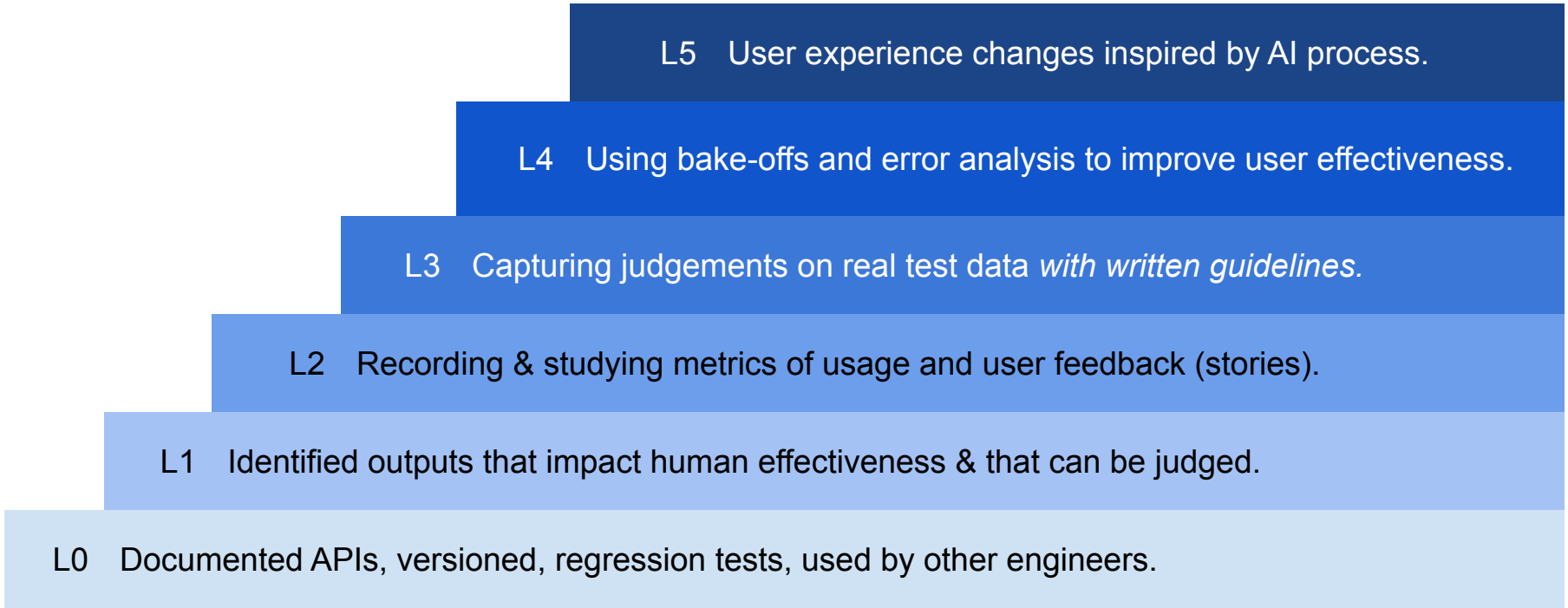
# AI Maturity Model for Production (AIM2prod)

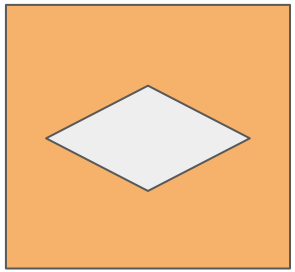
Enterprise Architecture AI Tutorial
Draft June 15, 2023
John R. Frank, PhD

