12.

Abductive Reasoning:
Explanation and Diagnosis

Outline

- Introduction to Abductive Reasoning
- Explanation & Diagnosis
- Computing Explanations
- Reading Material

Three Forms of Reasoning

- ***Deduction***
  - an analytic process based on the application of general rules to particular cases with the inference of a result
- ***Induction***
  - synthetic reasoning which infers the rule from the case and the result
- ***Abduction***
  - another form of synthetic inference but of the case from a rule and a result

Abduction

- Probational adoption of a hypothesis as explanation for observed facts results according to known laws
- Weak form of inference because we cannot say that we believe in the truth of the explanation

grass-is-wet $\leftarrow$ rained-last-night

grass-is-wet $\leftarrow$ sprinkler-was-on

shoes-are-wet $\leftarrow$ grass-is-wet

- Form of non-monotonic reasoning
- Multiple explanations might exist

Abductive Reasoning

Given KB, and an α that I do *not* believe,

what would be sufficient to make me believe that α was true?

or what else would I have to believe for α to become an implicit belief?

or what would *explain* α being true?

Deduction: given $(p \supset q)$, from p, deduce q

Abduction: given $(p \supset q)$, from q, abduce p

p is sufficient for q or one way for q to be true is for p to be true

Also induction: given $p(t_1), q(t_1), \ldots, p(t_n), q(t_n)$, induce $\forall x\, (p(x) \supset q(x))$

Can be used for causal reasoning: (*cause* $\supset$ *effect*)

Diagnosis

One simple version of diagnosis uses abductive reasoning

KB has facts about symptoms and diseases

including: $(Disease \land Hedges \supset Symptoms)$

Goal: find disease(s) that best explain observed symptoms

Observe: we typically do not have knowledge of the form

$(Symptom \land \ldots \supset Disease)$

so reasoning is not deductive

Example:

(tennis-elbow $\supset$ sore-elbow)
(tennis-elbow $\supset$ tennis-player)
(arthritis $\land$ untreated $\supset$ sore-joints)
(sore-joints $\supset$ sore-elbow $\land$ sore-hip)

Explain: sore-elbow

Want: tennis-elbow,
(arthritis $\land$ untreated),
...

Non-uniqueness: multiple equally good explanations

+ logical equivalences: (untreated $\land$ $\neg\neg$arthritis)

Adequacy criteria

Given KB, and β to be explained, we want an α such that

1. α is sufficient to account for β

   KB $\cup \{\alpha\} \models \beta$ or KB $\models (\alpha \supset \beta)$

2. α is not ruled out by KB

   KB $\cup \{\alpha\}$ is consistent or KB $\not\models \neg\alpha$

   otherwise $(p \wedge \neg p)$ would count as an explanation

3. α is as simple as possible

   parsimonious : as few *terms* as possible

   explanations should not unnecessarily strong or unnecessarily weak

   e.g. KB = $\{(p \supset q), \neg r\}$ and $\beta = q$

   $\alpha = (p \wedge s \wedge \neg t)$ is too strong

   $\alpha = (p \vee r)$ is too weak

4. α is in the appropriate vocabulary

   atomic sentences of α should be drawn from **H,** possible hypotheses in terms of which explanations are to be phrased

   e.g. diseases, original causes

   e.g. sore-elbow explains sore-elbow

   trivial explanation

   sore-joints explains sore-elbow

   may or may not be suitable

Call such α an explanation of β wrt KB

Example

- For the observation, shoes-are-wet,
  - the explanation grass-is-wet is not in the appropriate vocabulary
  - But it-rained-last-night, and sprinkler-was-on are in the appropriate vocabulary
  - {it-rained-last-night, sprinkler-was-on} is not a minimal explanation, but {it-rained-last-night} is a minimal explanation

Two Simplifications

- It is sufficient to consider
  - how to explain an atom or a literal
  - Conjunction of literals (or negation of a clause)

Some simplifications

From criteria of previous slide, we can simplify explanations in the propositional case, as follows:

- To explain an arbitrary wff β, it is sufficient to choose a new letter p, add $(p \equiv \beta)$ to KB, and then explain p.

