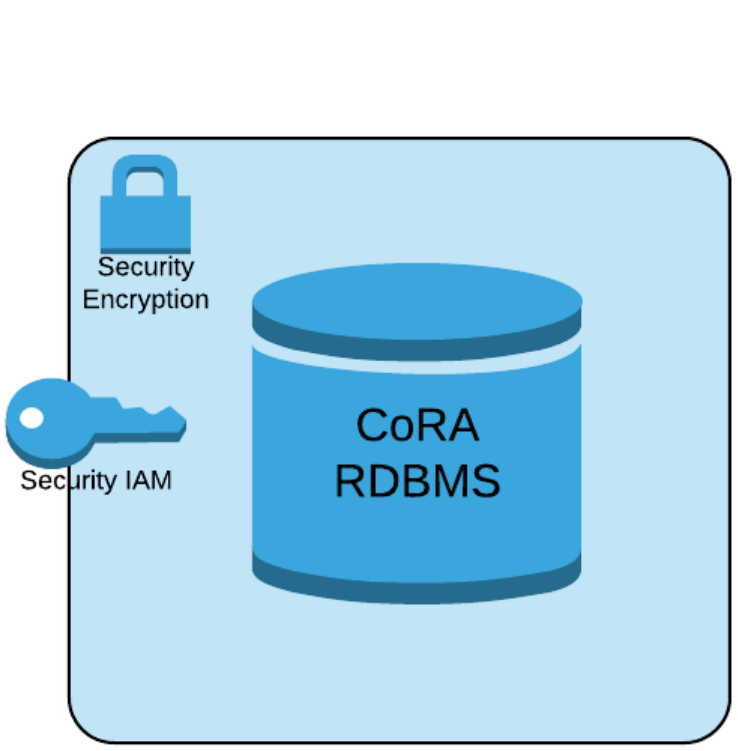
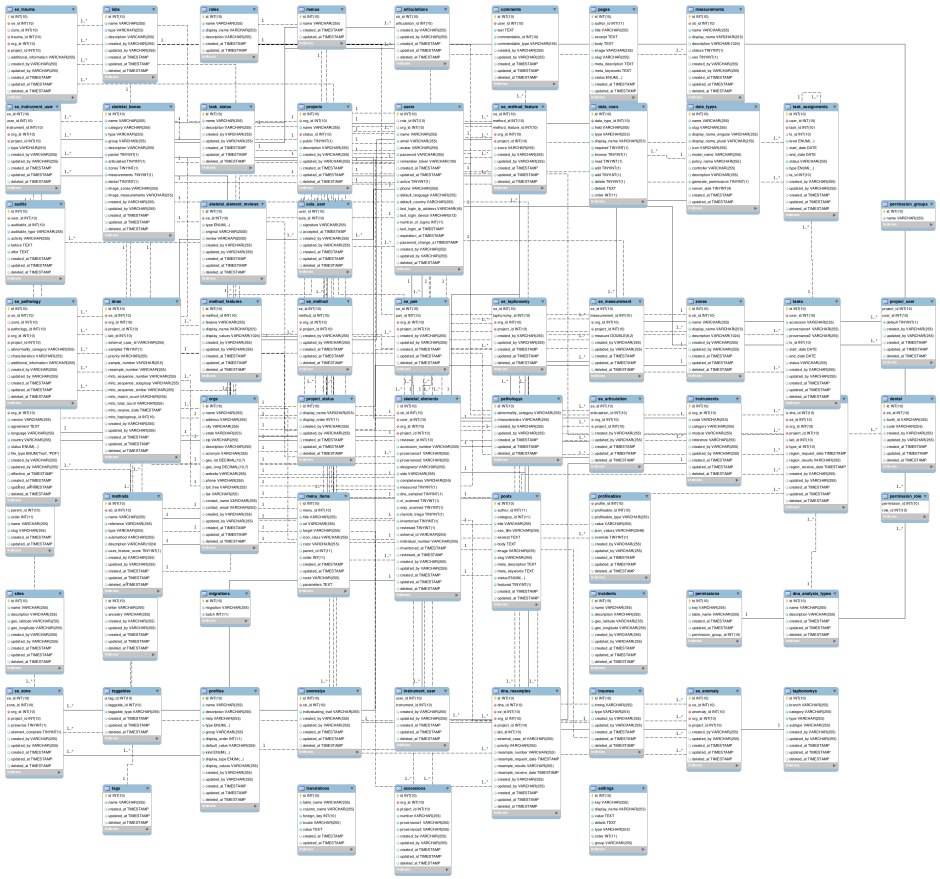
Information Management, Project Pipelines, and the Commingled Remains and Analytics Platform (aka Anthropologist as Data Scientist)