# AI Maturity Levels for Production Software Engineering

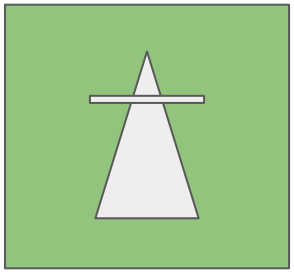


# Foundational AI Strategies for Software Engineering

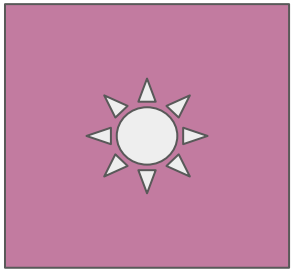
Detectors



Rankers



Mergers



Three Canonical Examples

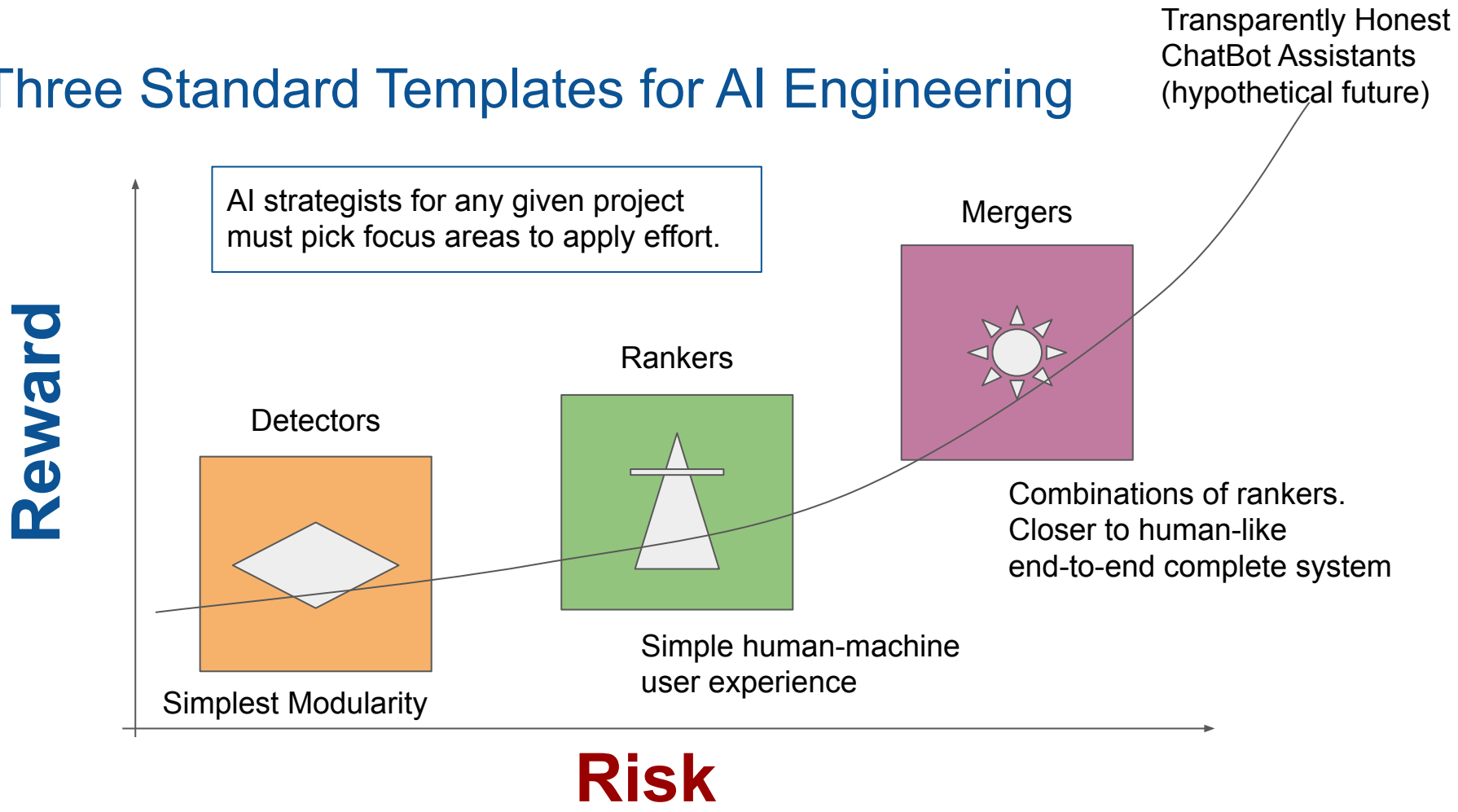
# AI Engineering Strategy Template (best practices)

