

Решении DM/ML задач

Нам дали данные.

Нам поставили задачу.

Мы знаем какой должен быть результат.



Нам дали данные.

Нам поставили задачу.

Мы знаем какой должен быть результат.



Давайте решим влоб!

RF/SVM, вдруг сработает

Нам дали данные.

Нам поставили задачу.

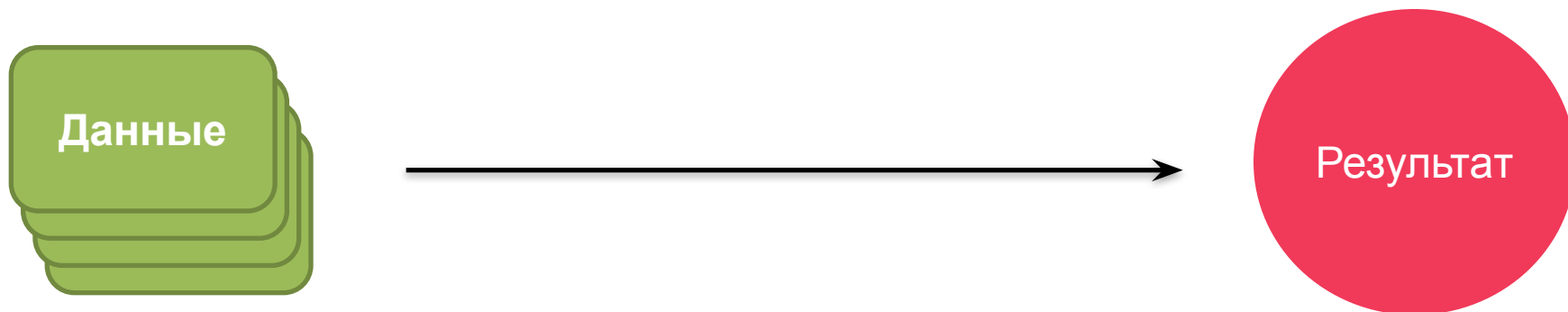
Мы знаем какой должен быть результат.



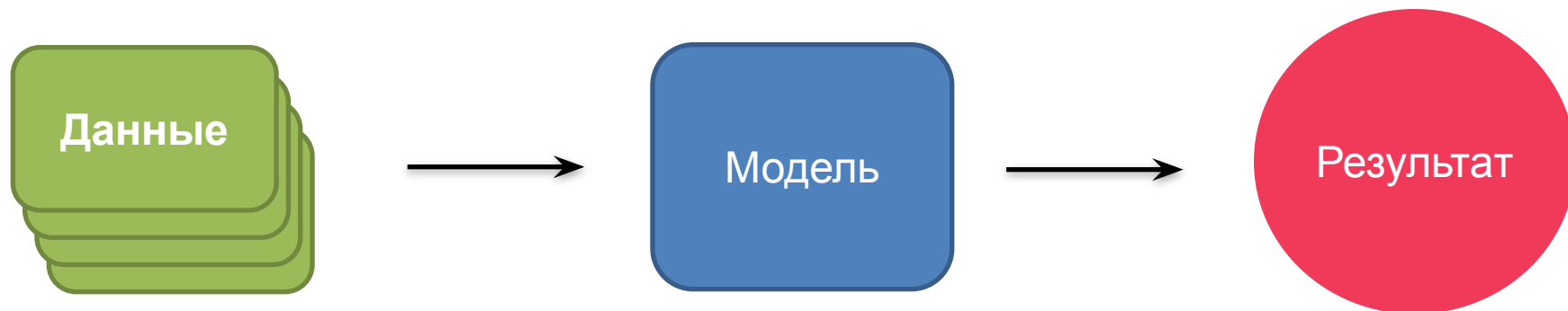
5 Что пошло не так?



Мы что-то где-то упустили.

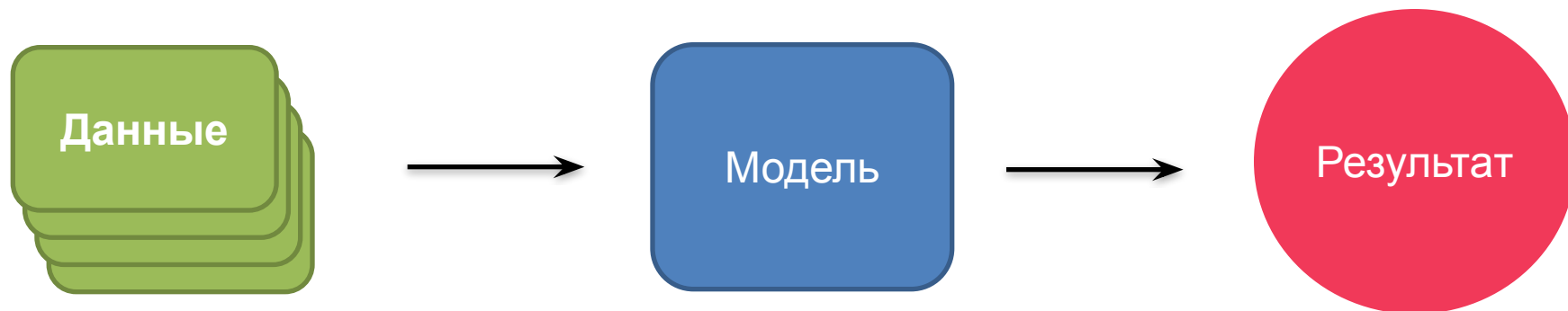


Мы поставили гипотезу: применим RF влоб, вдруг сработает.



