MT SUMMIT 2013

EUROPEAN ASSOCIATION FOR MACHINE TRANSLATION

Machine Translation Summit XIV, Nice, France
2 - 6 September 2013

Big Data Adaptation

DatAptor

Dr Khalil Sima'an

Institute for Logic, Language and Computation

University of Amsterdam

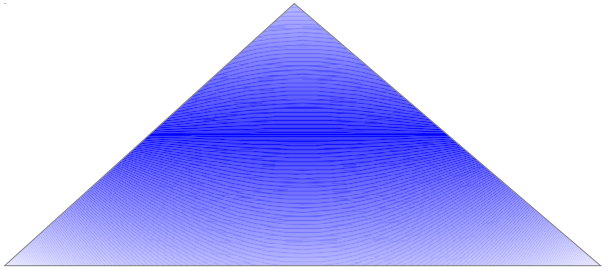
The Netherlands

UNIVERSITEIT VAN AMSTERDAM

MT at ILLC-UvA

Statistical Language Processing and Learning Lab.

Language Technology
Machine Translation

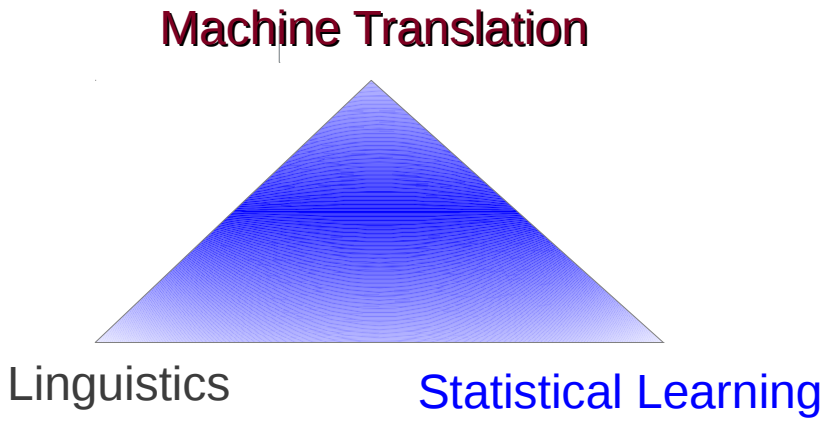


Linguistic structure

Statistical Learning

MT at ILLC-UvA

Statistical Language Processing and Learning Lab.



Main topics within SLPL

- Syntax-driven SMT (learning, decoding)
- Learning latent reordering for translation
- Hierarchical models with morph. And syntax
- Data-powered Adaptation
- ...

