Решении DM/ML задач

Задача

Нам дали данные.

Нам поставили задачу.

Мы знаем какой должен быть результат.

Задача

Нам дали данные.

Нам поставили задачу.

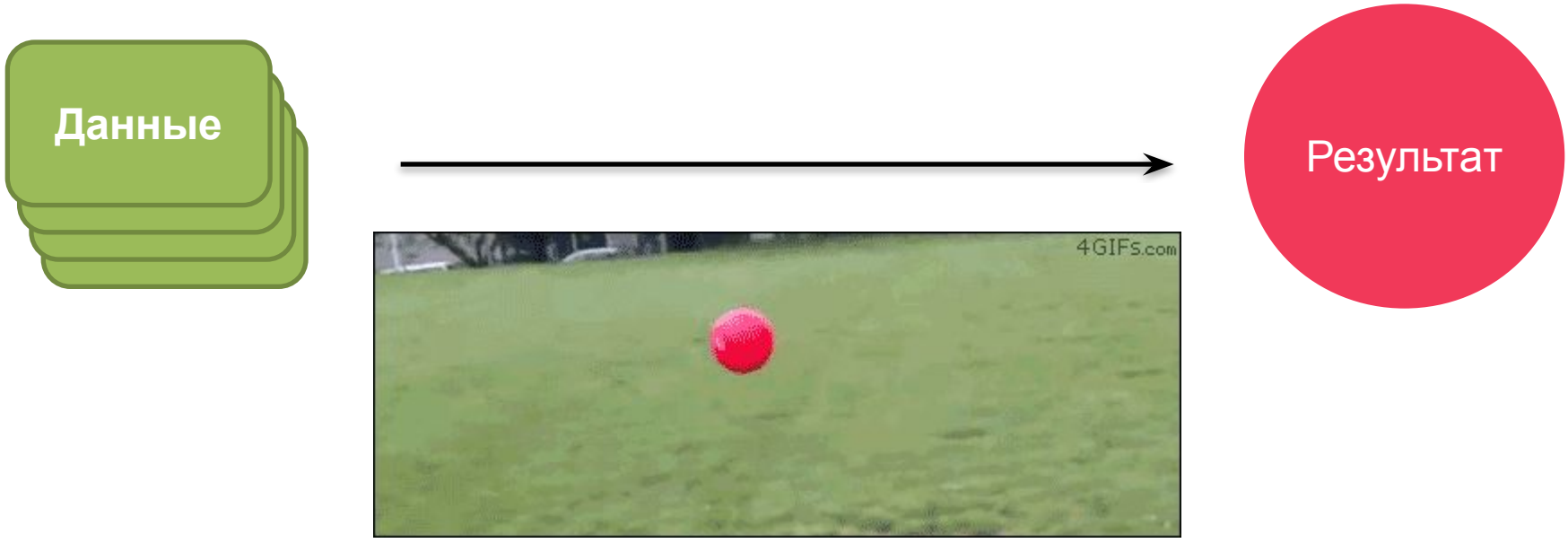
Мы знаем какой должен быть результат.

Задача

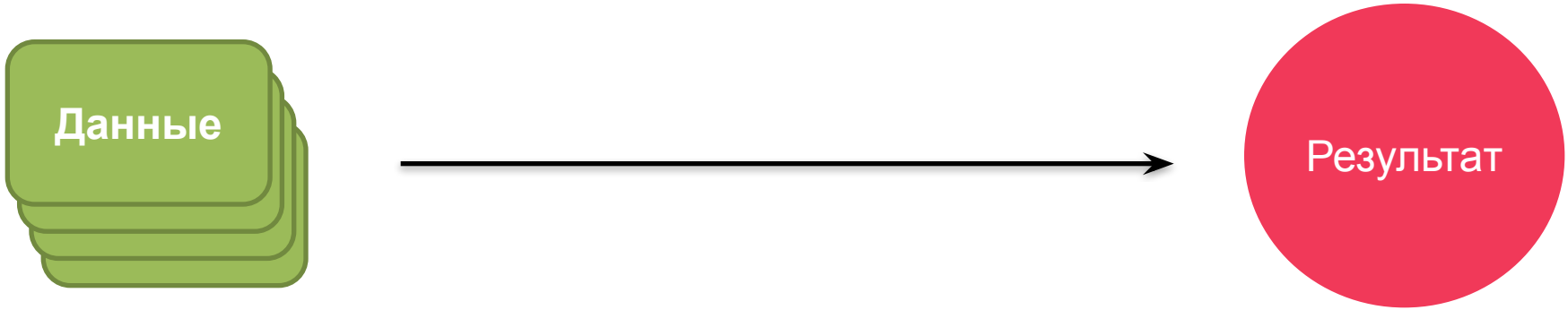
Нам дали данные.

Нам поставили задачу.

Мы знаем какой должен быть результат.

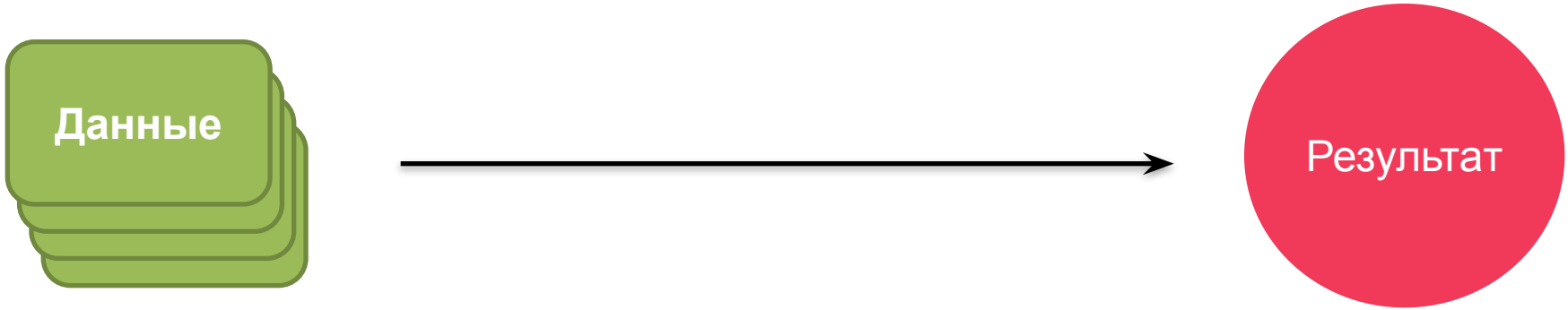


Что пошло не так?



