. Human Centered Design Des

   c. Domain Concentration (12 Credits) PRA
      i. Focus in a **Center or a "Collaboratory"**
      ii. Focus in a specific foundational element

   d. Capstone (6 credits) PRA

Move towards Analytics and Practice