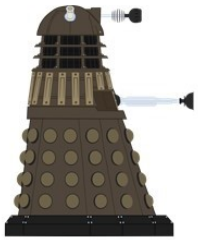
Thus Spoke the Dalek: Explain!
A Short Introduction to Explainable AI

Tarek R. Besold

City, University of London

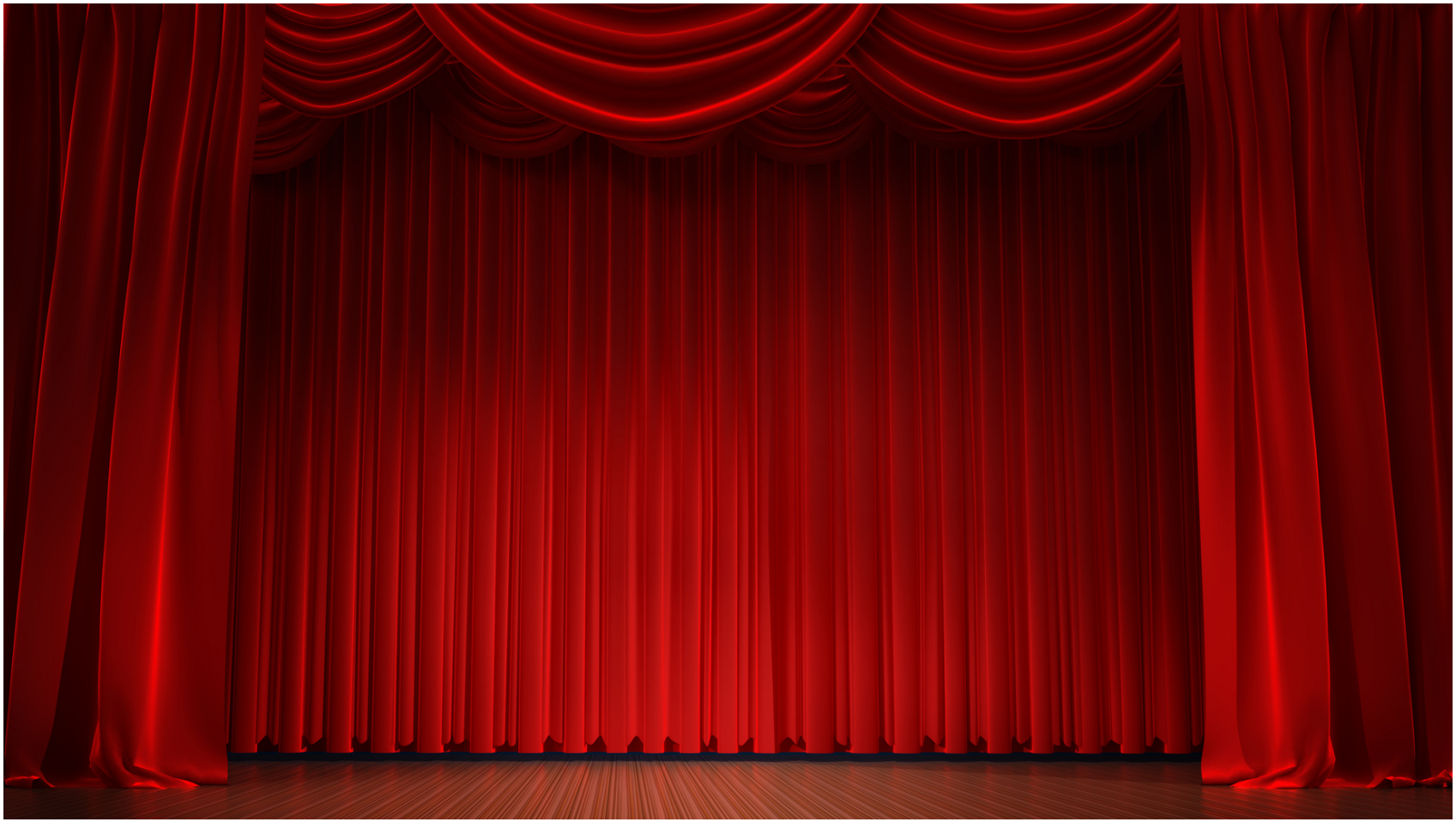
22. November 2017



Honour to whom honour is due...

Most of the following is taken from joint work with:

- **Derek Doran**, Dept. of Computer Science & Engineering, Wright State University (Dayton, Ohio, USA).
- **Sarah Schulz**, Institute for Natural Language Processing, University of Stuttgart (Stuttgart, Germany).



Explanation and Comprehensibility in AI and ML (the 1980s)

Comprehensibility and Michie's Criteria for ML

- **Comprehensibility** of symbolic knowledge as one of the major distinctions between logic-based and statistical/neural approaches in computer science.
- **Michie's criteria for Machine Learning (ML):**
  1. **Weak criterion:** ML occurs whenever a system generates an updated basis building on sample data for improving its performance on subsequent data.
  2. **Strong criterion:** Weak criterion + ability of system to communicate internal updates in explicit symbolic form.
  3. **Ultrastrong criterion:** Strong criterion + communication of updates must be operationally effective (i.e. user required to understand updates and consequences to be drawn from it).

