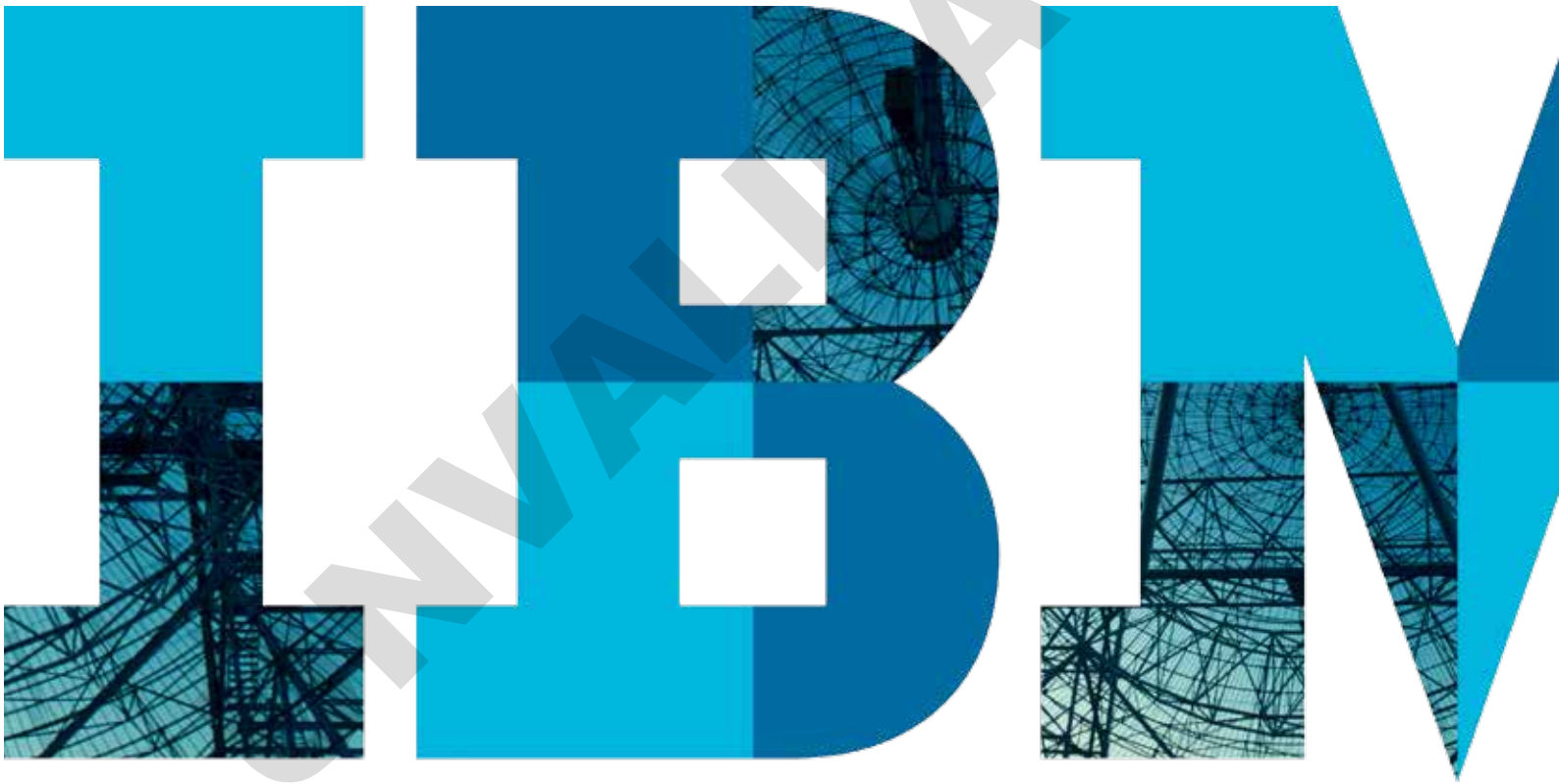
ノートブックの紹介:
データ・サイエンティスト向けの強力なツール



IBM

高速かつ柔軟な共同作業によるデータ探索と分析

データ探索と分析は、繰り返し行う反復的プロセスですが、ビジネスの需要を満たすために、データ・サイエンティストには常に長期の開発サイクルという贅沢が許されているわけではありません。データ・サイエンティストがもっと厳しい大きな質問にもっと早く答えることができたらどうなるでしょうか。もっと簡単かつ迅速に実験、テスト、仮説設定ができ、対話的分析でもっと共同作業ができたらどうなるでしょうか。