  KB $\models (E \supset \beta)$ iff KB $\cup \{(p \equiv \beta)\} \models (E \supset p)$

- Any explanation will be (equivalent to) a conjunction of literals (that is, the negation of a clause)

  Why? If α is a purported explanation, and DNF[α] = $(d_1 \vee d_2 \vee \ldots \vee d_n)$ then each d_i is also an explanation that is no less simple than α

A simplest explanation is then the negation of a clause with a *minimal* set of literals

So: to explain a literal ρ, it will be sufficient to find the minimal clauses C (in the desired vocabulary) such that

1. KB $\models (\neg C \supset \rho)$ or KB $\models (C \cup \{\rho\})$ sufficient
2. KB $\not\models C$ consistent

Prime implicates

A clause C is a <u>prime implicate</u> of a KB iff

1. KB $\models C$
2. For no $C^* \subset C$, KB $\models C^*$

Note: For any clause C, if KB $\models C$, then some subset of C is a prime implicate

Example: KB = $\{(p \wedge q \wedge r \supset g), (\neg p \wedge q \supset g), (\neg q \wedge r \supset g)\}$

Prime implicates:

$(p \vee \neg q \vee g)$,
$(\neg r \vee g)$, and
$(p \vee \neg p)$, $(g \vee \neg g)$, ...

Note: tautology $(a \vee \neg a)$ is always a prime implicate unless KB $\models a$ or KB $\models \neg a$

For explanations:

- want minimal C such that KB $\models (C \cup \{\rho\})$ and KB $\not\models C$
- so: find prime implicates C such that $\rho \in C$;
  then $\neg(C - \rho)$ must be an explanation for ρ

Example: explanations for g in example above

- 3 prime implicates contain g, so get 3 explanations: $(\neg p \wedge q)$, r, and g

Computing explanations

Given KB, to compute explanations of literal ρ in vocabulary **H**:

calculate the set $\{\neg(C - \rho) \mid C$ is a prime implicate and $\rho \in C\}$

prime implicates containing ρ

But how to compute prime implicates?

Can prove: Resolution is complete for non-tautologous prime implicates

KB |= C iff KB $\rightarrow C$ completeness for [] is a special case!

So: assuming KB is in CNF, generate *all* resolvents in language **H**, and retain those containing ρ that are minimal

Could pre-compute all prime implicates, but there may be *exponentially* many, even for a Horn KB

Example: atoms: $p_i,\ q_i,\ E_i,\ O_i,\ 0 \leq i < n\ +\ E_n,\ O_n$

wffs: $E_i \wedge p_i \supset O_{i+1},\ E_i \wedge q_i \supset E_{i+1},$
$O_i \wedge p_i \supset E_{i+1},\ O_i \wedge q_i \supset O_{i+1},$
$E_0,\ \neg O_0$

explain: E_n

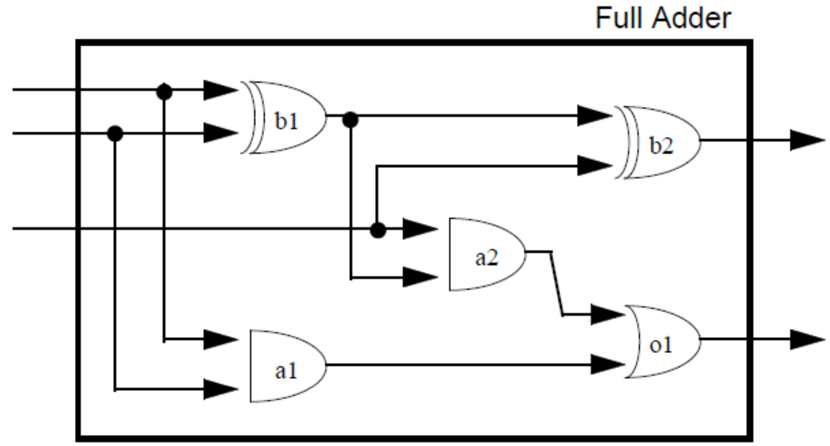
Circuit example

Components

Gate(x) $\equiv$ Andgate(x) $\vee$ Orgate(x) $\vee$ Xorgate(x)

Andgate(a1), Andgate(a2),
Orgate(o1),
Xorgate(b1), Xorgate(b2)

Fulladder(f) the whole circuit



Connectivity

in1(b1) = in1(f), in2(b1) = in2(f)

in1(b2) = out(b1), in2(b2) = in3(f)

in1(a1) = in1(f), in2(a1) = in2(f)

in1(a2) = in3(f), in2(a2) = out(b1)

in1(o1) = out(a2), in2(o1) = out(a1)

out1(f) = out(b2), out2(f) = out(o1)

Circuit behaviour

Truth tables for logical gates

and(0,0) = 0, and(0,1) = 0, ... or(0,0) = 0, or(0,1) = 1, ...
xor(0,0) = 0, xor(0,1) = 1, ...

