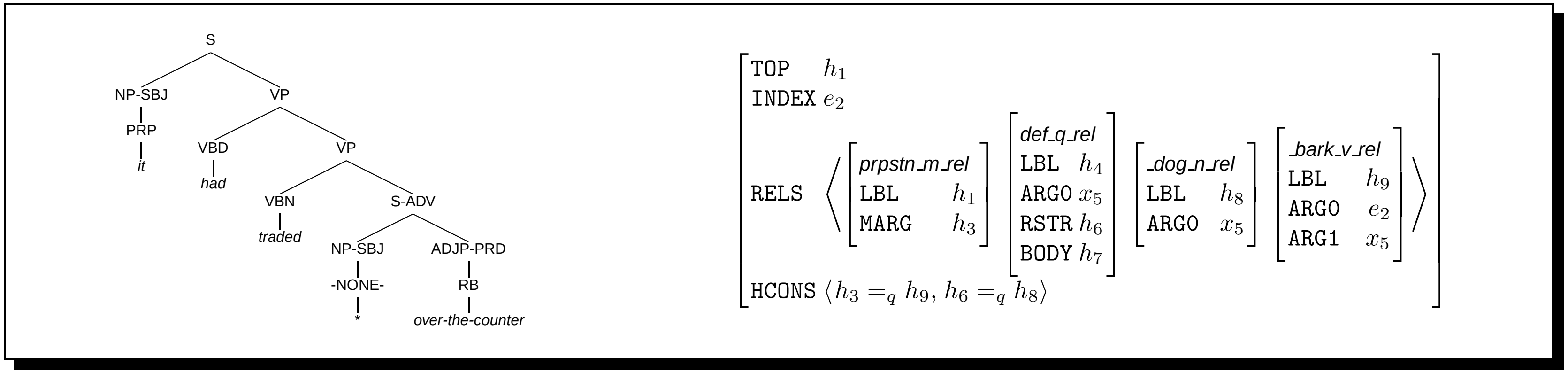


Large-Scale Computation for Language Technology

(In About Twenty Five Minutes)

Stephan Oepen

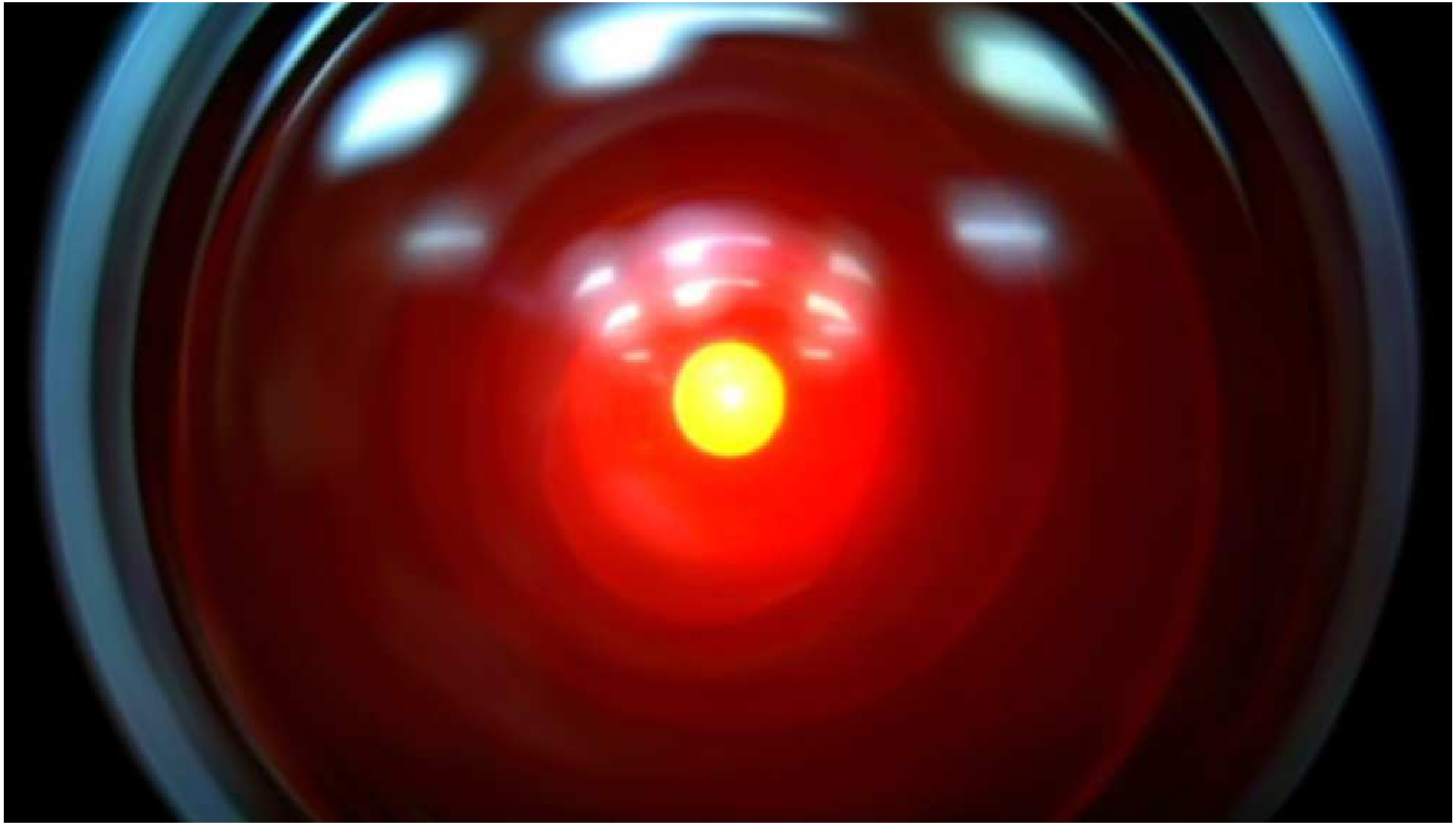
Universitetet i Oslo & CSLI Stanford

`oe@ifi.uio.no`

(Koenraad De Smedt and Christer Johansson, UiB)

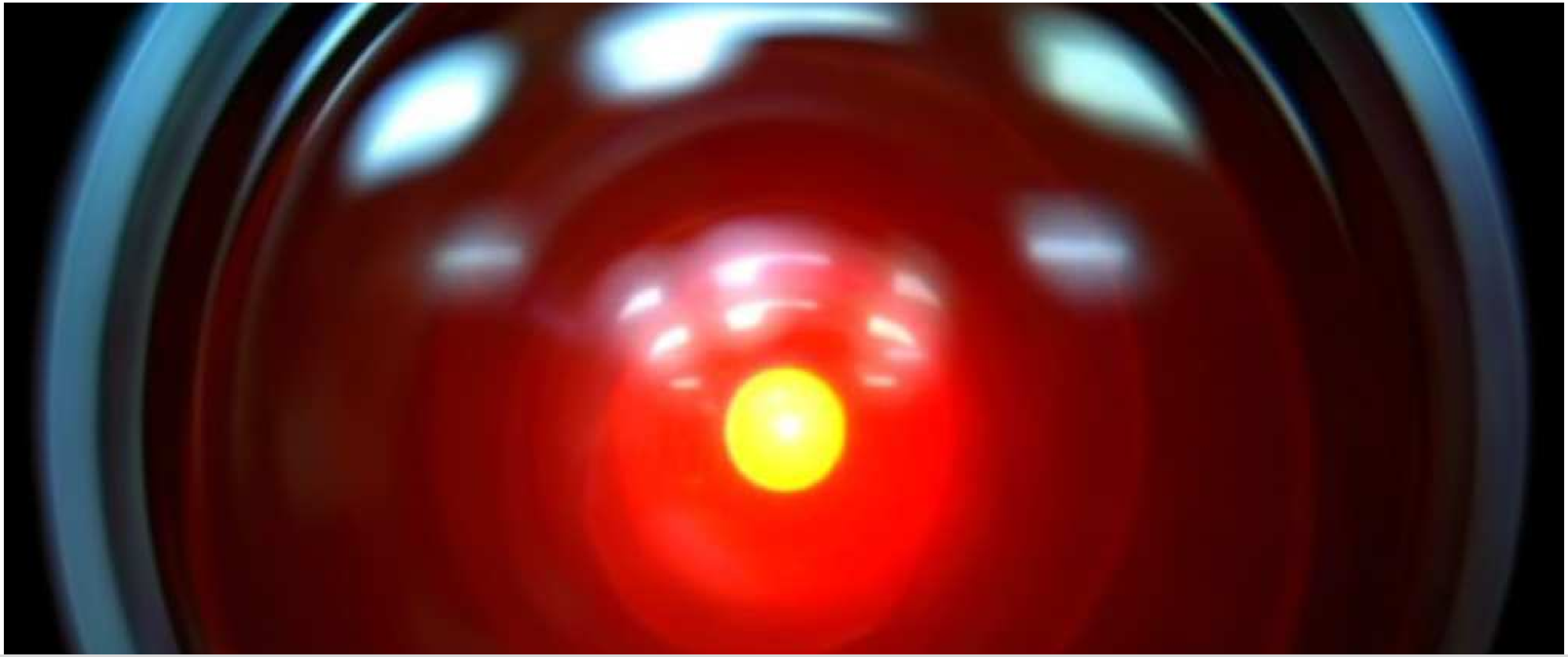
So, What Actually is Language Technology?

