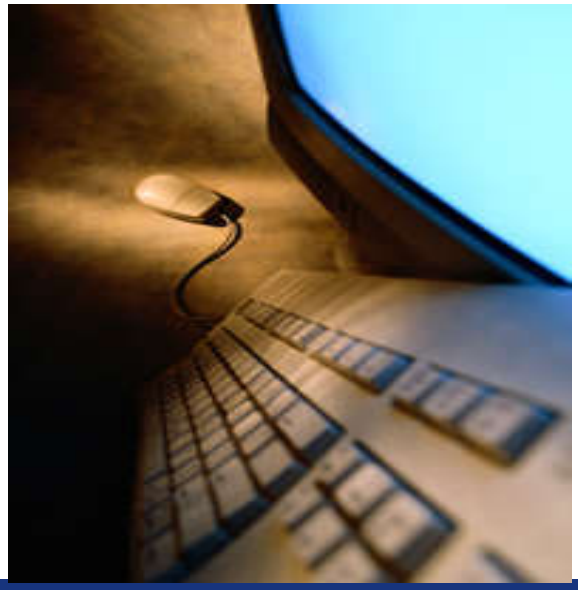
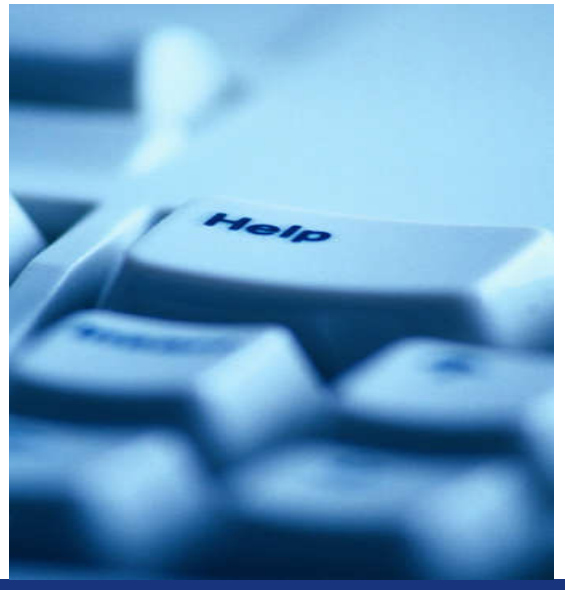
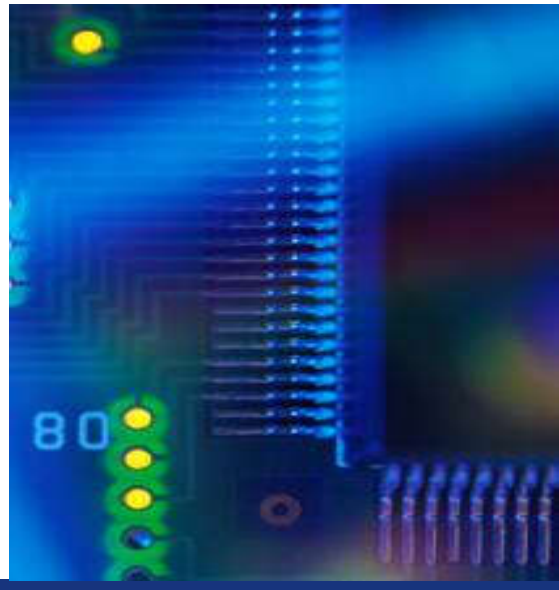
UNIVERSITAS
BUDI LUHUR





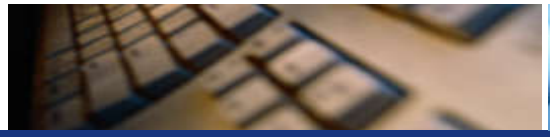


Pertemuan 2

DATA ANALYTICS LIFECYCLE

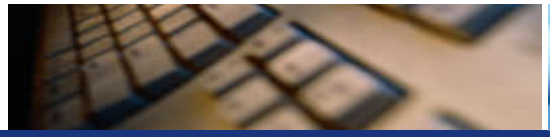
Data Analytics Lifecycle

- **Proyek Data Sains berbeda dari proyek BI**
  - Lebih bersifat eksplorasi
  - Sangat penting untuk memiliki proses proyek
  - Peserta harus teliti dan teliti
- **Memecah proyek-proyek besar menjadi potongan-potongan kecil**
- **Meluangkan waktu untuk merencanakan dalam lingkup pekerjaan**
- **Mendokumentasikan untuk menambah ketelitian dan kredibilitas**



Data Analytics Lifecycle

- Data Analytics Lifecycle Overview
- Phase 1: Discovery
- Phase 2: Data Preparation
- Phase 3: Model Planning
- Phase 4: Model Building
- Phase 5: Communicate Results
- Phase 6: Operationalize




Data Analytics Lifecycle Overview

- Siklus analitik data dirancang untuk masalah Big Data dan proyek ilmu data
- Dengan enam fase pekerjaan proyek dapat terjadi dalam beberapa fase secara bersamaan
- Siklus berulang untuk menggambarkan proyek yang nyata
- Pekerjaan dapat kembali ke fase sebelumnya jika terdapat informasi baru terungkap