Normal behaviour

$\text{Andgate}(x) \land \neg\text{Ab}(x) \supset \text{out}(x) = \text{and}(\text{in1}(x), \text{in2}(x))$
$\text{Orgate}(x) \land \neg\text{Ab}(x) \supset \text{out}(x) = \text{or}(\text{in1}(x), \text{in2}(x))$
$\text{Xorgate}(x) \land \neg\text{Ab}(x) \supset \text{out}(x) = \text{xor}(\text{in1}(x), \text{in2}(x))$

Abnormal behaviour: fault models

Examples

$[\text{Orgate}(x) \lor \text{Xorgate}(x)] \land \text{Ab}(x) \supset \text{out}(x) = \text{in2}(x)$ (short circuit)

Other possibilities ...

- some abnormal behaviours may be inexplicable
- some may be compatible with normal behaviour on certain inputs

Abductive diagnosis

Given KB as above + input settings

e.g. KB $\cup$ $\{\text{in1(f)} = 1, \text{in2(f)} = 0, \text{in3(f)} = 1\}$

we want to explain observations at outputs

e.g. $(\text{out1(f)} = 1 \;\wedge\; \text{out2(f)} = 0)$

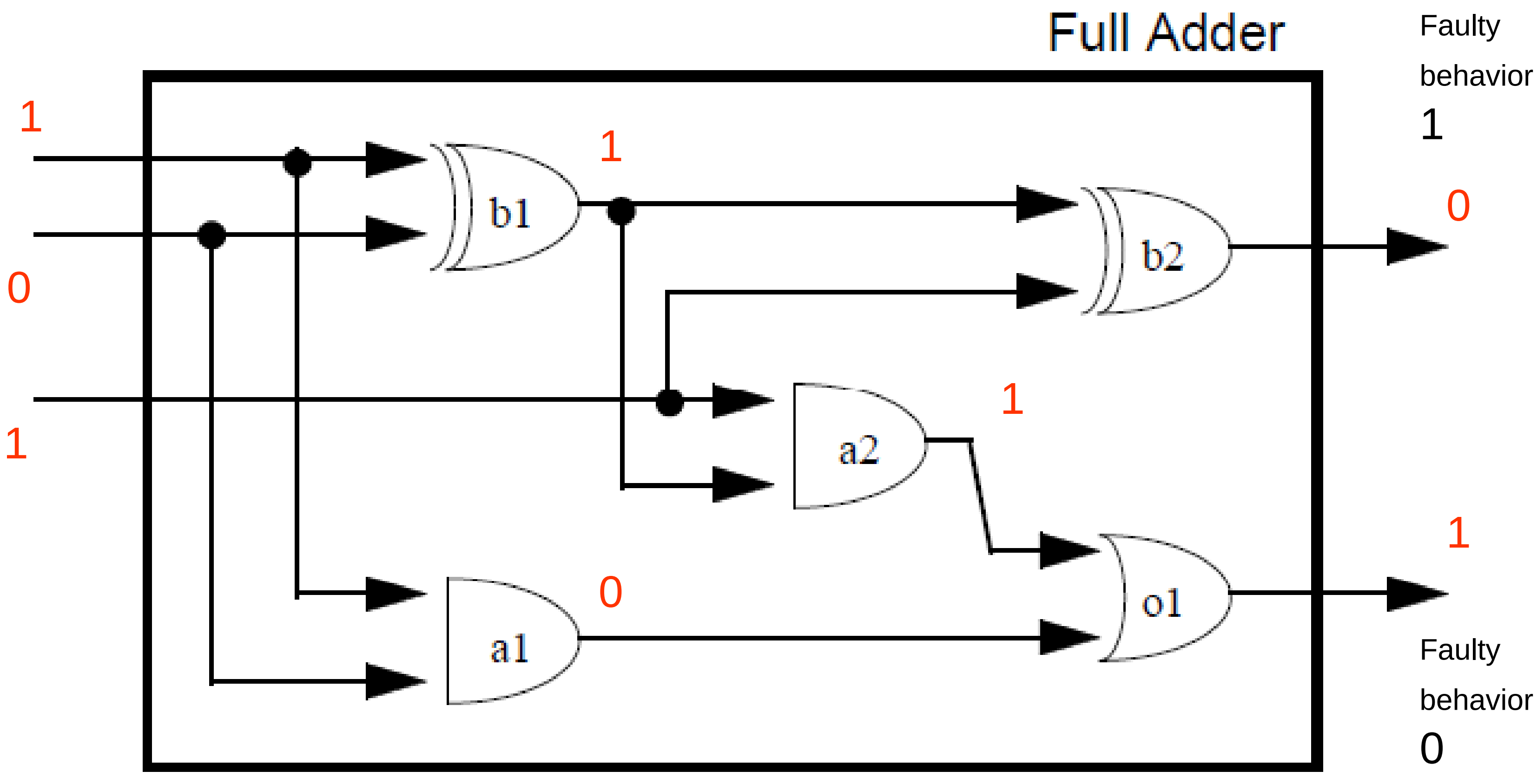
in the language of Ab

We want conjunction of Ab literals α such that
KB $\cup$ *Settings* $\cup$ $\{\alpha\}$ $\models$ *Observations*

Compute by "propositionalizing":

For the above, x ranges over 5 components and u, v range over 0 and 1.

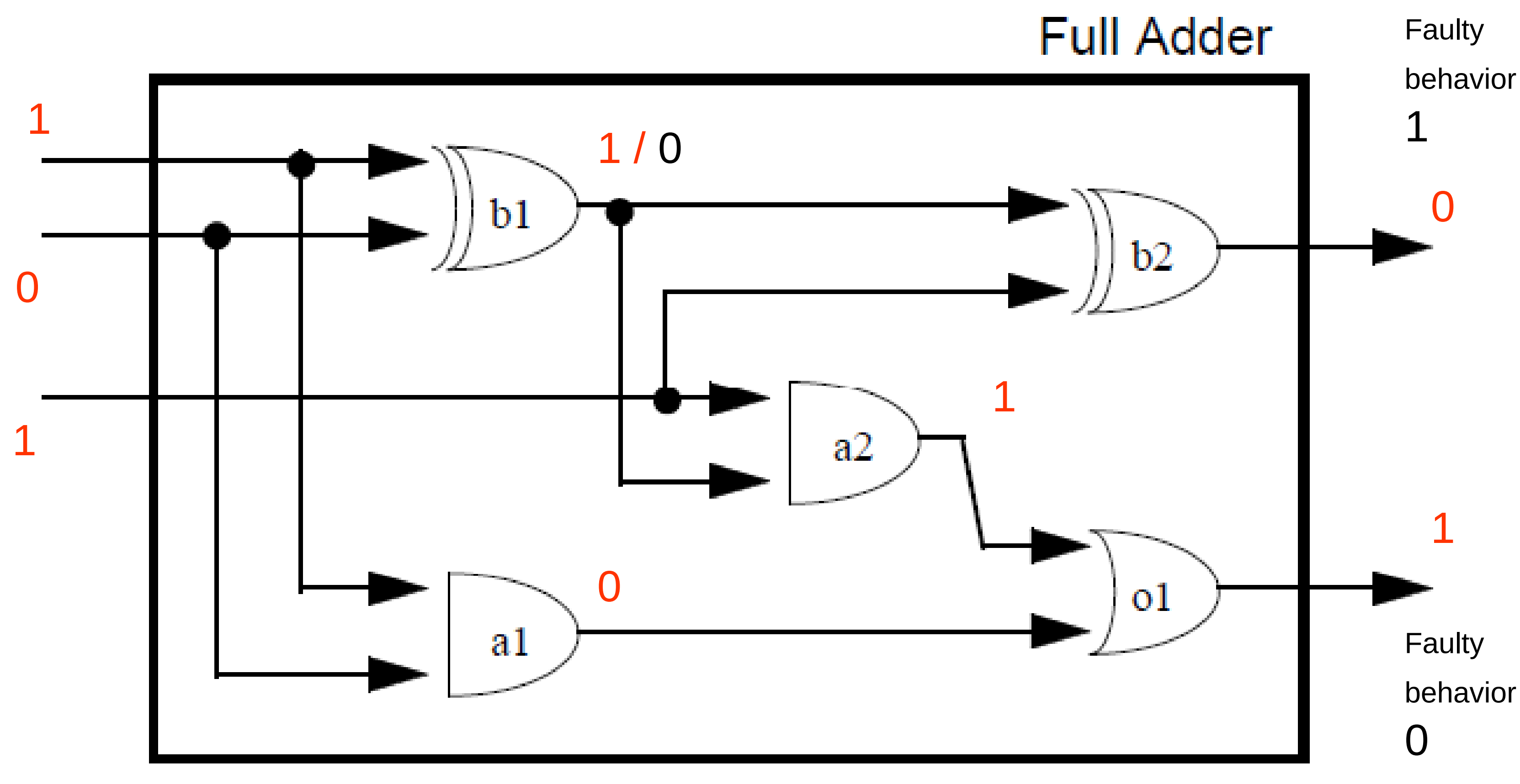
Easiest to do by preparing a table ranging over all Ab literals, and seeing which conjunctions entail the observations.