So, What Actually is Language Technology?



(2001: A Space Odyssey; HAL 9000; 1968)

So, What Actually is Language Technology?



→ (young) interdisciplinary science: language, cognition, computation;

→ (again) culturally and commercially relevant for 'knowledge society'.

Families of Language Processing Tasks

Speech Recognition and Synthesis

Summarization & Text Simplification

(High Quality) Machine Translation

Information Extraction — Text Understanding

Grammar & Controlled Language Checking

Natural Language Dialogue Systems

Families of Language Processing Tasks

Speech Recognition and Synthesis

Summarization & Text Simplification

(High Quality) Machine Translation

Information Extraction — Text Understanding

Grammar & Controlled Language Checking

Natural Language Dialogue Systems

Families of Language Processing Tasks

Speech Recognition and Synthesis

Summarization & Text Simplification

(High Quality) Machine Translation

Information Extraction — Text Understanding

Grammar & Cor … guage Checking

Natural Lang … ue Systems

Focus on Precision

Use of Semantics

(Mostly) English

What Makes Natural Language a Hard Problem?

```
< Den andre veien mot Bergen er kort. --- 12 x 30 x 25 = 25
> The other path towards Bergen is short. {0.58} (1:1:0).
> The other road towards Bergen is short. {0.56} (1:0:0).
> The second road towards Bergen is short. {0.55} (2:0:0).
> That other path towards Bergen is a card. {0.54} (0:1:0).
> That other road towards Bergen is a card. {0.54} (0:0:0).
> The second path towards Bergen is short. {0.51} (2:1:0).
> The other road against Bergen is short. {0.48} (1:2:0).
> The second road against Bergen is short. {0.48} (2:2:0).
  ...
> Short is the other street towards Bergen. {0.33} (1:4:0).
> Short is the second street towards Bergen. {0.33} (2:4:0).
  ...
```

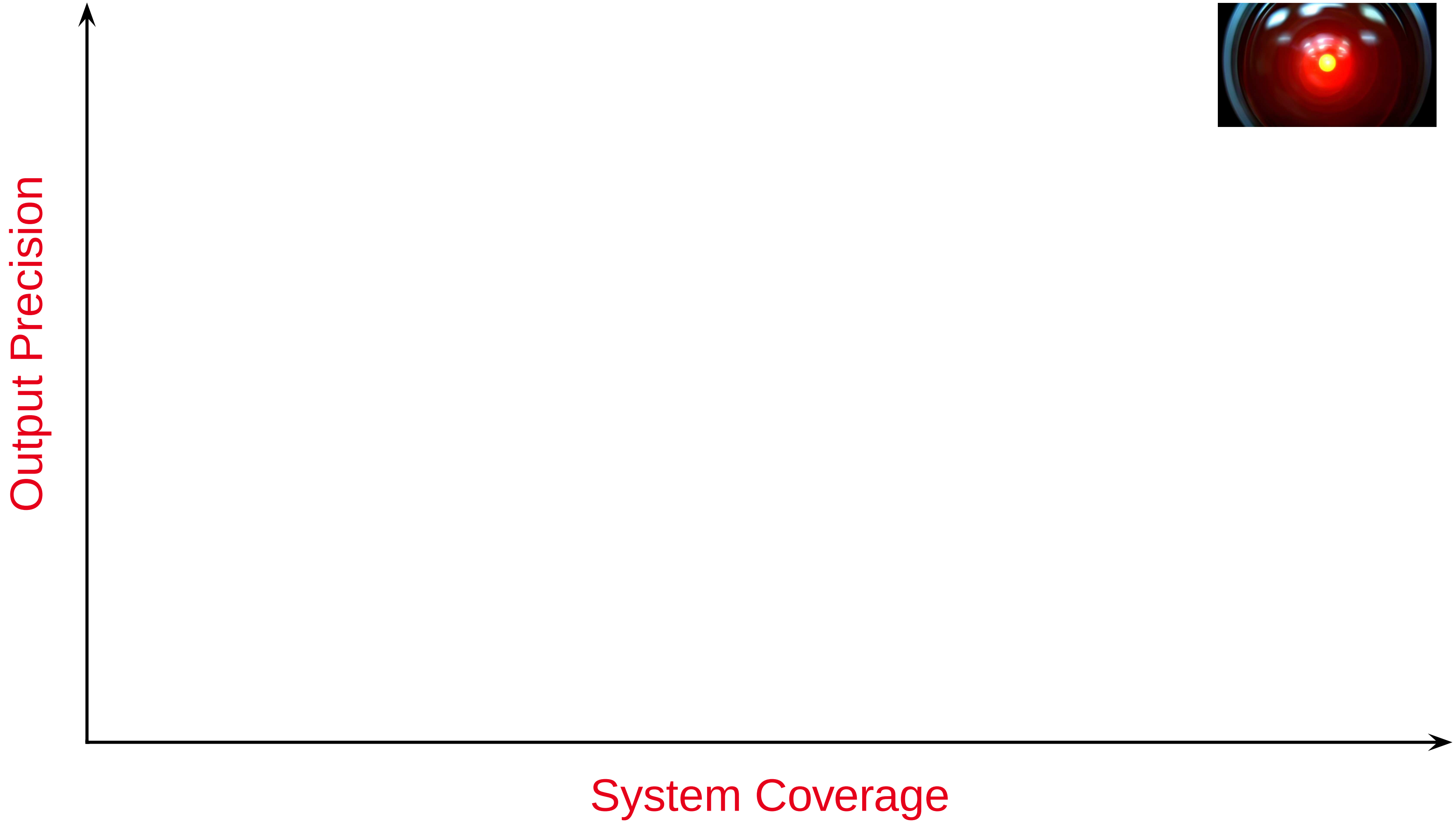
What Makes Natural Language a Hard Problem?

```
< Den andre veien mot Bergen er kort. --- 12 x 30 x 25 = 25
> The other path towards Bergen is short. {0.58} (1:1:0).
> The other road towards Bergen is short. {0.56} (1:0:0).
> The second road towards Bergen is short. {0.55} (2:0:0).
> That other path towards Bergen is a card. {0.54} (0:1:0).
> That other road towards Bergen is a card. {0.54} (0:0:0).
> The second path towards Bergen is short. {0.51} (2:1:0).
> The other road against Bergen is short. {0.48} (1:2:0).
> Th                                                     )).
  ...
> Sh                                                     0).
> Sh                                                     :0).
  ...
```

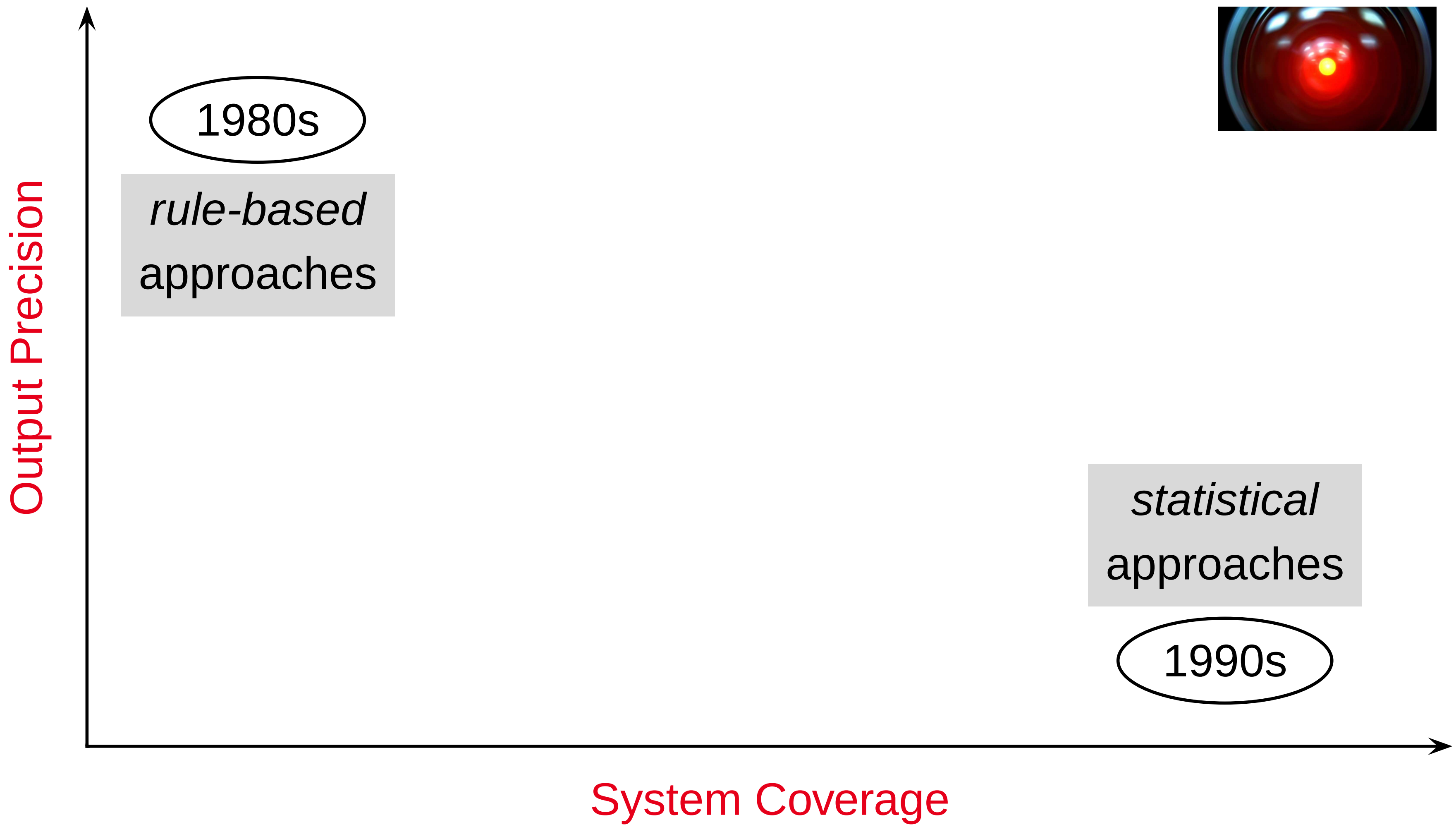
Scraped Off the Internet

```
the road to the other bergen is short .
Den other roads against Boron Gene are short.
Other one autobahn against Mountains am abrupt.
```