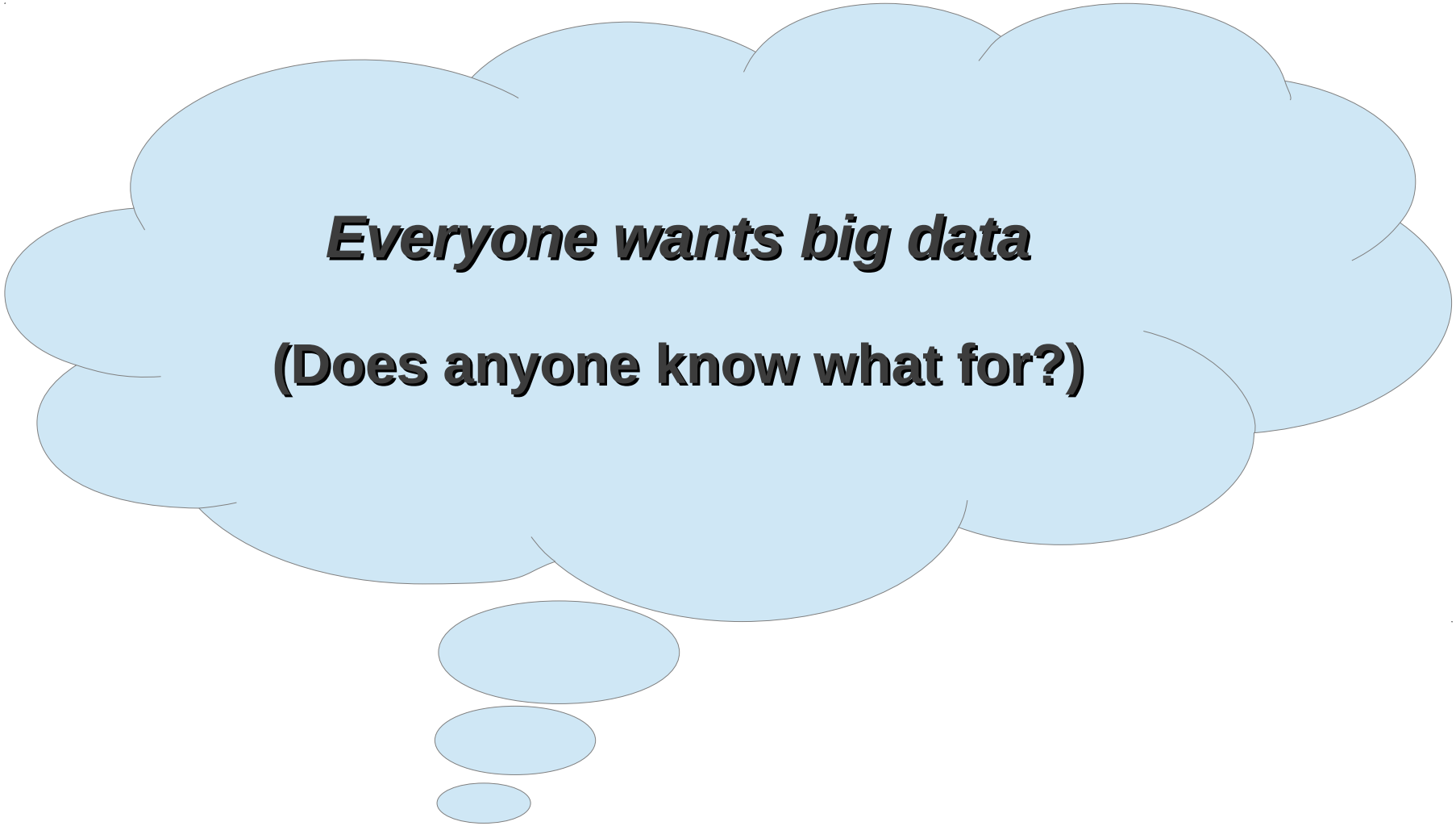
- **SLPL Lab** (Growing: 8 PhD students; 4 postdocs; Programmer)
- **Five projects on Statistical MT** (2012—2019)

This talk: Big Data and DatAptor (Feb 2013 - Feb 2017)

BIG DATA
What Comes to Mind?

BIG DATA

BIG DATA



BIG DATA
What comes to mind?

- **Efficient computing**

  Big storage + Fast search

- **Diversity**

  Quality differences

  Noise; Difficult statistics ...

- **Saturated statistics**

  Just count and divide

  Simple models are enough

  [cf. The Unreasonable Effectiveness of Data. Halevy, Norvig and Pereira 2009.]

BIG DATA
What comes to mind?

- **Efficient computing**

  Storage + search.

- **Diversity**

  Quality differences.

  Noise, difficult statistics ...

- **Saturated statistics**

  Just count and divide

  Simple models are enough.

  [cf. The Unreasonable Effectiveness of Data. Halevy, Norvig and Pereira 2009.]

Diversity also offers advantages

- Language use (different domains)
- Translators practice and guidelines

BIG DATA
What comes to mind?

- **Efficient computers**

  Storage + search.

- **Diversity**

  Quality differences.

  Noise, difficult statistics ...

- **Saturated statistics**

  Just count and divide

  Simple models are enough.

  [cf. The Unreasonable Effectiveness of Data. Halevy, Norvig and Pereira 2009.]

Diversity also offers advantages

- Language use (different domains)
- Translators practice and guidelines

DatAptor Project

DatAptor Project 2013-2017

Technology Foundation STW (NWO)

- Data-Powered
- Domain-Specific
- Translation Services
- on Demand

Partners (User Board)

- TAUS
- EC DGT
- Intel Inc.
- Symantec

Researchers (SLPL, ILLC, UvA)

- Dr Khalil Sima'an (principal investigator)
- Dr Christof Monz (senior researcher)
- Dr Bart Mellebeek (postdoc: Jan 2013)
- Milos Stanojevic (PhD student: March 2013)
- *Vacancies: postdoc + programmer*

DatAptor Motivation

Technology Foundation STW (NWO)

Versatile adaptation needed

- **Potential demand vs. current demand**
  - Continuously increasing text volumes
  - Large variability in kinds of texts (domains)

- **Changing translation market**
  - Changing domains, e.g. shifting international trade/cultural/... exchange
  - Changing acceptance for automatic translation

Versatile adaptation of MT Engines? How?