Key Roles for a Successful Analytics Project

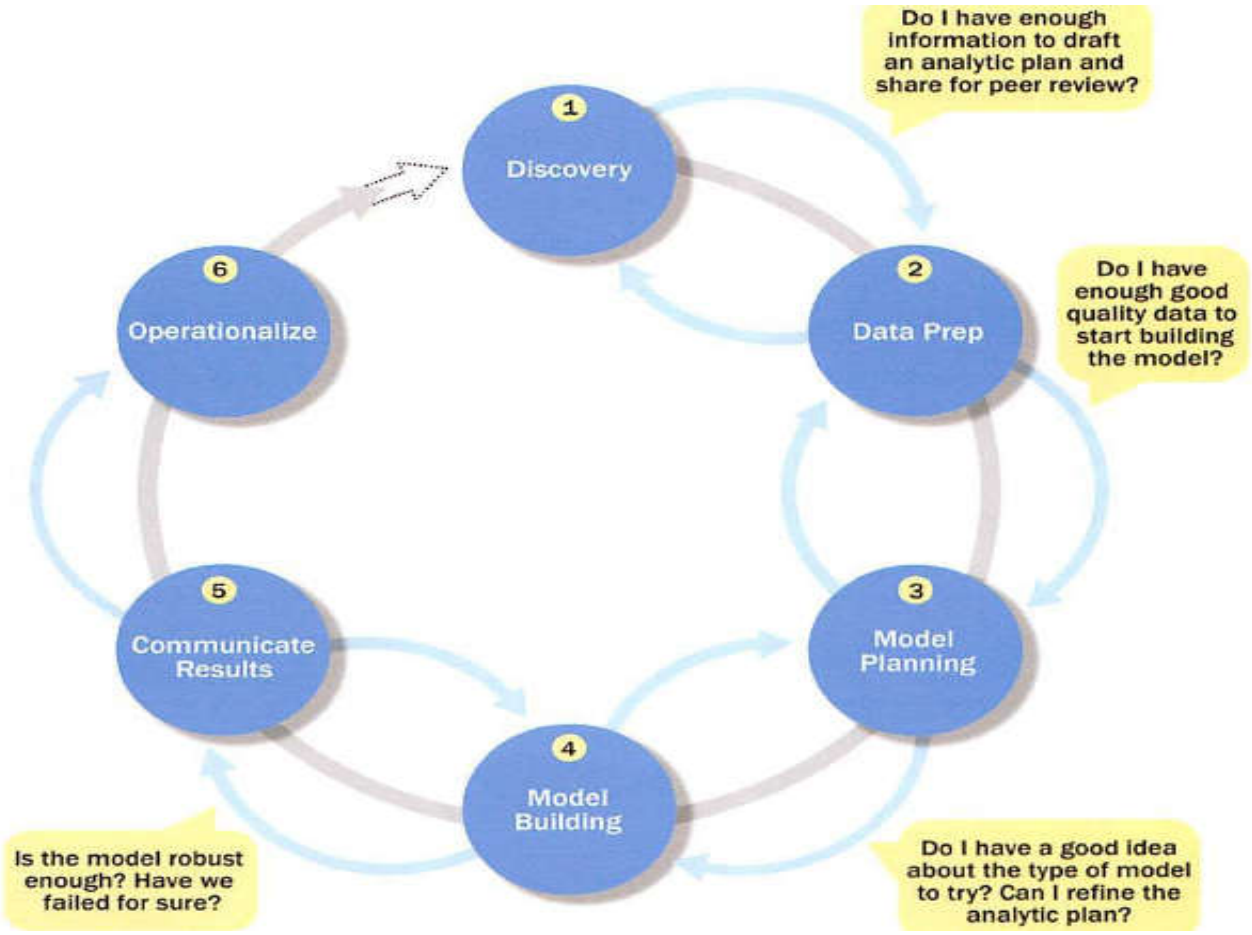

Key Roles for a Successful Analytics Project

- Business User – understands the domain area
- Project Sponsor – provides requirements
- Project Manager – ensures meeting objectives
- Business Intelligence Analyst – provides business domain expertise based on deep understanding of the data
- Database Administrator (DBA) – creates DB environment
- Data Engineer – provides technical skills, assists data management and extraction, supports analytic sandbox
- Data Scientist – provides analytic techniques and modeling

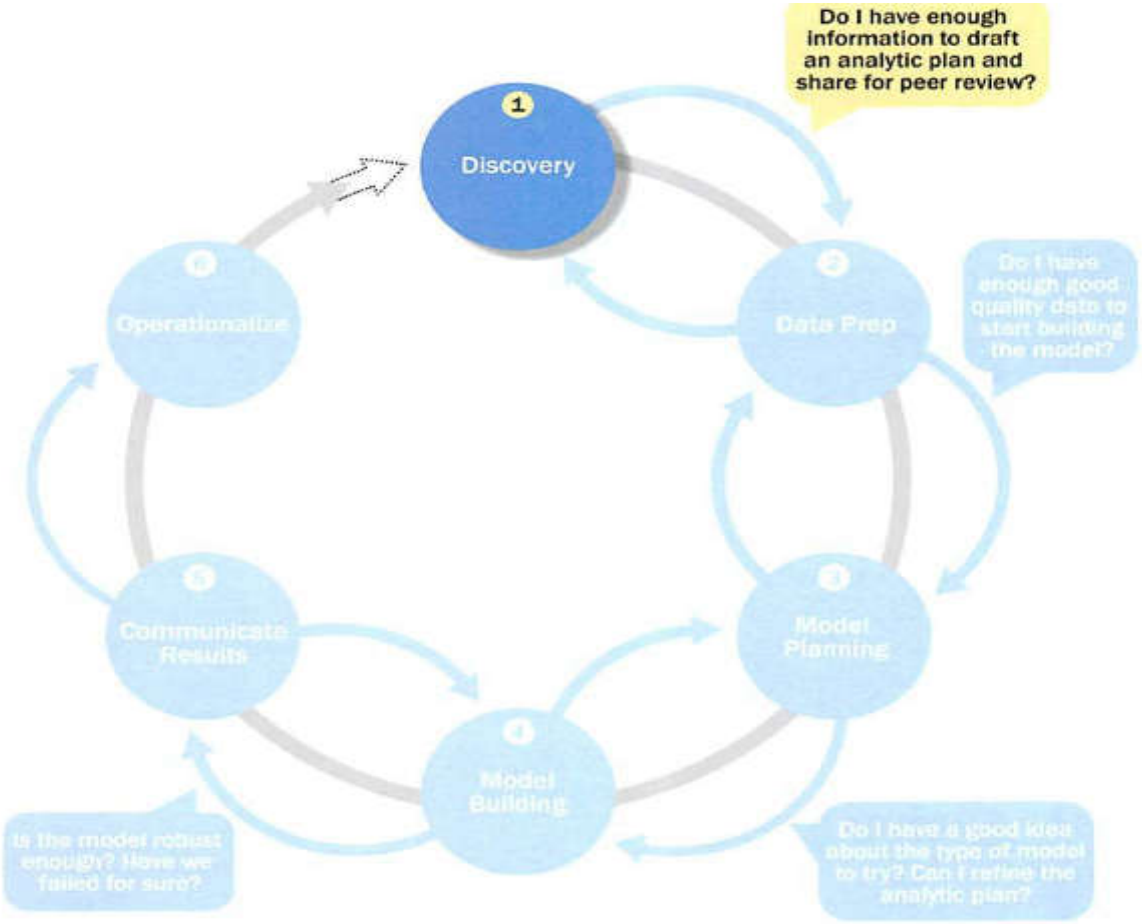
Background and Overview of Data Analytics Lifecycle

- **Data Analytics Lifecycle defines the analytics process and best practices from discovery to project completion**
- **The Lifecycle employs aspects of**
  - Scientific method
  - Cross Industry Standard Process for Data Mining (CRISP-DM)
    - Process model for data mining
  - Davenport's DELTA framework
  - Hubbard's Applied Information Economics (AIE) approach
  - MAD Skills: New Analysis Practices for Big Data by Cohen et al.