Comprehensibility and Michie's Criteria for ML

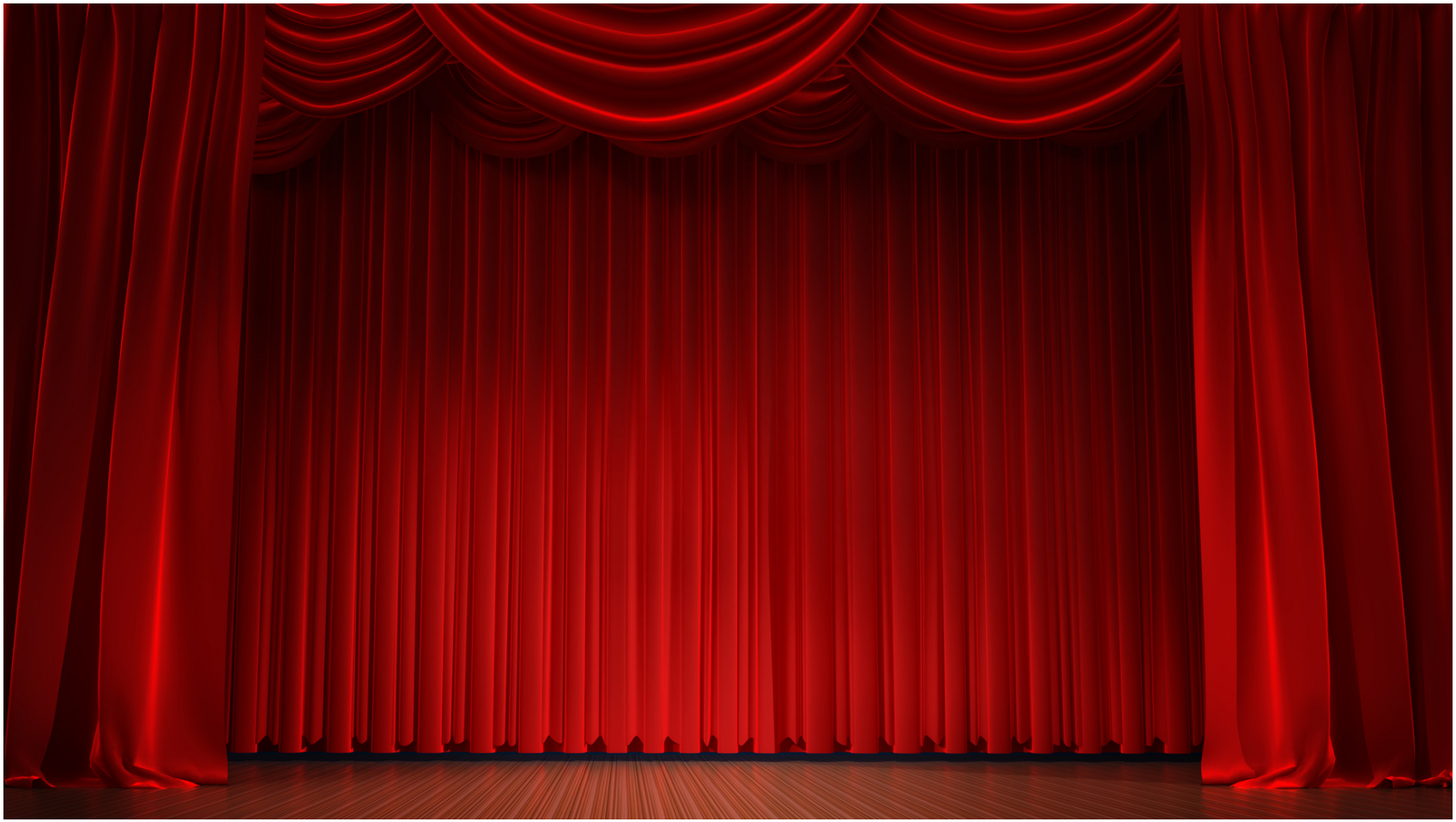
- **Comprehensibility** of symbolic knowledge as one of the major distinctions between logic-based and statistical/neural approaches in computer science.
- **Michie's criteria for Machine Learning (ML):**
  1. **Weak criterion:** ML occurs whenever a system generates an updated basis building on sample data for improving its performance on subsequent data.
  2. **Strong criterion:** Weak criterion + ability of system to communicate internal updates in explicit symbolic form.
  3. **Ultrastrong criterion:** Strong criterion + communication of updates must be operationally effective (i.e. user required to understand updates and consequences to be drawn from it).

Comprehensibility and Michie's Criteria for ML

- **Comprehensibility** of symbolic knowledge as one of the major distinctions between logic-based and statistical/neural approaches in computer science.
- **Michie's criteria for Machine Learning (ML):**
  1. **Weak criterion:** ML occurs whenever a system generates an updated basis building on sample data for improving its performance on subsequent data.
  2. **Strong criterion:** Weak criterion + ability of system to communicate internal updates in explicit symbolic form.
  3. **Ultrastrong criterion:** Strong criterion + communication of updates must be operationally effective (i.e. user required to understand updates and consequences to be drawn from it).

Comprehensibility and Michie's Criteria for ML

- **Comprehensibility** of symbolic knowledge as one of the major distinctions between logic-based and statistical/neural approaches in computer science.
- **Michie's criteria for Machine Learning (ML):**
  1. **Weak criterion:** ML occurs whenever a system generates an updated basis building on sample data for improving its performance on subsequent data.
  2. **Strong criterion:** Weak criterion + ability of system to communicate internal updates in explicit symbolic form.
  3. **Ultrastrong criterion:** Strong criterion + communication of updates must be operationally effective (i.e. user required to understand updates and consequences to be drawn from it).

