MSE

# Classification of developpers based on git activity

Moret Jérôme
Curty Pierre-Alain
Delessert Armand

# Sommaire

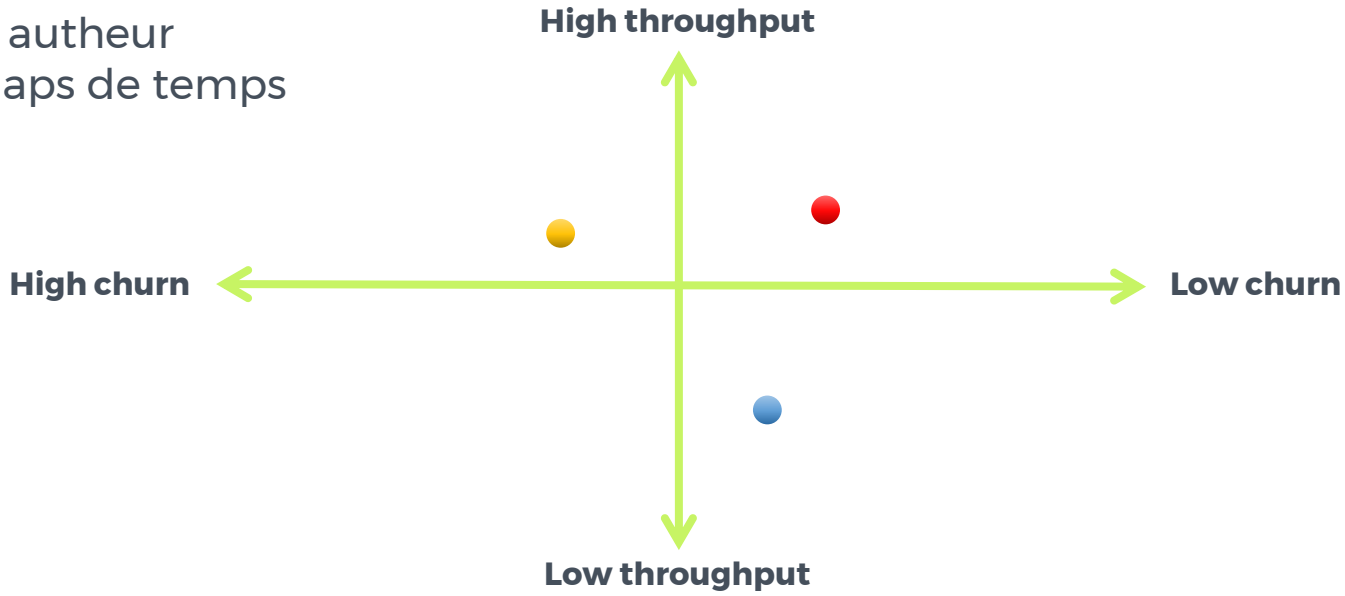
1. Objectif
2. Théorie
3. Pratique
4. Défauts & améliorations

# 1.

# Objectif

# Objectif

- Analyse des commits Git
- Classification développeur sur un espace à deux dimensions
  - **Throughput** quantité de code ajouté
  - **Churn** quantité de code réécrit
    - Même autheur
    - Court laps de temps



# 2.

# Théorie

# Théorie Extraction des évènements git

## git log --numstat

| Commit hash | Parent hash | Commit name | Author name | Commit date | Filename | Additions | Deletions | Churns |
|---|---|---|---|---|---|---|---|---|
| 1 | - | Write test.c | Jérôme Moret | 18:00 | Test.c | 14 | 0 | **0** |
| 2 | 1 | Modify test.c | Jérôme Moret | 19:00 | Test.c | 3 | **3** | **?** |

churn ssi
- Même autheur
- Dernière modif remonte à -3 semaines

# Théorie Récupération des lignes affectées

| Commit hash | Parent hash | Commit name | Author name | Commit date | Filename | Additions | Deletions | Churns |
|---|---|---|---|---|---|---|---|---|
| 1 | - | Write test.c | Jérôme Moret | 18:00 | Test.c | 14 | 0 | **0** |
| 2 | 1 | Modify test.c | Jérôme Moret | 19:00 | Test.c | 3 | **3** | **?** |

**git diff --unified=0**

**1:test.c 2:test.c**

**[2,3,8]**

# Théorie Vérification des conditions de churn

| Commit hash | Parent hash | Commit name | Author name | Commit date | Filename | Additions | Deletions | Churns |
|---|---|---|---|---|---|---|---|---|
| 1 | - | Write test.c | Jérôme Moret | 18:00 | Test.c | 14 | 0 | **0** |
| 2 | 1 | Modify test.c | Jérôme Moret | 19:00 | Test.c | 3 | **3** | **3** |