So many domains... so little time...

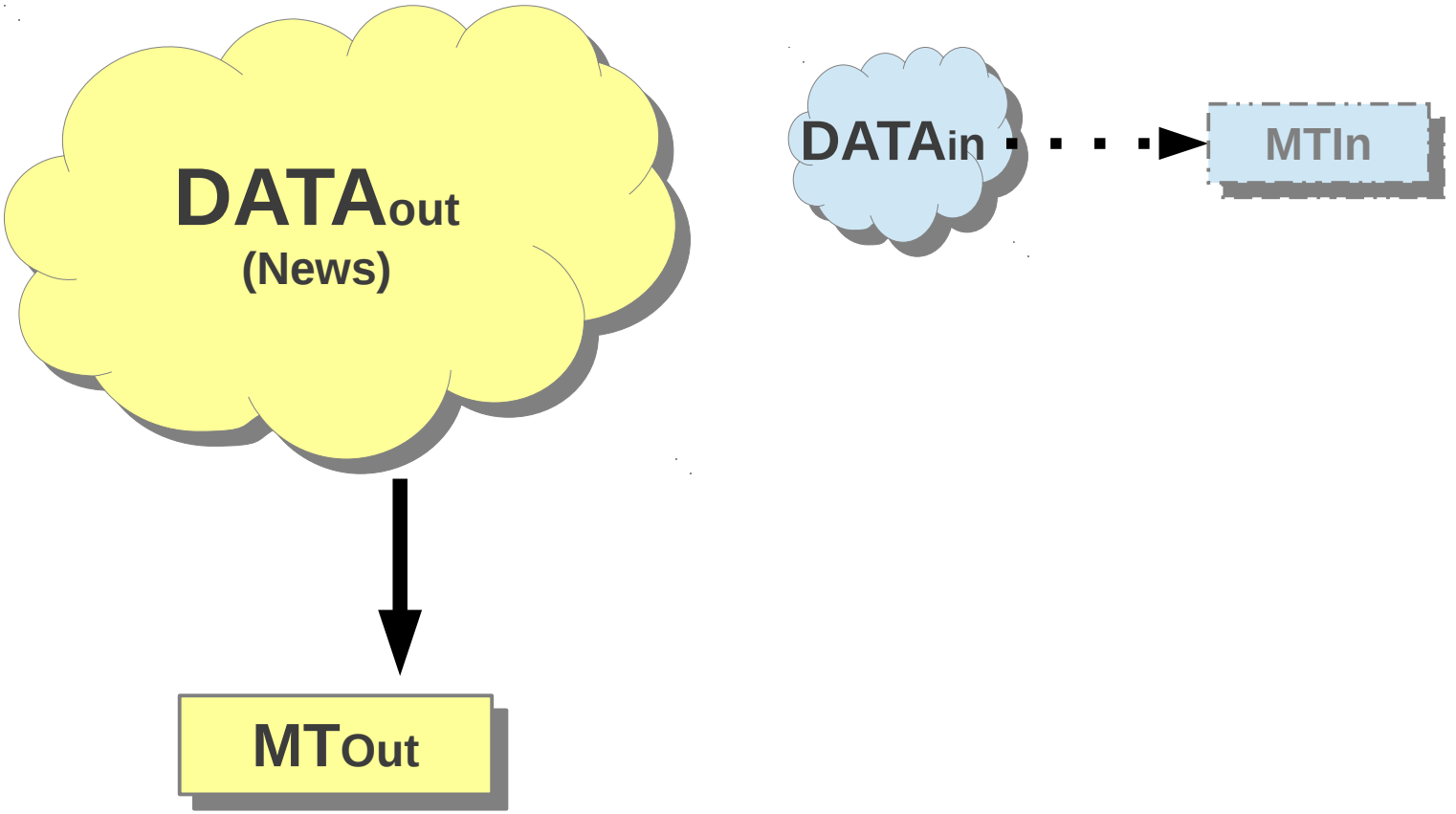
Versatile Build an MT System

- For every domain of language use → sports; news; politics; financial; banking; automotive; drugs; food; appliance manuals; hardware/software manuals; scientific articles; ....
- Automatically and rapidly → **Minimal human intervention**
- On demand: user specs. → **User supplied example texts to be translated**