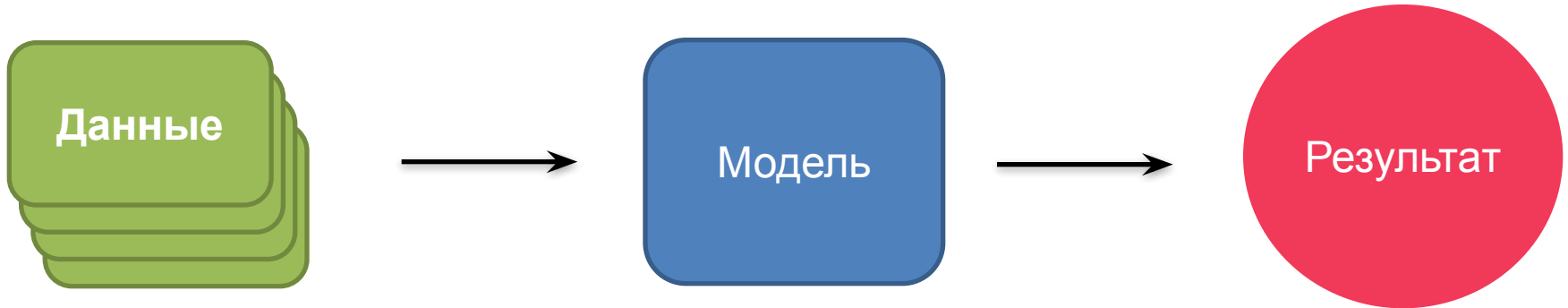
Что если мы ошиблись с RF/SVM?

То есть, все заработало, но результат был плох?



Что если мы ошиблись с RF/SVM?

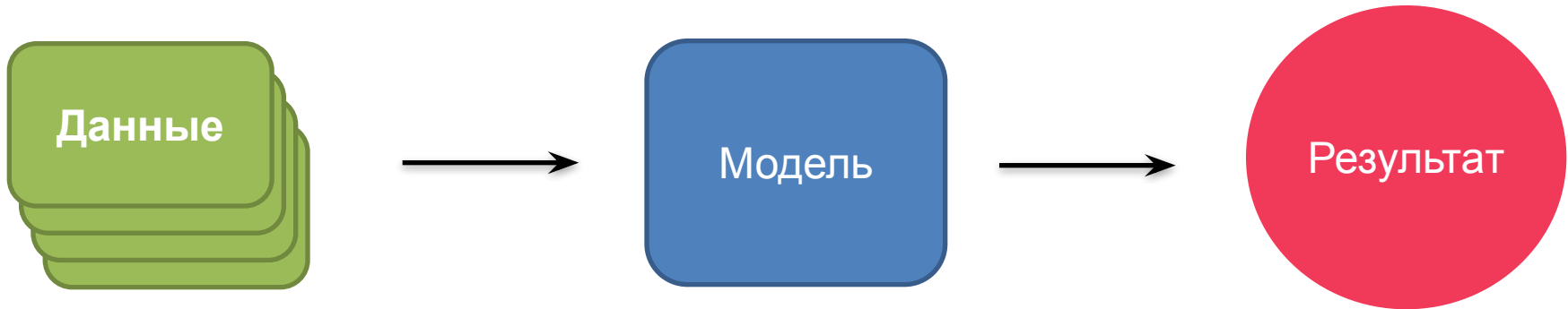
Мы же не накосячили с **тренировочным**, **валидационным** и **тестовым** множествами, а также CV?



Есть:

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





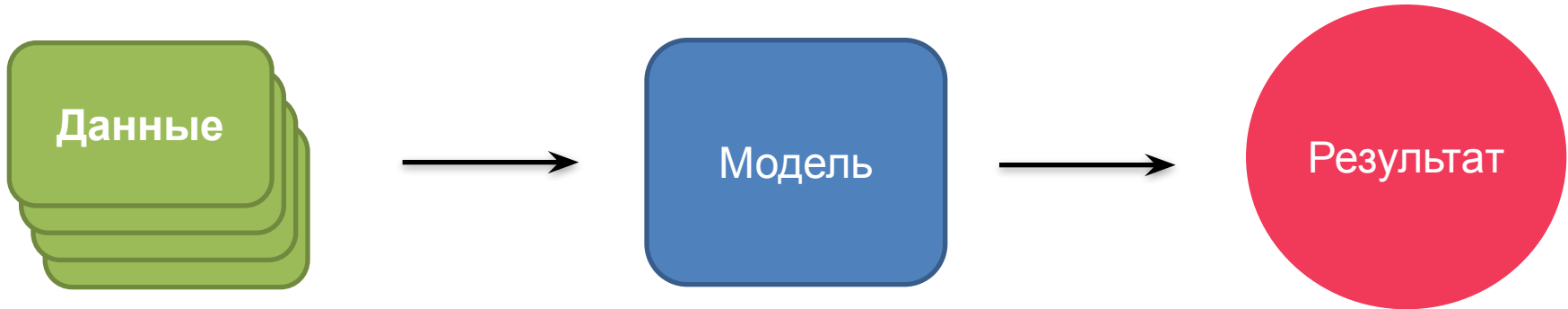
Есть:

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

Хотим:

- Улучшить (обобщающую) точность

Что мы можем сделать с моделями?



