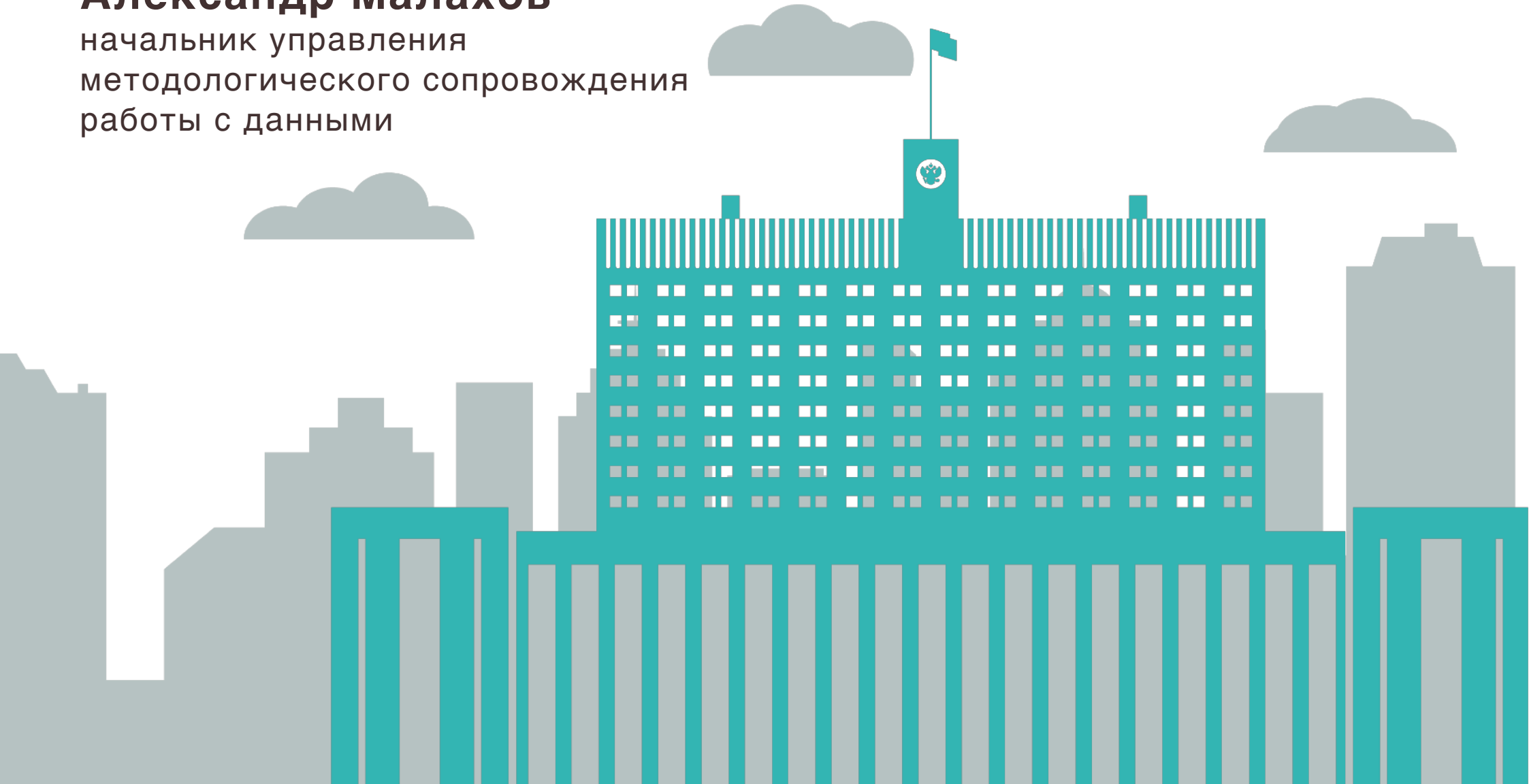
Проблемы сбора и анализа больших данных

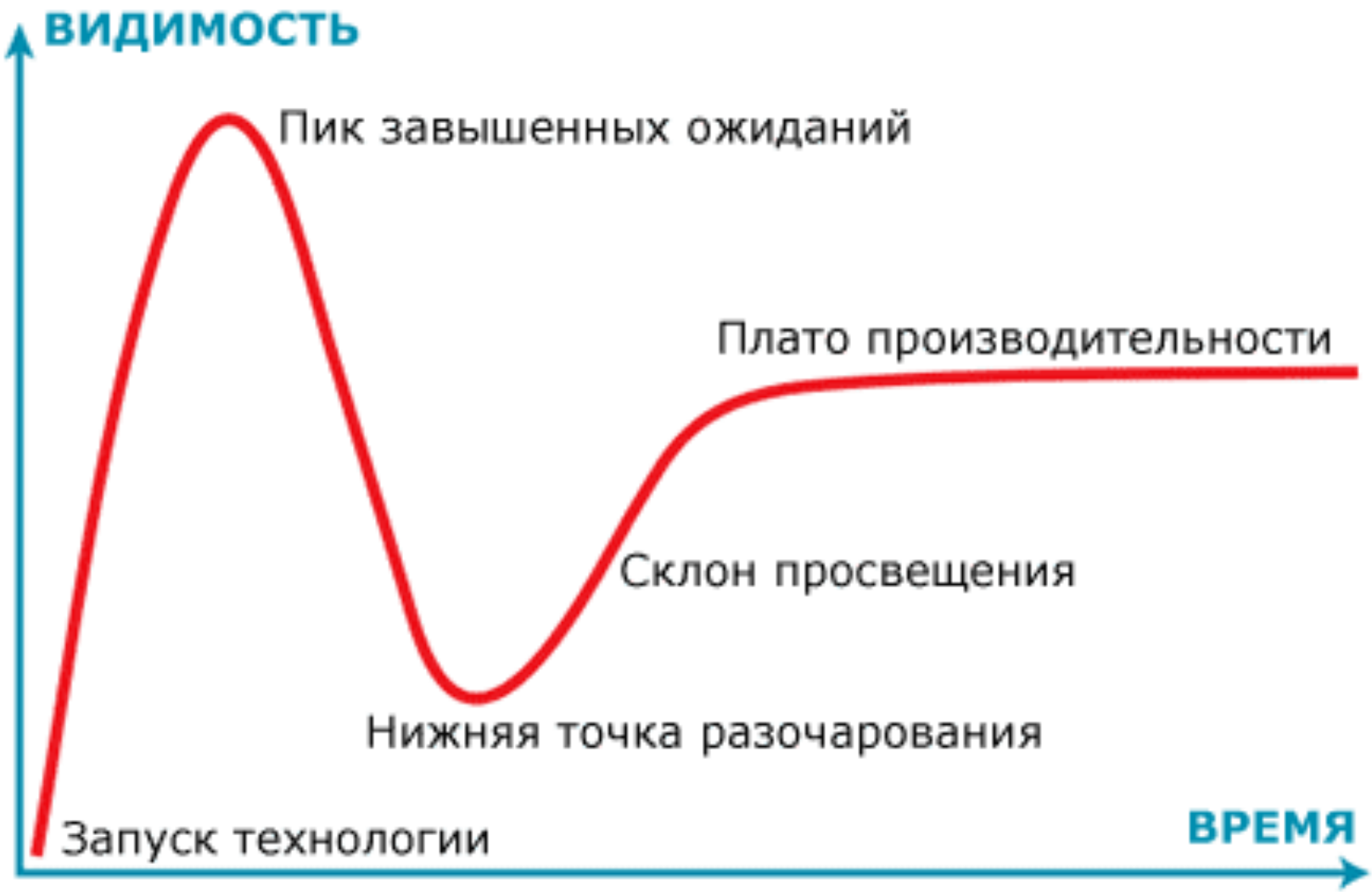
АНАЛИТИЧЕСКИЙ ЦЕНТР
ПРИ ПРАВИТЕЛЬСТВЕ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Александр Малахов
начальник управления
методологического сопровождения
работы с данными



ВТОРОЙ ПИК ХАЙПА БОЛЬШИХ ДАННЫХ

Кривая Gartner / кривая «Ажиотажа»



ЧТО ТАКОЕ БОЛЬШИЕ ДАННЫЕ

Способ обработки данных основанный на **поиске и построении верной корреляционной модели** данных, при условии, что данные подвергаемые такой обработки **не являются полно-нормализованными и полно-связными** При этом достоверность связности находится в каком-то узком коридоре от 70%-до 90%

100% коррелирующие данные обрабатываются традиционными методами BI и/или реляционными методами.

ПРОБЛЕМЫ РАБОТЫ С БД

Смешение понятий и технологий

при отнесении деятельности или проекта к анализу БД не следует смешивать Технологии БД (типа Hoodup / MapReduce / Hive), технологии машинного обучения и решаемые задачи

Технико-экономическая эффективность или Ошибка выбора технологии

обычное построение задачи: «Давайте всё соберем в data lake и проанализируем – получим новое знание и продадим его». Информационные технологии решают задачи Автоматизации технологических процессов – сначала должна быть поставлена задача – затем выбраны способы решения.