1. **Measure Quality:** Identify ***restricted*** **Turing tests** at module boundaries.
   - 1.1. *General* Turing tests: open-ended conversation → single judgement of is intelligent or not.
   - 1.2. *Restricted* Turing tests: **simplified story templates** → repeated judgements that we **count**.
   - 1.3. Leverage standard metrics, e.g. from NIST

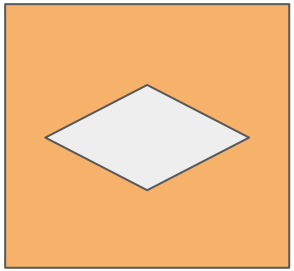
2. **Use Real Data:** Establish baseline measurements on operational data.
   - 2.1. Use existing operational system or simplest possible rule-based function.
   - 2.2. Make a test harness that records human judgements — usually new code.
   - 2.3. Crucial: document guidelines with examples of edge cases, so we can train more judges.

3. **Improve Quality:** Run **offline** contests or “bake offs”:
   - 3.1. Periodically upgrade input/output APIs of modules, so new AI approaches can run on tests.
   - 3.2. Crucial: save historical outputs of test runs for each iteration of new AI algorithms.
   - 3.3. Error Analysis: periodically study errors in past runs to identify opportunities for paths forward.

# Three Standard Templates for AI Engineering

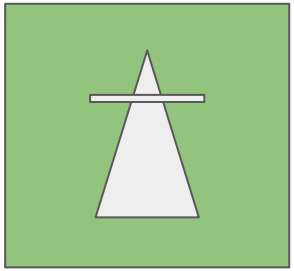


# Three Standard Templates for AI Engineering



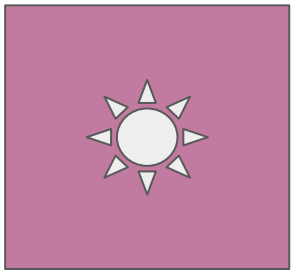
## Detectors:

- Input: *small* complex data **+ defined type of thing**, output is trivial yes/no (often plus/minus)
- **Restricted Turing Test:** human can read/see/hear input to determine yes/no.
- Examples:
  - Is this a phone number from region X?
  - Is this a photo of thing X?



## Rankers:

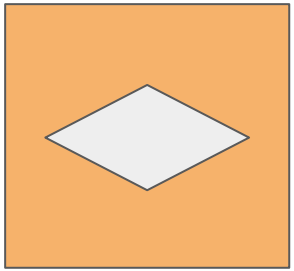
- Input: *massive* complex data **+ user request**, output is **ordered** list of responses.
- **Restricted Turing Test:** human reviews list, accepting/rejecting answers.
- Examples:
  - What should I read about topic X?
  - Give me new factoids of type X to add to my report on topic Z.



## Mergers:

- Input: *massive* complex data + user request **+ schema of ideal output** → populated schema
- **Restricted Turing Test:** human accepts/rejects data in each field.
- Examples:
  - Route me to points A, B, C efficiently given current traffic.
  - Organize my chemistry experiments to find best procedure for synthesizing X.

# Metrics for Detectors



**Detectors:**

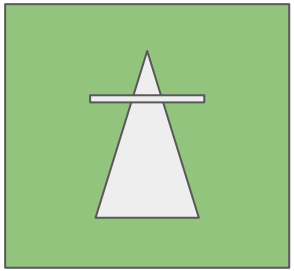
- Several standard metrics computed from building a “confusion matrix”
- Different metrics aim at different goals
- Often implemented with a numerical score computed for each input and then a threshold that decides whether an output is positive (above threshold) or negative (below threshold).

| Confusion Matrix (example) | | Human Judgement | | |
|---|---|---|---|---|
| | | True | False | Totals |
| System Output | Positive | 9,182 | 295 | 9,477 (above threshold) |
| | Negative | 6,485 | 48,279 | 54,764 (below threshold) |

Example:

- Rate of false positives
  3% = 295 / 9477 → Precision = 97%
- Rate of missed detections
  88% = 48279 / 54764 → Recall = 12%
- Aggregate “accuracy” is:
  24% = (9,182 + 6,485) / (9,477 + 54,764)
- F1 score
  harmonic mean of precision & recall:
  2 * prec * recall / (precision + recall)
  In this example: 21%