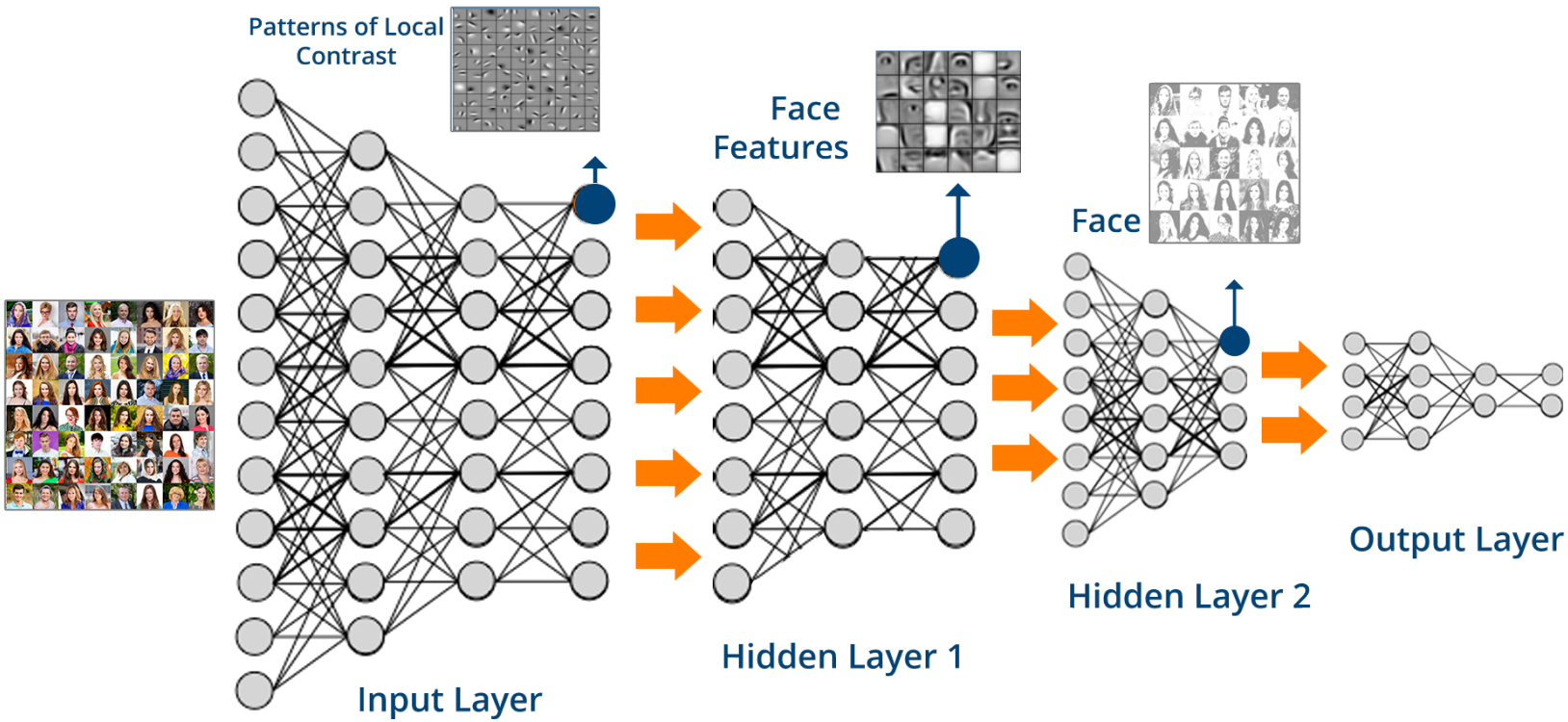


Explanation and Comprehensibility in AI and ML (22/11/2017)

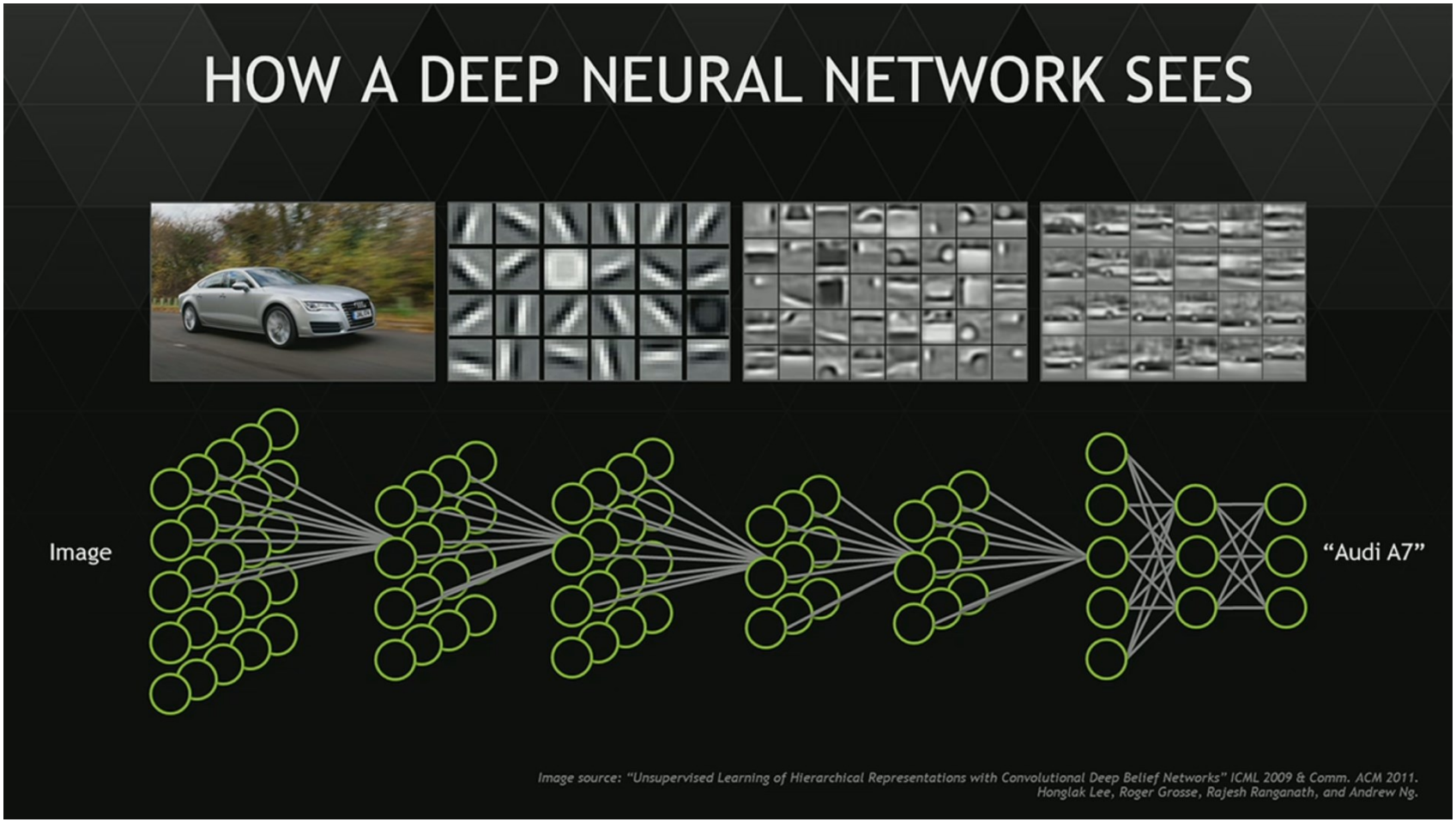
The Weak ML Empire

- Most state of the art systems implement *weak* ML approaches.
- Deep Learning (seems to be) almost necessarily of the weak type: Representations are internally learned, resulting in problem of “symbol reference” for hypothetical symbolic output.
- Laudable exception(s): Some systems in the ILP/IFP/SRL space.

