Один з прикладів Data Science. Коли усім людям будуть показувати однакову рекламу, то ймовірність того, що хтось нею зацікавить, клікне на посилання і купить в результаті товар, не є надто висока. Це так само як показувати усім однакову рекламу кави Jackobs. Якщо ви можете вгадати рекламу для користувачів, які каву не п'ють, показавши рекламу, наприклад чаю Lipton, тоді шанси, що реклама окупиться більші. І замовник наступного разу прийде до вас ще. Цей підхід є досить поширений.

Запитання до аудиторії: які ще приклади Ви знаєте?

- Прогноз погоди, кластеризація новин.

Якщо підійти до кластеризації новин більш ширше, то можна говорити про персоналізацію контенту. Медіа контент може бути різним, і ми його персоналізуємо (реклама автомобілів, політична реклама, прогноз погоди), адже різних людей будуть цікавити різні речі. Припустимо, якщо у вас є автомобіль, то для важливий факт чи буде ожеледиця чи ні. При цьому температура повітря не є вже настільки критичним фактором. Важливо, які пункти висвітлювати на дисплеї, від цього залежить чи буде людина задоволена тим, що вона бачить (особливо вранці поспішаючи не пропустити щось важливе).

- Розпізнавання зображень і голосу.

Речі, як стосуються обробки мови, розпізнавання зображень, голосу, вони стоять в стороні від Data Science по одній причині: людина дуже добре справляється з цими завданнями, адже ми дуже добре читаємо тексти, розуміємо зображення і т.д. І коли стоїть задача, наприклад, розпізнати зображення чи перевести голос в текст, тоді аналітик даних прекрасно розуміє, що в тих даних є і що з них можна вижати. Він, наприклад, розуміє: якщо людина не дуже чітко виговорила і система надала погану транскрипцію, то це є наслідком проблеми з дикцією. Тобто ці дані легко піддаються аналізу. Але є дані, які піддавати аналізу складно, наприклад, кліки в соціальних мережах (коли їх велика кількість і вони дуже різні).

- Перегляд фільмів онлайн.

Якщо ви створювали аккаунти на подібних сайтах, то напевно зауважували рекомендації для перегляду інших фільмів. Аналіз тут трохи складніший, адже кількість фільмів велетенська. На основі чого система робить рекомендації? – На основі того, що ви вже подивились, на основі різних жанрів. Та в першу чергу система аналізує тих людей, які переглянули такі самі фільми як і ви, з яким у вас велике перекриття, наприклад у вас 80% спільних переглядів. Це дозволяє системі зрозуміти, що у вас смаки збігаються. Але при цьому у вас 20% фільмів відрізняються. Тому з цих 20% вам пропонуватимуться різні пропозиції. Причому вибиратись фільми будуть розумно: якщо наприклад є 100 людей, схожих між собою, і 80 з них переглянути якийсь фільм, то решті двадцяти буде пропонуватись переглянути даний фільм. Є багато даних: рік виходу, жанр, країна-виробник і т.д. Такий підхід називається колаборативна фільтрація. Такі дані людині важко проаналізувати: велика кількість даних, велика кількість параметрів і це все важко скласти в голові. Тому це трохи інший варіант в аналітиці даних, коли дані є незрозумілими і аналітик витрачає багато часу, щоб зрозуміти, що є в цих даних і що з них можна витягнути.

Коли ви працюєте з даними, то найперше питання, яке потрібно ставити: що корисного можна витягнути з цих даних? І якщо можна, то що саме?

Застосування аналітики даних в медицині:

- Діагностика, обробка медичних зображень. Це заміняє роботу для рентгенологів, радіологів і т.д.