# Metrics for Rankers



**Rankers:**

- Metrics count effort required by user to read down the list and find good result(s)
- Nuanced judgment about what counts as a “good” result, e.g. not redundant with earlier hits.
- Like a detector with threshold lowered to allow higher recall with lower precision.
- Expected Reciprocal Rank (ERR) is a widely used scoring function.

| Q1 | Q2 | Q3 |
|---|---|---|
| R1.1 ✓ | R2.1 ✗ | R3.1 ✗ |
| R1.2 ✗ | R2.2 ✓ | R3.2 ✗ |
| R1.3 ✓ | R2.3 ✓ | R3.3 ✗ |
| R1.4 ✗ | R2.4 ✓ | R3.4 ✓ |
| R1.5 ✓ | R2.5 ✓ | R3.5 ✓ |

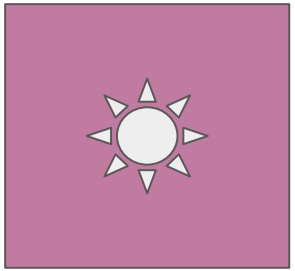
Ranking metrics tend to be precision focused.

Quantifying recall requires more judgments.

1 2 4 ⇨ ERR = avg( 1 + ½ + ¼ ) = 7/12

first good hit

# Metrics for Mergers

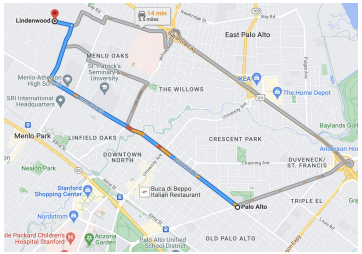


**Mergers:**

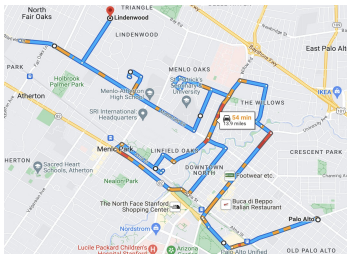
- Metrics based on comparison of output with a human-assembled work product.
- Detector-like: Count human accept/reject elements from merger?
- Quantify Recall:  What did the human find without the merger, what did the merger add?
- Gap analysis → Rumsfeld's Matrix of Epistemic Uncertainty

| | |
|---|---|
| Known Knowns | Known Unknowns |
| Unknown Knowns | Unknown Unknowns |

Merger




Human




Found only by merger

Found only by human

Many roads and bridges were destroyed, and international maritime travel was blocked by the 2022 Russian invasion of Ukraine.[241] Before that it was mainly through the Port of Odesa, from where ferries sailed regularly to Istanbul, Varna and Batumi. The largest ferry company operating these routes was Ukrferry.[258] There are over 1,600 km (1,000 mi) of navigable waterways on 7 rivers, mostly on the Danube, Dniester and Prypiat'. All Ukraine's rivers freeze over in winter, limiting navigation.[259]

Ukraine's rail network connects all major urban areas, port facilities and industrial centres with neighbouring countries.[citation needed] The heaviest concentration of railway track is the Donbas region.[260] Although rail freight transport fell in the 1990s, Ukraine is still one of the world's highest rail users.[261]

Ukraine International Airlines, is the flag carrier and the largest airline,[262] with its head office in Kyiv[263] and its main hub at Kyiv's Boryspil International Airport. It operated domestic and international passenger flights and cargo services to Europe, the Middle East, the United States,[264] Canada,[265] and Asia.

Energy in Ukraine is mainly from gas and coal, followed by nuclear then oil.[170] The coal industry has been disrupted by conflict.[265] Most gas and oil is imported, but since 2015 has prioritised diversifying energy supply.

About half of electricity generation is nuclear and a quarter coal.[170] The largest nuclear power plant in Europe, the Zaporizhzhia Nuclear Power Plant, is in Ukraine. Fossil fuel subsidies were US$2.2 billion in 2019.[267] Until the 2010s all of Ukraine's nuclear fuel came from Russia, but now most does not.[268]