TEXTY.ORG.UA

Посібник по роботі з даними

Київ, 2015

Зміст

Культура роботи з даними

У 2010 році юрист Майк Флауерс очолив створений міським головою Нью-Йорка Майклом Блумбергом відділ з аналітики даних у мерії міста (Mayor's Office of Data Analytics).

Разом із командою молодих аналітиків Флауерс працював над збором, структуруванням й осмисленням масивів даних від міських служб, а також опікувався тим, аби управлінські рішення щодо розвитку міста базувались на аналізі цих даних.

Серед іншого команда працювала над механізмами запобігання пожеж, зниження рівня злочинності, підвищення ефективності надання медичних послуг, зміцнення захисту від стихійних лих, викриття махінацій із фінансами та нерухомістю, боротьби із продажем контрафактних цигарок та виявлення джерел забруднення.

Флауерс контролював процес розкриття своїх даних всіма міськими службами з метою забезпечення прозорості діяльності і побудови на їх основі сервісів для жителів Нью-Йорку. Він також започаткував NYC DataBridge — платформу для обміну аналітичних матеріалів між міськими службами, активістами та населенням.

Нью-Йоркський Mayor's Office of Data Analytics став взірцем для наслідування для багатьох міських та національних служб не лише у США, але й у Європі. Він може стати прикладом і для українського уряду, перед яким наразі стоїть задача не лише відкриття даних, але й ефективного їх використання під час ухвалення важливих для держави рішень.

Як дані можуть допомогти в урядуванні?

На всіх рівнях урядування — державному, обласному, районному, місцевому — використання даних покликане забезпечити прозорість діяльності органів влади, зменшити рівень корупції, оптимізувати робочі процеси, виявити проблеми та допомогти у пошуку їх вирішень. Культивація щоденної роботи із даними в органах влади та урядових установах усіх рівнів здатна поліпшити економічну ситуацію, покращити якість медичного обслуговування, охорони правопорядку, освіти, містопланування.

При цьому важливо забезпечити обмін даними та співпрацю між міністерствами та урядовими установами. Так, скажімо, Укравтодор може скористатись даними ДАЇ про найбільш аварійні дороги та вулиці, комунальні служби — даними МВС про вулиці, на яких найчастіше відбуваються напади чи пограбування, Міністерство екології — даними медичних установ про захворюваність у різних регіонах, Мінагропром та комунальні служби — даними Гідрометцентру, Міносвіти — даними Державної служби зайнятості і т. д.

Ось лише кілька прикладів можливого використання даних у різних сферах.

Охорона правопорядку

Використання оперативних звітів про місця та час скоєння злочинів дозволяє правоохоронцям швидко реагувати на ситуацію і приділяти увагу найбільш проблемним районам міста у найбільш небезпечний час. Саме так працюють поліцейські у багатьох містах США. Найяскравішим прикладом використання даних у правоохоронній діяльності є Департамент поліції міста Мемфіс (Теннесі), де програма Blue CRUSH (Crime Reduction Utilizing Statistical History) продукує дані про зони потенційної загрози — в межах кількох кварталів та кількох годин конкретного дня тижня. Ці дані дозволяють правоохоронцям якнайкраще розподіляти обмежені ресурси. За деякими підрахунками, від початку використання Blue CRUSH у 2006 році кількість майнових та насильницьких правопорушень у місті скоротилась на чверть.

Медицина, охорона здоров'я

У сфері охорони здоров'я аналіз даних дозволяє прогнозувати сплески захворюваності, порівнювати ефективність різних способів лікування, виявляти групи ризику та проблемні місця (лікарні, поліклініки, станції швидкої допомоги) у системі, підвищувати ефективність роботи та зменшувати кількість помилок медичного персоналу. Принаймні саме в цих напрямках працює безліч державних установ та приватних компаній у США, де аналіз даних є одним з пріоритетів охорони здоров'я на національному рівні. Віднедавна аналогічні тенденції спостерігаються і в Європі.