Sachin Pawaskar, PhD, MBA, MS

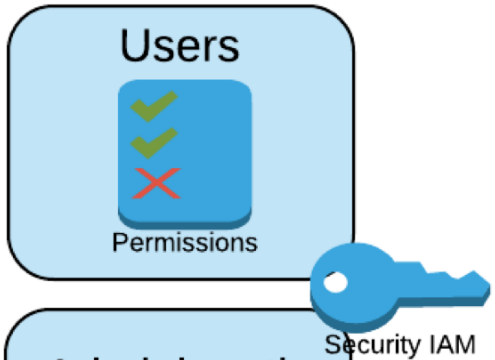
Franklin E. Damann, PhD, D-ABFA

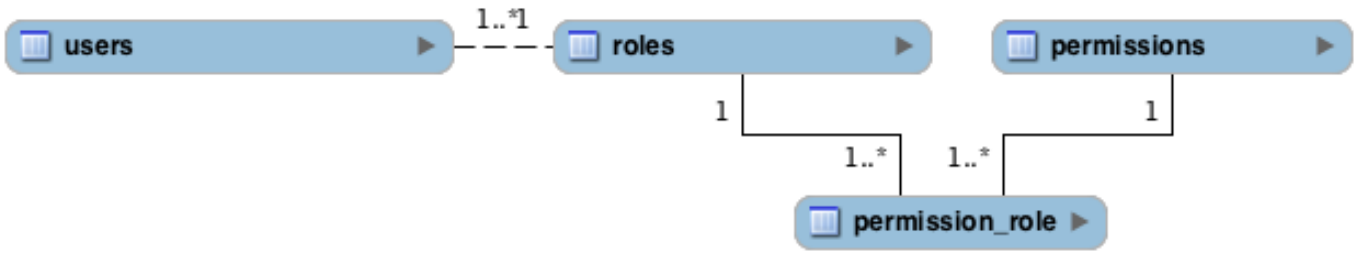
CoRA Database



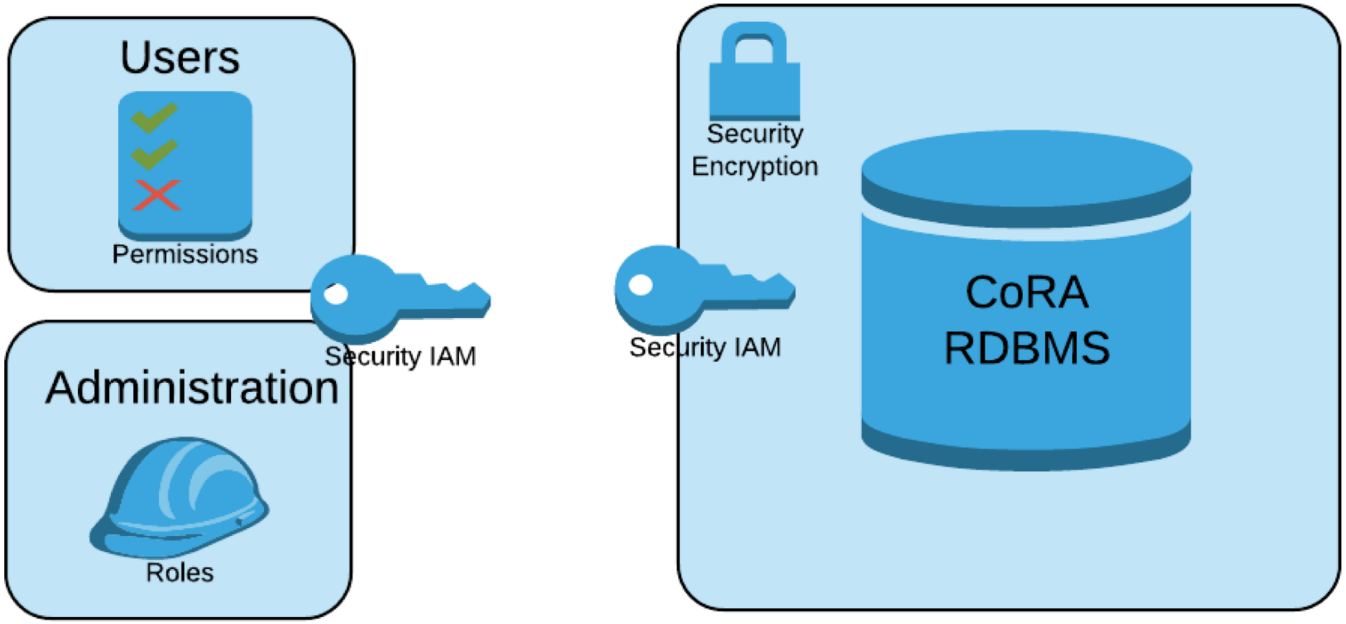


CoRA Database – User Authorization