Full Adder
Faulty
behavior
1
0
1
0
1
b1
1
b2
a2
1
o1
1
0
a1
Faulty
behavior
0

Table for abductive diagnosis

	Ab(b1)	Ab(b2)	Ab(a1)	Ab(a2)	Ab(o1)	Entails observation?
1.	Y	Y	Y	Y	Y	N
2.	Y	Y	Y	Y	N	N
3.	Y	Y	Y	N	Y	N
4.	Y	Y	Y	N	N	N
5.	Y	Y	N	Y	Y	Y
6.	Y	Y	N	Y	N	N
7.	Y	Y	N	N	Y	Y
8.	Y	Y	N	N	N	Y
9.	Y	N	Y	Y	Y	N
10.	Y	N	Y	Y	N	N
11.	Y	N	Y	N	Y	N
12.	Y	N	Y	N	N	N
13.	Y	N	N	Y	Y	Y
14.	Y	N	N	Y	N	N
15.	Y	N	N	N	Y	Y
...						
32.	N	N	N	N	N	N



Ab (b1) Ab (b2) Ab(a2) Ab(o1)

Example diagnosis

Using the table, we look for minimal sets of literals.

For example, from line (5), we have that

$$Ab(b1) \land Ab(b2) \land \neg Ab(a1) \land Ab(a2) \land Ab(o1)$$

entails the observations. However, lines (5), (7), (13) and (15) together lead us to a smaller set of literals (the first explanation below).

The explanations are

1. $Ab(b1) \land \neg Ab(a1) \land Ab(o1)$
2. $Ab(b1) \land \neg Ab(a1) \land \neg Ab(a2)$
3. $Ab(b2) \land \neg Ab(a1) \land Ab(o1)$

Note: not all components are mentioned since for these settings, get the same observations whether or not they are working

but for this fault model only

Can narrow down diagnosis by looking at a number of different settings

differential diagnosis

Diagnosis revisited

Abductive definition has limitations

- often only care about what is not working
- may not be able to characterize all possible failure modes
- want to prefer diagnoses that claim as few broken components as possible

Consistency-based diagnosis:

Assume KB uses the predicate Ab as before, but perhaps only characterizes the normal behaviour

e.g. $\mathrm{Andgate}(x) \wedge \neg\mathrm{Ab}(x) \supset \mathrm{out}(x) = \mathrm{and}(\mathrm{in1}(x), \mathrm{in2}(x))$

Want a minimal set of components D, such that

$\{\mathrm{Ab}(c) \mid c \in D\} \cup \{\neg\mathrm{Ab}(c) \mid c \notin D\}$

is consistent with KB $\cup$ *Settings* $\cup$ *Observations*

can use table as before with last column changed to "consistency"

In previous example, get 3 diagnoses: $\{\mathrm{b1}\}$, $\{\mathrm{b2}, \mathrm{a2}\}$ and $\{\mathrm{b2}, \mathrm{o1}\}$

Note: more complex to handle non-minimal diagnoses

Some complications

1. negative evidence
   - allow for missing observations

     e.g. ensure that KB $\cup \{\alpha\} \not\models$ fever

2. variables and quantification
   - same definition, modulo "simplicity", (but how to use Resolution?)
   - useful to handle open wffs also

     KB $\cup \{x = 3\} \models P(x)$ handles WH-questions

3. probabilities
   - not all simplest explanations are equally likely
   - also: replace (*Disease* $\wedge \ldots \supset$ *Symptom*) by a probabilistic version

4. defaults
   - instead of requiring KB $\cup \{\alpha\} \models \beta$, would prefer that given α, it is reasonable to believe β

     e.g. being a bird explains being able to fly

Other applications

1. object recognition

   what scene would account for image elements observed?

   what objects would account for collection of properties discovered?