Хотим:

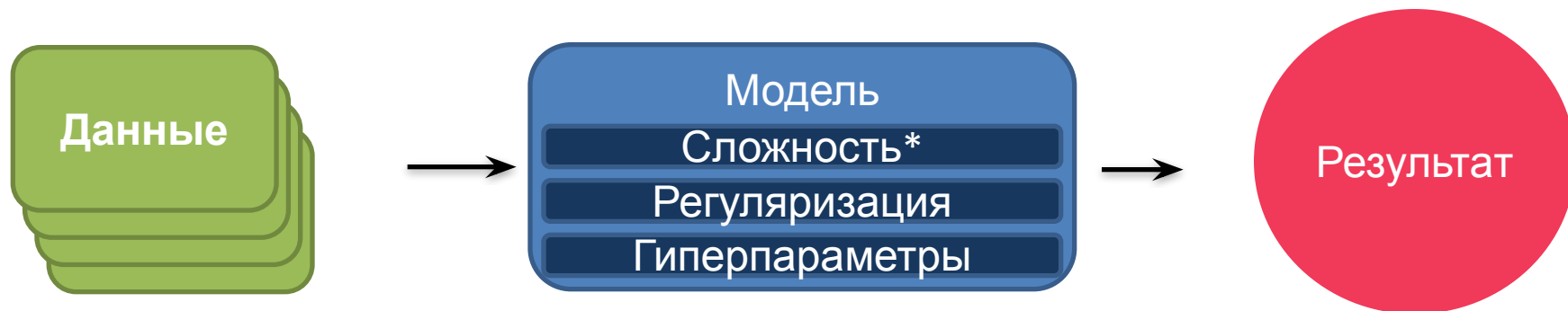
- Улучшить (обобщающую) точность

Чем можем управлять:

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

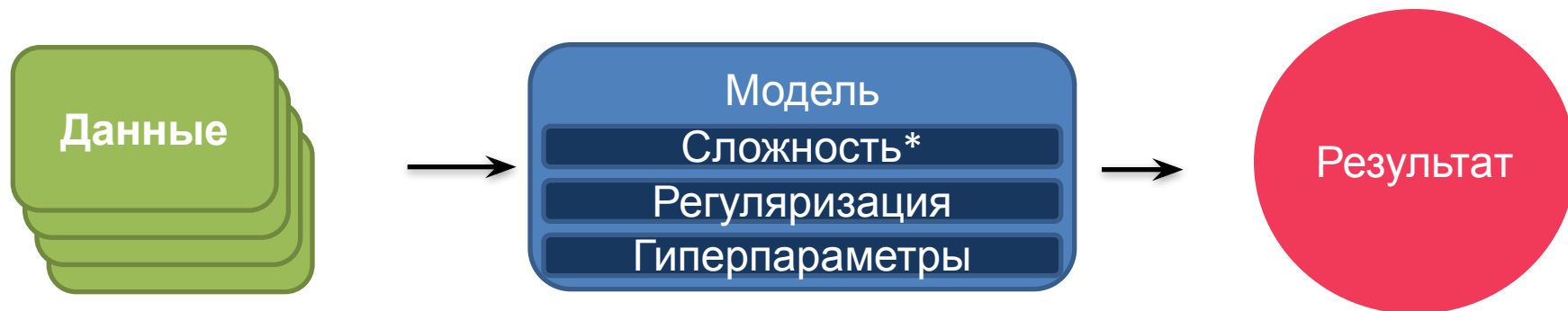
• Вид модели

Что мы можем сделать с моделями?

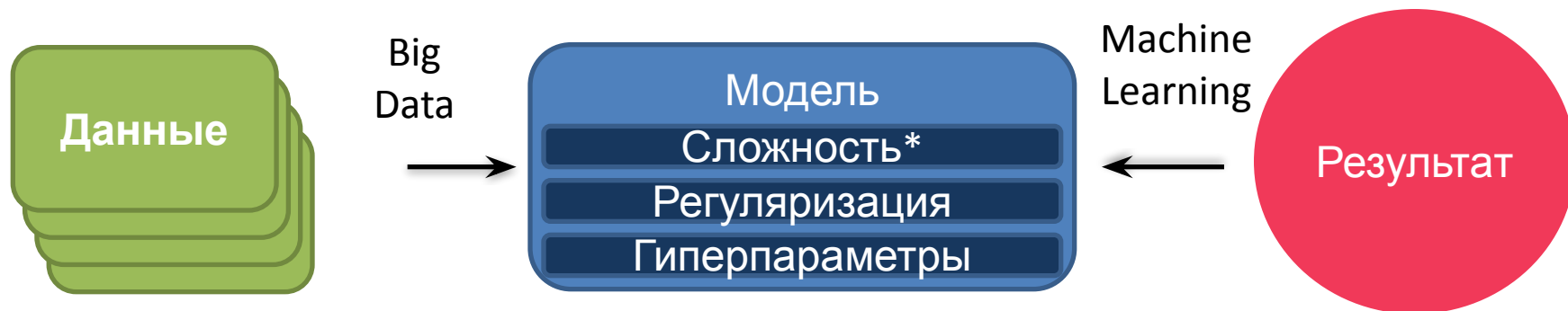


Выбор гиперпараметров тратит много машинного времени.

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



Где мы еще могли накосячить?