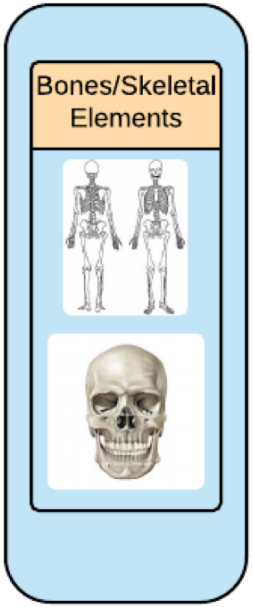


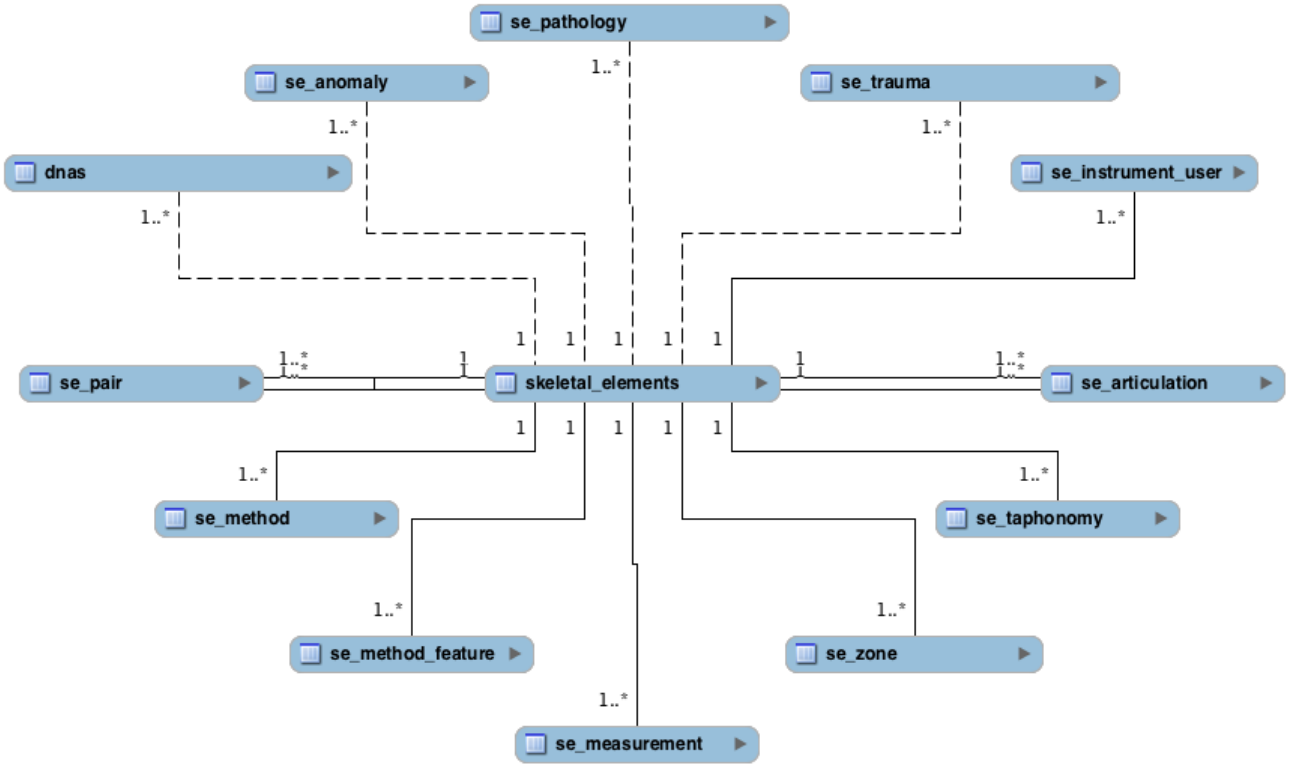


CoRA Database

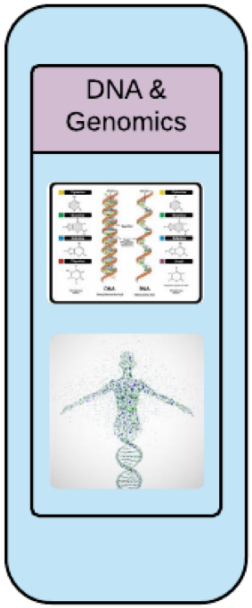


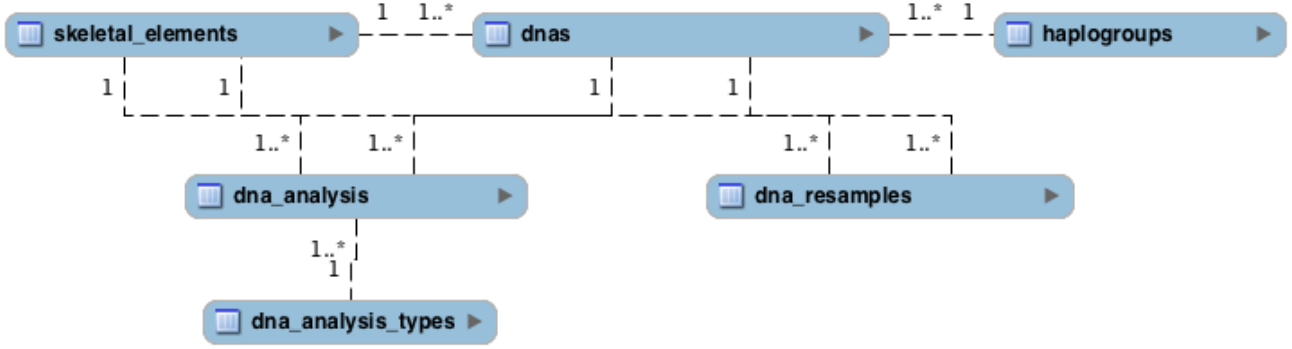
CoRA Database – SE Module



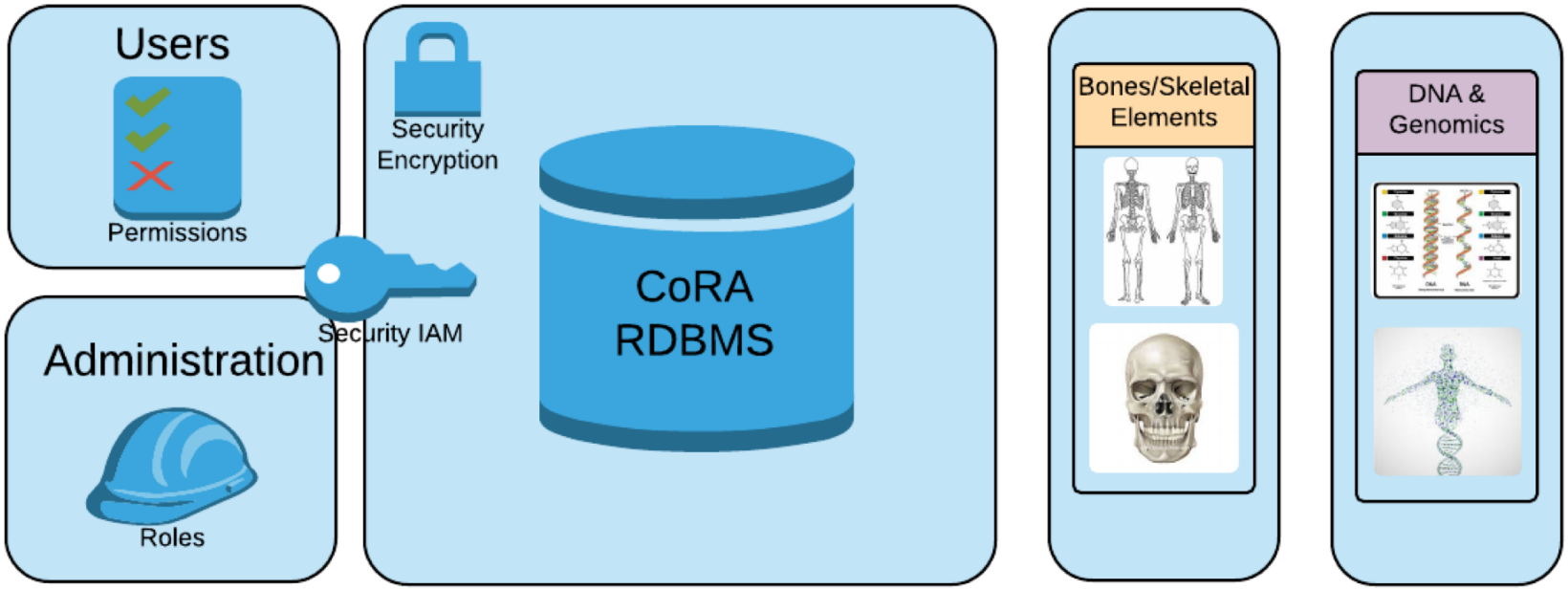