データ・サイエンス ― デジタル・ビスネスの変革と生き残りに必要

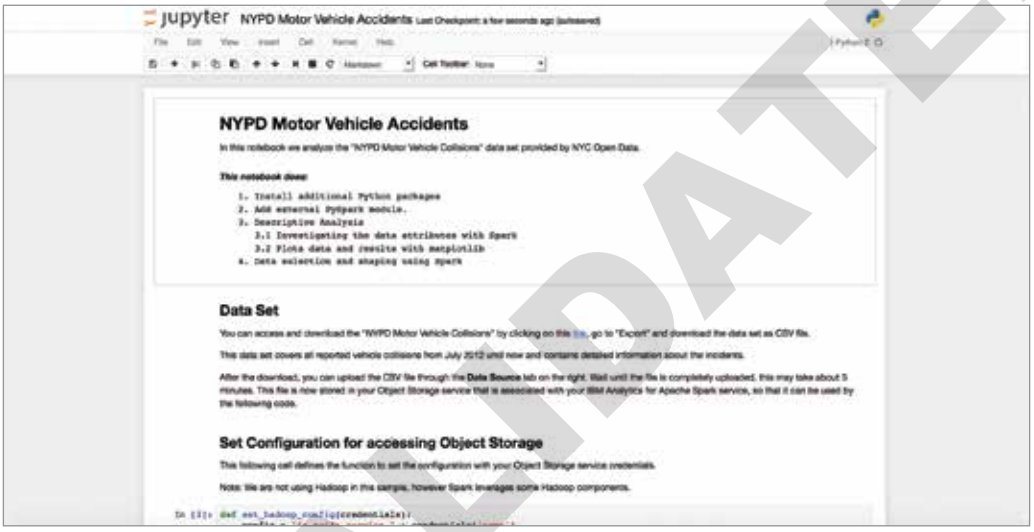
一流の会社はデータ・サイエンスを利用して市場を破壊しています。しかも俊敏で動きの速い環境でそれを行っています。この対応の速さというプレッシャーが、深い分析能力を持つ専門家、とりわけデータ・サイエンティストの肩にのしかかっています。データ・サイエンティストは困難な制約と戦って、企業の高い期待に応えるために、扱いにくい開発プロセスを改訂する必要があります。

データ・サイエンティストとビジネス上の利害関係者の双方にとって、道を間違えて行き詰まることは、多くの場合、例外的な出来事ではなく、いつものことです。必要なのは、すぐにミスを見つけ、コースを変えて、ビジネスの行方を変える可能性のある結論と知見に至る正しい道を見つけることです。

ノートブックを使った迅速で反復的な開発と共同作業

よくあるシナリオを考えてみましょう。会社の販売担当幹部が、最も人気のあるマシンの 1 つで交換部品の売上が急増していることに気付きました。幹部はこの異変の原因を明らかにするために、データ・サイエンティストの助けを借りて、動的データ探索を実施することにしました。 データ・サイエンティストはまず、データを抽出し整形してデータを探索します。最終目標は、企業の将来の販売戦略の指針となり、問題となっている特定の交換部品の販売予測の指針ともなるパターンとトレンドを発見することです。そのためには、データ・サイエンティストがビジネスを理解し、データに関する事実、知見、パターン、問題を簡単かつ理解しやすい方法で伝えることができる必要があります。たとえば、ビジネスの利害関係者向けに迅速にビジュアル・データ・モデルを提示したり、データ・サイエンティストと話し合って調査が正しい方向に進んでいるか確認することが考えられます。あいにく、多くの既存のツールはこの「早くて汚い」方法を実行できず、各種ツールを何度も切り替えて、データの抽出、整形、視覚化を行う必要があります。これは時間のかかる作業であり、必ずしもタイムリーに事態に対応できるわけでもありません。