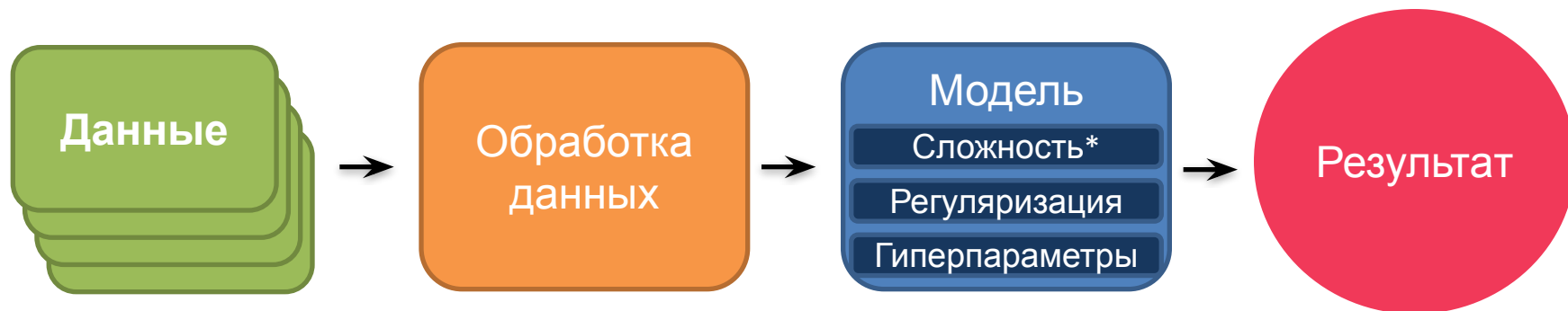


Где мы еще могли накосячить?

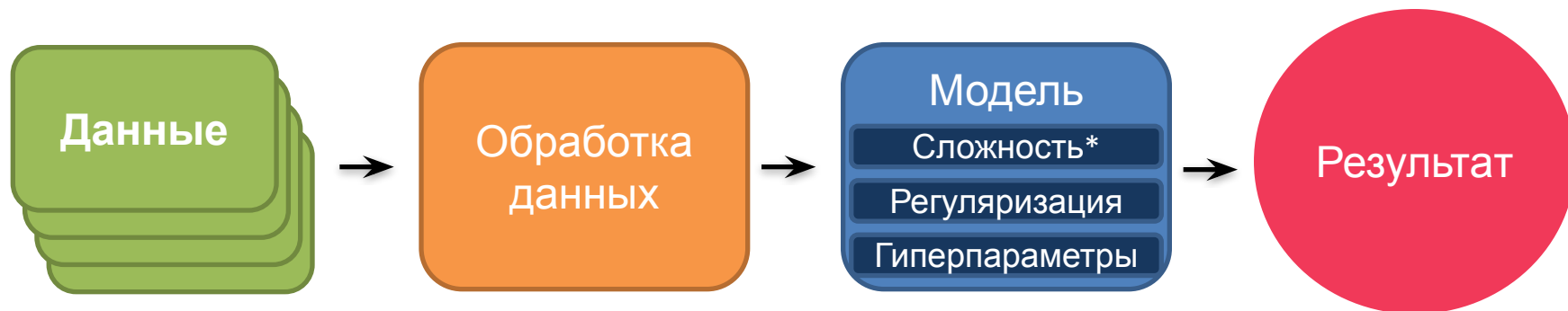


Оно вообще не посчиталось.

Иногда – феерично.

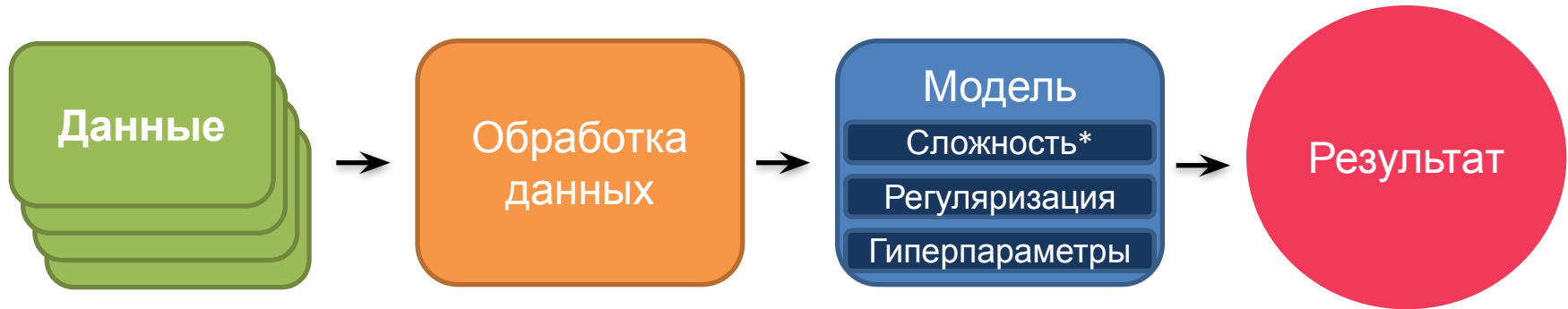


Тесно связана с моделью.



