Оценивание рисков распространения эпидемии по графу контактов методами имитационного моделирования и машинного обучения

Воронцов Константин Вячеславович
(д.ф.-м.н., проф. РАН, зав. лаб. машинного интеллекта МФТИ),

Зухба Анастасия Викторовна (к.ф.-м.н., доцент МФТИ),
Бишук Антон Юрьевич (студент МФТИ),
Рогозина Анна Андреевна (студент МФТИ)

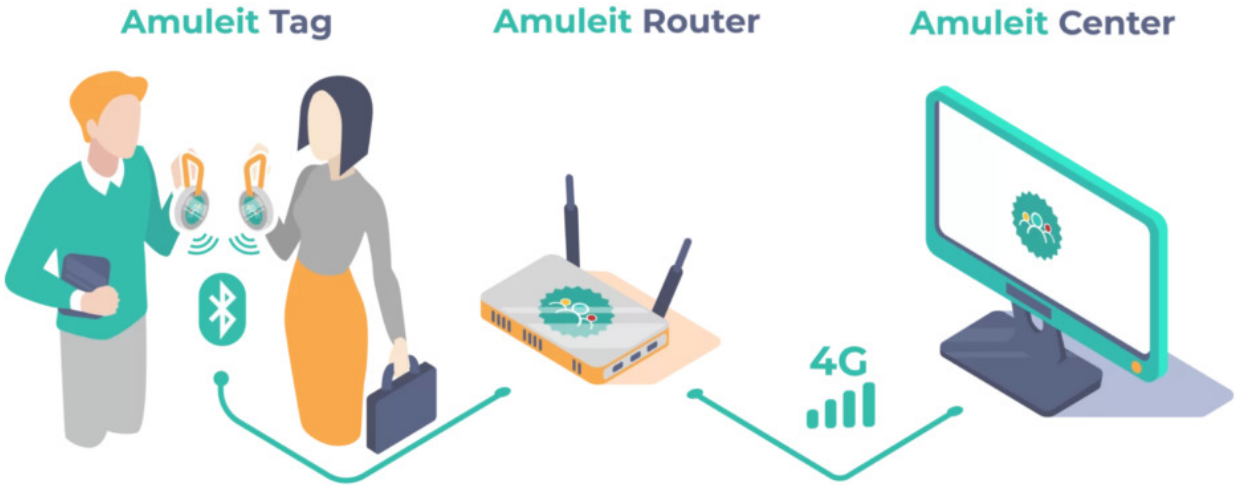
Круглый стол «Компьютерные симуляции
в исследовании макроэкономических процессов»
МИЭМ • 10 июня 2021

Содержание

Проект Amuleit: контроль эпидемии на предприятиях

Амулет — носимое устройство для сотрудников предприятий

Три основных элемента архитектуры системы Amuleit:



---

http://Amuleit.ru — сайт проекта

http://SoftTree.ru — сайт разработчика ООО Софттри (г.Пенза)

Как оценить риск инфицирования по данным о контактах?

Выборка данных:

- $\langle t : (u, v) \rangle$ — *контакт* индивидов u и v в момент времени t
- $\langle t : y(x) \rangle$ — *состояние* y индивида x в момент времени t

Состояния:

- S (susceptible) — восприимчивый здоровый
- E (exposed) — латентный инфицированный
- I (infected) — инфицированный больной
- R (recovered) — выздоровевший невосприимчивый

Задача — построить модель индивидуального риска:
$p(y|t, x)$ — вероятность состояния y индивида x в момент t

Наша мотивация — объединить два типа моделей:

- модели риска, обучаемые по выборке данных
- модели распространения эпидемии по графу контактов

Структура исходных данных по распространению Covid-19

Дана выборка контактов и состояний индивидов:

- $\langle t : (u, v) \rangle$ — *контакт* индивидов u и v в момент времени t
- $\langle t : y(x) \rangle$ — переход индивида x в *состояние* y в момент t

Найти модель индивидуального риска:

- $p(y|t, x)$ — вероятность состояния y индивида x в момент t

Критерий:

- максимум правдоподобия наблюдаемых состояний

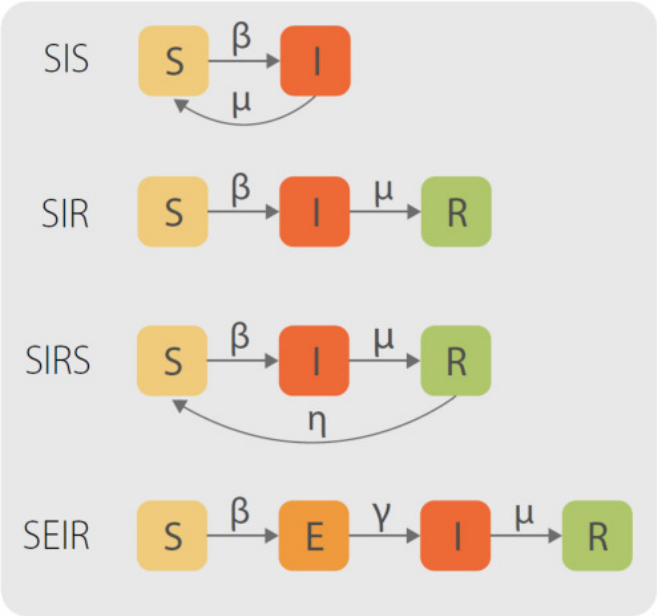
Обозначения и уточнения:

- $f(t, u, v)$ — информация о контакте, вектор признаков
- $y_t(x)$ — текущее состояние индивида x в момент t
- $y \in \{\mathtt{S}, \mathtt{I}\}$ — будем рассматривать только два состояния

Популяционные (компартментные) модели SIS, SIR, SIRS, SEIR

$S(t)$ — число восприимчивых
$E(t)$ — число латентных носителей
$I(t)$ — число инфицированных
$R(t)$ — число выздоровевших
$S + E + I + R = \text{const}$ — вся популяция

SIR — постоянный иммунитет
SIRS — временный иммунитет
SEIR — латентное инфицирование



$$\begin{cases} S' = -\beta IS + \mu I \\ I' = \beta IS - \mu I \end{cases}$$

$$\begin{cases} S' = -\beta IS \\ I' = \beta IS - \mu I \\ R' = \mu I \end{cases}$$

$$\begin{cases} S' = -\beta IS \\ E' = \beta IS - \gamma E \\ I' = \gamma E - \mu I \\ R' = \mu I \end{cases}$$

R. Pastor-Satorras et al. Epidemic processes in complex networks. 2014

Модель SIS для оценивания индивидуального риска $p_t(x)$

$p_t(x) = P(\mathtt{I}|t,x)$ — вероятность, что x инфицирован в момент t
$q_t(x)$ — вероятность, что x получил инфекцию в момент t

$$\frac{\partial}{\partial t} P(\mathtt{I}|t,x) = -\mu P(\mathtt{I}|t,x) + \beta\big(1 - P(\mathtt{I}|t,x)\big)\, q_t(x)$$

Пусть время дискретно с шагом 1. Конечно-разностный аналог:

$$p_t(x) = (1-\mu)\, p_{t-1}(x) + \beta\big(1 - p_{t-1}(x)\big)\, q_t(x)$$

Максимизация правдоподобия для оценивания параметров:

$$\sum_{t,x} [y_t(x) = \mathtt{I}] \ln p_t(x) + [y_t(x) \neq \mathtt{I}] \ln\big(1 - p_t(x)\big) \to \max$$

Параметры: μ, β и параметры вероятностной модели $q_t(x)$

R. Pastor-Satorras et al. Epidemic processes in complex networks. 2014

Модель логистической регрессии для вероятности $q_t(x)$

Гипотеза: вероятность, что x получил инфекцию в $[t-1, t]$

$$q_t(x) = \sigma\big(w_1 k_t(x) - w_0\big)$$

монотонно возрастает по числу его контактов

$$k_t(x) = \sum_{\langle t' : (x,v)\rangle} [t-1 \leqslant t' \leqslant t]$$

Недостатки этой модели — она не учитывает, что

1. контакты имеют различную вероятность передачи инфекции
2. индивиды v, контактирующие с x, имеют различную вероятность $p_t(v)$ быть инфицированными
3. при изменении состояния $y_t(x)$ индивида x должны измениться вероятности $p_t(u)$ для индивидов u, проконтактировавших с x незадолго до момента t

