УВИДЕТЬ НЕОПРЕДЕЛЕННОСТЬ

КАК С ПОМОЩЬЮ МЕТОДОВ ВИЗУАЛИЗАЦИИ ДАННЫХ ИЗУЧАТЬ НЕОПРЕДЕЛЕННЫЕ ПРОЦЕССЫ

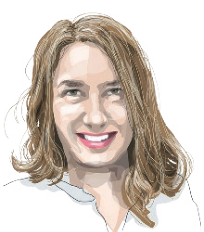
Джессика Халлман

Когда синоптики предупреждают нас о приближении урагана, они обычно показывают на карте так называемый конус неопределенности. Район формирования урагана отмечен точкой. От нее отходят две линии, очерчивая границы области, по которой разрушительная стихия должна промчаться в ближайшие дни. Считается, что наиболее вероятная траектория урагана проходит вдоль центральной линии, разделяющей конус на две половины; чем ближе к границам конуса, тем меньше вероятность появления урагана. Но проблема в том, что многие люди ошибочно думают, что размеры этого конуса совпадают с размерами урагана.

Чтобы избавиться от этого заблуждения, ученые предлагают вместо конуса показывать сразу несколько возможных траекторий движения урагана. Но и этот метод, как оказалось, не всегда правильно понимается. Многие почему-то думают, что вероятность самых больших разрушений увеличивается там, где траектории урагана пересекают сушу, а между траекториями вероятность разрушений якобы понижается.

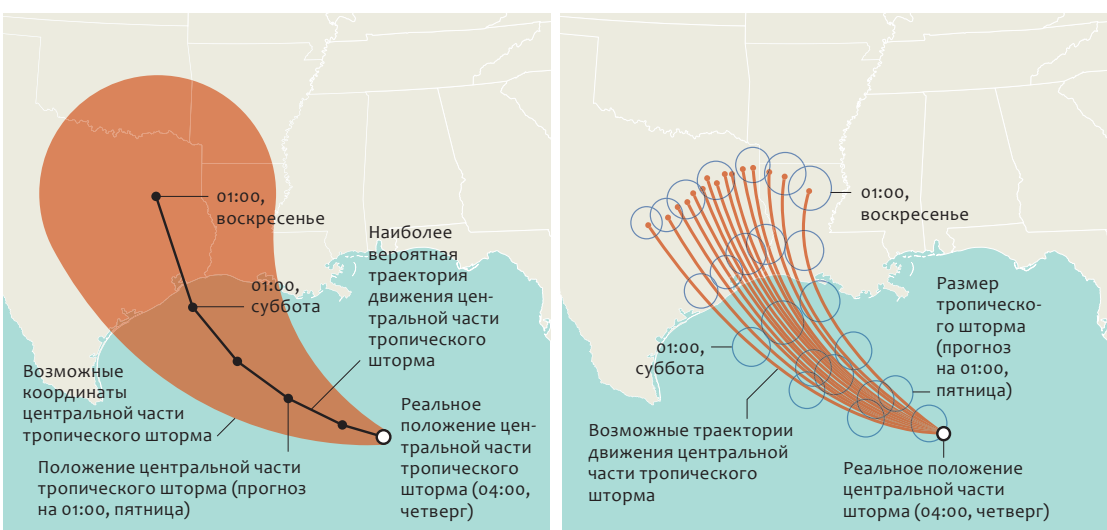
Данные, на основе которых ученые и различные организации принимают решения, тоже характеризуются неопределенностью. Методы визуализации данных позволяют в буквальном смысле слова «увидеть» неопределенность и выстраивать, осознанно или неосознанно, гипотезы о вероятности всевозможных исходов событий. Однако, как показывают многочисленные исследования особенностей человеческого мышления, если человека просят высказать вероятностное суждение, то он зачастую игнорирует фактор неопределенности. По мере развития информационного общества специалисты в области компьютерной графики пытаются понять, как лучше всего показать неопределенность в прямом смысле этого слова. В этой статье мы расскажем о методах визуализации, позволяющих отобразить процессы, в основе которых лежит неопределенность. Мы будем продвигаться от менее эффективного метода к более эффективному. Давайте подробнее рассмотрим, в каких случаях используются эти методы. Они обязательно помогут нам корректнее анализировать данные, характеризующиеся неопределенностью.