CoRA Database – DNA/Genomic Module

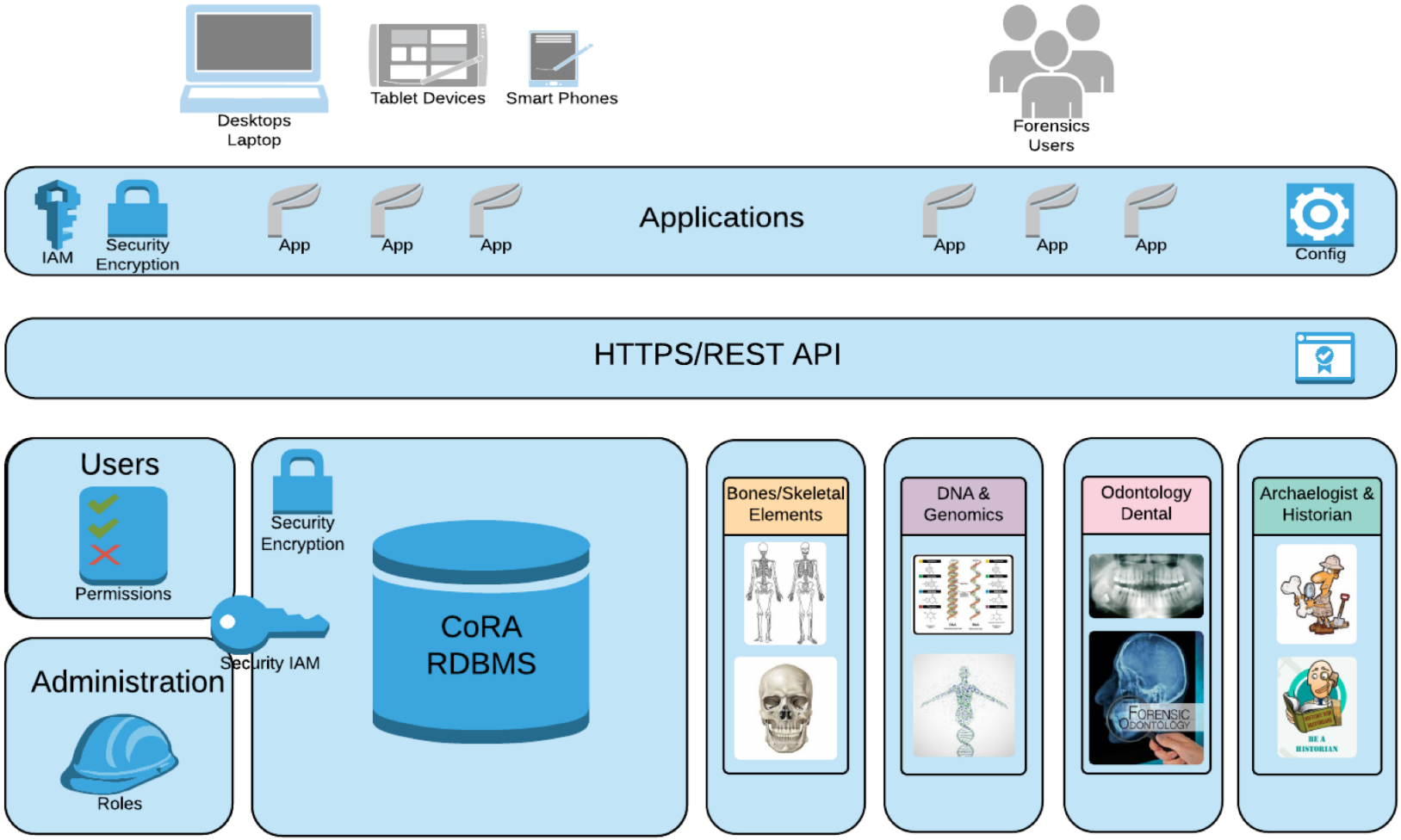




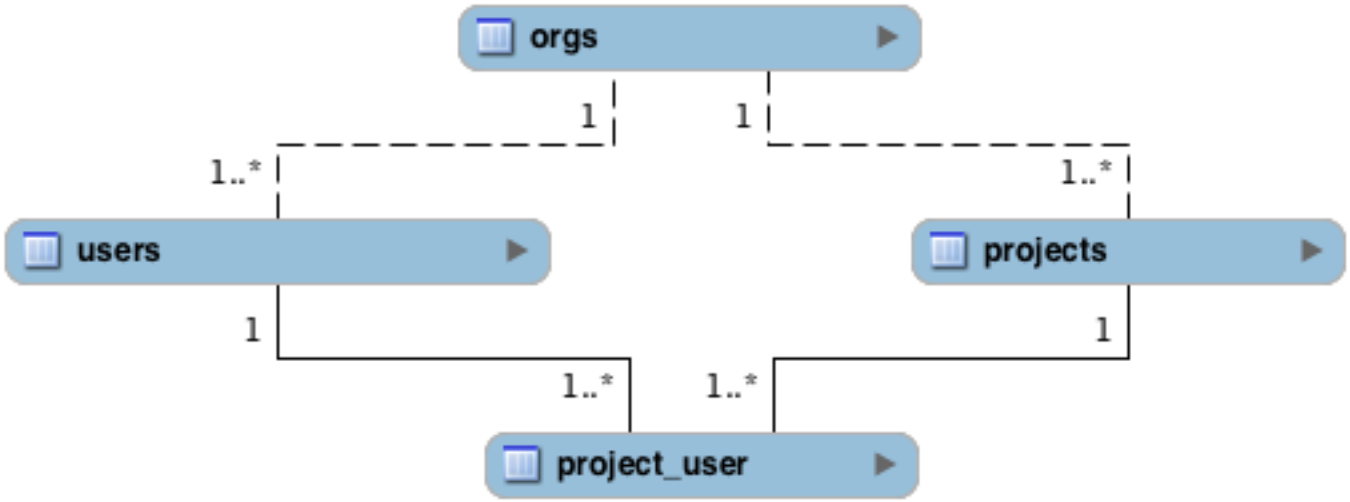
CoRA Database – Modules



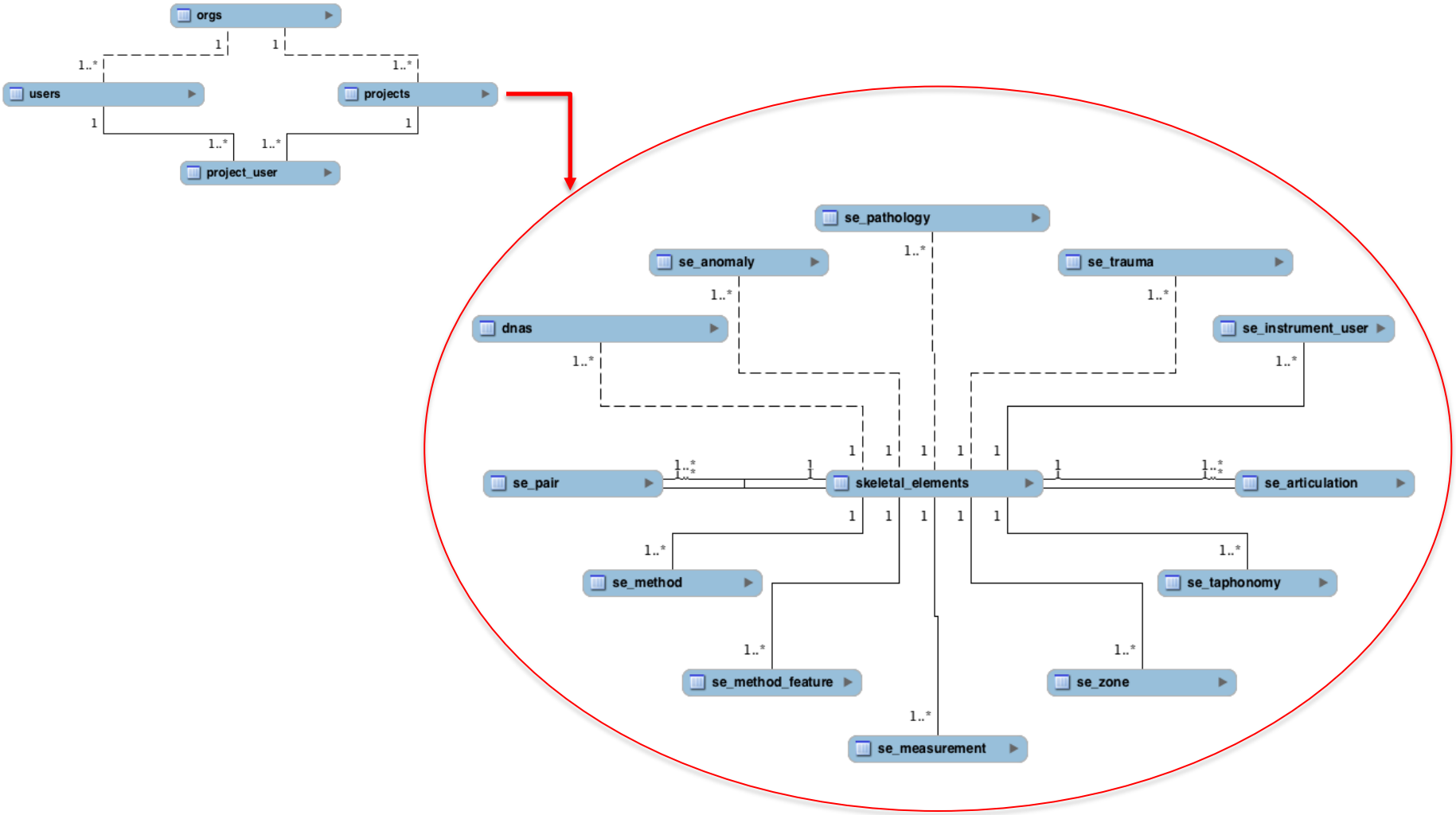
CoRA Database – Modules – HTTPS/REST API & Applications



