

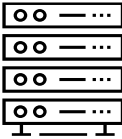



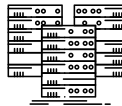

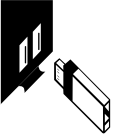

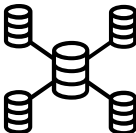



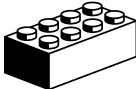


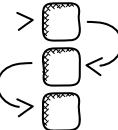

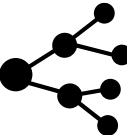
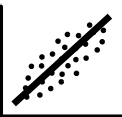
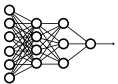

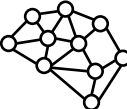


Applied Data Science Skills for Research

	 <p>A non-computational non-data driven researcher. Practices and methods used often prevent scaling research questions to meet ambitions.</p>	 <p>Desktop-based computational researcher able to apply local computer methods and techniques. Able to explore automation and computational modeling.</p>	 <p>HPC user or beginning Cloud user. Has a mental model for computing at-a-distance. Is able to integrate multiple remote systems into workflows.</p>	 <p>A skilled computational practitioner who is able to connect discrete systems for business and research purposes.</p>
Compute	 <p>Spreadsheet skills, mostly unstructured use of spreadsheets and other data sets. Beginning to think about data structures and models.</p>	 <p>Able to implement scripts and simple programs which aid the researcher in automating repetitive tasks.</p>	 <p>Able to think about discrete servers and systems and how they could contribute to a research or business workflow.</p>	 <p>Can leverage multiple cloud systems and or geographic locations to research or economic benefit. Understands DevOps and orchestration.</p>
Storage	 <p>Considers data and storage as something local and contained within a device carried around. Limited models for data-at-a-distance or backup</p>	 <p>Able to consider throughput and redundancy as important considerations to achieve research objectives.</p>	 <p>Able to make decisions about storage based on performance and latency and design systems to consider I/O workflows.</p>	 <p>Able to build/use tiered storage systems which trade off economics, timeliness and accessibility/persistence of data.</p>
Software	 <p>Mostly point and click interface usage. Limited models for reasoning about scripting and programming.</p>	 <p>Beginner use of version control and tracking changes in scripts and code. Collaboration beginning to draw upon these tools.</p>	 <p>Able to collaborate on version control systems and begin using continuous integration and automation workflows.</p>	 <p>Able to build integrated systems which are continuously integrated and invite external contribution and collaboration. Applies mature software practices.</p>
Workflows	 <p>Makes checklists with little to no integration into how items interelate or are scripted. Manual sets of point-click tasks.</p>	 <p>Able to connect multiple scripts or programs together in simple workflows. May not have abstracted scripts to general use.</p>	 <p>Able to consider asynchronous workflow systems and the use of queing systems to manage distributed infrastrucutre.</p>	 <p>Able to build systems of systems, interconnected and interoperable to build and extend capability.</p>
Methods	 <p>Applies statistical models over small and individual data sets. Limited ability to apply similar models over and repeatedly to 100s or 1,000s of data sets.</p>	 <p>Growing understanding of Machine Learning (ML) and ANNs. Able to apply stat methods across splits of larger data sets more capably.</p>	 <p>Advanced applications of ML and considerations of custom hardware, GPUs, FPGAs etc. Able to balance, compute & model complexity to achieve insight.</p>	 <p>Able to connect and interconnect flows of machine learning and statistical insight into complex reasoning networks.</p>