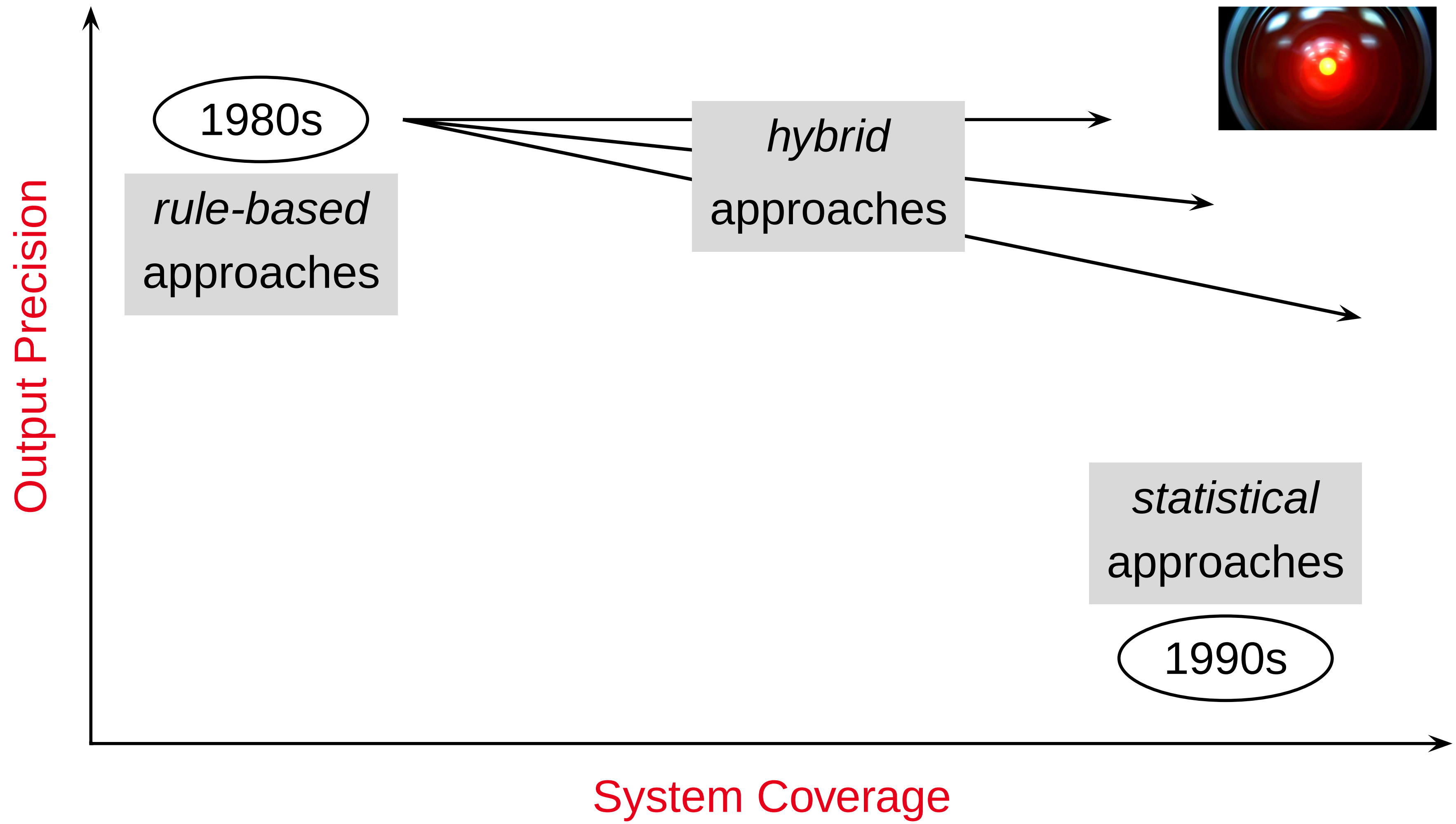
The Holy Grail: Balancing Coverage and Precision



The Holy Grail: Balancing Coverage and Precision



The Holy Grail: Balancing Coverage and Precision



A Tool Towards Understanding: (Formal) Grammar

Wellformedness

- *Kim was happy because ____ passed the exam.*
- *Kim was happy because ____ final grade was an A.*
- *Kim was happy when she saw ____ on television.*

A Tool Towards Understanding: (Formal) Grammar

Wellformedness

- *Kim was happy because ____ passed the exam.*
- *Kim was happy because ____ final grade was an A.*
- *Kim was happy when she saw ____ on television.*

Meaning

- *Kim gave Sandy the book.*
- *Kim gave the book to Sandy.*
- *Sandy was given the book by Kim.*

A Tool Towards Understanding: (Formal) Grammar

Wellformedness

- *Kim was happy because ____ passed the exam.*
- *Kim was happy because ____ final grade was an A.*
- *Kim was happy when she saw ____ on television.*

Meaning

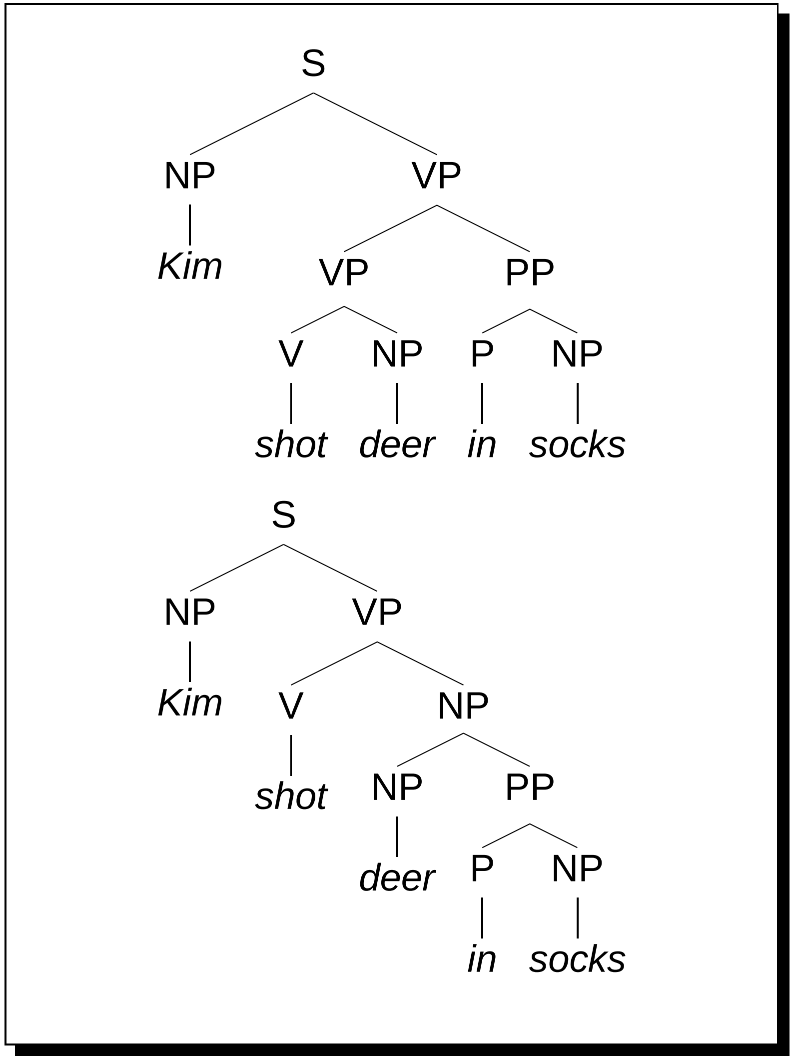
- *Kim gave Sandy the book.*
- *Kim gave the book to Sandy.*
- *Sandy was given the book by Kim.*

Ambiguity

- *I shot an elephant in my pyjamas.*
- *Have her report on my desk by Friday!*

Background: Context-Free Grammars

S → NP VP
VP → V NP
VP → VP PP
NP → NP PP
PP → P NP
NP → Kim | deer | socks
V → shot
P → in



All Complete Derivations

- are rooted in the start symbol S;
- label internal nodes with categories $\in C$, leafs with words $\in \Sigma$;
- instantiate a grammar rule $\in P$ at each local subtree of depth one.