Освіта

У сфері освіти аналіз процесу та результатів навчання допомагає оцінити ефективність програм та створити оптимальні умови для засвоєння інформації учнями. Крім того, аналіз результатів стандартизованих тестів дозволяє оцінити якість роботи вчителів та виявити випадки завищення ними оцінок. Робота з даними за зазначеними напрямками у США ведеться на національному рівні. Подібний підхід віднедавна також промотує для всіх без винятку країн Світовий банк.

Сільське господарство

Найновіші технології використання даних у сільському господарстві стосуються підвищення ефективності використання землі, збільшення врожайності, визначення найкращого часу для засівання на основі історичних та прогнозованих погодних даних, підбору насіння, комбінування добрив та налаштування режиму поливу. В цьому напрямку у світі працює одразу кілька компаній та стартапів.

Транспорт

Вивчення патернів розселення і переміщення жителів міста (району, області, країни) та щоденних пасажиропотоків на громадському транспорті дозволяє оптимізувати транспортну модель, скасувати неефективні маршрути та запровадити на їх заміну нові. Цей підхід активно практикується під час містопланування у багатьох країнах світу.

Енергетика

В енергетиці аналіз даних застосовується для дослідження і прогнозування споживання енергії домогосподарствами та підприємствами, промоції режиму економії, діагностування енергообʼєктів та виявлення випадків нелегального забору електроенергії. Цей підхід також уже є більш менш усталеним та активно застосовується у багатьох країнах світу.

Що включає в себе культура даних?

Організація збору даних

Від того, як організований процес збору даних, залежить якість і глибина аналізу. Допущені під час збору даних помилки унеможливлюють роботу з даними і зводять нанівець усі зусилля.

Збір даних має бути регулярним. У даних не має бути пропусків, кожна організація мусить дотримуватись розкладу збору та публікації наборів даних.

Дані слід збирати на найнижчому рівні агрегації — записи мають бути якомога більш деталізованими.

Форма збору даних має бути уніфікована. Відповідальні за збір даних на підприємствах і в організаціях мусять заповнювати однакові форми. Самі форми повинні містити чіткі інструкції із заповнення і не допускати можливості подвійного прочитання.

Інформація має надходити у структурованих форматах. Для вжитку рекомендуються формати CSV, JSON, XML, RDF. (Детальні рекомендації Кабінету міністрів щодо публікації даних прописані у постанові “Про затвердження Положення про набори даних, які підлягають оприлюдненню у формі відкритих даних).

Побудова інфраструктури збереження даних

Від інформації мало користі, якщо вона записана на аркуші та лежить на полиці у стосі паперів в офісі якоїсь філії державної компанії у забутому богом районному центрі. Інформація має збиратись в електронному вигляді і надходити з регіонів до центру, а згодом публікуватись на сторінках міністерств та порталі відкритих даних у машиночитаному вигляді. Побудова інфраструктури збереження даних, забезпечення серверних потужностей та програмних засобів — є ключовими задачами.

Портали відкритих даних

У 2009 році розпочав роботу портал відкритих даних США data.gov, який збирає в одному місці якомога більше даних від всіх міністерств, урядових агенцій та організацій у машиночитаному (придатному для автоматизованої обробки) форматі. Цей портал має забезпечити прозорість діяльності уряду та спонукати розробників до створення заснованих на даних сервісів. Відтоді подібні портали з'явились у десятках інших країн по всьому світу.

В Україні такий портал засновується постановою Кабміну «Про затвердження Положення про набори даних, які підлягають оприлюдненню у формі відкритих даних». Цією ж постановою затверджується регламент публікації даних міністерствами та урядовими установами, а також формати публікації даних.