Модель вероятности передачи инфекции

Вместо числа контактов будем оценивать сумму вероятностей передачи инфекции по всем контактам в интервале $[t-1, t]$:

$$k_t(x) = \sum_{\langle t' : (x,v) \rangle} \big[t - 1 \leqslant t' \leqslant t\big]\, a_{t'}(x, v)$$

Вероятность передачи инфекции при контакте $\langle t : (x, v)\rangle$ можно оценить с помощью логистической регрессии:

$$a_t(x, v) = \sigma\Big(-\alpha_0 + \sum_{j=1}^{m} \alpha_j f_j(t, x, v)\Big),$$

где f_j — признаки потенциальной опасности контакта, например:
— расстояние между устройствами по уровню сигнала,
— продолжительность контакта.

Коэффициенты α_j являются параметрами модели.

Модель с рекурсивным оцениванием риска

Добавим оценки вероятностей $\tilde{p}_{t'}(v)$, что v инфицирован:

$$k_t(x) = \sum_{\langle t' : (x,v) \rangle} \left[t-1 \leqslant t' \leqslant t \right] a_{t'}(x,v)\, \tilde{p}_{t'}(v)$$

$$\tilde{p}_t(v) = \begin{cases} 1, & y_t(v) = \mathtt{I}; \\ p_t(v), & y_t(v) \neq \mathtt{I}. \end{cases}$$

Модель становится рекуррентной. Варианты обучения:

- брать текущие значения $p_t(v)$ с предыдущей итерации
- распространять градиент через суперпозицию функций, как в рекуррентных нейронных сетях
- промежуточный вариант: ограничивать распространение градиента небольшой глубиной рекурсии

Модель распространения риска по сети

Изменение состояния $y_t(x)\colon \mathtt{S} \to \mathtt{I}$ увеличивает оценки риска $p_{t'}(x)$ и $p_{t'}(u)$ для всех u, контактировавших с x, и цепочек контактов $x \to u \to v \to \ldots$ в недавнем прошлом $t' \in [t-d, t]$.

Индикатор, что x будет инфицирован в интервале $(t, t+d]$:

$$b_t(x) = \big[\exists t'\colon\ t < t' \leqslant t+d \ \text{ и } \ y_{t'}(x) = \mathtt{I}\big]$$

Добавим его в логистическую модель $q_t(x)$:

$$q_t(x) = \sigma\big(w_1 k_t(x) + w_2 b_t(x) - w_0\big)$$

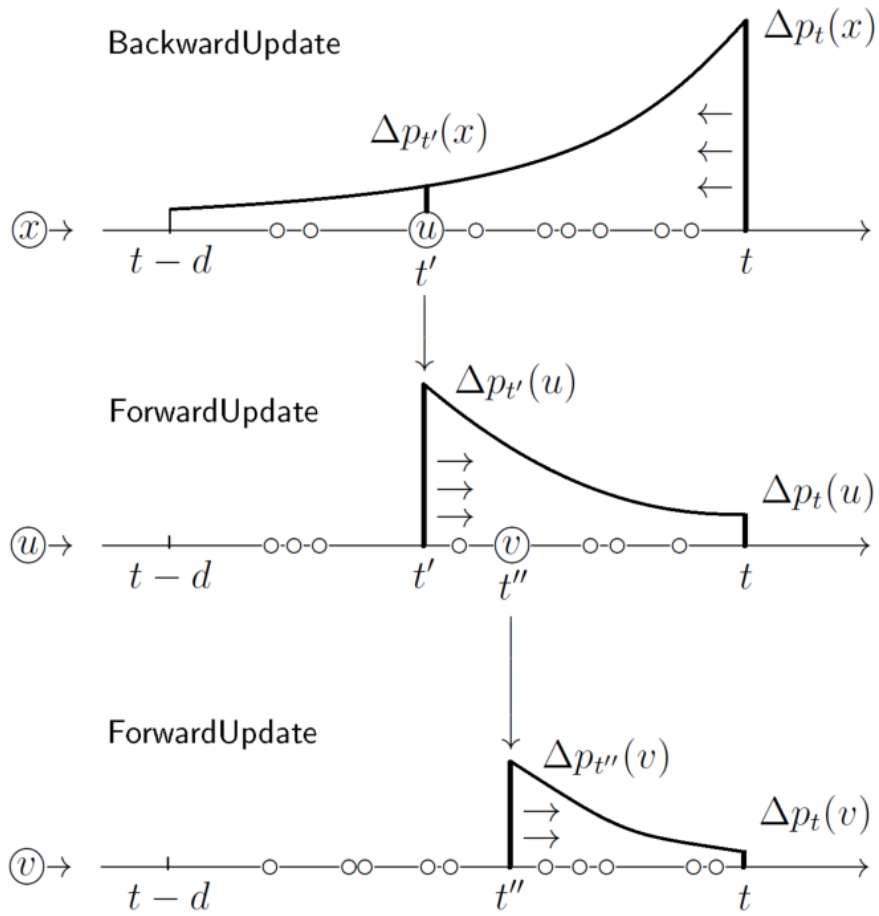
Алгоритм распространения риска по сети контактов запускается при переключении $y_t(x)\colon \mathtt{S} \to \mathtt{I}$, при этом риск скачком увеличивается до единицы, $\Delta p_t(x) = 1 - p_t(x)$, увеличиваются оценки $p_t(u)$, $p_t(v)$ и далее по цепочке контактов.