Overview Data Analytics Lifecycle



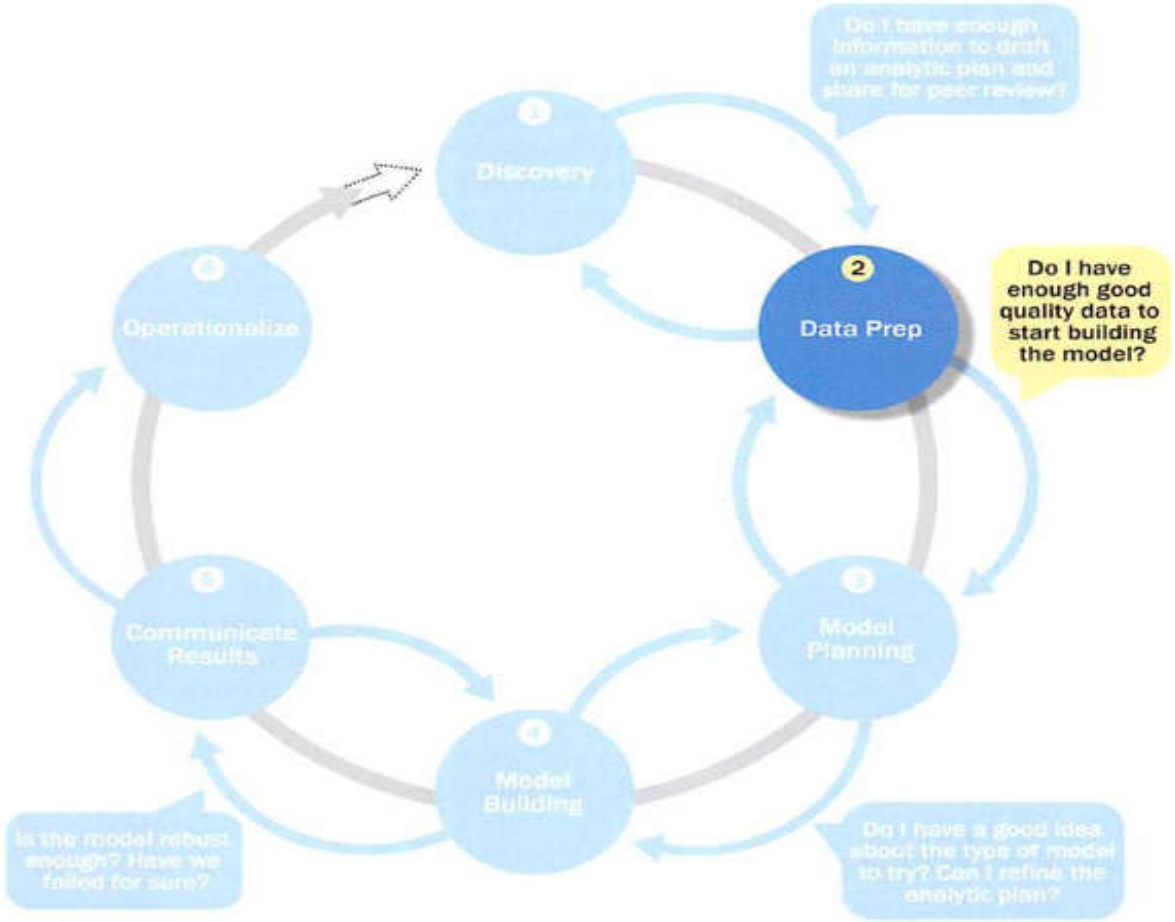
Phase 1: Discovery



Phase 1: Discovery

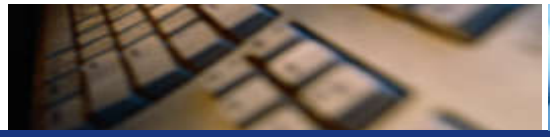
1. Mempelajari Domain Bisnis
2. Mengidentifikasi Sumber daya
3. Membingkai Masalah
4. Mengidentifikasi Pemangku Kepentingan Kunci
5. Mewawancarai Sponsor Analytics
6. Mengembangkan Hipotesis Awal
7. Mengidentifikasi Sumber Data Potensial

Phase 2: Data Preparation



Phase 2: Data Preparation

- Termasuk langkah-langkah untuk mengeksplorasi, prapemprosesan dan kondisi data
- Membuat lingkungan yang kuat - analytics sandbox
- Persiapan data cenderung menjadi langkah paling membutuhkan waktu yang cukup lama dalam siklus analitik
  - Seringkali setidaknya 50% dari waktu proyek data sains
- Tahap persiapan data umumnya yang paling berulang dan yang cenderung diremehkan oleh tim




Preparing the Analytic Sandbox

- Membuat analytic sandbox (juga disebut ruang kerja)
- Mengizinkan tim menjelajahi data tanpa mengganggu data produksi langsung
- Sandbox mengumpulkan semua jenis data (pendekatan ekspansif)
- Sandbox memungkinkan organisasi untuk melakukan proyek ambisius di luar analisis data tradisional dan BI untuk melakukan analisis prediktif canggih
- Meskipun konsep analytics sandbox relatif baru, konsep ini telah diterima oleh tim data sains dan grup TI


Performing ETLT (Extract, Transform, Load, Transform)

- Dalam pengguna ETL melakukan extract, transform, load
- Dalam Sandbox prosesnya ELT merupakan deteksi awal dalam mempertahankan data mentah yang dapat berguna untuk diperiksa
- Contoh - dalam deteksi penipuan kartu kredit, pencilan dapat mewakili transaksi berisiko tinggi yang mungkin secara tidak sengaja disaring atau diubah sebelum dimasukkan ke dalam basis data


Learning about the Data

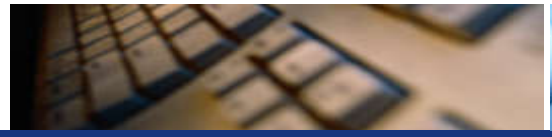
- **Mengenal data sangat penting**
- **Kegiatan ini mencapai beberapa tujuan:**
  - Menentukan data yang tersedia untuk tim di awal proyek
  - Menyoroti kesenjangan - mengidentifikasi data yang saat ini tidak tersedia
  - Identifikasi data di luar organisasi yang mungkin berguna


Learning about the Data Sample Dataset Inventory

Dataset	Data Available and Accessible	Data Available, but not Accessible	Data to Collect	Data to Obtain from Third Party Sources
Products shipped	●			
Product Financials		●		
Product Call Center Data		●		
Live Product Feedback Surveys			●	
Product Sentiment from Social Media				●

Data Conditioning

- **Pengkondisian data termasuk pembersihan data, menormalkan dataset, dan melakukan transformasi**
  - Sering dipandang sebagai langkah preproses sebelum analisis data, mungkin dilakukan oleh pemilik data, departemen TI, DBA, dll.
  - Terbaik untuk melibatkan ilmuwan data
  - Tim ilmu data lebih suka lebih banyak data daripada terlalu sedikit



Data Conditioning

- **Additional questions and considerations**
  - What are the data sources? Target fields?
  - How clean is the data?
  - How consistent are the contents and files? Missing or inconsistent values?
  - Assess the consistence of the data types – numeric, alphanumeric?
  - Review the contents to ensure the data makes sense
  - Look for evidence of systematic error


Survey and Visualize

- **Leverage data visualization tools to gain an overview of the data**
- **Shneiderman's mantra:**
  - "Overview first, zoom and filter, then details-on-demand"
  - This enables the user to find areas of interest, zoom and filter to find more detailed information about a particular area, then find the detailed data in that area