The Weak ML Empire



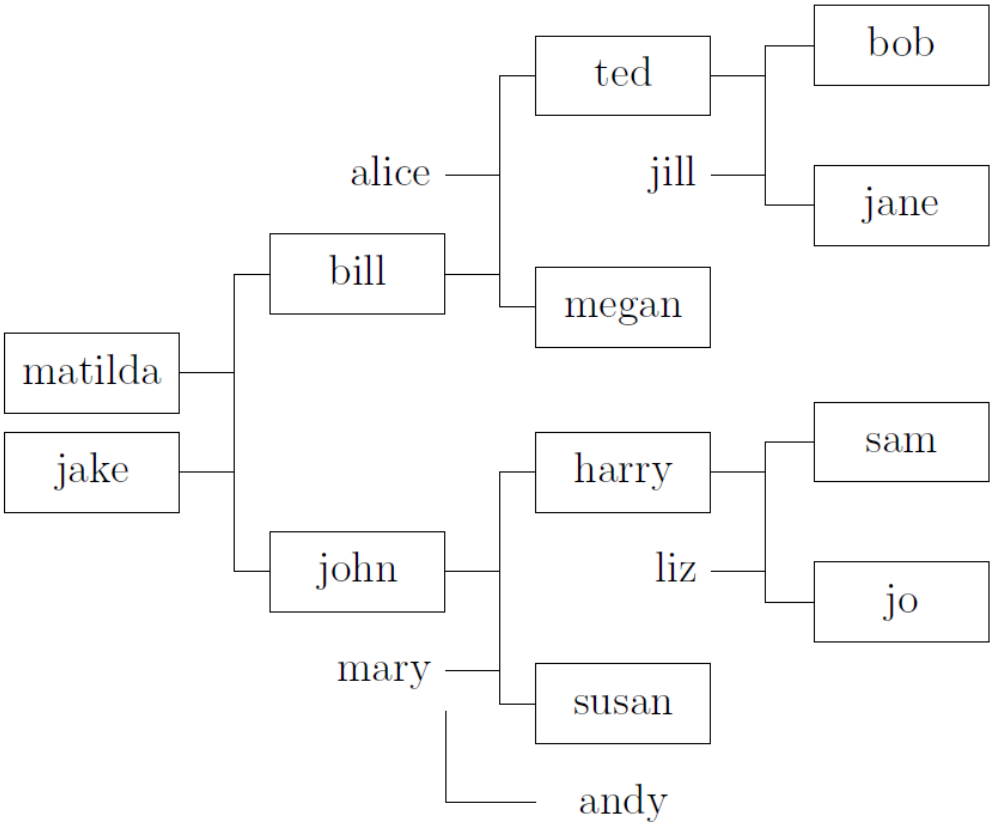
The Weak ML Empire



The Weak ML Empire

- Most state of the art systems implement *weak* ML approaches.
- Deep Learning (seems to be) almost necessarily of the weak type: Representations are internally learned, resulting in problem of “symbol reference” for hypothetical symbolic output.
- Laudable exception(s): Some systems in the ILP/IFP/SRL space.

A Very Quick ILP Primer




```
father(jake,bill).
father(jake,john).
father(bill,ted).
father(bill,megan).
father(john,harry).
father(john,susan).

father(ted,bob).
father(ted,jane).
father(harry,sam).
father(harry,jo).
```

```
mother(matilda,bill).
mother(matilda,john).
mother(alice,ted).
mother(alice,megan).
mother(mary,harry).
mother(mary,susan).
mother(mary,andy).
mother(jill,bob).
mother(jill,jane).
mother(liz,sam).
mother(liz,jo).
```

A Very Quick ILP Primer

; **grandparent** without PI
```
p(X,Y) :- father(X,Z), father(Z,Y).
p(X,Y) :- father(X,Z), mother(Z,Y).
p(X,Y) :- mother(X,Z), mother(Z,Y).
p(X,Y) :- mother(X,Z), father(Z,Y).
```

; **grandparent** with PI
```
p(X,Y) :- p1(X,Z), p1(Z,Y).
p1(X,Y) :- father(X,Y).
p1(X,Y) :- mother(X,Y).
```

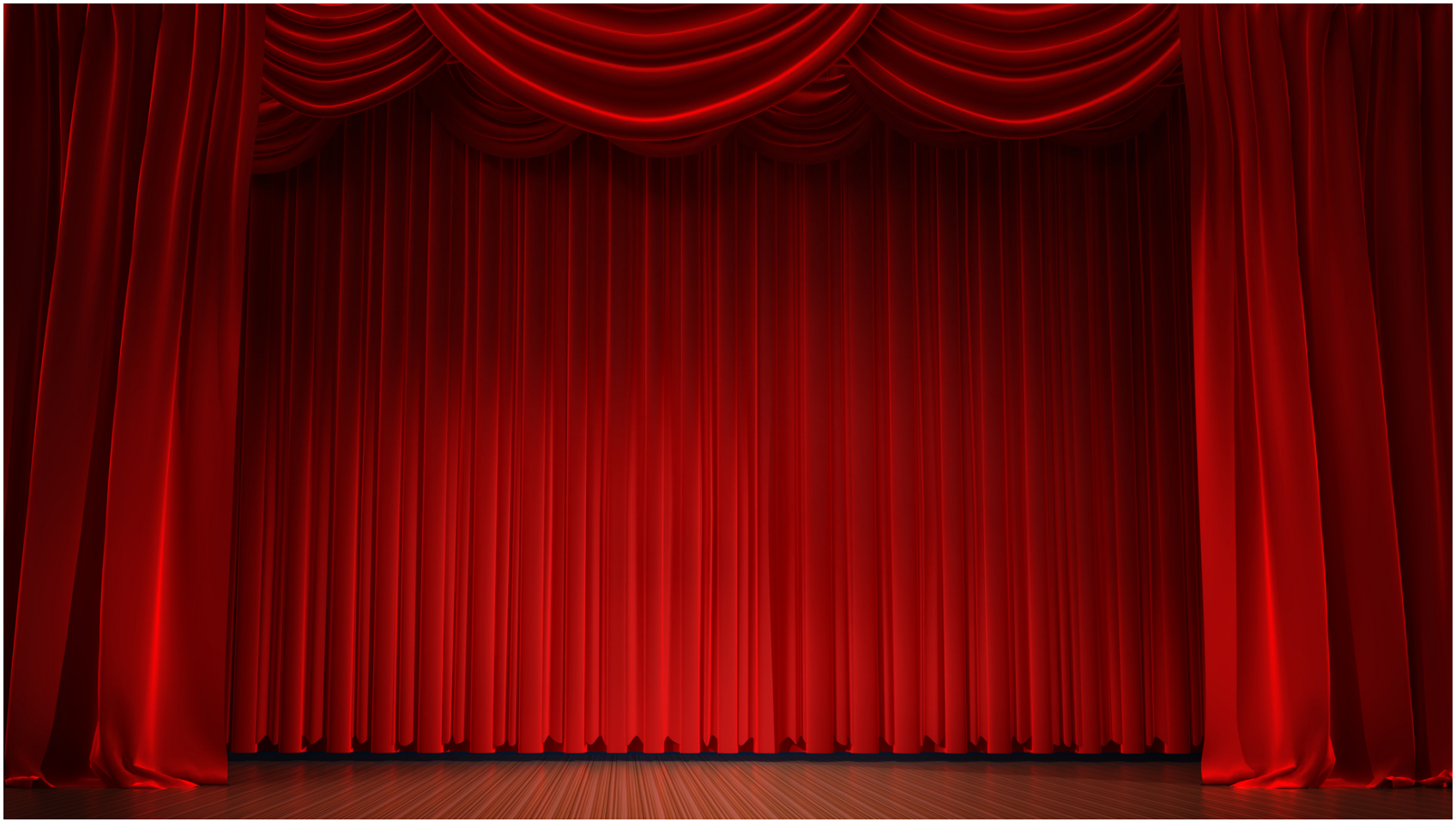
; **ancestor** without PI
```
p(X,Y) :- father(X,Y).
p(X,Y) :- mother(X,Y).
p(X,Y) :- father(X,Z), p(Z,Y).
p(X,Y) :- mother(X,Z), p(Z,Y).
```