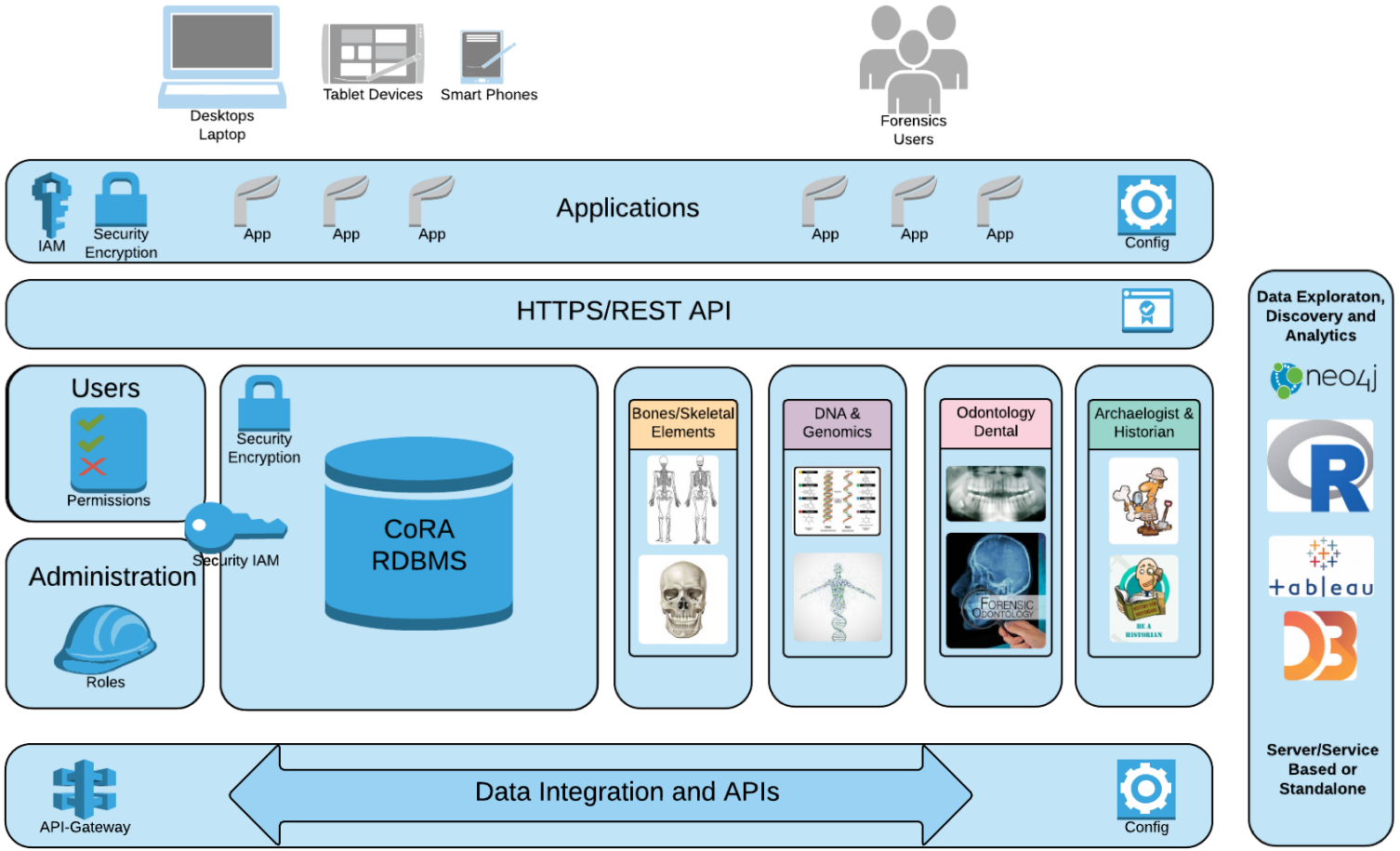
CoRA Database – Orgs-Users-Projects



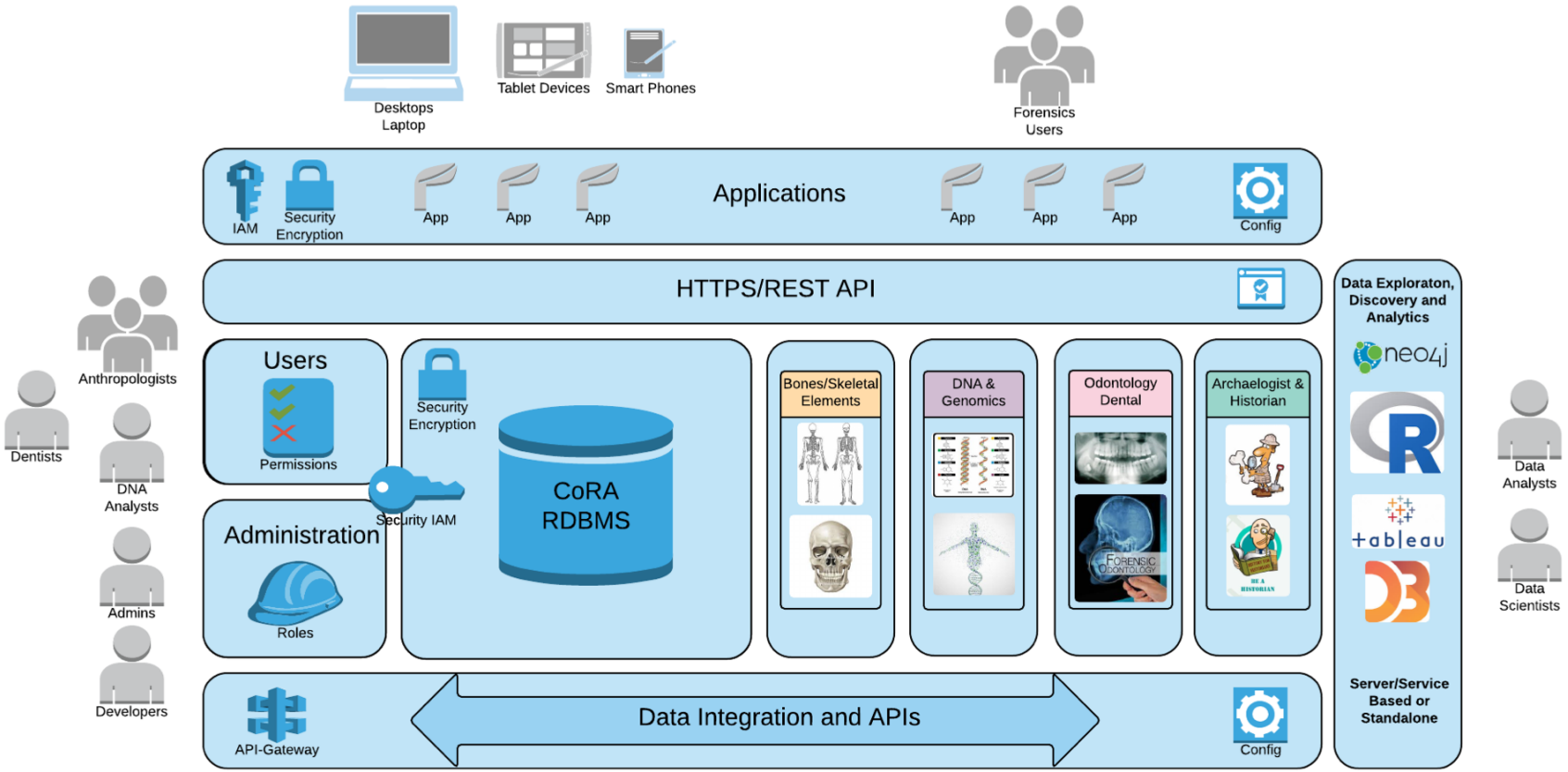
Projects – As Containers for SE



CoRA Database – Modules – HTTPS/REST API & Applications & Data API and Advanced Analytics



CoRA Database – Modules – HTTPS/REST API & Applications & Data API & Advanced Analytics



Anthropology Data

- Does Anthropology Data have Big Data Characteristics?

3 Key attributes for Big Data characteristics

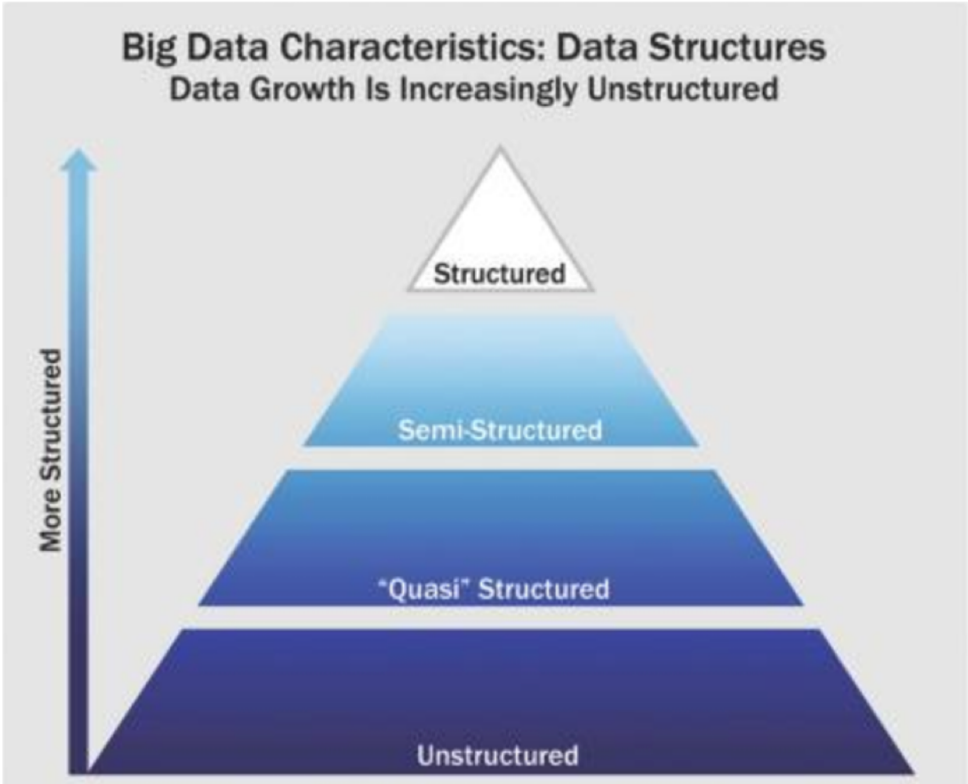
- Huge Volume
  - Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns
- Complexity of data types and structure
  - Variety of new data sources, formats and structures
- Speed of new data creation and growth
  - High velocity with
  - Rapid data ingestion, and
  - Real time analysis
- Other notable characteristics
  - Cannot be efficiently analyzed using traditional databases & methods
  - Require new tools & technologies to store, manage and realize benefits