Ложная корреляция

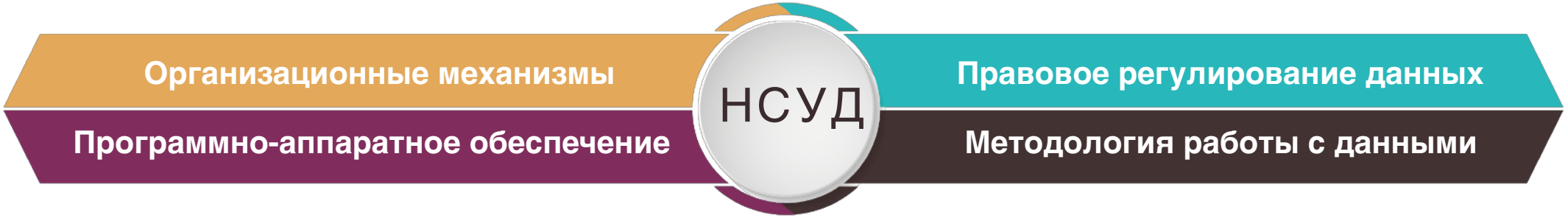
возникает, когда две независимых друг от друга величины меняются синхронно или почти синхронно. Попытка вычисление коэффициента корреляции даст очень высокое и часто очень достоверное значение. Это может подтолкнуть к ложным выводам о наличии причинно-следственной связи между явлениями.

Проблемы качества данных

в России отсутствуют практики накопления больших данных при этом качество данных оставляет желать лучшего из-за наличия искажений (выбросов) и недостаточной глубины. Таким образом, требуется значительно расширять наборы данных для анализа, но для этого нет возможности, т.к. в связи с защитой персональных данных в нашей стране практически отсутствует рынок купли/продажи информации в виде бирж данных (Data Exchange).

ГОСУДАРСТВЕННЫЕ ДАННЫЕ

В части государственных данных создается национальная система управления данными.



Основными продуктами НСУД востребованным при работе с большими данными являются:

- федеративная модель данных, которая в свою очередь представляет собой объединение моделей данных органов власти, моделей данных предметных областей и **корреляционных моделей высокого уровня достоверности**
- каталог моделей данных
- каталог источников данных (Реестры видов данных / метаданных)