; **greatgrandparent** without PI
```
p(X,Y) :- father(X,U), father(U,Z), father(Z,Y).
p(X,Y) :- father(X,U), father(U,Z), mother(Z,Y).
p(X,Y) :- father(X,U), mother(U,Z), father(Z,Y).
p(X,Y) :- father(X,U), mother(U,Z), mother(Z,Y).
p(X,Y) :- mother(X,U), father(U,Z), mother(Z,Y).
p(X,Y) :- mother(X,U), father(U,Z), father(Z,Y).
p(X,Y) :- mother(X,U), mother(U,Z), mother(Z,Y).
p(X,Y) :- mother(X,U), mother(U,Z), father(Z,Y).
```

; **greatgrandparent** with PI
```
p(X,Y) :- p1(X,U), p1(U,Z), p1(Z,Y).
p1(X,Y) :- father(X,Y).
p1(X,Y) :- mother(X,Y).
```

; **ancestor** with PI
```
p(X,Y) :- p1(X,Y).
p(X,Y) :- p1(X,Z), p(Z,Y).
p1(X,Y) :- father(X,Y).
p1(X,Y) :- mother(X,Y).
```

Figure: Prolog programs for *grandparent/2*, *greatgrandparent/2*, and *ancestor/2* with and without PI (*p1/2* equiv. *parent*).



Types of AI/ML Systems from Opaque to Comprehensible

Opaque Systems

Opaque Systems

- The mechanisms mapping inputs to outputs are invisible to the user.
- Can be seen as oracle making predictions over inputs, without indication of how and why.
- Examples:
  - Closed-source, proprietary AI.
  - Genuine “black-box approaches”: inspection of algorithm and implementation does not give insight into systems’s reasoning.

Interpretable Systems

- User cannot only see, but also study and understand input/output mapping.
- Implies model *transparency*, requires some understanding of technical details of mapping.
- Examples:
  - Regression model (interpreted by comparing covariate weights to realize relative importance of features).
  - SVMs and linear classifiers (data classes defined by location relative to decision boundaries).
- Counterexample: Deep ANNs with automatically learned input features transformed through non-linearities.

Comprehensible Systems

- Emits symbols along with its output (remember: Michie's *strong* and *ultra-strong machine learning*), allowing user to relate properties of inputs to outputs.
- User responsible for compiling and comprehending symbols, reasoning about them.
- Comprehensibility is graded notion, degree of comprehensibility corresponding to relative ease/difficulty of compilation and comprehension.
- Examples:
  - (Some) ILP systems.
  - Receptive field visualization on convolutional ANNs.