ノートブックの例



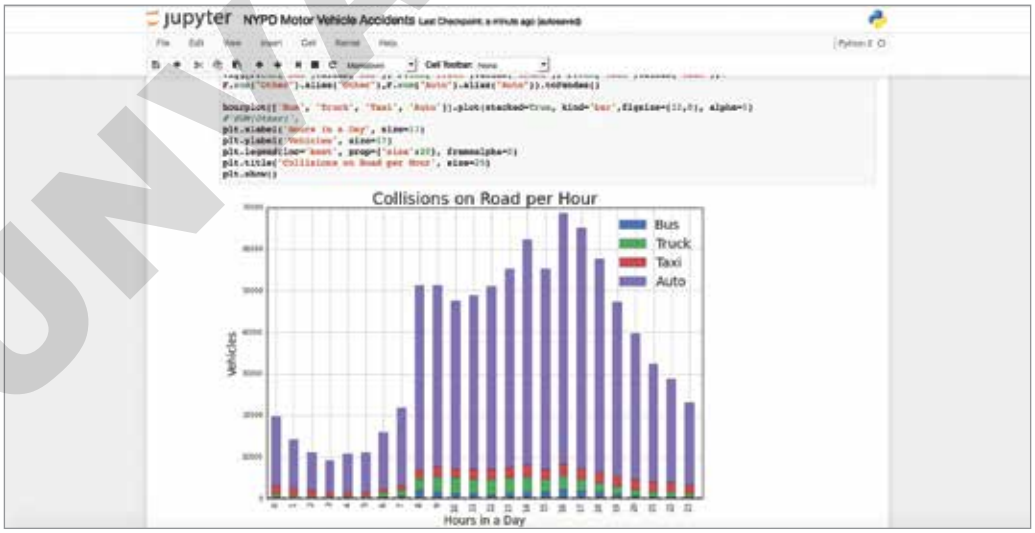


図 *1*:Jupyter Notebook の操作画面

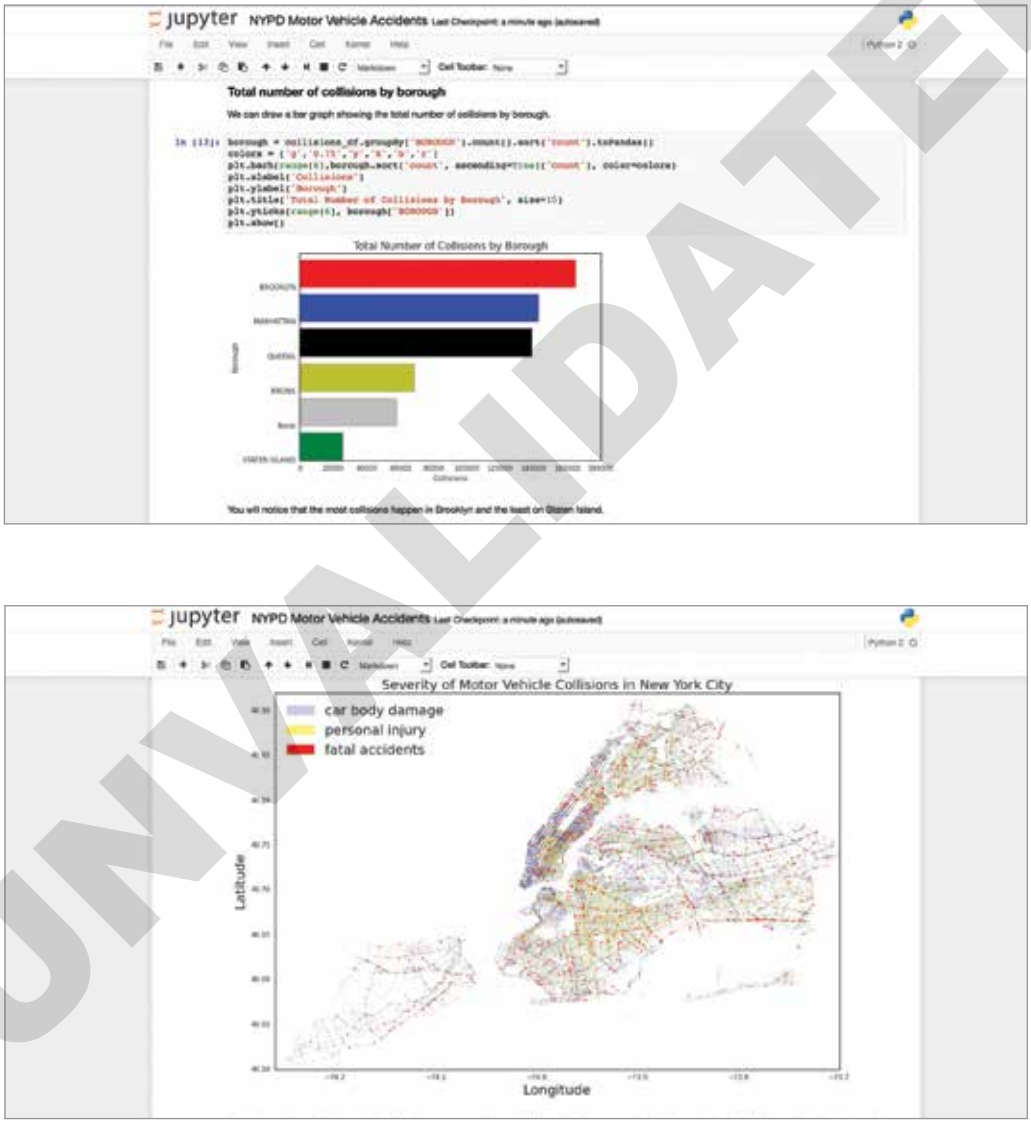


図 2:データを簡単に読み込んで、そのデータを柔軟な反復分析環境で分析します。

オープン・ソースの分析ノートブック―データ・サイエンスの流れを変えるソフト

分析ノートブックとは厳密には何なのでしょうか。よく思い浮かぶ最初の答えは、ノートブック・コンピューターや紙のノートですが、それではスピーディーで高度な分析というイメージを表してはいません。実際のところ、ノートブックは特定のデータ集約プロジェクトのコード、テキスト、画像やグラフをすべて含むWeb アプリケーションです。

ノートブックで作成された文書は共同作業者や利害関係者と簡単に共有することもできます。主としてデータ・サイエンティストがデータ診断、シミュレーション、統計的モデリング、機械学習のために使用します。ノートブックを使用すると、データ・モデルとフレームワークの試行プロセスと仮説のテスト・プロセスが高速化され、データ・サイエンス・チームと事業部門の担当者が迅速に反復して協力することが可能になります。

方法や結果を書き留め、簡単な図を描くのに使用する紙のノートと同じように、ノートブックはデータ・サイエンス・チームの唯一の作業文書という役割を果たし、試行錯誤が非常にしやすいツールです。データ・サイエンティストが扱うデータはノートブックで公開され、以前のステップを参照して、反復するごとに進捗を追跡できます。視覚化されたデータを見るだけでいいビジネス上の利害関係者とノートブックを共有するときは、コードを隠したり削除したりできます。

Jupyter Notebook は、IPython Notebook から進化したもので、現在入手できるオープンソースのデータ・サイエンス・ツールのなかでも屈指の広く採用された定評あるツールです。