Перевод: И.В. Ногаев

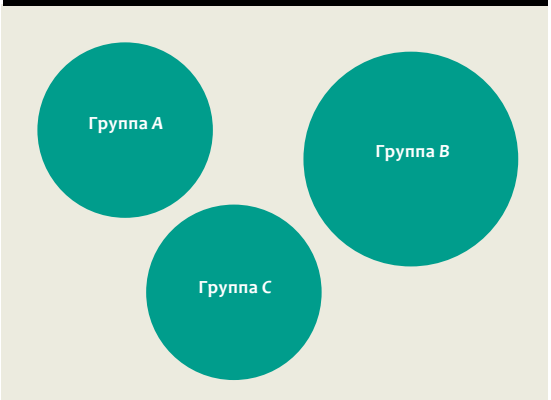


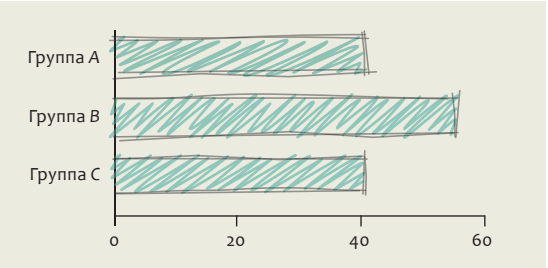
ОБ АВТОРЕ

Джессика Халлман (Jessica Hullman) — профессор информатики и журналистики в Северо-Западном университете (США). Группа ученых под ее руководством разрабатывает и тестирует методы визуализации данных и взаимодействия с данными, которые дают возможность лучше понять суть неопределенности.



«Конусом неопределенности» (*слева*) называется область, в которую может переместиться ураган; она определяется на основе множества прогнозов. При альтернативном подходе (*справа*) показывают каждую из траекторий указанного множества. Оба метода позволяют судить о риске появления урагана в данной местности. У каждого метода есть свои плюсы и минусы; метод справа еще раз показал, насколько вообще сложно предсказывать траектории урагана.





Без количественных показателей

Наименее эффективный способ представления неопределенности — это вообще ее не показывать. Иногда специалисты в области компьютерной графики пытаются компенсировать отсутствие данных с помощью этого метода — а для него неточность допустима. Здесь вместо количественных значений приводятся визуальные объекты, не обладающие какими-то конкретными параметрами: например, изображается не конкретная точка в декартовых координатах, а какой-нибудь круг в пространстве (*вверху*). При таком подходе интерпретация данных чревата ошибками. Кроме того, с помощью компьютерных средств можно создавать схематичные изображения от руки — это второй метод (*внизу*). Оба метода неточны.

Плюсы

- Если визуализированные образы сложно представить в количественном виде или они неточны, то пользователь, понимая это, станет с большей осторожностью делать на их основе выводы и принимать решения.

Минусы

- Пользователь может забыть, что данный метод визуализации априори неточен, отсюда — наличие грубых ошибок в выводах.
- И даже если пользователю известно, что данный метод визуализации по определению неточен, ему сложно сделать правильный вывод о вероятности изображенного события, если в этом возникнет потребность.