Relating Opaque/Interpretable/Comprehensible Systems

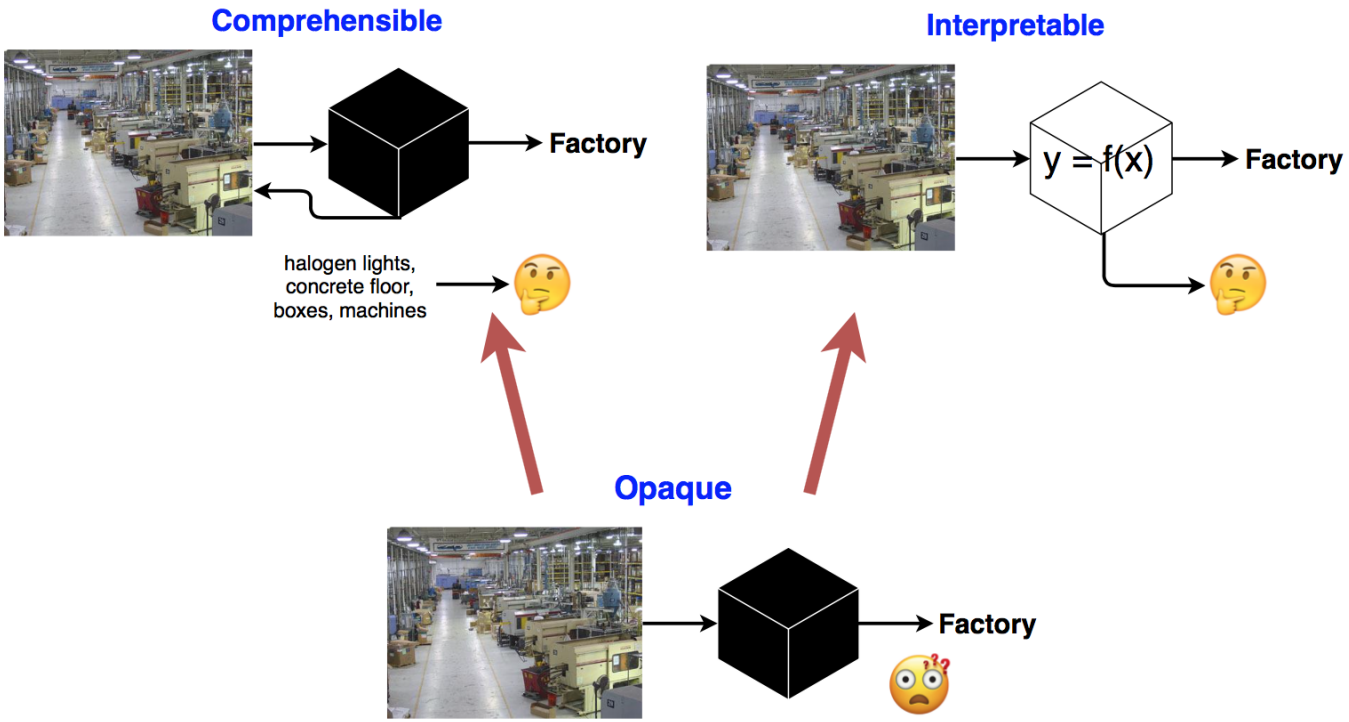


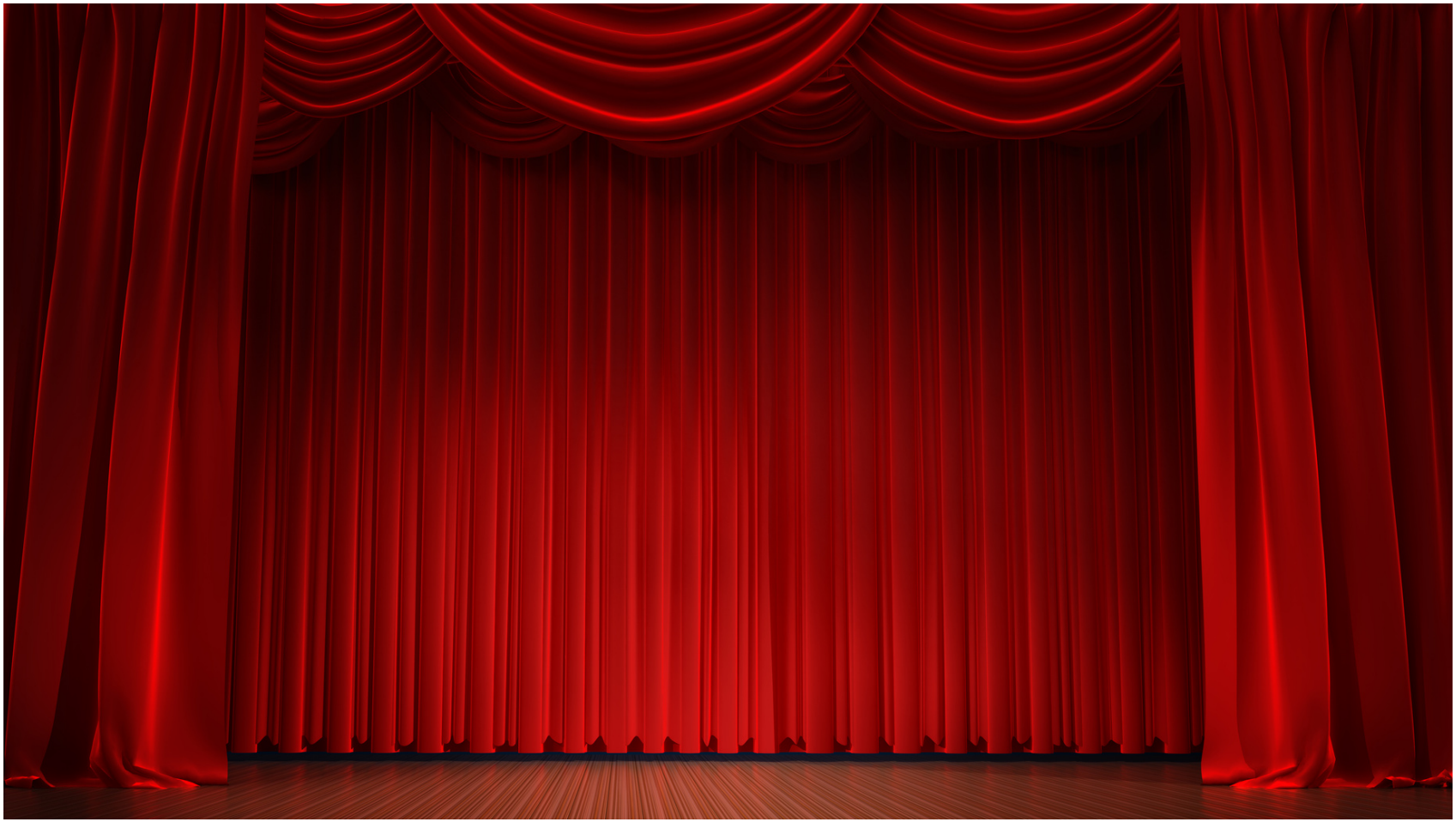
Figure: Notions of comprehension and interpretation are improvements over opaque systems, but remain distinct:
$\rightarrow$ Interpretation requires transparency in underlying mechanisms.
$\rightarrow$ Comprehensible systems are possibly opaque while emitting symbols usable for reasoning.

The Doctor Analogy: Interpretable vs. Comprehensible

- Towards patients, physicians should be like comprehensible models:
  - Deliver diagnosis by explaining high-level indicators revealed in tests (i.e. system symbols).
  - Not providing information about how medical tests and evaluations work.
- Towards other doctors/medical staff, physicians may be like interpretable models:
  - Sketch technical line of connecting patient symptoms and test results to particular diagnosis.
  - Other doctors and staff can interpret diagnosis, ensuring that conclusions are supported by reasonable evaluation functions and weight values for presented evidence.

The Doctor Analogy: Interpretable vs. Comprehensible

- Towards patients, physicians should be like comprehensible models:
  - Deliver diagnosis by explaining high-level indicators revealed in tests (i.e. system symbols).
  - Not providing information about how medical tests and evaluations work.
- Towards other doctors/medical staff, physicians may be like interpretable models:
  - Sketch technical line of connecting patient symptoms and test results to particular diagnosis.
  - Other doctors and staff can interpret diagnosis, ensuring that conclusions are supported by reasonable evaluation functions and weight values for presented evidence.



Beyond the Status Quo: Explainable Systems

Explanation(s)

- OED: *A statement or account that makes something clear; a* **reason** *or justification given for an action or belief.*
- Problem(s) for AI/ML:
  - ML algorithms producing rules about data features to establish a classification decision (e.g., decision trees) shed light into *how*, not *why*, decisions are made.
  - Visualizations/text provided along with a decision (e.g., in deep learning for CV) require human-driven post processing under their own line of reasoning.
- Lipton (2016): *"the term interpretability holds no agreed upon meaning, and yet machine learning conferences frequently publish papers which wield the term in a quasi-mathematical way".*

Explanation(s)

- OED: *A statement or account that makes something clear; a* **reason** *or justification given for an action or belief.*
- Problem(s) for AI/ML:
  - ML algorithms producing rules about data features to establish a classification decision (e.g., decision trees) shed light into *how*, not *why*, decisions are made.
  - Visualizations/text provided along with a decision (e.g., in deep learning for CV) require human-driven post processing under their own line of reasoning.
- Lipton (2016): *"the term interpretability holds no agreed upon meaning, and yet machine learning conferences frequently publish papers which wield the term in a quasi-mathematical way".*

