# Lý thuyết Bài 27: Máy tính và Khoa học dữ liệu

**Lý thuyết Tin học 12 Bài 27: Máy tính và Khoa học dữ liệu**  
**1. Vai trò của máy tính đối với sự phát triển của khoa học dữ liệu**  
Quy trình Khoa học dữ liệu là chuỗi các bước để nghiên cứu, phân tích và khám phá tri thức từ dữ liệu. Quy trình này bao gồm:  
-Thu thập và tiền xử lý dữ liệu  
-Khám phá tri thức  
-Phân tích, đánh giá, triển khai và báo cáo kết quả  
Máy tính đóng vai trò thiết yếu trong quy trình này với các chức năng chính:  
-Xử lý và lưu trữ dữ liệu: Cung cấp công cụ để xử lý, lưu trữ và quản lý dữ liệu lớn và phức tạp.  
-Phân tích và khai phá dữ liệu: Thực hiện các thuật toán và mô hình học máy để phân tích dữ liệu và khám phá tri thức.  
-Trực quan hóa dữ liệu: Tạo các biểu đồ và đồ thị giúp trình bày phát hiện một cách dễ hiểu.  
-Tự động hóa: Hỗ trợ tự động hóa các tác vụ lặp lại, tăng tốc và giảm lỗi trong quy trình.  
- Xử lý song song: Dùng bộ xử lý đa lõi và hệ thống phân tán để giảm thời gian phân tích dữ liệu lớn.  
-Điện toán đám mây: Cung cấp tài nguyên tính toán qua nền tảng đám mây, giảm nhu cầu đầu tư vào phần cứng.  
- Hợp tác và truyền thông: Hỗ trợ làm việc nhóm, chia sẻ thông tin và cộng tác hiệu quả.  
=>Máy tính có vai trò không thể thiếu trong mọi giai đoạn của quy trình khoa học dữ liệu. Nó cung cấp sức mạnh tính toán, khả năng lưu trữ và khả năng tự động hoá cần thiết để xử lí, phân tích và khám phá tri thức từ dữ liệu, góp phần vào sự phát triển và thành công của Khoa học dữ liệu, mở ra cơ hội làm việc với dữ liệu lớn mà trước đây không thể thực hiện được. Điều này đã thúc đẩy sự phát triển của lĩnh vực Khoa học dữ liệu, giúp tạo lập giá trị và tri thức từ nguồn dữ liệu lớn phong phú và đa dạng.  
**2. Tính ưu việt trong việc sử dụng máy tính và thuật toán hiệu quả để xử lí dữ liệu lớn**  
Máy tính và các thuật toán hiệu quả giúp xử lý dữ liệu lớn một cách nhanh chóng, nhất quán và hiệu quả. Ví dụ cụ thể về Dự án Hệ gene người (Human Genome Project - HGP) minh họa điều này rõ ràng.  
Giải trình tự gene:  
- Hệ gene người là chuỗi 3 tỉ ký tự A, C, G, T mang thông tin di truyền.  
- Máy giải trình tự gene chỉ xác định được các đoạn nucleotide ngắn, cần ghép hàng triệu đoạn ngắn này thành một hệ gene hoàn chỉnh.  
- Quá trình này đòi hỏi máy tính mạnh và thuật toán chính xác.  
Dự án Hệ gene người (HGP):  
- Kích thước dữ liệu: Chuỗi hệ gene có độ dài khoảng 107,8 tỉ km, tạo ra hàng trăm gigabyte dữ liệu thô.  
- Lưu trữ dữ liệu: Dữ liệu HGP ước tính chiếm khoảng một trăm nghìn gigabyte.  
- Sức mạnh xử lý: Dự án sử dụng mạng lưới siêu máy tính toàn cầu, tương đương hàng nghìn máy tính xách tay hiện đại.  
Lợi ích và tầm quan trọng:  
- Tốc độ và hiệu quả: Máy tính và thuật toán giúp đẩy nhanh quá trình phân tích dữ liệu di truyền, hoàn thành dự án nhanh hơn nhiều so với phương pháp thủ công.  
- Độ chính xác: Quy trình tự động giảm nguy cơ sai sót và đảm bảo tính chính xác của trình tự bộ gene.  
- Xử lý và lưu trữ dữ liệu: Cơ sở hạ tầng tính toán cho phép quản lý và lưu trữ dữ liệu gene lớn.  
- Tích hợp dữ liệu: Máy tính và thuật toán tích hợp dữ liệu từ nhiều nhóm nghiên cứu, đảm bảo tính nhất quán và hiệu quả hợp tác.  
- Giải thích dữ liệu: Các thuật toán giúp giải thích thông tin di truyền, xác định gene và các vùng chức năng trong bộ gene.  
- Phân tích thời gian thực: Khả năng phân tích dữ liệu nhanh chóng, hỗ trợ đưa ra quyết định nhanh.  
- Xử lý song song: Kỹ thuật tính toán song song tăng tốc việc phân tích dữ liệu.  
- Khả năng mở rộng: Cơ sở hạ tầng tính toán được thiết kế để xử lý quy mô và độ phức tạp của dữ liệu gene, cần thiết cho các dự án khoa học quy mô lớn.  
Sự thành công của HGP đã cung cấp nhiều thông tin về gene người, thay đổi hiểu biết về di truyền học và dẫn tới nhiều tiến bộ y học. Điều này cho thấy tầm quan trọng của máy tính và các thuật toán trong nghiên cứu bộ gene và xử lý dữ liệu lớn.  
**Trắc nghiệm Tin học 12 Bài 27: Máy tính và Khoa học dữ liệu**  
**PHẦN I.** **Câu trắc nghiệm nhiều phương án lựa chọn. Thí sinh trả lời từ câu 1 đến câu 10. Mỗi câu hỏi thí sinh chỉ lựa chọn một phương án.**  
**Câu 1:** Máy tính đóng vai trò gì trong quá trình thu thập và tiền xử lý dữ liệu trong khoa học dữ liệu?  
A. Chỉ lưu trữ dữ liệu  
B. Chỉ trực quan hóa dữ liệu  
C. Cung cấp công cụ xử lý và lưu trữ dữ liệu lớn  
D. Tạo ra dữ liệu mới  
**Hiển thị đáp án**  
  