Аналіз даних

Навіть за умови добре організованого збору та наявності інфраструктури збереження даних, без аналізу, без осмислення від них все ще мало користі. В міністерствах та інших урядових установах мають функціонувати аналітичні відділи, до обов'язків котрих входитиме оперативний аналіз даних та подання заснованих на цьому аналізі пропозицій і рекомендацій до керівного складу. Між такими відділами та керівним складом має бути налагоджена тісна співпраця: аналітики мусять розуміти потреби керівників в аналітичних продуктах та оперативно на них реагувати; керівники, у свою чергу, мають усвідомлювати потенціал та можливості своїх аналітиків та звертатись до них якомога частіше.

Імплементація заснованих на даних рішень

Всі попередні етапи — збір, побудова інфраструктури збереження та аналіз даних — не матимуть сенсу, якщо вищі чиновники ігноруватимуть або ж хибно інтерпретуватимуть результати цього аналізу. Рішення, що стосуються стратегій розвитку держави, економіки, охорони здоров'я — загалом будь-якої сфери — мають ґрунтуватись на аналізі даних. Більше того — у громадськості має бути доступ

до даних, на яких засновуються ті чи інші рішення, а також до розрахунків і результатів аналізу. Лише за таких умов діяльність уряду буде прозорою, а рішення — піддаватимуться перевірці і верифікації.

> ***Chief Data Officer***
>
> *У багатьох американських урядових агенціях зараз активно впроваджують посади chief data officer — себто директора по роботі з даними. Людина, яка обіймає цю посаду, відповідає за пошук найкращого застосування даним своєї агенції, культивує роботу з даними як із стратегічним активом, промотує найцікавіші набори даних на конференціях і хакатонах, співпрацює із розробниками та журналістами.*

Підготовка даних

Будь-хто з тих, кому доводилось працювати із даними, знає, що на підготовку даних для аналізу часто доводиться витрачати значно більше часу, ніж на сам аналіз. Аби уможливити швидкий і ефективний аналіз даних, а також побудову на їх основі сервісів, набори даних мають бути очищені і добре структуровані.

Що таке чисті дані?

«Чистими» можна вважати такі дані, які:

- не містять помилок та одруківок;
- у межах одного типу чи класу мають уніфікований формат: дати зазначаються в єдиному форматі (скажімо 2015—01—01), числа мають однаковий роздільник для десяткових (або кома, або крапка), в одному наборі даних не має зустрічатись два різних роздільника десяткових;
- назви всюди пишуться однаково (однаковий варіант напису однієї і тієї ж адреси, номера телефону, ПІБ людинитощо);
- в одній ячейці міститься лише один запис і кожен запис займає лише одну ячейку (тобто, в наборі немає об'єднаних ячейок).

Приклад «брудних» даних

Дата	Місто	Показник
2015—08—01	Київ	1 678 910
01—07—2015	Киев	1,567,890
1 черв 2015	Кийів	1 млн 456 тис 789

2015.06.01	Львів	1.345.678
2015/07/01	Львов	1 234 567

Приклад очищених даних

Дата	Місто	Показник
2015—08—01	Київ	1678910
2015—07—01	Київ	1567890
2015—06—01	Київ	1456789
2015—06—01	Львів	1345678
2015—07—01	Львів	1234567

З брудними даними неможливо працювати: в них присутній часовий вимір, однак ми не зможемо вибудувати часову лінію, оскільки всі дати мають різний формат; тут присутній географічний вимір, однак ми не зможемо агрегувати дані за змінною «Місто», оскільки кожен раз назва міста пишеться інакше; тут також присутній числовий показник, однак ми не зможемо здійснювати над ним ніяких арифметичних операцій, оскільки всі числа записані у різних форматах.

Чисті дані ми можемо фільтрувати, сортувати, здійснювати над ними арифметичні операції, візуалізувати та будувати на їх основі сервіси.

Що таке добре структуровані дані?

Добре структурованими даними вважаються такі дані, які можна швидко трансформувати, аналізувати і візуалізувати. Їх можна впізнати за двома простими ознаками:

- всі змінні містяться у колонках;
- всі спостереження містяться у рядках.