Background: Context-Free Grammars

S → NP VP

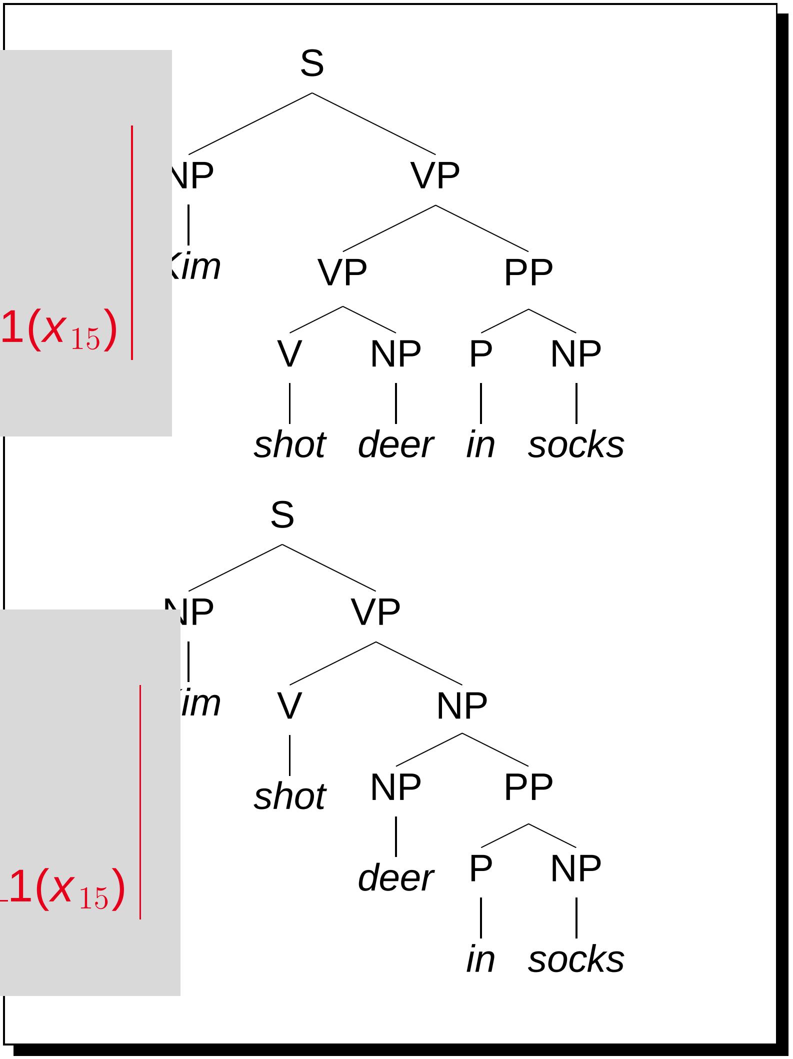
$$
\begin{aligned}
&\langle\, h_1, \\
&\;\big|\; h_3\text{:proper\_q}(x_5, h_4, h_6), h_7\text{:named}(x_5, \textit{Kim}), \\
&\;\big|\; h_8\text{:\_shoot\_v\_1}(e_2, x_5, x_9), \\
&\;\big|\; h_{10}\text{:udef\_q}(x_9, h_{11}, h_{12}), h_{13}\text{:\_deer\_n\_1}(x_9), \\
&\;\big|\; h_8\text{:\_in\_p}(e_{14}, e_2, x_{15}), h_{16}\text{:udef\_q}(x_{15}, h_{17}, h_{18}), h_{19}\text{:\_sock\_n\_1}(x_{15}) \;\big| \\
&\;\{\, h_4 =_q h_7, h_{11} =_q h_{13}, h_{17} =_q h_{19} \,\} \,\rangle
\end{aligned}
$$

V → shot

P → in

$$
\begin{aligned}
&\langle\, h_1, \\
&\;\big|\; h_3\text{:proper\_q}(x_5, h_4, h_6), h_7\text{:named}(x_5, \textit{Kim}), \\
&\;\big|\; h_8\text{:\_shoot\_v\_1}(e_2, x_5, x_9), \\
&\;\big|\; h_{10}\text{:udef\_q}(x_9, h_{11}, h_{12}), h_{13}\text{:\_deer\_n\_1}(x_9), \\
&\;\big|\; h_{13}\text{:\_in\_p}(e_{14}, x_9, x_{15}), h_{16}\text{:udef\_q}(x_{15}, h_{17}, h_{18}), h_{19}\text{:\_sock\_n\_1}(x_{15}) \;\big| \\
&\;\{\, h_4 =_q h_7, h_{11} =_q h_{13}, h_{17} =_q h_{19} \,\} \,\rangle
\end{aligned}
$$

each local subtree of depth one.



Complex Categories: Unification-Based Grammars

$$phrase\begin{bmatrix} \text{ORTH} & \textit{“Kim”} \\ \text{HEAD} & \textit{noun} \\ \text{SPR} & \langle\rangle \\ \text{COMPS} & \langle\rangle \end{bmatrix} \qquad phrase\begin{bmatrix} \text{ORTH} & \textit{“sleeps”} \\ \text{HEAD} & \textit{verb} \\ \text{SPR} & \left\langle \begin{bmatrix} \text{HEAD} & \textit{noun} \\ \text{SPR} & \langle\rangle \\ \text{COMPS} & \langle\rangle \end{bmatrix} \right\rangle \\ \text{COMPS} & \langle\rangle \end{bmatrix}$$

$$phrase\begin{bmatrix} \text{HEAD} & \boxed{1} \\ \text{SPR} & \langle\rangle \\ \text{COMPS} & \boxed{3} \end{bmatrix} \longrightarrow \boxed{2}\, phrase\begin{bmatrix} \text{SPR} & \langle\rangle \\ \text{COMPS} & \langle\rangle \end{bmatrix}, \quad phrase\begin{bmatrix} \text{HEAD} & \boxed{1} \\ \text{SPR} & \langle\boxed{2}\rangle \\ \text{COMPS} & \boxed{3} \end{bmatrix}$$

Symbolic Computation: A Few Figures

- Each category is a DAG of ~300 nodes; ~80 bytes average node size;
- ~4,000 top-level unifications per cpu second; accessing ~180 mbytes;
- parse time for ~20-word sentence ~4 seconds; thousands of analyses.

(Unification-Based) HPSG Parsing — Then and 'Now'

Version	Platform	Test Set	filter %	etasks ϕ	pedges ϕ	tcpu ϕ (s)	space ϕ (kb)
October 1996	PAGE	*'tsnlp'*	49·9	656	44	4·77	19,016
		'aged'	51·3	1763	**97**	**36·69**	79,093
August 2000	PET	*'tsnlp'*	93·9	170	55	0·03	333
		'aged'	95·1	753	**292**	**0·14**	1,435
		'fuse'	95·5	3084	1140	0·65	10,589

(generated by [incr tsdb()] at 5-nov-2000 (21:23 h))

Cumulative Break-Through in Parsing Efficiency

- Oldest comparable profiles: net speed-up of around 260 (excluding gc);
- grammar evolution: problem size (in edges) increased by factor of three;
- additional factors (hardware, packing): above four orders of magnitude;

→ Unification-based parsing nowadays applied at 'Web scale' (PowerSet).

A Typical Work Day in an Empirical Science

07:30 – 08:00	Check email; read detailed progress reports from others;
09:00 – 09:30	`cvs update`; **profile and validate** the standard test suites;
09:30 – 10:00	`cvs commit lingo`; **verify** latest release of HPSG system;
10:00 – 12:00	`cvs update english`; **evaluate** latest English generation;
13:00 – 14:00	MRS meeting with grammarians; write up new proposal;
14:00 – 14:30	`cvs update transfer`; **confirm** that bug 'fix' finally works;
14:30 – 15:00	upload head revision into on-line demonstrator; **validate**;
15:00 – 17:00	implement and **test** MRS for nominalization compounds;
17:00 – 17:30	update comments; `cvs commit`; progress report to others.