**[2,3,8]**

**git blame –L2,+1**

**1 -- test.c**

**Jérôme Moret** ✓

**18:00** ✓

# Théorie Aggrégation

| Commit hash | Parent hash | Commit name | Author name | Commit date | Filename | Additions | Deletions | Churns |
|---|---|---|---|---|---|---|---|---|
| 1 | - | Write test.c | Jérôme Moret | 18:00 | Test.c | 14 | 0 | **0** |
| 2 | 1 | Modify test.c | Jérôme Moret | 19:00 | Test.c | 3 | 3 | **3** |
| ... | ... | ... | Delessert Armand | ... | ... | 5 | 0 | **0** |
| ... | ... | ... | Delessert Armand | ... | ... | 5 | 3 | **3** |
| ... | ... | ... | Curty Pierre-Alain | ... | ... | 10 | 0 | **0** |
| ... | ... | ... | Curty Pierre-Alain | ... | ... | 32 | 14 | **12** |

| Author name | Additions | Churn | Churn rate |
|---|---|---|---|
| Jérôme Moret | 17 | 3 | 0.18 |
| Delessert Armand | 10 | 3 | 0.3 |
| Curty Pierre-Alain | 42 | 12 | 0.28 |

**Throughput**

**Churn**

# 3.

# Pratique

Méthodologie
**Démonstration**
Variantes

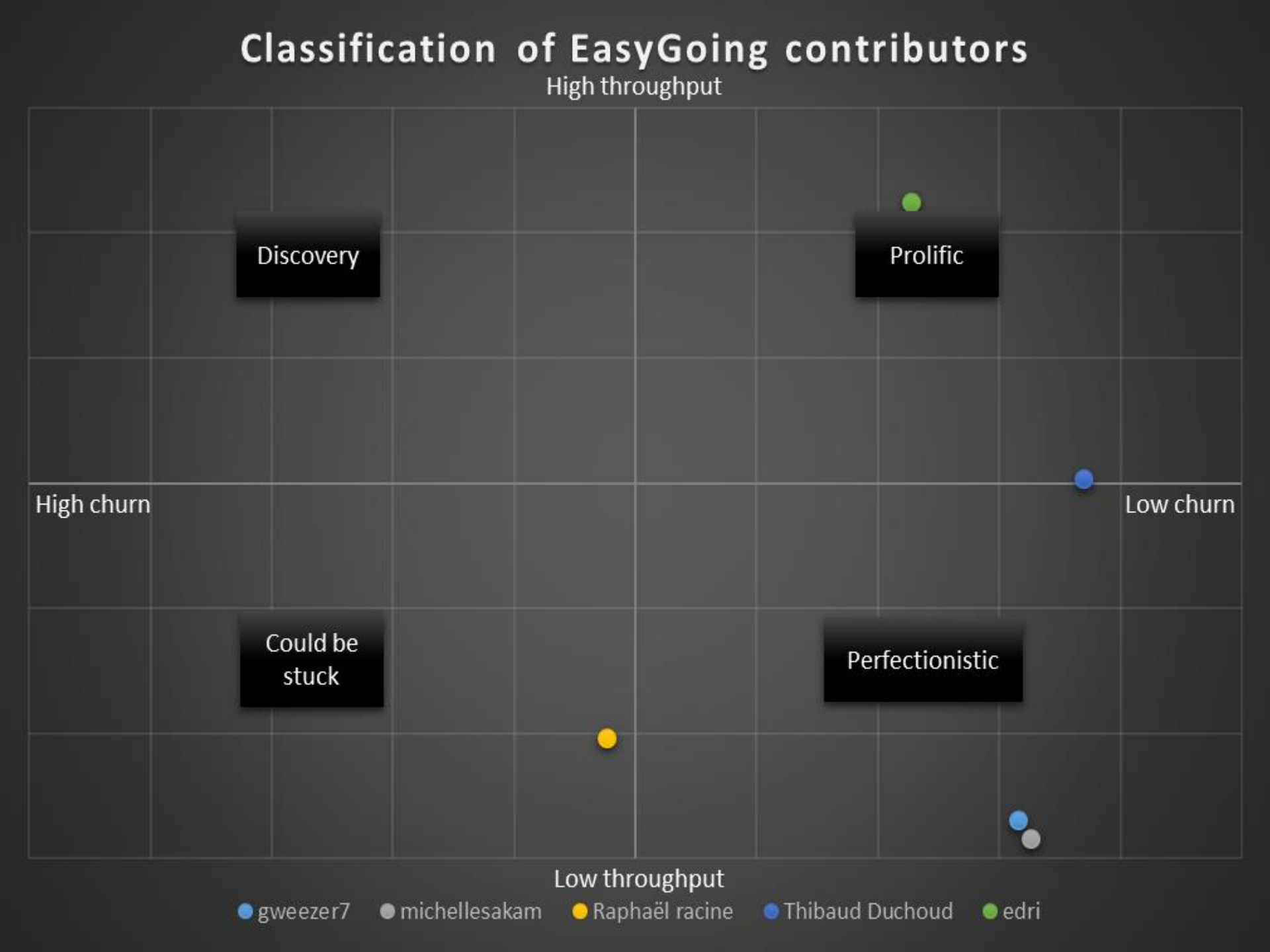
# **Pratique** Méthodologie

- Scripting Python
  - Librairie **GitPython**
  - Dataset.py
    - **Entrée :** chemin vers le dépôt
    - **Sortie :** dataset.csv
  - Aggregation.py
    - **Entrée :** dataset.csv
    - **Sortie :** result.csv

- Exploitation des résultats dans un tableur

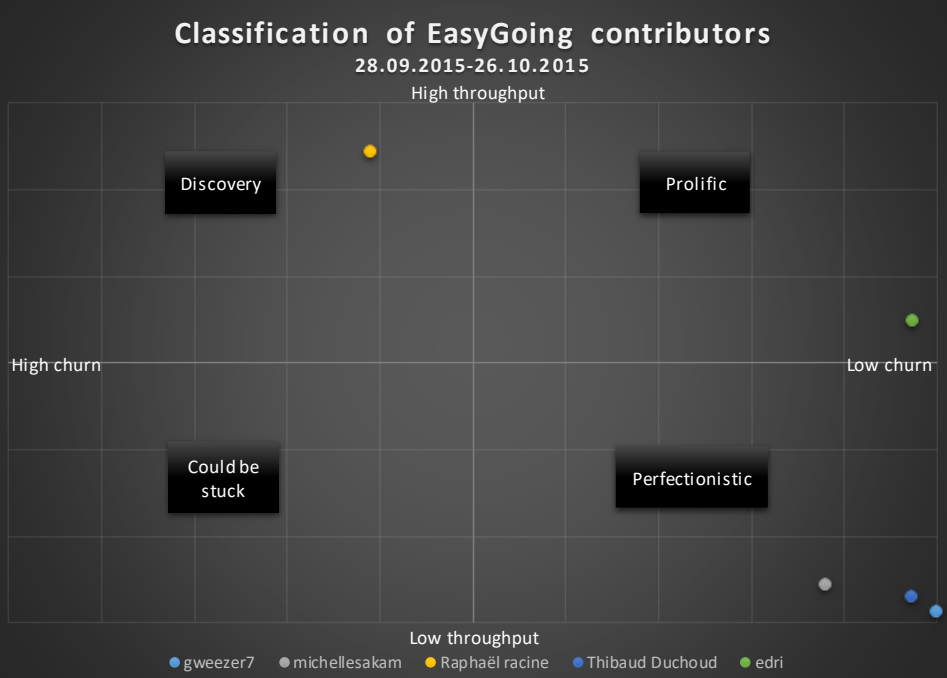
# **Pratique** Démonstration

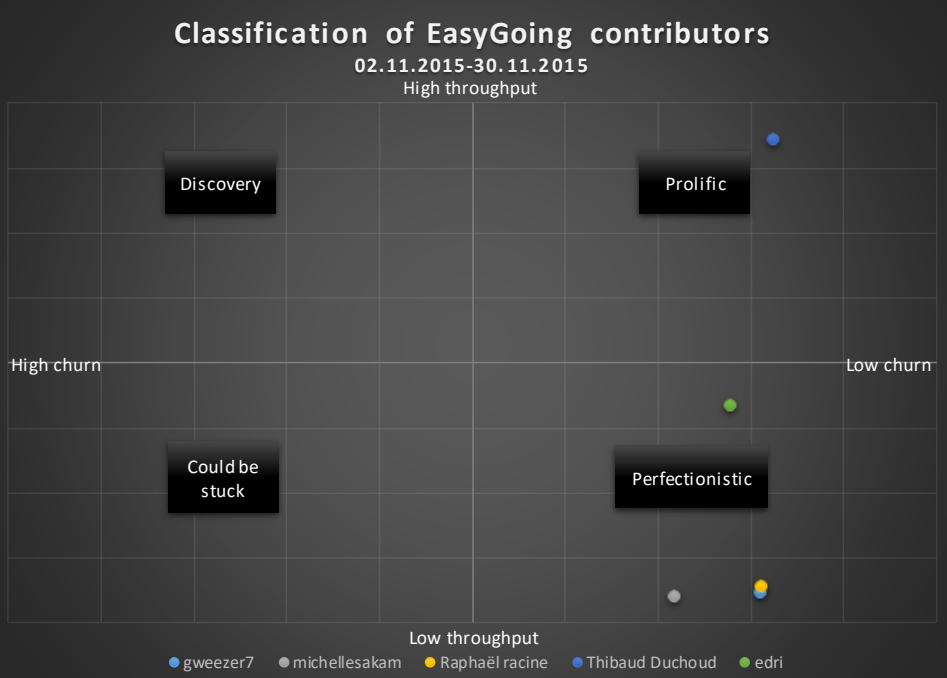
- Dépôt GIT d'un projet bachelor
  - **Nom du projet :** EasyGoing
  - 5 contributeurs
  - 775 commits