Проблеми із форматуванням і структуруванням найчастіше виникають через нерозуміння того, що у наборі даних є змінною. Розглянемо це на прикладі.

Погано структуровані дані

Доходи населення за регіонами України у 2012–2014 роках, млн грн

Область	2012	2013	2014
Вінницька	35441	37323	39184
Волинська	19546	20609	21971
Дніпропетровська	95349	99995	109545
Донецька	128767	135362	114135

...

Добре структуровані дані

Доходи населення за регіонами України у 2012—2014 роках

Область	Рік	Доходи населення, млн грн
Вінницька	2012	35441
Волинська	2012	19546
Дніпропетровська	2012	95349
Донецька	2012	128767
Вінницька	2013	37323
Волинська	2013	20609
Дніпропетровська	2013	99995

Донецька	2013	135362
Вінницька	2014	39184
Волинська	2014	21971
Дніпропетровська	2014	109545
Донецька	2014	114135

…

У першій таблиці в якості змінних подаються «Область», «2012», «2013» і «2014». Однак насправді, змінними в цьому наборі є «Область», «Рік» і «Дохід населення», як показано у другій таблиці. Саме другий варіант форматування є правильним.

Найліпше, звісно, не допускати помилок чи одруківок та дотримуватись певних стандартів форматування і структурування ще на етапі збору первинних даних. Втім, помилки — бодай незначні — будуть завжди. Під час очищення даних важливо, по-перше, зберігати оригінальний файл та працювати з копією, по-друге, записувати всі здійснені над даними операції в окремому файлі. Для очищення, форматування і структурування даних може стати в пригоді програма Open Refine.

> ***Open Refine***
>
> *Програма для обробки даних. Серед іншого дозволяє знаходити і виправляти помилки у наборі даних; виявляти різні варіанти написання однієї назви (кластеризувати текстові дані) чи дати та зводити їх до одного вигляду; транспонувати (повертати дані із рядків у стовпці і навпаки) таблиці, розділяти дані з однієї колонки на кілька, здійснювати прості арифметичні операції; об'єднувати кілька наборів даних.*
>
> *Програма безкоштовна і доступна для всіх основних операційних систем (Windows, Mac OS, Linux).*
>
> *openrefine.org*

Аналіз даних

Розпочинаючи роботу із певним набором даних, слід відповісти на кілька питань.

- Який розмір набору з даними? Йдеться не лише про розмір файлу, але й про те, яка ширина (кількість колонок/змінних) та довжина (кількість рядків/спостережень) набору даних? Від кількості змінних залежить й кількість варіантів аналізу.
- Які типи даних у наборі? Від того, які типи даних присутні у наборі, залежатиме, в яких вимірах можна здійснювати аналіз. Також важливо пересвідчитись, що всі типи даних інтерпретуються належним чином — себто, скажімо, дати сприймаються як дати, а не як текст чи число.
- Чи є в наявності метадані? Себто, чи існує детальний опис для всіх стовпчиків у наборі даних? Чи відомо, як і коли відбувався збір даних?
- Чи нема у даних пропусків і помилок? За наявності великої кількості пропусків певні види аналізу чи візуалізації втрачають сенс. Помилки та одруківки можуть ускладнити фільтрацію та агрегацію даних.

Базові операції із даними в Excel

Здебільшого робота з даними зводиться до кількох базових операцій — фільтрації, сортування, об'єднання даних, обчислення нових змінних, групування та агрегації, розбиття та об'єднання комірок.

Фільтрація

Добре структуровані дані легко фільтрувати, тобто вилучати з великого набору даних ту частину, яка представляє інтерес для дослідження. Фільтрувати дані можна, наприклад, за категоріальними показниками (країною, регіоном, компанією, особою і т.д.), часовими

(роком, місяцем, часовим інтервалом), а також кількісними показниками, які відповідають певним логічним умовам (більше, менше, дорівнюють чи не дорівнюють *n*).