Near-Instantaneous Regression Testing

Background

- Grammars, development and production systems evolve permanently;
- English Resource Grammar (ERG) continuously developed since 1993;
- seemingly simple changes will often have unexpected, global effects;
- automated regression testing, progress reporting, and error detection;

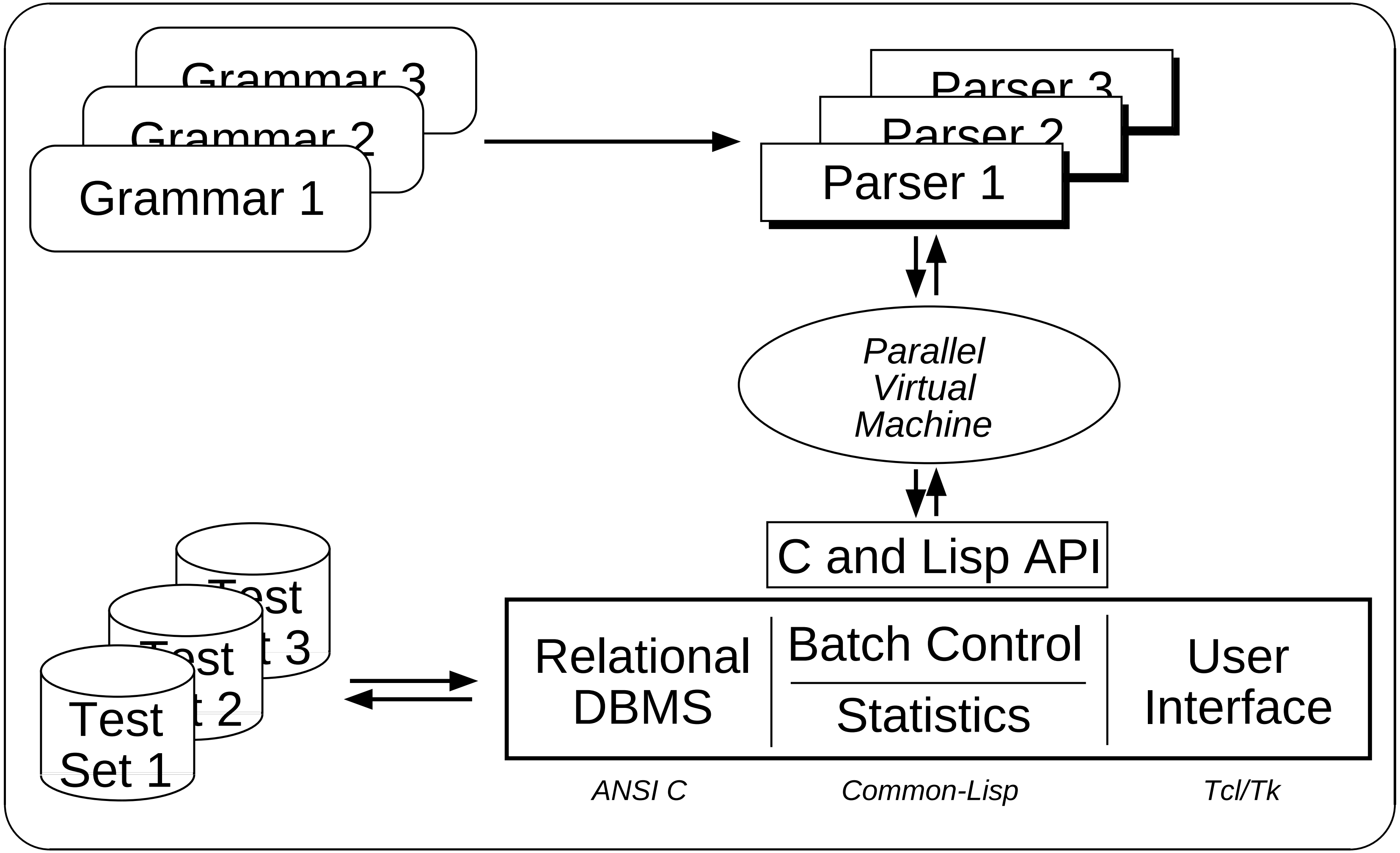
→ enable *all* developers to perform unit testing *and* end-to-end evaluation.

Technological Advances

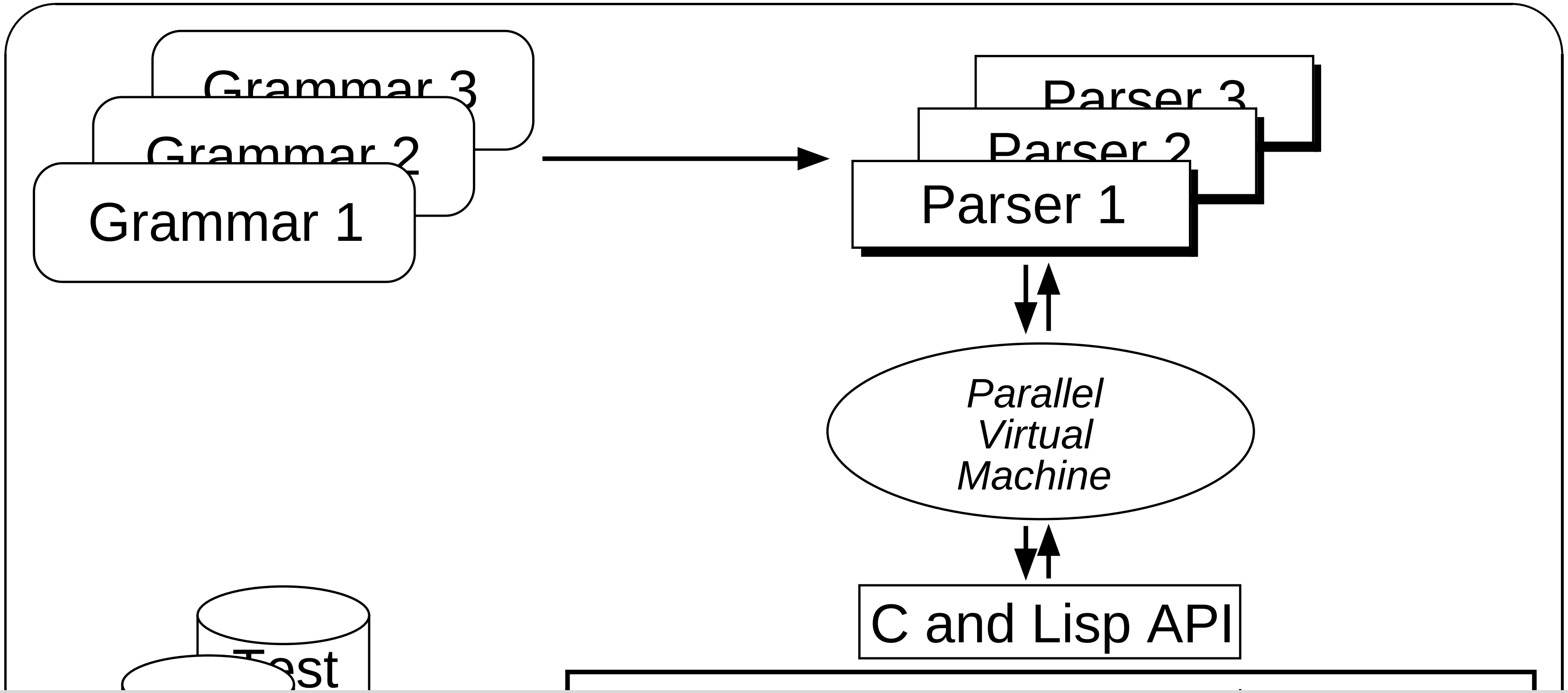
- Parallel processing allows dozens to hundreds of test runs each day;

→ *profiling* full grammar and software evolution recorded in database.