Интервалы

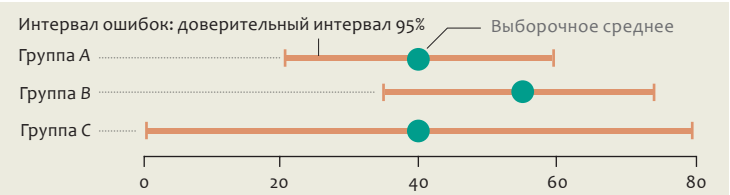
Использование интервалов — это, наверное, наиболее распространенный метод представления данных, которые характеризуются неопределенностью. Интервалы ошибок (*вверху*) и доверительные границы (*внизу*) широко известны. Но, несмотря на кажущиеся точность и понятность, их зачастую неправильно интерпретируют. Исследования показали, что эти методы зачастую неправильно понимают даже ученые.

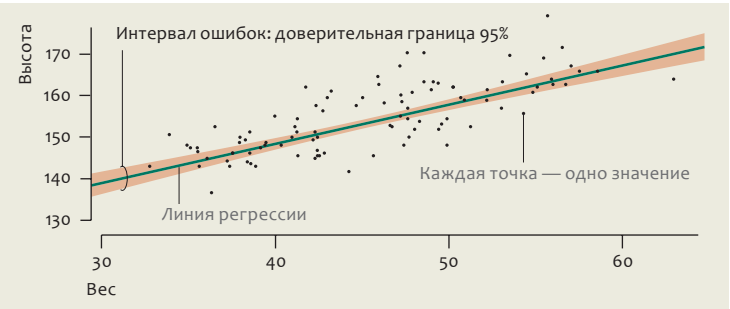
Плюсы

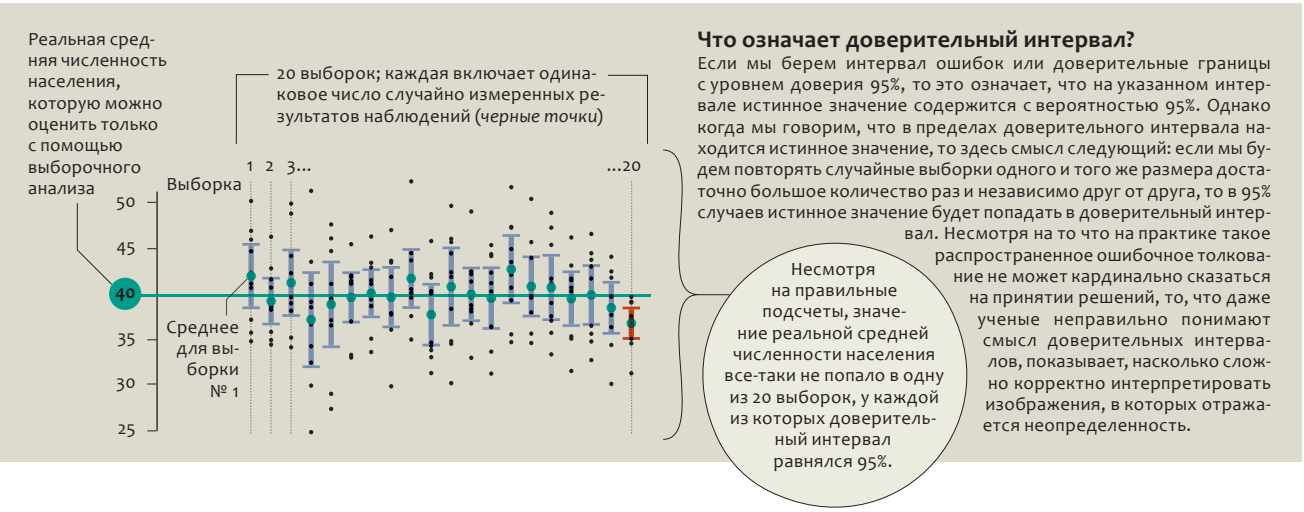
- Широко признан в качестве метода графического представления данных вероятностной природы.
- Простой формат представления вероятностей различных величин.
- Выбор интервала зависит от типа задач, заданных на одном и том же наборе данных. Например, если мы говорим о диапазоне значений в популяции, то интервалы означают среднеквадратическое отклонение; если мы говорим о диапазоне значений статистической величины (например, о среднем значении), то интервалы подразумевают стандартную ошибку.