Jupyter Notebook は、IPython Notebook から進化したもので、現在入手できるオープンソースのデータ・サイエンス・ツールのなかでも屈指の広く採用された定評あるツールです。コマンドラインで作業する代わりに、1 つのノートブック内で分析を実施し、画像やグラフを生成し、メディアと注釈を追加できます。またノートブックは共有可能です。つまり、データ・サイエンティストは会社の他の人と協力して作業することができ、それ自身が貴重な学習ツールとなります。データ・サイエンティストは、分析に注釈を入れて将来参照したり、データ・ディスカバリー・プロセスの各ステップを追跡したりできます。ノートブックはプロジェクトの意思疎通と文書作成の手段となり、社内の他のデータ・サイエンティストの学習手段にもなります。

データ・サイエンティストが Jupyter Notebook を利用してデータ探索とディスカバリーを劇的に改善している 3 つの主要分野を見ていきましょう。

滑らかな学習曲線

どの新しいツールでも同様ですが、ノートブックを使い始めるにあたって学習曲線が存在します。さいわい、データ・サイエンティストが参加して学習プロセスを容易にできる広大なユーザー・コミュニティが存在します。GitHub には 15 万件を超える Jupyter のノートブックが公開されており、経験豊富なノートブック・ユーザーの広範なネットワークにアクセスすれば、データ・サイエンティストがノートブックを使い始めるにあたって一人で悩んだり、すべてをゼロから構築したりする必要はありません。このレベルのサポートによって、データ・サイエンティストが技術的問題を解決する必要性が減り、データの操作や仮説の検証にもっと時間を使い、結論や実行可能なインサイトにより早く到達できるようになります。

迅速な反復

データ・サイエンティストは分析するデータの準備に相当の時間を費やします。これは気の弱い人向けの仕事ではありません。ソースが異なるデータはフォーマットが一貫していません。データは必ずしも完全なものでもありません。データ・サイエンティストが大量の汚れたデータの海で溺れることは珍しくないのです。ノートブックによってデータ洗浄の必要性がなくなるわけではありませんが、問題がどこにあるのかを素早く見つけて、適切な分野に注力してさらなる分析のためにデータを変換することで、データ・サイエンティストが「よりスマートに」働けるようになるという点で非常に貴重なツールです。先に触れたように、企業は迅速なインサイトを必要としています。