Big Data is data whose scale, distribution, diversity and/or timeliness require the use of new tools, technical architectures and analytics to enable insights that unlock new sources of business value

– McKinsey & Co, Big Data: The Next Frontier for Innovation, Competition and Productivity

Big Data Characteristics: Data Structure

- 80% - 90% of future data growth coming from non-structured data types

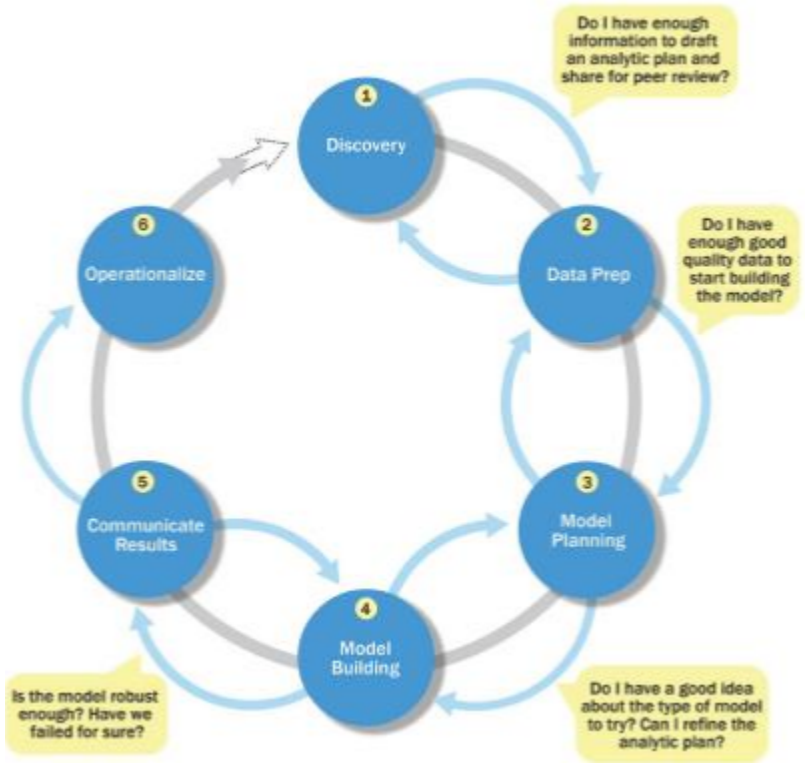


Big Data Characteristics: Data Structure

- Structured
  - Data containing a defined data type, format, and structure
    - Transaction data, OLAP data cubes, traditional RDBMS, CSV files.
- Semi – Structured
  - Textual data files with a discernable pattern that enables parsing
    - XML data files
- Quasi – Structured
  - Textual Data with erratic data formats that can be formatted with effort, tools and time.
    - Web click steams – may contain inconsistencies in value & format.
- Unstructured
  - Data that has no inherent structure
    - Text documents, PDFs, Images, Video