General Architecture: [incr tsdb()] Profiler

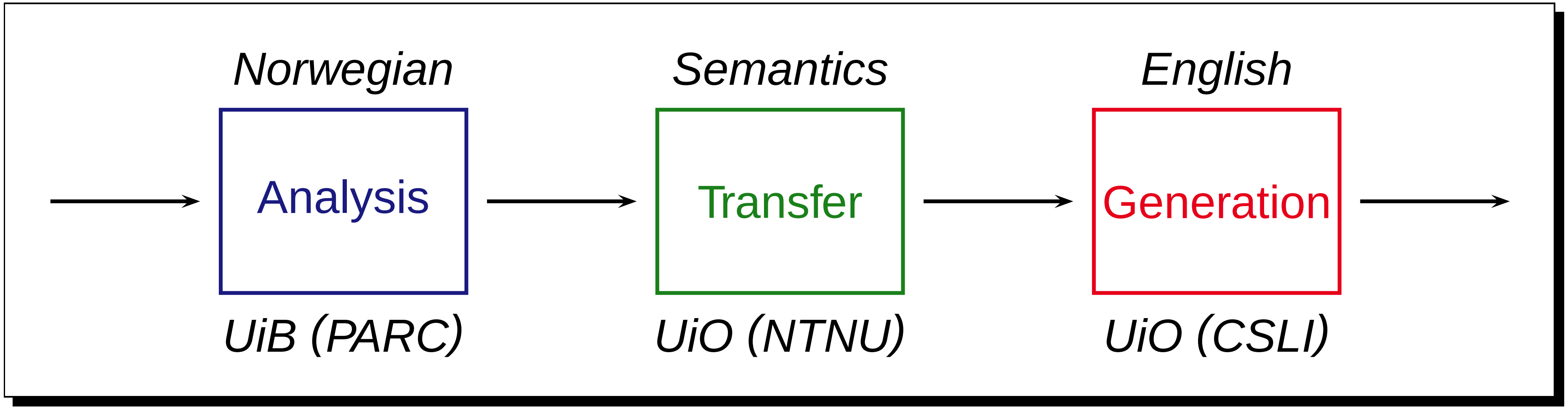


General Architecture: [incr tsdb()] Profiler

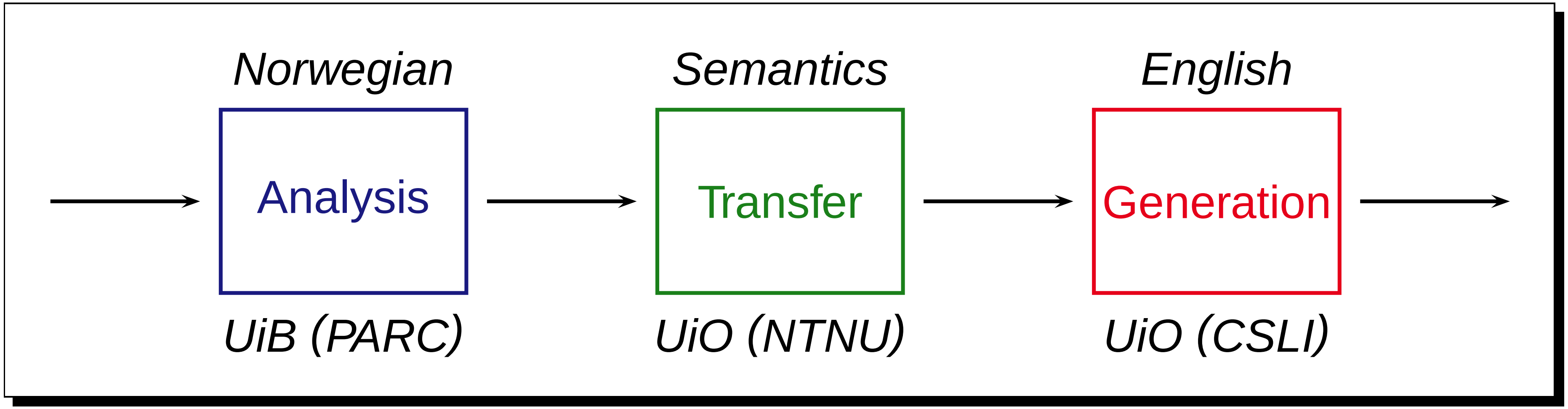


Originally developed in mid-1990s (based on PVM);
relatively large messages: full serialization of all parses;
scalability issues; unable to drive more than ∼20 clients.

An Example: Machine Translation in LOGON



An Example: Machine Translation in LOGON

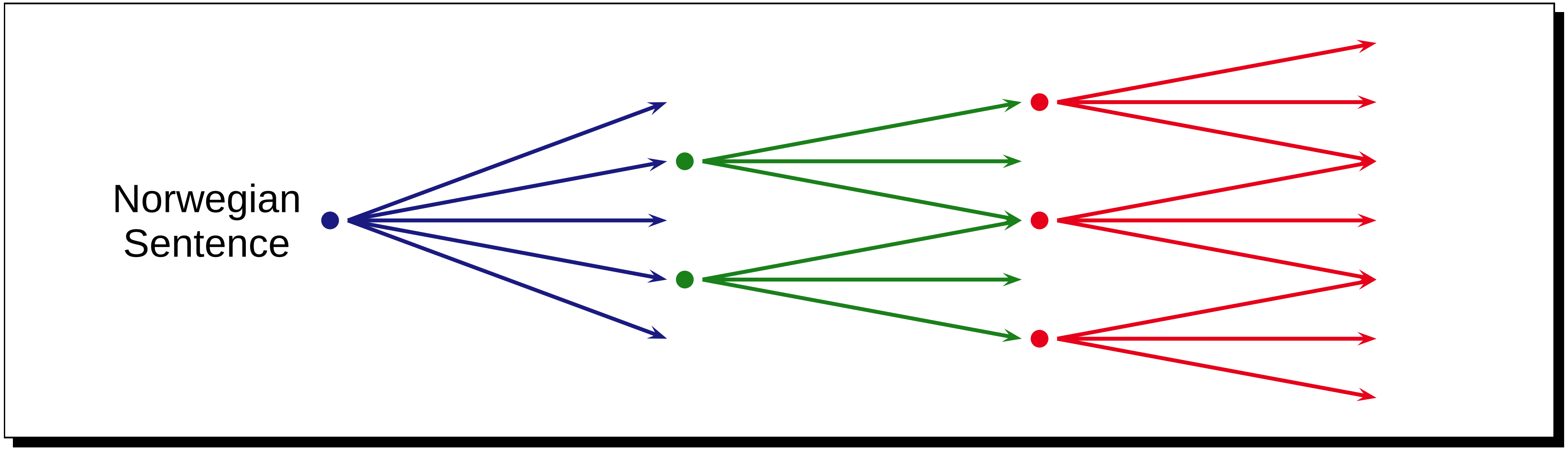


Some LOGON Highlights

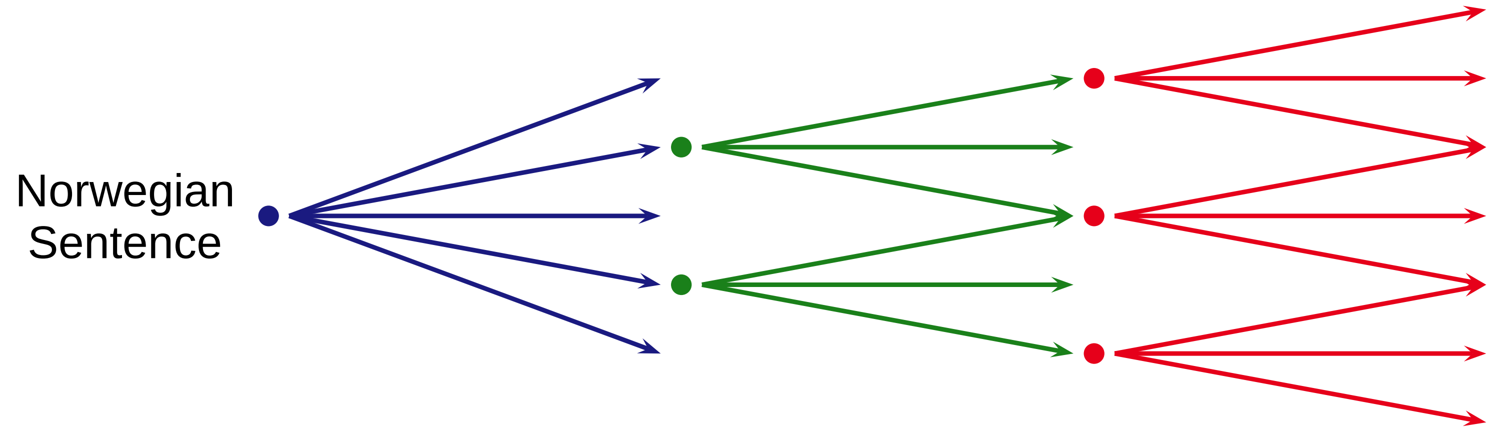
- Re-usable, mono-lingual precision grammars as linguistic back-bone;
- pipeline of three heavy-weight, unification-based search processes;

→ limited domain and vocabulary: very competitive with state-of-the-art.

Ambiguity Management: Stochastic Processes



Ambiguity Management: Stochastic Processes



Stochastic Elements in LOGON

- At each stage, rank alternate hypotheses — finally, re-rank globally;
- probabilistic modeling: model acquisition computationally intensive;

→ hybrid MT: linguistic back-bone combined with advanced statistics.

Ambiguity Resolution Remains a (Major) Challenge

The Problem

- With broad-coverage grammars, even moderately complex sentences typically have multiple analyses (tens or hundreds, up to tens of thousands);
- unlike in grammar writing, exhaustive parsing is useless for applications;
- identifying the 'right' (i.e. intended) analysis is an 'AI-complete' problem;
- inclusion of (non-grammatical) sortal constraints is generally undesirable.

Typical Approaches

- Design and use statistical models to select among competing analyses;
- for string s, some analyses t_i are more or less likely: maximize $P(t_i|s)$;

$\rightarrow$ Probabilistic Context Free Grammar (PCFG) is a CFG plus probabilities.

Probability Theory and Linguistics?

The most important questions of life are, for the most part, really only questions of probability. (Pierre-Simon Laplace, 1812)

Probability Theory and Linguistics?

The most important questions of life are, for the most part, really only questions of probability. (Pierre-Simon Laplace, 1812)

Special wards in lunatic asylums could well be populated with mathematicians who have attempted to predict random events from finite data samples. (Richard A. Epstein, 1977)

Probability Theory and Linguistics?

The most important questions of life are, for the most part, really only questions of probability. (Pierre-Simon Laplace, 1812)

Special wards in lunatic asylums could well be populated with mathematicians who have attempted to predict random events from finite data samples. (Richard A. Epstein, 1977)

But it must be recognized that the notion 'probability' of a sentence is an entirely useless one, under any known interpretation of this term. (Noam Chomsky, 1969)

Probability Theory and Linguistics?