**Đáp án:** C  
**Giải thích:** Máy tính cung cấp các công cụ để xử lý và lưu trữ dữ liệu lớn, giúp quản lý và phân tích dữ liệu một cách hiệu quả.  
  
  
**Câu 2:** Vai trò của máy tính trong việc trực quan hóa dữ liệu là gì?  
A. Tự động tạo ra dữ liệu mới  
B. Tạo các biểu đồ và đồ thị  
C. Lưu trữ dữ liệu  
D. Xử lý và phân tích dữ liệu  
**Hiển thị đáp án**  
  
**Đáp án:** B  
**Giải thích:** Máy tính giúp trực quan hóa dữ liệu thông qua việc tạo ra các biểu đồ và đồ thị, giúp trình bày kết quả một cách dễ hiểu.  
  
  
**Câu 3:** Trong dự án Hệ gene người (HGP), máy tính đã hỗ trợ gì cho quá trình giải trình tự gene?  
A. Lưu trữ dữ liệu một cách an toàn  
B. Đảm bảo độ chính xác của trình tự bộ gene  
C. Tạo ra các mô hình dữ liệu mới  
D. Phân tích dữ liệu bằng tay  
**Hiển thị đáp án**  
  
**Đáp án:** B  
**Giải thích:** Máy tính và các thuật toán đã giúp giảm nguy cơ sai sót và đảm bảo độ chính xác trong việc giải trình tự bộ gene.  
  
  
**Câu 4:** Một trong những ưu điểm của việc sử dụng điện toán đám mây trong khoa học dữ liệu là gì?  
A. Tăng chi phí phần cứng  
B. Giảm khả năng hợp tác  
C. Cung cấp tài nguyên tính toán mà không cần đầu tư vào phần cứng  
D. Giảm hiệu suất làm việc  
**Hiển thị đáp án**  
  
**Đáp án:** C  
**Giải thích:** Điện toán đám mây cho phép truy cập vào tài nguyên tính toán mạnh mà không cần đầu tư nhiều vào phần cứng, giúp tiết kiệm chi phí.  
  
  
**Câu 5:** Kỹ thuật nào được sử dụng để giảm thời gian phân tích dữ liệu lớn?  
A. Xử lý tuần tự  
B. Xử lý song song  
C. Xử lý bằng tay  
D. Xử lý dữ liệu nhỏ  
**Hiển thị đáp án**  
  
**Đáp án:** B  
**Giải thích:** Kỹ thuật xử lý song song sử dụng nhiều bộ xử lý để xử lý dữ liệu đồng thời, giúp giảm thời gian phân tích dữ liệu lớn.  
  