Минусы

- Неясность представления: интервалы могут соответствовать среднеквадратическому отклонению, стандартной ошибке и т.д. В каждом случае свой смысл.
- Пользователи могут допускать так называемые детерминированные ошибки толкования, то есть интерпретировать значения на концах интервала ошибок не как величину, обозначающую неопределенность, а как максимальное и минимальное значения результатов измерения.
- При использовании интервала ошибок, особенно на гистограммах, можно неверно понять смысл данных, заключенных в пределах интервала. Например, величину, лежащую справа от среднего значения на рисунке (*ниже*), пользователь может неверно интерпретировать как наиболее вероятную, а слева — как наименее вероятную.
- Пользователь может запросто пренебречь областями неопределенности и вместо этого сфокусировать свое внимание только на среднем значении, что приведет к неверным выводам.







Что означает доверительный интервал?

Если мы берем интервал ошибок или доверительные границы с уровнем доверия 95%, то это означает, что на указанном интервале истинное значение содержится с вероятностью 95%. Однако когда мы говорим, что в пределах доверительного интервала находится истинное значение, то здесь смысл следующий: если мы будем повторять случайные выборки одного и того же размера достаточно большое количество раз и независимо друг от друга, то в 95% случаев истинное значение будет попадать в доверительный интервал. Несмотря на то что на практике такое распространенное ошибочное толкование не может кардинально сказаться на принятии решений, то, что даже ученые неправильно понимают смысл доверительных интервалов, показывает, насколько сложно корректно интерпретировать изображения, в которых отражается неопределенность.

Maps by Tiffany Farrant-Gonzalez; Graphics by Jessica Hullman and Jen Christiansen

Диаграммы плотности распределения вероятности

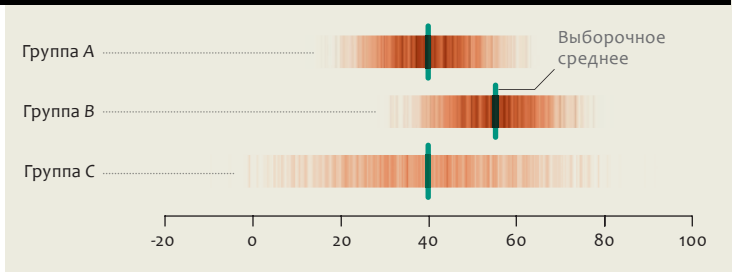
Вероятностные характеристики можно передать с помощью цвета, например показать в виде градиентной заливки на графике (*вверху*): здесь мы видим переход от темного цвета в центре (высокая вероятность) к более светлому по краям (малая вероятность). На диаграмме типа «скрипка» (*внизу*) выпуклые области означают более высокие значения вероятности. На таких диаграммах вероятность отображается более детально, чем с помощью интервальных методов (интервалов ошибок и доверительных границ), но их эффективность зависит от того, насколько хорошо пользователь способен различать изменения оттенков, высоты или других визуальных свойств.