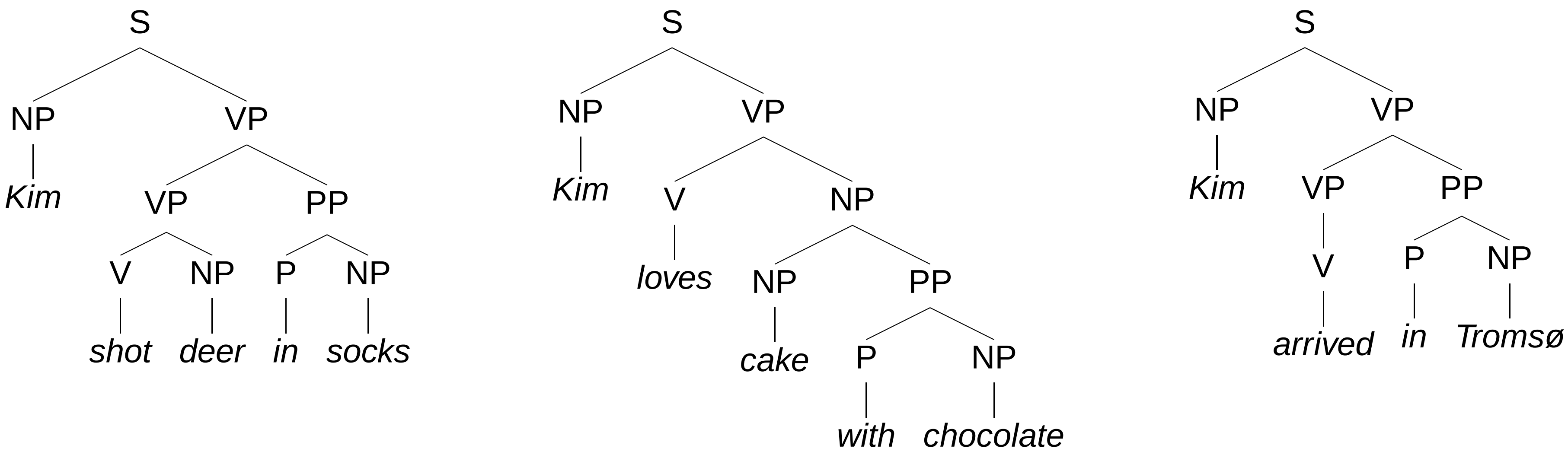
The most important questions of life are, for the most part, really only questions of probability. (Pierre-Simon Laplace, 1812)

Special wards in lunatic asylums could well be populated with mathematicians who have attempted to predict random events from finite data samples. (Richard A. Epstein, 1977)

But it must be recognized that the notion 'probability' of a sentence is an entirely useless one, under any known interpretation of this term. (Noam Chomsky, 1969)

Every time I fire a linguist, system performance improves. (Fredrick Jelinek, 1980s)

A (Simplified) PCFG Estimation Example



P(RHS|LHS)	CFG Rule
$3/3 = 1.00$	S $\rightarrow$ NP VP
$2/4 = 0.50$	VP $\rightarrow$ VP PP
$1/4 = 0.25$	VP $\rightarrow$ V NP
$1/4 = 0.25$	VP $\rightarrow$ V
$\vdots$	PP $\rightarrow$ P NP

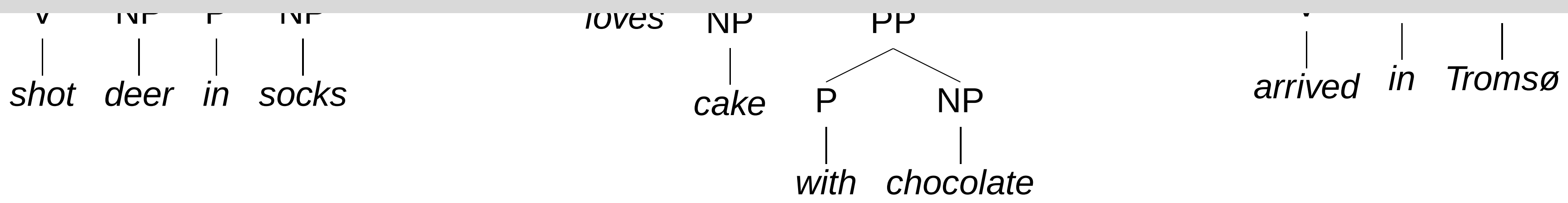
- Estimate rule probability from observed distribution;

$\rightarrow$ conditional probabilities:

$$\mathrm{P(RHS|LHS)} = \frac{C(\mathrm{LHS, RHS})}{C(\mathrm{LHS})}$$

A (Simplified) PCFG Estimation Example

Requires *treebank* as training data: collection of ‘correct’ trees;
can be annotated manually or semi-automatically (disambiguation);
Penn Treebank, for example: one million words of newspaper text.



P(RHS|LHS)	CFG Rule
$3/3 = 1.00$	S $\rightarrow$ NP VP
$2/4 = 0.50$	VP $\rightarrow$ VP PP
$1/4 = 0.25$	VP $\rightarrow$ V NP
$1/4 = 0.25$	VP $\rightarrow$ V
$\vdots$	PP $\rightarrow$ P NP

- Estimate rule probability from observed distribution;

$\rightarrow$ conditional probabilities:

$$P(\text{RHS}|\text{LHS}) = \frac{C(\text{LHS}, \text{RHS})}{C(\text{LHS})}$$

Stochastic Unification-Based Grammars

Conditional Parse Selection

- Local independence assumption is not true for unification grammars;
- PCFG unable to 'learn' from negative data, e.g. dis-preferred parses;

$\rightarrow$ *conditional* model: given some context, sample properties of events.

Maximum Entropy Ranking

Given a sentence s and a set of trees $\{t_1 \ldots t_n\}$ assigned to s by some grammar, find the tree t_i that maximizes $p(t_i|s)$. Assuming a set of features $\{f_1 \ldots f_m\}$ with corresponding weights $\{\lambda_1 \ldots \lambda_m\}$, the conditional probability for tree t_i is given by:

$$p(t_i|s) = \frac{\exp \Sigma_j \lambda_j f_j(t_i)}{\Sigma_{k=1 \ldots n} \exp \Sigma_j \lambda_j f_j(t_k)} \quad (1)$$

Stochastic Unification-Based Grammars

Conditional Parse Selection

- Local independence assumption is not true for unification grammars;
- PCFG unable to 'learn' from negative data, e.g. dis-preferred parses;

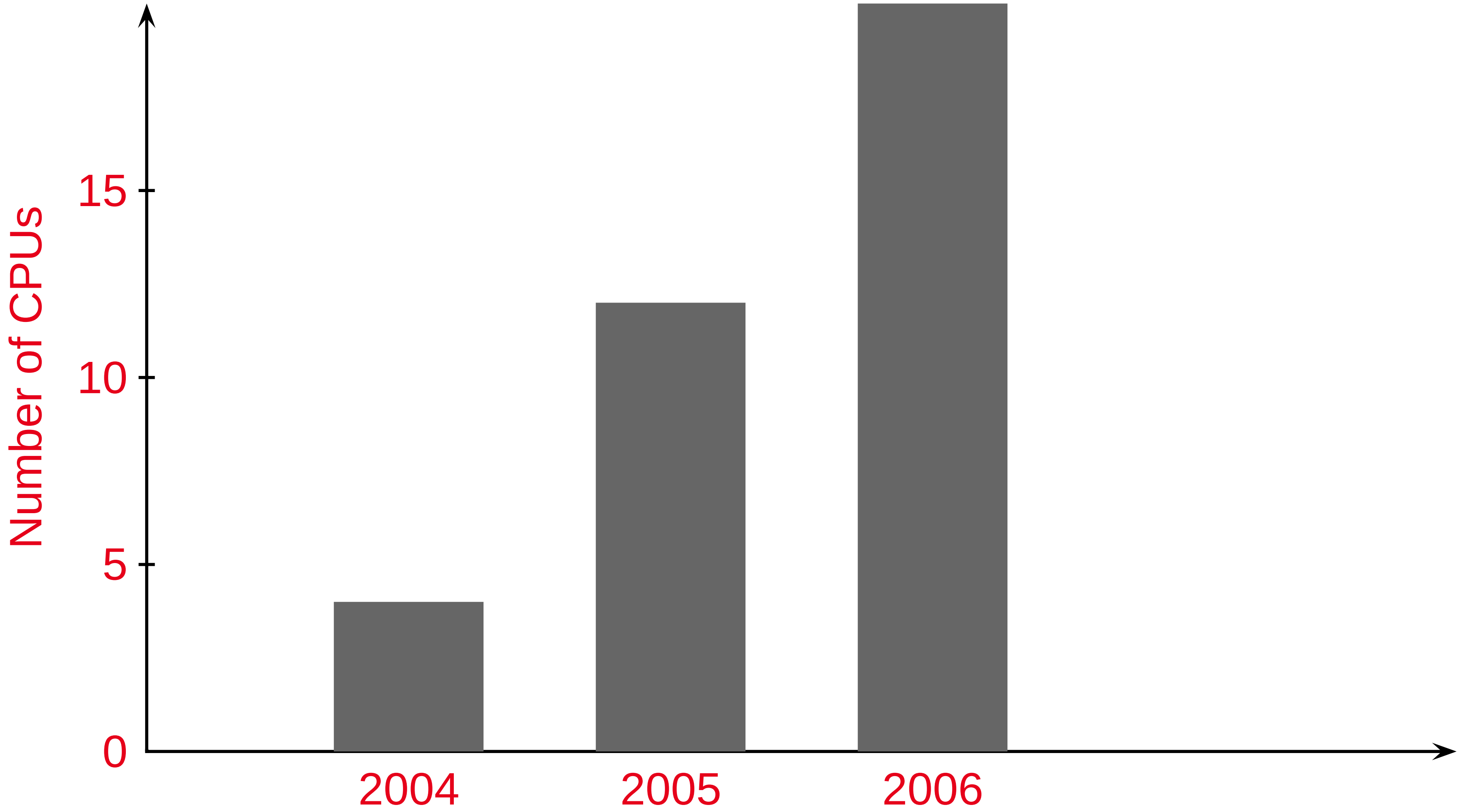
→ *conditional* model: given some context, sample properties of events.