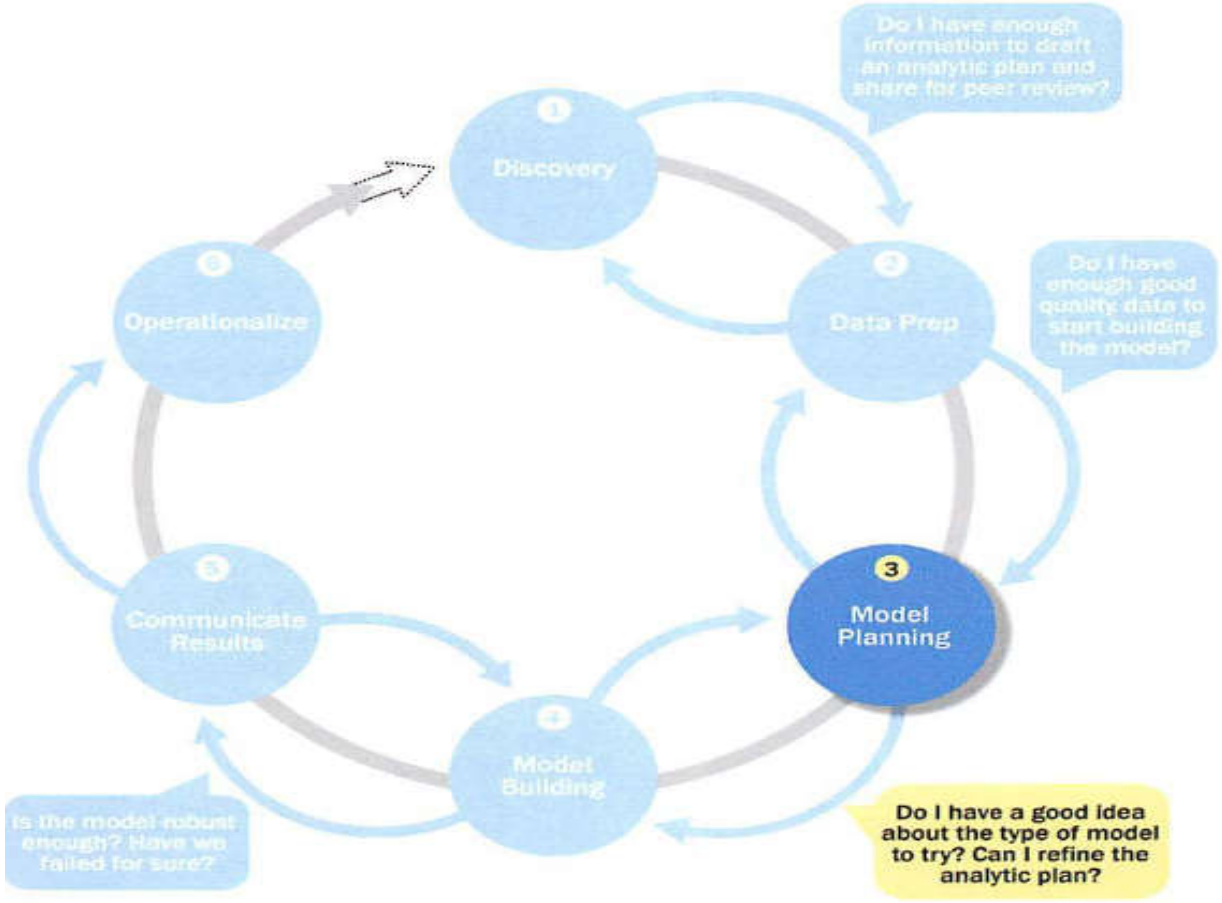

Survey and Visualize Guidelines and Considerations

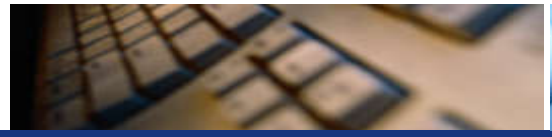
- Meninjau data untuk memastikan perhitungan konsisten
  Apakah distribusi data tetap konsisten?
- Menilai rincian data, kisaran nilai, dan tingkat agregasi data
  Apakah data mewakili populasi yang diminati?
- Periksa variabel terkait waktu - harian, mingguan, bulanan? Apakah ini cukup baik?
- Apakah data distandarisasi / dinormalisasi? Timbangan konsisten?
- Untuk dataset geospasial, apakah singkatan negara bagian / negara konsisten


Common Tools for Data Preparation

- Hadoop dapat melakukan penelaahan dan analisis paralel
- Alpine Miner menyediakan antarmuka pengguna grafis untuk membuat alur kerja analitik
- Open Refine (sebelumnya Google Perbaiki) adalah alat sumber terbuka gratis untuk bekerja dengan data yang berantakan
- Mirip dengan Open Refine, Wrangler Data adalah alat interaktif untuk pembersihan data transformasi

Phase 3: Model Planning






Phase 3: Model Planning

- **Aktivitas yang perlu dipertimbangkan**
  - Menilai struktur data - ini menentukan alat dan teknik analitik untuk fase selanjutnya
  - Memastikan teknik analitik memungkinkan tim untuk memenuhi tujuan bisnis dan menerima atau menolak hipotesis kerja
  - Menentukan apakah situasinya memerlukan model tunggal atau serangkaian teknik sebagai bagian dari alur kerja analitik yang lebih besar
  - Meneliti dan pahami bagaimana analis lain mendekati masalah seperti ini atau yang sejenis


Phase 3: Model Planning Model Planning in Industry Verticals

❑ Example of other analysts approaching a similar problem

Market Sector	Analytic Techniques/Methods Used
Consumer Packaged Goods	Multiple linear regression, automatic relevance determination (ARD), and decision tree
Retail Banking	Multiple regression
Retail Business	Logistic regression, ARD, decision tree
Wireless Telecom	Neural network, decision tree, hierarchical neurofuzzy systems, rule evolver, logistic regression


Data Exploration and Variable Selection

- Jelajahi data untuk memahami hubungan antar variabel untuk menginformasikan pemilihan variabel dan metode
- Cara umum untuk melakukan ini adalah dengan menggunakan alat visualisasi data
- Seringkali, para pemangku kepentingan dan pakar materi mungkin memiliki ide
  - Misalnya, beberapa hipotesis yang mengarah pada proyek
- Bertujuan untuk menangkap prediktor dan variabel yang paling penting
  - Ini sering membutuhkan iterasi dan pengujian untuk mengidentifikasi variabel kunci
- Jika tim berencana untuk menjalankan analisis regresi, identifikasi kandidat prediktor dan variabel hasil dari model