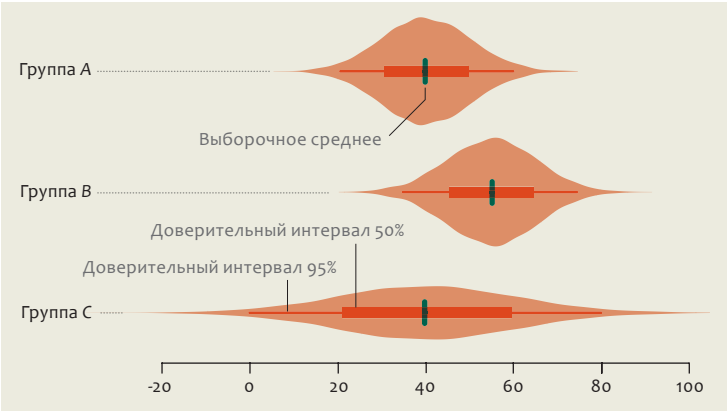
Плюсы

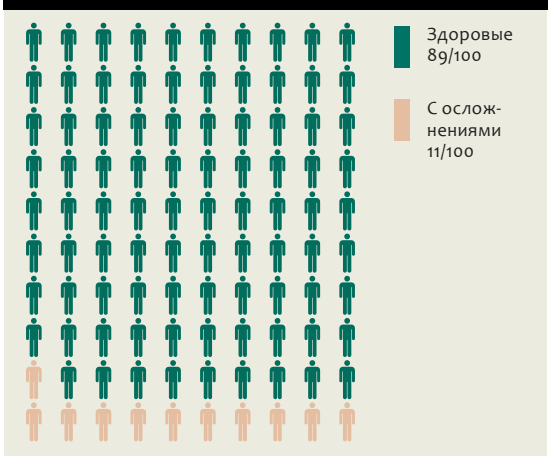
- Как правило, этот метод интуитивно понятен: затемнения или жесткие границы означают бóльшую определенность; более светлые цвета затенения или нечеткие границы — мéньшую.
- При использовании этого метода смысл данных понимается, как правило, корректно, чего не скажешь о методе интервалов.

Минусы

- Пользователь может не сразу догадаться, что плотность штриховки соответствует величине вероятности.
- Пользователи зачастую полагают, что в той части диаграммы, которая хорошо различима (наиболее темные или наиболее широкие области), содержатся фактические значения данных, а те части диаграммы, которые хуже различимы (очень светлый цвет или самые узкие области), — это и есть неопределенность, что неверно.
- Оценки могут быть смещены к самым темным цветам на диаграмме или самым высоким точкам.
- Могут возникать затруднения с определением конкретных значений вероятности.







Массивы иконок

Если выразить вероятность с помощью частоты (скажем, вероятность 30% будем интерпретировать как «три из десяти»), то пользователю будет легче воспринять такую информацию и, следовательно, он станет использовать ее корректнее. Вообще, человек лучше воспринимает вероятность в дискретном виде, поскольку сталкивается с ней каждый день.

Плюсы

- Этот метод более ясный, чем некоторые другие, ведь в этом случае пользователь без труда заметит, что вероятность соответствует количеству изображений.
- Если число используемых изображений невелико, то пользователь быстро обратит на это внимание, поскольку человеку вообще свойственно очень быстро замечать множества небольших размеров, для чего ему даже не нужно пересчитывать количество входящих элементов.

Минусы

- Предназначен для представления только одной вероятностной величины.

Пространственное представление выборки

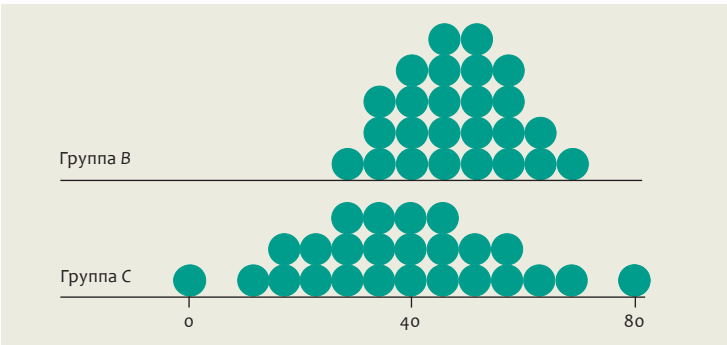
Пространственное представление выборочных данных вполне подойдет, если нужно отобразить вероятности одной или нескольких переменных величин в дискретном формате. Один из примеров такого подхода — квантильный точечный график. На нем изображены элементы распределения данных выборки (*в виде кружков*) таким образом, что их количество соответствует вероятности (*например, высота в два шарика, пять шариков — на рис. ниже*). Если существует неопределенность относительно значений параметров, на основе которых делаются оценки (например, начальные условия), то можно сгенерировать выборки, меняющие эти параметры, и показать в одной визуализации.