Хотим:

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



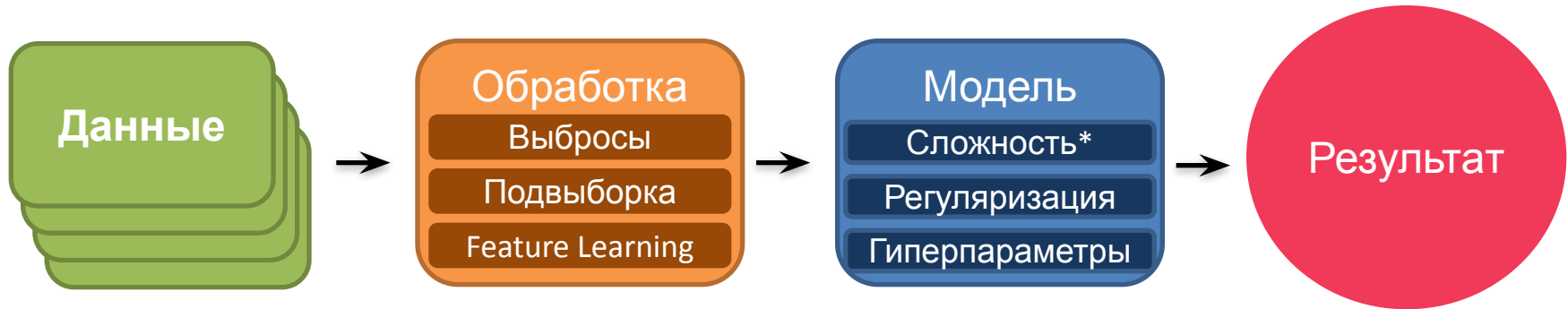
Хотим:

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

Чем можем управлять:

- Убрать выбросы
- Сделать подвыборку

- На уровне переменных:

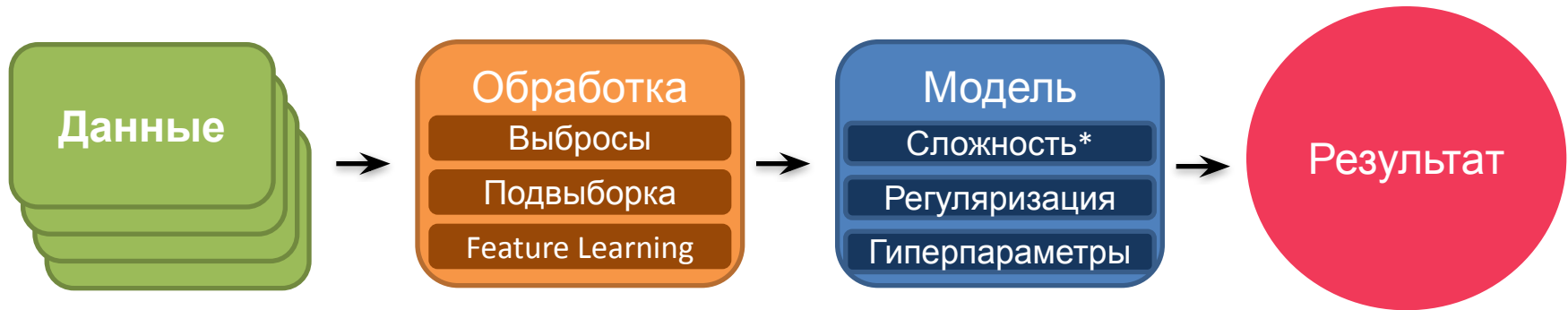


В подвыборки можно вкладывать смысл:
5% юзеров, записей, уникальных юзеров, последних записей...

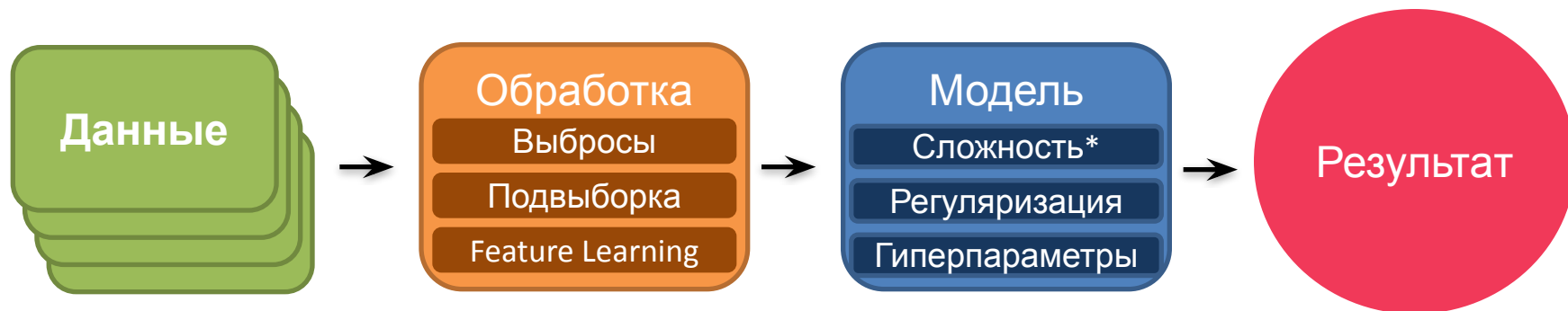
Feature learning – основа deep learning.