2. plan recognition

   what high-level goals of an agent would account for the actions observed?

3. hypothetical reasoning

   instead of asking: what would I have to be told to believe β?
   ask instead: what would I learn if I was told that α?

Dual of explanation: want β such that

- $KB \cup \{\alpha\} \models \beta$
- $KB \not\models \beta$
- simplicity, parsimony
- using correct vocabulary

Solution: you learn β on being told α
iff
$\neg\beta$ is an explanation for $\neg\alpha$

can use the abduction procedure

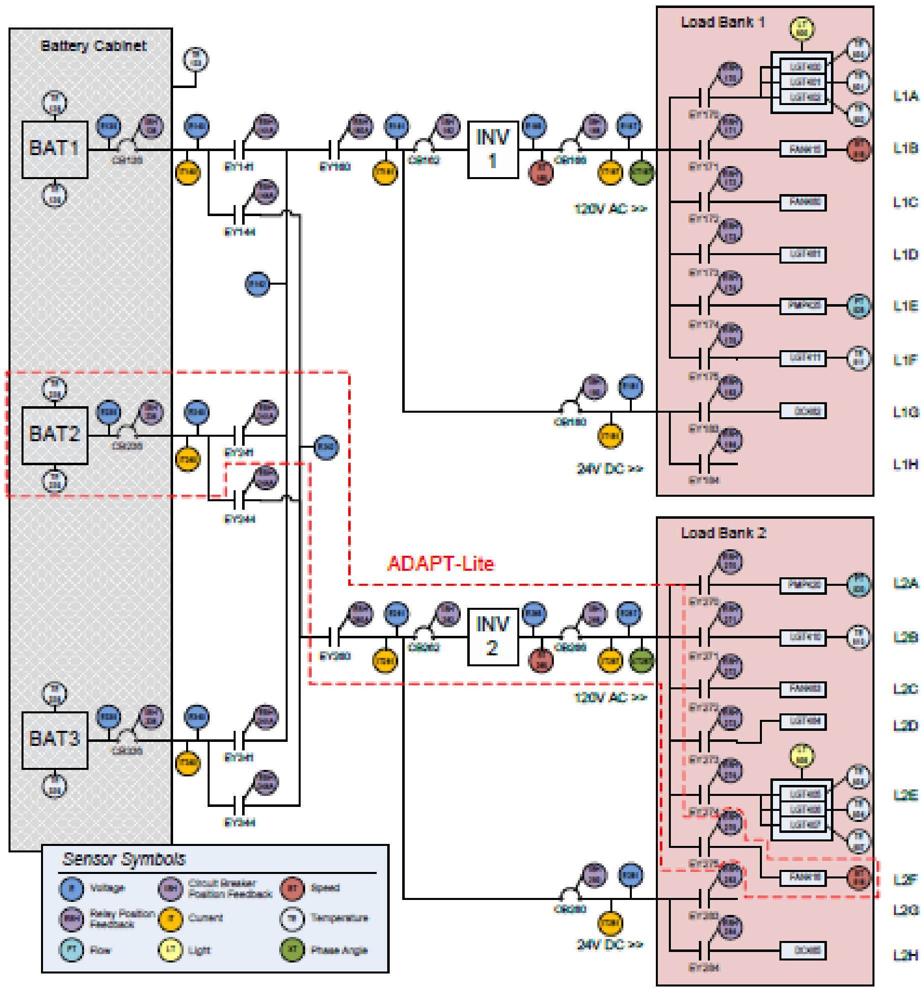
Related Conferences

- International Workshop on the Principles of Diagnosis

22nd International Workshop on Principles of Diagnosis

International Diagnostic Competition

Example Challenge Problem



Real World Diagnosis

- Key inferences
  - Fault detection
  - Fault isolation
  - Fault identification
  - Fault recovery
- Some challenges
  - Insufficient knowledge
  - Limited observability
  - Mixture of faults
  - Delayed effects
  - Fault Masking
- Prevalent techniques
  - Expert systems
  - Model based systems
  - Data driven systems
  - Stochastic systems
- Prevalent Algorithms
  - Mostly hybrid
  - Leverage model-based approaches
  - Consistency based inference

Suggested Reading

- Chapter 13 in Brachman & Levesque