Плюсы

- Создатель диаграммы может изобразить любое количество данных, достаточное для того, чтобы просматривался тип распределения, и в то же время их не должно быть слишком много, иначе читатель не сможет различить изображения отдельных элементов.

Минусы

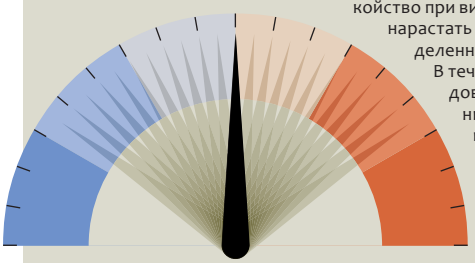
- Если количество элементов будет неоправданно большим, то читателю станет труднее производить корректные вычисления вероятности.
- Возможны неточности, особенно если исходное распределение сильно искажено так называемыми выбросами (резко выделяющимися значениями экспериментальных величин).



Graphics by Jessica Hullman and Jen Christiansen

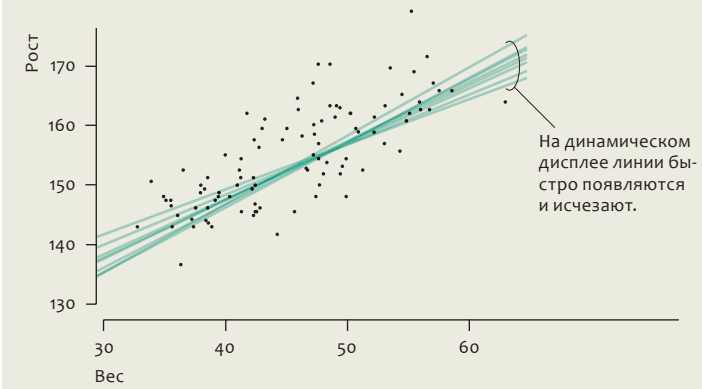
Стрелочный индикатор на президентских выборах 2016 г. в США

Методы визуализации иногда заставляют нас сильно поволноваться. Так было в 2016 г. В тот самый день, когда в США проходили президентские выборы, газета *The New York Times* поместила вечером на своем сайте динамический стрелочный индикатор, где отображались прогнозы результатов выборов. На полукруглом диске были секторы, окрашенные в разные цвета. Области слева соответствовали уверенной победе Клинтон, области справа — безоговорочной победе Трампа. Обновление индикатора происходило несколько раз в минуту по мере поступления новых результатов с избирательных участков. Стрелка прыгала то вправо, то влево, причем даже быстрее, чем обновлялись данные.



Вероятность победы на выборах

В тот вечер многие с волнением наблюдали, как непредсказуемо ведет себя стрелка индикатора. Беспокойство при виде пляшущей стрелки стало нарастать пропорционально неопределенности поступающих данных. В течение десятилетий граждане довольствовались лишь статичными прогнозами результатов выборов, при этом неопределенность упускалась из виду. И вот мы увидели резкий переход к визуализации неопределенности и, как результат, столкнулись с тревожными ожиданиями итогов выборов.



Динамическая визуализация во времени

Визуализация потока поступающих данных в виде динамического процесса дает нам возможность почувствовать, что такое неопределенность, — только в эти моменты понимаешь, что игнорировать ее не так-то просто. Название этого метода — диаграмма гипотетических результатов (*hypothetical outcome plot*). Его можно использовать для создания и простых, и сложных визуальных образов. Исследования механизмов человеческого восприятия говорят о том, что люди на редкость хорошо умеют судить о распределении данных, опираясь на частоту появления событий: чтобы оценить вероятность события, человеку нет необходимости подсчитывать, сколько раз оно наступает. Один из важных факторов — скорость наступления событий; она должна быть достаточно быстрой, чтобы человек смог увидеть меняющиеся события, и в то же время достаточно медленной, чтобы он имел возможность запомнить увиденное.