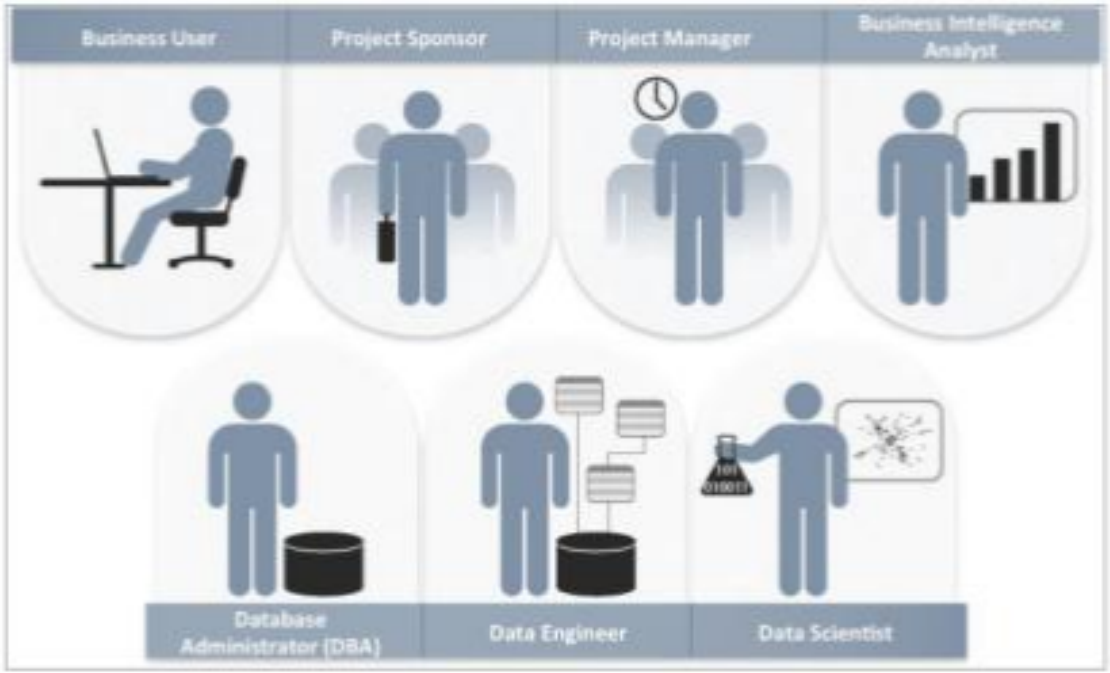
Data Analytics – Life Cycle

- Consists of 6 phases
- Teams commonly learn new things in a phase
  - Causes them to go back and refine work done in prior phases
  - Based on new insights and information that has been uncovered.
- It is an iterative process
  - Until the team members have sufficient information.
- Note that these phases are not formal stage gates.



Successful Analytics Project – Key Roles

- Most of these roles are not new, except for
  - Data Engineer and
  - Data Scientist



- These 2 have become more popular and in high demand due to the growing interest in Big Data.

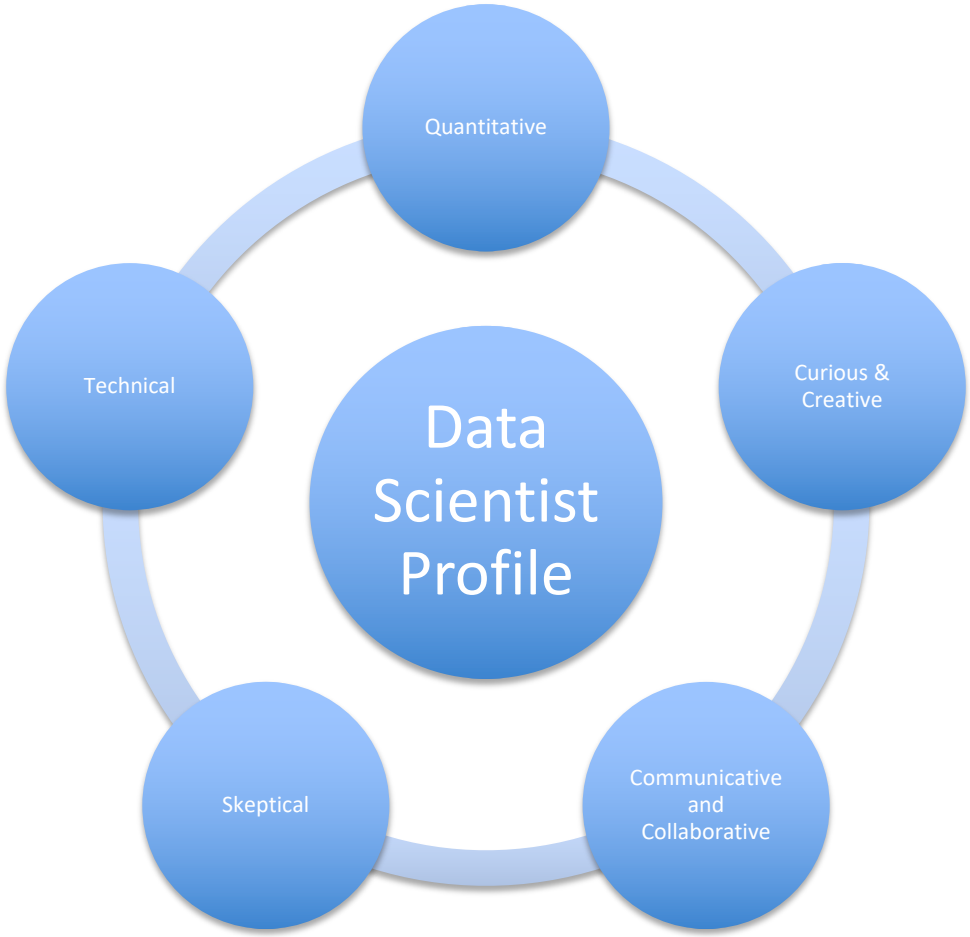
EMC Education Services

Data Scientist – 3 Recurring Activities

- Reframe business challenges as analytics challenge
  - This is a skill to diagnose business problems,
  - Consider the core of a given problem, and
  - Determine which kinds of analytical methods can be applied to solve it
- Design, implement, and deploy models and data mining techniques on Big Data
  - Apply complex or advanced analytical methods to a variety of business problems using data.
- Develop insights that lead to actionable insights and recommendations
  - Draw insights from the data and communicate appropriate recommendations effectively.