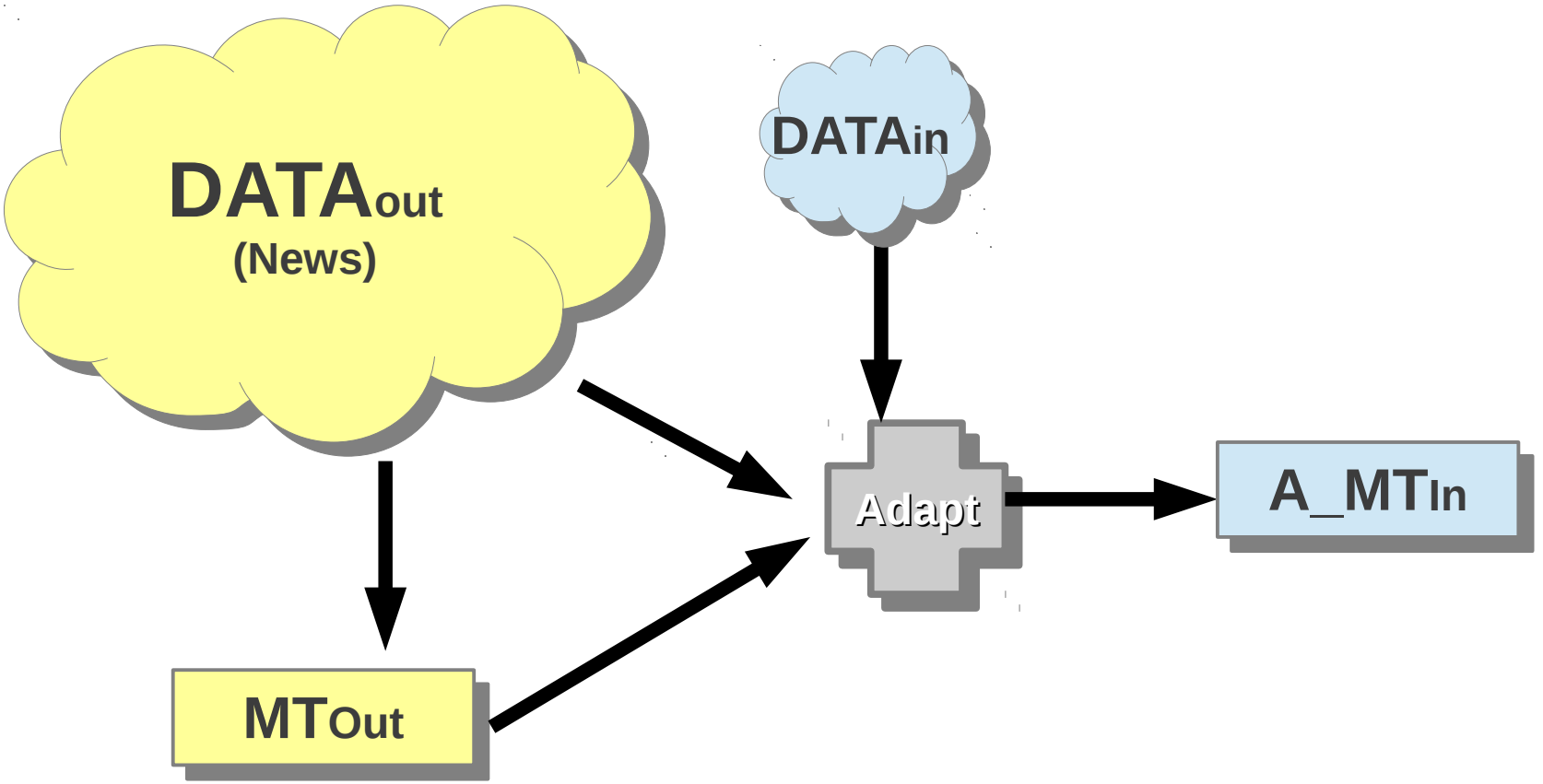
A population of MT Systems!