Плюсы

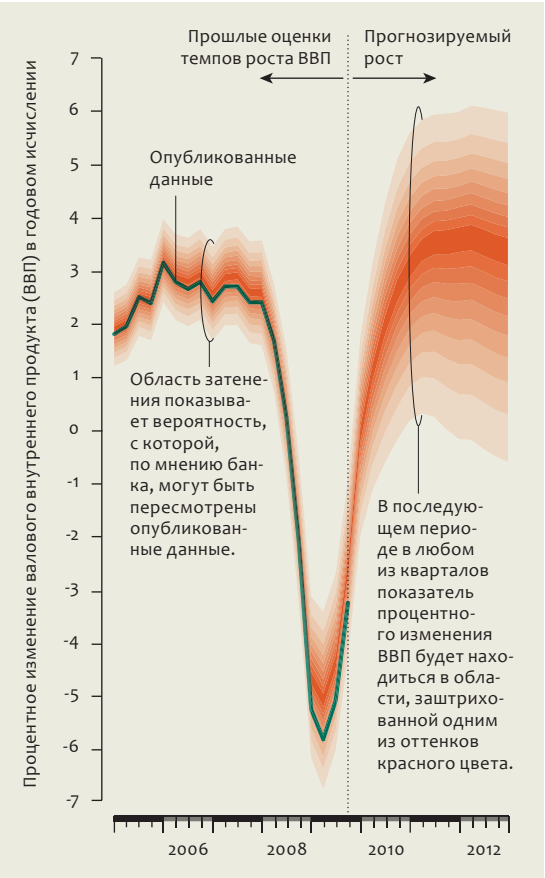
- Человеческий глаз способен довольно точно оценить вероятность событий; при этом человеку не нужно специально подсчитывать количество появляющихся объектов.
- Может широко применяться для различных типов данных и стилей визуализации.
- Динамическая визуализация позволяет оценивать вероятности сразу для многих переменных (на статическом графике это сделать трудно).

Минусы

- Неточность выборки, особенно если распределение сильно искажено выбросами.
- Сложно сказать, какую часть меняющихся данных заметит пользователь.
- Может оказаться, что некоторые форматы, например научные статьи, пока не поддерживают методы динамической визуализации.

Комбинированные подходы

Специалисты в области компьютерной графики могут создавать качественные рисунки и диаграммы с помощью комбинирования различных методов визуализации, не привязываясь к какому-то одному из них. В качестве примера мы приводим здесь известный клиновидный график, созданный Банком Англии. Слева от пунктирной линии (*находится в интервале между 2009 и 2010 гг.*) помещены данные за предыдущий период, справа — прогнозируемые данные. Неопределенность в предыдущем периоде — это важный фактор, обусловливающий неопределенность в будущем периоде. На этом графике области с более интенсивными цветовыми оттенками соответствуют более высокой вероятности, а с менее интенсивными — меньшей вероятности; разные оттенки соответствуют разным уровням достоверности (читатель сам их выбирает). Читатель может анализировать данные как по контуру полос, так и по их яркости. Некоторые из современных статистических программных пакетов, предназначенных для проектирования графиков и моделирования, позволяют комбинировать подходы визуализации.



ДОПОЛНИТЕЛЬНЫЕ ИСТОЧНИКИ

- Пентленд А. Как защитить большие данные от самих себя // ВМН, № 10, 2014.
- Picturing the Uncertain World: How to Understand, Communicate, and Control Uncertainty through Graphical Display. Harold Wainer. Princeton University Press, 2009.
- Visualizing Uncertainty. Claus O. Wilke in Fundamentals of Data Visualization. O'Reilly Media, 2019.

SOURCE: INFLATION REPORT. BANK OF ENGLAND, FEBRUARY 2010 (GDP chart)
Illustration by Tiffany Farrant-Gonzalez (election needle)