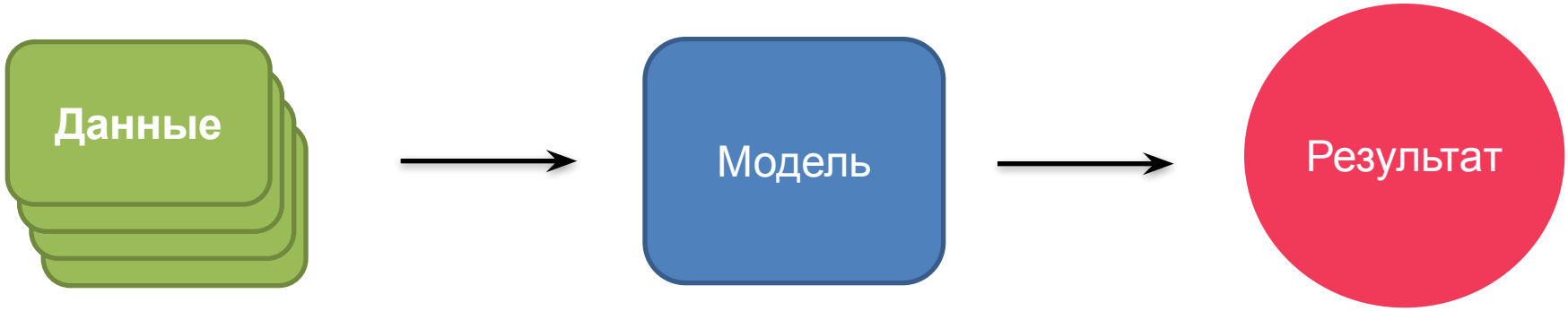
Мы что-то где-то упустили.

Что пошло не так?



Мы поставили гипотезу: применим RF влоб, вдруг сработает.

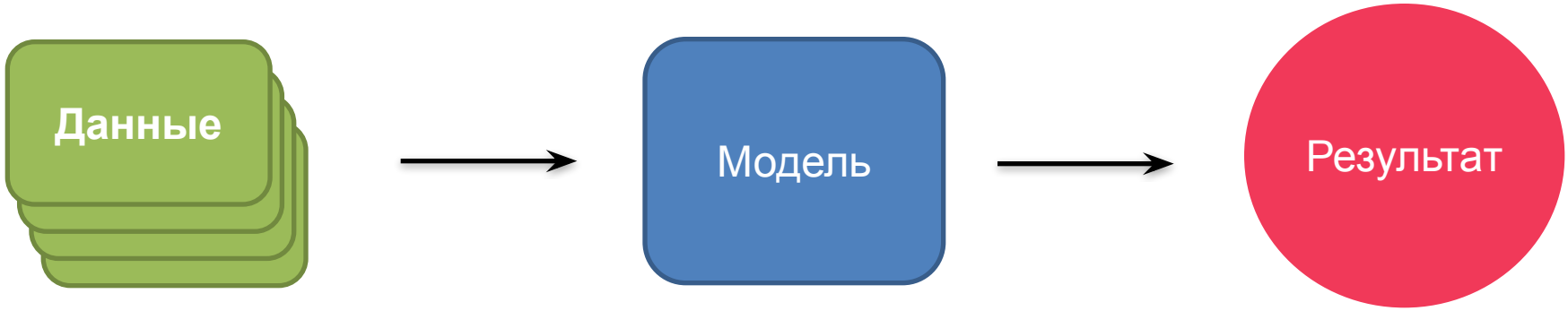
Модели



Что если мы ошиблись с RF/SVM?

Тоесть, все заработало, но результат был плох?

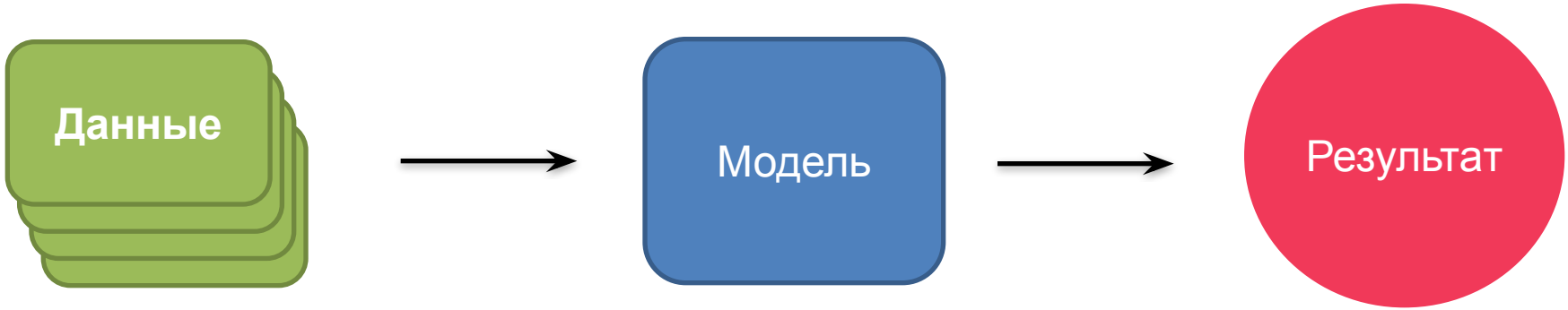
Модели. Обучающие множества



Что если мы ошиблись с RF/SVM?

Мы же не накосячили с тренировочным, валидационным и тестовым множествами, а также CV?

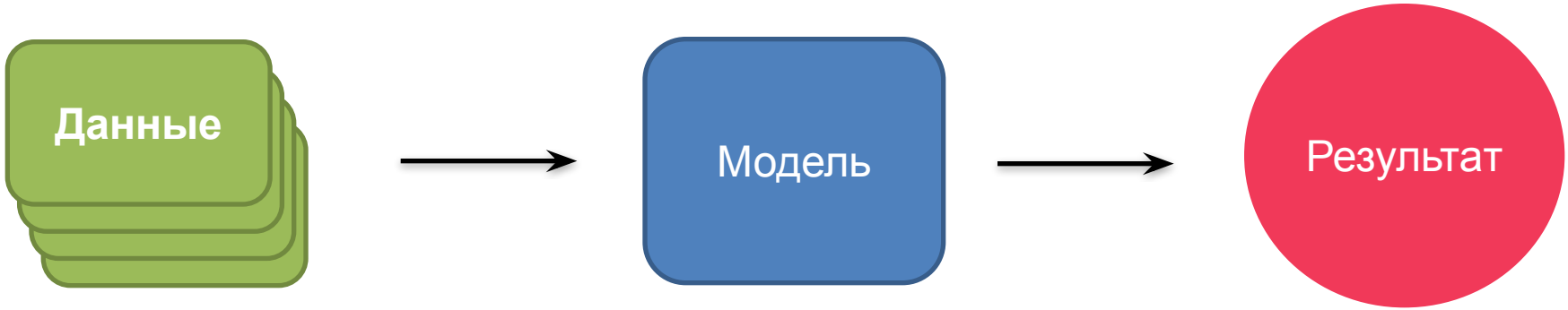
Что мы хотим от моделей?