Current Practice
Tiny Data Adaptation

Current Practice
Tiny Data Adaptation



Current Practice
Tiny Data Adaptation



Tiny Data Adaptation
Current Practice

Task

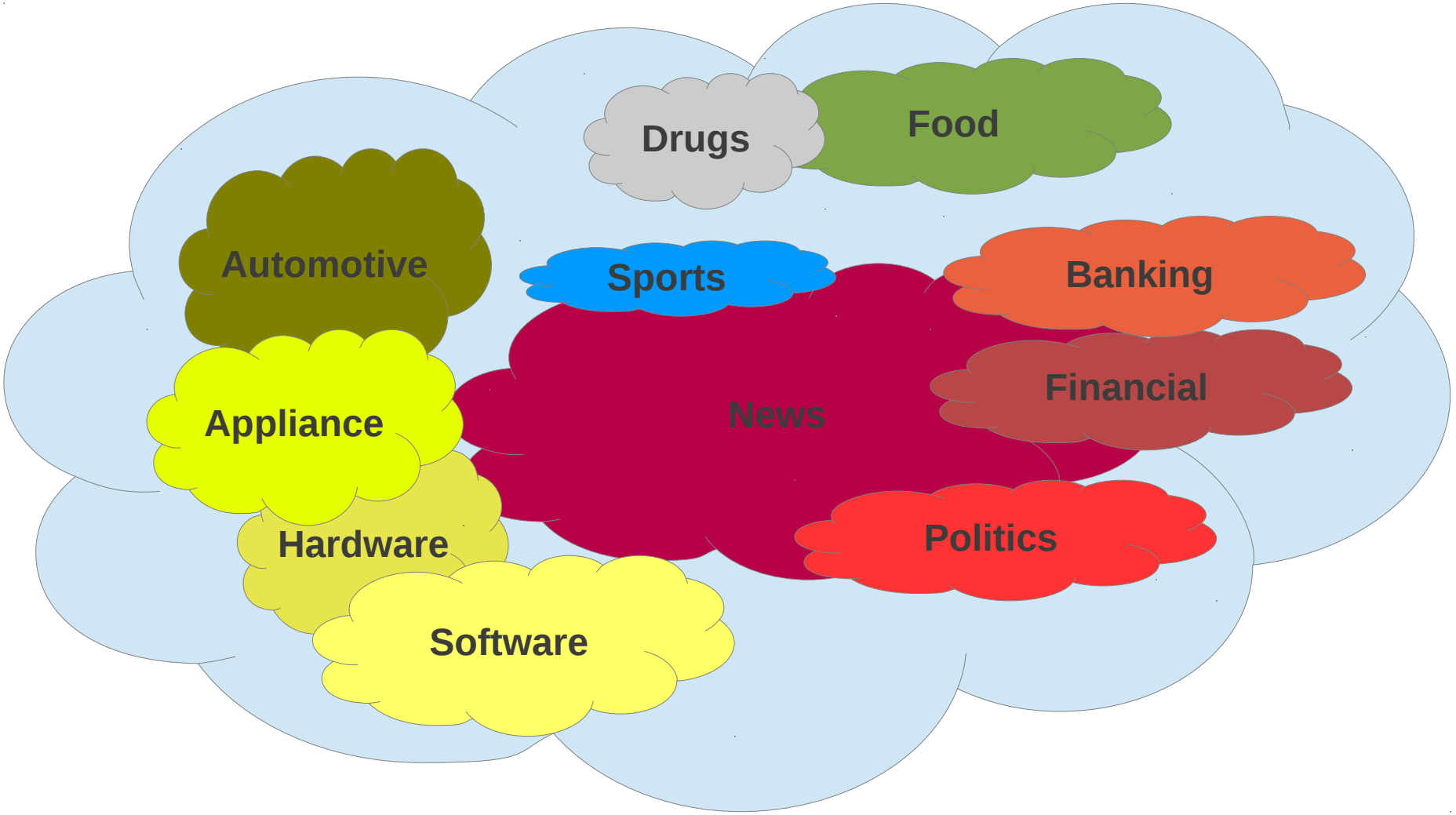
Build MT system from tiny in-domain data
using whatever out-of-domain data exists