Unsupervised, обычно отдельно:

<http://web.eecs.umich.edu/~honglak/nipsdluf10-AnalysisSingleLayerUnsupervisedFeatureLearning.pdf>

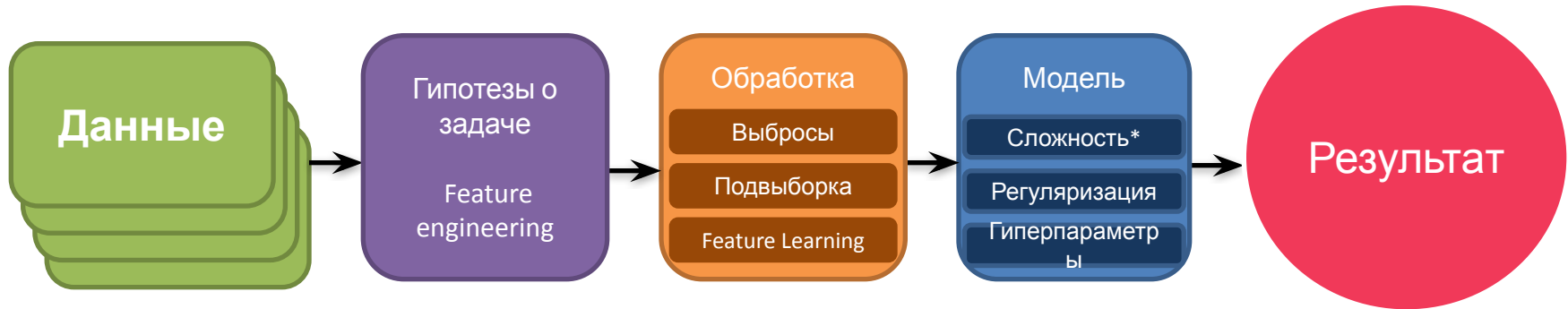


Что-то еще?



Что если все еще не взлетает.

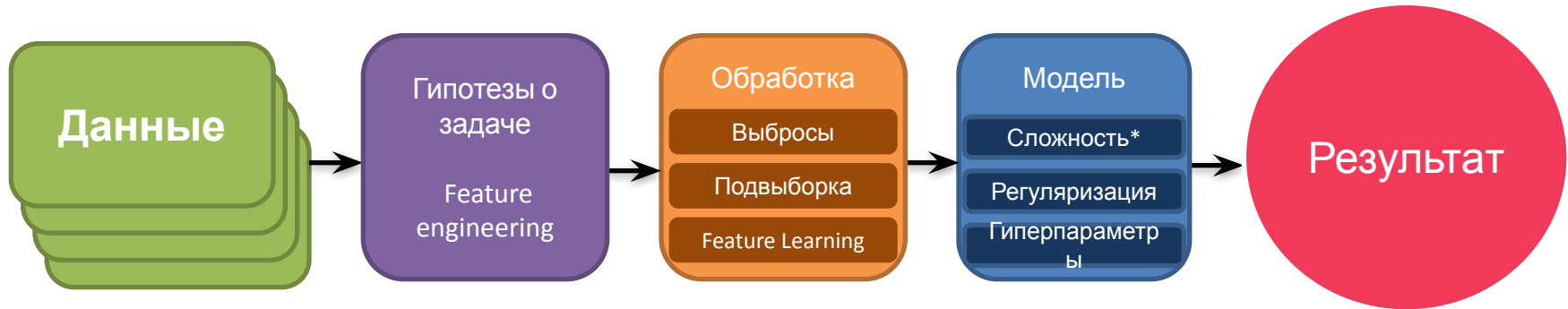




Самое сложное.

Понять что вообще происходит.

Неправильные фичи могут ни к чему не привести.



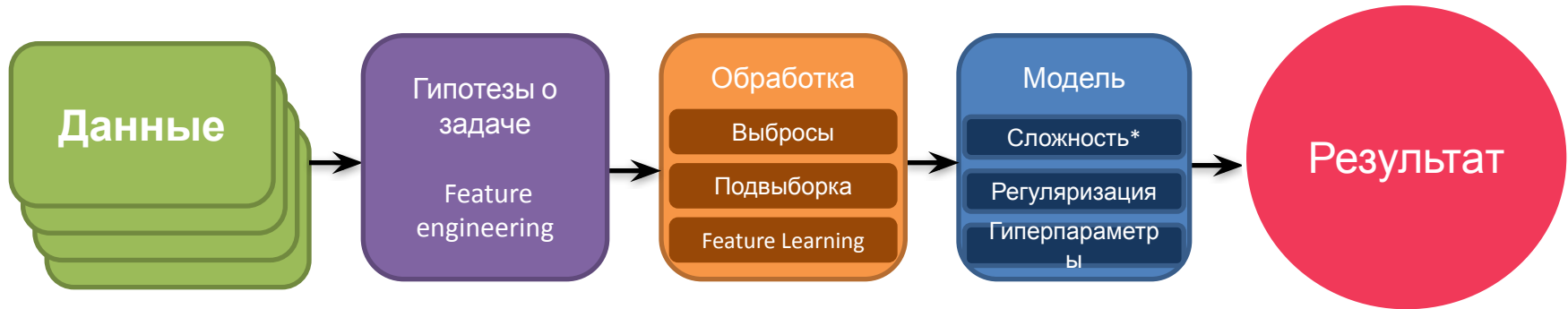
Есть:

- Мы перепробовали кучу моделей и кучу гипотез.