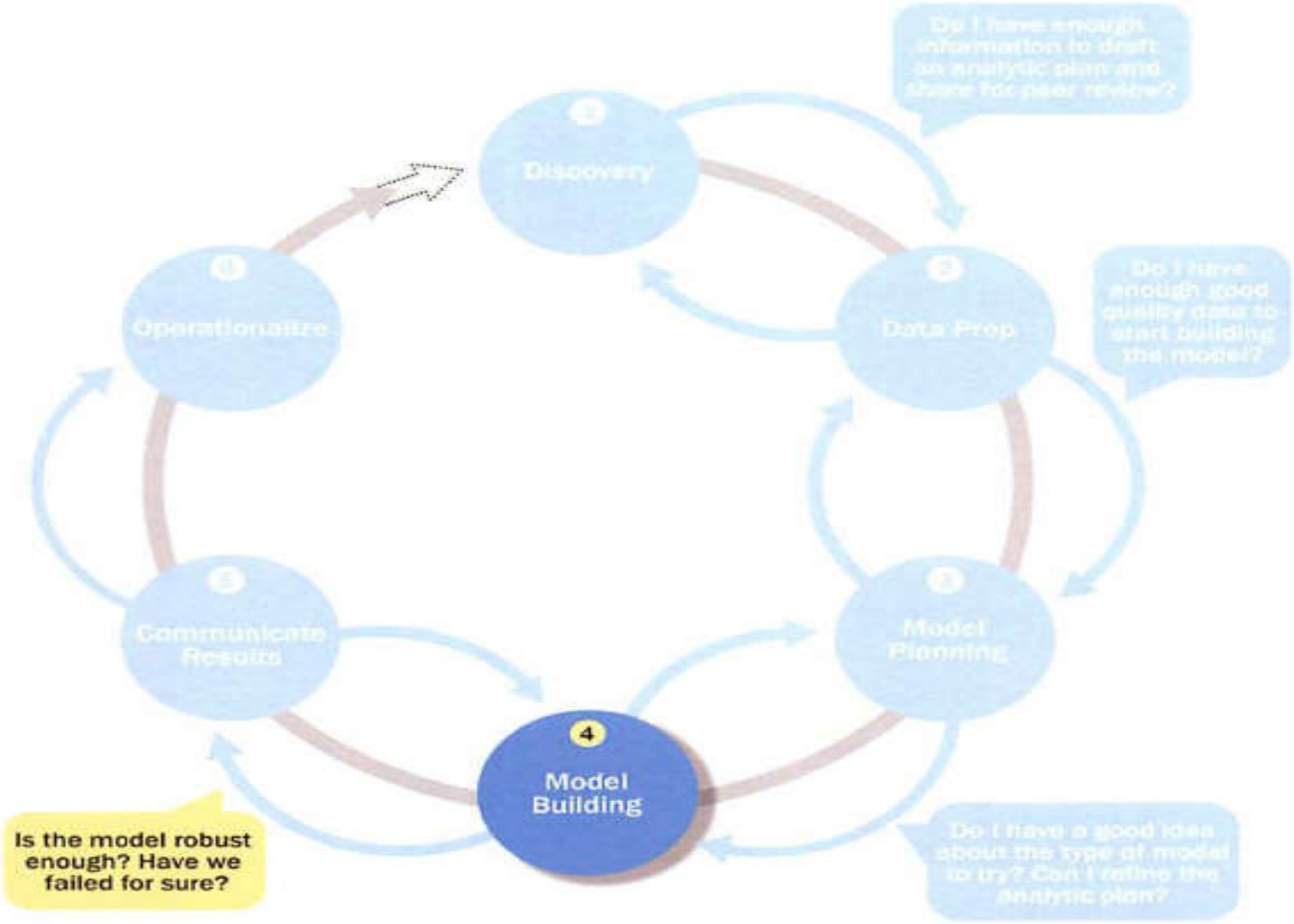

Model Selection

- **Tujuan utamanya adalah memilih teknik analisis, atau beberapa kandidat, berdasarkan tujuan akhir proyek**
- **Mengamati peristiwa di dunia nyata dan berupaya membangun model yang meniru perilaku ini dengan serangkaian aturan dan ketentuan**
  - Model hanyalah abstraksi dari kenyataan
- **Menentukan apakah akan menggunakan teknik yang paling cocok untuk data terstruktur, data tidak terstruktur, atau pendekatan hybrid**
- **Tim sering membuat model awal menggunakan paket perangkat lunak statistik seperti R, SAS, atau Matlab**
  - Yang mungkin memiliki keterbatasan ketika diterapkan pada dataset yang sangat besar
- **Tim bergerak ke fase pembangunan model setelah memiliki ide bagus tentang jenis model yang akan dicoba**

Common Tools for the Model Planning Phase

- R memiliki serangkaian kemampuan pemodelan yang lengkap
- R berisi sekitar 5000 paket untuk analisis data dan presentasi grafis
- Layanan Analisis SQL dapat melakukan analisis dalam-database dari fungsi-fungsi penambangan data umum, agregasi yang terlibat, dan model-model prediksi dasar
- SAS / ACCESS menyediakan integrasi antara SAS dan Sandbox Analitik melalui beberapa koneksi data

Phase 4: Model Building




Phase 4: Model Building

- Jalankan model yang didefinisikan dalam Fase 3
- Mengembangkan dataset untuk pelatihan, pengujian, dan produksi
- Kembangkan model analitik pada data pelatihan, tes pada data uji
- Pertanyaan untuk dipertimbangkan
  - Apakah model tampak valid dan akurat pada data uji?
  - Apakah output model / perilaku masuk akal bagi para pakar domain?
  - Apakah nilai parameter masuk akal dalam konteks domain?
  - Apakah model cukup akurat untuk memenuhi tujuan?
  - Apakah model menghindari kesalahan yang tidak dapat ditolerir?
  - Apakah dibutuhkan lebih banyak data atau input?
  - Apakah model yang dipilih akan mendukung lingkungan runtime?
  - Apakah diperlukan bentuk model yang berbeda untuk mengatasi masalah bisnis?


Common Tools for the Model Building Phase

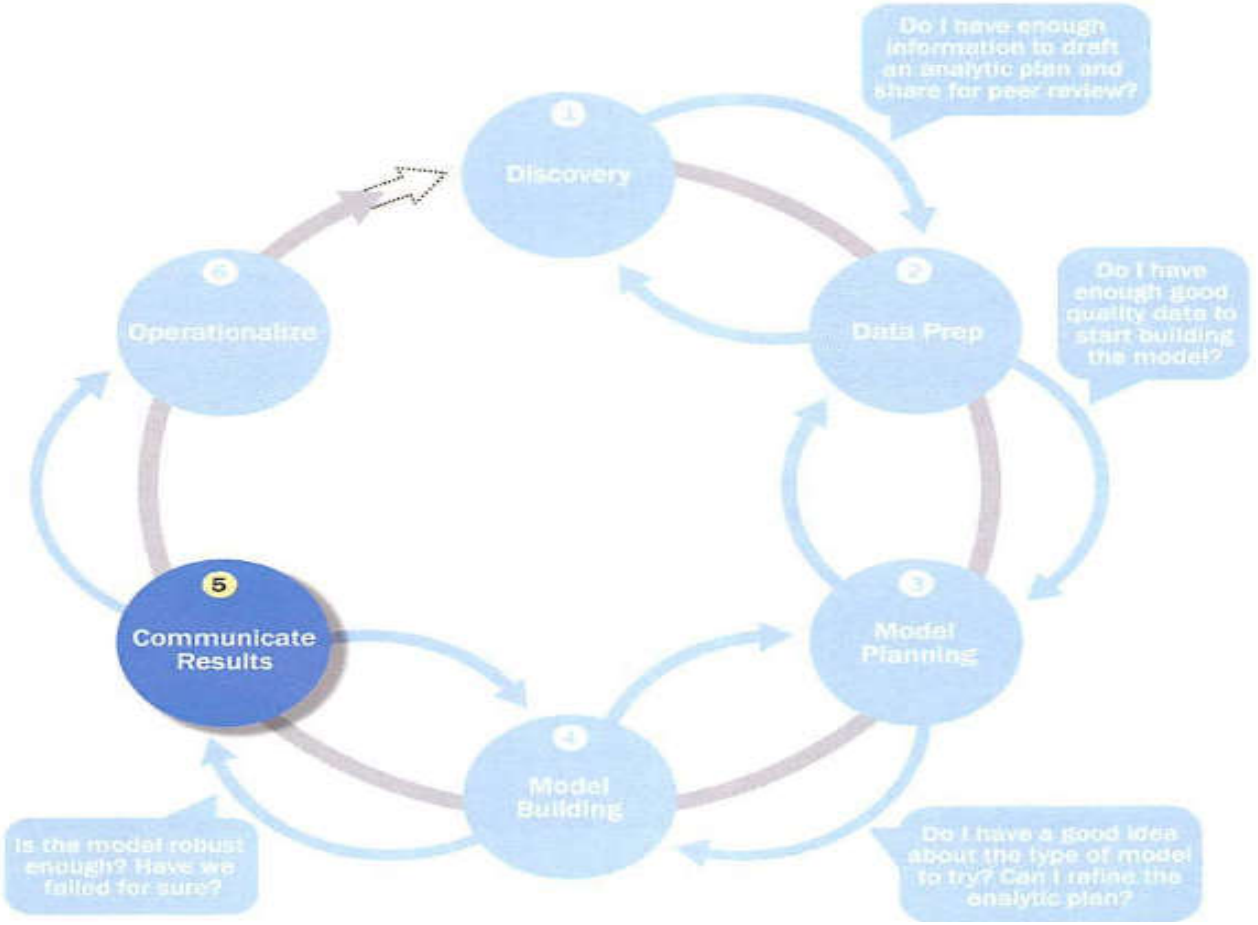
Commercial Tools

- SAS Enterprise Miner – built for enterprise-level computing and analytics
- SPSS Modeler (IBM) – provides enterprise-level computing and analytics
- Matlab – high-level language for data analytics, algorithms, data exploration
- Alpine Miner – provides GUI frontend for backend analytics tools
- STATISTICA and MATHEMATICA – popular data mining and analytics tools

Free or Open Source Tools

- R and PL/R - PL/R is a procedural language for PostgreSQL with R
- Octave – language for computational modeling
- WEKA – data mining software package with analytic workbench
- Python – language providing toolkits for machine learning and analysis
- SQL – in-database implementations provide an alternative tool (see Chap 11)