В більшості випадків рентген роблять здоровим людям (контрольний захід). Тому 90% знімків не потребують великої уваги з сторони лікаря, бо все гаразд. Тому таке дуже легко автоматизувати. Ті 10%, коли є підозра що щось не так, потрібно знімок показувати спеціалісту. Відповідно виграш з цього – ми можемо утримувати в 10 раз менше рентгенологів. Це саме стосується і радіологів, коли

мова йде про рак і т.д. Чому такого не відбувається? Потрібно зібрати великий обсяг даних (наприклад у нас є 2 млн. оцифрованих знімків і тут виникає питання техніки і грошей). Також нам потрібно незалежні спеціалісти, які б пояснили що є на знімках. Але час таких спеціалістів коштує досить дорого. Тому окупність такої системи є досить великою. Також є питання довіри. Тобто якщо вам робот скаже, що у Вас все гаразд, але ви можете доплатити, щоб знімок переглянув спеціаліст, то ви скоріш за все виберете варіант доплатити. Медицина – дуже консервативна галузь, адже ніхто не хоче ризикувати своїм здоров'ям, особливо якщо ціна питання не надто велика. І це є одним з моментів, які зупиняють автоматизацію. Адже люди більше довіряють медичним спеціалістам. Інша проблема, коли система, робот робить помилку і неправильно діагностує. Коли це робить медик, то ми маємо термін «лікарської помилки», але коли помилку робить робот, хто винен в цьому випадку: програміст чи медик, який неправильно розмітив дата-сет, чи компанія? Важливим є питання відповідальності: фінансову відповідальність оплатить компанія, а хто буде нести кримінальну відповідальність? Етичні питання штучного інтелекту частково стоять на заваді Data Science в медицині.

Ми підходимо до того, що стирається межа між healthcare і well-being. Приклад, коли вам лікар говорить після якоїсь травми робити вправи – це скоріш охорона здоров'я (бо реабілітація – це частина охорони здоров'я). А коли ви робите вправи, коли ви здорові, то це вже спосіб життя. Інший дотичний до цього момент – дієтологія. Наша компанія мала свого часу проект побудови персоналізованої дієти, куди люди вносила свої біологічні маркери, аналізи, і на основі цієї інформації утворювалась персоналізована дієта. Та зазвичай після формування таких дієт, вони переглядаються медиком, щоб система не «нарекомендувала» чогось незрозумілого.

Найбільш розвиненими галузями, коли є говоримо про data science в healthcare є:

1. Біоінформатика: аналіз геному, персоналізовані ліки і т.д. В цій галузі є багато невирішених завдань, адже багато речей, які з нами стаються, залежать не від нашого стилю життя, не від того, що споживаємо, де ми живемо, а від нашої генетики. Наприклад, схильність до раку великою мірою визначається не тільки шкідливими звичками, а й особливостями будови геному. Зробити висновок з генів про те, чи є чи нема схильності до певних хвороб – дуже складно. Геном не розшифрований. Тому одні з найбільш інтенсивних досліджень ведуться в цьому напрямку. «Зламати» (розшифрувати) геном або більшу його частину, прив'язати його до якихось практичних сценаріїв, це дозволяє різко комерціалізовувати. Якщо людині грунтовно доводити, що в неї є ризик і потрібно витрачати велику суму коштів на профілактику хвороби, то зазвичай людина буде витрачати ці гроші, буде витрачати її страхова компанія, бо профілактика може обійтись дешевше, ніж лікування.
2. Різні діагностики. Є об'єктивні перешкоди, що люди довіряють більше медикам і ще немає такого сплеску, хоча ведуться серйозні роботи і можливо через якийсь час це зміниться.
3. Коли ми намагаємось проаналізувати стиль життя людини, зробити її комплексний огляд. Якщо людина харчується досконало, але ми можемо знати які ліки вона приймала, якими хворобами хворіла, які в неї були травми і т.д. Ми маємо медичну історію, в багатьох країнах вона є в електронному вигляді. В Україні починають робити кроки в цьому напрямку. Є так званий термін EMR (electronic medical records) або HMR (health medical records). Це фактично база даних, в які внесені історії пацієнтів. Кожна хвороба кодується певним кодом.

Наша компанія працює з такими медичними даними пацієнтів. Звичайно, що таких даних є багато. Уявіть, що у вас є дані 100 млн. пацієнтів (чим хворіли, якими ліками лікувались і т.п.). І що можна зробити з такими даними? Можна вести статистику хто в якому віці якими хворобами хворіє, що призводило до цього, які звички призводили до захворювання. Можна вибирати лікування, базуючись на тому, на що у людини виникла алергія. Можна прогнозувати на які хвороби може захворіти пацієнт. Можна визначати які ліки ефективні, які ні, які лікарі ефективніші.