Алгоритм распространения рисков по сети контактов

Графики зависимости приращения риска Δp_t от времени t

функция **BackwardUpdate**(x, t);
$U := \varnothing$;
для всех $\langle t' \in [t-d, t] : (x, u) \rangle$
  пересчитать риск $p_{t'}(x)$;
ForwardUpdate$(x, t-d, t)$;

функция **ForwardUpdate**(x, t_0, t);
$U := U \cup \{x\}$;
для всех $\langle t' \in [t_0, t] : (x, u \notin U) \rangle$
  пересчитать риск $p_{t'}(u)$;
для всех $\langle t' \in [t_0, t] : (x, u \notin U) \rangle$
  если $\Delta p_{t'}(u) > \varepsilon$ **то**
    ForwardUpdate(u, t', t);



Модель индивидуального риска (собираем всё воедино)

Максимизация правдоподобия для оценивания параметров:

$$\sum_{t,x} [y_t(x) = \mathtt{I}] \ln p_t(x) + [y_t(x) \neq \mathtt{I}] \ln\bigl(1 - p_t(x)\bigr) \to \max$$

Модель SIS для вероятности, что x инфицирован в момент t:

$$p_t(x) = (1 - \mu)p_{t-1}(x) + \beta\bigl(1 - p_{t-1}(x)\bigr)q_t(x)$$

Модель вероятности, что x получил инфекцию в $[t-1, t]$:

$$q_t(x) = \sigma\bigl(w_1 k_t(x) + w_2 b_t(x) - w_0\bigr)$$
$$k_t(x) = \sum_{\langle t' : (x,v)\rangle} \bigl[t - 1 \leqslant t' \leqslant t\bigr]\, a_{t'}(x, v)\, \tilde{p}_{t'}(v)$$
$$b_t(x) = \bigl[\exists t' : t < t' \leqslant t + d \text{ и } y_{t'}(x) = \mathtt{I}\bigr]$$

Модель вероятности передачи инфекции при контакте (x, v) в t:

$$a_t(x, v) = \sigma\bigl(-\alpha_0 + \sum_{j=1}^{m} \alpha_j f_j(t, x, v)\bigr)$$

Замечания об Алгоритме распространения рисков

Особенности метода стохастического градиента:

- индивида x выбираем случайно
- по времени t проходим последовательно, формируя сбалансированную по классам $\{\mathtt{S}, \mathtt{I}\}$ выборку
- обновляем $p_t(u)$ при фиксированных параметрах
- только полностью обработав данные по индивиду x, делаем накопленный градиентный шаг по параметрам

Возможные модификации:

- Ограничение неотрицательности коэффициентов w_k, α_j
- Учёт большего числа состояний $\mathtt{S}$, $\mathtt{E}$, $\mathtt{I}$, $\mathtt{R}$ и др.
- Кластеризация индивидов по когортам

Имитационное моделирование для машинного обучения

Имитационное моделирование используется для расширения (аугментации) обучающих выборок:

- замена реальных данных $y_t(x)$ имитацией распространения инфекции по графу контактов
- имитация стартовых условий эпидемии (начальное множество инфицированных и вероятности $q_t(x)$)
- имитация различных страт внутри популяции по отношению к инфицированию извне
- имитация реалистичных графов контактов на основе имеющихся реальных графов