Phase 5: Communicate Results

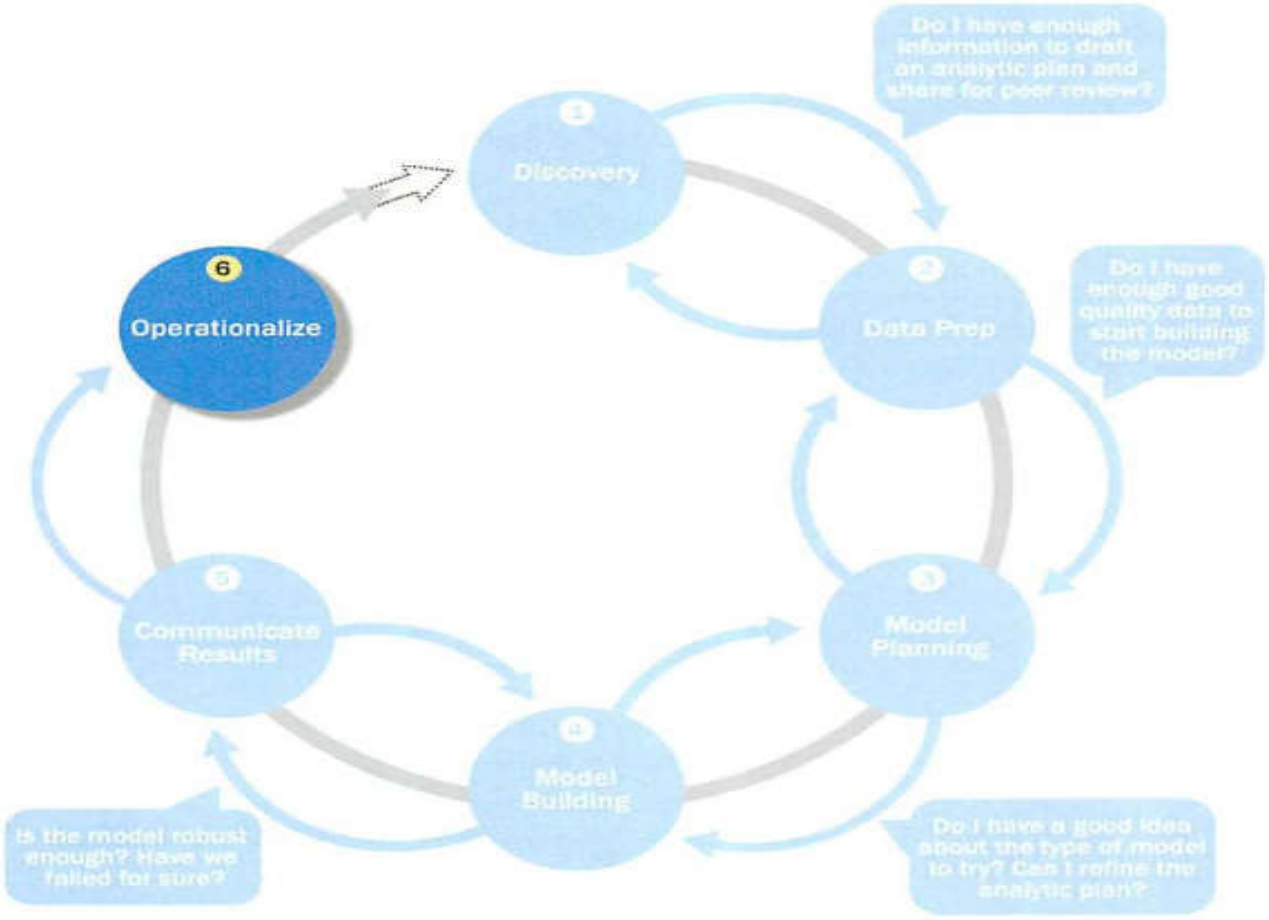


Phase 5: Communicate Results

- **Tentukan apakah tim berhasil atau gagal dalam tujuannya**
  Nilai jika hasilnya signifikan secara statistik dan valid
  Jika demikian, identifikasi aspek-aspek hasil yang menyajikan temuan yang menonjol
  Identifikasi hasil yang mengejutkan dan yang sesuai dengan hipotesis
  Komunikasikan dan dokumentasikan temuan kunci dan wawasan utama yang diperoleh dari analisis
  Ini adalah bagian proses yang paling terlihat bagi para pemangku kepentingan dan sponsor luar

- **Determine if the team succeeded or failed in its objectives**
- **Assess if the results are statistically significant and valid**
  - If so, identify aspects of the results that present salient findings
  - Identify surprising results and those in line with the hypotheses
- **Communicate and document the key findings and major insights derived from the analysis**
  - This is the most visible portion of the process to the outside stakeholders and sponsors

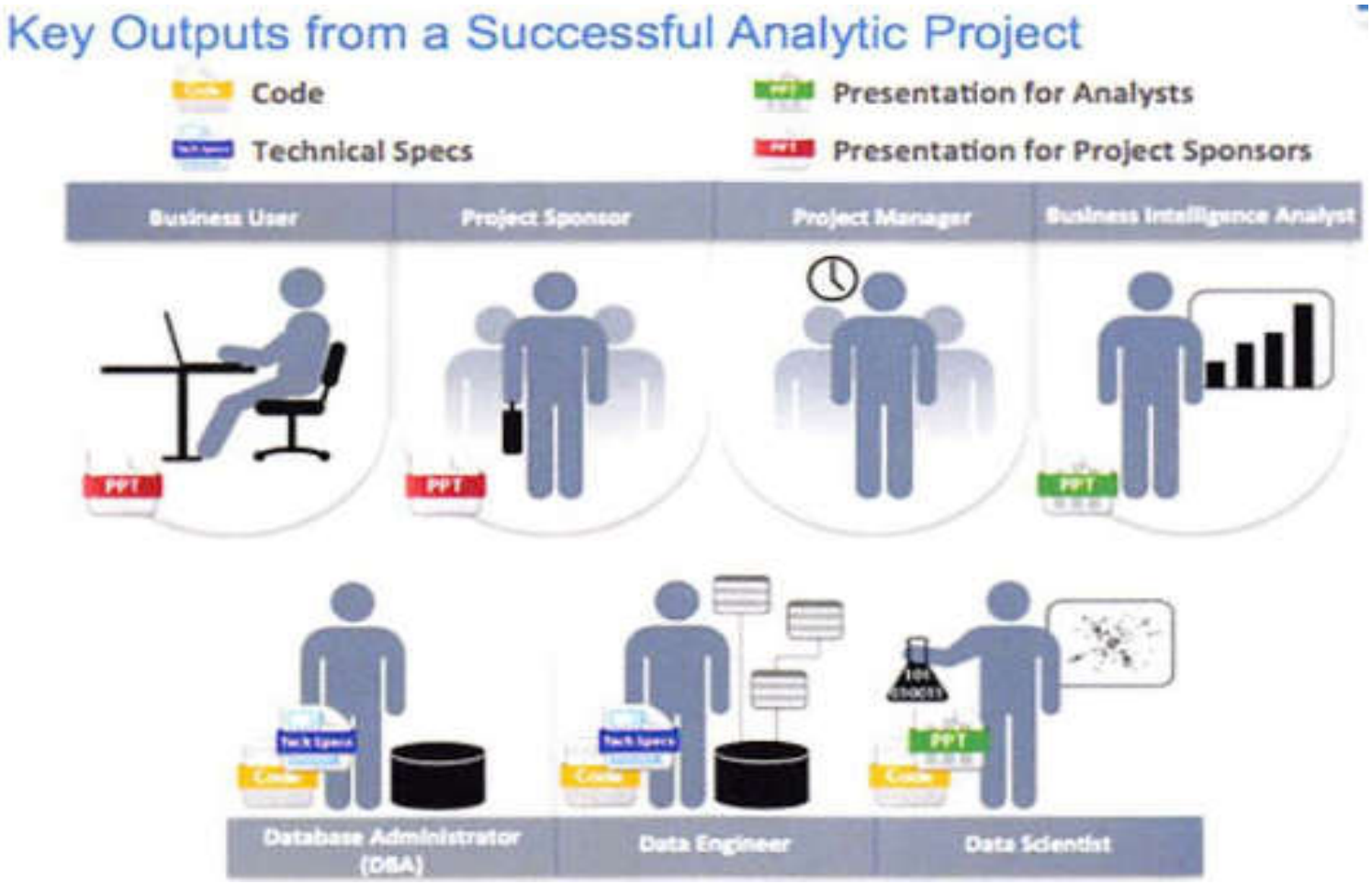
Phase 6: Operationalize




Phase 6: Operationalize

- Pada fase terakhir ini, tim mengkomunikasikan manfaat proyek secara lebih luas dan membuat proyek percontohan untuk menyebarkan pekerjaan secara terkendali.
- Risiko dikelola secara efektif dengan melakukan lingkup kecil, penempatan pilot sebelum peluncuran skala besar
- Selama proyek uji coba, tim mungkin perlu mengeksekusi algoritma lebih efisien dalam database daripada dengan alat dalam memori seperti R, terutama dengan set data yang lebih besar
- Untuk menguji model dalam pengaturan langsung, pertimbangkan menjalankan model dalam lingkungan produksi untuk satu set produk terpisah atau satu lini bisnis
- Monitor akurasi model dan latih kembali model jika perlu