Maximum Entropy Ranking

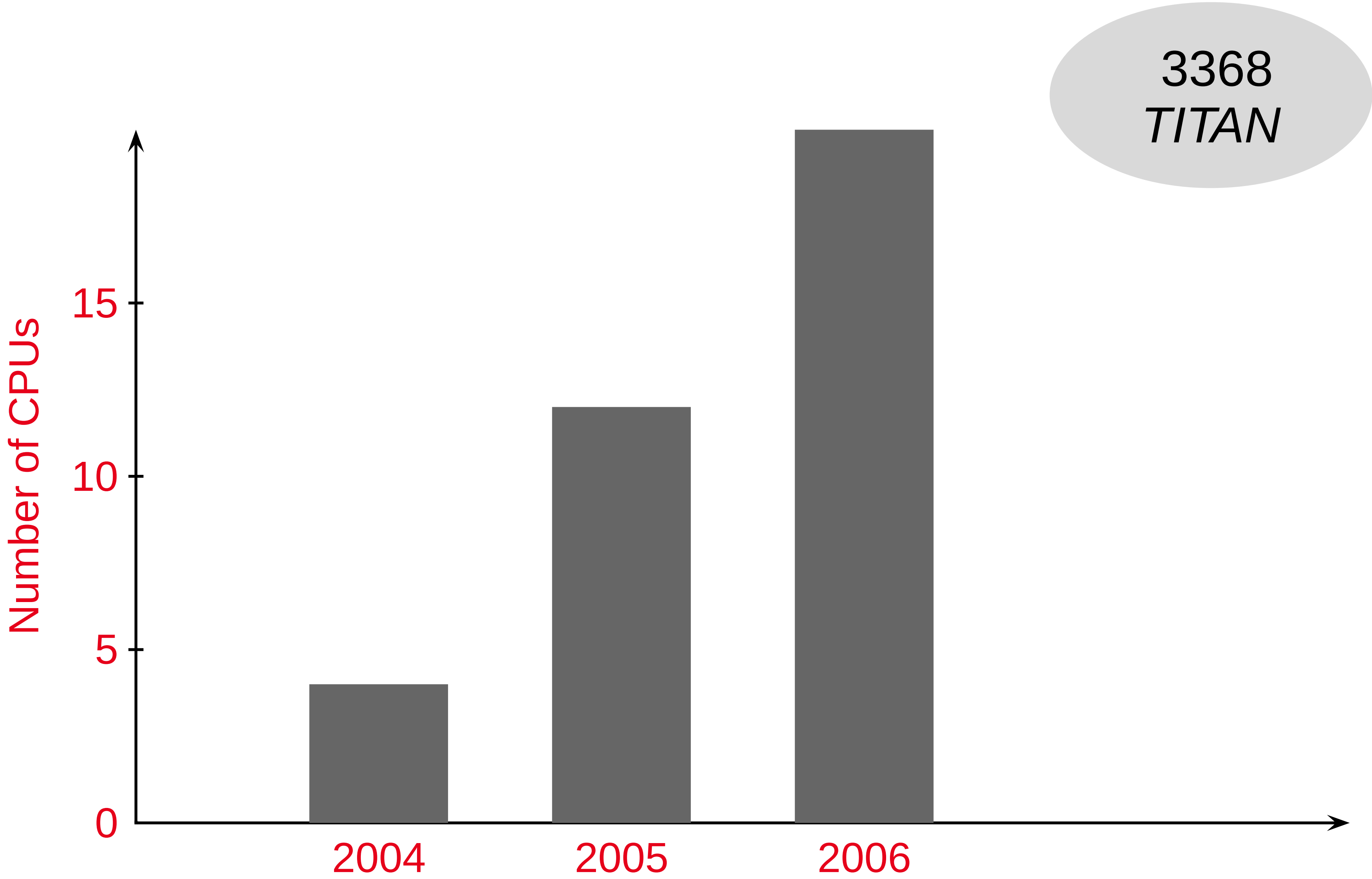
Given a sentence s and a set of trees $\{t_1 \ldots t_n\}$ assigned to s by some grammar, find the tree t_i that maximizes $p(t_i|s)$. Assuming a set of fea-

Hinoki Treebank: 65,424 sentences → 5,255,925 training instances;
3,310,202 distinct feature types (cached as 106-gbyte BerkeleyDB);
iterative parameter estimation *often* converges within a few hours;
cross validation, hyper-parameter search → hundreds of experiments.

High-Performance Computing in LOGON



High-Performance Computing in LOGON



Some Preliminary Conclusions — Outlook

Emerging Language Technology Trends

- eScience: information extraction from scholarly texts, e.g. Wikipedia;
- for example ontology learning, eventually reasoning; find contradictions;

→ Web-scale language technology, work towards *natural language search*.

Confluence of Approaches

- Fashion of the year: *hybridization*, balance of linguistics and statistics;
- large-scale data manipulation and computation increasingly important;

→ we still need to learn more about *large-scale* distributed computing.

Some Preliminary Conclusions — Outlook

Emerging Language Technology Trends

- eScience: information extraction from scholarly texts, e.g. Wikipedia;
- for example ontology learning, eventually reasoning; find contradictions;

→ Web-scale language technology, work towards *natural language search*.

Confluence of Approaches

- Fashion of the year: *hybridization*, balance of linguistics and statistics;
- large-scale data manipulation and computation increasingly important;

→ we still need to learn more about *large-scale* distributed computing

Please keep growing HPC infrastructure, we're about to discover it!

Some Preliminary Conclusions — Outlook

Emerging Language Technology Trends

- eScience: information extraction from scholarly texts, e.g. Wikipedia;
- for example ontology learning, eventually reasoning; find contradictions;

→ Web-scale language technology, work towards *natural language search*.

Confluence of Approaches

- Fashion of the year: *hybridization*, balance of linguistics and statistics;
- large-scale data manipulation and computation increasingly important;

→ we still need to learn more about *large-scale* distributed computing

(But please remember to put in insane amounts of memory!)

The IFI (Logic and) Natural Languages Group

Language Technology

Doctoral Fellow	Liv Ellingsen	Soft Grammatical Constraints
Professor	Tore Langhom	Logical-Form Semantics
Professor	*Jan Tore Lønning*	Computational Semantics
Professor	Stephan Oepen	Constraint-Based Processing
Doctoral Fellow	Erik Velldal	Machine Learning
Doctoral Fellow	Gisle Ytrestøl	Incremental Parsing
One PhD and Three MSc Candidates Finishing in 2008		

Computational Logic

Assistant Professor	Asbjørn Brændeland	Functional Programming
Professor	Herman Ruge Jervell	Proof Theory

Another (Albeit Somewhat Dubious) Vision

[Dave] *Open the pod bay doors, HAL.*

[HAL] *I'm sorry Dave, I'm afraid I can't do that.*

[Dave] *What's the problem?*

[HAL] *I think you know what the problem is just as well as I do.*

[Dave] *What are you talking about, HAL?*

[HAL] *This mission is too important for me to allow you to jeopardize it.*

...

[HAL] *Dave, this conversation can serve no purpose anymore. Goodbye.*

Some Sample Translations (And Errors)

1 *Velkommen til Jotunheimen!*
Welcome to Jotunheimen.

1037 *På vestbredden lå det der tre setre nesten ved siden av hverandre.*
On the west bank, 3 mountain pastures lay there almost beside each other.

1048 *Vil du ikke gå så langt, er Besstrondrundhø et utmerket alternativ.*
If you don't want to go so far, Besstrondrundhø is an excellent alternative.

1376 *Den toppen er et fint turmål om du bor på Bessheim eller Gjendesheim.*
That summit, a nice trip tongue is if you stay at Bessheim or Gjendesheim.

Some Sample Translations (And Errors)

1 *Velkommen til Jotunheimen!*
Welcome to Jotunheimen.

1037 *På vestbredden lå det der tre setre nesten ved siden av hverandre.*
On the west bank, 3 mountain pastures lay there almost beside each other.

1048 *Vil du ikke gå så langt, er Besstrondrundhø et utmerket alternativ.*
If you don't want to go so far, Besstrondrundhø is an excellent alternative.

1376 *Den* … *r Gjen-desh* …
That … r Gjen-desh …

Google Translate

```
Do not want to go so far,
is Besstrondrundhø an excellent alternative.
```