Якщо ви витрачаєте ресурси на аналіз, то маєте отримати щось таке, на чому можна заробити, тобто що можна використати, щоб окупились затрати, тобто надати якусь послугу пацієнтам або компаніям, які займаються охороною здоров'я.

В чому є трудність прогнозування захворювань: після аналізу даних Вам видасть уже всім відомі речі, наприклад «Після 60 років ризик інфаркту суттєво зростає» чи «Якщо Ви жінка, то отримати рак грудей у Вас вище, ніж у чоловіків», або «Якщо Ви жінка, то ризик раку простати у вас прямує до нуля». Буде багато тривіальних речей. З іншої сторони, внаслідок аналізу даних, ви отримаєте якісь гіпотези, які медики дуже довго будуть ставити під сумнів. Наприклад, «Якщо ви будете їсти багато яєць, то може розвинутись якась хвороба». Частина медиків скаже, що це якісь інші речі, це просто кореляції. І вони можуть бути праві. Дані зазвичай є неповні. Є ще якийсь фактор додатковий, якого ми не знаємо. І тому ми зазвичай кажемо: «Споживання чогось підвищує ризик». Наприклад як «куріння підвищує ризик отримати рак легень», але є багато курців, які проживають життя і не мають ніяких проблем з легенями, а є ті, хто не палять і мають проблеми. Загалом відома річ: якщо ви палите – ризик раку легень зростає. Та який є додатковий фактор, через який в одних людей є резистенція до тютюну,а інші не мають, цього поки до кінця невідомо. І таких прикладів багато. Ми з одної сторони отримаємо або відомі речі, або отримаємо якісь речі, які медики підозрюють, але не можуть до кінця пояснити через відсутність цих даних. Є певні аналізи, які ми не можемо зробити, щоб зрозуміти причину. Цей момент нас обмежує.

Заміна лікарів. Приблизно щось подібне з діагностикою. Мало хто погодиться, щоб комп'ютер йому виписав рецепт і буде його слідкувати. Люди будуть з обережність до цього ставитись, бо завжди є ризик, що система спрацювала неправильно, її дані, на відміну від лікарів, вони обмежені. Це складні речі. Система може наприклад порадити вам вживати інсулін, якщо у Вас діабет. Знову ж таки, речі, які мають мало цінності. Якщо мова йде про якусь складну хворобу, яку ми не можемо зрозуміти, найбільша складність в діагностиці, коли лікар не знає, що з людиною трапилося і як це лікувати. Дуже мала ймовірність, що алгоритм зможе це освоїти. Чому? Бо алгоритм добре працює на тих речах, які вони вже бачили, вони добре аналізують і узагальнюють. Наприклад, те що ми говорили про фільми: алгоритми добре працюють, бо багато людей дивиться одні й ті самі фільми, і можна шукати щось спільне. Та зазвичай в медицині дуже рідко є так, щоб люди хворіли абсолютно однаковими хворобами і мали схожі історії хвороби. Такі складні випадки, які складають найбільшу цінність, вони будуть унікальними. І зробити з цього висновок що і якклікувати, щоб лікування було ефективним, система не придумає нічого нового. Вона немає уявлень про медицину, про будову тіла. Вона має дані, ліки, хвороби, ними оперує і не може вийти подивитись на це трохи ширше. Це зазвичай працює в іншому напрямку, наприклад «80% пацієнтам виписували дані ліки і вони допомагала».

Ефективність ліків та лікарів. Це один з найбільш практичних моментів, коли це застосовується. В чому є проблема в медицині – ми користуємось слухами. Як Ви обираєте лікарів? Порадами і рекомендаціями. Тобто слухами. Щодо ліків, то проводиться статистика, хімічні дослідження, тому не часто ми використовуємо слухи при вборі ліків. Приклад: в одній лікарні після операцій виживає 80% пацієнтів, а у іншій – 60%, в яку ви підете? А якщо вам сказати, що лікарня з 80% - це районна лікарня, а з 60% - лікарня при медуніверситеті Нью-Йорку. Яку ви оберете? В лікарні з 60% роблять важкі операції, де смертність дуже висока. От що означає статистика і цифри, якщо з ними неакуратно поводитись. Інший приклад, який відсоток смертності у пацієнтів: ті, які самостійно прийшли до лікарні, чи ті, яких привезла швидка? - Ті що на швидкій. Тепер уявіть, ви написали алгоритм і припустимо людина зламала ногу. Бот в її телефоні каже: «Ні, не дзвони в швидку, бо якщо ти підеш на автобусі, то в тебе більші шанси вижити». В таких ситуаціях алгоритм буде робити безглузді речі. Він бачить дуже обмежений світ, який ми йому показали.