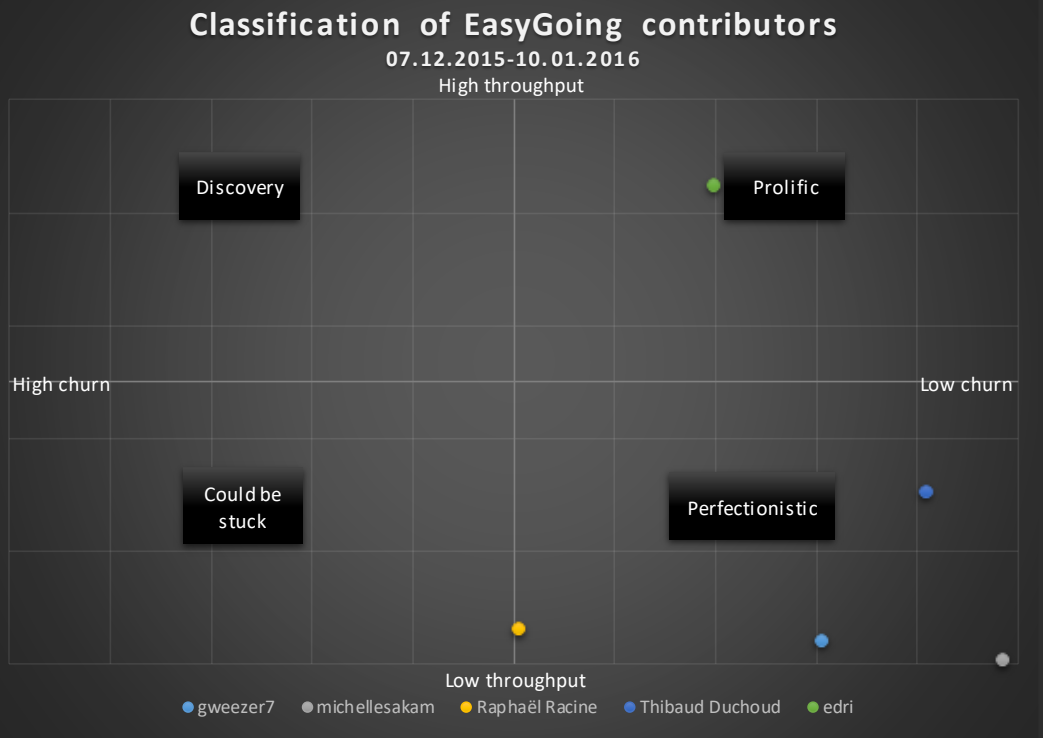

Classification of EasyGoing contributors
High throughput
Discovery
Prolific
High churn
Low churn
Could be stuck
Perfectionistic
Low throughput
gweezer7
michellesakam
Raphaël racine
Thibaud Duchoud
edri

# **Pratique** Variantes

- Utilisation incrémentale
  - Dataset.py
    –s, --since <since_date>
- Monitoring d’une équipe agile


Classification of EasyGoing contributors
28.09.2015-26.10.2015
High throughput
Discovery
Prolific
High churn
Low churn
Could be stuck
Perfectionistic
Low throughput
gweezer7
michellesakam
Raphaël racine
Thibaud Duchoud
edri


Classification of EasyGoing contributors
02.11.2015-30.11.2015
High throughput
Discovery
Prolific
High churn
Low churn
Could be stuck
Perfectionistic
Low throughput
gweezer7
michellesakam
Raphaël racine
Thibaud Duchoud
edri


Classification of EasyGoing contributors
07.12.2015-10.01.2016
High throughput
Discovery
Prolific
High churn
Low churn
Could be stuck
Perfectionistic
Low throughput
gweezer7
michellesakam
Raphaël Racine
Thibaud Duchoud
edri

# 4.

# Défauts & améliorations

# Défauts & améliorations

- Complexité
  - Temps d'exécution de dataset.py sur **EasyGoing** : ~15min

- Solutions
  - Réduire la complexité
    - Procédure
    - Algorithmes et structures de données
  - Parallélisation
  - Limite d'évènements

# Défauts & améliorations

- Développer l’approche incrémentale
- Concevoir une solution **tout-en-un**

# Merci !

## Des questions ?

Moret Jérôme
Curty Pierre-Alain
Delessert Armand