Чем можем управлять:

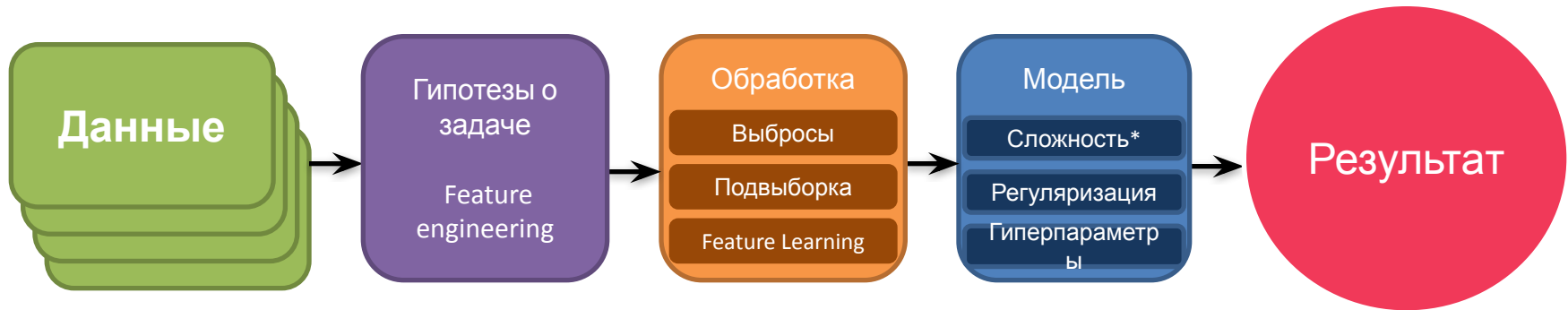
- Достать еще данных
- Feature engineering

- Посмотреть что делали другие:

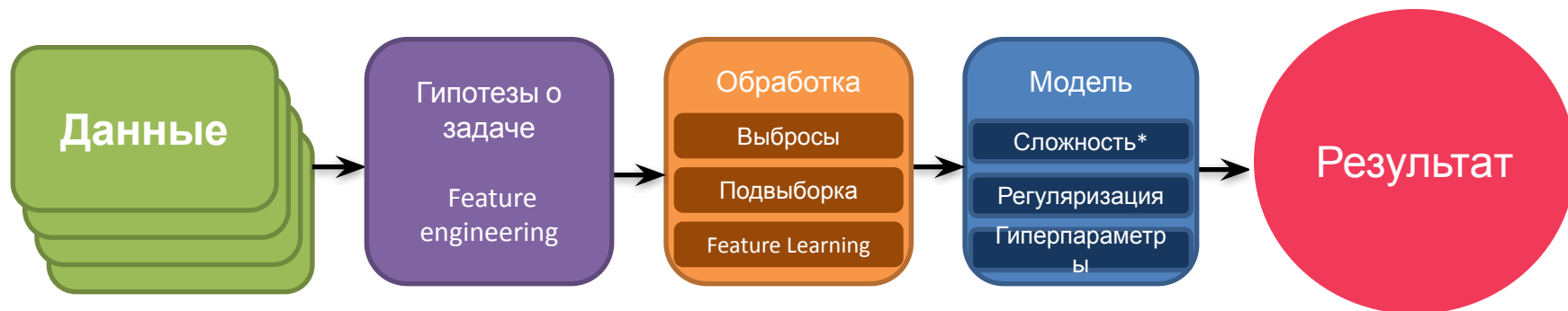


Пример:

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

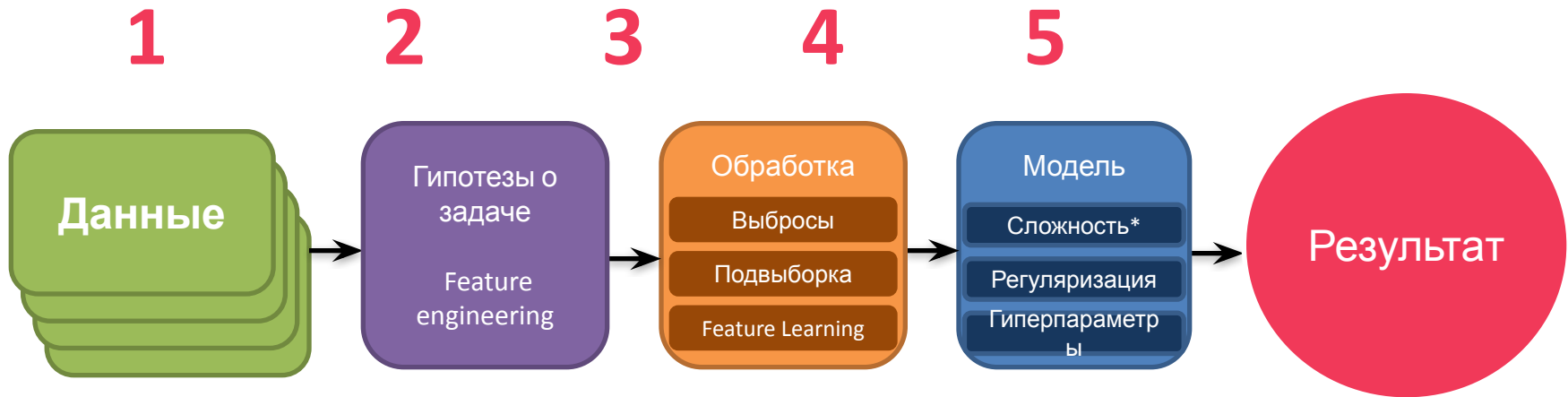


Как организовать команду?



Как организовать команду?





