INFÉRENCE DE SCHÉMAS **PRÉCIS** POUR JEUX DE DONNÉES JSON VOLUMINEUX

- Encadrement : Mohamed-Amine Baazizi (LIP6)
  - Collaboration avec : Dario Colazzo (Paris Dauphine), Giorgio Ghelli (Université de Pise) et Carlo Sartiani (Université Basilicata)

CONTEXTE DU STAGE

- JSON : modèle des données semi-structurés
  - API (Twitter, Facebook, …)
  - Capteurs (smartphone, capteurs, …)
- Applications d'analyse de données manipulent JSON
  - sentiment analysis, monitoring, villes intelligentes
- **Nouveau Paradigme NoSQL : *data first, schema later***
  - Avantages : production rapide des données, évolution facilité
  - Inconvénients : interrogation, spécification de programmes

CONTEXTE DU STAGE

- Problématique récemment découverte, interêt croissant (WWW2016, PVLDB 2016, *CIDR 2017*)
- Une première approche proposée : BDA 2016 - *EDBT 2017*
  - inférence de schémas concis mais relativement peu précis
- Une extension en cours : inférence de schémas concis enrichi d'information statistiques

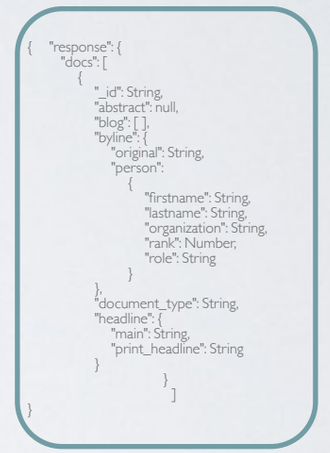
DATA MODEL

```
{
 "id": 1234,
  "abstract": "some text",
   "person":
     [
       {
         "firstname": "John",
          "lastname": "Smith",
          "phone": 1-1234-987
        }
      ]
}
```

- Basic values: null, bool, number, string
- Complex values:
  - Records: i.e set of attributes
  - Arrays: list of values

INFÉRENCE DE SCHÉMAS CONCIS MAIS RELATIVEMENT PEU PRÉCIS

Input dataset



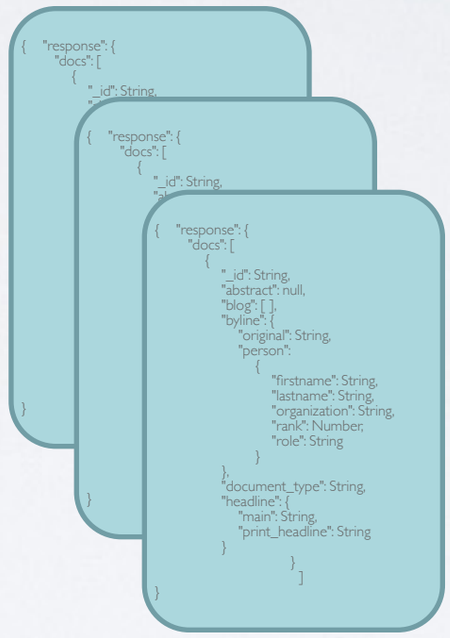
Global schema

INFÉRENCE DE SCHÉMAS CONCIS MAIS RELATIVEMENT PEU PRÉCIS

Input dataset

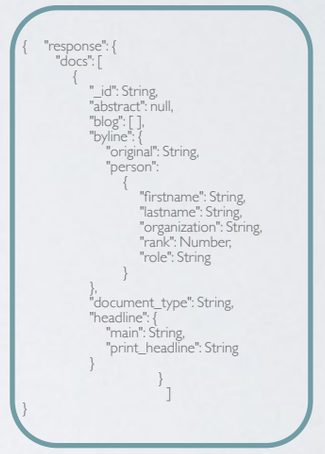
Map

Type inference



Reduce

Type fusion



Inferred types

Global schema

SCHEMA LANGUAGE

```
{
 "id": Number,
  "abstract": String,
  "person":
    [
      {
        "firstname": String,
         "lastname": String?,
         "phone": Number+String
       }
     ]
}
```

- Generalization of the data model
- Type union (+)
- Field optionality (?)

