Quick Reference Card Machine Learning Technology Readiness Levels (MLTRLs)

Lavin, A., Gilligan-Lee, C. M., Visnjic, A., Ganju, S., Newman, D., Ganguly, S., Lange, D., Baydin, A. G., Sharma, A., Gibson, A., Zheng, S., Xing, E. P., Mattmann, C., Parr, J., & Gal, Y. (2022). Technology readiness levels for machine learning systems. *Nature Communications*, *13*(1), 6039. https://doi.org/10.1038/s41467-022-33128-9

Level 0—First principles (Lavin et al., 2022, p. 2)

- **Motivation:** Novel idea, question, problem
- **Data:** "Building understanding", data readiness strategies
- **Work:** Literature discovery, math, white-boarding, coding
- **Quality:** Discover principles
- **Outcomes:** "concrete ideas with sound mathematical formulation" "principles, hypotheses, data readiness, and research plans"
- **Stakeholders:** Research lead / manager
- **Review:** "hypotheses and explorations for mathematical validity and potential novelty or utility"

Level 1—Goal-oriented research (p. 3)

- **Goal:** Estimate potential to reach goals, costs
- **Data:** "sample data need not be the full data", may be synthetic
- **Work:** "design and run low-level experiments to analyze specific model or algorithm properties"
- **Code:** "Research-caliber", "Hacky code is okay"
- **Config:** "start semantic versioning practices early in the project lifecycle, which should cover code, models, and datasets"
- **Outcomes:** "comparison studies and analyses" show "if/how/where the technology offers potential improvements and utility"
- **Stakeholders:** R&D group
- **Review:** May include iterative reviews, with feedback

Level 2—Proof of Principle (PoP) development (p. 3)

- **Goals:** "model-specific technical goals"
- **Data:** benchmarks / sampled / simulated
- **Work:** Build "testbeds: simulated environments and / or simulated data that closely matches the conditions and data of real scenarios
- **Outcomes:** "formal research requirements document (with well-specified V&V steps)"
- **Review:** "satisfy research claims made in previous stages (brought to bear by the aforementioned PoP data in both quantitative and qualitative ways) with the analyses well-documented and reproducible"

Level 3—System development (p. 4)

- **Goals:** "interoperability, reliability, maintainability, extensibility, and scalability"
- **Data:** Work on "data coverage and robustness" issues from L2. Test suites for default and "specific functionalities and scenarios"
- **Code:** "Prototype-caliber": more clean and robust, dataflow and interface architecture, unit and integration tests
- **Review:** "Teammates from applied AI and engineering are brought into the review to focus on sound software practices, interfaces and documentation for future development, and version control for models and datasets." Identify applicable standards.

Level 4—Proof of Concept (PoC) development (p. 4)

- **Goals:** "demonstrate the technology in a real scenario" … "explore candidate application areas", show performance
- **Data:** "real and representative data" "may include collecting new data or processing all available data using scaled experiment pipelines"
- **Quality:** "reveals specific differences between clean and controlled research data versus noisy and stochastic real-world data"
- **Ethics:** "Ethics conversations" as appropriate
- **Stakeholders:** Broader "working group" including "product engineering to help define service level agreements and objectives (SLAs and SLOs)" [see L3], "first touch-point with product managers and stakeholders beyond the R&D group."
- **Review:** "Demonstrate the utility towards one or more practical applications … taking care to communicate assumptions and limitations, and again reviewing data readiness: evaluating the real-world data for quality, validity, and availability", security, privacy

Level 5—Machine learning capability (p. 5)

- **Goals:** "more than an isolated model or algorithm, it is a specific capability" … "transition or handoff from R&D to productization"
- **Data:** scaling of data pipelines, data governance
- **Challenge:** *Valley of death* lies ahead: "Graduation from Level 5 should be difficult" … Need "resources to push … through [to] productization"
- **Quality:** QA, governance to "preempt related technical debt" [see L7]
- **Stakeholders:** "interdisciplinary working group", "demos, example scripts, and/or an API"
- **Review:** Create product-driven requirements, V&V

Level 6—Application development (p. 6)