Техник: 1, 5

данные должны быть всегда доступны, сабмит всегда делался

Шпион: 2, 3, идеи про 4

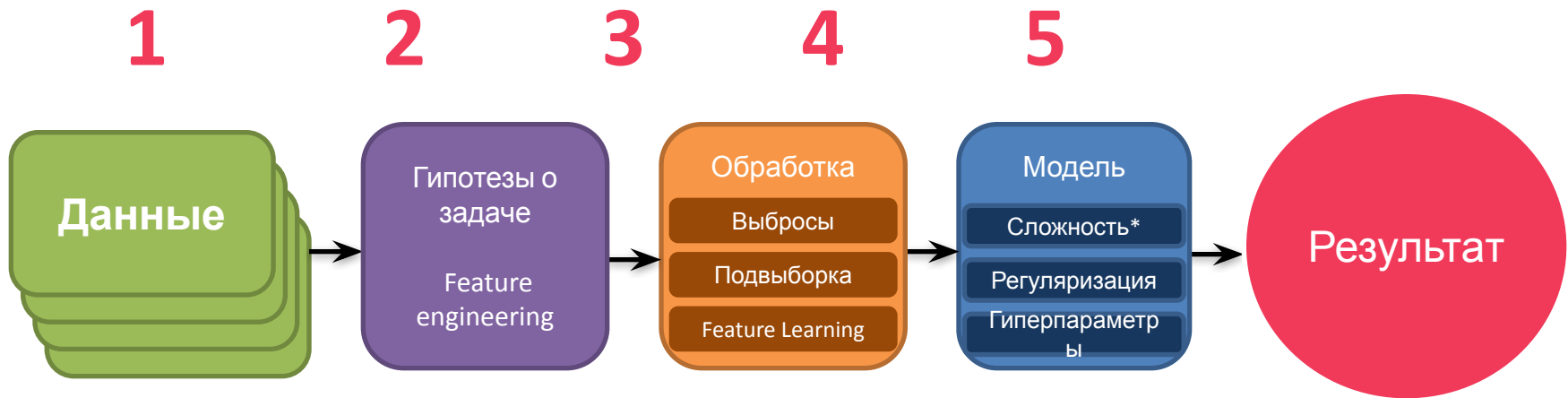
изучает литературу, статьи, форумы. Генерирует идеи

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



Техник: 1, 5

...

Шпион: 2, 3, идеи про 4

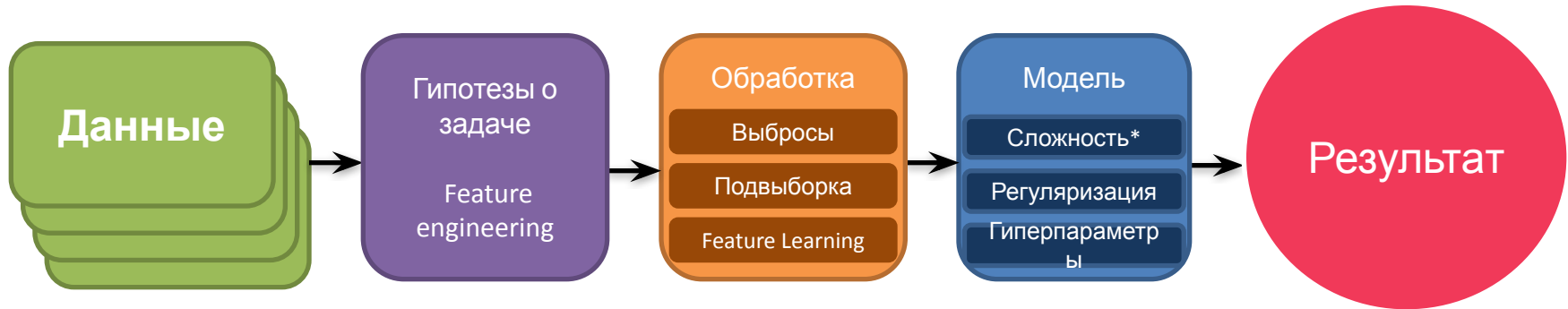
...

Спецназ: 3, 4

...

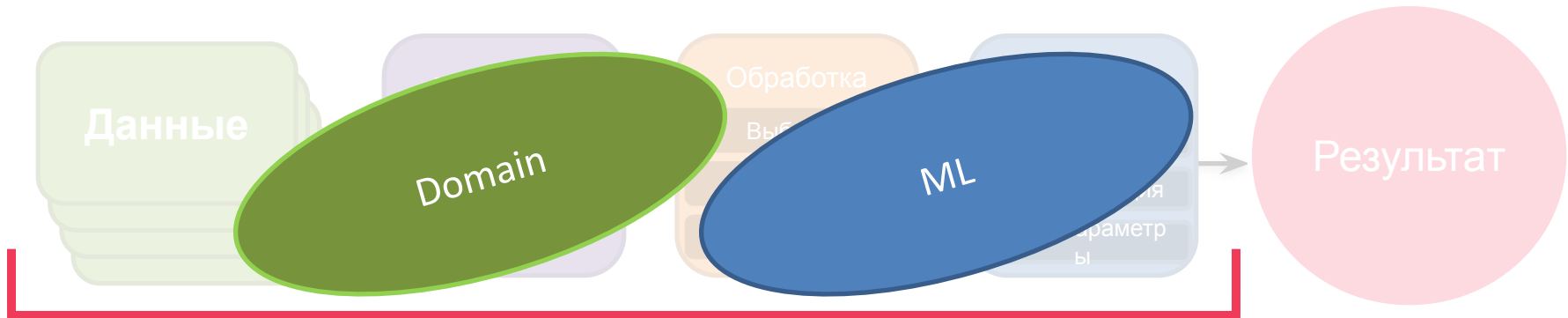
Капитан: 1, 2, 3, 4, 5

...



Где здесь результаты сообществ?

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



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, ...)

- ...

Стэкинг моделей:

Если вы вдруг сделали вообще все, можно ПОХИМИЧИТЬ