Data Scientist – Profile

- 5 main skills and behavioral characteristics
  - Quantitative skills
  - Technical aptitude
  - Critical Thinking & Skeptical Mindset
  - Curious & Creative
  - Communicative and Collaborative

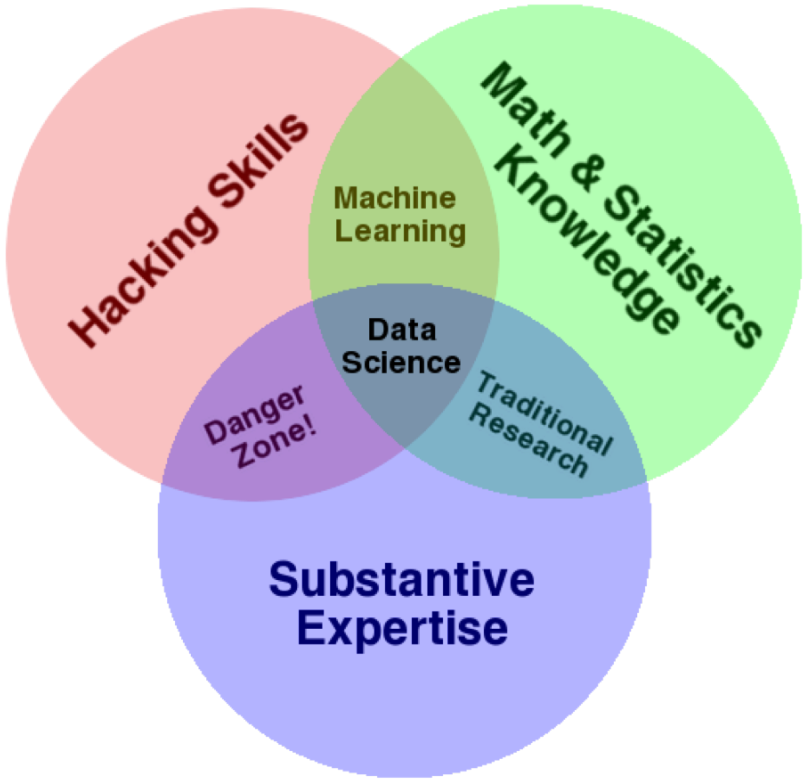


Data Scientist – Profile

- Quantitative skills
  - Such as mathematics and/or statistics
- Technical aptitude
  - Software engineering, machine learning and programming skills
- Critical Thinking & Skeptical Mindset
  - Ability to examine their own work critically rather than in a one-sided way.
- Curious & Creative
  - Passionate about data and finding creative ways to solve problems and portray or visualize information
- Communicative and Collaborative
  - Must be able to articulate the business value in a clear way and
  - Collaboratively work with groups, including sponsors and key stakeholders

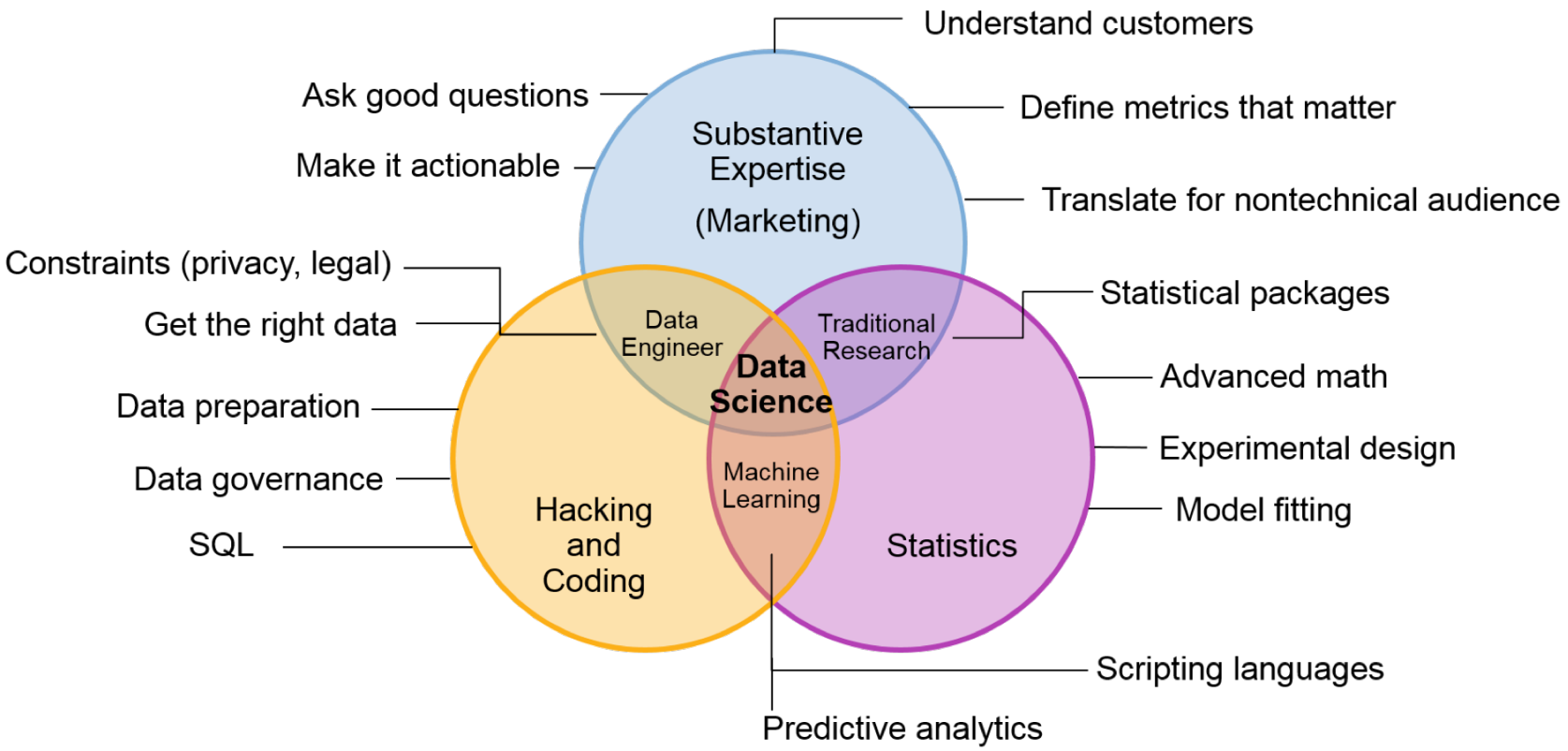
Data Science Venn Diagram – Conway 2010

- Inherent interdisciplinary nature of the these skills.
- None is discipline specific.
- More importantly, each of these skills are on their own very valuable,
- But when combined with only one other are at best simply not data science, or at worst downright dangerous.



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Data Science Venn Diagram – Gartner 2016



http://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html

Acknowledgments

- DPAA Scientific Staff
- UNO College of Information Science and Technology
- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services
- http://hadoop.apache.org/
- http://alpinedata.com/
- http://openrefine.org/
- http://vis.stanford.edu/wrangler/