Theoretically	Challenging, interesting, very difficult
Practice	Assumptions maybe too strong

Alternative Scenario

BIG DATA Adaptation

Big Data == Diversity



DatAptor Hypothesis

Big Data

Metaphore

Imagine a world of translators

- A translator: background + experience
- A translator for every situation

For every new translation order

Find the best suitable translator

DatAptor Hypothesis

Big Data == Diversity

Metaphore

Imagine a world of translators

- A translator for every situation
- A translator with own background and experience

For every new translation order

Find the best suitable translator

If (Big Data == ``A World of Domains")

Diversity enables rapid adaptation to new domain

``Find the most suitable MT system in the Data"

DatAptor Challenges I

INPUT: User documents from some domain + BIG DATA

OUTPUT: SMT system adapted for domain

- Distill from BIG DATA a suitable training data

  Weigh some documents as more relevant than others

- Train SMT system on distilled data.

Map of BIG DATA

- Efficiency for distilling suitable training data

  Map: the more related, the closer to each other on the map

- How to measure domain similarity?

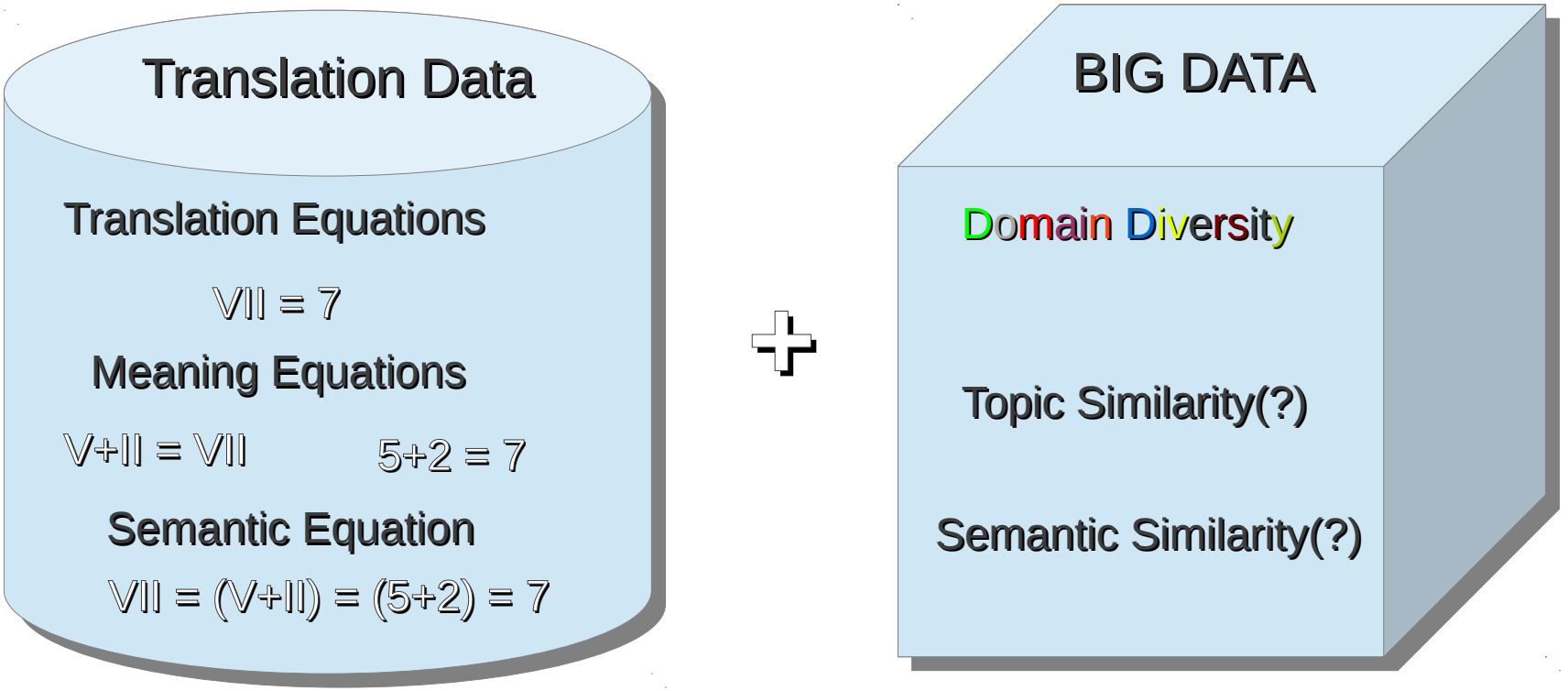
  Statistical (hierarchical-)topic similarity; translation-equivalence and instance weighting …