Сортування

Сортувати — впорядковувати дані або від найбільшого до найменшого, або ж від найменшого до найбільшого — можна не лише за одним показником, а й за кількома. В такому випадку, коли в сортованому стовпчику буде два чи більше тотожних значення, їхній порядок у наборі визначатиметься додатковим критерієм.

Об'єднання таблиць (Vlookup)

Об'єднання двох наборів даних здійснюється на основі спільного стовпчика в обох наборах — критерію об'єднання. Залежно від виду об'єднання до нової таблиці можуть потрапляти:

- всі дані з обох таблиць;
- всі дані з першої таблиці та лише ті дані з другої, які відповідають критерію об'єднання;
- всі дані з другої таблиці та лише ті дані з першої, які відповідають критерію об'єднання;
- лише ті дані з обох таблиць, які відповідають критерію об'єднання.

Обчислення нових змінних

Нові змінні можна створювати як за допомогою базових арифметичних операцій — додавання, віднімання, множення та ділення — та їхніх комбінацій, так і за допомогою умовних виразів (IF ELSE). Наприклад, можна створити категоріальну змінну шляхом віднесення даних до різних груп, залежно від виконання певної логічної умови (наприклад, якщо X > Y, то Z = «Перша група», інакше Z = «Друга група»).

Групування та агрегація (Pivot Table, Зведена таблиця)

Групуючи дані за певними категоріальними змінними (наприклад, за статтею), ми можемо обчислювати базові статистики (мінімум, максимум, середнє, медіану, квартилі, моду і т.д.) для утворених груп.

Розбиття та об'єднання ячейок

Розбивати одну ячейку на кілька або ж з'єднувати кілька ячейок в одну можна на основі певного роздільника — пробілу, коми, тире або ж навіть виразу.

Візуалізація даних

Візуалізація даних стає в пригоді як під час аналізу даних, так і під час презентації результатів цього аналізу. Від того, наскільки добре представлені результати аналізу, залежить, чи стануть вони поштовхом для ухвалення та імплементації певних рішень. Саме тому графіка, яка використовується для презентації результатів аналізу даних, має бути простою і ефективною, мусить чітко формулювати і доносити головні тези.

Гарна візуалізація уникає викривлення даних, вміщує множину чисел у малому просторі, логічно впорядковує великі обсяги інформації, уможливлює порівняння різних фрагментів даних та представляє дані на кількох рівнях масштабування — показує загальну картину та дозволяє розгледіти деталі.

Функції візуалізації даних

Найчастіше графіки та діаграми виконують наступні функції:

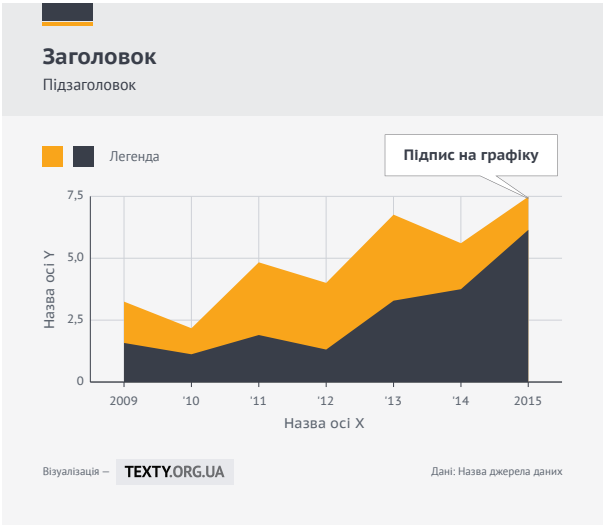
- порівняння величин,
- виявлення розподілу величин,
- виявлення зв'язків та залежностей,
- демонстрація зміни показників у часі,
- співвідношення частини і цілого.

Вибір типу візуалізації, отже, не є довільним, а залежить від поставленої задачі.