Phase 6: Operationalize Key outputs from successful analytics project



Phase 6: Operationalize Key outputs from successful analytics project

- Pengguna bisnis - mencoba menentukan manfaat dan implikasi bisnis
- Sponsor proyek - menginginkan dampak bisnis, risiko, ROI
- Manajer proyek - perlu menentukan apakah proyek selesai tepat waktu, sesuai anggaran, tujuan tercapai
- Analis intelijen bisnis - perlu tahu apakah laporan dan dasbor akan terkena dampak dan perlu diubah
- Insinyur data dan DBA - harus berbagi kode dan dokumen
- Ilmuwan data - harus membagikan kode dan menjelaskan model kepada rekan kerja, manajer, pemangku kepentingan

Phase 6: Operationalize Four main deliverables

- Meskipun ketujuh peran tersebut mewakili banyak kepentingan, kepentingan tersebut tumpang tindih dan dapat dipenuhi dengan empat hasil utama
  1. Presentasi untuk sponsor proyek - takeaways tingkat tinggi untuk pemangku kepentingan tingkat eksekutif
  2. Presentasi untuk analis - menjelaskan perubahan proses bisnis dan perubahan pelaporan, termasuk detail dan grafik teknis
  3. Kode untuk orang teknis
  4. Spesifikasi teknis penerapan kode


Case Study: Global Innovation Network and Analysis (GINA)

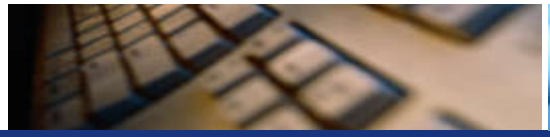
- **Pada 2012 direktur baru EMC ingin meningkatkan keterlibatan karyawan perusahaan di seluruh pusat keunggulan global (GCE) untuk mendorong inovasi, penelitian, dan kemitraan universitas**
- **Proyek ini dibuat untuk diselesaikan**
  - Menyimpan data formal dan informal
  - Lacak penelitian dari teknologi global
  - Tambang data untuk pola dan wawasan untuk meningkatkan operasi dan strategi tim


Phase 1: Discovery

- **Anggota dan peran tim**
  - Pengguna bisnis, sponsor proyek, manajer proyek - Wakil Presiden dari Kantor CTO
  - Analis BI - orang dari IT
  - Insinyur data dan DBA - orang-orang dari IT
  - Ilmuwan data - insinyur terkemuka

Phase 1: Discovery

- **Data terbagi dalam dua kategori**
  - Lima tahun pengajuan ide dari kontes inovasi internal
  - Risalah dan catatan yang mewakili inovasi dan kegiatan penelitian dari seluruh dunia
- **Hipotesis dikelompokkan menjadi dua kategori**
  - Analisis deskriptif tentang apa yang terjadi untuk memicu kreativitas lebih lanjut, kolaborasi, dan generasi aset
  - Analitik prediktif untuk memberi saran kepada manajemen eksekutif tentang di mana ia seharusnya berinvestasi di masa depan




Phase 2: Data Preparation

- Siapkan analytics sandbox
- Ditemukan bahwa data tertentu memerlukan pengkondisian dan normalisasi dan bahwa dataset yang hilang sangat penting
- Tim mengakui bahwa data berkualitas buruk dapat memengaruhi langkah-langkah selanjutnya
- Mereka menemukan banyak nama yang salah eja dan bermasalah dengan ruang ekstra
- Masalah yang tampaknya kecil ini harus diatasi

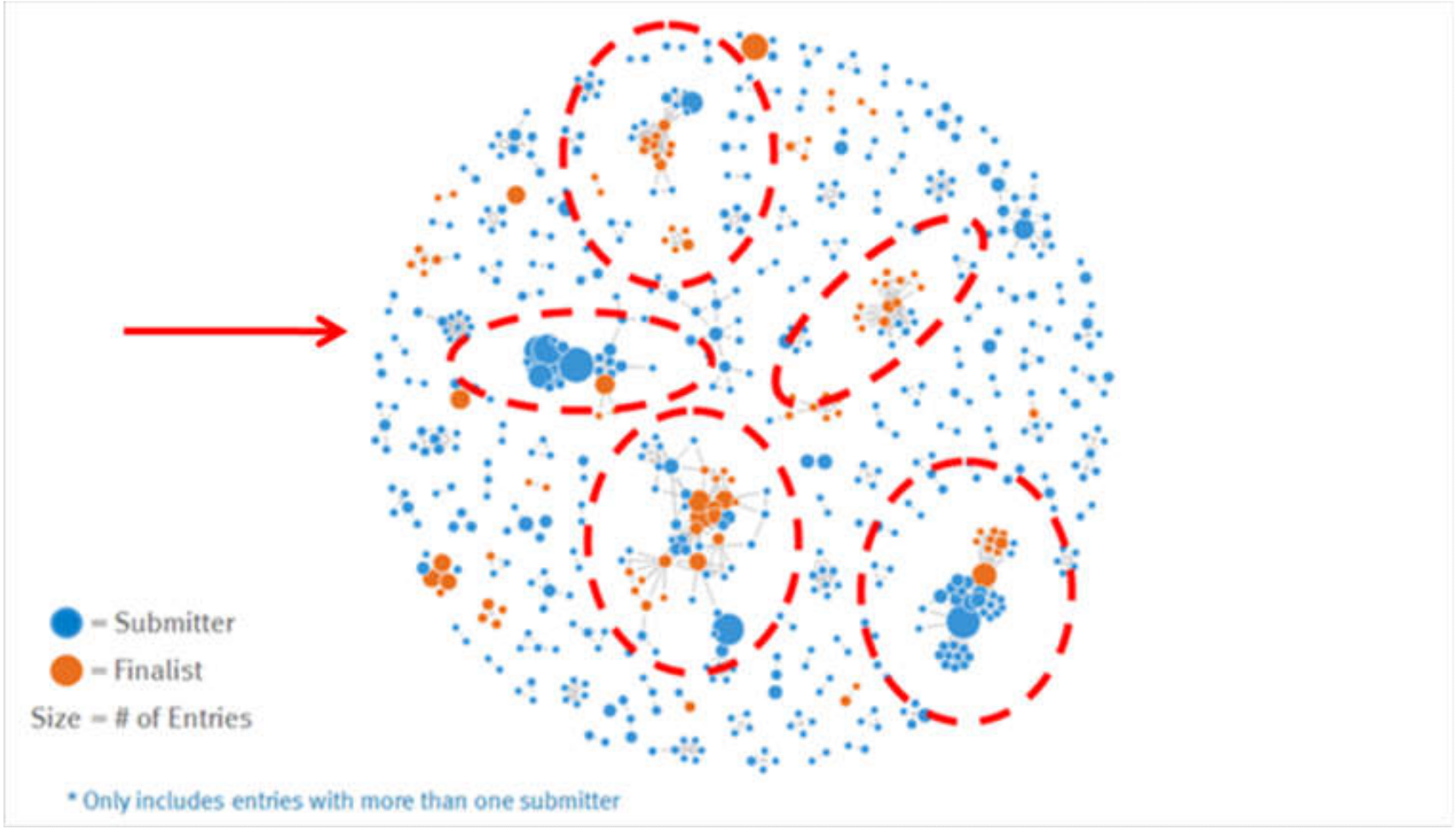
Phase 3: Model Planning

- **Studi ini mencakup pertimbangan berikut**
  - Identifikasi tonggak yang tepat untuk mencapai tujuan
  - Lacak bagaimana orang memindahkan ide dari setiap tonggak menuju tujuan
  - Trak ide-ide yang mati dan yang lain yang mencapai tujuan
  - Bandingkan waktu dan hasil dengan menggunakan beberapa metode berbeda