- **Goals:** "software engineering to bring the code up to product-caliber" "integrating the technology into existing production systems"
- **Data:** "robustifying" models, algorithms, and components, adversarial examples, perturbations, generalization to other data
- **Quality:** test suites, QA to "prioritize data governance: how data is obtained, managed, used, and secured by the organization."
- **Code:** "Product-caliber: specifications, test coverage, well-defined APIs, etc. for target use-cases. Any Explainability must be "built and validated alongside the ML model"
- **Review:** Code quality, product requirements, system SLA and SLO requirements, data pipelines spec, AI ethics / regulations

Level 7—Integrations (p. 6)

- **Goals:** Integrate and test
- **Quality:** Risk quantification table, test suites, QA, governance
- **Tests:** "use-case-specific critical scenarios", Golden dataset, "continuous integration and deployment (CI/CD)", metamorphic testing, tests of pipelines
- **Stakeholders:** infrastructure engineers and applied AI engineers. Reduce risks of latent model assumptions and failure modes.
- **Review:** "focus on the data pipelines and test suites; scorecards, ethical considerations

Level 8—Mission-ready (p. 7)

- **Goals:** "The technology is demonstrated to work in its final form and under expected conditions"
- **Data:** "mechanisms for automatically logging data distributions alongside model performance once deployed."
- **Tests:** Deployment focus … A/B tests, blue/green deployment tests, shadow testing, and canary testing … "CI/CD system should be ready to regularly stress test the overall system and ML components" to reveal "data quality issues, data drifts, and concept drifts"
- **Review:** "full slate of stakeholders" "A diligent walkthrough of every technical and product requirement, showing the corresponding validations" … "key decision is go or no-go for deployment, and when."

Level 9—Deployment (p. 7)

- **Goals:** "monitor the current version and explicit considerations towards improving the next version."
- **Data:** "Proper mechanisms for logging and inspecting data (alongside models) is critical for deploying reliable AI and ML—systems that learn on data have unique monitoring requirements"
- **Review:** "The review at this stage is unique, as it also helps in lifecycle management: at a regular cadence that depends on the deployed system and domain of use, owners and other stakeholders are to revisit this review and recommend *switchbacks* if needed (discussed in the Methods section). This additional oversight at deployment is shown to help define regimented release cycles of updated versions, and provide another "eye" check for stale model performance or other system abnormalities."

Notes on this quick reference card

Lavin, A., Gilligan-Lee, C. M., Visnjic, A., Ganju, S., Newman, D., Ganguly, S., Lange, D., Baydin, A. G., Sharma, A., Gibson, A., Zheng, S., Xing, E. P., Mattmann, C., Parr, J., & Gal, Y. (2022). Technology readiness levels for machine learning systems. *Nature Communications*, *13*(1), 6039. https://doi.org/10.1038/s41467-022-33128-9

Version 10, 24 May 2024, please send feedback to: John H. Ganter, Sandia National Laboratories, jganter@sandia.gov. SAND2024-06567O

Inspired by study aids, this quick reference card provides a compact, technically dense map and summary of the Lavin et al. (2022) paper. It encourages reading of the paper by extracting key quotes, concepts, and verbs of interest to developers and managers, and pointing to the page number in the paper. On a single page, the card reminds the reader of the ten levels, their ordering, and how they relate to each other. The reference card can also provide a quick refresher and discussion guide.

Content is either a direct quote in quotation marks or a paraphrase to make the card more compact and easier to scan visually.

For example, “This is a stage for greenfield AI research, initiated with a novel idea, guiding question, or poking at a problem from new angles” (Lavin et al., 2022, p. 2) is paraphrased here as “Novel idea, question, problem”

For each level in the paper, the authors add two subheadings: *Data* and *Review*. In iterating through the levels, we identified several additional themes to summarize, compare, and contrast the levels. For example, “Work” briefly summarizes activities at the level. These headings apply more or less to different levels, so they may be omitted in the card above.

On the quick reference card, there is an implicit ordering from the canonical sequence of a research project, e.g. *Data* is placed near the top because it exercises higher-order control over a project.

“[see L3]” means text is from a later section of the paper, in this case Level 3