  
**Câu 6:** Tính năng tự động hóa trong khoa học dữ liệu giúp đạt được điều gì?  
A. Tạo ra dữ liệu mới từ dữ liệu cũ  
B. Tăng tốc và giảm lỗi trong quy trình  
C. Loại bỏ sự cần thiết của con người trong quá trình phân tích  
D. Tăng độ phức tạp của các bài toán  
**Hiển thị đáp án**  
  
**Đáp án:** B  
**Giải thích:** Tự động hóa giúp tăng tốc quy trình và giảm thiểu lỗi do con người gây ra, đảm bảo tính nhất quán và hiệu quả.  
  
  
**Câu 7:** Máy tính hỗ trợ gì trong việc phân tích thời gian thực của dữ liệu?  
A. Chỉ lưu trữ dữ liệu mà không phân tích  
B. Phân tích dữ liệu với tốc độ chậm  
C. Hỗ trợ đưa ra quyết định nhanh chóng  
D. Tạo ra dữ liệu mới  
**Hiển thị đáp án**  
  
**Đáp án:** C  
**Giải thích:** Máy tính cho phép phân tích dữ liệu trong thời gian thực, giúp đưa ra các quyết định kịp thời và chính xác.  
  
  
**Câu 8:** Trong dự án HGP, vì sao cần sử dụng mạng lưới siêu máy tính?  
A. Để tiết kiệm chi phí  
B. Để tăng cường khả năng lưu trữ  
C. Để hoàn thành quá trình giải trình tự gene nhanh hơn  
D. Để giảm thiểu lỗi của con người  
**Hiển thị đáp án**  
  
**Đáp án:** C  
**Giải thích:** Mạng lưới siêu máy tính cung cấp sức mạnh tính toán lớn, giúp hoàn thành quá trình giải trình tự gene nhanh chóng hơn.  
  
  
**Câu 9:** Máy tính và các thuật toán có vai trò gì trong việc giải thích dữ liệu gene?  
A. Chỉ lưu trữ dữ liệu  
B. Chỉ trực quan hóa dữ liệu  
C. Xác định gene và các vùng chức năng trong bộ gene  
D. Tạo ra các dữ liệu gene mới  
**Hiển thị đáp án**  
  
**Đáp án:** C  
**Giải thích:** Các thuật toán giúp phân tích và giải thích thông tin di truyền, xác định các gene và vùng chức năng trong bộ gene.  
  
  
**Câu 10:** Lợi ích của việc tích hợp dữ liệu từ nhiều nhóm nghiên cứu trong dự án HGP là gì?  
A. Tăng chi phí lưu trữ  
B. Đảm bảo tính nhất quán và hiệu quả hợp tác  
C. Giảm khả năng phân tích dữ liệu  
D. Tăng khả năng lỗi trong dữ liệu  
**Hiển thị đáp án**  
  
**Đáp án:** B  
**Giải thích:** Việc tích hợp dữ liệu từ nhiều nhóm nghiên cứu giúp đảm bảo tính nhất quán của dữ liệu và nâng cao hiệu quả trong việc hợp tác nghiên cứu.  
  
  
**PHẦN II. Câu trắc nghiệm đúng sai.** **Thí sinh trả lời từ câu 1 đến câu 2. Trong mỗi ý a), b), c), d) ở mỗi câu, thí sinh chọn đúng hoặc sai**  
**Câu 1:** Phát biểu sau đúng hay sai về vai trò của máy tính trong khoa học dữ liệu ?  
a) Máy tính không cần thiết trong việc trực quan hóa dữ liệu.  
b) Máy tính hỗ trợ tự động hóa các tác vụ lặp lại, tăng tốc và giảm lỗi trong quy trình khoa học dữ liệu  
c) Máy tính không có khả năng xử lý song song dữ liệu lớn.  
d) Điện toán đám mây không cần thiết trong việc giảm nhu cầu đầu tư vào phần cứng.  
**Hiển thị đáp án**  
  
a) Sai: Máy tính rất quan trọng trong việc trực quan hóa dữ liệu, giúp tạo ra các biểu đồ và đồ thị để trình bày thông tin một cách dễ hiểu.  
b) Đúng: Máy tính giúp tự động hóa các tác vụ lặp lại, từ đó tăng tốc quá trình làm việc và giảm thiểu lỗi do con người gây ra.  
c) Sai: Máy tính có thể sử dụng bộ xử lý đa lõi và hệ thống phân tán để xử lý song song dữ liệu lớn, giúp giảm thời gian xử lý.  
d) Sai: Điện toán đám mây cho phép các tổ chức sử dụng tài nguyên tính toán mạnh mà không cần đầu tư vào phần cứng, từ đó giảm chi phí đầu tư.  
  