Есть:

- Обученная модель
- Результат ее работы на валидационном(тестовом) множестве (ошибка)

Что мы хотим от моделей?



Есть:

- Обученная модель
- Результат ее работы на валидационном(тестовом) множестве (ошибка)

Хотим:

- Улучшить (обобщающую) точность

Что мы можем сделать с моделями?

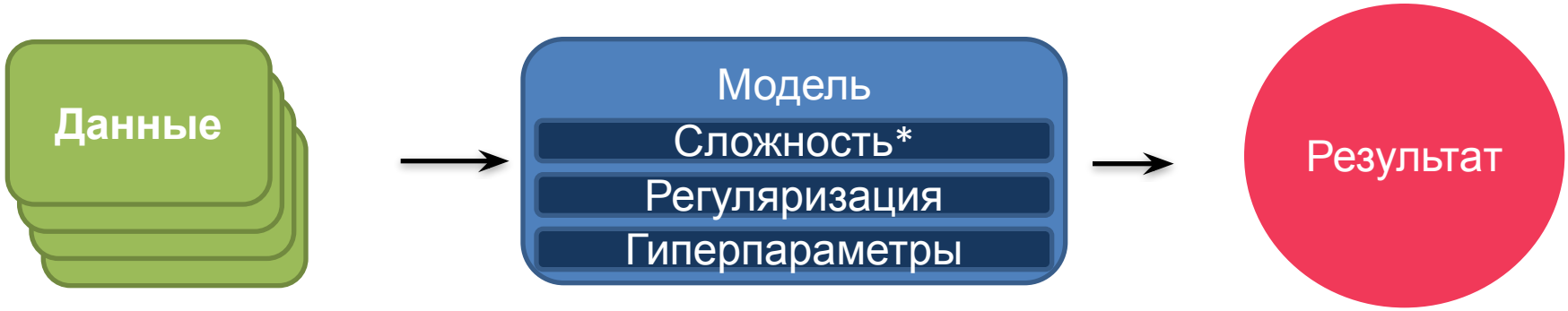
Хотим:

- Улучшить (обобщающую) точность

Чем можем управлять:

- Гиперпараметры
  - Лучший их выбор (CV, boot)
  - Регуляризация
- Отбор признаков на уровне модели

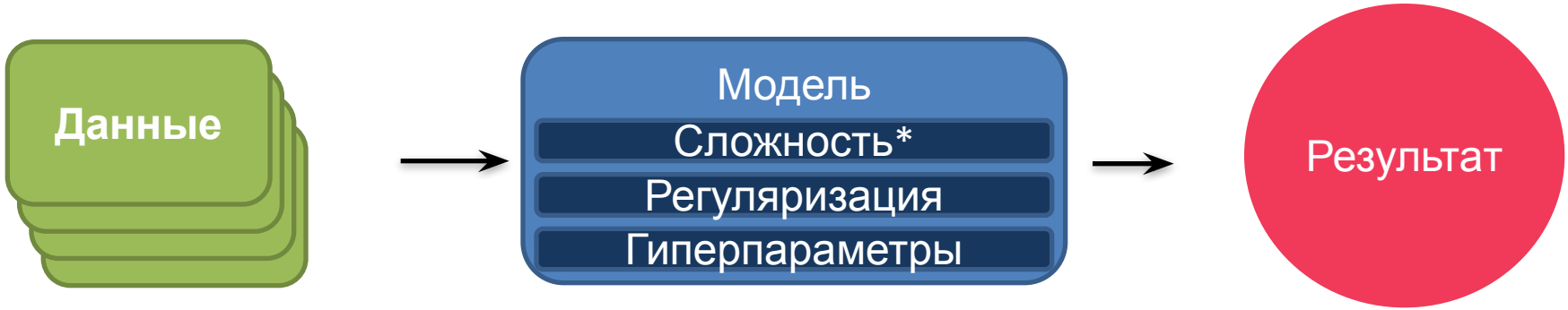
Что мы можем сделать с моделями?



Выбор гиперпараметров тратит много машинного времени.

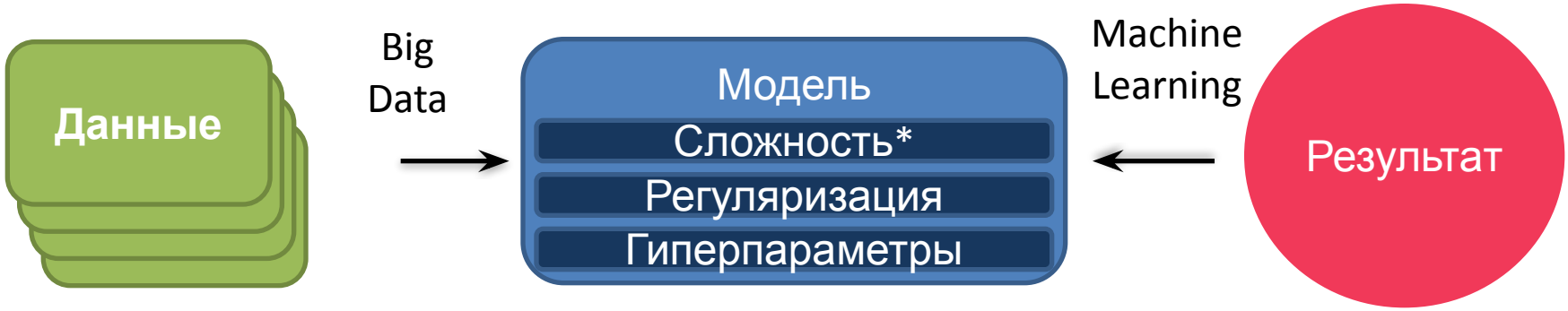
- **GBM**: #base-learners, lambda, RI, *prune, Loss,
- **SVM**: kernel, width, cost, nu-SVM, ...
- **GLMnet:** a.ridge, AIC, response family
- **RF:** … ?
- **Neural Net:** … ?