ノートブックは、企業の求めるスピードと並んで、データ・サイエンティストの求める反復作業もサポートします。

1 つのノートブックでは、文書内に埋め込まれた一連のコード・セル内で分析が実行されます。このコード・セルを使って、お互いを段階的に積み重ねて、データ・サイエンティストがデータに対して実行しているさまざまな操作を整然と並べることができます。データ・サイエンティストが実験内容を変更して分析の特定の段階を修正したくなったら、簡単に変更して適切なセルを再実行することができます。分析全体を手直しする必要は必ずしもありません。そして、ノートブックでは画像やグラフも生成でき、多くの場合、これによって何かを見つけやすく伝えやすくなるため、データ・サイエンティストは実験内容を素早く変更して、別のツールを使うことなく更新された画像やグラフを確認できます。

共同作業

データ・サイエンティストは多くの場合、IT 部門や業務部門、エンジニアリングなどの社内の従業員、また、エンジニア、開発者、アナリストなど、データ・サイエンス・エコシステム内の他の人と協力することが期待されています。情報共有と共同作業の能力は必須です。また、重要な注意点として、データ・サイエンスは必ずしも会社内で一元化されていません。多くの場合、部門や部署レベルに分散して存在しています。つまり、データの分析、探索、視覚化が社内の異なる分野で実施され共有されていないのです。
その結果、企業はデータ・サイエンスがもたらす恩恵を全面的に実現できていません。

ノートブックを共有しプロジェクトで共同作業をできることは、多様なスキルを持ったチームで作業する際には不可欠です。データ・サイエンティストは、数学、統計、コンピューター・サイエンスに関する相当なスキルを備えていますが、これは IT 部門や業務部門の誰にでも当てはまることではありません。さらに、データ・サイエンティストは同僚との話し合いを望むこともあります。一例として挙げられるのは、データ・サイエンティストがノートブックを使って開発者と共同作業をする場合です。開発者は、データ・サイエンティストと緊密に協力して分析モデルに慣れるにつれて、データ・サイエンティストが何を必要としているのかをよく理解し、それに応じてデータ・パイプラインを調整するようになります。このような機能 (ビッグ・データの分析、多言語サポート、共同作業) はすべて、データ・サイエンティストが高速に反復作業を実施し、次のプロジェクトに進むことを可能にしています。共有可能なプラットフォーム上で共同作業をすることで、データ探索プロジェクト (そして結果としてビジネス) は全員の力とスキルを全面的に発揮させることができます。

ノートブックの未来とデータ・サイエンティストに対する約束

ノートブックは強力なデータ・サイエンス・ツールであり、Apache Spark のようなビッグ・データ・テクノロジー・プラットフォームに埋め込まれるとさらに強力になります。Apache Spark は、大規模データ処理向けの超高速インメモリ・クラスタ・コンピューティング・エンジンです。Apache Spark のようなビッグ・データ・エンジンに接続されたノートブックは、従来の分散アーキテクチャよりも最大 100 倍速く出力を生成することができ、反復的探索がさらに高速化されます。Jupyter Notebook は、IPython カーネルに依存しており、このカーネルは Jupyter Notebook をインストールすると自動的にインストールされます。しかし、現在は、Jupyter Notebook にインストールできるカーネルが 65 以上存在しており、40 を超える言語がサポートされています。利用可能なカーネルのリストは増え続けています。最も人気のある分析ノートブックがオープンソースであるため、各種言語が利用でき、Python、R、Scala など、データ・サイエンスで最も一般的で発展している言語もサポートされています。ノートブックは、会社や組織のプログラムとアプリを接続する API として共有し利用することもできます。

まとめ

Jupyter のようなオープン・ソース・ノートブックは、データ・サイエンティストが学習曲線を管理しながら迅速に処理を反復して複数の利害関係者と協力できる環境をもたらします。管理されたビッグ・データ・サービスに接続されたノートブックはデータ・サイエンティストに最後の 1 マイルに注力する力、規模、時間を与えます。このような力を得ることで、データ・サイエンティストは道を切り開き、かつては地図のなかった場所を先頭に立って進み、予測可能な方法で事前対応的な運用を行います。

ノートブックと、いますぐ使い始める方法の詳細については、以下をご覧ください。

- Learn more about the IBM Data Science Experience
- Get started with notebooks in the Data Science Experience

日本アイ・ビー・エム株式会社〒103-8510東京都中央区日本橋箱崎町 19-21

Produced in Japan
June 2016

CDW12358-JPJA-00