Explanation(s)

- OED: *A statement or account that makes something clear; a* **reason** *or justification given for an action or belief.*
- Problem(s) for AI/ML:
  - ML algorithms producing rules about data features to establish a classification decision (e.g., decision trees) shed light into *how*, not *why*, decisions are made.
  - Visualizations/text provided along with a decision (e.g., in deep learning for CV) require human-driven post processing under their own line of reasoning.
- Lipton (2016): *"the term interpretability holds no agreed upon meaning, and yet machine learning conferences frequently publish papers which wield the term in a quasi-mathematical way".*

Truly Explainable AI Needs Reasoning

- Interpretable and comprehensible systems are lacking in ability to formulate **line of reasoning** explaining decision making process of model **using human-understandable features of input data**.
- Interpretable and comprehensible models *enable* explanations of decisions, but do not yield explanations themselves!
- **Reasoning** as critical step in formulating explanations about why or how some event occurred!

Truly Explainable AI Needs Reasoning

- Interpretable and comprehensible systems are lacking in ability to formulate **line of reasoning** explaining decision making process of model **using human-understandable features of input data**.
- Interpretable and comprehensible models *enable* explanations of decisions, but do not yield explanations themselves!
- **Reasoning** as critical step in formulating explanations about why or how some event occurred!

Truly Explainable AI Needs Reasoning

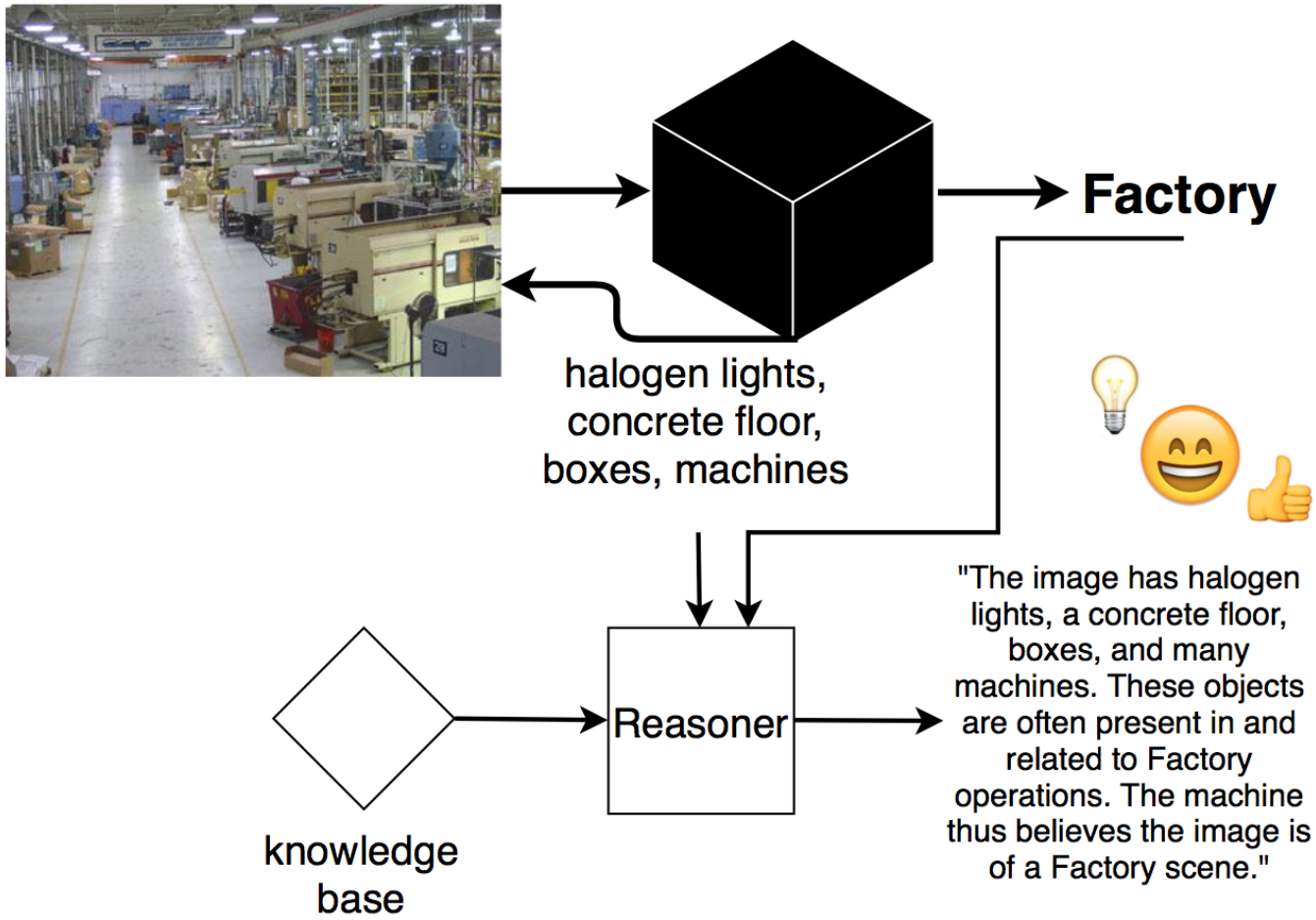


Figure: Combine symbols emitted by comprehensible machine with (domain specific) knowledge base encoding relationships between concepts represented by symbols. Relationships between symbols in knowledge based can yield logical deduction about relationship to machine's decision.