Параметры имитационной модели

Вероятности за сутки в модели SEIR:

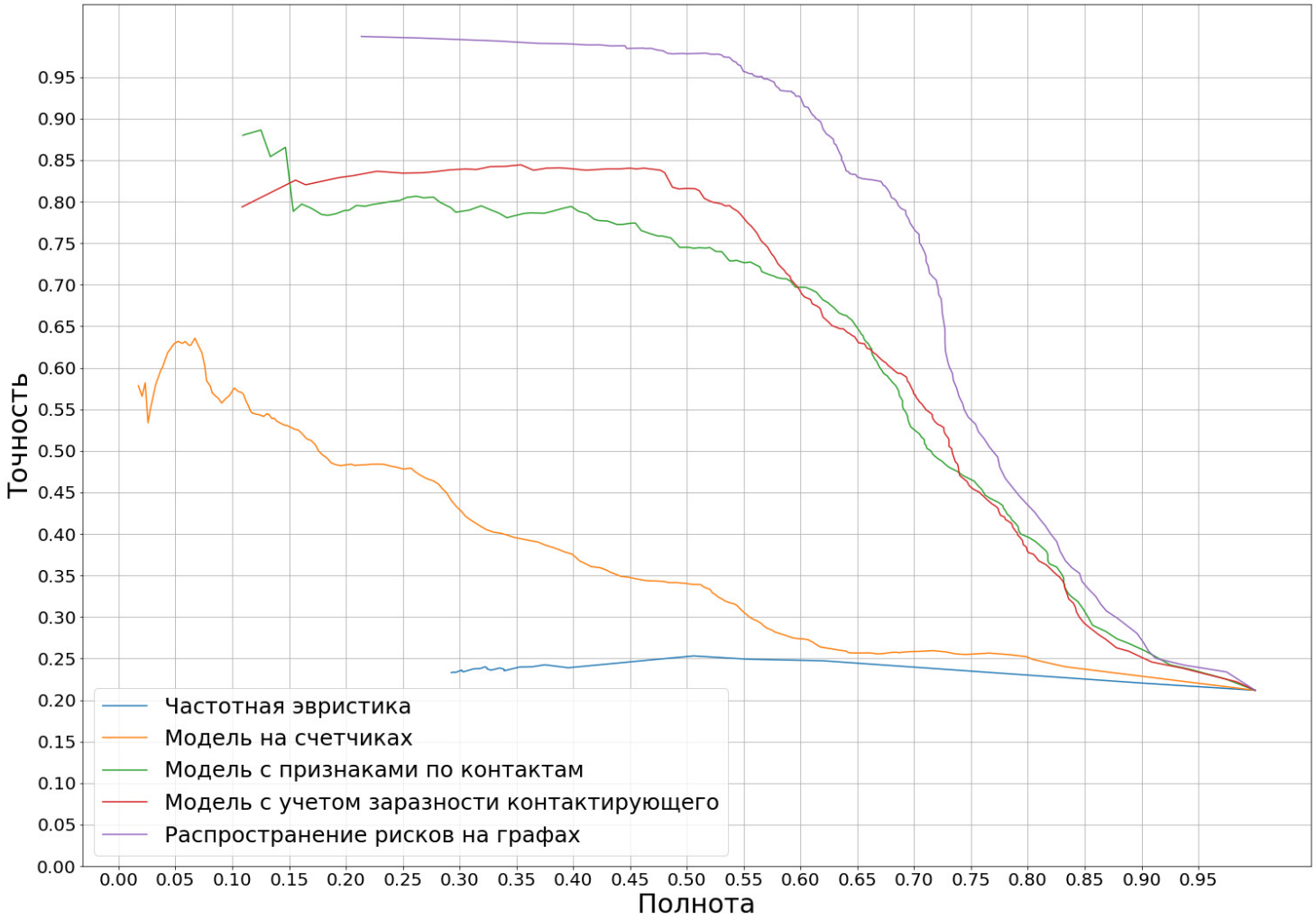
- $P(\mathtt{S} \to \mathtt{E}) = 0.05$ — инфицироваться
- $P(\mathtt{E} \to \mathtt{I}) = 0.4$ — заболеть
- $P(\mathtt{I} \to \mathtt{R}) = 1/16$ — выздороветь
- $P(\mathtt{R} \to \mathtt{S}) = 1/64$ — снова стать восприимчивым
- вероятность инфицирования зависит от времени контакта:

часы:	1	2	3	4	5
вероятность:	0.65	0.81	0.90	0.95	0.98

- задаётся доля инфицированных в начальный момент

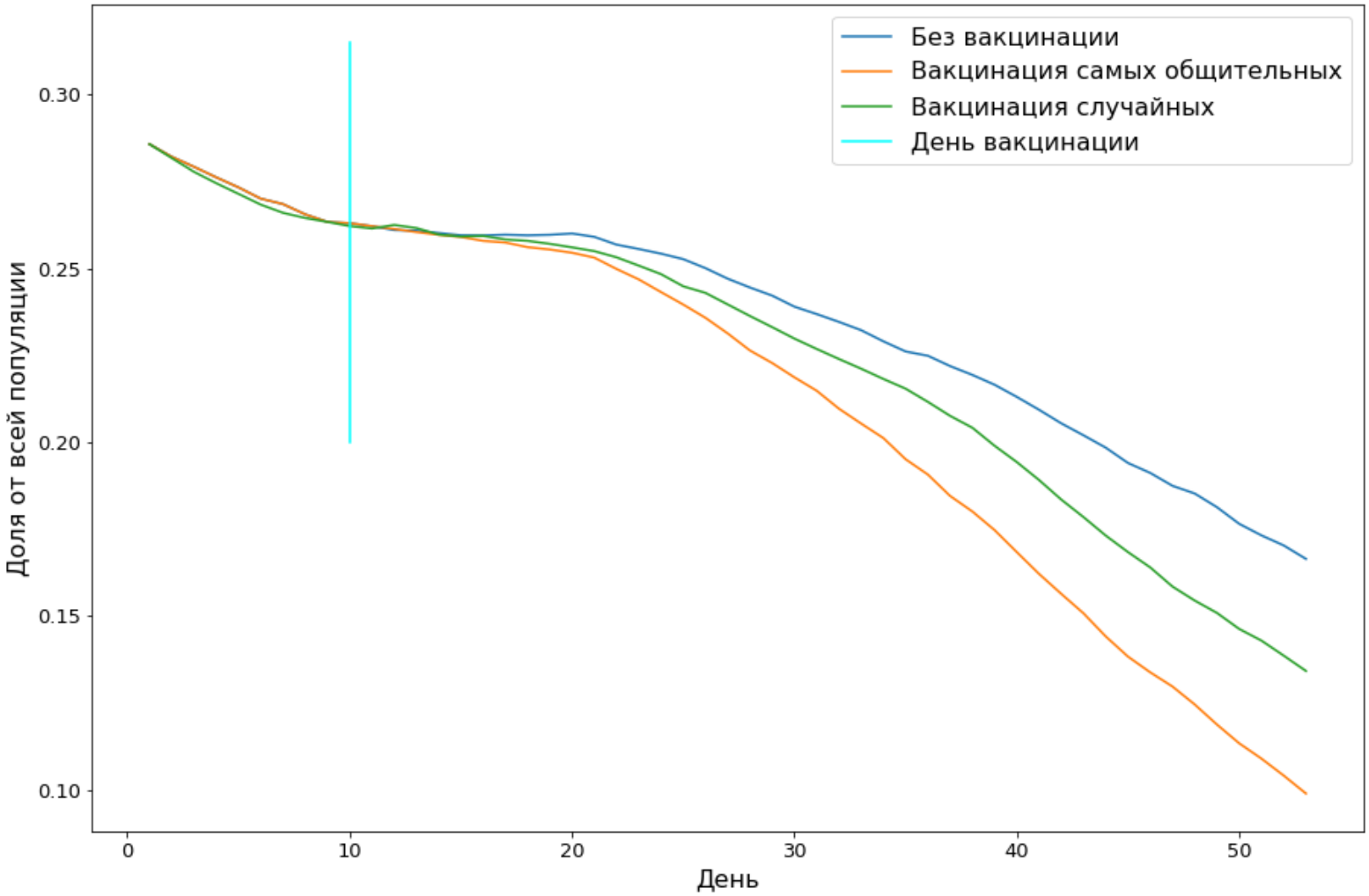
Эксперимент №1: точность и полнота вероятностных моделей