BIG DATA

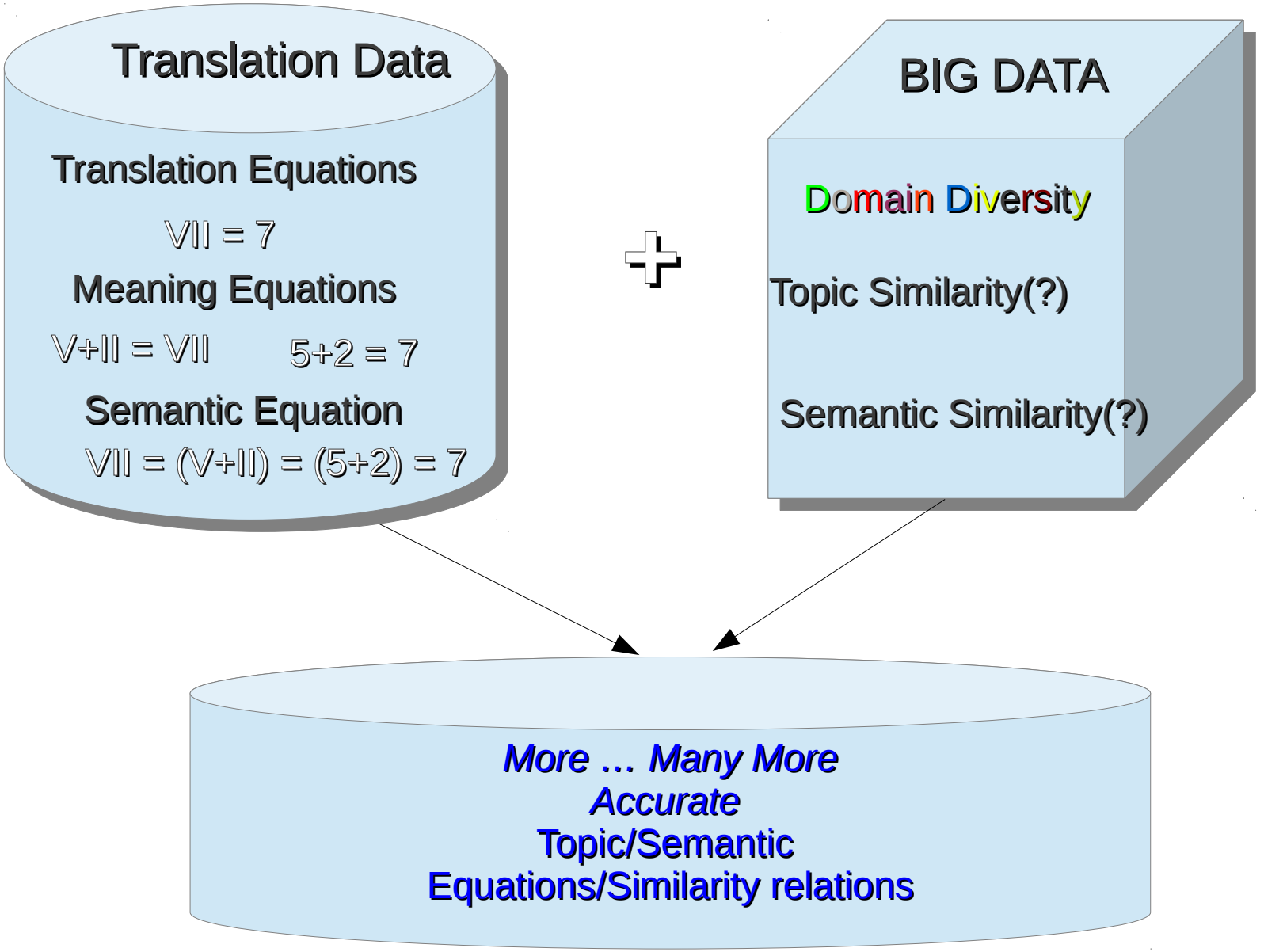
VS.