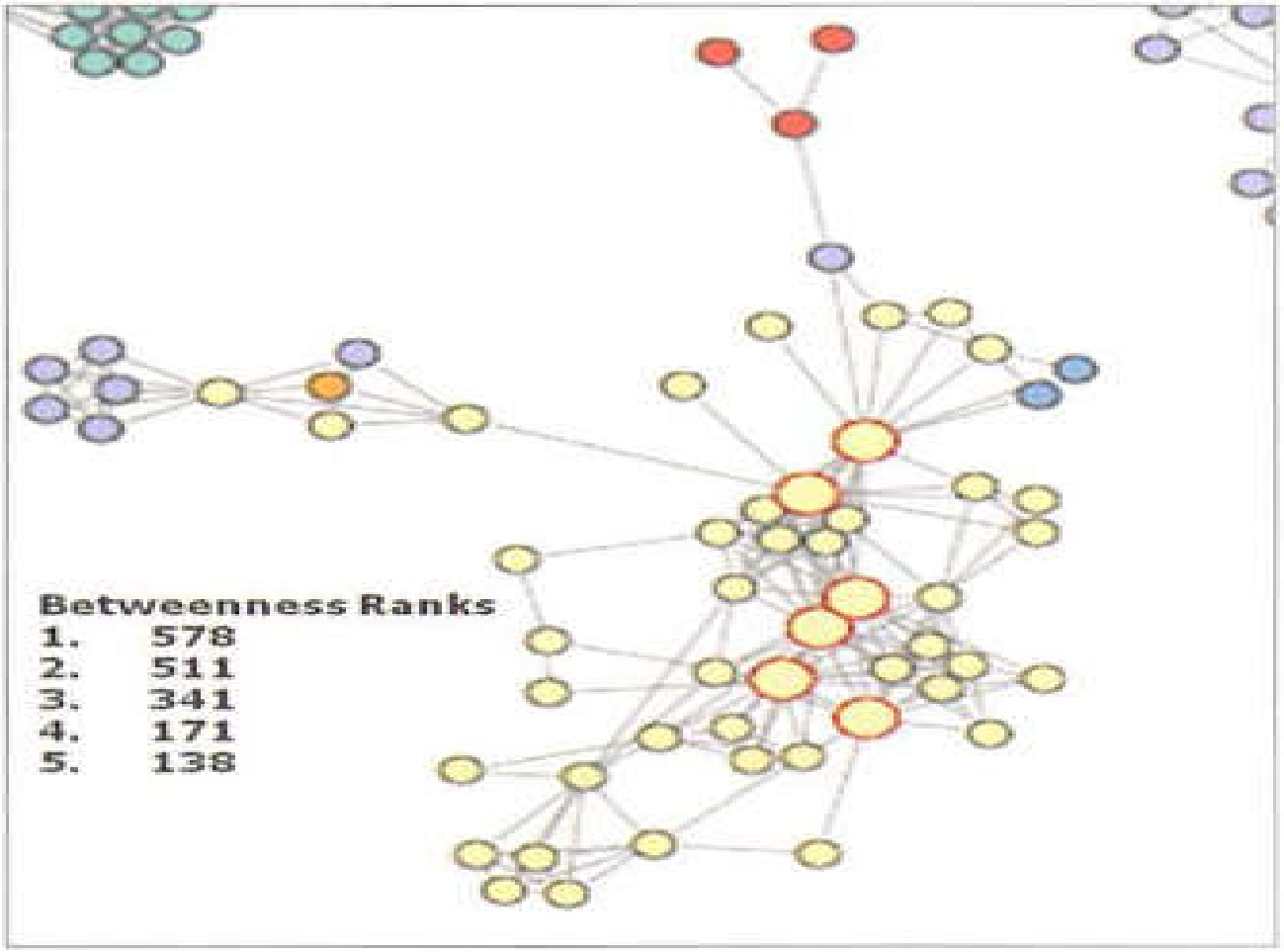

Phase 4: Model Building

- **Beberapa metode analitik digunakan**
  - NLP pada deskripsi tekstual
  - Analisis jaringan sosial menggunakan R dan Rstudio
  - Grafik dan visualisasi sosial yang dikembangkan
  - Several analytic method were employed

Phase 4: Model Building Social graph of data submitters and finalists



Phase 4: Model Building Social graph of top innovation influencers

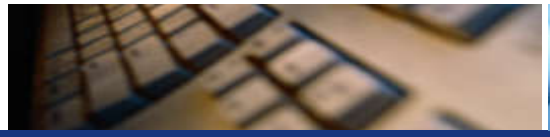


Phase 5: Communicate Results

- **Study was successful in in identifying hidden innovators**
  - Found high density of innovators in Cork, Ireland
- **The CTO office launched longitudinal studies**

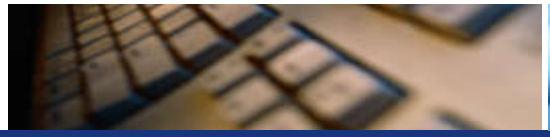
Phase 6: Operationalize

- **Deployment was not really discussed**
- **Key findings**
  - Need more data in future
  - Some data were sensitive
  - A parallel initiative needs to be created to improve basic BI activities
  - A mechanism is needed to continually reevaluate the model after deployment



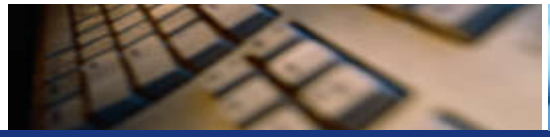

Phase 6: Operationalize

Components of Analytic Plan	GINA Case Study
Discovery Business Problem Framed	Tracking global knowledge growth, ensuring effective knowledge transfer, and quickly converting it into corporate assets. Executing on these three elements should accelerate innovation.
Initial Hypotheses	An increase in geographic knowledge transfer improves the speed of idea delivery.
Data	Five years of innovation idea submissions and history; six months of textual notes from global innovation and research activities
Model Planning Analytic Technique	Social network analysis, social graphs, clustering, and regression analysis
Result and Key Findings	1. Identified hidden, high-value innovators and found ways to share their knowledge 2. Informed investment decisions in university research projects 3. Created tools to help submitters improve ideas with idea recommender systems




Summary

- Data Analytics Lifecycle adalah pendekatan untuk mengelola dan melaksanakan proyek analitik
- Siklus hidup memiliki enam fase
- Sebagian besar waktu biasanya dihabiskan untuk persiapan - fase 1 dan 2
- Tujuh peran diperlukan untuk tim ilmu data
- Meninjau latihan



Focus of Course

- Fokus pada disiplin kuantitatif - mis., Matematika, statistik, pembelajaran mesin
- Berikan ikhtisar analisis Big Data
- Studi mendalam tentang beberapa algoritma kunci

SELESAI