Traits of Explainable Systems

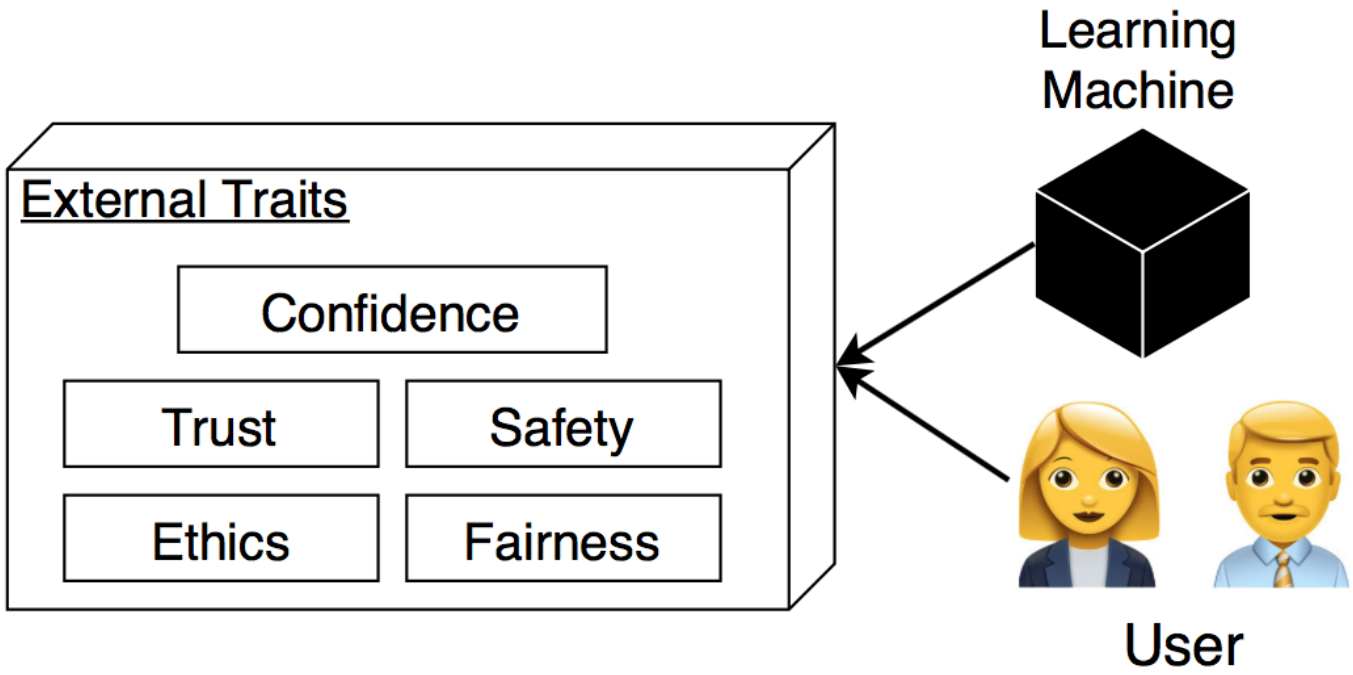
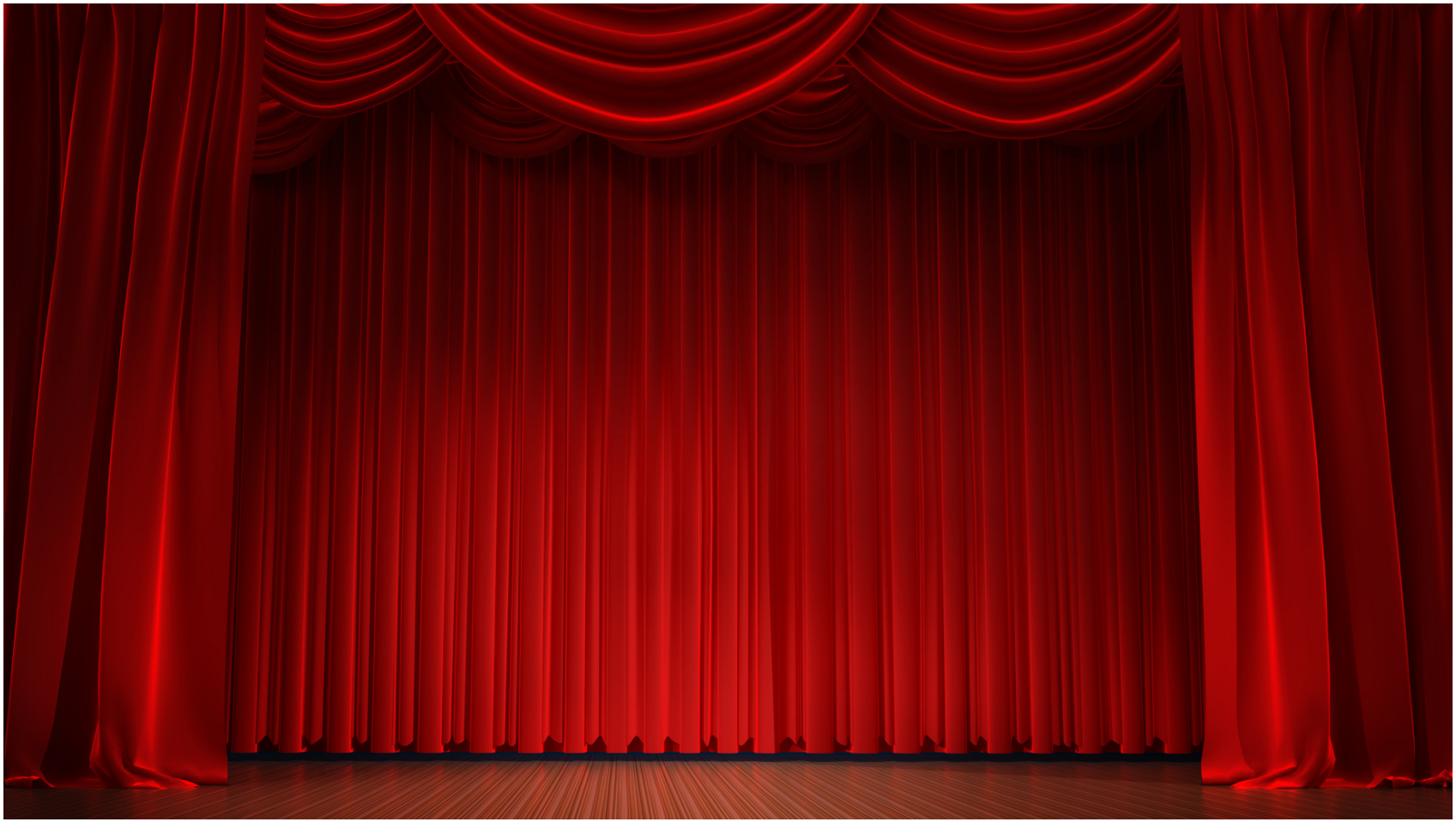


Figure: External system traits related to explainable AI. Traits depend not only on properties of learning machine, but also on users.



(Definitely Not) The End

Take Home Message

- Four general types of AI/ML systems:
  - Opaque.
  - Interpretable.
  - Comprehensible.
  - Explainable.
- Interpretability and comprehensibility are different!
  $\Rightarrow$ “White box” vs. “communicating black box”.
- Both are lacking in ability to formulate line of *reasoning* explaining decision making process of model *using human-understandable features of input data*.
- **Explanation requires reasoning!**