  
**Câu 2:** Tính ưu việt của máy tính và thuật toán trong xử lý dữ liệu lớn đúng hay sai?  
a) Máy tính và thuật toán có thể xử lý dữ liệu lớn một cách nhanh chóng và hiệu quả.  
b) Máy tính không cần thiết trong việc đảm bảo tính chính xác của trình tự bộ gene.  
c) Các thuật toán không thể tích hợp dữ liệu từ nhiều nhóm nghiên cứu khác nhau.  
d) Khả năng mở rộng của cơ sở hạ tầng tính toán không quan trọng trong các dự án khoa học quy mô lớn.  
**Hiển thị đáp án**  
  
a) Đúng: Máy tính và thuật toán giúp xử lý dữ liệu lớn nhanh chóng và hiệu quả, như được minh chứng qua Dự án Hệ gene người (HGP).  
b) Sai: Máy tính rất cần thiết để đảm bảo tính chính xác của trình tự bộ gene, nhờ vào khả năng xử lý tự động và giảm thiểu sai sót.  
c) Sai: Các thuật toán giúp tích hợp dữ liệu từ nhiều nhóm nghiên cứu, đảm bảo tính nhất quán và hiệu quả hợp tác.  
d) Sai: Khả năng mở rộng của cơ sở hạ tầng tính toán rất quan trọng trong các dự án khoa học quy mô lớn, giúp xử lý dữ liệu lớn và phức tạp một cách hiệu quả.  
  
  
**PHẦN III. Câu trả lời ngắn**. **Thí sinh trả lời từ câu 1 đến câu 3**  
**Câu 1:** Tại sao máy tính quan trọng trong quá trình phân tích dữ liệu trong khoa học dữ liệu?  
**Hiển thị đáp án**  
  
**Đáp án:** Máy tính quan trọng trong quá trình phân tích dữ liệu vì nó cung cấp khả năng tính toán mạnh mẽ để thực hiện các thuật toán phức tạp và mô hình học máy, đồng thời xử lý và lưu trữ khối lượng dữ liệu lớn.  
**Giải thích:** Máy tính giúp thực hiện các phân tích phức tạp và xử lý khối lượng dữ liệu lớn, điều này là thiết yếu trong quy trình phân tích dữ liệu để khám phá tri thức và đưa ra các kết luận chính xác.  
  
  
**Câu 2:** Những lợi ích chính của việc sử dụng điện toán đám mây trong khoa học dữ liệu là gì?  
**Hiển thị đáp án**  
  
**Đáp án:** Lợi ích chính của việc sử dụng điện toán đám mây bao gồm giảm nhu cầu đầu tư vào phần cứng, cung cấp tài nguyên tính toán linh hoạt và dễ dàng mở rộng khi cần thiết.  
**Giải thích:** Điện toán đám mây giúp giảm chi phí đầu tư vào phần cứng vì tài nguyên tính toán được cung cấp qua nền tảng đám mây. Ngoài ra, nó cung cấp khả năng mở rộng linh hoạt, cho phép điều chỉnh tài nguyên theo nhu cầu của dự án.  
  
  
**Câu 3:** Dự án Hệ gene người (HGP) đã minh họa tính ưu việt của máy tính và thuật toán trong xử lý dữ liệu lớn như thế nào?  
**Hiển thị đáp án**  
  
**Đáp án:** Dự án HGP minh họa tính ưu việt của máy tính và thuật toán qua việc sử dụng mạng lưới siêu máy tính để xử lý hàng trăm gigabyte dữ liệu gene nhanh chóng và chính xác, điều này không thể thực hiện được bằng phương pháp thủ công.  
**Giải thích:** Dự án HGP sử dụng máy tính và thuật toán để xử lý dữ liệu gene khổng lồ một cách nhanh chóng và chính xác, cho thấy sự cần thiết của công nghệ tính toán trong việc giải quyết các vấn đề dữ liệu lớn mà phương pháp thủ công không thể xử lý hiệu quả.  
  
  
Xem thêm