BIG Trans. DATA

BIG T-DATA


Translation Data
Translation Equations
VII = 7
Meaning Equations
V+II = VII
5+2 = 7
Semantic Equation
VII = (V+II) = (5+2) = 7
BIG DATA
Domain Diversity
Topic Similarity(?)
Semantic Similarity(?)

BIG T-DATA

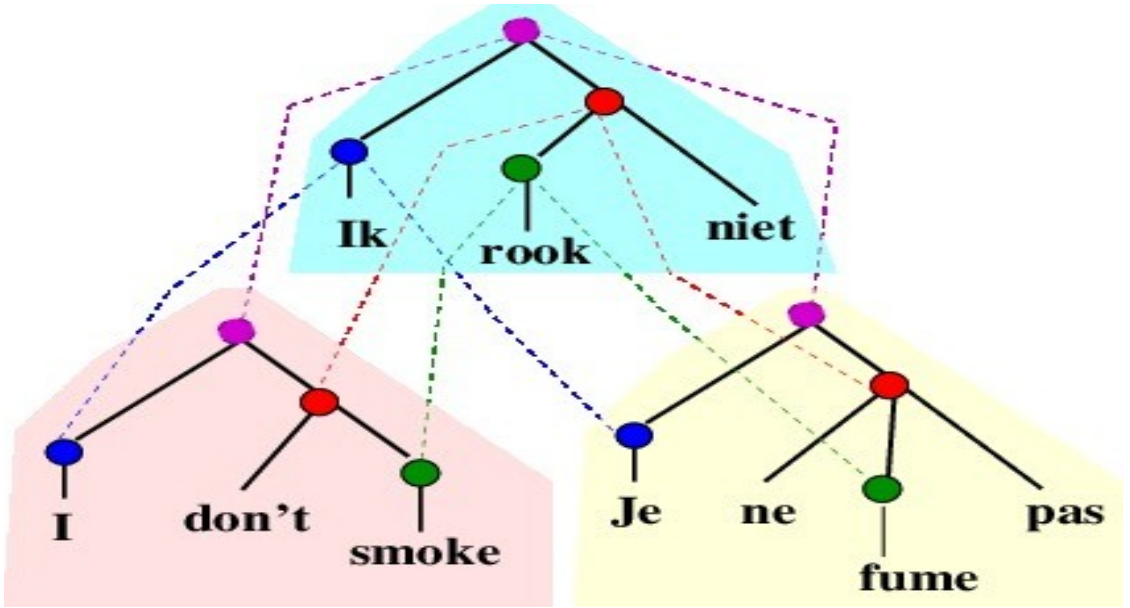

Translation Data
Translation Equations
VII = 7
Meaning Equations
V+II = VII
5+2 = 7
Semantic Equation
VII = (V+II) = (5+2) = 7
BIG DATA
Domain Diversity
Topic Similarity(?)
Semantic Similarity(?)
More … Many More
Accurate
Topic/Semantic
Equations/Similarity relations

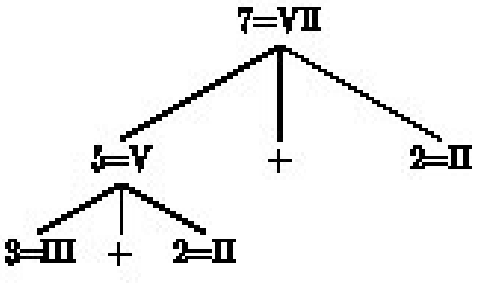
DatAptor Challenges II

- *User-Driven Data-Powered Adaptation*
- *Recursive, Hierarchical Meaning Equations*

  Structures translation equations; Better reordering

  Better fit with morpho-syntax; ``Deeper" meaning equations





To conclude

Big Trans. Data

- Enables Data-Powered Adaptation (DatAptation)
- Statistics over ``Meaning Equations”
- More than MT? Language understanding!

Thank you!

k.simaan@uva.nl

MT SUMMIT 2013

Machine Translation Summit XIV, Nice, France
2 - 6 September 2013