Функція	Тип графіку
Порівняння величин	горизонтальна стовпчикова діаграма
Розподіл величин	гістограма «ящик з вусами»
Частина і ціле	секторна діаграма
Зміна у часі	лінійний графік вертикальна стовпчикова діаграма
Зв'язок між величинами	діаграма розсіяння

Структура графіку

Практично кожен графік, незалежно від типу, мусить мати заголовок, підзаголовок, підписи осей, легенду та джерело даних. У деяких випадках, однак, легенда може бути замінена прямими підписами, а позначення осей – задані у підзаголовку.



Заголовок — коротко формулює основну думку графіку.

Підзаголовок — більш детально пояснює, що зображено на графіку, задає контекст для розуміння та інтерпретації.

Назви осей — пояснюють, які саме дані (і в яких розмірностях) представлені на графіку.

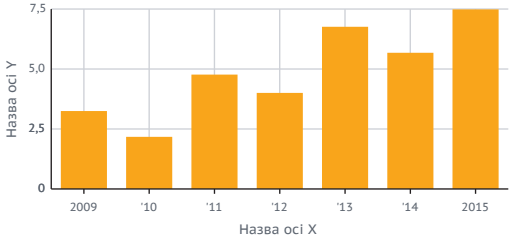
Легенда — роз'яснює використані у графіку умовні позначення та способи кодування інформації (розміри, кольори, форми).

Підписи на графіку — звертають увагу читача на найважливіші дані.

Джерело даних — вказує, звідки походять дані, за якими побудований графік, аби зацікавлені читачі завжди могли перевірити їх.

Види візуалізацій

Стовпчикова діаграма



Структура. Вертикальна стовпчикова діаграма найчастіше по осі X має часові інтервали (дні, тижні, місяці, роки), а по осі Y — певний числовий показник, який змінюється впродовж часу. Горизонтальна стовпчикова діаграма може мати по осі X різні категоріальні показники (список людей, компаній, країн і т.д.), а по осі Y — певний числовий показник, за яким ці категорії ранжовані.

Функції. Демонстрація зміни певного показника в часі або ж порівняння кількох категоріальних показників.

Особливості використання. Відстань між стовпчиками не має бути більшою за третину товщини самого стовпчику. Вісь Y обов'язково має починатись від нуля.

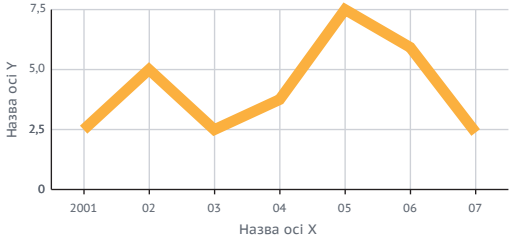
Гістограма

Структура. По осі X — інтервали з набору даних, по осі Y — кількість елементів набору даних, що потрапляють до того чи іншого інтервалу.

Функції. Виявлення розподілу величин.

Особливості використання. Форма гістограми залежить від ширини інтервалу групування по осі X. Змінюючи цей інтервал, можна отримати геть різні діаграми.

Лінійна діаграма

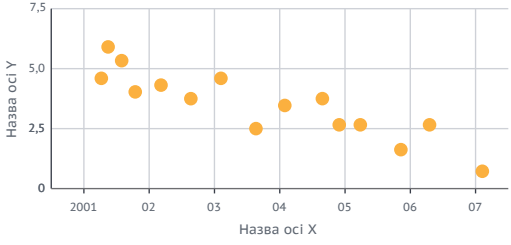


Структура. По осі X — часові інтервали (дні, тижні, місяці, роки), по осі Y — певний числовий показник, який змінюється у часі.

Функції. Демонстрація зміни певного показника у часі.

Особливості використання. Вісь Y не обов'язково починати від нуля. У випадку використання на графіку більш ніж чотирьох ліній, є сенс замислитись про те, аби рознести їх по окремих графіках.