Что еще могло пойти не так?



Где мы еще могли накосячить?

Что еще могло пойти не так?



Где мы еще могли накосячить?

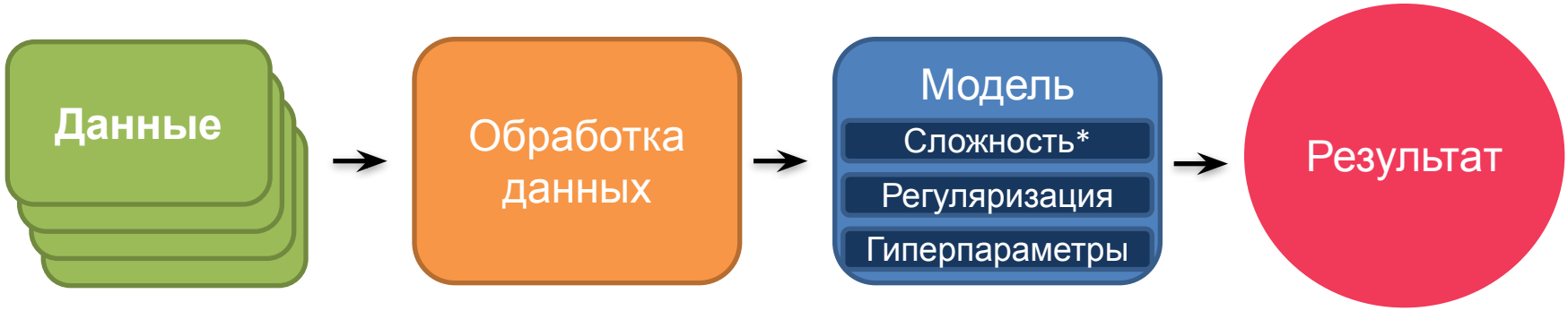
Что еще могло пойти не так?

DM Labs

Оно вообще не посчиталось.

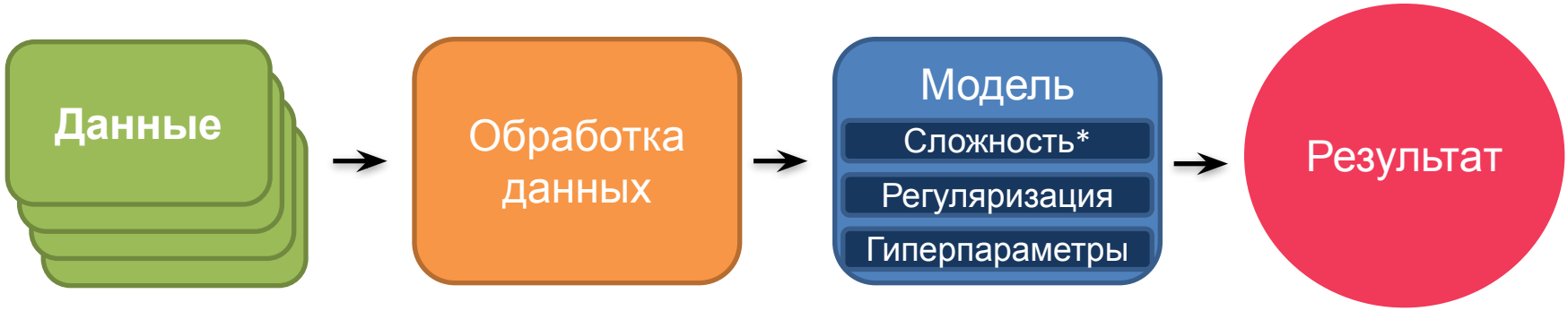
Иногда – феерично.

Обработка данных



Тесно связана с моделью.

Обработка данных

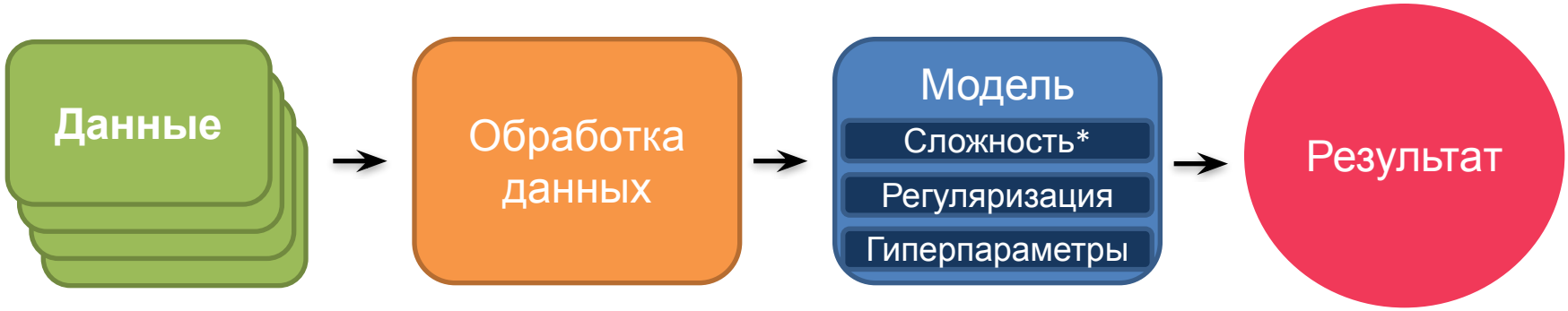


Хотим:

- Заставить что-то работать ☹
- Уменьшить количество переменных\данных
- Увеличить скорость вычисления\обучения (!=)

Возможности обработки данных

DM Labs



Хотим:

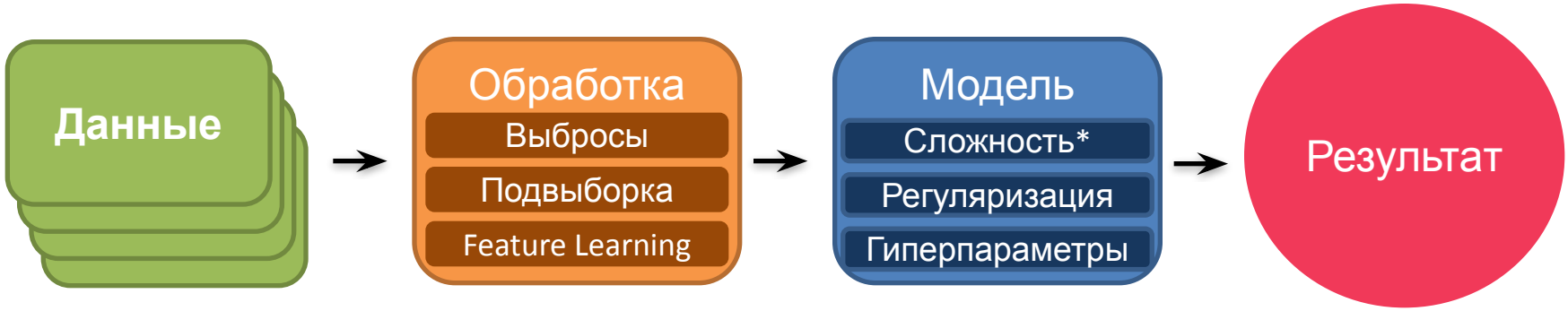
- Заставить что-то работать ☹
- Уменьшить количество переменных\данных
- Увеличить скорость вычисления\обучения (!=)

Чем можем управлять:

- Убрать выбросы
- Сделать подвыборку

- На уровне переменных:

Обработка данных. Feature Learning



В подвыборки можно вкладывать смысл:

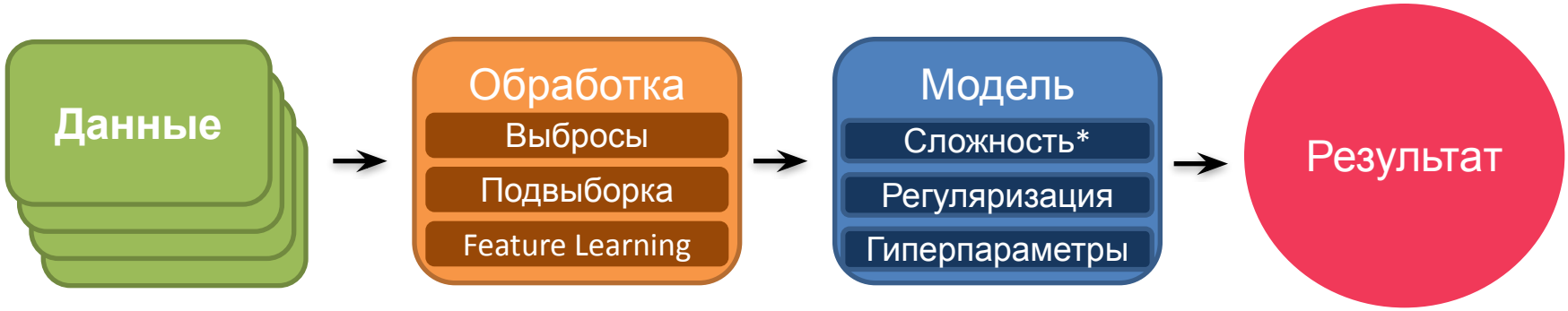
5% юзеров, записей, уникальных юзеров, последних записей...

Feature learning – основа deep learning.

Unsupervised, обычно отдельно:

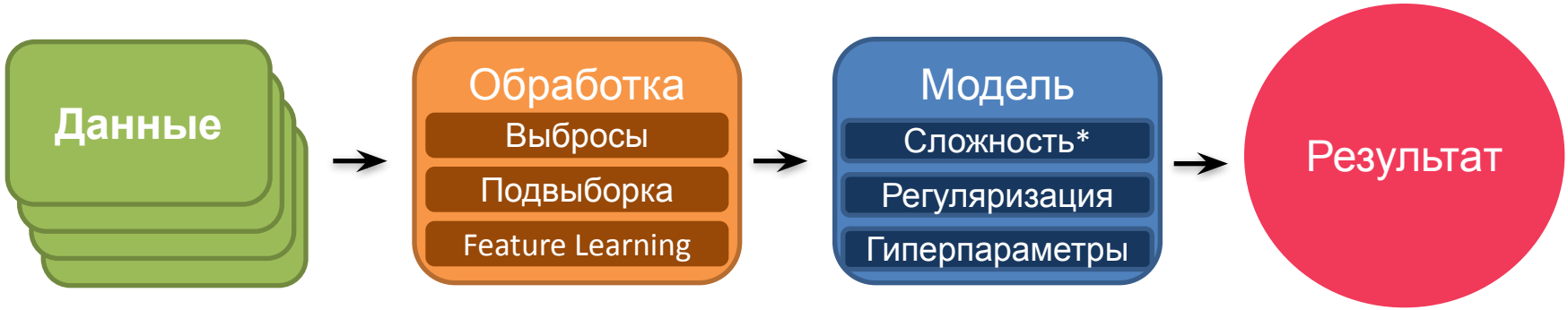
http://web.eecs.umich.edu/~honglak/nipsdlufl10-AnalysisSingleLayerUnsupervisedFeatureLearning.pdf

Что еще могло пойти не так?



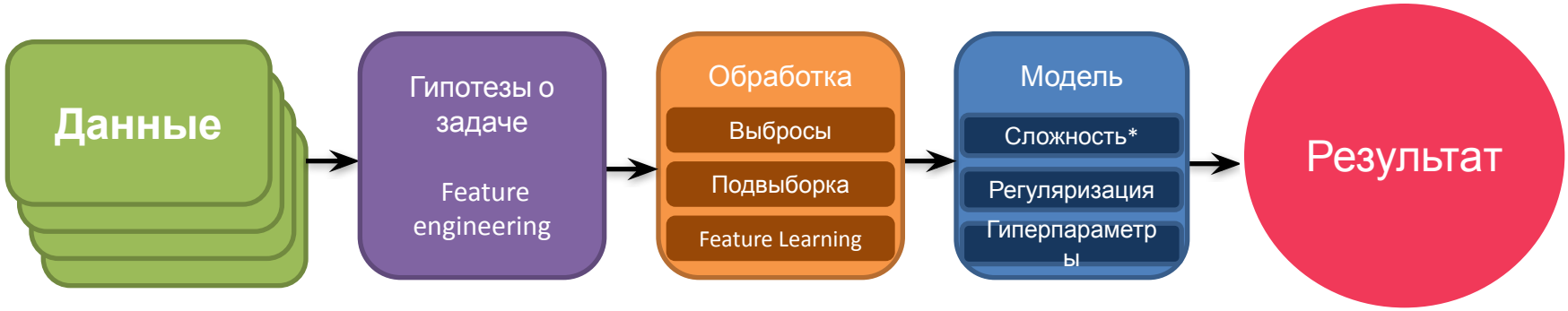
Что-то еще?