Повернемось про ефективність ліків. Цим займається наша компанія. Вийшли одні ліки, вийшли інші. Як порівняти їхню ефективність? Перед тим, як ліки вийшли на ринок, поводять багато клінічних досліджень. Як порівняти ефективність ліків, які вже достатній час на ринку і багато людей їх прийняли, сотні тисяч, і у нас є ця інформація? Найпростіше – взяти відсоток тих, хто приймав одні

ліки, які прийняли інші ліки і порахувати який від них ефект. Що в цьому підході неправильно? Зазвичай, коли існує 2 види ліків, ми не говоримо про назву брендів, ми говоримо про складні інгрідієнти цих ліків. Важливо, щоб вибірка була репрезинтативна. Є таке поняття як неупередженість: одні види ліків виписують в одній клініці, інші - в іншій. Або одні виписують багатшим пацієнтам, а тим, хто на державній страховці виписують дешевші ліки. У випадку з США можна отримати перевагу однієї раси в групах. І якщо є певні генетичні особливості хвороби і в її розвитку, то ви отримаєте зовсім різні дані і необ'єктивну оцінку дії ліків. Важливо, щоб дані були збалансовані. Це можна зробити так: забрати інформацію про расову приналежність і просити систему співставити яке з двох ліків випишуть цьому пацієнту. І ви заставляєте алгоритм прогнозувати ймовірність того, випишуть перші або другі ліки. Якщо ви отримаєте пацієнтів, які з ймовірністю 90%, що йому випишуть або перші ліки або другі, то таких пацієнтів не можна порівнювати, адже скоріш за все це дуже різні люди і є серйозні причини, чому їм виписують або одні або інші ліки. Ваша цільова група – це група біля 50%, яким би могли виписати або одне або інше, вони між собою подібні. Серед цієї групи є зміст порівнювати і можна уникнути загроз. І зазвичай дуже часто в таких ситуаціях ніякої різниці немає, вона в межах статистичної похибки.

Такі задачі data science в healthcare: порівняння ефективності ліків, ефективності лікування, можуть бути і порівняння ефективності клінік. До цифр потрібно підходити дуже обережно, адже цифри можуть помилятись. І найважливіше, data science – бізнес-орієнтована галузь. Фактично це є перетин математики, програмування і предметної галузі. В нашому випадку предметна галузь – це медицина. Ви повинні розуміти як вона працює, які є законодавчі обмеження, як працюють фармакологічні компанії, мати якісь базові уявлення. Можна навіть отримати дуже цікавий результат, який не можна буде застосувати і комерціалізувати. Важливим є той факт, що результати, які ви даєте можна використати. Тому потрібно розуміти як працює галузь з точки зору не чисто медичної, а з точки зору її фінансування, її доходів. Також важливим є навики візуалізації, представлення та пояснення,адже після розрахунків ви отримаєте різні графіки, гістограми і т.д. Data science часто дітися на 2 групи: тих хто виконує більш технічну роботу і тих хто представляє все клієнтам. Коли вам потрібно спілкуватись з клієнтом, то вам непотрібно використовувати страшних термінів, потрібно говорити простою мовою і при цьому залишатись переконливим. Візуалізація – це один з інструментів, адже всі люблять гарні графіки, але цього не є достатньо. Ви повинні вміти відповідати на запитання клієнтів так, щоб в них не закрадались сумніви, що ви некомпетентні, але разом з тим, щоб вони не відчували, що чогось не розуміють. Вміння представити результат, вміння його донести – одні з базових компетенцій. Після завершення навчання, коли вас будуть брати на роботу, у вас буде виникати нестача цих навиків. Тому, там, де у вас були можливості, я б на вашому місці, звертав на це увагу і їх освоював.