Діаграма розсіяння

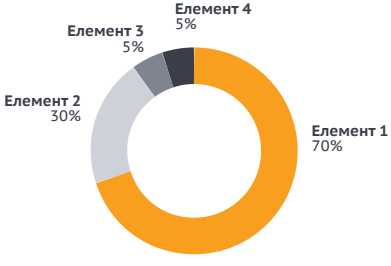


Структура. На координатній площині розкидані точки, положення котрих задається двома координатами — X та Y. На осях X та Y — показники, зв'язок між котрими потрібно дослідити.

Функції. Виявлення зв'язку та залежності між двома показниками.

Особливості використання. На діаграмах розсіяння доволі часто окрім точок використовують також лінію тренду, яка пролягає максимально близько до всіх точок на площині і демонструє загальну тенденцію.

Секторна діаграма



Структура. Коло, поділене на сектори. Все коло представляє собою 100%, розмір сектору — частку від цих 100%.

Функції. Порівняння частини та цілого.

Особливості використання. Строго кажучи, цей вид візуалізації не слід використовувати в принципі. Для порівняння частини та цілого краще підходить, скажімо, стовпчикова діаграма із накопиченням. Однак, якщо ви все ж наважились використати секторну діаграму, подбайте про те, аби на ній було не більше трьох секторів.

> ***The Data Visualisation Catalogue***
>
> *Каталог графіків та діаграм, із детальним поясненням структури та сфери застосування кожного виду графіку.*
>
> *http://www.datavizcatalogue.com/*

Епілог

Міністерствам та урядовим організаціям годі скаржитись на нестачу даних у тій чи іншій сфері. Щодня підпорядковані їм підприємства та установи продукують силу силенну інформації. Заповнюють форми, пишуть аналітичні записки або ж здають звіти.

Інша справа, що доволі часто ці форми, звіти та записки існують лише на папері, а отже доступ до них закритий не лише для широкого загалу, а й для урядових же посадовців. Та навіть ті дані, які існують в електронному вигляді, здебільшого лежать без діла.

Задекларований урядом перехід до електронного документообігу і заплановане відкриття даних міністерствами та підпорядкованими їм організаціями і підприємствами – безперечно поліпшить ситуацію. Отримавши доступ до масивів даних, активісти і програмісти зможуть продукувати численні продукти і сервіси, для розробки котрих у держави нема часу, ресурсів і компетенції.

Однак не менш важливо, аби дані використовували у самому уряді. Щоб прийняття рішень диктувалось не політичною кон'юнктурою чи бізнес інтересами, а базувалось на аналізі даних.

Для цього слід плекати культуру роботи з даними в міністерствах та відомствах. Аналітичні відділи у міністерствах мусять виробляти цікавий і актуальний продукт, урядовці, у свою чергу, мають цим продуктом активно користуватись. Між аналітиками та чиновниками має бути налагоджена тісна співпраця.

Для держави відкриття даних і засноване на даних врядування може забезпечувати прозорість та підзвітність влади перед населенням, економію бюджетних коштів, збільшення доходів та оптимізацію процесів. Для бізнесу – появу нових, заснованих на даних, стартапів, створення нових робочих місць. Для населення – поліпшення якості інфраструктури, продуктів і послуг, рівня життя загалом.

Підготовано TEXTY.org.ua за сприяння
Міжнародного фонду Відродження

Посібник видано за підтримки «Ініціативи з розвитку аналітичних центрів в Україні», яку виконує Міжнародний фонд «Відродження» (МФВ) у партнерстві з Фондом розвитку аналітичних центрів (TTF) за фінансової підтримки Посольства Швеції в Україні (SIDA).

Думки та позиції викладені у цій публікації є позицією автора та не обов'язково відображають позицію уряду Швеції.

Підписано до друку 28.09.2015.
Формат сторінки: А5. Кількість сорінок: 24.
Папір офсетний. Друк цифровий.

СПД Марченко, ДК №4841 від 29.01.2015
04080, Київ, вул. Нижньоюрківська, 3 к. 25.
Телефон 044 467–5322

Брошуру ви можете завантажити тут:
texty.org.ua/pdf/data2015.pdf