Что еще могло пойти не так?



Что если все еще не взлетает.

«Спросить у эксперта»

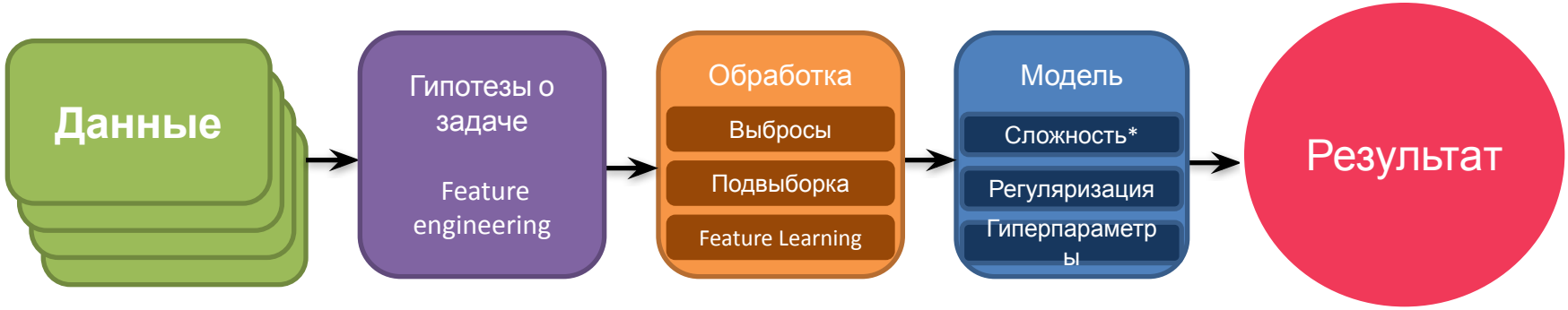


Самое сложное.

Понять что вообще происходит.

Неправильные фичи могут ни к чему не привести.

Гипотезы о задаче



Есть:

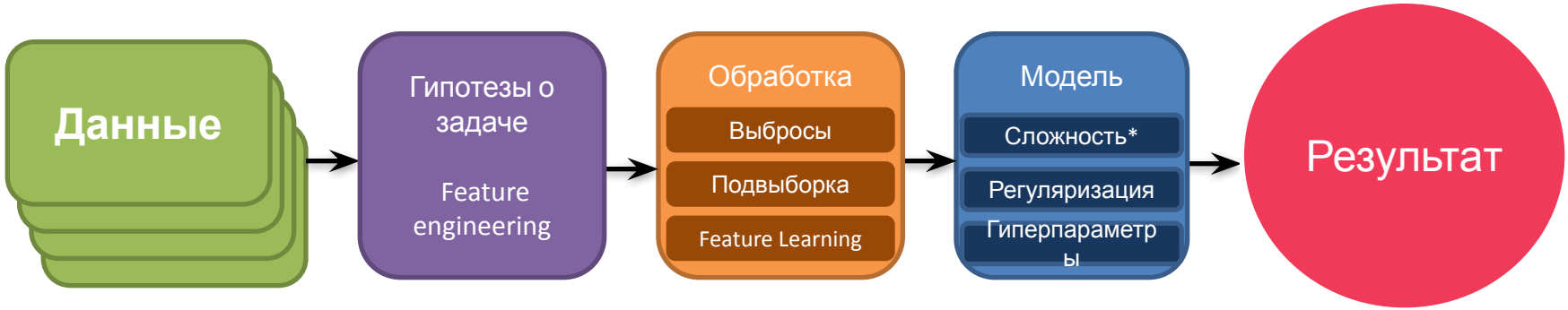
- Мы перепробовали кучу моделей и кучу гипотез.

Чем можем управлять:

- Достать еще данных
- Feature engineering

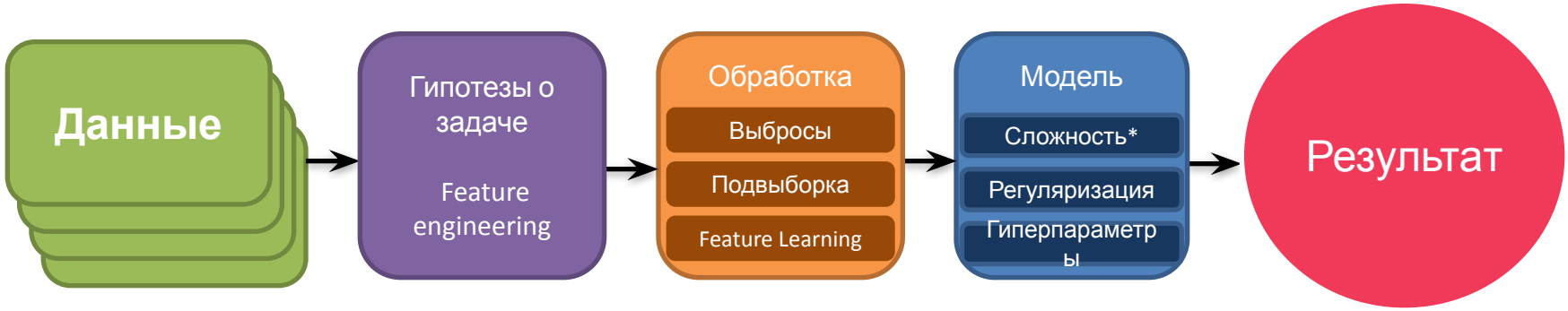
- Посмотреть что делали другие:

Гипотезы о задаче



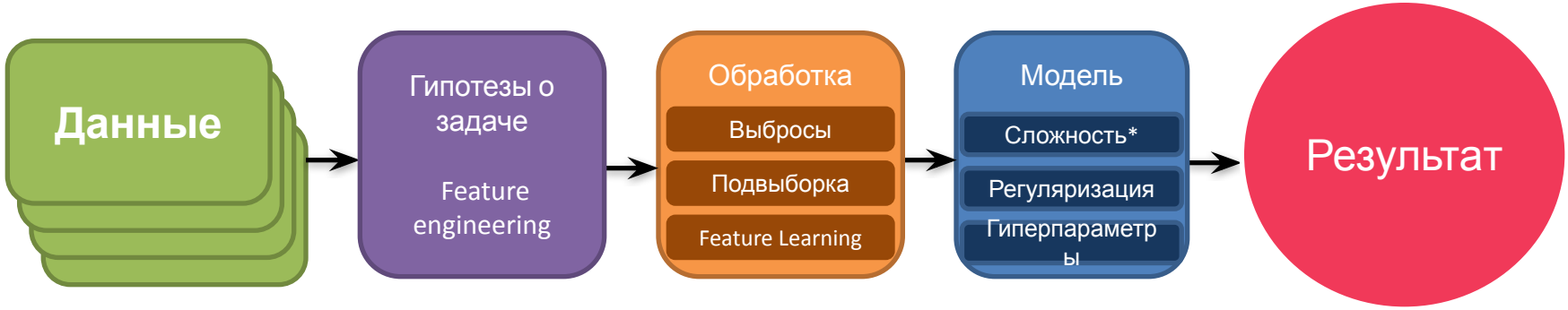
Пример:

- Изображения: сверточные признаки, Haar-признаки
- Временные ряды: fft разложения, моменты с лагом, MA
- Пользователи: признаки из графа(betweenness, degree, centrality, page rank), гео-специфика
- ...



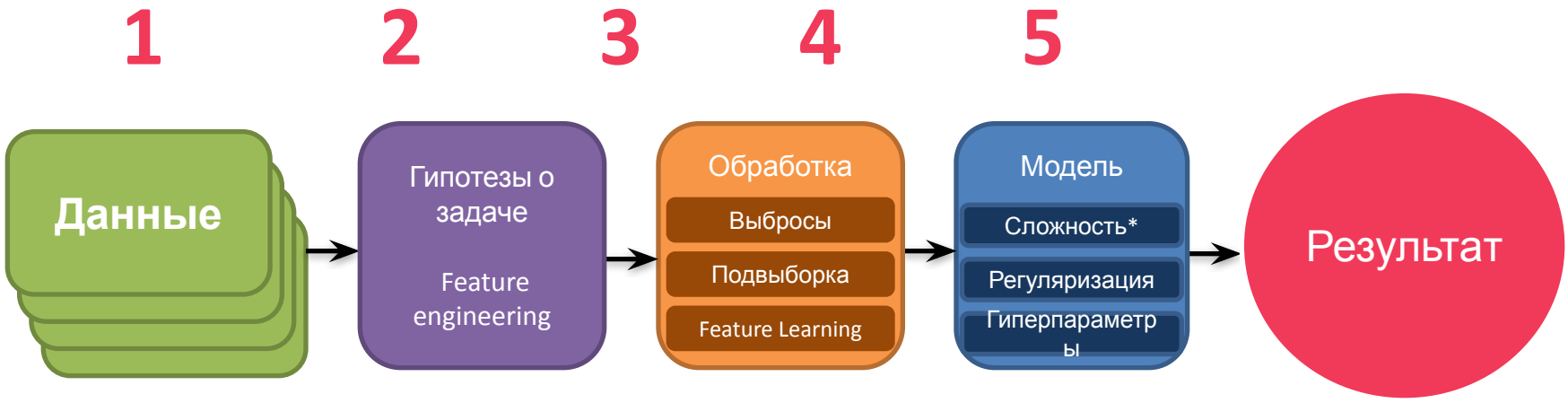
Как организовать команду?

Kaggle



Как организовать команду?

Kaggle



Техник: 1, 5

данные должны быть всегда доступны, сабмит всегда делался

Шпион: 2, 3, идеи про 4

изучает литературу, статьи, форумы. Генерирует идеи

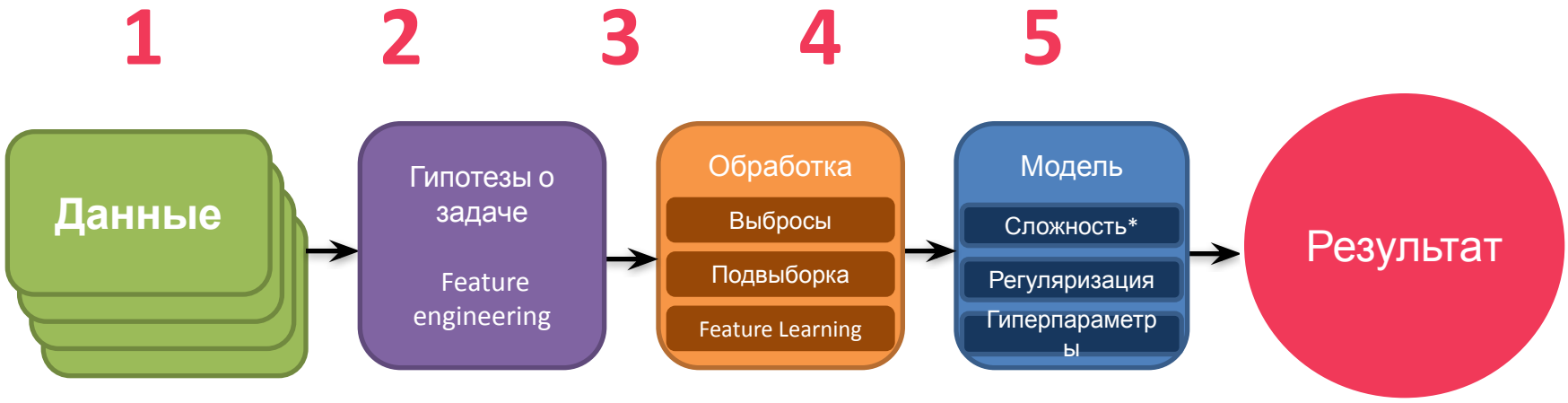
Спецназ: 3, 4

заняты основной работой, не парятся об 1, 2, 5 !!!