TYPE FUSION

```
{
 "id": Number,
  "abstract": String,
   "person":
    [
     {
       "firstname": String,
        "lastname": String,
        "phone": Number
      }
    ]
}
```

T

```
{
 "id": Number,
  "abstract": String,
"person":
    [
     {
       "firstname": String,

         "phone": String
      }
    ]
}
```

U

```
{
 "id": Number,
  "abstract": String,
   "person":
    [
     {
       "firstname": String,
        "lastname": String?,
        "phone": Number+String
      }
    ]
}
```

Merge(T,U)

RÉSULTATS EXPÉRIMENTAUX

	Description	Record	Array	Depth	Structural Variation
GitHub	metadata of pull requests	yes	no	4	null+string null+record
Twitter	tweet and delete	yes	yes	3	null+string null+number null+record
Wikidata	facts	yes	yes	6	*poor design*
NYTimes	news articles + noise	yes	yes	7	null+string record+array

TYPE FUSION : PERTE DE PRÉCISION

```
{
 "id": Number,
  "abstract": String,
   "person":
     [
       {
         "firstname": String,
          "lastname": String,
          "phone": Number
        }

     ]
}
```

T

```
{
 "id": Number,
  "abstract": String,
"person":
     [
       {
         "nom": String,
          "prenom": String,
          "mobile": Number
        }
     ]
}
```

U

```
{
 "id": Number,
  "abstract": String,
   "person":
     [
       {
         "firstname": String ?,
          "lastname": String?,
          "phone": Number?,
          "nom": String ?,
          "prenom": String?,
          "mobile": Number?,
        }
     ]
}
```

Merge(T,U)

TYPE FUSION : PERTE DE PRÉCISION

```
{
 "id": Number,
  "abstract": String,
   "person":
     [
       {
         "firstname": String,
          "lastname": String,
          "phone": Number
        }

     ]
}
```

T

```
{
 "id": Number,
  "abstract": String,
"person":
     [
       {
         "nom": String,
          "prenom": String,
          "mobile": Number
        }
     ]
}
```

U

```
{
 "id": Number,
  "abstract": String,
   "person":
     [
       {
         "firstname": String ?,
          "lastname": String?,
          "phone": Number?,
          "nom": String ?,
          "prenom": String?,
          "mobile": Number?
           }
      ]
  }
```

Merge(T,U)

```
{
 "id": Number,
  "abstract": String,
   "person":
     [
       {
         {“firstname": String ,
          "lastname": String,
          "phone": Number } +
          {“nom": String ?,
          "prenom": String?,
          "mobile": Number?}
            }
     ]
}
```

CONTEXTE DU STAGE

- Problématique récemment découverte, interêt croissant (WWW2016, PVLDB 2016, *CIDR 2017*)
- Une première approche proposée : BDA 2016 - *EDBT 2017*
  - inférence de schémas concis mais relativement peu précis
- Une extension en cours : inférence de schémas concis enrichi d'information statistiques
- **Prochaines étapes :**
  - **inférence de schémas concis et précis**
  - **caractérisation des jeux de données du web**

DÉROULEMENT DU STAGE

- Partie théorique
  - Formalisation et conception de nouvelle technique d'inférence
  - Preuves de correction et de complétude
- Partie pratique
  - Prise en main de l'implantation courante : code Scala, packaged pour Spark et **extension avec la nouvelle inférence**
  - Utilisation des datasets existants : Twitter, GitHub, NYTimes (~20GB), Wikidata (74GB par dump !). Crawl d'autres DS
  - **visualisation**

POUR PLUS D'INFORMATION :

mohamed-amine.baazizi@lip6.fr