Обоснование постепенного усложнения моделей:



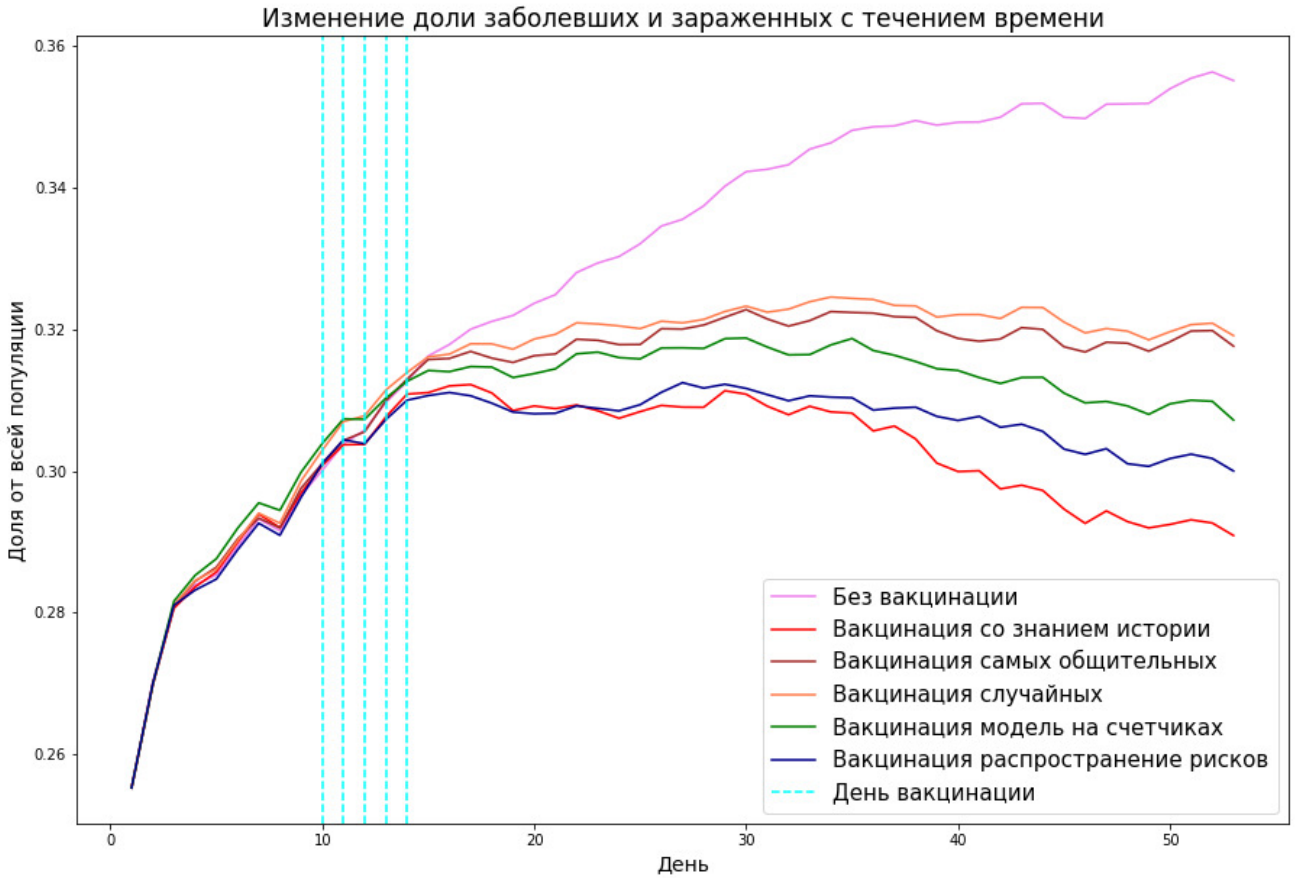
Эксперимент №2: сравнение стратегий вакцинации

Вакцинация 10% сотрудников:



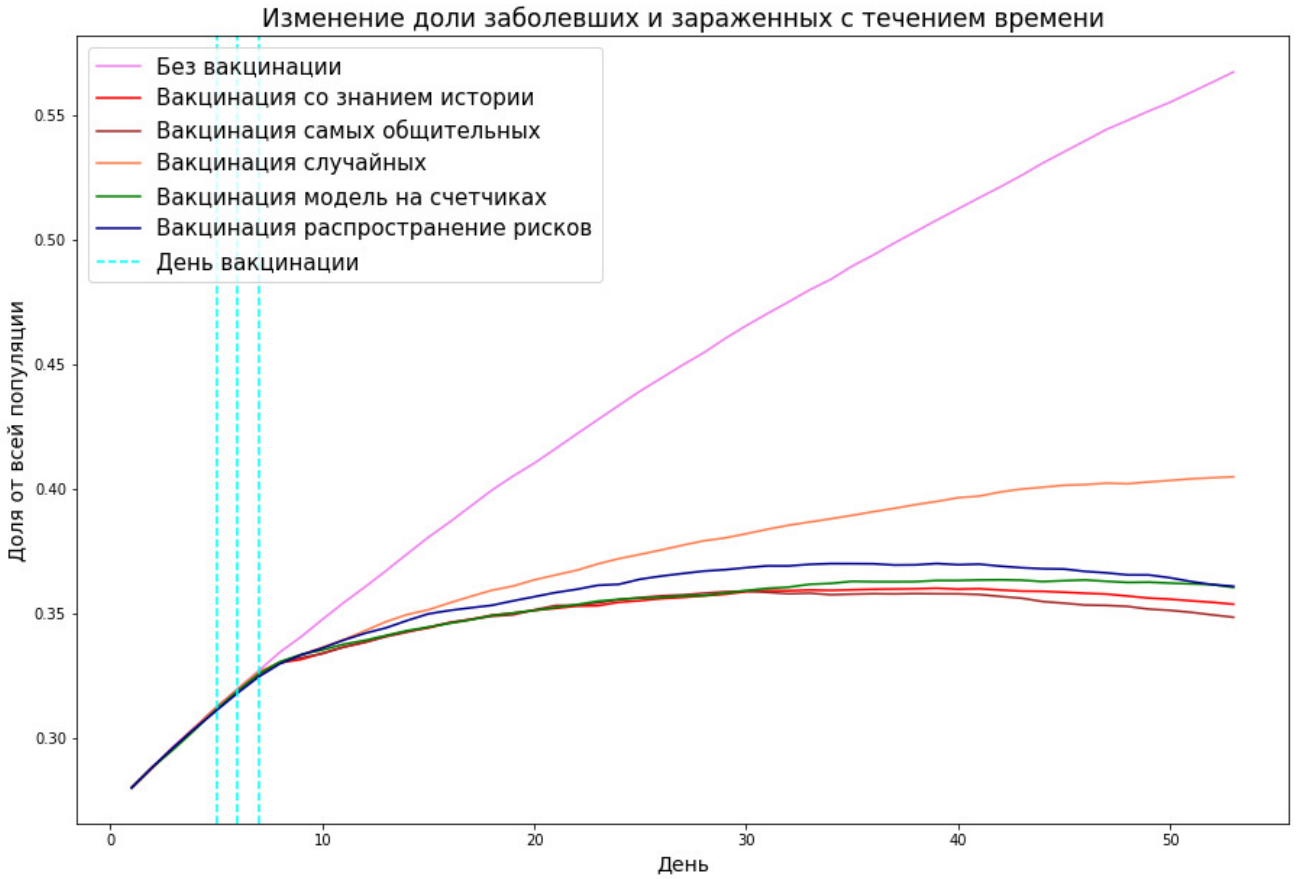
Эксперимент №3: сравнение стратегий вакцинации

Модель распространения рисков приближается к «идеальной»
Вакцинация 10% сотрудников в течение пяти дней:



Эксперимент №4: сравнение стратегий вакцинации

Вакцинация 50% сотрудников может быть недостаточна:



Резюме

Мы объединили три подхода к моделированию:

- эпидемиологические модели на уровне индивидов
- машинное обучение для оценивания моделей риска
- алгоритмы распространения информации по графу

Новизна таких моделей в математической эпидемиологии:

- большие данные о контактах появились только недавно, в связи с пандемией Covid-19
- распространение информации в прошлое по графу контактов
- использование методов машинного обучения