Капитан: 1, 2, 3, 4, 5

координирует работу всех участников, следит за всем сразу

Kaggle



Техник: 1, 5

…

Шпион: 2, 3, идеи про 4

…

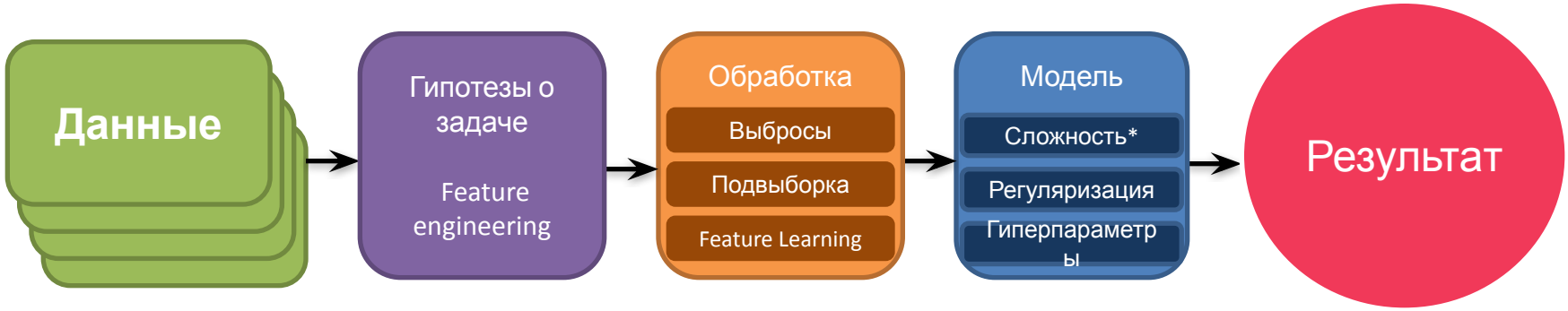
Спецназ: 3, 4

…

Капитан: 1, 2, 3, 4, 5

…

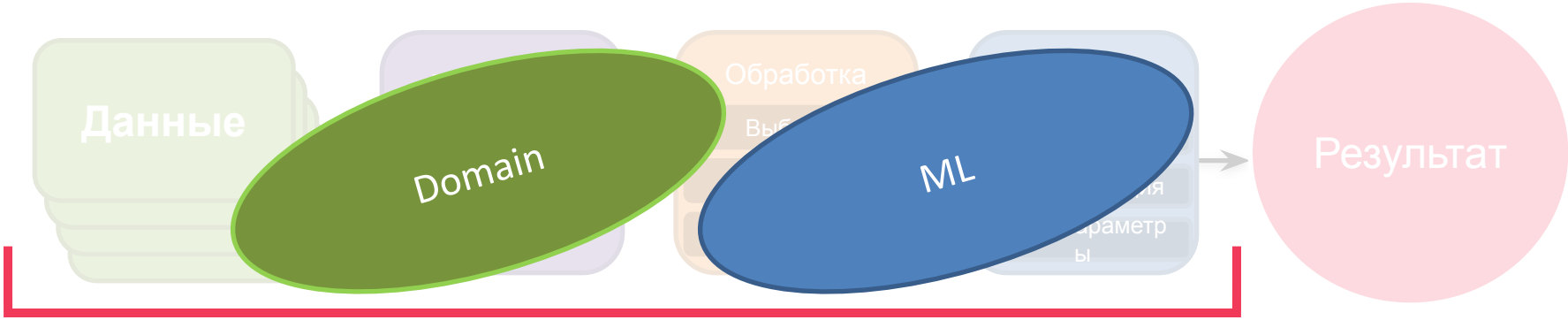
Как вообще решают задачи?



Где здесь результаты сообществ?

- Machine Learning
- Data Mining
- Специфичных областей (видео, биология, поиск, ...)

Как вообще решают задачи?



Data Miner’ами себя зовут вообще все. Даже те кто выложил данные...

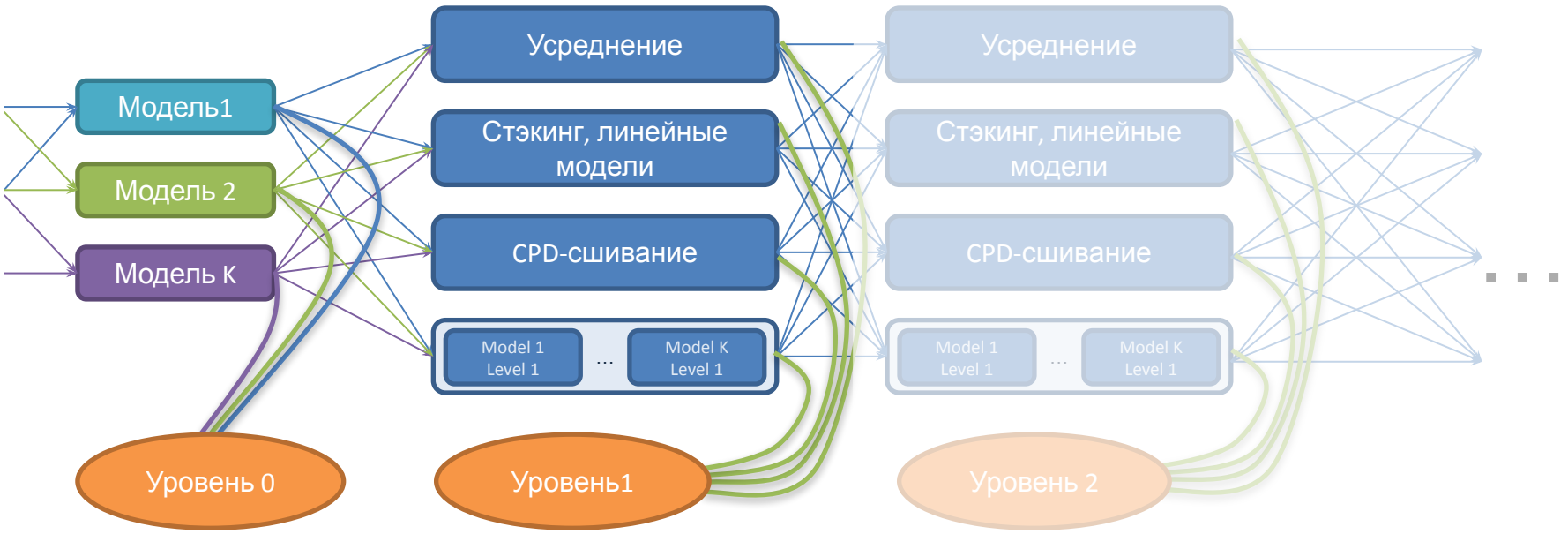
Где здесь результаты сообществ?

- Machine Learning
- Data Mining
- Специфичных областей (видео, биология, поиск, ...)

Summary

- ...

Прочие штуки

Стэкинг моделей:

Если вы вдруг сделали вообще все, можно похимичить